저자 식별을 위한 자질 비교

Features for Author Disambiguation

강인수, 이승우, 정한민, 김평, 구희관, 이미경, 성원경, 박동인
한국과학기술정보연구원, 정보기술개발단

In-Su Kang, Seungwoo Lee, Hanmin Jung, Pyung Kim, HeeKwan Goo, MiKyung Lee, Won-Kyung Sung, DongIn Park
Korea Institute of Science and Technology Information

요약

학술 정보에서 저자는, 실세계의 한 저자가 형태적으로 둘 이상의 저자명으로 출현할 수 있으며, 서로 다른 저자들이 동일한 저자명을 공유하기도 한다. 이는 각각 학술 정보에 대한 검색 및 탐색에 있어, 재현율과 정확률을 저하시키는 요인이다. 이 연구에서는 후자에 해당하는 저자의 동명이인 문제에 있어, 그 중의성 해소를 위한 자질의 특성에 집중하고자 한다. 최근까지, 저자 식별을 위한 자질로, 공저자, 논문 제목, 게재지명과 같은 서지 내적 자질과, 논문 원문 텍스트로부터 획득되는 전자메일주소, 소속기관, 논문의 토픽 등과 같은 서지 외적 자질이 사용되어 왔다. 그러나, 이러한 자질들이 저자 식별에 미치는 영향에 대한 비교 분석 연구는 찾아 보기 힘들다. 이 연구에서는, 한글 저자명에 대해 원문과 연계된 대용량 저자 식별 평가 셋을 구축하여, 동명 저자 중의성 해소에 있어 다양한 자질들의 특성을 비교한다.

Abstract

There exists a many-to-many mapping relationship between persons and their names. A person may have multiple names, and different persons may share the same name. These synonymous and homonymous names may severely deteriorate the recall and precision of the person search, respectively. This study addresses the characteristics of features for resolving homonymous author names appearing in citation data. As disambiguation features, previous works have employed citation-internal features such as co-authorship, titles of articles, titles of publications as well as citation-external features such as emails, affiliations, Web evidences. To the best of our knowledge, however, there has been no literature to deal with the influences of features on author disambiguation. This study analyzes the effect of individual features on author resolution using a large-scale test set for Korean.

I. 서 론

사람과 관련된 정보가 웹 상에 증가하면서, 인명을 질의로 하는 인물 검색에 대한 수요가 점차 증가하고 있다 (Guha & Garg, 2004). 그러나, 실세계의 사람과 인명 간에는 다대다 관계가 성립하므로 인명으로 사람을 찾는데 있어 어려움이 발생한다. 즉, 동일 인명을 가진 다수의 사람이 존재할 수 있으며, 또한 한 사람이 자신의 이름을 여러 형태(예: David Johnson vs. Johnson, D.) 로 표기하기도 한다. 이는 사람에 대한 검색에 있어 정확률과 재현율을 저하시키는 요인이 되며, 최근 주목받는 사회망의 질적 측면에도 악영향을 미친다. 따라서, 동명이인을 식별하는 것과 동일 이름의 다른 표현을 하나로 묶는 것은 사람 검색 및 표현에 있어 반드시 다루어져야 할 문제이다.

본 연구는, 인명 출현의 도메인을 학술 정보로 제한하여, 저자명을 실세계 사람으로 대응시키는 저자 식별 문제를 다루고자 한다. 기존 연구에서는, 저자 식별을 위해 공저자, 논문제목, 게재지와 같은 서지 내적 자질과, 저자의 전자메일주소/소속, 논문 원문 텍스트, 웹 상의 저자 홈페이지와 같은 서지 외적 자질을 활용해 왔다. 그러나, 최근까지 주로 수행된 저자 식별의 방법론적 연구들은 다른 분류 문제의 범주를 크게 벗어나지 못했고, 저자 식별에 있어 고유한 자질 특성에 대한 연구는 찾아보기 힘들다. 본 논문은, 저자 식별의 다양한 자질들이 저자 신원 결정에 미치는 영향을 대용량 평가셋을 통한 실험을 통해 비교 분석하고자 한다.

II. 관련 연구

학술 서지에 출현한 저자명을 실세계의 저자로 구분하는 문제는, 동일 저자명 매칭(name matching)과 동명 저자 식별(name disambiguation)의 문제로 세분된다. 관점에 따라, 저

자 식별 문제가 저자명 매칭의 문제를 포함하는 것으로 볼 수도 있겠으나, 이 논문에서는 저자명 매칭을 저자 식별의 전단계로 고려함으로써 두 문제를 분리하는 관점을 취할 것이다.

동일 저자명 매칭은, 같은 저자명에 대한 서로 다른 표기들을 하나의 클래스로 그룹화하는 것으로, 레코드 링키지(record linkage) 분야에서 오랫동안 다루어온 토픽의 하나이다 (Winkler, 2006). 이는, 저자명을 적는 스타일의 차이, 입력 오류, 로마자 표기 변환의 비일관성 등의 문제들로 인해, 영어 저자명 처리에서 피할 수 없는 과정이다. 다행히, 한중일 언어의 경우는 단일의 이름 표기법을 사용하므로, 동일 저자명 매칭의 문제를 거의 겪지 않는다. 대부분의 저자명 매칭 방법은, 임의의 두 저자명에 대해 계산되는 형태적 유사도가 임계치를 넘는 지를 검사하는 절차를 밟는다. 형태적 유사도는, 이름 토큰(first/middle/last name)의 보편성, 이름 스트링 간 문자 오버랩 및 발음 유사 정도 등의 저자명 내부 자질을 주로 사용하여, 편집거리, Soundex, Jaro, 코사인 유사도 기법 등을 적용하여 계산된다. 이 논문에서는 저자명 매칭에 대해 더 깊이 다루지 않으며, 저자 식별 문제와 그 자질에 집중하고자 한다.

일반적으로, 저자 식별은, 저자명 매칭을 통해 하나로 묶인 동일 저자명들을 입력으로 받아, 먼저 임의의 두 저자명 사이의 의미적 유사도[1](이후, 저자 유사도)를 계산하고, 그 결과를 바탕으로 저자명들을 군집화하는 방식을 취한다. 저자 식별을 위한 기본적 자질로는, 저자명이 출현한 논문 서지 레코드를 구성하는 공동 저자명, 논문 제목, 게재지명, 게재 연도 등이 사용된다. 이외에도, 저자의 논문 원문으로부터 해당 저자의 전자메일주소 및 소속기관을 추출하여 이용하거나 (Culotta et al., 2007; Huang et al., 2006; Kanani et al., 2007), 웹 검색을 통해 얻어지는 저자 홈페이지(예: publications pages)를 활용하기도 한다 (Aswani et al., 2006; Kanani et al., 2007; McRae-Spencer & Shadbolt, 2006; Tan et al., 2006; Yang et al., 2006). 일부 연구에서 타자질들에 비해 공동 저자 자질이 저자 식별에 보다 효과적임을 지적하고 있으나 (Han et al., 2004; Torvik et al., 2005; Yang et al., 2006), 다양한 저자 식별 자질들의 특성을 비교한 논문은 찾아 보기 힘들다.

상기의 자질들에 기반한 의미적 유사도 계산 방법으로는, 지도식과 비지도식이 있다. 지도식은, 학습 데이터로부터, 출현된 두 저자가 동일인인지 아닌지를 판별하는 이진 분류기를 학습하고, 분류기가 출력하는 분류 신뢰도 점수를 두 저자명 사이의 유사도로 사용한다 (Huang et al., 2006; Yang et al., 2006; Kanani et al., 2007; Culotta et al., 2007). 비지도식은, 대응하는 자질들 사이의 유사도로부터 저자 유사도를 계산하는 함수를 정의하여 사용한다 (Aswani et al., 2006; Lee et al., 2005; Torvik et al., 2005). 저자 유사도를 계산한 이후, 저자 군집화가 수행되며, 최근까지 응집형 군집법 (Culotta et al., 2007; Song et al., 2007; Tan et al., 2007), DBSCAN (Huang et al., 2006), 통계적 그래프 분할법 (Kanani et al., 2007) 등이 사용되었다.

Ⅲ. 저자 식별 자질

저자 식별을 위한 자질은, 서지 내적 자질과 서지 외적 자질로 구분할 수 있다. 서지 내적 자질은, 저자 리스트, 논문제목, 게재지명, 연도 등으로 구성되는 논문 서지 레코드 내에서 추출될 수 있는 자질을 가리킨다. 서지 외적 자질은, 논문 서지 레코드 외부로부터 얻어지는, 저자 신원 결정에 도움이 되는, 모든 종류의 자질을 의미한다. 이는 논문 원문으로부터 획득될 경우, 저자의 전자메일주소/소속정보, 초록, 키워드 리스트, 논문 전문 텍스트, 관련 과제 정보, 감사의 글, 참고문헌 등을 포함할 수 있다. 또한, 웹으로부터 획득될 경우 저자의 출판 논문들을 정리해 둔 웹 페이지 URL이 유용한 서지 외적 자질에 해당된다.

저자 식별을 위해 이들 자질을 사용하는 기본 가정은 다음과 같다. 즉, 실세계의 특정한 한 저자의 신원은, 일정 기간 동안 유지되는 그 저자의 공동연구자나, 그 저자의 관심 토픽들로부터 결정될 수 있다는 것이다 (Han et al., 2003).

저자의 관심 토픽은, 논문제목/게재지명 뿐만 아니라, 초록, 키워드 리스트, 논문 전문 텍스트로부터도 추출될 수 있다. 이와 관련하여, 참고문헌 내의 인용 논문들은, 한 저자가 특정 토픽에 대한 연구를 수행하는 일정 기간 동안 주요 인용 논문들을 자신의 논문의 참고문헌에 거의 빠짐 없이 포함시킬 것이라는 가정하에 사용될 수 있을 것이다. 이는, 논문제목, 게재지명, 초록, 키워드, 논문 전문 등에서 얻어지는 용어 기반의 관심 분야보다, 저자의 연구 토픽을 더 정확하고 구체적으로 표현할 수 있을 것으로 판단된다. 또한, 인용 기반의 토픽 표현에 비해, 용어 기반의 토픽 표현은 토픽 셋 통제의 어려움을 겪는다. 기존 연구에서, 참고문헌 내 인용 논문 레코드는 저자 식별 자질로 사용된 적이 없다.

전자메일주소는 특정 시점에 주민등록번호와 유사하게 개인의 고유식별문자열로 기능할 수 있다. 그러나, 메일주소는 일정 기간 동안에는 여러 사람에 의해 재사용이 가능하므로 항상 한 저자를 고유하게 식별할 수 있는 것은 아니다. 저자의 소속 기관은, 한 기관 내에 동명 이인이 많지 않을 것이라는 가정하에, 서로 다른 기관에 소속된 동명 저자들은 각기 다른 사람이라는 추론을 가능케 한다. 과제 정보나 감사의 글에는 특정 논문에 기여한 과제의 명칭이나 번호가 기록되어 있으므

1) 동일 저자명에 대응하는 실세계 저자 각각을 하나의 의미 표지로 고려하여 만든 용어임.

로, 동일한 과제번호/명칭이 사용된 논문들의 동명 저자들은 실세계의 동일인임을 강하게 추측케 할 수 있다.

저자의 논문 정보가 기록된 웹 페이지 URL은, 그것이 연구자 개인의 논문업적페이지를 가리키는 것일 경우, 저자 식별에 절대적 영향을 미칠 수 있다. 그러나, 웹을 통해 저자명과 논문제목이 포함된 웹페이지를 검색했을 때, 여러 전자도서관들의 URL도 얻어진다. 따라서, 웹 URL을 저자 식별에 사용하기 위해서는, 그 URL이 개인의 논문업적페이지인지 아닌지를 가려내는 별도의 분류기가 요구된다.

이 연구에서는, 상기 자질들 중 그것의 즉시 활용 가능성과 통제성을 고려하여, 공동 저자, 논문제목, 게재지명, 전자메일주소, 소속, 참고문헌의 여섯 가지 자질들에 집중하여 저자 식별에 미치는 영향을 살펴 볼 것이다.

Ⅳ. 실 험

저자 식별 평가를 위해, 1999년부터 2006년까지 출판된 IT 관련 국내 주요 학술대회 발표 논문 8,675편에 대해 수작업으로 논문 서지 메타데이터(저자명, 논문제목, 게재지명, 연도, 전자메일주소, 저자소속 등)를 구축하고, 23,177개의 출현 저자명에 대해 수작업으로 저자 식별을 수행하였다. 수작업 저자 식별은, 먼저 동명 저자들에게 모두 다른 식별자를 부여한 다음, 국가과학기술인력 종합정보시스템[2], 홈페이지 검색, 서지 메타데이터 등을 참조하여 동일인임이 판명된 저자들을 하나의 식별자 아래로 병합하는 과정을 거쳤다. 그 결과, 전체 5,332개의 동명 저자 그룹에 대해, 9,133명의 실세계 저자가 발견되었다.

저자 식별 자질로는, 공동 저자, 논문제목, 게재지명, 전자메일주소, 소속, 참고문헌의 여섯 가지 자질을 사용하여, 개별 및 통합 자질의 저자 식별 성능을 비교하였다. 저자 유사도는, 대응하는 자질 간 유사도의 일차결합으로 정의하였다. 자질 간 유사도는, 공동 저자, 게재지명, 메일주소, 소속, 참고문헌 인용[3]의 경우 일치하는 자질값이 하나라도 존재할 경우 1, 그렇지 않은 경우 0의 값을 부여하였다. 게재지명과 소속은 전거통제를 하였다. 논문제목의 경우, 미등록어 문제를 고려하여 조사/어미 절단 방식으로 용어 추출 후 음절바이그램을 생성하고, 두 논문제목 사이의 음절바이그램 오버랩 비율이 임계치 0.08 이상일 때 자질 유사도 값 1을 그렇지 않은 경우 0을 부여하였다. 임계치 0.08은, 임계치 후보로 0.01, 0.02, ..., 0.1, 0.2, ..., 0.9, 1에 대해 저자 식별 성능을 테스트했을 때 최고 성능을 보였다.

2) http://www.hrst.or.kr/

3) 동일 논문에 대한 서로 다른 인용 레코드의 매칭(citation matching)을 위해, 인용 논문의 제목 스트링만을 사용하였다.

저자 군집을 위해, 단일 링크 응집형 군집법(single-link agglomerative clustering)을 적용하였다. 군집화 과정에서 군집 병합 판단을 위한 군집간 유사도는 전술한 저자 유사도가 사용된다.

저자 식별 성능의 평가를 위해, F1 지표를 사용하였다. F1은, 동일 정답 군집 내에 있는 임의의 저자 개체쌍이 동일 시스템 군집 내에서 발견되는 비율인 재현율과, 동일 시스템 군집 내에 있는 임의의 저자 개체쌍이 동일 정답 군집 내에서 발견되는 비율인 정확률의 조화평균을 계산한 것이다. 또한, 군집 오류 평가를 위해, 과다군집오류(over-clustering error: OE)와 과소군집오류(under-clustering error:UE) 지표를 사용하는데, 이는 임의의 저자 개체쌍 전체 중에서 각각 동일 시스템 군집과 동일 정답 군집에서만 발견되는 비율을 의미한다.

[표 1] 단일 자질의 저자 식별 성능 (저자 유사도 임계치=1)

자질	F1	OE	UE
게재지명 (P)	57.58%	6.17%	34.24%
참고문헌 (R)	31.07%	0.24%	61.6%
논문제목 (T)	54.47%	3.59%	38.01%
소속 (A)	83.66%	4.74%	11.19%
전자메일 (E)	77.32%	5.58%	16.39%
공저자 (C)	83.8%	1.2%	14.09%

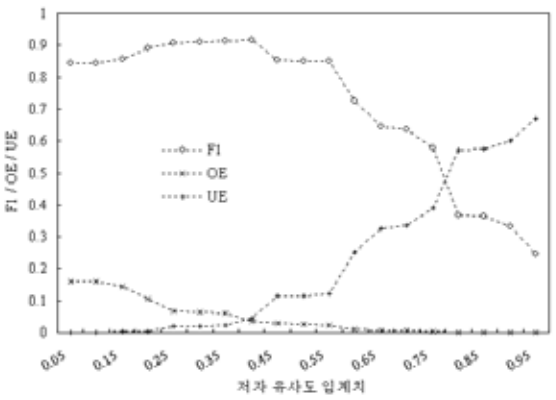
표 1은 여섯 개 개별 자질의 저자 식별 성능을 보여 준다. 단일 자질로써는, 저자 중의성 해소 측면에서, 공저자 자질이 가장 효과적이었으며, 소속, 전자메일이 그 뒤를 이었다. 이 실험에서의 자질 사용은 순전히 임의의 두 군집을 병합하기 위한 용도임을 감안할 때, 표 1의 과소 및 과다군집오류는 각각 자질의 보편성과 위험성으로 해석될 수 있다. 즉, 소속 자질은, 약 5%의 저자쌍에 대해 군집화 오류의 위험을 안고서 약 90%의 저자쌍에 대해 군집화를 적용할 수 있는 것이다. 자질의 위험성이 낮다는 것은, 그 자질이 적용될 때, 저자 식별 능력이 높음을 의미하는 것으로, 실험에서는 참고문헌, 공저자, 논문제목 순으로 나타났다. 표 1에서 소속이나 전자메일주소 자질의 위험성이 적지 않은 것은, 일반적 직관과 상치되는 결과인데 이에 대해서는 보다 깊은 데이터 분석이 필요한 것으로 판단된다. 자질의 보편성 측면에서는, 소속, 공저자, 전자메일이 높았다. 그러나, 소속, 전자메일 자질의 사용을 위해서는, 원문 텍스트의 존재 여부 및 자동 추출의 오류를 감안해야 한다. 따라서, 서지 내적 자질 중에서는 공저자가 가장 보편적이며 식별 오류의 위험성이 낮은 자질로 평가할 수 있다. 자질의 보편성이 가장 낮은 자질은 참고문헌이었는데, 이는 본 실험의 대상인 학술대회 논문의 경우 지면 제한으로 인해 충분한 인용이 포함되지 않아 발생한 것으로 판단된다.

[표 2] 이중 자질 결합의 저자 식별 성능

(각 셀은 행과 열의 자질을 병합 사용했을 때의 F1값이다. F1 값에 붙은 두 위첨자들은 차례로 그 셀에 대응하는 이중 자질의 사용이 행 자질 및 열 자질의 사용과 비교하여 통계적 유의미한 차이[4]가 있음을 가리킨다. 저자 유사도 임계치=1)

	E	A	T	R	P
C (83.8%)	90.4%++	91.4%++	85.6%++	85.2%++	84.5%++
E (77.3%)		87.5%++	82.4%++	78.9%++	80.1%++
A (83.7%)			85.9%++	85.1%++	84.8%++
T (54.5%)				57.6%++	73.4%++
R (31.1%)					61.9%++
P (57.6%)					

표 2는 여섯 개 후보 자질 중 가능한 모든 이중 자질의 결합을 통한 저자 식별의 성능을 보이고 있다. 표에서 C, E, A, T, R, P는 각각 공저자, 전자메일, 소속, 논문제목, 참고문헌, 게재지 자질을 의미하며, 저자 유사도는 이중 자질 유사도 중 하나라도 1의 값을 가지면 1을 그 외는 0을 부여하였다. 표에서 알 수 있듯이, 이중 자질의 사용은 단일 자질의 사용과 비교하여 통계적 유의미한 차이를 보이면서 저자 식별의 성능을 향상시켰다. 이는 여섯 가지 자질들이 저자 식별에 기여하는 측면이 상호 간에 상이함을 나타낸다. 이중 자질의 사용에 있어, 90% 이상의 성능을 보인, 공저자와 소속, 혹은 공저자와 전자메일의 결합이 가장 좋았다.



▶▶ 그림 1. 다중 자질 결합의 저자 식별 성능

그림 1은, 여섯 개 자질을, 표 1의 개별 자질이 보인 F1 성능에 비례하는 값을 자질 유사도 가중치로 고려하여 결합했을 때, 저자 유사도 임계치의 변화에 따른 저자 식별 성능을 보여준다. 군집화에 사용된 군집 병합 임계치에 해당하는 저자 유사도 임계치가 증가할수록 저자 식별력은 서서히 증가하다가 0.4를 정점으로 가파르게 감소하고 있으며, 과소군집오류의 추이 또한 0.4를 기점으로 급상승하고 있다. 이는, 표 2와 그림 1의 최고 성능을 비교하여 감안할 때, 주요한 두세 자질이 저자 중의성 해소에 결정적 기여를 하고 있으며, 나머지 자질들의 추가 사용은 평균적으로 잉여적임을 암시한다. 그러나, 이것이 공저자, 소속, 전자메일 등 주요 자질의 저자 식별력이 나머지 자질들의 저자 식별 능력을 완전히 포함하고 있음을 의미하는 것은 아니다. 이에 대해서는 별도의 연구가 요구된다.

4) 99% 유의수준에서 Wilcoxon signed rank test 수행하였다.

V. 결 론

본 연구는 저자 식별에 있어 서로 다른 자질이 미치는 상관관계를 대용량 평가셋을 통해 살펴보았다. 실험을 통해, 자질들은 출현 보편성과 적용 위험성이 상이함을 알 수 있었으며, 타자질에 비해 공저자, 소속, 전자메일이 저자 중의성 해소에 크게 기여하는 것으로 나타났다.

▌참 고 문 헌▐

[1] Aswani, N., Bontcheva, K., & Cunningham, H. (2006). Mining information for instance unification. ISWC-2006, Nov. 5-9, GA:USA, pp.329-342.

[2] Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. IIWeb-2007, Jul. 23, Vancouver:Canada.

[3] Guha R., & Garg, A. (2004). Disambiguating people in search. WWW-2004, May 17-20, NY:USA.

[4] Han, H., Giles, C. L., & Zha, H. (2003). A model-based k-means algorithm for name disambiguation. Semantic Web Technologies for Searching and Retrieving Scientific Data, Oct. 20, Florida:USA.

[5] Han, H., Giles, C. L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. JCDL-2004, Jun. 7-11, AZ:USA, pp.296-305.

[6] Huang, J., Ertekin, S., & Giles, C. L. (2006). Efficient name disambiguation for large scale databases. PKDD-2006, Sep. 18-22, Berlin:Germany, pp.536-544.

[7] Kanani, P., McCallum, A. & Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the Web. IJCAI-2007, Jan. 6-12, Hyderabad:India, pp.429-434.

[8] Lee, D. W., On, B. W., Kang, J. W., & Park, S. H. (2005). Effective and scalable solutions for mixed and split citation problems in digital libraries. IQIS-2005, Jun. 17, Maryland:USA, pp.69-76.

[9] McRae-Spencer, D. M., & Shadbolt, N. R. (2006). Also by the same author: AKTiveAuthor, a citation graph approach to name disambiguation. JCDL-2006, Jun. 11-15, NC:USA, pp.53-54.

[10] Song, Y., Huang, J., Councill, I., Li, J., & Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. JCDL-2007, Jun. 18-23, Vancouver:Canada.

[11] Tan, Y. F., Kan, M. Y., Lee. D. W. (2006). Search engine driven author disambiguation. JCDL-2006, Jun. 11-15, NC:USA, pp.314-315.

[12] Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: a model for author name disambiguation. Journal of the American Society for Information Science and Technology, 56(2):140-158.

[13] Yang, K. H., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2006). Extracting citation relationships from Web documents for author disambiguation. Technical Report, TR-IIS-06-017, Institute of Information Science, Academia Sinica, Taipei:Taiwan.

[14] Winkler, W. E. (2006). Overview of record linkage and current research directions. Research Report Series #2006-2, Statistical Research Division, U.S. Census Bureau.