학술 문헌 기반 전문기관 추천 시스템 설계

Design of a Professional Institution Recommendation System Based on Academic Documents

정상준*, 송상호*, 김종훈*, 오영호*,
Retiti Diop Emane Christopher**, 장준혁**, 김상혁**,
이서희**, 전종우**, 최도진***, 유재수*†
충북대학교 정보통신공학부*,
충북대학교 빅데이터 협동과정**,
창원대학교 컴퓨터공학과***

Sangjun Chung*, Sangho Song*, Jonghun Kim*,
Youngho Oh**, Retiti Diop Emane Christopher**,
Junhyuck Jang**, Sanghyeuk Kim**, Seoheui Lee**,
Jongwoo Jeon**, Dojin Choi***, Jaesoo Yoo*†
Chungbuk National University*,
Changwon National University***

요약

연구자의 논문, 특허, 보고서 등과 같은 연구 결과물들은 전문가를 선정하기 위한 객관적인 주요 지표로써 사용된다. 연구자의 소속 또한 전문가 선정에 중요한 지표로써 사용될 수 있다. 하지만, 다양한 학술 분야에서 분야별 전문기관에 대한 객관적인 평가가 어려워, 신뢰성 있는 전문기관 추천에는 한계가 있다. 본 논문에서는 학술 문헌 기반 전문기관 추천 시스템을 제안한다. 온라인 학술 사이트에서 데이터를 수집하여, 전문기관 랭킹을 계산하고, LDA 모델을 이용한 주요 키워드와 연구 실적에 대한 결과를 함께 제공하여, 전문가 선정 시 의사결정에 도움을 주고자 한다.

I. 서론

국내외에서 다양한 분야의 연구가 활발히 진행됨에 따라, 해당 연구에 도움을 줄 수 있는 전문가의 필요성은 나날이 커지고 있다. R&D 과제에 대한 심사위원을 위촉하거나, 특정 분야의 전문지식에 관한 자문을 구하고자 할 때, 해당 분야의 전문가를 필요로 하는 경우들이 많이 발생한다. 연구자의 논문, 특허, 보고서 등과 같은 연구 결과물들은 연구자의 전문성을 판단하는데 중요한 지표로써 사용된다.

전문기관이란 연구 개발 사업에 대한 기획, 관리, 평가 및 활용 등의 업무를 진행하기 위한 기관을 뜻하며, 연구기관과 기업 연구소, 대학교 등이 이에 포함된다. 연구자의 소속 또한 해당 연구자의 전문성을 판단하는데, 중요한 지표가 된다. 하지만, 학술 분야는 다양하고, 분야별 전문기관에 대한 객관적인 평가가 어려워, 신뢰성 있는 전문기관 추천에는 한계가 있다.

† 교신저자 : yjs@cbnu.ac.kr
이 (성과)는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2022R1A2B5B02002456)과 중소벤처기업부 '산업전문인력역량강화사업'의 재원으로 한국산학연협회(AURI)의 지원(2021년 기업연계형연구개발인력양성사업, 과제번호:S3047889)과 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성(Grand ICT연구센터) 사업의 연구결과로 수행되었음 (IITP-2022-2020-0-01462)

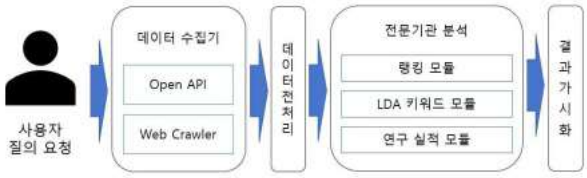
본 논문에서는 학술 문헌 데이터를 기반으로 전문기관을 추천하는 시스템을 제안한다. 온라인 학술 사이트에서 제공하는 데이터를 추출하여, 사용자에게 전문가를 선정하는데, 주요 지표로써 작용할 수 있는 전문기관에 대한 랭킹 정보를 제공한다. 이뿐만 아니라, 연구 실적과 LDA 모델을 이용하여, 질의 결과에 해당되는 전문기관의 주요 키워드 등의 정보들을 함께 제공한다.

Ⅱ. 학술 문헌 기반 전문기관 추천 시스템 설계

본 논문에서는 학술 문헌 기반으로 전문기관 추천 시스템을 제안한다. 사용자에게 전문기관을 추천하기 위해 저자의 소속 정보를 이용한다. 온라인 학술 사이트에서 수집한 데이터를 이용하여 신뢰성을 높이고, 해당 데이터를 기반으로 소속에 대한 랭킹 결과와 랭킹 외에 추가적인 정보들을 제공하여 객관적인 전문기관 추천에 기여하고자 한다.

[그림 1]은 제안하는 전문기관 추천 시스템의 전체 구조도를 나타낸다. 먼저, 사용자가 원하는 분야에 대해 질의하면, 질의와 관련된 논문, 연구 과제, 특허, 보고서와 같은 연구 결과물들에 대한 데이터를 수집한다. 이 때, 데이터 수집은 온라인 학술 사이트에서 제공하는 Open API 및 자체 제작한 Web Crawler를 통해 이뤄진다 [1,2,3]. 수집된 데이터는 전처리를 통해 전문기관 랭킹에

필요한 필수 데이터만을 추출하고, 중복 처리 및 불용어 처리가 함께 이뤄진다. 전처리 과정이 완료되면, 전문기관 분석을 위한 단계로 전환된다. 저자의 소속 정보에 대한 빈도수를 통해 랭킹 정보를 계산한다. 이 뿐만 아니라, LDA 모델을 이용하여 질의에 대한 소속별 잠재 키워드 결과를 추출하고, 수집된 데이터를 통해 연구 실적을 계산한다. 최종적으로 가시화를 통해 통합된 결과를 사용자에게 제공한다.



▶▶ 그림 1. 제안하는 전문기관 추천 시스템 구조도

[그림 2]는 전문기관 추천에 대한 가시화 예제이다. 전문기관 랭킹은 학술 문헌 기반의 데이터에서 연구자의 소속 정보를 추출하여 계산한다. 예를 들어, 논문의 경우, 교신저자를 포함하는 해당 논문의 모든 저자의 소속 정보를 수집하고, 연구 과제의 경우, 책임 연구원을 포함하는 해당 연구 과제의 모든 참여 연구원의 소속 정보를 수집한다. 추출한 연구원들의 소속 정보에 대한 빈도수를 계산하여, 해당 질의에 대한 소속별 랭킹 결과를 도출한다. 도출된 정보는 순위에 따라 가시화되어 사용자에게 제공된다.

Rank	기관명	연구 키워드	연구실적			
			논문수	피인용 합계	공동 연구 건 수	과제수
1	A	키워드A_1, 키워드A_2, 키워드A_3, ...	57건	26건	9건	22건
2	B	키워드B_1, 키워드B_2, 키워드B_3, ...	27건	37건	11건	4건
3	C	키워드C_1, 키워드C_2, 키워드C_3, ...	11건	26건	1건	8건

▶▶ 그림 2. 전문기관 추천 가시화 예제

제안하는 전문기관 추천 시스템은 랭킹 정보를 제공할 때, 다른 유용한 정보들을 함께 제공한다. 먼저, LDA 모델을 이용한 소속별 잠재 키워드 정보를 제공한다. 해당 기관에 속하는 연구자의 논문, 보고서, 특허 같은 연구 결과물들에 대해 LDA 모델링을 수행한다. 이 때, 연구 결과물에서 명사만을 추출하도록 전처리를 진행한다. 전처리된 데이터를 통해 기관별 잠재 키워드를 추출한다. 추출한 잠재 키워드는 연구 키워드라는 항목으로 사용자에게 제공된다[4]. [그림 3]은 LDA 모델을 통해 추출한 잠재 키워드들을 시각화한 예제이다. 잠재 키워드 출현빈도에 따라 워드 클라우드 기법을 이용하였고, 이는 해당 기관의 잠재 키워드들을 파악하는데 용이하다.

▶▶ 그림 3. LDA 잠재 키워드 시각화 예제

LDA 모델을 이용한 기관별 잠재 키워드 외에도 연구 실적에 대한 정보를 함께 제공한다. 연구 실적에는 사용자 질의에 따라, 온라인 학술 사이트에서 수집한 해당 기관 연구자들의 논문 수, 피인용 합계, 공동 연구 건 수, 과제 수가 포함된다. 결국, 가시화를 통해 랭킹 결과를 사용자에게 제공할 때, 전문기관 추천을 위한 유의미한 정보를 함께 제공할 수 있게 된다.

Ⅲ. 결론

본 논문에서는 학술 문헌 데이터를 기반으로 전문기관을 추천하는 시스템을 제안하였다. 학술 문헌 기반 전문기관 추천 시스템은 온라인 학술 사이트에서 수집하고, 전처리한 연구 결과물로 전문기관에 대한 랭킹 결과를 제공하고, 이뿐만 아니라, 사용자에게 LDA 모델을 이용한 기관별 잠재 키워드와 연구 실적에 대한 정보를 함께 제공한다. 향후 연구에서는 효율적인 전문기관 랭킹 수식을 도입하고, 사용자의 의사결정을 돕기 위한 더 많은 정보를 제공할 예정이다.

▌참 고 문 헌▐

[1] https://www.ntis.go.kr/
[2] https://www.scienceon.kisti.re.kr/
[3] https://www.dbpia.co.kr/
[4] Blei, David. M. "Probabilistic topic models." Communications of the ACM 55.4 (2012): 77-84.
[5] Bilir, Selçuk, et al. "A new ranking scheme for the institutional scientific performance." arXiv preprint arXiv:1508.03713 (2015).