Joint Workshop of the 5th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2024) and the 4th AI + Informetrics (AII2024)

Quantifying scientific novelty of doctoral theses with Bio-BERT model

Presenter: Meijun Liu
Assistant professor, Institute of Global Public Policy, Fudan University

Changchun, China, 2024-04-24

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*

In a nutshell

- 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.

Why do we care?

NEW INNOVATOR AWARD

- 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

- Scientific novelty possesses the potential to become a breakthrough on its own and trigger subsequent advancements that may have far-reaching impacts (Criscuolo et al., 2017; Ulnicane, 2022).

- Early-career scientists are encouraged to generate novel knowledge and bring new ideas to science.

- Doctoral theses are widely recognized as a critical demonstration of independent research contributions and serve as the primary research output for junior scientists.

Conceptualizing scientific novelty

- The creation of any sort of novelty in art, science, or practical life-consists to a substantial extent of a recombination of conceptual and physical materials that were previously inexistence.

—Nelson and Winter(1982)

- Combinatorial novelty: combining existing scientific components in an unprecedented fashion.
- Economists (Schumpeter, 1939; Nelson & Winter, 1982); psychologists(Mednick, 1962; Simonton, 2004);sociologists (Latour & Woolgar, 1986).

- Combinatorial novelty is just one dimension of novelty.

Quantifying scientific novelty of doctoral theses: a five-step process

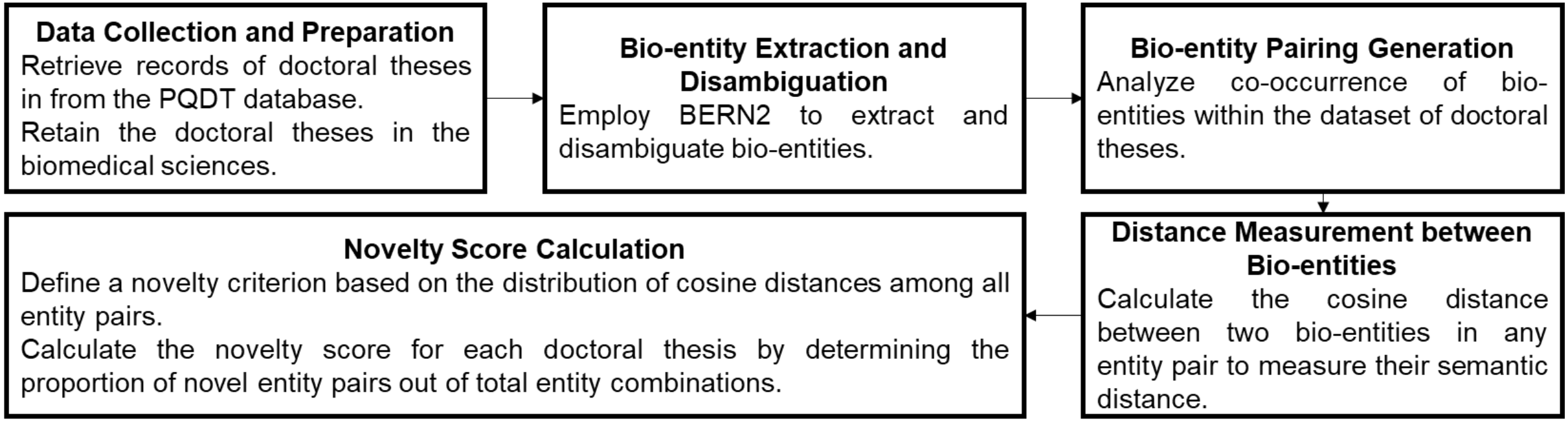
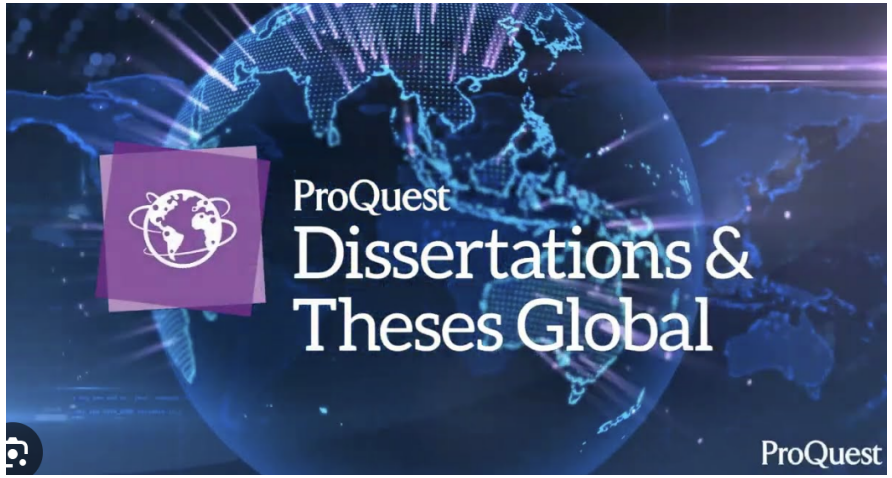


Figure 1: Steps of quantifying scientific novelty of doctoral theses.

Data collection and preparation



- 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.
- Each thesis is associated with one or more subjects chosen by the author, which can be mapped to 22 broader disciplines.
- We analyze doctoral theses published from 1980 to 2016, retaining 313,274 theses in the biomedical sciences encompassing biological science, health, and medical science.

Bio-entity extraction and disambiguation

- We utilize BERN2, an advanced neural biomedical tool, to extract biomedical entities from a corpus comprising 313,274 doctoral theses. We opt to extract bio-entities from the titles and abstracts of doctoral theses rather than relying on full texts for several reasons.
- 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.

Bio-entity pairing generation

- 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.

Measuring the distance of two bio-entities and calculate novelty score

- 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.

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

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

- 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.
- 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.

Measuring the distance of two bio-entities

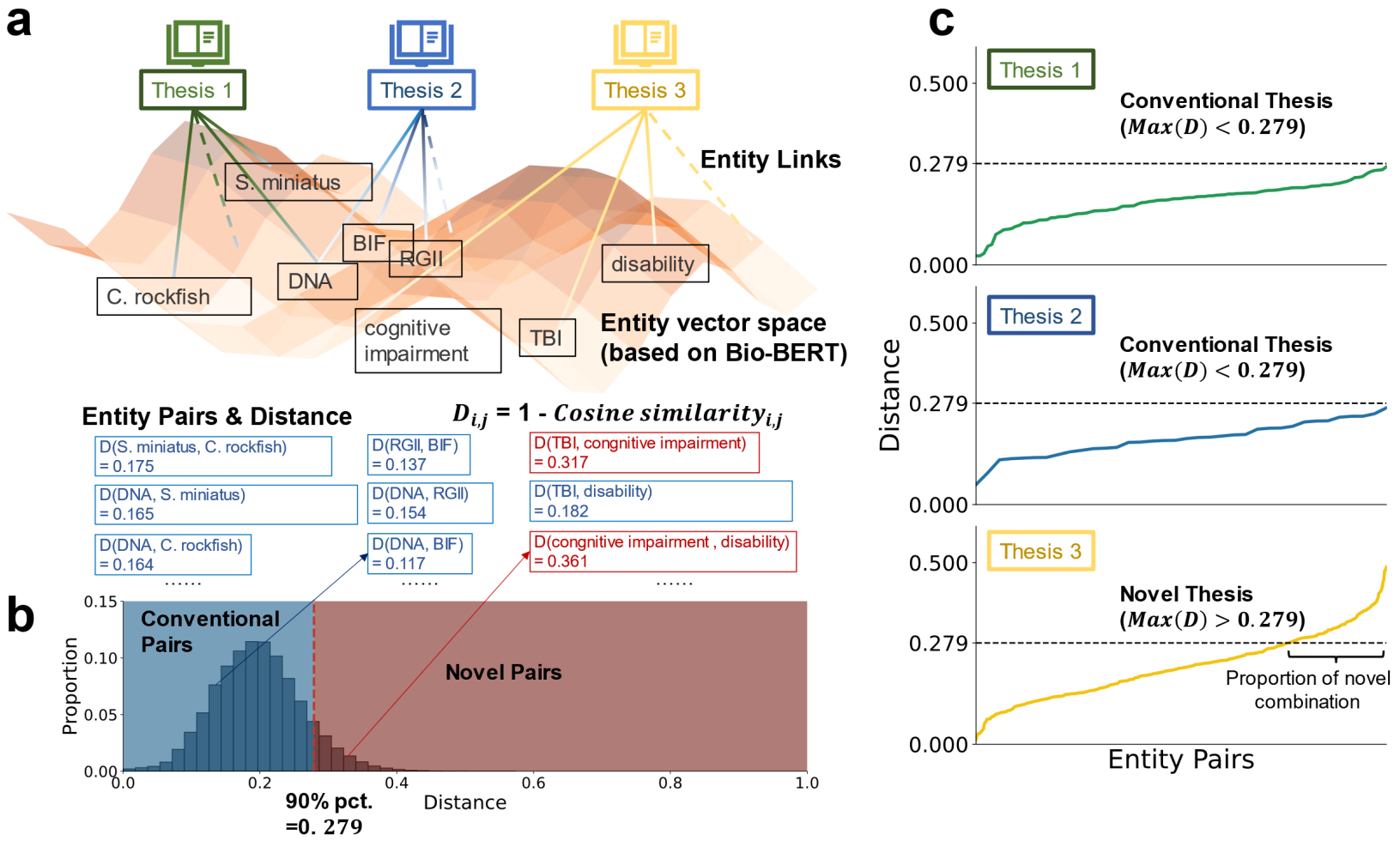


Figure 2: The illustration of how to measure novelty scores for doctoral theses using the Bio-BERT model.

The next step:

Applying this indicator to investigate how early-career scientists pursue novel research path.

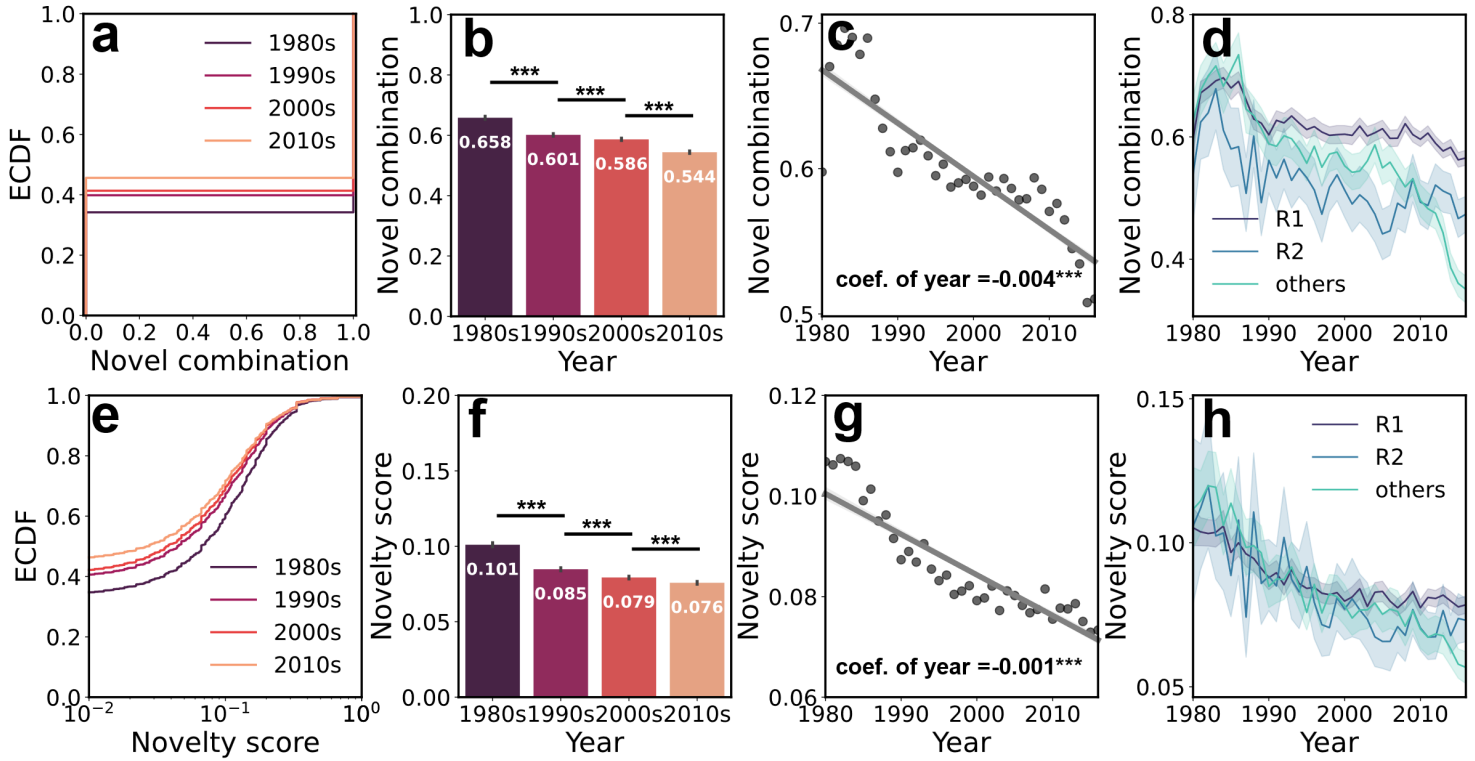


Figure 3. The temporal evolution of scientific novelty in doctoral theses.

meijunliu@fudan.edu.cn