저자 식별을 위한 전자메일의 추출 및 활용

Email Extraction and Utilization for Author Disambiguation

강인수
경성대학교 컴퓨터정보학부

In-Su Kang(dbaisk@ks.ac.kr)

요약

논문의 저자는 일반적으로 저자명으로 표현되며, 저자명을 통한 저자의 표현 및 관련 논문의 검색은 해당 시스템의 정확률과 재현율을 저하시키게 된다. 이는 같은 저자명을 적는 여러 다른 형태가 존재할 뿐만 아니라, 같은 저자명으로 논문에 기술되었으나 실제 서로 다른 사람일 수 있기 때문이다. 이 문제의 해결을 위해서는, 논문의 저자로 출현하는 동일한 인명 표현을 실세계의 서로 다른 개체로 구분하는 저자 식별 처리가 필요하다. 기존 저자 식별의 자질로, 논문의 기본 서지 항목들인 저자, 논문제목, 출처 등이 사용되었으나, 저자 식별 성능 개선을 위해서는 새로운 자질의 도입이 요구된다. 이 연구에서는 한 개인의 고유 식별자로 기능할 수 있는 저자의 전자메일주소 자질을 저자 식별 문제에 적용하고자 한다. 이를 위해 논문 원문으로부터의 저자 메일주소의 추출 문제를 다루고, 추출된 메일주소 자질이 저자 식별에 미치는 영향을 대용량 테스트셋을 통해 평가하고 분석한다.

■ 중심어 : | 저자 식별 | 동명저자 | 전자메일 |

Abstract

An author of a paper is represented as his/her personal name in a bibliographic record. However, the use of names to indicate authors may deteriorate recall and precision of paper and/or author search, since the same name can be shared by many different individuals and a person can write his/her name in different forms. To solve this problem, it is required to disambiguate same-name author names into different persons. As features for author resolution, previous studies have exploited bibliographic attributes such as co-authors, titles, publication information, etc. This study attempts to apply email addresses of authors to disambiguate author names. For this, we first handle the extraction of email addresses from full-text papers, and then evaluate and analyze the effect of email addresses on author resolution using a large-scale test set.

■ keyword : | Author Disambiguation | Same-Name Authors | E-mail |

I. 서 론

논문, 연구보고서 등의 학술 정보를 검색하는 여러 응용에 있어, 탐색키로써 저자명의 사용은 검색의 성능

* 본 논문은 2008학년도 경성대학교 학술연구비지원에 의하여 연구되었습니다.
접수번호 : #080324-001
접수일자 : 2008년 03월 24일
심사완료일 : 2008년 04월 15일
교신저자 : 강인수, e-mail : dbaisk@ks.ac.kr

을 저하시키는 주요 요인이다. 이는 동일 저자명을 갖는 다수의 연구자가 존재할 뿐만 아니라, 한 연구자의 이름 표기가 단일하지 않을 수 있기 때문이다. 예를 들어, James E. Smith라는 동일 이름을 갖는 서로 다른 연구자가 미국의 하버드 대학이나 영국의 옥스퍼드 대학에 있을 수 있다. 또한, 미국 하버드 대학의 James E. Smith라는 연구자의 이름은 James E. Smith, J. E. Smith, J. Smith 등으로 다양한 표기가 가능하다. 이러한 상황에서 논문의 저자를 저자명으로 표현하고 저자명으로 논문을 검색하는 것은 검색의 정확률과 재현율을 동시에 떨어뜨리게 된다.

이의 해결을 위해서는 같은 이름을 의미하는 다양한 저자명 표기를 정규화하고, 같은 이름을 갖는 서로 다른 저자들을 구분할 필요가 있다. 전자의 문제 즉 이름 매칭(name matching)은 동일 개체를 지칭하는 서로 다른 레코드를 연결시키고자 하는 레코드 링키지 (record linkage) 분야에서 오랫동안 연구되어 왔으며[9], 문자열 부분 매칭의 다양한 기법들이 사용된다. 후자의 문제는 저자 식별(name disambiguation)이라는 이름으로 최근 들어 본격적으로 다루어지고 있는 분야이다. 이 연구에서는 후자의 문제를 다룬다.

저자 식별의 문제는 저자명, 논문 제목, 게재지명, 게재 연도 등의 다양한 저자 식별 자질들로 표현된 저자 개체들을 그들 간의 유사도를 계산한 다음 군집화 하는 방식으로 다루어진다. 저자 식별 자질에 있어 기존 연구에서는 공동 저자명, 논문 제목, 게재지 정보(게재지명, 권호, 연도) 등의 기본 서지 정보 외에도, 논문 원문으로부터 획득 가능한 저자의 전자메일주소나 소속 정보 등을 활용해 왔다. 이러한 자질 사용의 기본 가정은 해당 자질들이 저자의 실세계 신원을 결정하는데 도움을 줄 것이라는 데서 출발한다. 개인의 신원 결정에 있어, 상기의 자질들은 그 식별력의 차이가 있으며, 이 중 가장 정확한 개인 식별 정보를 제공하는 자질은 전자메일주소이다. 그러나, 전자메일주소는 일반적으로 기본 서지 레코드에 제공되지 않으므로, 그 사용에 앞서 논문원문텍스트로부터 추출될 필요가 있다.

본 연구에서는 저자 식별 문제에 있어서 전자메일주소 자질이 미치는 영향의 분석에 집중할 것이다. 이를 위해, 논문 원문으로부터의 전자메일주소의 추출 과정과 추출된 메일주소의 저자 식별로의 활용을 다룬다. 기존에 전자메일주소 자질을 사용한 연구는 있으나 [2][4][5], 원문으로부터의 메일주소 자동 추출이나 메일주소의 저자 식별력 분석에 대한 연구는 찾아볼 수 없다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개한다. 3장에서는 논문 원문으로부터의 전자메일주소의 추출 과정을 기술한다. 4장에서는 전자메일주소를 이용한 저자 식별의 실험 결과를 기술하고, 5장에서 결론을 맺는다.

II. 관련연구

학술 정보에 출현한 저자의 이름을 실세계의 저자로 구분하는 저자 식별의 문제는, 보다 큰 문제인 인명 검색 (person search)의 특별한 경우에 해당한다. 인명 검색 연구는, 최근 들어 전체 웹 문서의 검색에서 인명 검색이 차지하는 비율이 적지 않은 데 비해[3], 문자열에 기반한 현재의 웹 문서 색인 기술이 불필요하고도 부정확한 인명이 포함된 엄청난 양의 웹 문서들을 쏟아내고 있는 현실을 개선하기 위해 출발되었다. 즉 웹 문서에 출현한 인명의 경우에도 한 인명을 공유하는 다수의 사람이 존재하며, 한 사람의 인명에 대한 다양한 표기가 가능하기 때문이다.

저자 식별을 위해서는 이름 변이형들을 정규화하는 이름 매칭의 단계가 선행된다. 이름 매칭은 레코드 링키지 분야[9]의 전통적인 토픽이며, 임의의 두 이름을 문자열로 고려하여 그들의 형태적 유사도가 미리 정해진 임계 조건을 통과하는 지 검사한다. 형태적 유사도 계산에는, Soft tf/idf, Jaro, Edit distance 등이 사용되어 왔다. 한글의 경우 필명의 사용이나 오탈자를 제외하면 이름 변이형이 거의 존재하지 않으므로, 이름 매칭의 단계를 건너 뛸 수 있는 장점이 있다.

저자 식별은 이름 매칭 단계를 거쳐 동일 저자명으로 정규화된 저자 개체에 대해 군집화 기법을 적용함으로서 수행된다. 각 저자 개체는 저자 식별 자질들의 벡터

로 표현되며, 임의의 두 저자 개체쌍에 대해 저자 유사도를 계산한 다음 군집화 알고리즘을 적용한다. 저자 유사도는 식별 자질별 자질 값 유사도로부터 SVM[4] 등의 기계학습 기법이나 자질 가중치들의 일차결합수식[1][6][8]등을 통해 계산되며, 자질 값 유사도 계산에는 이름 매칭에서 사용된 형태적 유사도 계산 기법들이 주로 적용된다. 군집화 알고리즘으로는 응집형 군집법[2][7]이 주로 사용되었다.

저자 식별 자질로 논문 서지 레코드에 기술된 공동저자명, 논문 제목, 게재지명, 게재연도 등은 기본적으로 사용되며, 논문의 원문으로부터 얻어지는 저자의 전자메일주소나 소속기관 정보가 사용되기도 하였다[2,4,5]. 또한, 연구자는 자신의 홈페이지에 논문 출판 정보를 제공한다는 점을 활용하여 저자의 논문 제목을 질의어로 사용하여 검색된 해당 저자의 홈페이지 URL이나 홈페이지 내의 논문 서지 레코드를 저자 식별 자질로 활용하기도 하였다[1].

III. 논문 원문으로부터의 전자메일주소 추출

전자메일을 통한 정보교환이 보편화되면서 논문의 형식에 있어서도 저자의 전자메일주소를 기술하도록 요구하고 있으며, 최근 출판되는 거의 모든 논문들은 저자의 전자메일주소를 포함하고 있다. 메일주소는 인터넷 공간에서 한 개인에게 고유하게 부여된 메시지 배달의 목적지 주소이므로 개인의 고유 식별자로 사용될 수 있다. 저자 식별 측면에서 전자메일주소와 같이 저자 개인의 고유한 식별자로써 기능하는 다른 식별 자질은 없을 것이다.

저자의 전자메일주소는 논문의 기본 서지 레코드에 포함되어 있지 않으므로 논문 원문으로부터 추출되어야 한다. 전자메일주소는 @라는 특수 기호를 중심으로 좌우에 사용자 아이디와 도메인명으로 이루어진 문자열로써 그 형식은 인터넷 표준 규정집 RFC-2822[1]에 기술되어 있다. 따라서 원문으로부터 전자메일의 추출은, 논문의 원문 텍스트 내에서 @문자를 포함하는 연속된 문자열(들)을 추출함으로써 쉽게 획득될 것으로 판단할 수 있다. 그러나 논문 원문의 바이너리 파일을 원문 텍스트로 변환하는 과정에 오류가 발생할 수 있으며, 논문 저자들이 논문에 기술하는 메일주소에는 다양한 기호들이 추가로 부착되는 경우가 많아, 단순한 추출 기법으로는 실효를 거두기 어렵다.

이 연구에서는 논문의 원문 텍스트에 실제 기술되는 전자메일의 형태적 특성을 반영하여 아래와 같은 추출 기법을 사용하였다.

- 기본규칙(Baseline): 아래의 정규식으로 매치되는 문자열(들)을 전자메일주소로 추출하고, 후처리로 부적절한 '.'의 사용이 포함된 추출 결과를 수정(예: abc@xyz.com.-> abc@xyz.com)하거나 메일주소 후보에서 제거
  (예: .abc@xyz.com, abc..cde@xyz.com)한다.

  [0-9a-zA-Z-._]+@[0-9a-zA-Z-.]+

- 단일화규칙(Individual): 아래와 같은 복합 메일 주소 기술 형식을 처리하여 개별 메일 주소들로 생성해 낸다. 복합 메일 주소에서 @ 문자 앞의 그룹화 시작 기호로 '{', '[', '('와 대응되는 종료 기호로 '}', ']', ')'를 사용한다.

  {abc, bcd, def}@xyz.com

- 각주처리규칙(Superscript): 아래와 같이 저자명과 대응되는 메일주소를 표시하기 위한 지정자 문자(o, a, b 등)을 추출된 메일 주소에서 제거한다. 이는 아래 예에서처럼 원문 바이너리 파일과 달리 원문 텍스트 파일에서는 윗 첨자와 같은 서식 정보가 누락되기 때문이다.

  <논문 원문 바이너리 파일>

  홍길동[o], 김철수[a], 이영희[b]

  {gdhong[o], cskim[a]}@xyz.com, yhlee[b]@abc.edu

  <논문 원문의 텍스트 파일>

  홍길동0, 김철수a, 이영희b

  {gdhongo, cskima}@xyz.com, yhleeb@abc.edu

1) http://tools.ietf.org/html/rfc2822/

• 포맷팅규칙(Format): 아래와 같이 논문 원문이 텍스트 파일로 변환되는 과정에 메일주소의 @ 문자 앞이나 뒤에서 줄바꿈이 발생되는 현상을 처리하여, 원 메일주소를 복원한다.

<논문 원문 바이너리 파일>

yhlee@abc.edu

<논문 원문의 텍스트 파일>

yhlee

@abc.edu

IV. 실험

원문으로부터의 전자메일 추출과 이후 저자 식별의 성능 평가를 위해 한국과학기술정보연구원에서 저자 식별을 위해 구축한 평가셋을 사용하였다. 이 평가셋은 1999년부터 2006년까지 개최된 국내 IT 관련 주요 9개 학회 29개 학술대회에 발표된 논문들을 대상으로 한 것이며, 각 학술대회 발표논문집의 원문 CD를 입수하여 논문의 원문 바이너리 파일(예: HWP, DOC, PDF 등)로부터 다음과 같은 논문 서지 레코드를 수작업으로 구축한 것이다.

• 논문서지레코드 = (논문제목, 게재지명[2], 저자 순번, 저자명, 저자식별ID, 저자 전자메일주소, 원문파일패스)

위의 논문서지레코드 항목에서 알 수 있듯이 개별 저자명에 대해, 저자의 전자메일주소와, 실세계에서의 서로 다른 저자에 대해 서로 다른 저자식별ID가 부여되어 있다. 실세계 저자 식별을 위해, 먼저 동명 저자들에게 모두 다른 식별자를 부여한 다음 국가과학기술인력 종합정보시스템[3], 홈페이지 검색, 서지 메타데이터 등을 참조하여 동일인임이 판명된 저자들을 하나의 식별자 아래로 병합하는 과정을 거친 것이다.

이 연구에서는 상기의 평가셋 중 원문파일로부터 텍스트 변환이 가능한 논문 7,677편을 추출하여 사용하였다. 그 이유는 이 연구의 메일주소추출기법을 적용하기 위해서는 원문의 텍스트 파일이 요구되기 때문이다.

표 1. 테스트셋 통계

논문수	저자 개체 출현수	동명 저자 그룹수	실세계 저자수
7,677	20,614	5,164	8,307

[표 1]은 실험 대상 논문 7,677편에 대한 통계치를 보여준다. 즉 7,677편의 논문에 대해, 8,307명의 실세계 저자들이 20,614개의 저자 개체와 5,164개의 동명 저자 그룹으로 출현하였다. 이 테스트셋은 한글에 국한된 것이어서 범용성이 떨어지긴 하나, 저자가 아는 범위에서 저자 식별 분야에서 전세계적으로 가장 큰 규모이다.

표 2. 전자메일추출 성능

추출방법	Paper단위			Email단위		
	Rec.	Pre.	F1	Rec.	Pre.	F1
B	42.22%	44.43%	43.30%	31.48%	73.61%	44.10%
B+I	74.51%	65.00%	69.43%	71.63%	70.96%	71.29%
B+I+S	81.82%	72.57%	76.92%	80.28%	80.69%	80.48%
B+I+S+F	83.56%	73.74%	78.35%	82.64%	79.28%	80.92%

[표 2]는 논문 원문으로부터의 전자메일주소 추출의 성능을 보여준다. 표에서 기호 B, I, S, F는 각각 앞 장에서 기술한 메일주소 추출기법들 Baseline, Individual, Superscript, Format에 해당한다. Paper단위는 개별 논문 단위로 전자메일 추출의 성능(Recall, Precision, F1)을 계산하여 평균을 취한 것이다. Email단위는 전체 논문에 대해 시스템이 추출한 전자메일의 전체 개수, 정답 전자메일의 전체 개수, 그리고 시스템 추출 메일주소 중 정답과 일치하는 개수(정답과의 일치 여부는 개별 논문 단위로 검사된다)로부터 계산된 것이다. 실험은, 원문파일이 존재하는 전체 7,677편의 논문 중 전자메일주소가 기술된 논문 6,727편(87.6%)에 대해 추출 성능을 평가한 것이다.

[표 2]의 실험 결과에서 B와 B+I를 비교하면, 논문에서의 복수 저자들의 전자메일 기술 형태가 동일 도메인명을 갖는 전자메일들을 그룹화하여 표기하는 경우가

2) 게지지명은 전거통제를 하였다.

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

많음을 알 수 있다. 이러한 그룹화로 인해 복수 저자명의 출현 순서가 전자메일의 출현 순서와 일치하지 않는 현상이 발생된다. 이러한 이유로 저자명과 전자메일을 대응시키기 위해 위첨자 기호들이 쓰이게 된다. 표의 실험 결과에서 B+I과 B+I+S를 비교하면, 이러한 위첨자 기호들의 제거 또한 전자메일의 원형태를 복원하는데 효과적임을 알 수 있다. 물론 위첨자 기호의 사용은 저자명과 소속의 대응을 위해 사용되기도 하지만, 이 연구에서는 소속기관은 저자식별 자질로 다루지 않았다. 비정상적으로 삽입된 줄바꿈 문자의 처리 기법(F)의 적용은 B+I+S 기법의 성능을 크게 향상시키지 못했는데, 이는 그러한 오류의 발생 비율이 드물기 때문인 것으로 판단된다.

다음으로 자동 추출된 전자메일 자질을 사용하여 저자 식별의 성능을 실험하였다. 이를 위해 5,164개의 동명 저자 개체 그룹 각각에 대해 단일 링크 응집형 군집법을 적용하여 저자 군집을 생성하여 테스트셋의 정답과 비교하였다. 응집형 군집법은 초기에 개별 저자 개체 각각을 하나의 군집으로 가정하고 출발하여, 군집간 유사도의 최대값이 임계치를 넘지 않을 때까지 최대 유사도를 갖는 군집쌍을 하나의 군집으로 병합시키는 과정을 반복하게 된다. 군집 병합시 각 군집이 표현하는 저자 벡터들이 병합되는데 대응되는 자질 별로 자질값들을 유니언하는 방식으로 병합된다.

각 저자 개체는 공동저자, 논문제목, 게재지명, 전자메일주소의 네 자질 벡터로 표현되며, 각 자질은 복수개의 자질 값을 가질 수 있다. 자질 유사도는 대응하는 자질에서 일치하는 자질 값이 하나 이상 존재할 경우 1 그렇지 않은 경우 0의 값을 부여하였다. 저자 유사도는 자질 유사도들의 가중치 기반 일차결합수식을 사용하여 계산하였다. 최적 가중치 결정을 위해 먼저 각 자질에 대해 0, 0.1, 0.2, ..., 0.8, 0.9, 1의 가중치 값들 중 하나를 할당하여 저자 유사도를 계산한 다음, 0.05, 0.1, 0.15, 0.2, ..., 0.9, 0.95의 저자 유사도 임계값 범위에서 최고의 저자 식별 성능을 보이는 가중치 값들의 조합을 선택했다. 군집 유사도는 각 군집이 한 명의 저자를 표현하고 있으므로, 상기의 저자 유사도 정의와 동일하다.

다른 자질들과 달리 논문제목의 경우 자질값을 추출할 필요가 있다. 논문제목 자질의 값으로 논문 제목의 각 어절에 대해 조사/어미 사전을 통해 최장 조사/어미를 절단한 다음, 음절 바이그램을 추출하여 사용하였다. 두 논문제목의 음절 바이그램 공유 비율이 임계치 0.0 84)이상일 때 자질 유사도 1을, 그렇지 않은 경우 0을 할당하였다.

저자 식별 성능 평가를 위해서는 F1 지표를 사용하였다. F1은 동일 정답 군집 내에 있는 임의의 저자 개체쌍이 동일 시스템 (출력) 군집 내에서 발견되는 비율인 재현율(recall)과, 동일 시스템 (출력) 군집 내에 있는 임의의 저자 개체쌍이 동일 정답 군집 내에서 발견되는 비율인 정확률(precision)의 조화평균을 계산한 것이다. 또한, 군집 오류 분석을 위해 과다군집오류(over-clustering error, OE)와 과소군집오류(under-clustering error, UE) 지표를 사용한다. OE는 전체 개체쌍 중 다른 군집에 속해야 함에도 시스템이 같은 군집으로 묶은 개체쌍의 비율이며, UE는 전체 개체쌍 중 같은 군집에 속해야 함에도 시스템이 다른 군집으로 분리시킨 개체쌍의 비율이다.

표 3. 저자 식별 성능

자질	OE	UE	Recall	Precision	F1
C	1.10%	14.20%	82.70%	86.70%	83.78%
T	4.49%	32.38%	60.04%	64.77%	61.07%
P	5.61%	35.67%	54.44%	63.85%	56.89%
aE	0.95%	31.95%	63.47%	68.09%	**64.78%**
aE+C	1.72%	6.32%	91.10%	92.53%	**91.47%**
aE+T	5.74%	14.25%	79.61%	79.95%	78.97%
aE+P	7.63%	13.5%	78.72%	79.37%	78.03%
C+T	5.85%	7.50%	87.03%	86.55%	86.00%
C+P	8.16%	6.39%	86.23%	84.96%	84.78%
T+P	10.67%	13.43%	76.91%	75.80%	75.23%
C+T+P	2.60%	10.52%	85.87%	87.93%	**86.22%**
aE+C+T+P	3.22%	4.85%	91.83%	91.93%	**91.48%**
mE	0.65%	33.10%	62.56%	67.46%	63.94%
mE+C	1.74%	4.05%	93.73%	94.52%	93.86%

[표 3]은 자동 추출된 전자메일 자질을 사용한 저자 식별 성능을 보여준다. 표에서 기호 C, T, P, aE, mE는 저자식별자질들로 각각 공동저자(Coauthor), 논문제목(Title), 게재지명(Publication), 자동추출 전자메일주소

4) 이는 별도 실험에서 최적화된 임계치이다.

(automatically-extracted Email), 수동추출 전자메일주소(manually-extracted Email)를 의미한다.

단일 자질 측면에서의 저자 식별력은 C, aE, mE, T, P 순이었다. 메일주소보다 공동저자 자질이 더 효과적이었던 것은, 메일주소가 주민등록번호처럼 개인의 고유ID 역할을 할 것이라는 관점에서는 이해하기 힘든 부분이다. 그러나, 전술한 바에 따르면 메일주소 자질이 출현하지 않은 논문이 전체 논문의 12.4%를 차지하고 있다. 또한, 소속의 변경, 이중 소속, 혹은 웹메일주소 사용 등의 이유로 한 저자가 복수개의 메일주소를 소유할 수 있다. 이러한 이유들로 인해 메일주소 자질이 한 저자의 유일한 고유 ID로써의 역할을 충분히 발휘하지 못한 것으로 판단된다.

또한, 예상과 달리 자동추출 메일주소 자질(aE)의 사용이 수동추출 메일주소(mE)의 경우보다 더 성능이 좋았다. 그 이유는 특정한 한 저자 개체에 대해, aE 자질값으로 그 저자개체의 논문에서 추출된 모든 메일주소를 사용하였고, mE 자질값으로는 그 저자 개체에 대응되는 하나의 메일주소만을 사용했기 때문이다. 이는 aE 자질의 경우 한 논문에서 추출된 복수개의 메일주소 중 특정 저자 개체에 대응되는 메일주소를 정확히 추출하기 힘든 이유 때문이었다. 그러나, 결과적으로 한 논문에서 추출된 모든 메일주소를 사용하는 것이 80%정도의 추출 성능[표 2]을 감안할 때, 한 저자 당 단일 메일주소의 사용보다 월등히 효과적임을 보였다. 그 이유는 한 논문에서 추출된 전체 메일주소들은, 한 저자 개체의 공동저자 자질을 표현하는 또 다른 형태이기 때문이다. 즉 aE는 전자메일주소자질과 공동저자자질을 동시에 사용한 효과를 보인 것이며, 그렇다 하더라도 여전히 추출된 메일주소의 오류로 인해 공동저자와 메일주소의 온전한 결합인 mE+C의 성능에는 크게 못 미치고 있다.

[표 3]은 메일주소와 타자질의 이중 결합이 저자 식별에 미치는 효과도 보이고 있다. 이중 결합에 있어 단일 자질에서 상위 저자식별력을 보인 공동저자와 메일주소 자질의 결합(aE+C)이 가장 효과적이었으며, 전자메일과 다른 자질의 이중 결합의 경우도 개별 단일 자질의 사용과 비교할 때 월등히 좋은 효과를 보였다. 특히 전자메일과 공동저자의 결합은 자동 추출된 전자메일(aE)을 사용하더라도 90%이상의 성능을 보이고 있는데 이는 실용적인 저자 식별기의 개발 가능성을 시사하는 결과로 판단된다. 또한 자동 추출된 전자메일의 사용은 수동 추출된 전자메일의 사용과 비교하여, 저자 식별력 측면에서 크게 뒤떨어지지 않음을 보여 준다. 이는 적어도 공저자자질과 결합될 경우 원문으로부터의 메일주소 추출성능이 저자 식별의 전체 성능에 민감하게 작용하지 않을 수 있음을 의미한다. 즉 단일 저자 논문의 경우 그 저자의 전자메일이 저자 식별에 있어 가장 확실한 실마리가 될 수 있겠으나, 다수 저자 논문의 경우 추출된 전자메일이 정확치 않더라도 공동저자(명) 자질이 부가적인 저자 식별력을 제공할 수 있기 때문이다. 이 연구에서 사용한 테스트셋의 경우 단일 저자 논문이 전체의 4.6%에 불과한 것도 메일주소 자질의 충분한 효과를 보일 수 없는 이유 중 하나이다. 그러나, 단일 저자 논문이 적은 것은 학제간의 연구와 공동연구의 증가라는 최근의 현실을 반영하는 것이기도 하다.

저자 식별 성능의 오류 정도를 가늠하는 지표인 OE와 UE 관점에서, 메일주소와 타자질의 이중 결합은 개별 자질들과 비교했을 때 OE를 소폭 희생(0.62~2.02%)하면서 UE를 대폭 감소(7.88~22.17%)시키는 효과가 있음을 보이고 있다. 이는 메일주소의 사용이 저자 식별의 자질 부족 문제를 완화시키는 데 있어 부정적 측면보다 긍정적 효과가 더 큼을 보이는 것이다. 메일주소 사용이 보이는 과다군집오류(OE)의 유형에 대해서는 향후 추가 분석이 요구된다. 상기의 메일주소가 보인 효과는 [표 3]의 다른 다중 자질 결합들에서도 관찰되는데 이는 저자 식별에서 자질 부족 문제가 심각함을 의미한다.

또한, [표 3]에서 보인 이중 자질의 결합 성능들은 논문의 기본 서지 항목들인 C, T, P 자질에 기반한 저자 식별 성능이 논문 원문파일로부터 추출된 메일주소 자질의 사용을 통해 성능차 5%이상으로 향상될 수 있음을 보인다(C+T, C+P, T+P 등과 aE+C를 비교하라).

원문으로부터의 메일주소추출의 성능이 저자 식별의 성능에 민감하게 작용하지 않을 수 있다는 이전 단락의

주장에 대한 실험적 증거를 더 찾기 위해, 서로 다른 성능을 보이는 자동추출 메일주소들을 저자 식별에 적용하여 비교해 보았다. [표 4]는 그 결과이다.

표 4. 메일주소추출방법 vs. 저자 식별 성능

메일주소 추출방법	Email 단위 추출 성능 (F1)	저자 식별 성능 (F1)	
		단일자질 (aE)	이중자질 (aE+C)
B	44.10%	38.68%	87.21%
B+I	71.29%	61.51%	90.94%
B+I+S	80.48%	62.81%	91.24%
B+I+S+F	80.92%	64.78%	91.47%

표에서 알 수 있듯이, 복합메일주소표기에 대한 처리기법(I)의 적용 이후부터 메일주소 추출의 F1 성능의 변화는 단일자질로 메일주소를 사용한 저자 식별의 성능에 민감한 영향을 미치지 못하고 있다. 공저자 자질과 메일주소 자질을 이중 결합한 저자 식별의 경우에도 비슷한 결과를 보였다. 이러한 결과는 전술한 바와 같이, 한 저자 개체에 대한 aE 자질값으로 그 저자의 논문에서 추출된 모든 메일주소를 사용했기 때문인데, 근원적으로는 한 논문 내의 메일주소들이 공동저자 자질로써 동시에 기능했기 때문이다. 즉 적어도 이 논문의 테스트셋에 한해서는 전체적으로 메일주소 추출의 성능이 낮더라도 논문 한 편 내에서 평균적으로 70% 이상의 메일주소를 정확히 추출해 낸다면 그러한 메일주소들이 공저자 자질로 충실히 기능한다는 것이다.

더욱이 44%에 불과한 추출 성능을 보인 베이스라인 추출기법(B)을 통해 얻어진 메일주소 자질의 경우에도 공저자와 결합될 때 87%대의 저자 식별 성능을 보였다. 이는 공저자와 메일주소의 이중결합을 사용한 저자 식별의 성능은 메일주소 추출의 성능 변화에 둔감하게 반응함을 의미한다. 그러나, 이것이 메일주소 추출의 성능 개선의 불필요성을 주장하는 것은 아니며, 실용적 수준의 저자 식별기 개발에 있어 자동 추출된 메일주소의 유용성과 필요성이 적지 않음을 의미한다.

마지막으로 [표 3]은 다수 자질의 결합(aE+C+T+P)이 저자 식별에 미치는 효과도 보이고 있는데, 본 실험에서는 논문제목(T)이나 게재지명(P) 자질이 공동저자와 전자메일 자질의 결합이 갖는 저자 식별에 있어 잉여적 기여를 하고 있음을 보였다. 또한, 저자 식별에 있어 논문제목(T)의 사용은 한 저자는 유사한 토픽을 다룬다는 가정에서 출발한 것인데, 이 가정을 충실히 달성하기 위해서는 논문제목으로부터의 토픽 추출 및 표현을 본 실험의 경우보다 더 개선된 방식으로 수행해야 할 것이다. 게재지명(P)의 사용은 한 저자는 그 연구 결과를 제한된 몇 개의 게재지에 제출한다는 가정에서 출발한 것인데, 본 실험에서 사용한 학술대회 논문들의 경우 한 학술대회의 투고분야별로 게재지 표현을 세분한다면 자질의 효과를 더 높일 수 있을 것이다. 정리하면 전자메일과 공동저자자질이 결합된 상황에서 논문제목과 게재지명을 추가 자질로 결합시키는 것은, 향후 다수 자질의 결합 기법이나 논문제목 및 게지지명 자질의 세련된 표현 기법에 대한 추가 연구를 통해 개선시킬 필요가 있다.

V. 결론

이 연구는 전자메일주소가 한 개인의 고유한 식별자로 기능할 수 있다는 가정 하에 논문 저자의 식별을 위한 자질로 전자메일 주소의 사용을 시도하였다. 이를 위해, 저자의 논문 원문으로부터 전자메일주소를 추출하는 기법을 개발하였고, 최고 80%대의 추출 성능을 보였다. 또한 추출된 메일주소 자질은 저자 식별에 적용됐을 때 기본 서지 항목들로부터 얻을 수 있는 저자 식별의 성능을 5%이상의 차이로 향상시킴으로써 메일주소 자질의 효용성을 보였다.

메일주소 추출 성능의 개선이 저자 식별의 성능에 유의미한 차이를 가져올 지에 대해서와 저자 유사도를 계산하기 위한 다중 자질의 결합 기법과 관련하여 추가 연구가 요구된다. 최근 Song[7]은 논문 원문 텍스트로부터의 토픽 추출에 PLSA, LDA 등의 기법을 적용하여 저자 식별 성능을 개선했다고 보고하였다. 이와 관련하여 향후 논문제목, 초록 및 원문 텍스트 전문으로부터 토픽을 추출하는 기법을 통해 저자 식별을 위한 토픽 자질의 적용에 대한 연구도 절실히 요구된다.

참 고 문 헌

[1] N. Aswani, K. Bontcheva, and H. Cunningham, "Mining information for instance unification," ISWC-2006, pp.329-342, 2006.

[2] A. Culotta, P. Kanani, R. Hall, M. Wick, and A. McCallum, "Author disambiguation using error-driven machine learning with a ranking loss function," IIWeb-2007, 2007.

[3] R. Guha and A. Garg, "Disambiguating people in search," WWW-2004, 2004.

[4] J. Huang, S. Ertekin, and C. Giles, "Efficient name disambiguation for large scale databases," PKDD-2006, pp.536-544, 2006.

[5] P. Kanani, A. McCallum, and C. Pal, "Improving author coreference by resource-bounded information gathering from the Web," IJCAI-2007, 2007.

[6] D. Lee, B. On, J. Kang, and S. Park, "Effective and scalable solutions for mixed and split citation problems in digital libraries," IQIS-2005, pp.69-76, 2005.

[7] Y. Song, J. Huang, I. Councill, J. Li, and C. Giles, "Efficient topic-based unsupervised name disambiguation," JCDL-2007, 2007.

[8] V. Torvik, M. Weeber, D. Swanson, and N. Smalheiser, "A probabilistic similarity metric for Medline records: a model for author name disambiguation," J. of the American Society for Information Science and Technology, Vol.56, No.2, pp.140-158, 2005.

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

저 자 소 개

강 인 수(In-Su Kang) 정회원

- 1995년 2월 : 경북대학교 컴퓨터공학과(공학사)
- 1999년 2월 : 포항공과대학교 컴퓨터공학과(공학석사)
- 2006년 2월 : 포항공과대학교 컴퓨터공학과(공학박사)
- 1995년 ~ 1997년 : (주)포스데이타
- 1999년 ~ 2001년 : 포항공과대학교 학술정보원
- 2006년 ~ 2008년 : 한국과학기술정보연구원
- 2008년 ~ 현재 : 경성대학교 컴퓨터정보학부

<관심분야> : 자연어처리, 시맨틱 웹, 정보검색