How to Measure Information Cocoon in Academic Environment

Jia Yuan[1], Guoxiu He[1] and Yunhan Yang[2,*]

[1]*School of Economics and Management, East China Normal University, Shanghai, China*
[2]*Faculty of Education, The University of Hong Kong, Hong Kong, SAR, China*

Abstract

When individuals face an abundance of information, they often selectively choose data that reinforces their existing beliefs, ignoring opposing views and creating an 'information cocoon'. This phenomenon is not limited to social media; it is also relevant in academic circles. This study introduces a novel method for measuring information cocoons in academia from two main perspectives: depth and breadth. We analyze a broad dataset of academic papers, employing BERTopic for topic modeling and Sentence-BERT for semantic similarity. The results of the study show that the degree of information cocoon in the overall citation network is on a decreasing trend, and the information exchange in academia is gradually open and innovative. Secondly, there are significant differences in the information cocoon between disciplines, and disciplines with different cocoon sizes have their own characteristics, whose uniqueness and complexity need to be taken into full consideration in the assessment. In addition, the study also found that there is a non-linear pattern between the number of citations of scholarly literature and its information cocoon performance. These results stress the need to understand and address information cocoon dynamics in academia, promoting strategies for a more inclusive and diverse scholarly collaborations.

Keywords

information cocoon, academic environment, research depth, research breadth

1. Introduction

In the era of big data, the explosive growth and overload of information have led to increased network dependence, fragmentation, and selective exposure in people's information behavior [1]. In information dissemination, the public only pays attention to what they choose and the field that makes them happy. Over time, they will confine themselves to a cocoon like *cocoon room* [2]. When people in a positive feedback loop, they are mainly exposed to content they have already agreed with, which affects the diversity of information acceptance [3].

The concept of the information cocoon applies to any environment where information is generated, including academia. Within this system, scholars' interaction with information can lead to the formation of an information cocoon. This manifests when researchers excessively consume similar information over time, resulting in issues like information narrowing, group polarization, reduced innovation, and research bottlenecks. This prompts questions: How prevalent is the information cocoon in academia? How can it be measured? And what variations exist among different groups?

Previous studies have extensively examined the formation, impact, and ways to break out of information cocoons. However, there has been limited systematic research on measuring information cocoons, especially within academic environments. Furthermore, most studies have focused on social media platforms, with few addressing academic settings.

Motivated by these research gaps, we aim to propose a method for measuring the information cocoon within academic environments. Our objective is to quantify the extent to which scholars are constrained by homogeneous information, analyze variations in cocoons among different groups. To caution academic researchers against the risks associated with information cocoons, and to actively broaden the focus of academic work and explore research innovation, thereby enhancing awareness of optimizing the information environment within academia and emphasizing the improvement of research competence.

To achieve a scientific measurement, we decompose the factors influencing the cocoon into two primary dimensions: depth and breadth. We employ a topic modeling approach based on BERTopic and Sentence-BERT for similarity calculation. This enables us to assess the relationship between articles and themselves or their references.

Our analysis reveals significant disparities in the value of information cocoon among various groups. Through thorough measurement and analysis, we offer comprehensive and practical insights, providing a clearer visualization of the extent of the information cocoon. This serves as a reminder to scholars to reflect on their own perspectives and actively work to avoid falling into the trap of information cocoon.

2. Theoretical Foundation

2.1. Metrics

In the academic career of scholars, continuous horizontal and vertical development is crucial. Horizontal development refers to the ability to engage with a wide range of different fields and topics, while vertical development emphasizes in-depth exploration of specific fields or subjects. These two modes of development complement each other and mutually reinforce. If a scholar focuses solely on horizontal development, they may have a superficial understanding and lack specialized knowledge in various fields. Conversely, pursuing only vertical development may result in deep research in a particular area but fails to integrate knowledge from different fields, thus limiting the breadth of research.

Scholars need to balance these two modes of development throughout their careers. They should maintain in-depth research in specific fields while also maintaining a broad understanding of other areas. Such comprehensive development aids in enhancing scholars' problem-solving abilities for complex issues, fostering innovation, and advancing academic research. Moreover, the depth and breadth of

Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024) and the 4th AI + Informetrics (AII2024), April 22-26, 2024, Changchun, China and Online

[*]Corresponding author.

yunhany@connect.hku.hk (Y. Yang)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

research directly impact scholars' research outcomes. Indepth research enables scholars to grasp the essence and intrinsic mechanisms of problems, providing deeper insights and analysis. On the other hand, breadth of research helps scholars acquire diverse information and perspectives from different angles and fields, thereby offering comprehensive and diversified research results.

Scholars with a high degree of academic cocoon may exhibit one of the following characteristics: prolonged focus on a single topic without achieving breakthrough innovation, leading to academic stagnation; or an excessive pursuit of research breadth, spanning multiple fields but struggling to generate valuable research outcomes.

To this end, evaluating the extent of information cocoons requires a comprehensive consideration of both depth and breadth. Only by simultaneously addressing these two dimensions can researchers better transcend the constraints imposed by information cocoons. Conversely, focusing solely on one dimension or conducting superficial analyses may lead to limitations and misconceptions regarding information. Thus, this paper is grounded in this rationale to devise methodologies and propose corresponding metrics.

2.2. Pretrained Language Model

2.2.1. Sentence-BERT

Sentence BERT is a modification of the pre-trained BERT network that use siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity [6]. In particular, Sentence-BERT is derived from a deep learning model called Bidirectional Encoder Representations from Transformers (BERT) which has revolutionized NLP applications over the last couple of years by capturing the meaning of a sentence at an unprecedented quality [7, 8].

Therefore, this study employs Sentence-BERT to extract sentence features from document titles, facilitating similarity calculation for the assessment of information correlation among documents. This approach enables a more precise measurement of document relevance, thereby furnishing a more dependable foundation for subsequent analyses.

2.2.2. BERTopic

BERTopic is a topic model, which uses a pre-trained transformer based language model to generate document embeddings, clusters these embeddings, and finally uses a class based TF-IDF process to generate topic representations. BERTopic can generate coherent topics, involving classic models and maintaining competitiveness in subject modeling [9].

Utilizing BERTopic technology enables us to identify the themes covered in scholarly literature titles with greater precision. Determining the types, quantity, and probability distribution of these themes facilitates the effective assessment of the breadth of research within the literature.

3. Methodology

3.1. Data

This study collected data from the Semantic Scholar Open Research Corpus (S2ORC), which covers 81.1 million academic papers from multiple disciplines. The corpus contains rich metadata, abstracts, parsed references, and structured full text for 8.1 million open access users. In S2ORC, it aggregates users from hundreds of academic publishers and digital archives into a unified source, creating the largest publicly available collection of machine-readable academic texts to date [4].

We extracted papers published between 2010 and 2021 from S2ORC, capturing data such as article titles, first authors, reference titles, publication dates, and citation counts. After eliminating duplicates and incomplete entries, we further refined our sample by excluding authors with sporadic publication patterns, focusing on those who averaged more than 2 articles per year. Our final dataset comprises 107,775 articles.

3.2. Measures

3.2.1. Research Depth

Citing literature is crucial in academic research, serving to honor past work and foster innovation. Authors must skillfully reference previous research while providing fresh insights. This study introduces the concept of "citation depth," indicating the substantive variance between literature and cited sources, thus highlighting innovation. We employ the Sentence-BERT model to quantify this variance by assessing title similarities between papers and citations. This method enhances our understanding of the paper-citation relationship and deepens insights into research content and depth. Through this approach, we gain a more comprehensive understanding of academic research and explore its inherent value. $R\left(a, b_i\right)$ represents the similarity between the paper and each reference, while n denotes the total number of references to this paper. The calculation formula is:

$$\text{Ref\_depth } = 1-\frac{\sum_{i=1}^n R\left(a, b_i\right)}{n} \quad (1)$$

The depth of research is evident in the evolving trajectory and intensity of individual scholars' pursuits. Each presentation of research findings represents a journey of continuous self-challenge and breakthrough. Put simply, scholars who achieve breakthroughs in research depth often showcase distinct differences from prior research. Such differentiation may manifest in the exploration of new topics or the attainment of fresh insights within the same problem domain. To more precisely gauge this depth of inquiry, this paper employs Sentence-BERT to quantify the divergence of each piece of literature authored by the same individual from their previous research. The literature dataset is organized by author and arranged chronologically in reverse order of publication. By calculating the similarity between each piece of literature and its first three publications, the initial trio of works by each author in the dataset is excluded from the statistical analysis.$\operatorname{AvgSim}(i)$ is the similarity of two designated papers of the same author. The calculation formula is as follows:

$$\operatorname{AvgSim}(i)= \begin{cases}\frac{1}{3} \sum_{j-1}^3 R\left(a_i, b_{i+j}\right), & \text{if } i+3 \leq n \\ 0, & \text{otherwise}\end{cases} \quad (2)$$

$$\text{self\_depth } = 1-\operatorname{AvgSim}(i) \quad (3)$$

3.2.2. Research Breadth

The number of topics covered in references is a key indicator of the breadth of literature research. We consider all reference titles collected within this time frame as a complete dataset. By applying the BERTopic model to classify each reference title into topics, we obtain detailed topic information. Throughout this process, we record the frequency of each topic type appearing in the references, referred to as "ref_topic_counts." This approach not only allows us to comprehensively understand the distribution of topics within the research field but also provides a reliable data foundation for further analysis.

$$\text{Ref\_breadth} = \frac{\text{ref\_topic\_counts}}{10} \quad (4)$$

Research topics chosen on a personal level can embody the diversity of research interests. Utilizing BERTopic modeling, each paper is assigned probabilities for being grouped into various topics. In this study, the Gini coefficient is employed to quantify the breadth of one's research interests. The Gini coefficient is commonly utilized to gauge the level of inequality within a dataset or distribution. A higher coefficient indicates a more uneven distribution of probabilities among literature topics, leaning towards a singular topic and indicating narrower breadth. Conversely, a lower coefficient signifies a more evenly distributed probability of theme allocation, suggesting a broader range of diverse themes.

The calculation formula is as follows:

$$Gini = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}} \quad (5)$$

$$\text{Self\_breadth} = 1 - \text{Gini} \quad (6)$$

where n is the number of observations, X_i represents the i-th observation, and $\bar{x}$ is the average of all observations.The distribution of topics within each article is quantified through the calculation of the Gini coefficient, representing the probability set.

3.3. Cocoon Value

The four metrics presented in the previous section aim to break the information cocoon. In this regard, a decrease in the depth and breadth values indicates that the literature is limited by a single piece of information, thus leading to an increase in the cocoon value. Conversely, an increase in the depth and breadth values indicates that the literature has the potential to break out of the information cocoon. Therefore, the expression for cocoon value is as follows.M_i is the four indicators calculated above.

$$\text{Cocoon} = \text{Avg}\{(1 - \text{M}_\text{i})\} \quad (7)$$

4. Results and Analysis

4.1. The Whole Academic Environment

After measuring the depth and breadth values, this paper conducted a comprehensive analysis of the overall data and classified it according to the publication year. It is observed that the changes in the two depth indicators are relatively stable; however, there is a noticeable increase in the “ref_breath" indicator. Meanwhile, the overall cocoon value shows a decreasing trend year by year, indicating a continuous opening up and innovation of information in the academic environment, which is a positive phenomenon. The specific results are displayed in Figures 1 and 2.

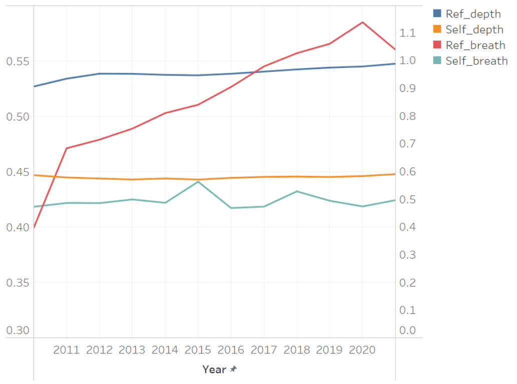

Figure 1: The trend of changes in the depth and breadth of the overall academic environment.

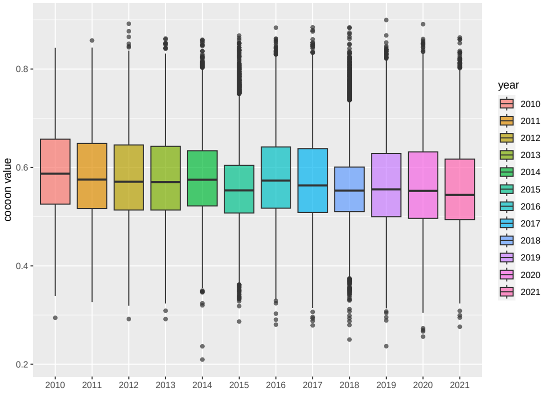

Figure 2: The trend of information cocoon value in the overall academic environment.

4.2. Different disciplines

After analyzing the results shown in Figure 3 and Figure 4, it is obvious that disciplines with smaller information cocoons tend to show a higher level of interdisciplinary and openness, such as computer science and art. On the contrary, traditional disciplines tend to have large information cocoons. In addition, some disciplines can break through the information cocoon room to a certain extent through their extensive research fields. Considering the unique complexity of each discipline, in-depth analysis is essential. When dealing with the challenge of information cocoon, scholars should focus on disciplines with innovative advantages. We should carry out self reform and optimization, strengthen exchanges, and integration among disciplines, and open up a new academic track.

4.3. Citations Classification

The results are depicted in Figures 5 and 6. We grouped disciplines based on their co-citation values and selected those with the highest, lowest, and near-average co-citation values for analysis. These disciplines were then categorized into four groups based on their citation frequency: Group A, with more than 300 citations; Group B, with between 100 and 300 citations; Group C, with between 10 and 100 citations; and Group D, with less than 10 citations.

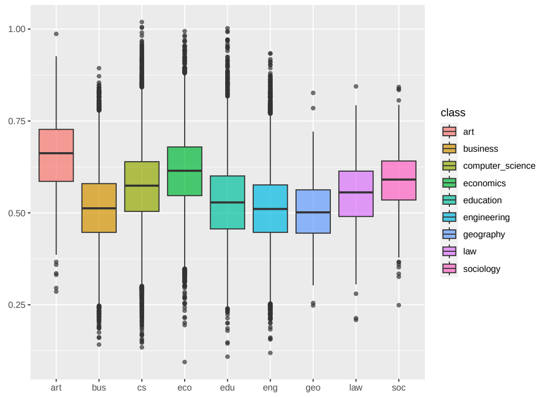

Figure 3: Information cocoon value of different disciplines.

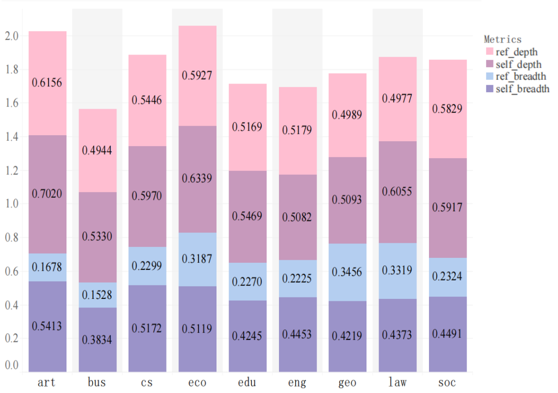

Figure 4: Depth and breadth of different disciplines.

The analysis found that the literature in groups A, B, C, and D showed similar trends in the performance of metric values in the three disciplines, and there was a pattern between the number of citations and the degree of information cocooning. The most cited literature usually has higher research depth and breadth, indicating that it explores a specific area in depth and covers related areas extensively. Highly cited literature in group B focuses on academic hotspots and attracts scholars with great breadth, but the depth may be insufficient. Whereas the less-cited literature is limited in research breadth but high in depth, it may be difficult to be accepted because it explores a niche issue or has a high degree of depth. In conclusion, the number of citations is not the only goal of scientific research; extensive research can get a certain number of citations, but research with both depth and breadth is more influential. Fewer citations do not mean fewer results, but the research may be too in-depth or niche, and still has some potential for development. Therefore, scholars should not only focus on the number of citations, but also on the quality of research. In research evaluation, we should consider the depth, breadth and actual impact of research rather than a single indicator.

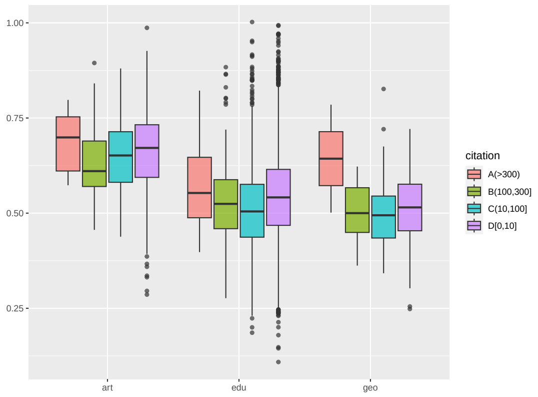

Figure 5: The depth value of different citations.

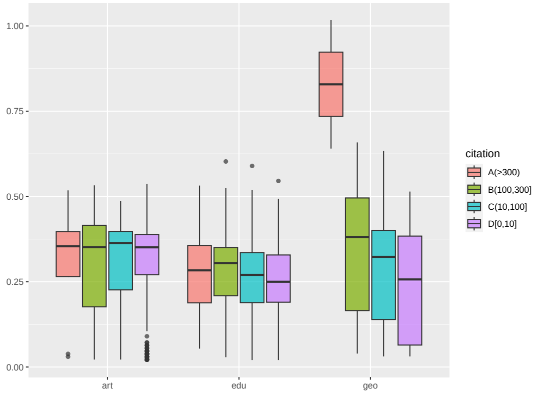

Figure 6: The depth and breadth value of different citations.

5. Discussion

In our study, we propose an index and methodology for measuring the scale of information cocoon within academic environments and classify them accordingly. The main findings of this paper can be summarized into three points. Firstly, we observe a gradual breakdown of information cocoon within the overall academic environment, presenting a trend towards greater comprehensiveness and innovation. Secondly, significant disparities exist in terms of depth, breadth, and cocooning across traditional, creative, and technological disciplines. Lastly, we select three representative disciplines and divided them based on citation frequency. Results indicate variations in breadth and depth among different citation groups, with literature possessing greater breadth and lesser depth often garnering wider acceptance. This suggests that scholars may not necessarily prioritize citation frequency to overcome cocooning constraints. Therefore, researchers should adeptly utilize extensive and intricate academic information, continually assessing whether their research processes are constrained by information cocoon.

References

[1] Yuan, X., and Wang, C. (2022). Research on the Formation Mechanism of Information Cocoon and Individual Differences among Researchers Based on

Information Ecology Theory, Frontiers in Psychology (13).

[2] Sanstan, (2008).Information Utopia: How People Produce Knowledg", Translated by Bi Jingyue Beijing: Law Publishing House(8).

[3] Falck, A., and Boyer, K. 2022. Online Filters and Social Trust: Why We Should Still Be Concerned about Filter Bubbles.

[4] Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The Semantic Scholar Open Research Corpus, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (eds.), Online: Association for Computational Linguistics, July, pp. 4969–4983.

[5] Loet, L., Caroline S., W., & Lutz, B. (2019) Interdisciplinarity as Diversity in Citation Patterns among Journals: Rao-Stirling Diversity, Relative Variety, and the Gini coefficient., arXiv: Digital Libraries, 13.1: 255-269.

[6] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Conference on Empirical Methods in Natural Language Processing.

[7] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.

[8] Rath, S., & Chow, J.Y. (2022). Worldwide city transport typology prediction with sentence-BERT based supervised learning via Wikipedia. ArXiv, abs/2204.05193.

[9] Grootendorst, M.R. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. ArXiv, abs/2203.05794.