Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 1, pp. 55-69, February 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.1.055

FindCoResearcher: Retrieval 기반 협력 연구자 추천 시스템

성시열[1] · 김재희[2] · 천재원[1] · 손준영[1] · 강필성[2†]

[1]고려대학교 산업경영공학과 / [2]서울대학교 산업공학과

FindCoResearcher: Retrieval-based Collaborative Researcher Recommendation System

Siyul Sung[1] · Jaehee Kim[2] · Jaewon Cheon[1] · Junyeong Son[1] · Pilsung Kang[2]

[1]Department of Industrial Management Engineering, Korea University
[2]Department of Industrial Engineering, Seoul National University

Researchers primarily depend on personal networks or manual exploration when searching for collaborative researchers. This approach presents significant limitations in identifying researchers suitable for their research topics. While Information Retrieval (IR) approaches can address these limitations, two major challenges arise: datasets with limited expression diversity and the gap between realistic scenarios and IR research. We propose FindCoResearcher, a collaborative researcher recommendation system that incorporates a query generation methodology for increasing expression diversity and a new evaluation metric for bridging realistic scenarios. To ensure query diversity, we construct query sets for each passage by diversifying query styles and specificity levels based on augmented passages. We introduce a researcher-unit Top-k Accuracy evaluation approach that better reflects realistic scenarios. We fine-tuned a dense encoder BGE-M3 using DPR achieving superior researcher search performance. FindCoResearcher is expected to promote industry-academia collaboration and expand their collaborative networks.

Keywords: Information Retrieval, Collaborative Researcher Recommendation, Passage-based Query Generation, Evaluation Metric

1. 서 론

공동 연구 및 산학 협력은 전문성과 자원의 결합을 통해 실무적 문제 해결과 혁신적 연구 성과 창출에 기여한다. 이러한 장점 덕분에 다양한 분야에서 협력 연구가 활발히 이루어지고 있다. 하지만 협력 연구자를 탐색할 때 대부분 인적 네트워크나 메뉴얼한 정보 탐색에 의존하여 원하는 연구 주제나 목적에 부합하는 협력 연구자를 체계적으로 찾는 것이 어렵다. 이러한 한계를 극복하기 위해서는 연구자의 요구에 맞는 협력 연구자를 추천해주는 검색 시스템이 필요하다.

이러한 검색 시스템 구축에는 정보 검색(Information Retrieval, IR) 기술이 효과적으로 활용될 수 있다. IR 기술은 검색 엔진, 추천 시스템, 전문 정보 검색 등 다양한 분야에서 사용자의 질문과 관련된 정보를 자동으로 찾아주는 데 적용되고 있다(Abass *et al.*, 2017; Hambarde *et al.*, 2023; Nadkarni, 2002). 또한, IR 기술의 성능을 증대시키기 위해 거대언어모델(Large

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(RS-2025-02214591, 자율 제조 구현을 위한 현장 작업자 친화적 혁신 AI 에이전트 개발), (RS-2024-00460011, 인류세의 기후테크 역량 강화를 위한 기후환경 데이터 구축 처리 플랫폼), (RS2021-II211343, 인공지능 대학원 프로그램(서울대학교)). 또한 이 연구는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2024-00407803, RS-2025-23523657). 또한 이 논문은 교육부의 재원으로 BK21 FOUR 프로그램(산업혁신 애널리틱스 교육연구단)의 지원을 받아 수행된 연구임(No. 4120240214912).
† 연락저자 : 강필성 교수, 08826 서울시 관악구 관악로1 서울대학교 공과대학 산업공학과, Tel : 02-880-7172, Fax : 02-889-8560,
E-mail : pilsung_kang@snu.ac.kr
2025년 7월 29일 접수; 2025년 9월 2일 수정본 접수; 2025년 10월 9일 게재 확정.

Language Model, LLM)을 이용한 학습 데이터 생성 기술도 활발히 사용되고 있다(Kim *et al.*, 2025; Lee *et al.*, 2024). 하지만 기존 IR 기술을 협력 연구자 추천 시스템에 그대로 적용하는 경우, 다음과 같은 두 가지 현실적 제약이 존재한다.

첫째, 제한된 정보 환경에서의 질의어(query) 생성 문제가 존재한다. 검색 모델을 학습하기 위해서는 문서와 관련된 질의어 데이터가 필요하나, 일반적으로 각 연구자별 연구 실적(이하 '문서')만 확보 가능한 상황에서 실제 사용자 질의어를 확보하는 것은 현실적인 어려움이 따른다. 이처럼 문서와 관련된 질의어 데이터가 존재하지 않는 경우에는 문서에서 직접 질의어를 생성하여 검색 모델을 학습시키는 접근법이 필요하지만(Kim *et al.*, 2025; Lee *et al.*, 2024; Yu *et al.*, 2023), 이 과정에서 단순 프롬프팅 기법을 사용할 경우 질의어 표현의 다양성이 부족하다는 한계를 지니게 된다(Kang *et al.*, 2025; Bacciu *et al.*, 2024).

둘째, 기존 IR 시스템에서 활용하는 문서 단위 평가 방식에는 한계가 존재한다. 전통적으로 IR 평가 방식은 개별 문서 단위를 표준으로 하며(Cleverdon, 1967; Harman, 1992), 하나의 질의어에 대해 하나의 관련 문서가 연결된다고 가정한다. 하지만 이 방식은 여러 문서에 분산된 정보를 반영하지 못해 현실 상황에 적합하지 않다(Wang *et al.*, 2023). 특히 협력 연구자 추천 시스템의 경우, 사용자는 개별 문서가 아닌 '연구자' 단위로 결과를 얻고자 하기 때문에 목적을 충분히 반영하는 연구자 단위의 확장된 평가 방식이 필요하다.

본 연구에서는 FindCoResearcher(Find Collaborative Researcher)라는 정보 검색 기반 협력 연구자 추천 시스템을 제안하며, 두 가지 핵심 방안을 통해 기존 IR 기술의 한계를 극복하고자 한다. 첫째, 문서 기반 질의어 데이터셋 생성 시, 문서 증강과 질의어 스타일 및 구체화 수준을 다양화함으로써 질의어 표현의 풍부함을 증가시키고 효과적인 학습을 추구한다. 둘째, 문서 단위가 아닌 연구자 단위에서의 적절한 평가 방법론을 도입함으로써 목적에 맞는 정확한 평가 체계를 구축하고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 연구의 기반이 되는 관련 연구를 서술한다. 제3장에서는 표현 다양성을 향상시키는 데이터셋 구축 방법과 연구자 단위 평가 방식을 소개한다. 제4장에서는 실험 구성, 제5장에서는 실험 결과와 추가적인 분석 사항에 대해 서술한다. 마지막으로 제6장에서는 본 연구의 기여점과 한계를 정리하고, 향후 연구방향을 제안한다.

2. 관련 연구

2.1 과학 정보 검색

과학 정보 검색(Scientific Information Retrieval, SIR) 분야는 학술 문서와 같은 특수한 도메인을 대상으로 사용자에게 관련된 정보를 효과적으로 제공하는 것을 목표로 한다. 학술 문서는 일반적인 도메인과 다르게 전문 용어 사용 빈도가 높고 복잡한 학술적 개념들이 포함되어 있다. 그러므로 이를 효과적으로 처리하기 위한 특화된 방법론이 필요하다(Beltagy *et al.*, 2019; Wang *et al.*, 2023).

Beltagy *et al.*(2019)은 과학 도메인의 텍스트가 전문 용어 사용 빈도가 높고 복잡한 학술적 개념을 포함하여 기존의 범용 언어 모델로는 도메인의 특성을 충분히 반영하기 어렵다고 주장하였다. 이러한 문제를 해결하기 위해 114만 개의 과학 논문으로 구성된 대규모 코퍼스에서 사전 훈련된 SCIBERT를 제안하였다. SCIBERT는 과학 도메인의 다양한 자연어 처리 과업에서 우수한 성능을 보이며 도메인 특화 학습의 효과를 입증하였다. Wang *et al.*(2023)은 동일한 연구에 대해 연구 방법론, 실험 결과, 응용 분야 등 여러 측면에서 포괄적으로 해석될 수 있다는 학술 질의어의 다면적 특성에 주목하였다. 이러한 복잡성을 해결하기 위해 질의어를 여러 측면(aspect)과 하위 측면(sub-aspect)으로 분해하여 처리하는 DORIS-MAE를 제안하였다.

이러한 연구들은 주로 과학 도메인의 복잡성을 특수하게 관리하고 있지만, 여전히 실제 사용자의 검색 의도와 표현 다양성을 충분히 반영하지 못하는 공통적인 한계를 지닌다. 본 연구에서는 제한된 정보 환경에서도 실제 사용자 요구를 효과적으로 충족시키는 시스템을 구축하고자 한다.

2.2 문서 기반 질의어 생성

문서 기반 질의어 생성은 주어진 문서와 관련되어 있으면서 사용자의 요구 사항을 반영하는 질의어를 생성하는 기술로, 검색 모델 학습 시 문서 데이터만 존재하는 경우에 필수적인 과정이다. 이러한 배경에서 최근 LLM의 발전과 함께 합성 질의어 생성이 검색 모델 성능 향상에 효과적임을 입증하였으며(Kang *et al.*, 2025; Bacciu *et al.*, 2024), 특히 문서 기반 질의어 생성에서 표현 다양성을 높이는 것이 핵심적인 과제로 대두되면서 이를 위한 다양한 연구가 활발히 진행되고 있다(Sinha *et al.*, 2024).

Kang *et al.* (2025)은 제한된 문서 정보에서 단순 프롬프팅 기법으로 질의어를 생성할 경우, 생성된 질의어들이 유사한 표현 패턴을 반복하게 되어 실제 사용자의 다양한 검색 의도를 충분히 반영하지 못한다는 한계점을 지적하였다. 이러한 문제를 해결하기 위해 이전 질의어에서 부족하게 다뤄진 개념을 식별하고 이를 후속 질의어 생성에 반영함으로써 질의어 표현의 다양성을 체계적으로 확보하였다. Bacciu *et al.*(2024)은 LLM을 활용한 생성 기반 질의어 추천 시스템이 사전 학습된 표현 지식을 바탕으로 효과적인 질의어 생성을 수행할 수 있음을 보였다. 예시를 통해 프롬프트에 다양한 주제와 표현을 포함함으로써 모델의 일반화 성능을 높이고, 유사한 표현

의 반복을 방지하여 사용자 선호도를 개선할 수 있음을 실험적으로 확인하였다. 특히, 의미적으로 유사하지만 어휘적으로 다양한 추천 질의어를 생성하는 것이 검색 품질과 사용자 만족도에 기여함을 보여주었다.

이러한 연구들은 질의어 표현의 다양성 부족이 검색 모델의 일반화 성능과 견고성을 저해하는 핵심 요인임을 보여준다. 하지만 단일 문서를 기반으로 효과적인 표현 다양성 확보는 여전히 큰 도전 과제이다. 본 연구에서는 제한된 문서 정보로 인한 질의어 표현의 다양성 한계를 해결하는 방식을 통해 검색 모델의 성능과 강건성을 향상시키고자 한다.

결론적으로 지금까지 과학 정보 검색과 문서 기반 질의어 생성 분야에서 여러 연구가 진행되었지만, 학술 문서를 활용한 협력 연구자 추천 시스템 구축에 있어 여전히 개선 사항이 존재한다. 이러한 한계를 극복하기 위해 본 연구는 세 가지 기여점을 제시한다. 첫째, 표현 다양성이 부족한 기존 질의어 생성 방식의 문제점을 개선하기 위해 풍부하고 다양한 질의어 집합 구축 방법론을 제안한다. 둘째, 기존의 문서 단위 평가 방식이 사용자의 실제 목적인 연구자 탐색을 적절히 반영하지 못하는 문제를 해결하기 위해 연구자 단위 평가 지표를 새롭게 도입한다. 셋째, 위 개선점들을 포함하여 실제 사용자의 요구를 효과적으로 충족시키는 협력 연구자 추천 시스템을 구축하고자 한다.

3. 방법론

3.1 개요

본 연구에서는 사용자의 요구에 맞는 협력 연구자를 추천하는 검색 시스템인 FindCoResearcher을 제안한다. FindCoResearcher을 구축하고 평가하기 위한 전체 과정은 <Figure 1>과 같으며, 크게 데이터셋 구축, 모델 학습, 평가의 세 단계가 <Figure 1>의 Step 1, 2, 3로 표현된다. Step 1에서는 질의어 표현의 다양성을 높이기 위해 문서 증강과 체계적인 질의어 생성이라는 두 개의 세부 과정을 거쳐 문서 기반 질의어 데이터셋을 구축하며, 이 과정에 대한 자세한 설명은 3.2장에서 다룬다. Step 2에서는 구축한 데이터셋을 기반으로 검색 모델을 학습한다. Step 3에서는 기존의 문서 단위 평가 방식 대신 FindCoResearcher의 목적을 정확히 반영하기 위해 연구자 단위 평가 방식을 새롭게 도입하며, 구체적인 방식은 3.3장에서 자세히 설명한다. 마지막으로 3.4장에서 목적에 맞게 구성한 전체 파이프라인을 사용자 관점에서 설명한다.

(1) 문제 정의

본 연구에서 사용하는 데이터셋 구성 요소는 수식 (1), (2)와 같다. 수식 (1)의 P는 전체 문서 집합, Q는 전체 생성 질의어 집합을 의미한다. 이 때, (q_i,p_i)는 i번째 질의어-문서 쌍을 의미하며 M은 전체 질의어-문서 쌍의 개수를 의미한다. 수식 (2)의 $PROF$은 전체 연구자 집합이며, W는 전체 연구자 인원수를 의미한다. 전체 문서는 연구자 단위로 그룹화될 수 있으므로, 전체 연구자 인원수 W는 전체 문서 개수 M보다 작거나 같다. 모든 문서에는 연구자 정보가 존재하고, i번째 문서 p_i의 연구자는 $prof(p_i)$로 표현할 수 있다. 또한, $p(prof_w)$는 w 연구자의 문서들의 집합이며, 이는 전체 문서 집합 P에 반드시 포함된다. 수식 (3)의 θ_f는 인코더 모델의 파라미터를 의미하고, $f(\,)$는 θ_f로 파라미터화 된 인코더 모델을, d는 임베딩 차원 수를 나타낸다. 질의어와 문서 간의 유사도는 수식 (4)의 $sim(\,)$을 통해 산출할 수 있다.

$$P=\{p_1,p_2,\cdots,p_i,\cdots,p_M\},\ Q=\{q_1,q_2,\cdots,q_i,\cdots,q_M\}, \tag{1}$$

$$PROF=\{prof_1,prof_2,\cdots,prof_w,\cdots,prof_W\}\ (W\leqq M). \tag{2}$$

$$v_{q_i},v_{p_j}=f(q_i;\theta_f),f(p_j;\theta_f)\in R^d, \tag{3}$$

$$sim(q_i,p_j)=v_{q_i}^{T}v_{p_j}. \tag{4}$$

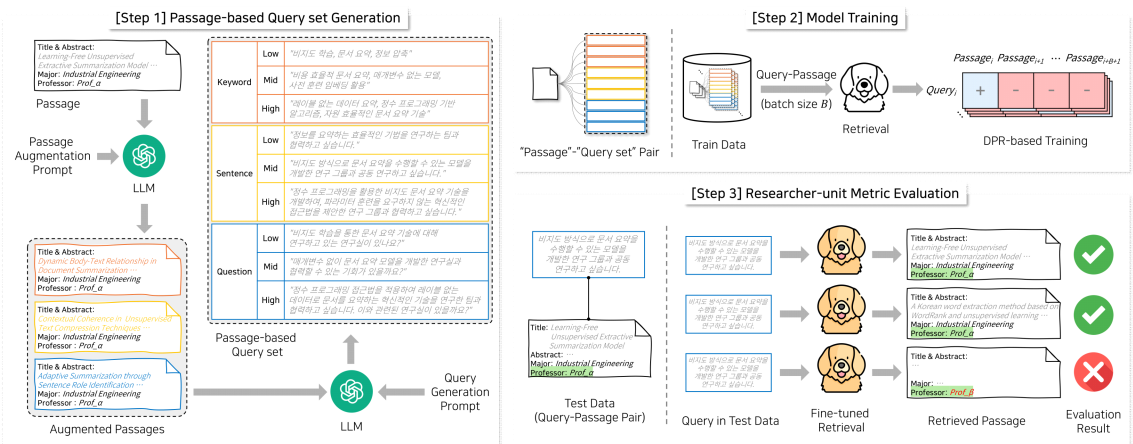

Figure 1. Overall Process for FindCoResearcher

FindCoResearcher의 목적은 질의어와 관련된 문서의 연구자를 효과적으로 탐색하는 것이다. 이를 구현하기 위한 목적 함수는 수식 (5)와 같다.

$$L(\theta_f) = -\sum_{i=1}^{M} \log \frac{\sum_{p_j \in p(prof(p_i))} \exp(sim(q_i, p_j))}{\sum_{p_j \in P} \exp(sim(q_i, p_j))}. \quad (5)$$

이는 질의어 q_i가 정답 연구자 집합에 속한 문서들에 대해 전체 후보 문서보다 상대적으로 더 높은 유사도를 가지도록 만드는 구조이다. 수식 (5)의 분자는 질의어 q_i와 정답 연구자의 문서 집합 $p(prof(p_i))$ 간 유사도를 계산한 뒤, 지수 함수로 변환하여 모두 합한 값이다. 이는 정답으로 간주되는 여러 문서들이 질의어와 얼마나 가깝게 연결되는지를 나타낸다. 분모는 질의어 q_i와 모든 문서 집합 P 간 상대적인 유사도의 총합을 나타낸다. 즉, 정답 연구자의 문서 집합 $p(prof(p_i))$에 포함되지 않은 문서와 질의어 간 유사도가 낮아질수록 분모 대비 분자의 비율이 커져 손실 $L(\theta_f)$이 작아진다. 본 시스템의 목적 함수은 손실 $L(\theta_f)$를 최소화 시키는 방향으로 학습하기 위해 분자의 값이 커지고 분모의 값이 작아지도록 설계되었으며, 이는 각 질의어가 정답 연구실의 문서 집합 간 유사도를 높이며, 이 외 문서 간 유사도를 낮추는 방향으로 유도된다.

3.2 데이터셋 구축 및 모델 학습

FindCoResearcher 구축을 위해 질의어-문서 쌍 형태의 데이터셋이 필요하다. 이에 본 연구에서는 실제 연구 실적 데이터를 수집하여 문서 데이터베이스를 구축한 뒤, 질의어와 문서가 연결된 데이터셋을 생성한다. 이후 학습 데이터셋을 활용하여 검색 모델을 구현한다.

(1) 문서 기반 질의어 집합 생성

먼저 연구자 별 연구 실적에 대한 정보를 수집한다. 이렇게 수집된 문서는 크게 학과명, 제목, 요약의 세 가지 요소로 구성된다. 학과명은 연구의 도메인 정보를 반영하며, 제목은 연구의 핵심 주제를 압축하여 제공한다. 요약은 방법론, 실험 결과 등의 핵심 내용을 포함한다.

이러한 상황에서 문서 기반 질의어 생성 시, 질의어 표현의 다양성을 확보하기 위해 질의어 간 표현 다양성을 고려해야 한다. 실제 검색 환경에서는 하나의 문서가 다양한 사용자 질문에 대한 적절한 답이 될 수 있다. 이러한 질의어 표현의 다양성을 반영하기 위해 데이터셋 구축 시 하나의 문서를 기반으로 여러 개의 질의어를 생성해야 한다. 이 때, 생성된 질의어들이 서로 유사하다면, 모델이 특정 표현 패턴에 특화된 학습으로 인해 실제 사용자가 다른 방식으로 표현한 검색어에는 대응하지 못하게 된다. 따라서 동일 문서에서 파생된 질의어 간에도 충분한 어휘적, 구문적 차별화가 이루어져야 한다.

질의어 스타일 및 구체화 수준 다양화: 질의어의 어휘적 다양성을 높이기 위해 하나의 문서를 기반으로 여러 개의 질의어를 생성하여 질의어 집합-문서 쌍을 구성한다. 이 때, 질의어 집합 내 질의어 간 다양성을 확보하기 위해 실제 사용자의 검색 패턴을 반영한 체계적인 다양화 전략을 사용한다. 구체적으로 3가지 질문 스타일(키워드 기반, 문장형, 질문형)을 설정하여 정보 탐색 시 사용하는 다양한 표현을 반영하였다. 또한, 각 스타일 별로 검색의 구체화 수준(간단, 보통, 구체적)을 달리함으로써 사용자의 다양한 검색 성향을 반영하였다. i번째 문서 p_i에 대해 생성된 질의어 집합 $\hat{q_i}$은 수식 (6)과 같이 표현할 수 있다.

$$\hat{q_i} = (\widehat{q_i^{s,l}} | s \in \{keyword, sentence, question\}, \; l \in \{low, mid, high\}). \quad (6)$$

문서 증강: 제한된 문서 정보를 기반으로 질의어 집합을 생성할 경우, 질의어-문서 간 다양성이 저하된다. 이를 해결하기 위해 원본 문서의 맥락을 유지하되, 도메인, 응용 분야 등의 일부 요소를 변형하여 증강 문서 3개를 생성한다. i번째 문서 p_i를 기반으로 생성한 증강 문서 집합 $\hat{p_i}$은 수식 (7)과 같이 나타낼 수 있다. 이는 제한된 문서 정보의 한계를 극복하는 주요 해결책이며, 증강된 문서를 바탕으로 질의어 집합을 생성한다.

$$\hat{p_i} = (\widehat{p_{i,j}} | j \in \{1,2,3\}). \quad (7)$$

질의어의 다양성을 고려한 문서 기반 질의어 집합 생성 과정은 <Figure 1>의 Step 1과 같다. 먼저 LLM을 이용해 원본 문서를 기반으로 표현 다양성을 높인 3개의 증강 문서를 생성한 뒤, 각 증강 문서마다 3가지 질의어 스타일을 일대일로 지정하고, 증강 문서 별 3가지 구체화 수준을 가지는 질의어를 생성한다. 결론적으로 하나의 원본 문서로부터 다양성을 확보한 9개의 질의어를 생성한다. 이 과정을 모든 문서에 적용하면, 전체 데이터셋을 구축할 수 있으며 그 과정에 대한 의사 코드는 부록 A의 <Table A1>를 통해 확인 가능하다.

(2) 모델 학습

전체 데이터셋이 구축되면 평가용 데이터 일부를 제외하고, 모두 학습 데이터셋으로 사용한다. 검색 모델 학습은 IR 학습에 활용되는 대표적인 방식인 DPR(Karpukhin *et al.*, 2020)을 기반으로 수행되며, 학습 과정은 <Figure 1>의 Step 2와 같다. DPR의 핵심 아이디어는 연결된 질의어 - 문서 쌍은 가깝게, 연결되지 않은 쌍은 멀게 임베딩하도록 파라미터 θ_f를 조정하는 것이다. 이를 위해서 수식 (3)을 활용해 질의어와 문서를 인코더 f로 임베딩하여 차원 d의 고정된 벡터로 변환한다. 이후 수식 (4)를 통해 유사도를 산출하고, 하나의 질의어에 대해 여

러 후보 문서의 점수를 소프트맥스로 정규화한다. 이 때, 연결된 질의어-문서 쌍의 유사도는 높아지도록, 질의어와 전체 문서 간 유사도는 낮아지도록 학습을 설계함으로써 입력 질의어와 유사한 문서가 검색되도록 유도한다. 본 연구의 궁극적인 목표는 연구자 단위 검색 성능 향상이지만, 실제 학습은 문서 단위로 수행된다. 즉, 이상적인 상황은 수식 (5)와 같이 정답 연구자 집합에 속한 문서들을 전체 후보 문서보다 상대적으로 더 높게 평가하도록 만드는 것이지만, 실제 학습 데이터의 입력은 질의어-문서 쌍으로 구성되며, 이러한 학습 방식에 대한 자세한 분석은 5.3절에서 진행한다. 학습은 사전학습된 인코더를 기반으로 수행되며, 미니배치에 포함된 다른 문서들을 연결되지 않은 문서로 구성하는 in-batch negative 방식을 사용하여 계산 효율성을 확보한다. 또한, 각 문서 p_i에 대응되는 질의어 q_i는 질의어 집합 $\hat{q}_i$에서 매 에폭(epoch)마다 무작위로 선택하여 다양한 표현의 질의어 - 문서 매핑을 학습한다.

3.3 연구자 단위 평가 방식

FindCoResearcher의 궁극적인 목표는 적합한 협력 연구자를 탐색하는 것이다. 이에 본 연구에서는 연구자 단위의 검색 성능을 평가하기 위한 방식을 새롭게 도입하고자 한다. 기존 평가 방식과 차이점은 <Figure 2>와 같다.

기존의 문서 단위 평가에서 사용하는 Top-k Accuracy는 수식 (8)을 통해 산출할 수 있다.

$$Top-k\ \ Acc = \frac{1}{T}\sum_{t=1}^{T} I[p_t \in p_{retrieved}^{:k}]. \tag{8}$$

이 때, $I[]$는 지시함수를 나타내며, T는 평가 데이터셋의 개수를 의미한다. 수식 (4)의 $sim(\,)$을 통해 평가용 질의어와 문서 데이터베이스 내 각 문서 간 유사도를 산출했을 때, 유사도 상위 k개의 문서 집합을 $p_{retrieved}^{:k}$로 표현한다. 평가는 평가용 질의어 q_t에 대한 검색 문서 집합 $p_{retrieved}^{:k}$ 중 정답 문서 p_t가 포함되어 있는지 여부로 수행된다. 해당 평가 방식은 협력 연구자 탐색 시, 사용자 의도에 적합한 연구자가 적절히 탐색되어도 일대일로 연결된 정답 문서가 아닌 경우 검색이 잘못되었다고 평가한다. 이를 오답으로 처리하는 것은 협력 연구실을 추천하는 FindCoResearcher의 목적에 맞지 않으므로 문서 단위의 평가 지표가 아닌 연구자 단위의 평가 지표가 필요하다.

연구자 단위의 평가 지표는 검색 문서 집합 $p_{retrieved}^{:k}$가 아닌 검색된 문서의 연구자 집합 $prof_{retrieved}^{k}$를 활용한다. 검색 문서 집합 $p_{retrieved}^{:k}$에서 검색된 문서의 연구자 집합 $prof_{retrieved}^{k}$을 추출하는 과정은 수식 (9)로 표현할 수 있다. 검색 문서 집합 $p_{retrieved}^{:k}$ 내에 동일한 연구실의 문서가 여러 개 존재할 경우 검색된 문서의 연구자 집합 $prof_{retrieved}^{k}$ 의 원소 개수는 k보다 작아질 수 있다. 이를 기반으로 제안하는 연구자 단위의 Top-k Accuracy는 수식 (10)을 통해 산출할 수 있다.

$$prof_{retrieved}^{k} = \{prof(p_i) | p_i \in p_{retrieved}^{:k}\}, \tag{9}$$

$$Top-k\ \ Acc_{prof} = \frac{1}{T}\sum_{t=1}^{T} I[prof(p_t) \in prof_{retrieved}^{k}]. \tag{10}$$

Table 1. Actual Relevance of Retrieved Passages in Researcher-unit Evaluation Metric

	The Number of Pairs (Query-Top-5 Passage)
Relevant Content in Top-5 Passages	52 (74.29%)
Relevant Content not in Top-5 Passages	18 (25.71%)
Total	70

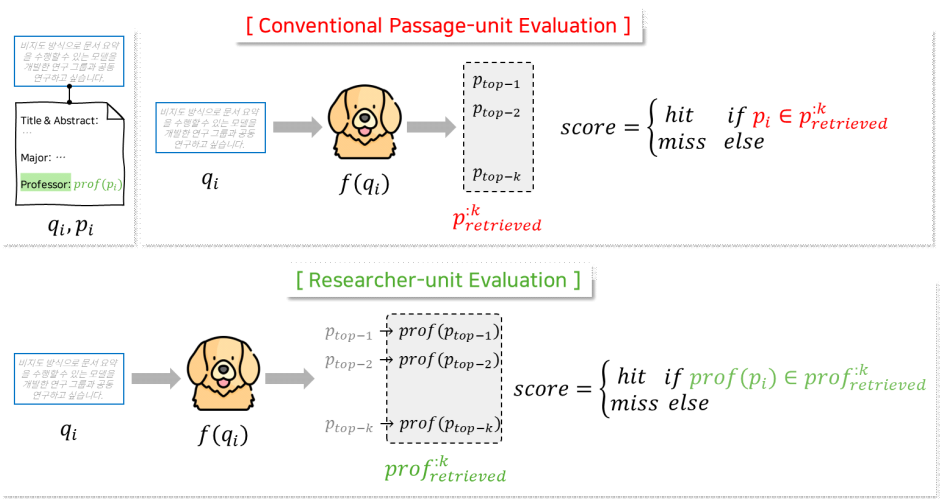

Figure 2. Researcher-unit Evaluation Metric

연구자 단위의 평가는 기존 문서 단위의 평가와 동일하게 평가용 질의어 q_t에 대한 검색 문서 $p^{:k}_{retrieved}$를 추출하되, 검색 문서의 연구자 집합 $prof^{k}_{retrieved}$ 내에 정답 문서의 연구자 $prof(p_t)$가 포함되었는지 여부로 수행된다. 결과적으로 동일 연구자의 여러 문서가 정답 후보로 인정되며 연구자 단위 검색이라는 목표에 부합하는 평가가 가능해진다.

문서 단위 평가 방식의 한계점을 실증적으로 검증하기 위해 추가 분석을 수행하였다. 구체적으로, 문서 단위 평가에서는 오답으로 처리되나 연구자 단위 평가에서는 정답으로 판단되는 케이스들을 대상으로, 검색된 문서들이 실제로 질의어와 관련성이 있는지 정성적 평가를 진행하였다. 평가는 자동화된 정성 평가를 위해 널리 활용되는 G-Eval(Liu *et al.*, 2023) 방식을 차용하였다. GPT-4o(Hurst *et al.*, 2024)를 활용한 LLM-as-a-Judge 방식을 적용하였으며, 각 질의어에 대해 Top-5로 검색된 문서들이 해당 질의어와 의미적으로 관련이 있는지 판단하도록 하였다. <Table 1>는 평가 결과를 보여준다.

전체 평가 데이터에서 문서 단위로는 오답이나 연구자 단위로는 정답인 70개의 "질의어-Top-5 검색 문서" 쌍을 분석한 결과, 52개(74.29%)의 쌍이 검색된 문서가 실제로 질의어와 관련이 있는 것으로 확인되었다. 이는 현재 모델이 사용자의 정보 요구에 부합하는 적절한 연구자의 문서를 검색하고 있음에도 불구하고, 기존의 문서 단위 평가 방식이 이를 오답으로 판정하는 구조적 한계를 지니고 있음을 보여준다. 따라서 협력 연구자 탐색이라는 시스템의 목적을 고려할 때, 연구자 단위 평가 방식이 더욱 적합한 평가 체계임을 알 수 있다.

연구자 단위 평가의 부가적인 지표로 Mean Reciprocal Rank (MRR)를 활용한다. MRR은 정답이 처음 등장하는 순위의 역수를 평균낸 값으로, 상위 순위에서 정답이 출현하는지를 민감하게 평가할 수 있는 지표이다. 연구자 단위의 Top-k MRR은 Accuracy와 다르게 검색 문서의 연구자 집합 내 정답 문서의 연구자가 존재하는 경우에도 순위에 따라 점수 차이를 두어 검색 모델에 대한 정밀한 평가가 가능하다. 연구자 단위의 Top-k MRR은 수식 (11)과 같이 정의된다.

$$Top-k\ \ MRR_{prof} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{rank(prof(p_t), prof^{k}_{retrieved})}. \quad (11)$$

여기서 $rank(\,)$는 검색 문서의 연구자 집합 $prof^{k}_{retrieved}$ 내에 정답 문서의 연구자 $prof(p_t)$가 처음 등장하는 순위를 산출하는 함수이다. 예를 들어, 검색 결과의 3번째 문서가 정답 연구자의 첫 번째 문서라면 $rank(prof(p_t), prof^{k}_{retrieved}) = 3$이 된다.

3.4 사용자 관점의 협력 연구자 탐색 파이프라인

사용자 관점에서 FindCoResearcher은 적절한 협력 '연구자'를 검색하는 도구이며, 연구자 단위의 관련 정보를 습득하는 것이 중요하다. 이러한 사용자 관점을 반영한 파이프라인은 다음과 같다.

먼저 사용자의 질의어에 대해 모델을 활용하여 전체 문서 중 유사도가 높은 상위 k개 문서를 검색한다. 이후 수식 (9)를 활용하여 검색 문서 $p^{:k}_{retrieved}$ 를 연구자 단위로 그룹화한다. 이 때, 사용자에게 상위 z명의 연구자를 추천하기 위해 $k \gg z$로 설정한다.

이후 각 연구자마다 사용자 질의어와 가장 유사한 상위 z'개의 문서 정보를 함께 제공하여 사용자의 의사결정에 효과적인 판단 근거를 제공한다. 이 과정은 수식 (4)의 $sim(\,)$을 활용하여 사용자 질의어에 관련된 문서 $p^{:z'}_{retrieved}$ 를 추출하되, 유사도 산출 대상을 검색된 연구자 별 문서 집합 $p(prof_w | prof_w \in prof^{k}_{retrieved})$으로 제한하여 각 연구자 별 상위 z'의 문서 집합을 산출한다.

이 때, 사용자는 추천받을 협력 연구자 개수 z와 추천받은 협력 연구자 별로 제공받고 싶은 문서 정보 개수 z'를 선택할 수 있으며, 해당 파이프라인 기반으로 구성한 사용자 인터페이스 예시는 부록 B의 <Figure A1>을 통해 확인할 수 있다.

4. 실험

4.1 데이터셋

국내 종합대학의 특정 단과대학 소속 교수 451명의 연구 실적 정보를 수집하여 문서 데이터베이스를 구성하였다. 총 문서는 42,142개를 수집하였으며, 이 중 평가 데이터 확보를 위해 100개의 문서를 구분한다. 질의어 데이터셋은 gpt-4o-mini-2024-07-18 모델을 활용하여 각 문서당 9개의 질의어를 생성하였다.

생성된 데이터셋에 대해 직접 검수를 수행하여 오탈자, 번역 오류 등을 정제하였다. 이후 평가 데이터셋의 품질을 체계적으로 검증하기 위해 LLM-as-a-Judge 방식으로 질의어-문서 관련성과 표현 다양성을 평가하였다. 평가 시, 사용한 모델은 데이터 생성 시 사용한 gpt-4o-mini-2024-07-18보다 매우 우수한 GPT-4o를 사용하였다. 또한, LLM이 생성한 질의어와 연구자가 직접 작성한 질의어를 동일한 기준으로 비교 분석하였으며, 상세한 분석 결과는 부록 C와 <Figure A2>에서 확인할 수 있다.

문서 증강과 질의어 생성 시 사용한 프롬프트는 https://bit.ly/4mfcRvM에서 확인할 수 있으며, 최종적으로 구축한 학습 및 평가 데이터셋의 통계량은 부록 D의 <Table A2>와 같다.

4.2 모델 학습 및 추론

FindCoResearcher의 기반 모델은 MIRACL 벤치마크(Zhang

et al., 2022)에서 높은 한국어 성능을 보이며, MKQA 벤치마크(Longpre *et al.*, 2021)에서 교차 언어 검색 성능이 우수하다고 알려진 BGE-M3(Chen *et al.*, 2024)을 사용하였다. 자세한 학습 환경은 부록 E의 <Table A3>에서 확인할 수 있다. 추론 과정에서는 각 입력 질의어에 대한 유사도 상위 k개의 문서를 산출해야 하므로, 효율적이고 빠른 검색을 위해 FAISS(Douze *et al.*, 2024)를 사용한다.

4.3 비교 방법론

비교 방법론은 임베딩 추출 방식에 따라 크게 희소 검색 모델과 밀집 검색 모델로 나눌 수 있다. 희소 검색 모델은 키워드 매칭 기반 검색 방식 중 가장 대표적인 BM25(Robertson *et al.*, 1994)를 비교 대상으로 선정하였다. 주요 비교 대상인 밀집 검색 모델은 한국어 성능이 뛰어나며, 검색 과업 수행 능력이 좋은 모델 중, 제안하는 모델의 파라미터 수(569M)와 같거나 유사한 모델인 E5-multilingual-Large(Wang *et al.*, 2024), KURE-v1(Jang *et al.*, 2024), BGE-M3(Chen *et al.*, 2024)를 비교 대상으로 선정하였다.

5. 실험 결과 및 분석

5.1 주요 실험 결과

주요 실험 결과는 <Table 2>와 같다. 희소 검색 모델과 비교했을 때, 밀집 검색 모델이 우월한 협력 연구자 검색 성능을 보였다. 이는 전통적인 키워드 매칭 기반의 희소 검색 방식이 연구자 검색 성능에 한계를 가지고 있음을 시사한다. 이러한 성능 제한은 사용자가 동일한 연구 주제에 대해서도 다양한 표현으로 질의어를 작성하는 질의어-문서 간 어휘적 다양성에 기인하는 것으로 분석된다. 또한, 밀집 검색 모델 간 성능을 비교했을 때, 모두 성능이 유사함을 알 수 있다. 이를 통해 FindCoResearcher 구축 시, 기반 모델에 따른 차이가 크지 않음을 알 수 있다.최종적으로 본 연구에서 제안하는 FindCoResearcher이 연구자 단위 Top-5 Accuracy 기준 0.4573을 달성하여 모든 비교 방법론 대비 우수한 성능을 달성하였다. 이는 실제 사용자가 연구자 추천 개수 z를 5로 설정했을 때, FindCoResearcher이 약 45.73%의 확률로 적절한 협력 연구자를 추천함을 의미한다. 따라서 제안 방법론 적용 시, 사용자 질의어에 적합한 연구자를 추천하는 우수한 모델을 구축할 수 있다. 이러한 경향은 연구자 단위 Top-k Accuracy와 MRR이 동일하게 나타났으며, 밀집 모델 사용 및 제안 방법론 적용 시, 더 상위 순위에 정답이 출현함을 알 수 있다.

5.2 문서 기반 질의어 집합 생성 시, 증강 문서 사용 여부에 따른 표현 다양성 변화

본 연구에서는 생성 질의어 표현의 다양성 확보를 위해 증강 문서 기반 질의어 집합 생성 방식을 제안한다. 정보가 제한된 단일 문서를 기반으로 여러 개의 질의어를 생성하는 방식은 질의어 다양성 측면의 한계를 가질 수 있기 때문에 본 연구에서는 문서 증강 절차를 추가하였으며, 해당 절차에 대한 효용성을 3가지 측면에서 분석하고자 한다. 3.2장에서 서술하였듯, 질의어 표현의 다양성을 확보하기 위해 질의어 간 표현 다양성을 고려해야 한다. 이에 (1)에서 문서 증강 절차에 따른 2가지 다양성의 변화를 분석하고, (2)에서는 생성 질의어 예시를 정성적으로 비교 분석한다.

다양한 질의어 스타일 및 구체화 수준을 반영하여 문서 기반 질의어 집합을 생성할 때, Original은 문서 증강 없이 원본 문서만을 사용한 경우를 의미하고, Augmented (3)는 3개의 증강 문서를 사용한 경우, Augmented (9)는 9개의 증강 문서를 사용한 경우를 의미한다. 추가 실험에서는 증강 문서 개수의 영향을 심층적으로 분석하기 위해 9개 증강 문서 조건을 추가하였으며, 이는 3개 증강 문서 생성 과정을 3회 반복하여 구현하였다.

(1) 질의어 간 표현 다양성 변화 분석

각 문서로부터 생성된 질의어 간 표현 사용의 중복 여부는

Table 2. Main Result. Comparison of Researcher-unit Top-k Accuracy, MRR (k=1, 5, 20). The best performance is shown in bold and underlined

Model	$Top-kAcc_{prof}$			$Top-kMRR_{prof}$		
	k=1	k=5	k=20	k=1	k=5	k=20
Sparse						
BM25	0.0246	0.0947	0.1801	0.0246	0.0508	0.0589
Dense						
E5-multilingual-Large	0.1544	0.3158	0.5380	0.1544	0.2098	0.2312
KURE-v1	0.1485	0.3415	0.5602	0.1485	0.2197	0.2425
BGE-M3	0.1649	0.3614	0.5637	0.1649	0.2342	0.2550
Ours	**0.1988**	**0.4573**	**0.6339**	**0.1988**	**0.2894**	**0.3082**

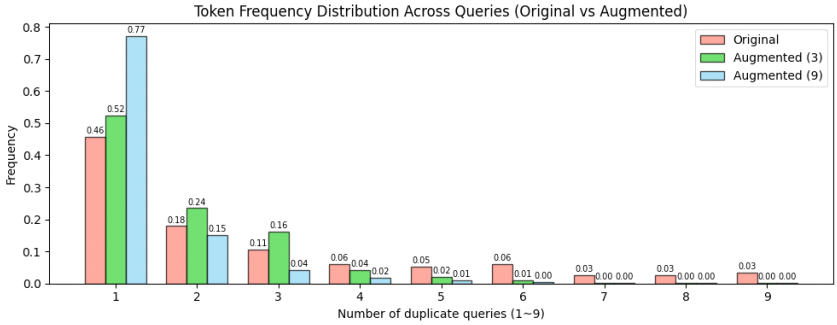

Figure 3. Comparison Distributions of Word Duplication between Queries

질의어의 다양성을 확인할 수 있는 지표 중 하나이다. 이를 측정하기 위해 각 문서에 대한 총 9개의 질의어 중 각 단어들이 사용된 질의어의 개수를 측정하여 '중복 표현 등장 빈도'를 분석하였다. <Figure 3>의 x축은 각 단어가 몇 개의 질의어에 동시에 등장했는지를 나타내며, y축은 해당 조건을 만족하는 단어의 상대적 빈도를 의미한다. 이 때, 1개에서 3개 정도로 비교적 적은 개수의 질의어에서 중복하여 등장하는 표현이 많을수록 질의어 표현의 다양성이 높다고 할 수 있다. 반대로 비교적 많은 개수의 질의어에서 중복하여 등장하는 표현이 많을수록 질의어 표현의 다양성이 낮으며, 특히 7개 이상의 질의어에서 중복하여 등장하는 표현이 많은 것은 대부분의 질의어에서 같은 표현이 사용된다는 의미이므로 어휘 다양성이 매우 낮음을 의미한다.

측정 결과, 증강 문서를 적용한 경우 Original 대비 적은 개수의 질의어에서 등장하는 표현이 증가함을 알 수 있다. 특히 Augmented(9)의 경우, 중복 쿼리의 개수가 1개인 비율이 0.77에 달하며, 질의어 간 중복이 없는 고유 토큰의 비율이 압도적으로 높음을 확인할 수 있다. 반대로, 다수의 질의어에서 등장하는 표현이 감소되며, 특히 7개 이상의 질의어에서 등장하는 표현은 Augmented(3)와 Augmented(9) 모두에서 거의 완전히 제거됨을 확인하였다. 이를 통해 문서 증강을 포함한 데이터셋 생성 방식이 질의어 간 표현 다양성을 향상시키며, 증강 문서 개수가 증가할수록 이러한 효과가 더욱 강화됨을 확인할 수 있다.

(2) 문서 증강 방법론의 표현 다양화 효과에 대한 정성적 분석

문서 증강 방법 도입이 질의어의 어휘 다양화에 기여하는지 확인하기 위해, Original 방식과 Augmented 방식으로 생성된 질의어를 정성적으로 비교분석 하였다. <Table 3>에서 확인할 수 있듯, Original 방식은 기존 문서에 포함된 키워드를 그대로 사용하여 획일화된 질의어가 생성되며, 질의어 간의 다양성도 낮은 것을 볼 수 있다. 반면 Augmented 방식은 '법률 문서 및 과학 논문의 요약' 같은 기존 연구와 연결되는 새로운 분야가 생성되거나, '정보를 요약하는 효율적인 기법'이라는 단순한 키워드가 '요약 과정에서 맥락의 일관성 유지'로 풍부해지는 현상을 확인할 수 있다.

즉, 기존 문서에 포함된 키워드를 거의 사용하지 않지만, 여

Table 3. Comparison Diversity of Passage-based query. Example queries at middle specificity level generated from a single passage. Words overlapping with the passage are underlined.

Passage	... we propose a model called Learning Free Integer Programming Summarizer (LFIP-SUM), which is an unsupervised extractive summarization model ...	
Query (Specificity level: Mid)		
Original	Augmented (3)	Augmented (9)
"비용 효율적 문서 요약, 매개변수 없는 모델, 사전 훈련 임베딩 활용" "정보를 요약하는 효율적인 기법을 연구하는 팀과 협력하고 싶습니다." "정수 프로그래밍 접근법을 적용하여 레이블이 없는 데이터로 문서를 요약하는 혁신적인 기술을 연구한 팀과 협력하고 싶습니다. 이와 관련된 연구실이 있을까요?"	"상황적 단서 활용, 동적 추출 기준, 문서 구조적 특성" "문서의 요약 과정에서 맥락의 일관성을 유지하는 방법을 연구할 연구팀을 찾고 있습니다." "법률 문서 및 과학 논문의 요약 시 중요한 정보의 선정을 위한 기법을 개발하는 데 관심이 큰데, 이에 대한 공동 연구 협력에 참여하고 싶으신가요?"	"자기 지도 학습, 맥락 요약, 대규모 텍스트" "다양한 멀티미디어 데이터를 요약하기 위한 맥락 인식 압축 기법을 연구하고 있습니다. 협력할 연구팀을 찾습니다." "심층 의미 분석을 통한 비지도 문장 선택에 대해 함께 연구하시겠습니까?"

전히 문서와 관련된 질의어가 생성되는 것을 확인할 수 있다. 따라서 증강 문서를 기반으로 질의어를 생성하는 방식이 문서 정보를 반영하면서 풍부한 질의어를 생성함을 확인하였다.

(3) 문서 증강 기법 및 증강 문서 개수 설정

결과적으로 증강 기법을 적용하는 경우 질의어 간 토큰 중복이 감소하며, 정성적으로 확인했을 때도 표현이 다양화됨을 확인하였다. 또한, 증강 문서 개수가 늘어날수록 질의어 표현 공간이 확장되고, 다양성이 더욱 증대됨을 확인하였다.

다만, 증강 문서의 개수를 늘릴 경우, LLM의 생성이 불안정해지는 문제가 발생한다. 실제로 단일 프롬프트로 9개의 증강 문서를 동시에 요청할 경우, 9개를 증강하지 못하고 생성을 종료하는 경우가 대다수였으며, 이는 실제 적용 시, 비효율적인 토큰 생성 반복 및 생성 비용 증가로 이어질 수 있다. 또한, 데이터 생성 시 필요한 토큰 수가 증가한다. 전체 문서의 개수가 M개이고, 증강 문서 1개당 발생하는 평균 생성 토큰 수를 tok_p , 생성 쿼리 1개당 발생하는 평균 토큰 수를 tok_q 라고 할 때, 3개의 증강 문서를 사용하는 경우 총 $M(3tok_p+9tok_q)$ 개의 토큰이 생성되며, 9개의 증강 문서를 사용하는 경우 총 $M(9tok_p+9tok_q)$ 개의 토큰이 생성된다. 두 토큰 수 간의 격차는 $M(6tok_p)$이며, 이는 전체 문서 데이터셋이 많을수록, 기준 문서의 길이가 길수록 극대화된다.

결론적으로 본 연구에서 제안하는 증강 문서 기반 질의어 데이터셋 구축 방식이 질의어 다양성을 향상시키며, 데이터 품질 향상에 기여하였음을 확인하였다. 해당 방식으로 생성한 데이터셋의 예시는 부록 F의 <Table A4>을 통해 확인할 수 있다.

5.3 연구자 단위 문서 유사도 향상을 위한 학습 방법

본 연구에서는 연구자 단위로 평가하면서 문서 단위로 학습을 수행하였다. 이러한 접근 방식의 타당성을 검증하기 위해 학습 전 후의 모델을 활용하여 문서 간 유사도를 산출한다. 학습 전 모델은 사전학습된 BGE-M3을 학습 후 모델은 본 연구에서 제안하는 FindCoResearcher을 사용하였으며, 동일 연구자와 다른 연구자 간의 문서 유사도를 구분하여 측정함으로써 연구자 단위 문서 유사도에 대한 분포 변화를 확인한다.

<Figure 4>의 상단이 학습 전 모델, 하단이 학습 후 모델을 사용하여 측정한 유사도 분포이다. 문서 단위 학습은 동일 연구자가 다양한 분야의 연구를 수행하는 경우, 동일 연구자 내 문서간 유사도를 낮아지게 할 수 있다. 하지만 측정 결과, 동일 연구자와 다른 연구자의 문서 간 유사도 분포의 차이가 학습 전 대비 학습 후에 더욱 극대화됨을 알 수 있다. 이는 문서 단위의 대조 학습을 수행하였음에도 자연스럽게 동일 연구자 내 문서 간 유사도는 높이고 서로 다른 연구자 간 문서들의 유사

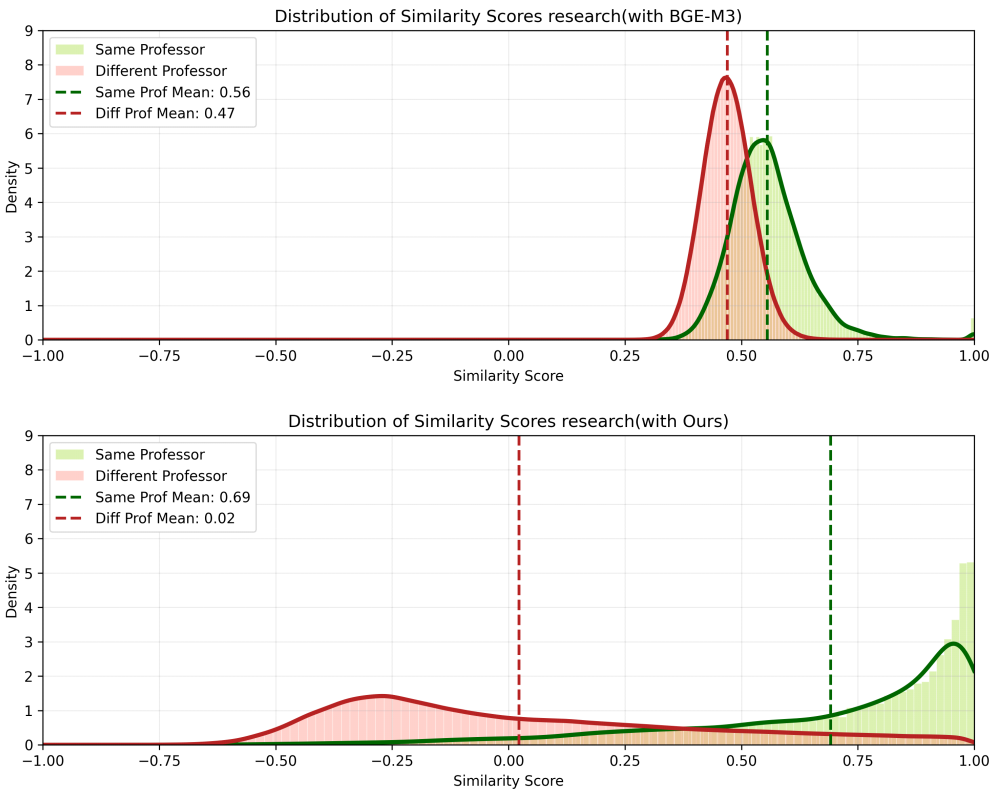

Figure 4. Comparison Distribution of Similarity Score. Similarity score distributions extracted using BGE-M3 (top) and FindCoResearcher (bottom).

도는 낮추는 방향으로 학습됨을 의미한다. 이는 평균적으로 동일 연구자가 유사한 주제의 연구를 수행한다는 현실적인 경향을 반영한 것으로 해석할 수 있으며, 특정 단과대학을 대상으로 수행하는 본 연구의 특성상, 이러한 경향이 두드러짐을 알 수 있다. 결과적으로 제안 방법론의 단순화된 학습 방식으로도 충분히 연구자 간 구분이 가능함을 알 수 있다.

6. 결 론

본 연구에서는 협력 연구자 탐색의 효율성을 향상시키기 위한 정보 검색 기반 협력 연구자 추천 시스템인 FindCoResearcher을 제안하였다. FindCoResearcher을 통해 기존의 인적 네트워크와 수동 탐색에 의존하던 협력 연구자 탐색 과정의 비효율성을 해결하며, 연구자들이 자신의 연구 주제에 대해 적합한 협력 파트너를 효과적으로 찾을 수 있도록 지원한다.

본 연구는 세 가지의 기여점이 존재한다. 첫째, 제한된 정보 환경에서도 풍부한 표현의 질의어를 생성할 수 있는 체계적인 방법론을 제시하였다. 문서 증강과 질의어 스타일 및 구체화 수준 다양화라는 두 가지의 순차적인 전략을 통해 표현 다양성을 크게 향상시켰다. 이는 5.2장의 추가 분석을 통해 제안 방법론이 질의어 간 다양성 및 질의어-문서 간 다양성을 향상시킴을 확인하였다. 둘째, 기존의 문서 단위 평가 방식의 한계를 극복하기 위해 연구자 단위 평가 방식을 새롭게 도입하였다. 이 방식을 통해 '연구자' 라는 정보를 종합적으로 고려할 수 있게 되어 목적에 맞는 정확한 시스템 평가가 가능해졌다. 셋째, 기존 공개 모델들과 비교하여 연구자 검색 성능이 개선된 모델을 개발하였다. 사전학습된 모델을 활용하여 DPR 기반 훈련을 수행한 결과 Top-1 Accuracy 0.1988, Top-5 Accuracy 0.4573을 달성하여 기존 공개 모델보다 우수한 성능을 보였다.

본 연구는 다음과 같은 한계점이 존재한다. 첫째, 문서 데이터베이스가 특정 단과대학으로 제한되어 다양한 학문 분야와 대학으로의 확장 가능성을 충분히 검증하지 못하였다. 둘째, 충분한 연구 이력을 축적하지 못한 신진 연구자들이 검색에서 상대적으로 불리한 상황에 놓인다. 정보 검색 모델의 특성상, 검색 대상이 되는 문서가 많을수록 사용자의 다양한 검색 표현과 매칭될 가능성이 높아진다. 이는 신진 연구자들은 상대적으로 적은 수의 문서만을 보유하고 있어 검색 성능에서 구조적으로 불리할 수밖에 없다.

향후 연구 방향은 다음과 같다. 첫째, 더 많은 대학 및 연구기관의 데이터를 수집하여 시스템의 범용성과 확장성을 검증할 필요가 있다. 특히, 대상 학과가 확장되면서 다양한 분야의 문서가 추가될 때에도 검색 성능이 강건하게 유지될 수 있는 방안을 모색해야 한다. 둘째, 신진 연구자의 문서 또한 동등하게 검색될 수 있도록 학습 방법론과 데이터 구축 전략을 개발해야 한다. 이를 통해 연구 경력이 짧은 연구자들의 가시성을 높이고, 협력 연구 참여 기회를 확대할 수 있을 것이다.

참고문헌

Abass, O. A., and Arowolo, O. A. (2017), Information retrieval models, techniques and applications, *International Research Journal of Advanced Engineering and Science*, **2**(2), 197-202.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... and Kivlichan, I. (2024), Gpt-4o system card, ArXiv, abs/2410.21276.

Bacciu, A., Palumbo, E., Damianou, A., Tonellotto, N., and Silvestri, F. (2024), Generating query recommendations via LLMs, ArXiv, abs/2405.19749.

Bascur, J. P., Verberne, S., van Eck, N. J., and Waltman, L. (2022), Academic information retrieval using citation clusters: in-depth evaluation based on systematic reviews, *Scientometrics*, **128**, 2895-2921.

Beltagy, I., Lo, K., and Cohan, A. (2019), SciBERT: A Pretrained Language Model for Scientific Text, *Conference on Empirical Methods in Natural Language Processing*.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024), BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, *Annual Meeting of the Association for Computational Linguistics*.

Cleverdon, C. (1967), The Cranfield Tests on Index Language Devices, *Aslib Proceedings*, **19**(6), 173-194.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. (2024), The Faiss library, ArXiv, abs/2401.08281.

Hambarde, K. A. and Proenca, H. (2023), Information retrieval: Recent advances and beyond, *IEEE Access*, **11**, 76581-76604.

Harman, D. (1992), Evaluation issues in information retrieval, *Information Processing & Management*, **28**(4), 439–440.

Iyer, G., Dziugaite, G. K., and Rolnick, D. (2024), Linear weight interpolation leads to transient performance gains Transactions on Machine Learning Research.

Jang, Y., Son, J., Park, C., Choi, S., Lee, B., Lee, T., and Lim, H. (2024), KoE5: A New Dataset and Model for Improving Korean Embedding Performance, In *Annual Conference on Human and Language Technology*, Human and Language Technology, 239-244.

Kang, S., Jin, B., Kweon, W., Zhang, Y., Lee, D., Han, J., and Yu, H. (2025), Improving scientific document retrieval with concept coverage-based query set generation, *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, 895-904.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-T. (2020), Dense passage retrieval for open-domain question answering, ArXiv, abs/2004.04906.

Kim, M. and Baek, S. (2025), Syntriever: How to train your retriever with synthetic data from LLMs, *North American Chapter of the Association for Computational Linguistics*.

Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J.R., Hui, K., Boratko, M., Kapadia, R., Ding, W., Luan, Y., Duddu, S.M., Abrego, G.H., Shi, W., Gupta, N., Kusupati, A., Jain, P., Jonnalagadda, S.R., Chang, M., and Naim, I. (2024), Gecko: Versatile text embeddings distilled from large language models,

ArXiv, abs/2403.20327.

Lim, S., Kim, M., and Lee, J. (2019), KorQuAD1.0: Korean QA Dataset for Machine Reading Comprehension, ArXiv, abs/1909.07005.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023), G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, *Conference on Empirical Methods in Natural Language Processing*.

Longpre, S., Lu, Y., and Daiber, J. (2021), MKQA: A linguistically diverse benchmark for multilingual open domain question answering, *Transactions of the Association for Computational Linguistics*, **9**, 1389-1406.

Nadkarni, P. M. (2002), An introduction to information retrieval: applications in genomics, *The Pharmacogenomics Journal*, **2**(2), 96-102.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1994), Okapi at TREC-3, In *Proceedings of the Third Text REtrieval Conference*.

Sinha, A., Mall, P. R., and Roy, D. (2024), Exploring the nexus between retrievability and query generation strategies, ArXiv, abs/2404.09473.

Wang, J. A., Wang, K., Wang, X., Naidu, P., Bergen, L., and Paturi, R. (2023), Scientific document retrieval using multi-level aspect-based queries, *Advances in Neural Information Processing Systems*, **36**, 38404-38419.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024), Multilingual E5 Text Embeddings: A Technical Report, ArXiv, abs/2402.05672.

Yu, L., Miao, J., Sun, X., Chen, J., Hauptmann, A. G., Dai, H., and Wei, W. (2023), DocumentNet: Bridging the Data Gap in Document Pre-Training, *Conference on Empirical Methods in Natural Language Processing*.

Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., and Lin, J. (2022), Making a MIRACL: Multilingual Information Retrieval Across a continuum of languages, ArXiv, abs/2210.09984.

<부록>

A. 증강 문서 기반 질의어 집합 구축 알고리즘

Table A1. Pseudo Code about Passage-based Query set Generation

Input:
- Passage database $P=\{p_1,p_2,...,p_M\}$
- Prompt for augmenting passage $prompt_{aug}(p)$
- Prompt for generating query-set by query styles $prompt_{gen^{s,l}}(p)$
- Query styles, specificity level $s\in\{keyword,sentence,question\}$, $l\in\{high,mid,low\}$

Output:
- Passage-based query-set database $Q=\{\hat{q_1},\hat{q_2},...,\hat{q_M}\}$

for p_i in P :

Augment Passage

$\widehat{p_{i,1}},\widehat{p_{i,2}},\widehat{p_{i,3}}=LLM(prompt_{aug}(p_i))$

Generate Query-set using augmented passages

$prompt_{gen}=prompt_{gen^{keyword,l}}(\widehat{p_{i,1}})\bigcup prompt_{gen^{sentence,l}}(\widehat{p_{i,2}})\bigcup prompt_{gen^{question,l}}(\widehat{p_{i,3}})$

$\hat{q_i}=LLM(prompt_{gen})$

Get Passage-based Query set

$Add\ \hat{q_i}\ to\ Q$

B. 사용자 인터페이스 예시

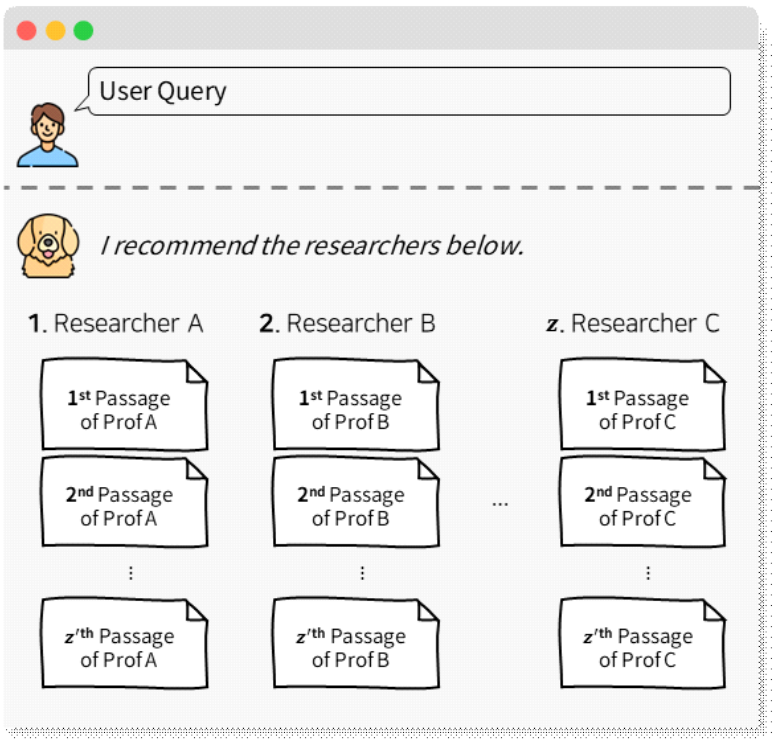

Figure A1. User Interface. When users input a query, they receive recommendations for relevant researchers along with research records for each researcher. Users can customize parameters z and z' (z: the number of researchers to be recommended, z': the number of passage information provided for each researcher)

C. 평가 데이터셋 품질 검증

평가 데이터셋의 품질을 체계적으로 검증하기 위해 (1) 표현 다양성(expression diversity)과 (2) 질의어-문서 관련성(query-document relevance) 두 가지 측면에서 정성적 평가를 수행하였다. 각 항목은 1~5점 척도로 평가하였으며, LLM이 생성한 질의어 기반 평가 데이터셋(llm_generated)과 연구자가 직접 작성한 질의어(human_generated)를 비교·분석하였다.

human_generated 데이터셋은 각 문서와 관련된 전공을 이수했거나 연구 중인 6명의 연구자에게 요청하여 구축하였다. 연구자들은 문서별로 1~9개의 질의어를 작성하였으며, 최종적으로 총 62개 문서로부터 142개의 질의어를 수집하였다. <Figure A2>는 human_generated와 llm_generated 간의 품질 점수 분포를 보여준다.

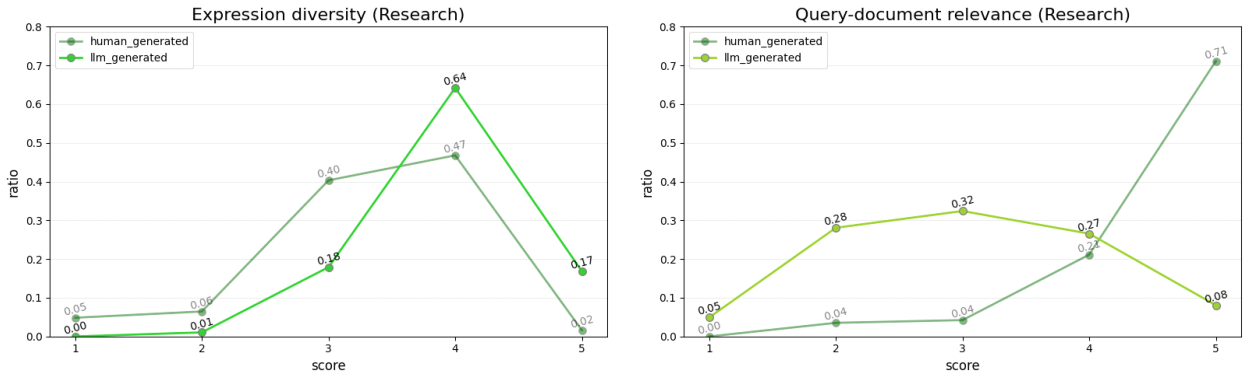

Figure A2. Quality Evaluation of human_generated vs llm_generated using LLM-as-a-Judge

평가 결과, 표현 다양성 측면에서는 두 데이터셋이 전반적으로 유사한 분포를 보였으며, human_generated는 평균 3.43점, llm_generated는 평균 3.97점으로 모두 다양한 표현 확보에 성공한 것으로 나타났다.

반면, 질의어-문서 관련성에서는 뚜렷한 차이가 관찰되었다. human_generated는 평균 4.60점으로 주로 4~5점의 높은 점수대에 분포한 반면, llm_generated는 평균 3.05점으로 3점 부근에 집중되는 양상을 보였다. 이는 연구자가 직접 작성한 질의어가 연구 내용과 더욱 밀접하게 연결되어 있음을 시사한다. 그러나 이러한 현상은 전문 분야 지식을 가진 연구자가 작성했기 때문에 가능한 결과로 해석된다. 실제로 본 연구의 주요 사용자 집단 중, 산학 협력 기업은 해당 연구에서 사용되는 전문 개념에 익숙하지 않을 수 있으며, 오히려 전문 용어가 포함되지 않은 포괄적이고 일반적인 검색어를 사용하는 것이 보다 현실적인 상황이라고 판단된다.

결론적으로, LLM 기반 데이터 생성은 다음과 같은 실용적 장점을 가진다. 먼저 각 문서에 대한 질의어를 단기간에 생성할 수 있는 대규모 데이터 구축의 효율성을 보인다. 그리고 체계적인 스타일 및 구체화 수준 적용으로 균일한 품질 유지할 수 있으며, 평균 3점 이상의 관련성 점수로 학습 및 평가에 적합한 충분한 품질 수준을 갖춘 현실적인 데이터를 생성할 수 있다. 따라서 LLM을 활용한 평가 데이터셋은 일부 분포 차이에도 불구하고 본 연구의 목적인 협력 연구자 추천 시스템 구축 및 평가에 충분히 타당하다고 판단하였다.

D. 데이터셋 정보

Table A2. Dataset Information. For training, queries corresponding to each passage are randomly selected at each epoch. For testing, evaluation is conducted on all queries corresponding to each passage.

Dataset	Train		Test	
	Passage	Query	Passage	Query
	42,042	42,042 * 9	95	95 * 9

E. 학습 환경

Table A3. Training Configurations (GPU: NVIDIA V100 32GB × 1)

Epoch	20	Batch Size	128
Micro Batch Size	4	Gradient Cache Size	32
Loss Temperature	1	LR Schedular	Linear
Learning Rate	5e-5	Warmup Ratio	0.1
Optimizer	AdamW	Beta 1, Beta 2, Epsilon	0.9, 0.999, 1e-08

F. 증강 문서 기반 질의어 집합 데이터셋 예시

Table A4. Example of Passage-based Query Set

Passage		
"Learning-Free Unsupervised Extractive Summarization Model Text summarization is an information condensation technique that abbreviates a source document to a few representative sentences with the intention to create a coherent summary containing relevant information of source corpora. This promising subject has been rapidly developed since the advent of deep learning. However, summarization models based on deep neural network have several critical shortcomings. First, a large amount of labeled training data is necessary. This problem is standard for low-resource languages in which publicly available labeled data do not exist. In addition, a significant amount of computational ability is required to train neural models with enormous network parameters. In this study, we propose a model called Learning Free Integer Programming Summarizer (LFIP-SUM), which is an unsupervised extractive summarization model. The advantage of our approach is that parameter training is unnecessary because the model does not require any labeled training data. To achieve this, we formulate an integer programming problem based on pre-trained sentence embedding vectors. We also use principal component analysis to automatically determine the number of sentences to be extracted and to evaluate the importance of each sentence. Experimental results demonstrate that the proposed model exhibits generally acceptable performance compared with deep learning summarization models although it does not learn any parameters during the model construction process 산업공학과"		
Query		
Style	Specificity level	Example
Keyword	Low	"비지도 학습, 문서 요약, 정보 압축"
	Mid	"비용 효율적 문서 요약, 매개변수 없는 모델, 사전 훈련 임베딩 활용"
	High	"레이블 없는 데이터 요약, 정수 프로그래밍 기반 알고리즘, 자원 효율적인 문서 요약 기술"
Sentence	Low	"정보를 요약하는 효율적인 기법을 연구하는 팀과 협력하고 싶습니다."
	Mid	"비지도 방식으로 문서 요약을 수행할 수 있는 모델을 개발한 연구 그룹과 공동 연구하고 싶습니다."
	High	"정수 프로그래밍을 활용한 비지도 문서 요약 기술을 개발하여, 파라미터 훈련을 요구하지 않는 혁신적인 접근법을 제안한 연구 그룹과 협력하고 싶습니다."
Question	Low	"비지도 학습을 통한 문서 요약 기술에 대해 연구하고 있는 연구실이 있나요?"
	Mid	"매개변수 없이 문서 요약 모델을 개발한 연구실과 협력할 수 있는 기회가 있을까요?"
	High	"정수 프로그래밍 접근법을 적용하여 레이블 없는 데이터로 문서를 요약하는 혁신적인 기술을 연구한 팀과 협력하고 싶습니다. 이와 관련된 연구실이 있을까요?"

저자소개

성시열 : 인하대학교 산업경영공학과에서 2024년 학사학위를 취득하고 고려대학교에서 산업경영공학과 석사과정에 재학 중이다. 연구분야는 Information Retrieval, Anomaly Detection, Vision-Langauge이다.

김재희 : 성균관대학교 소비자학과에서 2022년 학사, 고려대학교 산업경영공학과에서 2024년 석사학위를 취득하고 서울대학교에서 산업공학과 박사과정에 재학 중이다. 연구분야는 Information Retrieval, Question Answering이다.

천재원 : 고려대학교 미디어학부에서 2023년 학사학위를 취득하고 고려대학교에서 산업경영공학과 석사과정에 재학중이다.

연구분야는 LLM, Efficient Transformer, Inference Acceleration 이다.

손준영 : 인하대학교 산업경영공학과에서 2024년 학사학위를 취득하고 고려대학교에서 산업경영공학과 석사과정에 재학 중이다. 연구분야는 Computer Vision, Anomaly Detection, Vision-Langauge이다.

강필성 : 서울대학교 산업공학과에서 2003년 학사, 2010년 박사 학위를 취득하였다. 이후 현대카드 과장, 서울과학기술대학교 조교수, 고려대학교 정교수로 근무하였으며, 현재는 서울대학교 산업공학부 부교수로 재직 중이다. 연구 분야는 정형 및 비정형 데이터를 활용한 데이터마이닝 및 기계학습 알고리즘 개발 및 제조/IT/공공분야 응용이다.