토픽모델링을 통한 저자명 식별 성능 비교

A Comparison of Author Name Disambiguation Performance through Topic Modeling

김하진, 연세대학교 문헌정보학과, hajin_228@yonsei.ac.kr
정효정, 연세대학교 문헌정보학과, hjung4582@yonsei.ac.kr
송민, 연세대학교 문헌정보학과, min.song@yonsei.ac.kr
Ha Jin Kim, Dept. of Library & Information Science, Yonsei Univ.
Hyo-jung Jung, Dept. of Library & Information Science, Yonsei Univ.
Min Song, Dept. of Library & Information Science, Yonsei Univ.

본 연구에서는 저자명 모호성 해소를 위해 토픽모델링 기법을 사용하여 저자명을 식별 하였다. 기존의 토픽모델링은 용어 자질만을 고려하였지만 본 연구에서는 제 3의 메타데이터 자질을 활용하여 ACT(Author-Conference Topic Model) 모델과 DMR(Dirichlet-multinomial Regression) 토픽모델링을 대상으로 저자명 식별 성능을 평가, 비교하였다. 또한 수작업으로 저자 식별 작업을 한 데이터셋을 기반으로 저자 당 논문 수와 토픽 수에 차이를 두고 연구를 진행하였다. 그 결과 저자명 식별에 있어 ACT 모델보다 DMR 토픽모델링의 성능이 더 우수한 것을 알 수 있었다.

1. 서 론

1.1 연구 배경 및 목적

최근 빅데이터 시대로 진입함에 따라 데이터의 양이 방대해졌으며 무분별하게 많은 양의 데이터 중 양질의 데이터를 선별하는 것에 대한 중요성이 더욱 높아지고 있다. 학술 데이터 역시 그 양이 점점 증가하고 있으며 양질의 학술 데이터를 선별하는데 있어 동명이인과 같은 저자명 모호성에 대한 문제의 해결은 필수적이라고 할 수 있다.

기존의 저자명 식별 기술은 문헌이 가지고 있는 메타데이터를 활용하여 문헌 쌍 간의 유사도에 기반을 둔 응집적 클러스터링 기법을 사용하거나 문헌 자질을 이용한 학습기반 분류기법을 사용한 연구가 대부분이다(Liu et al., 2014; Han et al., 2005; Fan et al., 2011). 반면 대용량의 컬렉션의 숨겨진 구조 즉, 주제 분포도를 발견해내기 위한 기법인 토픽모델링(Blei, 2012)은 클러스터링 기법이나 학습기반 분류기법과 함께 사용하여 식별하고자 하는 저자의 주제 분야를 표현하는데 주로 적용되었다(Song et al., 2007).

본 연구는 토픽모델링 기법을 통해 저자명 식별 성능을 평가하고자 하였다. 기존의 저자명 모호성 해소에 사용된 토픽모델링은 용어 자질만을 고려한 LDA(Latent Dirichlet Allocation)의 기법이었으나 본 연구에서는 제 3의 메타데이터를 자질로 고려한 ACT(Author-Conference Topic Model; Tang et al., 2008) 모델과 DMR (Dirichlet-multinomial Regression; Mimno & McCallum, 2012) 토픽모델링을 사용하

여 저자명 모호성 해소를 하였다. 또한 저자당 논문 수 및 토픽 수에 따른 차이를 통해 두 가지 토픽모델링 기법의 저자명 식별 성능을 평가, 비교하였다.

1.2 선행연구

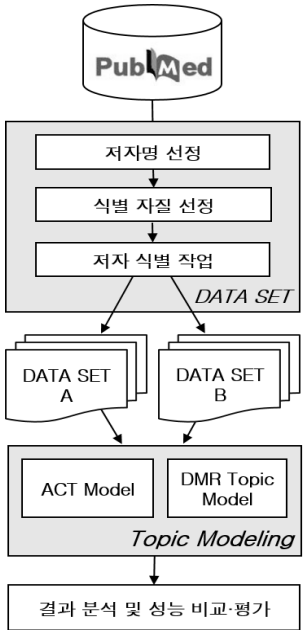
저자명 모호성은 동일한 저자가 복수 개의 이름을 사용하였을 경우와 복수의 저자가 동일한 이름을 가지고 있는 경우에 발생한다. 저자명 모호성 해소를 위한 저자명 식별과 관련된 연구는 주로 학술 데이터베이스인 PubMed와 DBLP를 대상으로 다양한 연구가 진행되어 왔다(Liu et al., 2014; Han et al., 2005; Fan et al., 2011).

저자명 식별을 위해 토픽모델링을 적용한 대표적인 연구로는 Song 등(2007)의 연구가 있다. 이들은 LDA(Latent Dirichlet Allocation)와 PLSA(Probablistic Latent Semantic Analysis)를 활용하여 저자명을 식별하였는데 이 때 토픽모델링의 각 토픽을 고유의 저자로 표현되는 자질로 간주하고 이 자질을 대상으로 클러스터링 기법을 사용하여 저자명 식별을 하였다. 더 나아가 Tang 등(2008)이 제안한 ACT(Author-Conference Topic Model) 모델 기법은 토픽모델링의 파라미터 값의 변수에 컨퍼런스명을 추가한 것으로 학술 저자들의 논문정보를 자동으로 수집하여 저자들의 공저자 네트워크와 연구 분야를 식별하여 제공하는 ArnetMiner 라는 모델을 구축하였다(http://arnetminer.org).

기존 연구에서는 토픽모델링 기법이 저자명 모호성 해소의 보조 수단으로써 저자의 주제 분야를 파악하기 위해 활용되었고 토픽모델링 기법 자체의 저자명 식별 성능에 대해서는 알 수가 없었다. 본 연구에서는 기존의 연구와 달리 토픽모델링 기법 자체의 저자명 식별 성능을 살펴보았다. 또한 토픽모델링 기법의 저자명 식별 성능을 비교하기 위하여 ACT(Author-Conference Topic Model) 모델과 DMR(Dirichlet-multinomial Regression) 토픽모델링을 동일한 조건 아래에서 비교하여 성능을 평가하였다.

2. 연구 방법 및 결과

2.1 연구 방법



<그림 1> 연구방법

토픽모델링을 통해 저자명 식별 성능을 비교하기 위한 본 연구의 연구방법은 <그림 1>과 같다. 먼저, PubMed에서 동서양 이름을 포함한 상위 빈도 저자 67개의 그룹 군을 수집하였다. 다음으로 67개의 수집된 저자 그룹 군을 대상으로 2번의 저자명 식별 작업을 거쳤다. 1차 저자명 식별 작업의 자질로는 affiliation, co-author list, e-mail, Scopus ID를 선택해 식별하였고 2차 저자명

식별 작업은 키워드 조합으로 웹에서 저자가 직접 입력한 Publication list를 수작업으로 수집하여 식별 작업을 진행하였다. 2차 식별 작업을 통해 최종적으로 식별된 저자에 대해 인덱스를 부여하여 최종 데이터 셋을 구축하였다.

이와 같이 인덱스를 부여한 최종 데이터 셋을 대상으로 토픽모델링 기법 즉, ACT 모델과 DMR 토픽모델링에 대한 실험을 진행하였다. ACT 모델과 DMR 토픽모델링은 모두 기존 LDA 토픽모델링을 발전시킨 알고리즘으로 새로운 파라미터를 추가한 기법이다. ACT 모델의 경우 저자-컨퍼런스명을 새로운 자질로 주었고, DMR 토픽모델링의 경우 저자명을 제 3의 파라미터로 지정하였다. 각 토픽모델링 ACT 모델과 DMR 토픽모델링의 저자명 식별 성능을 비교하기 위해 저자가 식별된 최종 데이터 셋을 저자 당 논문수가 5개 이상인 DATA SET A와 한 저자 당 논문 수가

<표 1> 데이터 셋의 Name Variants

DATA SET A		DATA SET B	
Martin C	Zhang D	Gupta R	Miller M
Williams S	Miller M	Evans M	Agarwal R
Gupta R	Agarwal R	Roy S	Banerjee S
Evans M	Taylor J	Klein R	Ghosh S
Liu F	Sun X	Patel S	Jain S
Roy S	Banerjee S	Cohen J	Nakamura H
Klein R	Ghosh S	Schmidth H	
Patel S	Brown J		
Cohen J	Jain S		
Schmidth H	Nakamura H		
Smith R			
Name Variants: 21 Total Variants: 146 Total record: 1,197		Name Variants: 13 Total Variants: 40 Total record: 504	

10개 이상인 DATA SET B로 나누었다. <표 1>은 데이터 셋에 대한 통계로 저자 당 논문 수가 5개 이상인 저자들의 Name Variants는 21명으로 총 레코드 수는 1,197개이며, 저자 당 논문 수가 10개 이상인 저자들의 Name Variants는 13명으로 총 레코드 수는 504개이다. 이와 같이 구축된 두 데이터 셋를 대상으로 각 토픽모델링의 토픽 수를 20, 30, 40으로 차이를 두어 ACT 모델과 DMR 토픽모델링에 대한 저자명 식별 성능 평가 및 비교 분석을 진행하였다.

2.2 연구 결과

ACT 모델과 DMR 토픽모델링을 통한 저자명 식별 성능은 <표 2>와 같다. 저자명 식별 평가에 있어서 정보검색의 정확률 공식을 기반으로 수행하였다.

실험 결과 ACT 모델보다 DMR 토픽모델링의 저자명 식별 성능이 전반적으로 우수한 것으로 나타났으며 ACT 모델의 경우 토픽의 수가 30일 때 최대 정확률로 53%, DMR 토픽모델링의 경우 토픽의 수가 20일 때 최대 정확률로 74%가 나타났다. 또한 분석 결과, 저자 당 논문 수가 5개 이상인 데이터를 대상으로 토픽모델링을 수행하였을 때보다 저자 당 논문 수가 10개 이상인 데이터를 대상으로 토픽모델링을 수행하였을 때 두 토픽모델링 모두 저자명 식별 성능이 전반적으로 높은 것을 알 수 있었으며, 두 토픽모델링 모두 상위 10위까지의 결과에 높은 식별 정확률을 보이는 것을 발견할 수 있었다. 그러나 모든 조건에 있어 DMR 토픽모델링의 저자 식별 성능이 ACT 모델 성능보다 더 높은 것을 알 수 있었다.

3. 결 론

본 연구는 두 가지 토픽모델링인 ACT 모델과 DMR 토픽모델링을 통해 저자명 식별의 성능을 비교하고자 하였으며, 토픽모델링의 저자명 식별 성능 비교를 위해

<표 2> ACT 모델과 DMR 토픽모델링에 의한 저자명 식별 정확률

	저자 당 논문 수 5개 이상인 DATA SET A						저자 당 논문 수 10개 이상인 DATA SET B					
	ACT			DMR			ACT			DMR		
	Topic 20	Topic 30	Topic 40	Topic 20	Topic 30	Topic 40	Topic 20	Topic 30	Topic 40	Topic 20	Topic 30	Topic 40
상위 10	0.280	0.333	0.320	0.460	0.450	0.490	0.435	0.527	0.525	0.735	0.660	0.683
상위 20	0.245	0.287	0.299	0.338	0.350	0.370	0.400	0.387	0.325	0.548	0.467	0.449

식별대상이 되는 저자 당 논문 수와 토픽 수 변화에 차이를 두고 이에 따라 두 토픽 모델링 기법의 저자명 식별 성능 차이를 비교하고자 하였다.

연구 결과 전반적으로 토픽 수의 변화나 저자 당 논문 수의 차이와 관계없이 DMR 토픽모델링의 저자명 식별 성능이 ACT 모델보다 우수하였다. 더 나아가 저자 당 논문 수가 많을 때 두 토픽모델링 모두 저자명 식별 성능이 높다는 것을 알 수 있었다. 그러나 본 연구에서는 단순히 토픽모델링의 저자 할당 확률의 차이가 미미하기 때문에 상위 순위를 파악하는데 어려움이 있었다. 향후 연구로 저자명 식별을 위한 데이터를 확장하고 정교화 하여 토픽모델링 결과 나타난 저자 할당 확률의 미미한 차이의 문제를 보완하면 더 좋은 분석 결과를 얻을 수 있을 것으로 기대된다.

참고문헌

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84.

Fan, X., Wang, J., Pu, X., Zhou, L., & Lv, B. (2011). On graph-based name disambiguation. Journal of Data and Information Quality (JDIQ), 2(2), 10.

Han, H., Zha, H., & Giles, C. L. (2005, June). Name disambiguation in author citations using a k-way spectral clustering method. In Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on (pp. 334-343). IEEE.

Liu, W., Islamaj Doğan, R., Kim, S., Comeau, D. C., Kim, W., Yeganova, L., & Wilbur, W. J. (2014). Author name disambiguation for PubMed. Journal of the Association for Information Science and Technology, 65(4), 765-781.

Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv preprint arXiv:1206.3278

Song, Y., Huang, J., Councill, D., Li, J., & Giles, C .L. (2007). Efficient topic-based unsupervised name disambiguation. In JCDL, 342-351.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008, August). Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 990-998). ACM.