이종 학술 문헌 사이트 기반 동명이인 판별 기법

Name Disambiguation Scheme Based on Heterogeneous Academic Document Sites

장 준 혁*, 최 도 진**, 김 종 훈*, 전 종 우*, 정 상 준*, 이 서 희*, 김 상 혁*, Retiti Diop Emane Christopher*, 오 영 호*, 송 상 호*, 김 윤 아*, 이 현 병*, 임 종 태*, 복 경 수***, 유 재 수*†
충북대학교*, 창원대학교**, 원광대학교***

Jun-Hyeok Jang*, Do-Jin Choi**, Jong-Hun Kim*, Sang-Jun Chung*, Seo-Heui Lee*, Sang-Hyeuk Kim*, Retiti Diop Emane Christopher*, Young-Ho Oh*, Sang-Ho Song*, Yun-A Kim*, Hyeon-Byeong Lee*, Jong-Tae Lim*, Kyoung-Soo Bok***, Jae-Soo Yoo*†
Chungbuk National Univ*.
Changwon National Univ.** Wonkwang Univ.***

요약

학술 문헌 사이트에 저장되어 있는 연구물은 매우 다양하고 방대하다. 이런 상황에서 내가 찾는 연구자의 모든 연구물을 한 번에 찾을 수 있는 방법이 필요하다. 본 논문에서는 이종의 학술 문헌 사이트에서 제공하는 학술 데이터를 모두 수집하고 동명이인을 판별하는 규칙기반 동명이인 판별 기법을 제안한다.

I. 서론

연구자들은 자신의 연구 결과물을 논문, 특허, 연구보고서 등의 다양한 형태의 연구물로 작성한다. 따라서 다양한 학술 문헌 사이트에서 연구자를 검색할 때 연구자의 모든 연구 결과물을 통합된 데이터로 보여주는 것이 중요하다. 국내 최대 학술 데이터베이스 DBPIA는 무수히 많은 논문을 제공하지만, 논문이 아닌 다른 연구물은 제공하지 않는다. 국가과학기술 지식정보 서비스(NTIS)는 연구자를 검색하면 연구결과물을 R&D과제 보고서, 논문, 지식재산권 등의 연구결과물을 한 번에 확인이 가능하다. 하지만 NTIS에 등록하지 않은 연구물은 찾을 수가 없다. 이런 이종의 학술 문헌 사이트들로부터 학술 문헌들을 수집하여 통합 서비스를 제공하기 위해서 동명이인의 연구자들을 정확하게 판별해야 할 필요가 있다.

동명이인 판별 문제를 해결하기 위해 다양한 연구들이 진행되었다. 그들은 규칙 기반 기법과 심층 학습 기법으로 구분할 수 있다[1, 2]. 하지만 기존의 동명이인 판별 기법은 동명이인 판별을 위해 정형화된 데이터셋을 이용하므로 이종의 학술 문헌 사이트들로부터 수집한 다양한 형태의 데이터를 그대로 사용하기 어렵다. 또한 기존의 동명이인 판별기법은 영어를 대상으로 만들어져 국문 연구결과물을 고려하지 않는다. 이런 이유로 이종의 학술 문헌 사이트의 학술 문헌들을 통합하기 위한 동명이인 판별 기법이 필요하다.

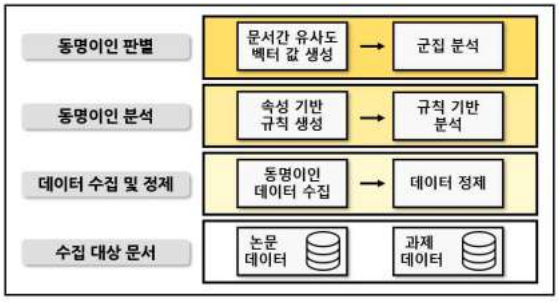
본 논문에서는 이기종 학술 문헌 사이트에서 데이터를 수집하고 같은 이름의 연구자를 판별하는 동명이인 판별 시스템을 제안한다. 제안하는 기법은 연구물을 수집하여 동명이인을 판별에 필요한 속성을 찾고, 그 속성에 가중치를 부여하여 군집 분석을 통해 저자를 구별한다. 또한 제안하는 기법의 우수성을 보이기 위해 성능평가를 수행한다.

Ⅱ. 제안하는 동명이인 판별 기법

본 논문에서는 이종의 학술 문헌 사이트에서 데이터를 수집하여 동명이인을 판별하는 기법을 제안한다. 여러 학술 문헌 사이트에서 데이터를 수집하기 때문에 연구자의 연구결과물이 여러 사이트에 존재할 수 있고, 이름이 같은 다른 연구자의 연구결과물이 검색될 수 있기 때문에 동명이인 판별 기법이 필요하다. 그림 1은 제안하는 기법의 전체 시스템 구조도를 나타낸다. 학술 문헌 사이트에서 수집된 모든 연구결과물 중에서 본 논문에서 고려할 대상인 동명이인이 존재하는 데이터를 수집한 뒤

† 교신저자 : yjs@chungbuk.ac.kr
이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2022R1A2B5B02002456), 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성(Grand ICT연구센터) 사업(IITP-2022-2020-0-01462) 그리고 중소벤처기업부 '산업전문인력역량강화사업'의 재원으로 한국산학연협회(AURI)의 지원(2021년 기업연계형연구개발인력양성사업, 과제번호 : S3047889)을 받아 수행된 연구임.

동명이인 판별에 필요한 속성을 찾아 정제한다. 정제한 속성을 임베딩하고, 속성 값을 이용한 동명이인 판별 규칙을 생성하여 문서 간 유사도를 생성한다. 생성된 유사도를 거리 값의 형태로 변환하여 군집 분석을 통해 동명이인의 모든 저자의 연구결과물을 저자별로 클러스터링하여 동명이인을 판별한다.



▶▶ 그림 1. 전체 시스템 구조도

본 논문의 수집 대상 문서는 이종의 학술 문헌 사이트에서 제공하는 논문과 과제 보고서와 같은 연구결과물이다. 학술 문헌 사이트에 존재하는 모든 연구결과물을 한 번에 수집할 수 없으므로 유의미한 키워드를 검색하여 결과로 나오는 모든 연구결과물을 수집하여 데이터베이스에 저장한다. 본 논문에서 제시하는 기법인 동명이인 판별을 위해 먼저 키워드 검색결과로 나타난 연구결과물을 비교하여 같은 저자의 이름이 있는 2개 이상의 모든 연구결과물을 수집한다.

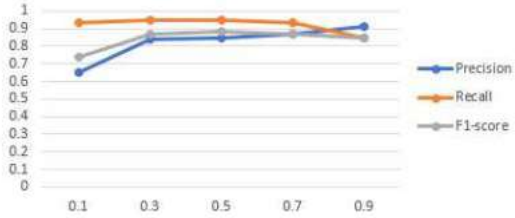
연구결과물에 각기 다르게 표시되는 속성을 고려하여 정제한다. 본 논문에서는 저자의 소속 테이블을 이용하여 저자의 소속을 정제한다. 예를 들어 포항공과대학교, 포스텍, POSTECH등으로 각기 다르게 기재된 소속 정보를 표준 이름으로 조정하여 속성으로 사용한다.

동명이인 분석 단계에서는 수집 대상의 모든 학술 문헌 사이트에 존재하는 다양한 속성(소속, 공동저자, 출판연도, 학회 등)을 고려하여 동명이인을 판별하기 위한 최적의 속성을 찾는다. 고려할 속성의 중요도를 확인하여 각각의 속성에 가중치를 부여하는 규칙을 생성한다. 이 규칙의 가중치를 이용하여 같은 이름을 가진 저자 집합의 모든 문서의 문서 간 유사도를 생성한다.

마지막으로 동명이인 판별 단계에서는 HAC 클러스터링 알고리즘을 이용하여 같은 이름을 갖는 문헌들을 군집화 한다. 즉 동명이인 분석 단계에서 생성된 문서 간 유사도를 문서 간 거리 값으로 변환하고 이 값을 이용하여 군집화 하여 동명이인을 판별한다.

Ⅲ. 성능 평가

제안하는 기법의 우수성을 보이기 위해 성능평가를 수행하였다. 성능평가는 Intel(R) Core(TM) i5-9600K CPU @ 3.70GHz 프로세서와 32GB 메모리를 갖는 시스템에서 진행하였다. 이종의 학술 문헌 사이트에서 인공지능, 검증의 키워드 검색결과로 나타나는 연구결과물 중 동명이인이 존재하는 9,792개의 연구물을 테스트 데이터로 사용하였다. 규칙기반 기법으로 문서 간 거리 값을 생성하고 그 거리 값을 비교하여 데이터들을 계층적으로 군집을 형성하는 기법인 HAC(Hierarchical Agglomerative Clustering)를 이용하여 군집 분석을 실행하였다. 〈그림 2〉는 HAC 중지 기준(0.1~0.9)에 따른 성능 평가 결과이다. 거리 값에 따라 군집에 묶이는 기준을 중지 기준으로 설정하고 정답셋과 비교하여 정밀도(Precision), 재현율(Recall), F1-score를 계산한다.



▶▶ 그림 2. 성능 평가 결과

정밀도는 연구물의 고유한 저자로 확인된 저자의 수를 나타내며, 정밀도가 높을수록 동일 저자의 연구결과물을 군집화 한다. 재현율은 실제 연구결과물의 저자 중 확인된 저자가 몇 명인지를 나타낸다. 따라서 두 성능이 높게 나타난 0.5를 본 논문의 중지 기준으로 사용한다.

Ⅳ. 결론

본 논문은 이종의 학술 문헌 사이트에서 데이터를 수집하고 동명이인을 판별하는 규칙기반 동명이인 판별 시스템을 제안하였다. 제안하는 기법은 학술 문헌 사이트에서 다양한 연구물을 수집하고 연구물의 속성을 이용하여 규칙을 이용한 가중치로 군집 분석을 실행하여 같은 이름을 가진 저자들의 연구물 간 동명이인을 판별하였다. 향후 다양한 성능평가를 통해 제안하는 기법의 우수성을 입증할 예정이다.

■ 참 고 문 헌 ■

[1] Protasiewicz, J., & Dadas, S. "A hybrid knowledge-based framework for author name disambiguation.", p.000594-000600. IEEE International Conference on Systems, Man, and Cybernetics. 2016.

[2] Chen, Y., Yuan, H., Liu, T., & Ding, N. "Name Disambiguation Based on Graph Convolutional Network." Scientific Programming. 2021.