학술 문헌 기반 동명이인 처리 연구 동향 분석

Analysis of Name Disambiguation Research Trends Based on Academic Documents

최도진, 장준혁*, 송상호*, 오영호*,
RETITI DIOP EMANE Christopher*, 김상혁*, 이서희*,
정상준*, 전종우*, 김종훈*, 복경수**, 유재수*1)
창원대학교, 충북대학교*, 원광대학교**

Choi do-jin, Jang jun-hyeok*, Song sang-ho*,
Oh young-ho*, RETITI DIOP EMANE Christopher*,
Kim sang-hyeuk*, Lee seo-heui*, Chung sang-jun*,
Jeon jong-woo*, Kim jong-hun*, Bok kyoung-soo**,
Yoo jae-soo*
Changwon National University, Chungbuk National University*, Wonkwang University**

요약

연구자들의 다양한 연구 성과물들에 대해서 다른 연구자와 구별하기 위해서 동명이인 분석 연구가 진행되고 있다. 본 논문에서는 기존의 동명이인 분석에 대한 연구들의 동향과 문제점을 분석한다. 마지막으로 분석 된 내용을 바탕으로 향후 연구 방향을 제시한다.

I. 서론

연구자들의 연구 성과물은 다양한 방법과 형태로 생성된다. 서로 다른 형태의 연구 성과물들은 하나의 정형화된 포맷으로 만들기 어렵고, 이러한 이유로 인해서 서로 다른 연구 성과물들이 한 연구자에 의해 생성되었다는 의미를 부여하기 위해서 다양한 연구와 방안들이 제시되고 있다. 기존에는 본인의 연구 성과물을 인정받기 위해서 연구자 고유의 번호를 발급받는 방법이 주로 활용되고 있다.

그러나 논문과 같은 연구 성과물에서는 모든 저널에서 연구자 고유 번호를 제공하지 않고, 제공하더라도 논문의 모든 저자가 이를 활용하지는 않는다. 이러한 이유로 인해 다양한 저널에서 제공하는 논문의 단순 메타데이터만을 이용하여 동일한 저자(연구자)를 판별하는 동명이인 판별 기법이 활발하게 연구되고 있다. 기존 동명이인 연구는 규칙 기반 동명이인 분석 연구와 딥 러닝 기반 분석 연구로 나뉜다[1-6]. 본 논문에서는 두 가지 방법의 동명이인 연구들을 분석하고, 이에 대한 특징과 문제점을 분석한다.

본 논문에서는 기존에 활발하게 연구되고 있는 동명이인 판별 기법들의 동향 분석을 수행한다. 뿐만 아니라 현재 판별 기법들의 문제점을 분석하여 향후 연구 방향에 대한 제시를 수행한다.

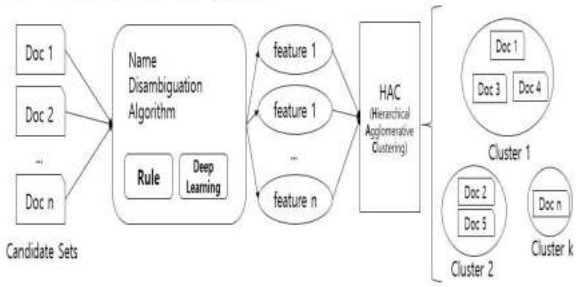
1) 교신저자 : yjs@cbnu.ac.kr
이 (성과)는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2022R1A2B5B02002456)과 중소벤처기업부 '산업전문인력역량강화사업'의 재원으로 한국산학연협회(AURI)의 지원(2021년 기업연계형연구개발인력양성사업, 과제번호 : S3047889)과 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성(Grand ICT연구센터) 사업의 연구결과로 수행되었음" (IITP-2022-2020-0-01462)

Ⅱ. 동명이인 처리 연구 분석

동명이인 분석 연구는 크게 두 가지로 나뉜다. 첫 번째는 규칙 기반의 통합 기법으로, 학술 문헌의 정보는 공통된 메타데이터 정보를 담고 있기 때문에, 이러한 특징을 활용한 통합을 수행하는 연구이다. 두번째는 학습 기반의 통합 연구이다. 최근 기계 학습 기술의 발달로 인해 이러한 연구가 가장 많이 진행되고 있다. 메타데이터 정보를 기반으로 특징 벡터를 생성하고 특징 벡터 기반의 통합을 수행하는 연구들이 주로 이루어져있다.

그림 1은 동명이인 분석 방법의 전체 흐름을 나타낸다. 먼저, 동명이인 분석 대상의 후보 집합 (동일한 저자의 이름이 포함 된 학술 문헌들)을 생성한다. 그 이후에, 동명이인 분석기로 입력이 되는데, 내부적으로 규칙 기반 혹은 딥 러닝 기반의 분석을 수행한다. 최종적으로 분석된 내용은 각 문서별 특징 벡터가 추출되고, 특징 벡터 기반의 HAC (Hierarchical Agglomerative Clustering)를 수행하여 문서별 벡터간의 거리 값을 기반으로 모든 후보 문서를 클러스터링한다. 클러스터링 결과가 최종적으로 동명 이인 분석 된 결과이며, 각 클러스터에는 동일한 저자가 포함 된 논문들을 나타낸다. 전체적인 시스템을

보면 알 수 있듯이 동명 이인 분석 연구는 후보군 생성과 최종 HAC 클러스터링 적용은 규칙 기반, 딥 러닝 기반 연구 모두 동일한 방법을 수행하고, 클러스터링에 입력될 특징 벡터를 생성하는 방법이 동명이인 분석 연구의 핵심으로 볼 수 있다.



▶▶ 그림 1. 동명이인 분석 방법

표 1은 기존 학술 문헌 기반의 동명이인 연구들의 특징을 나타낸다. [1]은 사전에 정의한 휴리 스틱한 규칙 기반의 동명이인 분석을 수행한다. 결과 생성에서 HAC를 수행하기 전에 규칙에 의해서 사전 클러스터링되는 정보가 존재한다. [2]는 이름을 기반으로 사전에 클러스터링을 수행하고, 최종 결과를 생성하기 위한 클러스터링 단계에서 이름 특징을 일부 반영한 준지도 학습을 수행하는 특징이 존재한다. [3]은 SVM(Support Vector Machine)을 이용한 최종 결과 생성과 더불어 개인 정보 보호를 위해 익명화된 데이터만으로 동명 이인 분석을 수행한다.

[4-8]은 문서의 이름과 문서간의 관계성을 이용하여 그래프로 모델링 한 후, 그래프 학습을 통해 최종적은 HAC를 수행하는 연구들이다. [4]는 그래프 모델링을 통해 모든 문서의 특징 벡터를 학습시킨 후, HAC를 통해 결과를 생성한다. HAC 결과에 대해서 Human Annotator라고 하는 전문 집단이 피드백 정보를 시스템에 반환한다. 시스템은 반환된 피드백 정보 기반의 HAC 학습을 수행하여 지속적으로 향상 가능한 분석을 수행한다. [5]는 그래프 분석의 속도 문제를 해결하기 위해서 빅데이터 분산 처리 플랫폼인 Spark 기반의 연구를 수행하였다. 마지막으로 [6]은 공저자 그래프, 문서 간 그래프, 문서와 저자간의 그래프 3가지의 그래프를 생성한다. 생성 된 3가지의 그래프에 대해 GCN(Graph Convolutional Network)을 통해 특징 벡터를 학습하고, 최종적으로 추출 된 문서 별 Hidden Feature를 통해 클러스터링을 수행한다.

표 1. 기존 동명 이인 분석 연구

저자	특징	결과	구분
Protasiewicz [1]	휴리스틱 규칙	HAC+ 규칙	규칙
Louppe [2]	준지도 클러스터링	HAC	딥 러닝
Zhang [3]	정보보호 특징	SVM	
Zhang [4]	Human Annotator	HAC	
Du [5]	Spark GraphX		
Chen [6]	GCN		

Ⅲ. 결론

본 논문에서는 기존 동명이인 처리 연구들을 분석하였다. 현재 동명 이인 분석 연구는 주로 딥 러닝 기반의 연구들을 수행하고 있으며, 그래프 모델링 기반의 GCN을 활용하는 추세로 이어지고 있다. 기존 연구들은 성능 평가 데이터 셋만을 활용하여 타당성을 입증하고 있다. 실 생활에 적용 할 수 있는 연구를 수행하기 위해서는, 실제 데이터와 분석 시간에 대한 고려도 많이 이루어져야 실효성있는 연구가 될 것이다.

▌참 고 문 헌▐

[1] Protasiewicz, Jaros ł aw, and S ł awomir Dadas. "A hybrid knowledge-based framework for author name disambiguation." 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2016.

[2] Louppe, Gilles, et al. "Ethnicity sensitive author disambiguation using semi-supervised learning." international conference on knowledge engineering and the semantic web. Springer, Cham, 2016.

[3] Zhang, Baichuan, and Mohammad Al Hasan. "Name disambiguation in anonymized graphs using network embedding." Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017.

[4] Zhang, Yutao, et al. "Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop." Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018.

[5] Du, Hongliang, Zhiyi Jiang, and Jianliang Gao. "Who is who: Name disambiguation in large-scale scientific literature." 2019 International Conference on Data Mining Workshops (ICDMW). IEEE, 2019.

[6] Chen, Ya, et al. "Name Disambiguation Based on Graph Convolutional Network." Scientific Programming 2021 (2021).