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 utilised two models, BERTopic and Sentence-BERT, to help quantify the depth and breadth of the study. 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 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].

Any environment that generates information is likely to have an information cocoon, 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[3][4]. 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[5][6].

Motivated by the existing research gaps, our primary objective is to propose a comprehensive method for measuring the information cocoon within academic environments. Specifically, we aim to quantify the evolutionary changes in the value of information cocoon within academia. To achieve this, we decompose the information cocoon into two key components: research depth and research breadth. In order to accurately quantify these aspects, we utilize BERTopic and sentence-BERT techniques. Furthermore, we intend to analyze the variations in information cocoons across different groups, encompassing various disciplines and citation levels.

Our analysis uncovers a downward trend in the value of the information cocoon, accompanied by disparities among different groups. These findings provide comprehensive and practical insights into the phenomenon of information cocoon within academia. It serves as a timely reminder for scholars to critically examine their perspectives and take proactive measures to avoid succumbing to the pitfalls of an information cocoon. By doing so, scholars can effectively optimize the information environment within academia for enhanced research outcomes.

2. Theoretical Foundation

2.1. Information cocooning in an academic context

In the academic career of scholars, it is crucial to balance continuous horizontal and vertical development. Horizontal development allows scholars to cover a wide range of fields, while vertical development allows them to conduct in-depth research in specific areas. Focusing only on horizontal development may lead to a superficial understanding of fields and a lack of expertise, while pursuing only vertical development may limit the breadth of knowledge. Therefore, scholars need to maintain in-depth study of specific fields throughout their careers, while gaining a broad understanding of other fields, to enhance their ability to solve complex problems, foster a spirit of innovation, and promote the all-round development of academic research. Such a balance not only captures the essence of the problem and provides insights, but also integrates knowledge from different fields to produce comprehensive and diverse results for academic research[7].

Scholars with larger information cocoons tend to perform poorly in terms of depth or breadth of research, as evidenced by their inability to break through to innovation in a particular research direction, or limitations in their research areas.

To this end, evaluating the extent of information cocoons requires a comprehensive consideration of both depth and

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)

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 modified version of the pre-trained BERT network that incorporates siamese and triplet network structures. By leveraging these structures, Sentence-BERT generates semantically meaningful sentence embeddings that can be compared using cosine-similarity [8]. In recent years, Sentence-BERT has brought about a significant transformation in NLP applications by capturing sentence meaning with unprecedented accuracy [9, 10]. Building upon this advancement, our study utilizes Sentence-BERT to extract valuable sentence features from document titles. This facilitates the calculation of similarity, enabling a comprehensive assessment of information correlation among documents. By employing this approach, we achieve a more precise measurement of document relevance, providing a robust foundation for subsequent analyses.

2.2.2. BERTopic

BERTopic is a topic modeling technique that leverages a pre-trained transformer-based language model to generate document embeddings. These embeddings are then clustered, and a class-based TF-IDF process is employed to generate topic representations[11]. BERTopic has demonstrated its ability to generate coherent topics, incorporating both traditional models and retaining competitiveness in subject modeling. By harnessing the power of BERTopic, we can accurately identify the themes addressed in scholarly literature titles, enabling a more precise understanding of the research scope. This allows us to assess the distribution of these themes effectively, thereby facilitating an in-depth evaluation of the research breadth in our study.

3. Methodology

3.1. Data

For data collection, this study utilized the Semantic Scholar Open Research Corpus (S2ORC), an extensive dataset comprising 81.1 million academic papers from diverse disciplines. The corpus includes comprehensive metadata, abstracts, and parsed references. S2ORC serves as a centralized repository that aggregates papers from hundreds of academic publishers and digital archives, resulting in the largest publicly available collection of machine-readable academic text to date [12].

From the S2ORC dataset, we extracted papers published within the timeframe of 2010 to 2021. The data collection process encompassed capturing various information, including article titles, first authors, reference titles, publication dates, and citation counts. To ensure a sufficient level of academic expertise among scholars, we removed duplicate and incomplete entries. Additionally, first authors with fewer than six publications during the specific period were excluded. As a result of this rigorous selection process, our final dataset consists of 107,775 articles.

3.2. Measures

3.2.1. Research Depth

To measure the research depth, we developed two distinct metrics named "re_depth" and "self_depth". The "re_depth" metric quantifies the research depth by assessing the disparity between the target paper and its reference list. On the other hand, the "self_depth" metric quantifies the research depth by evaluating the distinction between the target paper and previous studies published by the same author.

In the first place, citing relevant literature is of utmost importance for authors, as it allows them to build upon existing knowledge and propose new insights. The disparities in knowledge between their research and the sources they cite indicate the level of innovation within their study. This concept of aggregated knowledge at the topic or field level allows us to observe the macro-scale evolution of knowledge, emphasizing the critical nature of citing behavior. We believe that a greater difference between the target literature and the cited sources reflects a higher level of innovation in the target paper, indicating a deeper level of research. Therefore, we utilize the variance between the target publication and the cited sources as an indication of the research depth within the target paper. To quantify this variance, we utilized the Sentence-BERT model to assess title similarities between a paper and its reference list. The calculation formula is:

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

Here, $R\left(p, r_{i}\right)$ represents the similarity between the paper and each reference, while n denotes the total number of references to this paper. p refers to an article, r_i refers to the i_th reference of p.

Furthermore, the depth of research becomes evident through the evolving trajectory and intensity of individual scholars' pursuits. Each presentation of research findings signifies a continuous journey of self-challenge and breakthrough. Scholars who achieve breakthroughs in research depth often showcase distinctions from prior research. These distinctions can manifest in the exploration of new topics or the acquisition of fresh insights within the same problem domain. To precisely evaluate this depth of inquiry, Sentence-BERT was employed in this study to quantify the divergence of each publication authored by the same individual from their prior research. The dataset was organized accordingly, categorized by author, and arranged chronologically in reverse order of publication. Subsequently, the similarity of each paper to the three papers preceding its publication time was calculated. The calculation formula is as follows:

$$\text{Self\_depth } = \left\{ \begin{array}{c} 1-\frac{\sum_{j=i+1}^{i+3} R\left(p_{i}, p_{j}\right)}{3}, i+3 \leq n \\ 0, i+3>n \end{array} \right. \quad (2)$$

In this context, $R\left(p_{i}, p_{j}\right)$ represents the similarity between two specific papers authored by the same individual. The variables p_i and p_j correspond to distinct documents authored by the same individual.

3.2.2. Research Breadth

To assess the research breadth, we introduced two distinct metrics: "ref_breadth" and "self_breadth". The "ref_breadth" metric quantifies the research breadth by examining the number of topics within the references of the target paper. Conversely, the "self_breadth" metric quantifies the research breadth by evaluating the diversity of topics addressed within the target paper.

Initially, we hypothesized that the number of topics covered by the references serves as an indicator of the research breadth within the literature[13]. To capture this valuable information, we collected the reference titles associated with each paper and employed the BERTopic model to classify each reference title into specific topics. Consequently, we recorded the number of topics for each group of references as "ref_topic_counts". The formula for calculating the "ref_breadth" metric is as follows:

$$\text{Ref\_breadth} = \frac{\text{ref\_topic\_counts}}{10} \tag{3}$$

By dividing the ref_topic_counts by 10, we harmonized this value with the scale of other indicators utilized in this study.

Furthermore, the "ref_breadth" metric offers insights into whether scholars have explored diverse fields of knowledge throughout their research endeavors. Through the utilization of BERTopic modeling, each paper is assigned probabilities for belonging to various topic groups. In this study, the Gini coefficient is employed to quantify the breadth of research interests. The Gini coefficient is a widely used measure to assess the level of inequality within a dataset or distribution[14]. A higher coefficient indicates a more uneven distribution of probabilities among literature topics, implying a focus on a singular topic and suggesting a 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 for the Gini coefficient is as follows:

$$Gini = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}} \tag{4}$$

Subsequently, the "self_breadth" metric is derived as follows:

$$\text{Self\_breadth} = 1 - \text{Gini} \tag{5}$$

In the formulas, n represents the number of papers, X_i denotes the i-th paper, and $\bar{x}$ represents the average value across all papers.

3.2.3. Cocoon Value

In accordance with our definition of the information cocoon within an academic context, a reduction in both research depth and breadth indicates that an article is confined to a singular aspect of information, thereby increasing the likelihood of information cocooning. Conversely, an expansion in both research depth and breadth implies that an article holds the potential to transcend the information cocoon. Therefore, the expression for the cocoon value is as follows: M_i represents the sum of the four aforementioned indicators.

$$\text{Cocoon} = \text{Avg}\{(1 - \text{M}_\text{i})\} \tag{6}$$

4. Results and Analysis

4.1. The Evolved Information Cocoon in Academic Context

Our initial objective is to examine the phenomenon of information cocooning within the entire academic environment over time. To achieve this, we calculated the research depth and breadth values on an annual basis, followed by computing the cocoon value for each year. The corresponding findings are illustrated in Figures 1 and 2. Figure 1 demonstrates that the changes in two depth indicators (represented by the blue and orange lines) remain relatively stable, whereas there is a noticeable increase in the “ref_breath" indicator (depicted by the red line). Furthermore, Figure 2 presents the overall cocoon value, which exhibits a decreasing trend over the years. This trend signifies a continuous opening up and innovation of information in the academic environment, reflecting a positive phenomenon.

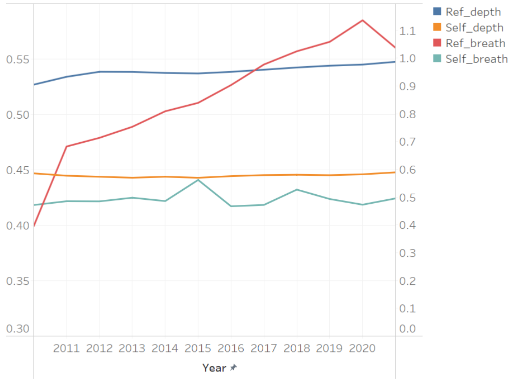

Figure 1: Evolution of Research Depth and Breadth in the Dataset Over Time. The right vertical axis represents the range of values for "ref_breadth", while the left vertical axis corresponds to the remaining indicators

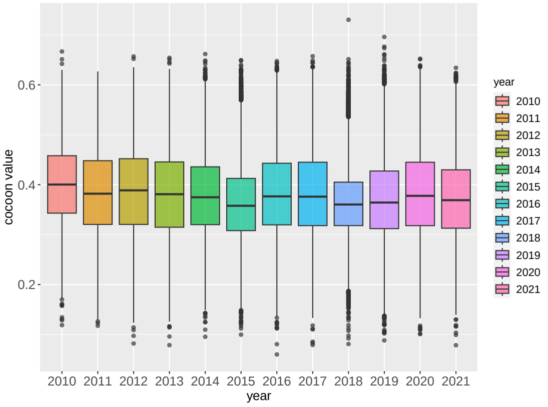

Figure 2: Evolution of Information Cocoon Value in the Dataset Over Years.

4.2. Information Cocoon at Different Disciplines

Subsequently, we conducted an investigation into the variability of information cocooning across different research fields and presented our findings in Figures 3 and 4. To ensure the reliability of our results, we excluded disciplines with limited data and focused on disciplines with larger volumes for analysis. Our analysis reveals notable trends within specific disciplines. In Figure 4, art, economics, and computer science exhibit the lowest levels of information cocooning. This is evident from their higher values in research depth and self_breadth indicators, as depicted by the pink, shallow purple, and dark purple bars in Figure 3. On the other hand, geography, business, and engineering tend to have larger information cocoons, as indicated by their lower research depth and breadth values in Figure 3. Furthermore, disciplines such as education, law, and sociology demonstrate the ability to partially break through the information cocoon, thanks to their relatively higher values on one or more of the four metrics. For example, the discipline of law exhibits a higher "ref_breadth" value, albeit with a lower "ref_depth" value.

These findings emphasize the importance of considering the uniqueness and complexity of each discipline when developing strategies or policies to overcome the information cocoon at the field level. Scholars within each field should also take into account the specific characteristics of their discipline's information cocoon when designing their research studies.

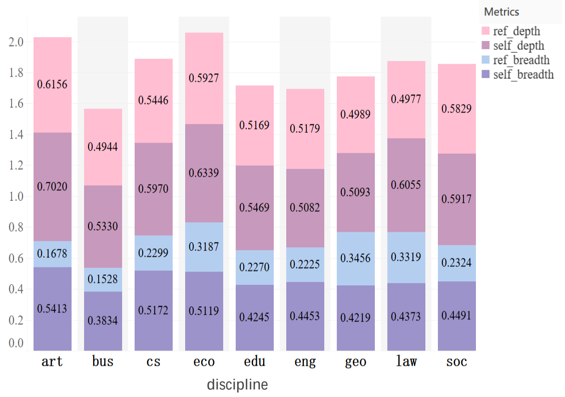

Figure 3: Research Depth and Breadth Values across Different Disciplines.

4.3. Information Cocoon at Different Citations Levels

Finally, our objective is to examine potential variations among papers at different citation levels. To accomplish this, we gathered citation data for each paper in three representative fields: art, education, and geography, which exhibit high-level, mid-level, and low-level information cocoons, respectively. Subsequently, we classified all papers into four groups based on their citation counts: Group A comprises papers with the highest number of citations over 300; Group B consists of papers with citations ranging from 100 to 300; Group C includes papers with citations between 10 and 100; and Group D encompasses papers with the lowest citation count, less than 10. Then, we calculated the research depth value and research breadth value within each group and presented the results in Figures 5 and 6.

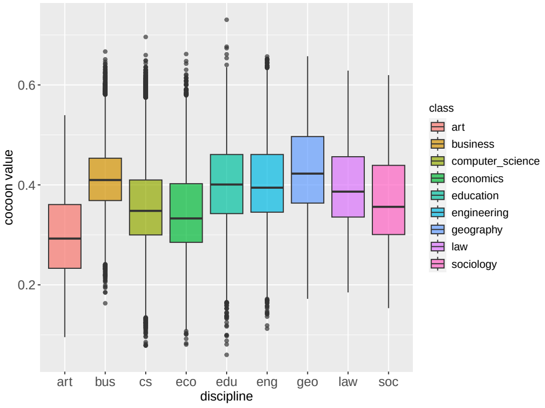

Figure 4: Information Cocoon Value across Different Disciplines.

The analysis revealed consistent trends in the metric values across papers in groups A, B, C, and D. A clear pattern emerged between the number of citations and the degree of information cocooning. It was observed that the most highly cited papers generally exhibited higher levels of research depth and breadth, indicating their comprehensive exploration of a specific area along with extensive coverage of related domains. In group B, which comprised highly cited papers, there was a focus on academic hotspots, attracting scholars with broad interests; however, the depth of analysis may have been comparatively limited. On the other hand, less-cited papers demonstrate a narrower research breadth but exhibit a significant level of depth. These papers, which often delved into niche issues or possessed a high degree of depth, may have faced challenges in gaining acceptance due to their specialized nature.

In conclusion, our findings suggest that while extensive research can lead to a considerable number of citations, studies that exhibit both depth and breadth tend to have a greater impact. It is crucial to recognize that a lower number of citations does not necessarily imply lower quality. Instead, such papers may possess a high level of depth or focus on niche topics, holding potential for further development. Therefore, instead of solely emphasizing citation counts, evaluating research based on both research depth and breadth can provide more informative insights. Consequently, research depth and breadth can serve as indicators for scholars to reflect upon the information cocoons, as well as for the academic community to assess the influence of research.

5. Discussion

In our study, we present an index and methodology for quantifying the scale of information cocoons within academic environments and classify them accordingly. The key findings of this paper can be summarized as follows. Firstly, we observe a gradual breakdown of information cocoon within the overall academic landscape, indicating a trend towards greater comprehensiveness and innovation. Secondly, disparities exist in terms of research depth, breadth, and information cocooning across different

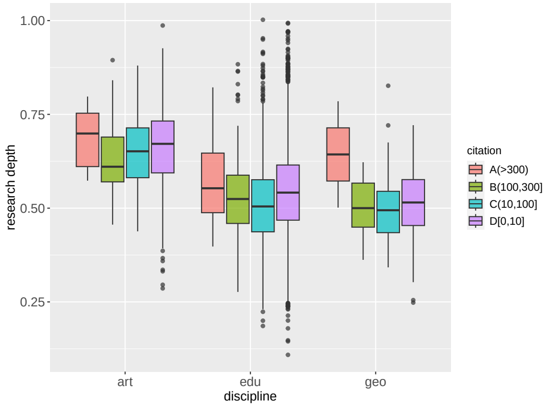

Figure 5: Research Depth Value across Different Citation Levels. Depth value was presented by the average value of "ref_depth" and "self_depth".

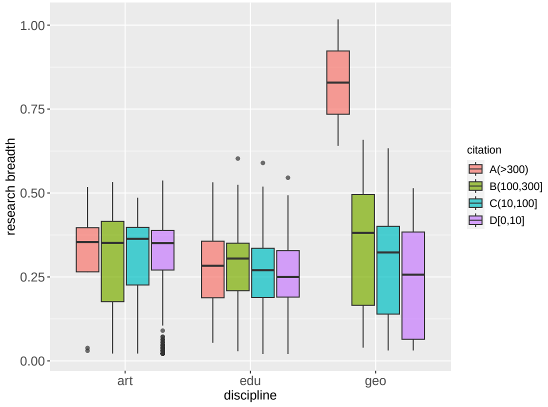

Figure 6: Research Breadth Value across Different Citation Levels. Breadth value was presented by the average value of "ref_breadth" and "self_breadth".

disciplines. Lastly, it is worth noting that while some papers may accumulate citations through diverse research, it is the papers that possess both research depth and breadth that have the potential to truly become influential. Additionally, it is important to consider that papers with a low citation count may be a result of delving deeper into niche topics, rather than indicating lower research quality. Therefore, scholars should adeptly leverage extensive and intricate academic information, continuously evaluating whether their research processes are constrained by information cocoons. Communities can utilize the research depth and breadth metrics proposed in this study to effectively assess the impact of the research.

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] "Bursting Your (Filter) Bubble: Strategies for Promoting Diverse Exposure," in Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion, San Antonio Texas USA: ACM, February 23, pp. 95–100.

[5] Nikolov, D., Oliveira, D. F. M., Flammini, A., and Menczer, F. 2015. "Measuring Online Social Bubbles," PeerJ Computer Science (1), p. e38.

[6] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. 2021. "The Echo Chamber Effect on Social Media," Proceedings of the National Academy of Sciences (118:9), p. e2023301118.

[7] Sutherland, K. A. (2018). Holistic academic development: Is it time to think more broadly about the academic development project? International Journal for Academic Development, 23(4), 261–273.

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

[9] 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.

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

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

[12] 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.

[13] Yang, S., and Han, R. 2015. "Breadth and Depth of Citation Distribution," Information Processing & Management (51:2), pp. 130–140.

[14] 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.