Quantifying scientific novelty of doctoral theses with Bio-BERT model

Alex J. Yang[1], Yi Bu[2], Ying Ding[3], and Meijun Liu[4,*]

[1] *School of Information Management, Nanjing University, Nanjing, China*
[2] *Department of Information Management, Peking University, Beijing, China*
[3] *School of Information, University of Texas at Austin, Austin, TX, USA*
[4] *Institute for Global Public Policy, Fudan University, Shanghai, China*

Abstract

Scientific novelty plays a pivotal role in advancing scholarly endeavors, driving the evolution of knowledge across various disciplines. In this paper, we present a methodology for quantifying the scientific novelty of biomedical doctoral theses utilizing the Bio-BERT model. Leveraging BERN2 for bio-entity extraction and normalization, we analyze a dataset comprising 305,693 doctoral theses to generate unique bio-entity combinations. Employing Bio-BERT, we calculate the semantic distance between bio-entities in entity pairs and establish a criterion for identifying novel entity pairings. We introduce a novelty score to assess the scientific novelty of each thesis. Our findings contribute to the discourse on scientific novelty assessment, offering insights into the evolving landscape of biomedical research and providing a framework for enhanced analysis of scholarly innovation for early-career scientists based on their doctoral theses.

Keywords

Biomedical research, Bio-BERT model, Doctoral theses, BERN2, Scientific novelty

1. Introduction

Scientific novelty serves as a cornerstone in scholarly pursuits, driving the progression of knowledge across diverse fields. Originating from Schumpeter's seminal insights on business cycles in the 1930s, the concept of scientific novelty underscores the transformative nature of innovation, wherein novel theories, methodologies, data, or discoveries emerge to shape subsequent investigations (1). Over time, this perspective has become integral to the examination of innovation, permeating scholarly discourse and guiding inquiries into the novelty of scientific artifacts such as publications, patents, and grant proposals (2-6).

With the exponential growth of scientific data, researchers have turned to various methodologies to operationalize and quantify scientific novelty, often leveraging textual

* Corresponding author.
alexjieyang@outlook.com (A. J. Yang); buyi@pku.edu.cn (Y. Bu); ying.ding@ischool.utexas.edu (Y. Ding)
meijunliu@fudan.edu.cn (M. Liu)
0000-0002-5385-5761 (A. J. Yang); 0000-0003-2549-4580 (Y. Bu); 0000-0002-2800-5511 (M. Liu)

information or citation data to delineate knowledge elements and their combinations (7, 8). For instance, Fleming (2001) proposes evaluating novelty in patents by identifying unexplored technology classes (2), while Boudreau et al. (2016) advocate for assessing grant proposals based on unique MeSH keyword combinations (7). Despite these endeavors, challenges persist in accurately capturing the intricate interplay of knowledge components.

In this context, recent advancements aim to refine methodologies for gauging scientific novelty, drawing inspiration from combinatorial approaches that consider the semantic relationships between knowledge elements (9). Liu et al. (2022) propose an innovative methodology for assessing scientific novelty in biomedical publications related to coronavirus (10), utilizing bio-entities as fundamental knowledge units and employing a pre-trained Bio-BERT model to measure their semantic distance. By scrutinizing entity pairs and identifying novel combinations based on a semantic distance threshold, this approach offers a nuanced perspective on scientific novelty, surpassing traditional methods reliant solely on textual or citation-based analyses.

Building upon this pioneering framework, our study endeavors to evaluate the scientific novelty of biomedical doctoral theses through a comprehensive five-step method. By adopting the approach outlined by Liu et al. (2022) (10), which integrates domain-specific contexts and semantic analysis, we aspire to enhance the precision and depth of our analysis, providing invaluable insights into the evolving landscape of biomedical research. Through this endeavor, we contribute to the ongoing discourse on scientific novelty assessment, advancing methodologies to better encapsulate the richness and complexity of scholarly innovation.

The primary data source for this study is the Sciences and Engineering Collection of The ProQuest Dissertations & Theses Citation Index (PQDT). PQDT stands as the world's largest multidisciplinary dissertation database, housing over 5.5 million dissertations from universities worldwide and serving as an official repository for the US Library of Congress. From a compilation of US higher education institutions provided by the Carnegie Commission on Higher Education, we gather records of doctoral theses from the Science and Engineering collection of PQDT. This dataset encompasses 1,109,491 theses from 828 US institutions, spanning publication years 1960 to 2016. PQDT offers comprehensive information about dissertations, including author details, advisors, universities, subjects, and publication years. Each thesis is associated with one or more subjects chosen by the author, which can be mapped to 22 broader disciplines. Prioritizing data accuracy, we analyze doctoral theses published from 1980 to 2016, retaining 313,274 theses in the biomedical sciences encompassing biological science, health, and medical science. The steps of quantifying scientific novelty of doctoral theses are shown in Figure 1.

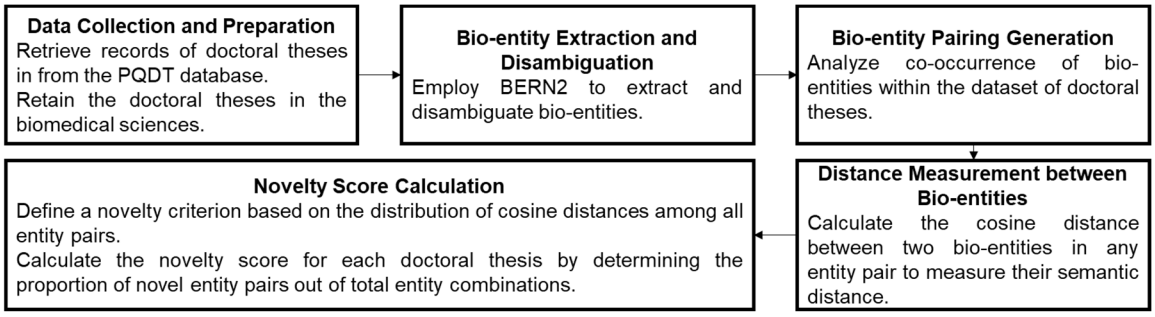

Figure 1: Steps of quantifying scientific novelty of doctoral theses.

2. Extracting and disambiguating bio-entities

We utilize BERN2 (11), an advanced neural biomedical tool, to extract biomedical entities from a corpus comprising 313,274 doctoral theses. BERN2 comprises two principal models: (1) Named Entity Recognition (NER), which discerns nine types of biomedical entities—gene/protein, disease, drug/chemical, species, mutation, cell line, cell type, DNA, and RNA—employing a multi-task NER model; and (2) Named Entity Normalization (NEN), which associates annotated entities with concept unique identifiers using a combination of rule-based and neural network-based NEN models. BERN2's superiority over existing biomedical text mining tools (12) lies in its ability to provide more efficient annotations.

We opt to extract bio-entities from the titles and abstracts of doctoral theses rather than relying on full texts for several reasons. Firstly, although the PQDT database offers access to 3 million full texts of doctoral dissertations added since 1997, a download limit is imposed. However, titles and abstracts are available for nearly all doctoral theses added since 1980. The title succinctly encapsulates the main topic addressed by the author, while the abstract provides a summary of the substantive content. Utilizing titles and abstracts instead of full texts ensures higher data accessibility, a denser concentration of relevant vocabulary reflecting the publication's topic, as well as advantages such as reduced computation time and simplified data preprocessing processes.

Utilizing BERN2, we extract 1,519,599 annotated bio-entity names from the titles and abstracts of 305,693 doctoral theses from the final dataset. In 2.42% of the 313,274 doctoral theses, we fail to extract any bio-entity, leading to the exclusion of these theses from further analyses, resulting in a remaining subset of 305,693 doctoral theses. The 1,519,599 annotated bio-entity names were disambiguated and linked to 118,349 unique bio-entity IDs. The standard name for each ID was determined as the most frequently occurring bio-entity name associated with it in the biomedical doctoral theses. In cases of multiple associated names with unequal occurrences, one was randomly designated as the standard name.

Subsequently, we establish pairings among the 118,349 distinct bio-entity IDs by analyzing their co-occurrence in the dataset comprising 305,693 doctoral theses. Among these theses, 8.45% exclusively mentioned a single bio-entity, rendering the generation

of any bio-entity combinations impossible. Consequently, these instances were excluded from subsequent analyses, leaving us with 277,288 doctoral theses and resulting in the generation of 68,949,061 unique bio-entity combinations.

3. Measuring the distance of two bio-entities

Using the standard names associated with the 118,349 unique bio-entity IDs obtained in the previous step, we convert each standard bio-entity name into a vector representation using a Bio-BERT model. We then calculate the distance between two bio-entities that are denoted by i and j, $D_{i,j}$, for any entity combination that is generated from the doctoral theses using Equation 1.

$$D_{i,j}=1-CosSim_{i,j} \quad (1)$$

where $CosSim_{i,j}$ is the cosine similarity between entities i and j based on their corresponding vector representations that are obtained from the Bio-BERT model. The examples of an entity vector space for three theses based on the Bio-BERT model are shown in Figure 2a-b.

We develop a criterion to determine what qualifies as a novel combination of entities. To do this, we analyze the distribution of cosine distances among all pairs of entities in our dataset. If the cosine distance between the two constituent entities of a pair falls within the top 10% of this distribution, we consider it as a novel entity pairing. The 90th percentile of the distribution corresponds to a cosine distance of 0.279 (Figure 2c). Any entity pair with a cosine distance greater than 0.279 is considered to be a novel combination. We further define a novel thesis as a doctoral thesis that includes at least one novel entity combination/pair.

To provide a nuanced evaluation of each doctoral thesis's scientific novelty, we introduce the novelty score. This score is calculated by determining the proportion of novel entity pairs out of the total number of entity combinations generated within a given thesis. As an illustration, let us consider a thesis that mentions three bio-entities: a, b, and c. Within this thesis, the number of generated entity combinations is calculated as $C_3^2=3$. Out of these three entity pairs, only the combination of a and b meets our novelty criterion, which requires the cosine distance between the two bio-entities to be greater than 0.279. Accordingly, the novelty score for this particular thesis is 1/3. The novelty score is bounded between 0 and 1, with a higher score indicating a greater degree of novelty. This metric provides a precise and continuous measure of the unique combinations of entities present in each thesis.

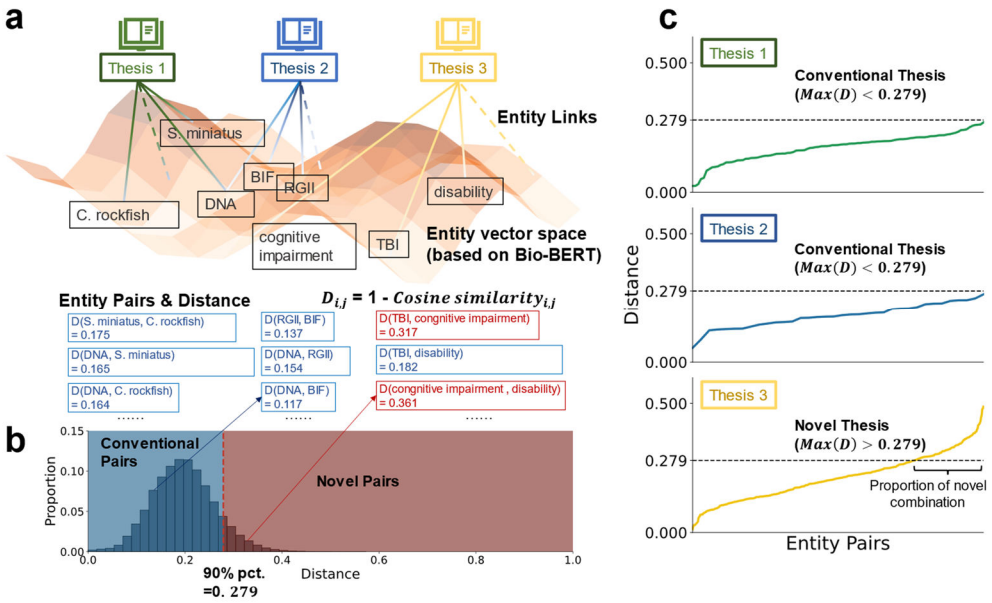

Figure 2: The illustration of how to measure novelty scores for doctoral theses using the Bio-BERT model. (a) An entity vector space containing all entities extracted from three sample doctoral theses based on Bio-BERT. (b) The distribution of cosine distances between entities for all entity pairs extracted from the three sample doctoral theses. Within each thesis, the entity pairs are ordered from left to right based on their cosine distance values. (c) The distribution of cosine distance for all entity pairs extracted from all doctoral theses in this study. If the cosine distance between the two constituent entities of an entity pair falls within the upper 10th percentile of this distribution, it is considered a novel entity pair.

Acknowledgements

This study is sponsored by the National Natural Science Foundation of China (72104054, 72104007), the Shanghai Pujiang Talent program (21PJC026), and the Key Project of the National Natural Science Foundation of China (72234001). We acknowledge the technical support provided by Mr. Grant Guo.

References

1. J. A. Schumpeter, *Business cycles* (Mcgraw-hill New York, 1939), vol. 1.
2. L. Fleming, Recombinant uncertainty in technological search. *Management Science* **47**, 117-132 (2001).
3. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468-472 (2013).
4. M. L. Weitzman, Recombinant growth. *The Quarterly Journal of Economics* **113**, 331-360 (1998).

5. D. K. Simonton, Scientific creativity as constrained stochastic behavior: the integration of product, person, and process perspectives. *Psychological Bulletin* **129**, 475 (2003).
6. J. Wang, S. Shibayama, Mentorship and creativity: Effects of mentor creativity and mentoring style. *Research Policy* **51**, 104451 (2022).
7. K. J. Boudreau, E. C. Guinan, K. R. Lakhani, C. Riedl, Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science* **62**, 2765-2783 (2016).
8. S. Chai, A. Menon, Breakthrough recognition: Bias against novelty and competition for attention. *Research Policy* **48**, 733-747 (2019).
9. P. Azoulay, J. S. Graff Zivin, G. Manso, Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics* **42**, 527-554 (2011).
10. M. Liu *et al.*, Pandemics are catalysts of scientific novelty: Evidence from COVID-19. *J Assoc Inf Sci Technol* **73**, 1065-1078 (2022).
11. M. Sung *et al.*, BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics* **38**, 4837-4839 (2022).
12. D. Kim *et al.*, A neural named entity recognition and multi-type normalization tool for biomedical text mining. *IEEE Access* **7**, 73729-73740 (2019).