

연구주제 탐색과 연구 경력 형성:

딥러닝 기반 연구자 데이터 분석

2026. 06. 23.

최세영

서울대학교 박사과정
기술경영경제정책전공

핵심 질문

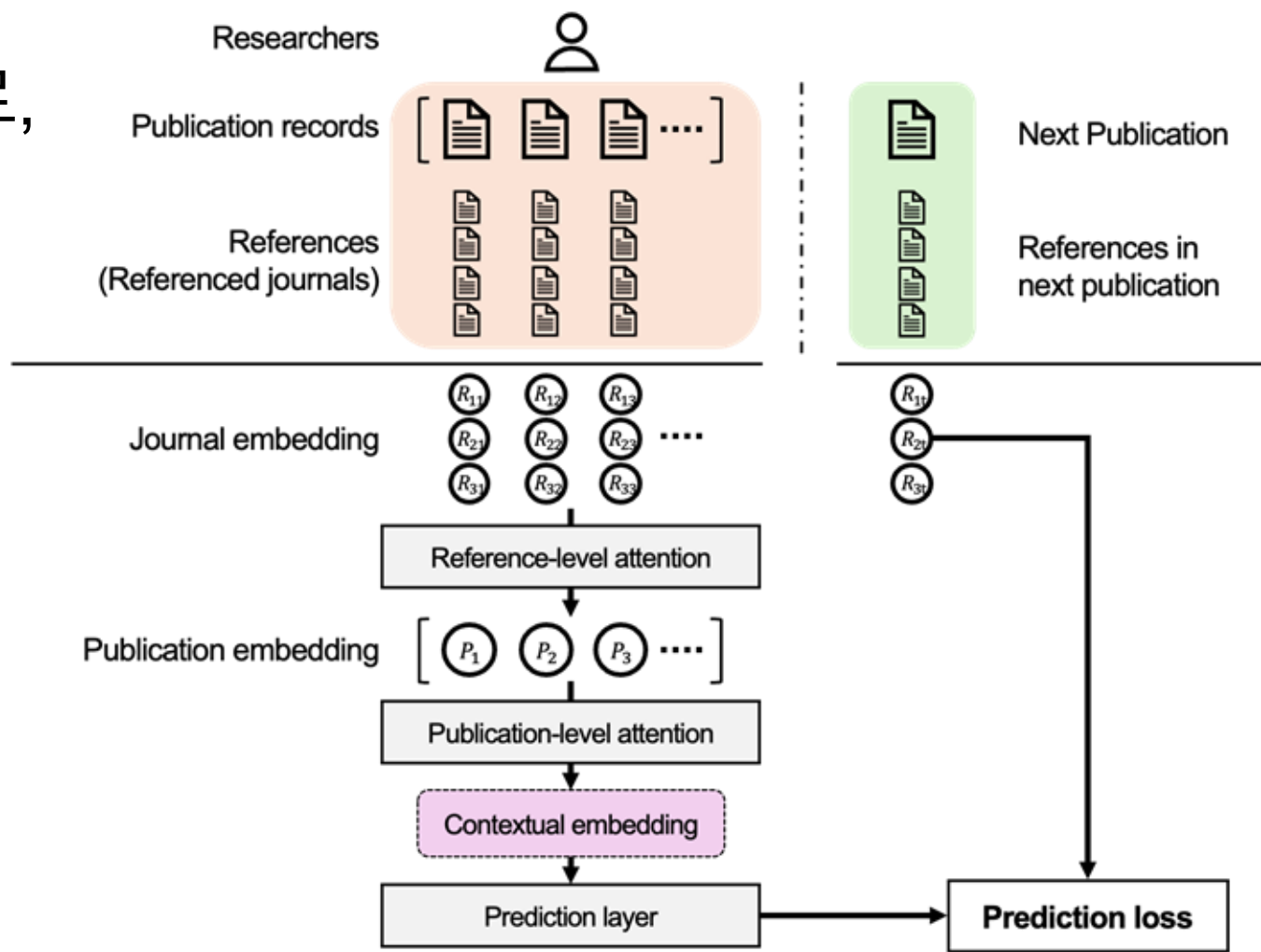
- 연구자들은 익숙한 주제에 몰두하는가, 새로운 주제를 탐색하는가?
 - 연구자들이 “얼마나 탐색하는지”, “탐색의 비용/보상은 어떻게 나타나는지”에 관한 선행연구는 많이 이루어짐
 - 그러나, 탐색을 통해 새로 얻은 지식을 이후 얼마나 (자주/오래) 활용하는지에 대한 종단적 연구는 거의 이루어지지 않음
- 새로운 지식을 탐색하였다면, 그것을 얼마나 활용하는가? 그것이 이후 연구 경력에 어떤 영향을 미치는가?
 - 1) 연구주제 탐색 강도를 측정
 - 2) 연구주제 선택이 이후 연구 경력에 미친 영향을 측정

OpenAlex 데이터 구조화

- **연구자의 연구 경력** = 논문 시퀀스 $H_a = [p_1, p_2, \dots, p_T]$
 - 연구자가 작성한 논문의 시퀀스
 - 논문-저자 쌍 데이터 (OpenAlex의 name disambiguation 가공 데이터 활용)
*단, 퀄리티 문제에 유의 — 동명이인이 하나의 프로필로 존재하는 등의 오류 있음
- **논문의 주제** = 인용한 저널의 집합 $p_t = \{j_1, j_2, \dots, j_n\}$
 - 각 논문이 어떤 저널에 실린 논문들을 인용하였는지
 - 논문-논문 인용 데이터 + 논문-저널 source 데이터
 - (대안) 각 논문에 할당된 topic, 혹은 abstract의 embedding 또한 활용 가능

예측 모델 학습

- 연구자의 이전 기록을 바탕으로,
다음 논문의 연구주제를 예측
- $\Pr(p_t | p_1, p_2, \dots, p_{t-1})$
- 어텐션 기반 순차 예측 모델
 - SASRec + hierarchical attention
 - 연도별 분할 학습



예측 모델 기반 추정

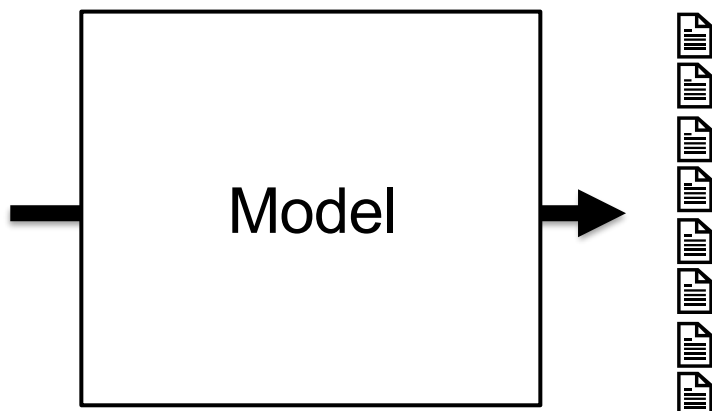
- 연구자의 실제 연구 경력 $[p_1, p_2, \dots, p_{t-1}]$ 이 주어졌을 때,
- 다음에 어떤 연구주제가 어떤 확률로 등장할 것인가: $\Pr(\cdot | p_1, \dots, p_{t-1})$

Saeyoung Choi

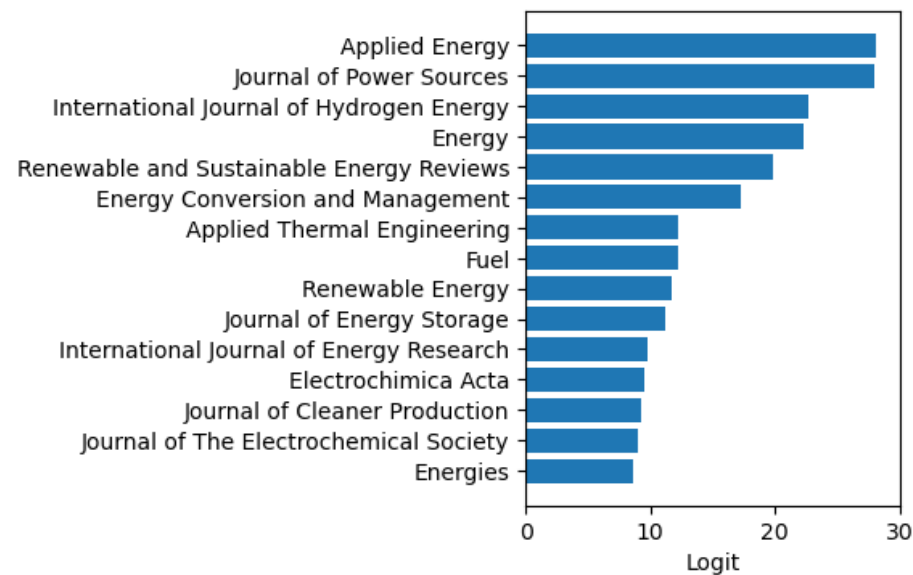


Previously cited:

Applied Energy, Fuel Cells, Fuel, Energy,
Journal of Power Sources,
Energy conversion and management, ...



Top-K logits of citing journals



예측 모델 기반 추정

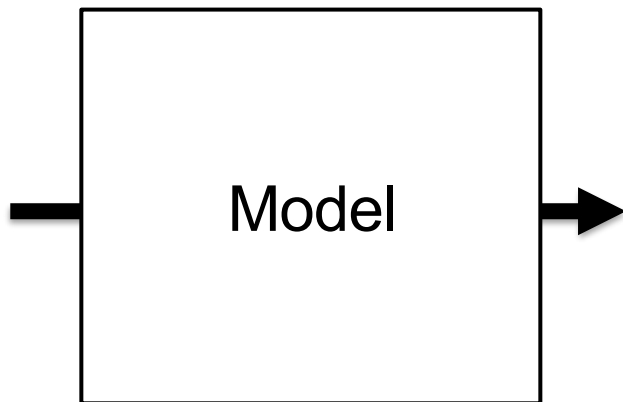
- 모델 예측에 따르면, 실제 (관측된) 다음 연구주제가 등장할 확률은 얼마였는가: $\Pr(p_t | p_1, \dots, p_{t-1})$

Saeyoung Choi



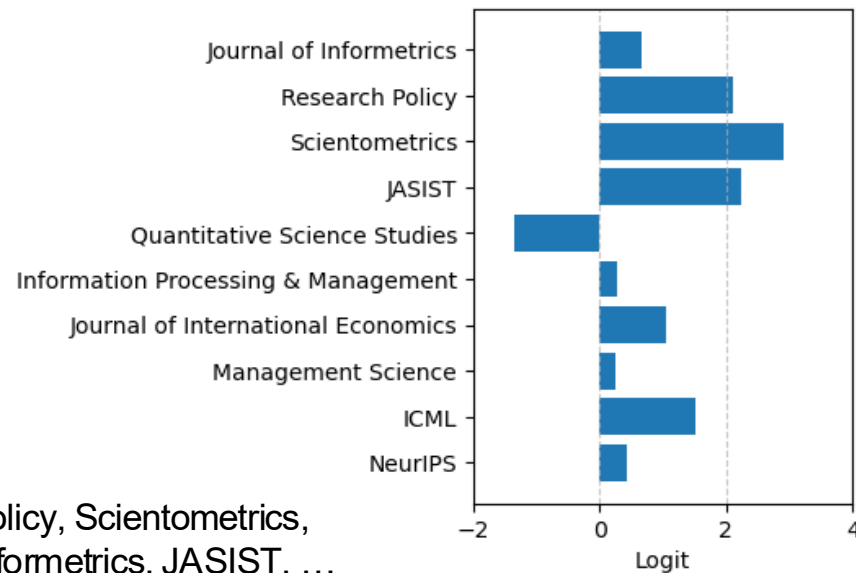
Previously cited:

Applied Energy, Fuel Cells, Fuel, Energy,
Journal of Power Sources,
Energy conversion and management, ...



Research Policy, Scientometrics,
Journal of Informetrics, JASIST, ...

Logits for actual cited journals

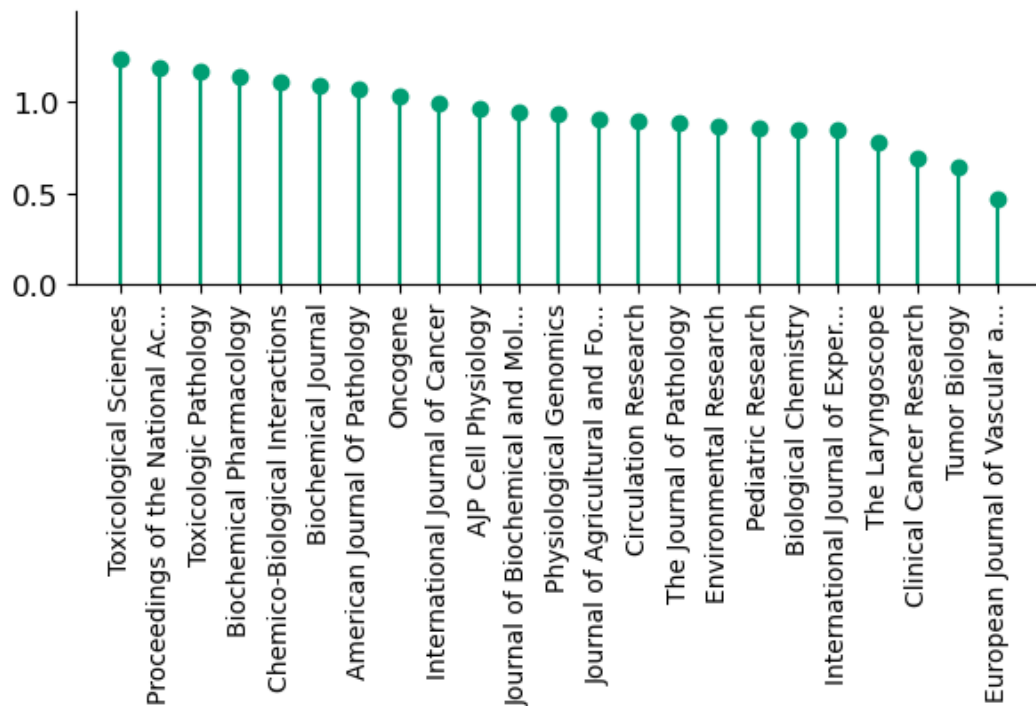


측정: 연구주제 탐색 강도

- **연구주제 탐색 강도:** $-\log \Pr(p_t | p_1, \dots, p_{t-1})$
 - 이전과 비슷한 연구주제를 반복하는 경우,
 - $= [p_1, \dots, p_{t-1}]$ 다음에 p_t 가 나타날 확률이 높음
 - = 탐색 강도가 낮음
 - 이전과 다른 새로운 연구주제를 탐색하는 경우,
 - $= [p_1, \dots, p_{t-1}]$ 로부터 p_t 를 예상하기 어려움 (나타날 확률이 낮음)
 - = 탐색 강도가 높음
- *연구주제마다 전역 인기도 $\Pr(p_t)$ 가 다름 (예: AI vs. 역사학) → 사후 보정 필요

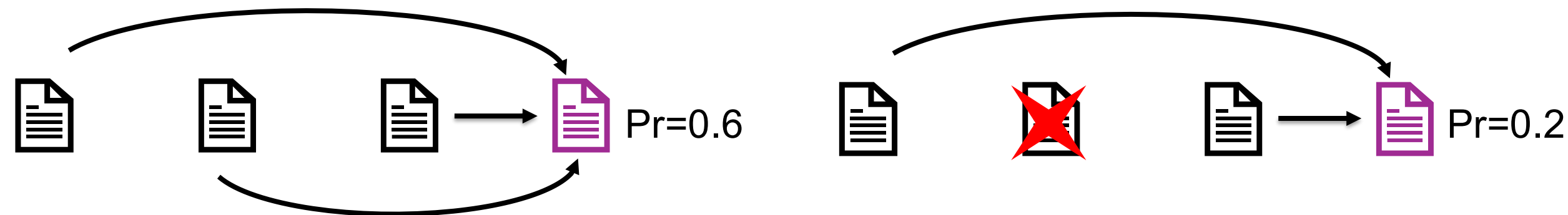
측정: 연구주제 탐색 강도

- 연구주제 탐색 강도: $-\log \Pr(p_t | p_1, \dots, p_{t-1})$
 - *논문이 여러 연구주제의 합으로 표현되므로 $p_t = \{j_1, j_2, \dots, j_n\}$, 연구주제 $j \in p_t$ 에 대해 계산된 $-\log \Pr(j | p_1, \dots, p_{t-1})$ 를 집계해야 함



측정: 연구 경력 내 의존성

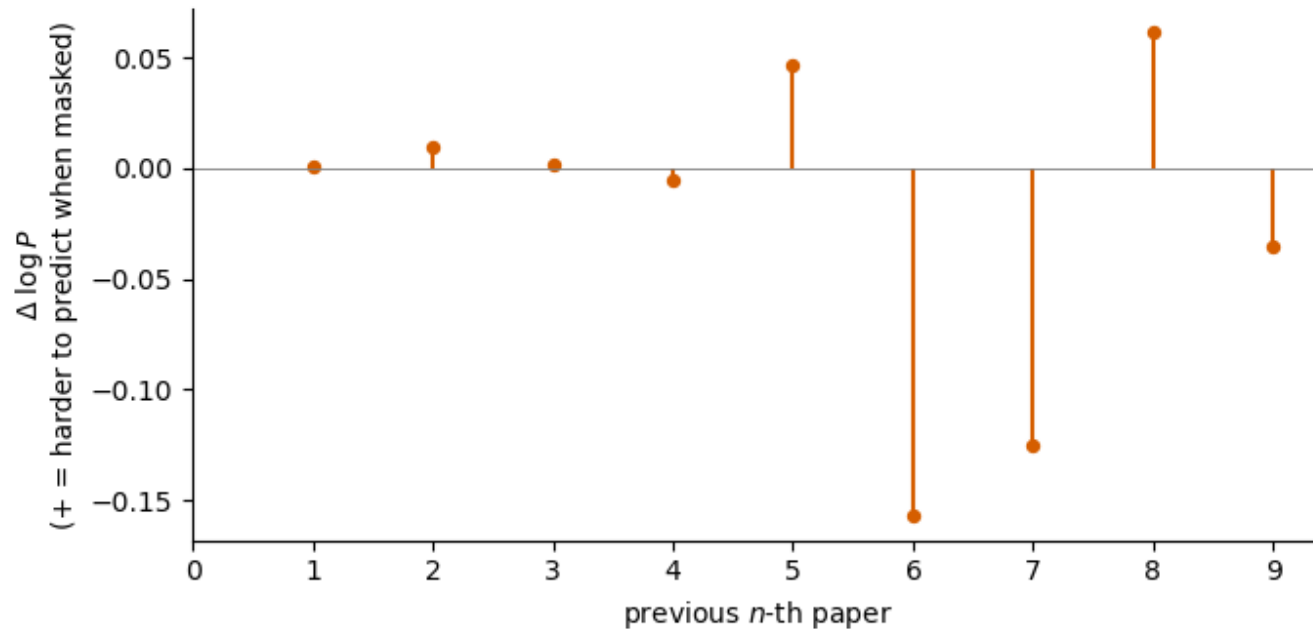
- 모델 기반의 반사실적 추론
 - 모델이 순차적 의존성을 충분히 학습하였다고 가정
 - 시퀀스 내 특정 요소를 **임의로 제거**하였을 때, 예측 확률의 변화를 계산



- 연구 경력 내 의존성: $\log \Pr(p_t | p_1, p_2, \dots, p_{t-1}) - \log \Pr(p_t | p_2, \dots, p_{t-1})$
 - 연구 p_1 를 수행한 것이, 이후 연구 p_t 가 나타날 확률을 높이는가/낮추는가?

측정: 연구 경력 내 의존성

- 연구 경력 내 의존성: $\log \Pr(p_t | p_1, p_2, \dots, p_{t-1}) - \log \Pr(p_t | p_2, \dots, p_{t-1})$
 - 연구 p_1 를 수행한 것이, 이후 연구 p_t 가 나타날 확률을 높이는가/낮추는가?



- 이후 논문들이 높은 의존성을 보이는 논문 = 연구 경력 내 핵심 논문!

향후 분석 구상

- “(단기적) 연구주제 전환” 뿐 아니라, “(장기적) **커리어 전환**”을 식별
 - **탐색 강도가 높은** 논문에 대해, 경력 내 후속 논문들이 **높은 의존성**을 가짐
 - = 이전 경력과 다른 새로운 주제이면서, 이후 경력 형성에 큰 영향을 미친 논문
- “커리어 전환”에 영향을 미치는 요인 탐색
 - **연구 경력 단계**: 어느 경력 단계에서 커리어 전환이 더 자주 일어나는가?
 - **연구과제**: 지원(사사 표기)을 받은 경우 커리어 전환이 더 자주 일어나는가?