https://doi.org/10.51408/issi2025_209

How Does Knowledge Source Novelty Influence Knowledge Output Novelty? Evidence from 269,569 PLOS Articles

Yi Xiang[1], Chengzhi Zhang[2]

[1] xiangyi@njust.edu.cn, [2] zhangcz@njust.edu.cn
Department of Information Management, Nanjing University of Science and Technology, 210094, Nanjing (China)

Introduction

Novelty is a key criterion in evaluating the innovativeness of academic research. As academic literature expands rapidly, effectively measuring novelty has become a critical research focus. Existing methods for assessing the novelty of academic papers can be classified into two categories based on the knowledge components they employ: (1) citation-based methods and (2) word-level knowledge unit methods, such as MeSH terms and entities. Citation-based approaches capture novelty in knowledge sources, while entity-based approaches focus on research content. However, prior studies often examine these methods in isolation, neglecting their interconnections. Investigating their relationship can deepen our understanding of measurement discrepancies and correlations, providing a theoretical basis for integrating multiple novelty dimensions to improve accuracy.

Addressing the limitations of existing research, this study models academic paper writing as a production process. We then apply the Cobb-Douglas production function (Douglas, 1928), commonly used in economics to model the relationship between input and output, to examine the relationship between the novelty of knowledge sources and the novelty of knowledge output in academic papers.

Methodology

Dataset

We collected 362,269 papers published between 2003 and September 2024 from the PLOS database. After extracting reference records and analysing corresponding journals, we excluded papers with missing reference lists, resulting in a final dataset of 330,966 papers. We then retrieved MeSH term lists from OpenAlex[1] and excluded records with missing data, yielding 269,569 papers. As MeSH terms pertain to biomedical fields, this filtering indicates that the study focuses primarily on biomedical literature. A basic statistical analysis of the dataset revealed that, on average, each paper cites 23 different journals and contains 17 MeSH terms.

Novelty Measurement

We propose a graph representation learning approach to measure novelty, based on combinatorial innovation theory (Uzzi et al., 2013). For papers published in year Y, we first compile prior papers, extracting knowledge components (reference journals or MeSH terms) as network nodes, with edges linking co-occurring units. We then apply the LINE algorithm (Tang et al., 2015) to generate vector representation of nodes.

Given a focal paper with N knowledge units, each represented by a vector V_i, we construct all possible knowledge unit combinations. The novelty of each combination $Comb_{i,j}$ is then quantified using the following formula:

$$Novelty_{i,j}^{comb} = 1 - \frac{|V_i||V_j|}{V_i \cdot V_j} \#(1)$$

The overall novelty of the paper is the sum of the novelty scores for all combinations. Since this study considers two types of knowledge units—references and MeSH terms—we distinguish between them by denoting reference-based novelty as $Novel_J$ and MeSH-based novelty as $Novel_M$.

[1] https://openalex.org/

Cobb-Douglas production function

The Cobb-Douglas function is widely used in economics to model the relationship between inputs (e.g., capital and labor) and outputs in production activities. The writing of academic papers can also be viewed as a production process, where scholars accumulate raw experience by reading references, invest time and effort to validate research ideas, and ultimately produce research papers. Therefore, we consider the novelty of references $Novel_J$ as capital input, and the number of authors (L) as labor input. The novelty based on MeSH terms $Novel_M$ serves as the output. We then model the relationship between these variables using the transcendental logarithmic model (Christensen et al., 1973), an extension of the Cobb-Douglas function that accounts for interactions between input factors:

$$\begin{aligned} ln\ ln\ Novel_M = & ln\ ln\ A + \alpha\ ln\ ln\ Novel_J \\ & + \beta\ ln\ ln\ L \\ & + \gamma\left(ln\ ln\ Novel_J \cdot ln\ ln\ L\right) + \varepsilon \#(2) \end{aligned}$$

Where, A and ε represent the intercept and the error term, respectively.

Result

Descriptive Statistics

Figure 1 illustrates the distribution of paper novelty calculated using two methods: MeSH term-based and reference-based novelty. Both methods reveal a clear right-skewed distribution, indicating that the majority of papers exhibit low novelty, while only a small proportion are classified as highly novel.

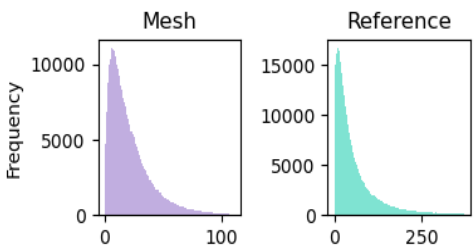

Figure 1. The distribution of papers' novelty.

Analysis of the relationship between two types of novelty

Table 1 presents the regression results based on Equation (2). First, regarding the novelty of knowledge sources, when the number of authors is held constant, each unit increase in $Novel_J$ is associated with an average increase of 0.0359 in the novelty of the output $Novel_M$. The impact of the number of authors (L) on output novelty is even more pronounced. For each additional author, $Novel_M$ is expected to increase by 0.5133, consistent with prior research. We also examined the quadratic term for the number of authors and found its coefficient to be negative, suggesting that beyond a certain threshold, additional authors may diminish output novelty. Furthermore, the interaction term between $Novel_J$ and L has a negative effect on the dependent variable, indicating that the influence of knowledge source novelty and the number of authors on output novelty may counteract each other.

These findings demonstrate that the novelty of knowledge sources positively influences the novelty of a paper's content. However, the number of authors also plays a crucial role in knowledge flow. The negative interaction between $Novel_J$ and L suggests that while an increase in the number of authors may introduce diverse perspectives and enhance novelty, excessive collaboration can lead to higher coordination costs. Additionally, researchers from different backgrounds may have varying perceptions of novelty, potentially hindering the effective translation of knowledge source novelty into novel research output.

Table 1. The regression results.

	(1)	(2)
$Novel_J$	0.0359***	0.0279***
	(0.004)	(0.004)
L	0.5133***	0.7944***
	(0.008)	(0.014)
L^2		-0.0871***
		(0.004)
$Novel_J * L$	-0.0275***	-0.0241***
	(0.002)	(0.002)
Constant	1.8048***	1.6166***
	(0.015)	(0.017)
Observations	269,569	269,569
Pseudo R^2	0.061	0.063

Note: ***: p<0.001.

Conclusion

This study examines the relationship between the novelty of knowledge sources (references) and outputs (MeSH terms) in academic papers. We propose a graph representation learning method to measure novelty and use the Cobb-Douglas function to model idea transformation as a production process. Findings reveal that source novelty significantly impacts output novelty, advancing our understanding of knowledge flow. However, factors such as team diversity and funding may influence this relationship. Future research should explore these variables and assess the findings' generalizability across disciplines.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No.72074113).

References

Christensen, L. R., Jorgenson, D. W., & Lau, L. J. (1973). Transcendental logarithmic production frontiers. The review of economics and statistics, 28-45.

Douglas, P. (1928). Cobb douglas production function. The Quarterly Journal of Economics, 42(3), 393-415.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. Science, 342(6157), 468-472.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015, May). Line: Large-scale information network embedding. In Proceedings of the 24th international conference on world wide web (pp. 1067-1077).