국내 학술논문의 동명이인 저자명 식별을 위한 방법

A Method for Same Author Name Disambiguation in Domestic Academic Papers

신 다 예 (Daye Shin)*
양 기 덕 (Kiduk Yang)**

초 록

저자명 식별이란 다른 이름으로 표기된 한 명의 개인을 식별하는 것과 같은 이름을 가진 서로 다른 저자들을 각기 구별된 개인으로 분류하는 것으로, 저자의 연구 목록 및 연구 업적 평가, 특정 분야의 전문가를 검색하거나, 인용색인과 같은 학술 정보 서비스의 원활한 운영을 위해 반드시 해결해야 할 문제이다. 본 연구는 단순 머신러닝만을 사용한 실험 결과와 휴리스틱 방식으로 데이터 셋의 오류 수정 및 정규화 작업을 이후 머신러닝의 처리 과정에 룰 베이스 기반의 규칙을 부여한 저자명 식별 실험의 결과의 비교를 통하여, 인간의 개입이 머신러닝의 단점을 보완하고 저자명 식별 성능을 향상시킬 수 있는지 알아보았다. 그 결과 F-measure 0.1 이상 향상시킨 정규화 된 email기반의 룰 베이스 저자식별 결과로 정규화 과정과 휴리스틱 설정에 필요한 인간의 패턴인식과 추론능력이 머신러닝의 단점을 보완해줄 수 있음에 대한 가능성을 나타내었다.

ABSTRACT

The task of author name disambiguation involves identifying an author with different names or different authors with the same name. The author name disambiguation is important for correctly assessing authors' research achievements and finding experts in given areas as well as for the effective operation of scholarly information services such as citation indexes. In the study, we performed error correction and normalization of data and applied rules-based author name disambiguation to compare with baseline machine learning disambiguation in order to see if human intervention could improve the machine learning performance. The improvement of over 0.1 in F-measure by the corrected and normalized email-based author name disambiguation over machine learning demonstrates the potential of human pattern identification and inference, which enabled data correction and normalization process as well as the formation of the rule-based diambiguation, to complement the machine learning's weaknesses to improve the author name disambiguation results.

키워드: 저자명 식별, 머신러닝, 룰 베이스 방법, 휴리스틱
Author Name Disambiguation, Machine Learning, Rule-based Method, Heuristic

---

* 경북대학교 문헌정보학과 대학원(sunset1007@knu.ac.kr) (제1저자)
** 경북대학교 문헌정보학과 교수(yangkiduk@gmail.com) (교신저자)
논문접수일자 : 2017년 11월 27일 논문심사일자 : 2017년 12월 16일 게재확정일자 : 2017년 12월 19일
한국비블리아학회지, 28(4): 301-319, 2017. 〔http://dx.doi.org/10.14699/kbiblia.2017.28.4.301〕

1. 서 론

1.1 연구 목적 및 배경

저자명 식별은 학술 정보 내에서의 인명을 검색하는 것으로(강인수 2008b), 다른 이름으로 표기된 한 명의 개인을 식별하는 것뿐만 아니라 동일 이름을 가진 서로 다른 저자들을 각기 구별된 개인으로 분류하는 것을 말한다.

이러한 저자명 식별은 학술 정보 관련 기관이 저자와 논문을 효율적으로 관리하기 위해 반드시 해결해야 하는 요소로, 저자를 정확히 식별하는 것은 각 저자의 논문 목록과 각 인용 지수 등의 분별적 파악이 가능함을 의미하고, 연구업적 평가와 계량서지학적 지표의 확실한 추출을 위한 기반이 된다. 또한 저자명 키워드 검색 결과의 정확도 향상 측면에서도 검색의 정확률을 저하 문제를 해결하는 수단이 될 수 있다. 그러나 전자 저널이 본격화되기 전 출판된 대부분의 학술 논문에 저자의 소속기관, email 등 저자명 식별에 필요한 서지 정보가 제대로 기재되어 있지 않는 점 등의 메타데이터 불완전성과 현재까지도 학술지 또는 출판사마다의 서지 정보 기재 형식과 종류가 통일되어 있지 않은 경우가 다소 존재하여, 데이터 추출 시 정확도가 하락하고 높은 비용이 발생한다는 점, 이외 학술 논문의 폭발적 증가 및 국제적인 연구 성과 교류 등을 통한 저자명 모호성 문제는 현재 당면한 도전 과제이다.

저자명 식별 문제는 크게 표기법과 철자 변형, 철자 오류, 결혼 등으로 인한 이름 변경 및 필명 사용으로 인해 한 명의 개인이 다양한 이름으로 출판하는 경우와 서로 다른 개인들이 같은 이름을 갖는 동명이인 문제로 나눌 수 있다(Smalheiser and Torvik 2009). 이미 "영미권 출판계, 도서관계 등에서는 모호한 이름 식별이라는 의미를 갖는 'Name Disambiguation'을 매우 중요한 과제로 인식(조재인 2013)"하고 있으며, 효율적인 저자명 식별을 위한 다양한 모델을 제시해왔다. 국내에서는 도서관계는 전거통제를 중심으로, 과학기술 분야에서는 자동분류와 클러스터링을 활용한 머신러닝을 중심으로 저자명 식별을 위한 연구가 이루어지고 있다.

전통적으로 도서관에서는 사서들이 인명 전거를 통해 저자의 한자이름, 생년월일 등으로 동명의 저자를 구분하고 필명, 이명 등을 하나의 표목으로 검색될 수 있도록 참조를 작성하여 저자를 식별한다. 수동으로 수행되는 전거통제는 폭발적으로 증가하는 정보 발생량을 감당할 수 없을 뿐만 아니라 저자명 식별의 대상이 도서관 소장 자원에만 제한되어 있는 등의 문제점이 있다. 전거통제에서 확장된 방식으로는 국제표준이름식별자(International Standard Name Identifier, ISNI)나 ORCID(Open Researcher and Contributor)와 같이 개별 저자에게 주민등록번호처럼 중복되지 않은 고유 식별자를 부여하는 방법이 있다. 이상적으로는 가장 완전한 저자명 식별 방법인 저자 고유 식별자 부여 방식은 저자들의 저조한 참여율과 저자 정보를 관리하고 보존할 영구적인 조직의 부재, 비용 대비 효용성 등으로 인해 실현가능성이 낮다.

한편 머신러닝 분야에서는 논문의 메타데이터 또는 원문 텍스트에서 추출한 논문제목, 저자명, 공저자명, 학술지명, 발행연도, 전자메일

주소, 소속기관, 인용논문 등의 저자명 식별 자질(feature)들을 수집한 데이터 셋(data set)을 구축하고, 자질 간 유사도를 계산하여 유사도가 임계치 이상인 데이터를 군집화 하는 방식으로 저자명을 식별한다. 그러나 다양한 머신러닝 기법과 유사도 계산 방법을 적용한 여러 저자명 식별 연구가 진행되었으나 아직까지 모든 환경에 적용할 수 있는 식별 모델은 부재한 상태이다.

학술 정보 시스템에서는 특정 저자명으로 검색했을 때 검색한 저자명과 문자열이 같은 모든 동명이인의 연구 결과가 출력되는 불편함을 줄이기 위해 발행연도, 주제, 학술지명, 자료 유형과 같은 패싯을 통해 검색결과를 줄여주는 패싯 내비게이션 방식을 사용하고 있다. 저자명 식별 문제가 해결되기 전까지의 대안일 뿐 저자명 식별 문제를 근본적으로 해결할 수 있는 방안은 아니다.

저자명 식별 문제에 대한 기존의 연구는 반자동으로 저자명을 식별하는 전거통제 및 저자 고유 식별자 부여 방식 또는 자질값의 식별 성능을 비교하거나 머신러닝 모델별 식별 성능을 비교한 머신러닝 방식에 대한 연구가 대부분이다. 특히 머신러닝의 결과를 향상시키기 위해 자질별, 유사도 함수별, 머신러닝 모델별 비교 연구는 많이 진행되었으나, 머신러닝 자체의 한계점을 돌아보고자 하는 연구는 없었다.

본 연구에서는 인간이 저자명 식별 과정에 적극적으로 개입하여 휴리스틱 방식으로 데이터 셋의 오류 수정 및 정규화 작업을 진행하고 머신러닝의 처리 과정에 규칙을 부여한 저자명 식별 실험을 진행하는 것이 머신러닝의 단점을 보완하고 저자명 식별 성능을 향상시킬 수 있는지 알아보고자 한다. 이 연구의 목적은 머신러닝이 극복할 수 없는 데이터 문제를 인간의 패턴인식과 지식에 기반한 추론으로 도출한 룰(heuristic)로 보완하여 저자명 식별의 결과를 향상시키는 것이다.

1.2 연구 방법

머신러닝은 데이터를 스스로 학습하여 문제를 해결해나가지만 데이터 자체에 오류가 있을 경우 그 오류를 식별하지 못하며 같은 데이터라도 표기법이 다른 경우 같은 데이터로 식별하지 못한다. 이를 보완하기 위해 저자명 식별을 위한 데이터 셋에 적용할 수 있는 휴리스틱 기반의 데이터 클리닝 방법 제시하고, 저자명 식별 실험을 통해 그 효용성을 평가하였다.

또한 단순히 자질 간의 유사도를 계산하여 일정 임계치 이상의 유사도를 갖는 저자명 데이터를 군집화 하는 머신러닝 방식에 문제 해결을 위한 전제를 설정하고 그것을 기반으로 결론을 도출하는 룰 베이스 방식을 결합한 저자명 식별 실험을 진행하여 머신러닝만을 활용한 저자명 식별 실험과 그 결과를 비교한다. 이때, 룰 베이스 방식을 활용한 저자명 식별 실험은 'email 주소는 고유하므로, 같은 email 주소를 사용하는 저자는 동일인이다.'는 본 연구에서 설정한 기본 가정을 바탕으로 하며, email 주소 자질이 같은 데이터를 동일저자로 식별한 후(i.e., 유사도=1) 클러스터링을 하는 방식을 말한다.

추가로 룰 베이스 방식에서 email 자질의 정확도가 저자명 식별 성능을 향상 시키는지 알아보기 위해 원래의 email 주소 자질과 오류 수

정 및 정규화 작업을 마친 클린 email 주소 자질을 원 데이터 셋과 클린 데이터 셋 각각에 적용하여 실험하고 그 결과를 비교하였다. 종합하면, 머신러닝의 약점을 최소화하기 위해 데이터 셋 처리 과정부터 저자명 식별 과정 전반에 인간이 지식을 습득하고 처리하는 능력과 패턴을 파악하는 인지 능력 및 직관력을 적극 활용하여 데이터 셋의 패턴을 분석하고, 이를 데이터 클리닝 작업에 적용하여 머신러닝의 성능을 향상시키고자 한다.

2. 선행연구

저자명 식별에 관한 전통적인 연구 분야로는 전거통제와 저자 고유 식별자 연구가 있다. 이석형, 곽승진(2010)은 논문 메타정보 입력부터 색인, 서비스 과정에서 활용될 수 있는 반자동 전거데이터 구축 시스템을 개발하였다. 이미화(2014)는 국내에 맞는 국제표준이름식별자 체제인 KISNI를 구축하고, 서지레코드 및 전거레코드에 국제표준이름식별자 식별자를 기술할 수 있도록 KORMARC를 확장해야 한다고 주장한다. 조재인(2013)은 국내에 ORCID를 도입하고 ORCID Identity system의 파트너 시스템으로 등록함으로써 해외 학술 논문에 기재된 국내 저자명 식별까지 도모해야 한다고 주장한다.

머신러닝을 통한 저자명 식별 실험 전에 실험을 위한 기본이 되는 평가 셋 구축 연구가 진행되었다. 정한민 외(2006)는 9,484건의 논문을 대상으로 클러스터링을 활용한 대상 집합과 수작업으로 구축한 정답 집합을 비교함으로써 인력정보 구축 과정에서의 실수나 누락을 재확인한 한국어 저자명 식별 데이터 셋을 구축하였다. 강인수 외(2009)는 DBLP로부터 87만여 편의 논문서지레코드를 다운로드 받은 후, 고빈도 순으로 상위 1,000개의 저자명을 추출하여 저자명 식별을 위한 영어 이름 평가 셋을 구축하였다.

머신러닝을 활용한 저자명 식별 연구는 크게 학습된 정보를 바탕으로 데이터를 식별하도록 하는 지도 학습법과 사전 학습 없이 컴퓨터가 스스로 패턴을 찾아 군집화 하는 비지도 학습법으로 나뉜다. 또한 자질별 가중치를 다르게 부여하거나 여러 자질 결합하여 저자명 식별 성능을 높이는 방식으로 연구가 진행된다. 국내에서는 강인수와 한국과학기술정보연구원 팀에서 관련 연구가 주로 이루어졌으며, 자질별 식별 성능 비교, 유사도 계산 함수별 성능 비교, 분류 또는 클러스터링 방법 등을 다양하게 적용하여 저자명 식별 연구를 진행하였다.

이승우 외(2006)는 공저자 자질의 효용성을 증명하기 위해 소속기관과 email 주소만을 활용한 클러스터와 소속기관과 email 주소와 공저자를 모두 활용한 클러스터, 공저자만 활용한 클러스터를 생성하여 성능 비교 결과, 소속기관과 email 주소만을 사용한 경우 발생한 #under-clustering(동일인으로 그룹 짓지 못한 불일치 쌍의 수) 17,267건이 소속기관과 email 주소, 공저자를 모두 활용한 경우 2,266건으로 13%가 줄어 학술 논문에서 저자의 동명이인 식별을 위해 공저자 관계를 활용하는 것이 유용함을 증명하였다. 강인수 외(2006)는 논문의 원문 텍스트를 저자명 식별 자질로 활용하고, 동명이인을 식별하기 위해 문서 클러스터링 방

식을 활용하였다. 실험 결과, CLUTO를 활용한 클러스터링에서 군집의 개수가 3개이고 동명저자가 2회 출현했을 때 Rand-Index 값이 78.4%로 가장 성능이 좋았으며, 동명의 저자가 작성한 논문의 내용이 저자명 식별을 위한 자질로서 효과가 있음을 증명하였다.

한편 강인수(2008a, 2008b, 2009, 2011a, 2011b)는 논문의 저자 식별을 위하여 머신러닝을 활용한 다양한 연구를 진행하였다. 저자식별의 대표적 자질인 email 주소, 공동저자, 논문제목, 학술지명 등을 추출하여 머신러닝의 저자식별을 분석하였으며, 저자식별을 위한 계층적 군집법 7가지와 개체거리 함수 6가지의 저자명 식별 성능을 비교하고 평가하였다. 또한 2011년 연구에서는 선행연구를 기반 하여 동시인용 자질을 공저자와 논문제목, 학술지명 자질과 결합을 통해 저자명 식별력을 향상 시키는 새로운 알고리즘과, 저자를 노드로 하고 공저자 관계를 이용하여 노드 간 링크를 연결한 저자 그래프를 생성하는 알고리즘을 제안하였다.

신동욱(2009)은 저자를 정점으로 하고 공저자 관계를 연결선으로 하는 사회망을 구축하여 저자명을 식별하였다. 김하진, 정효정, 송민(2014)은 토픽모델링에 제3의 메타데이터를 추가한 ACT 모델과 DMR 모델을 활용하여 저자명을 식별하고 성능을 비교, 평가하였다. 김일환과 이도길(2015)은 초기 현대소설 70편에서 저자별 어휘 사용 양상을 통계적으로 분석하여 계량적 방법으로 저자의 문체적 특성을 밝힐 수 있음을 밝혔다.

저자명 식별을 위한 기존의 연구는 식별 성능이 높은 자질이나 머신러닝 모델, 기타 식별 방법을 찾는 것에 주목하였다. 본 연구에서는 머신러닝이 갖고 있는 한계점에 주목하고 고유 식별자인 email 주소 자질에 초점을 맞춘 룰 베이스 방식의 저자명 식별 방법을 제안하고 그 효과를 알아보고자 한다.

3. 데이터 셋 분석

본 연구에서는 한국과학기술정보연구원(KISTI)에서 배포(강인수 2008b, 31)한 한국어 저자명 식별 평가 셋을 활용하였다. 데이터 셋에는 23개의 학술대회 문헌집 중 텍스트를 추출할 수 없거나 원본이 존재하지 않는 논문을 제외한 데이터가 포함되어 있다. 제공 받은 데이터 셋은 텍스트 파일 형식이었으며, 데이터 셋의 구성은 다음과 같다(정한민 외 2006).

> 구분/논문식별자/논문제목/출판정보/논문파일명/논문파일확장자/저자총수/저자순번/저자명/저자식별자/Dummy/기관명/기관코드/부서명/부서코드/전자메일주소/저자명리스트

본 연구에서는 '논문제목', '저자명', '공저자(저자명리스트)', '출판정보('연도+학술지명'으로 표기되어 있음)', '소속기관(기관명)', '소속부서(부서명)', 'email 주소(전자메일주소)' 자질을 사용한다. '논문식별자', '저자식별자', '기관코드', '부서코드'는 데이터 셋 구축 시 연구자가 자체적으로 부여한 식별 정보이며, '저자식별자'는 실험 결과와 비교할 정답 데이터로 사용하고 '논문식별자', '기관코드', '부서코드', '구분', '논문파일명', '논문파일확장자', 'dummy' 필드는 본 연구에서 활용하지 않는다.

데이터 셋의 필드별 레코드 개수를 살펴보면 논문의 기본 서지사항인 논문제목, 저자명, 출판정보, 저자명리스트는 모든 레코드마다 기재되어 있어 각 필드마다 구축된 레코드 개수와 총 레코드 개수가 같으나, 원문 추출 과정이 필요한 소속기관, 소속부서, email 주소 필드는 원문의 정보 제공 정도와 원문 추출 과정에서의 누락에 따라 구축된 개수가 다르다. 데이터 셋의 총 레코드 수는 20,614개이며 소속기관은 18,161개, 소속부서는 15,402개, email 주소는 15,586개의 행에 기재되어 있다.

〈표 1〉 필드별 레코드 개수

소속기관	18,161개(88.1%)
소속부서	15,402개(74.7%)
email 주소	15,586개(75.6%)
총	20,614개(100%)

총 20,614개의 데이터 중에서 연구자가 임의로 부여한 코드를 제외한 기본 식별자인 논문제목, 출판정보, 저자총수, 저자순번, 저자명, 기관명, 부서명, 전자메일주소, 저자명리스트가 모두 같은 데이터가 377건 존재하여 원문 검색을 해본 결과 중복 입력되었음을 확인하고 두 번 이상 중복 입력된 데이터를 하나의 데이터만 남겨두고 삭제하였다. 한편, 데이터 추출 시 오류가 발생하여 논문제목, 출판정보, 저자명, 기관명, 부서명, 전자메일주소가 모두 같으나, 공저자가 일부 누락되어 저자총수, 저자순번, 저자명리스트가 다른 데이터가 48건 발견되었고, 원문 검색 후 같은 논문이라 판단되어 중복 레코드 중 하나의 레코드를 제외한 24개의 레코드를 삭제하였다. 따라서 분석에 활용된 데이터의 개수는 총 20,396개이다.

〈표 2〉 중복 삭제 후 레코드 개수

소속기관	17,957개(88%)
소속부서	15,226개(74.7%)
email 주소	15,384개(75.4%)
총	20,396개(100%)

중복 데이터를 삭제한 데이터 셋 자질값의 중복을 제외한 고유 데이터 개수를 살펴보면 고유 저자명 수는 5,164개, 고유 저자식별자 수는 8,304개, 공저자를 모두 포함한 고유 저자명 수는 9,372개, 고유 기관명 수는 535개, 고유 부서명 수는 856개, 고유 email 주소 수는 7,575개, 고유 도메인 수는 1,164개로 나타났다.

〈표 3〉 고유 자질 개수

필드명	개수
고유 저자명 수	5,164개
고유 저자식별자 수	8,304개
공저자를 포함한 고유 저자명 수	9,372개
고유 기관명 수	535개
고유 부서명 수	856개
고유 email 주소 수	7,575개
고유 도메인 수	1,164개

4. 데이터 클리닝

본 단락에서는 데이터 셋의 오류를 찾아 수정하고, 데이터 클리닝 차원에서 동일한 대상을 서로 다른 표기로 인해 일치시킬 수 없는 사례에 대해 정규화 작업을 수행하였다.

4.1 자질 수정 및 정규화

원문으로부터 데이터를 추출하는 과정 또는 원문 자체에 오타가 있을 경우에 데이터 값에 오류가 발생할 수 있으며, 기관명과 부서명, 영어 공저자명 등 기관이나 사람의 이름의 경우에 다양한 표기법으로 인해 정규화를 하지 않으면 컴퓨터 같은 기관과 사람을 같은 데이터로 인지하지 못한다. 데이터 셋 분석 결과, 데이터의 오류가 발견된 자질은 email 주소 자질이며, 정규화가 필요한 자질은 기관명 및 부서명, 영어 공저자명, 논문제목, email 주소로 나타났다. 정규화 작업은 데이터 셋에서 각 자질별로 중복을 제외한 고유 데이터를 추출한 뒤 패턴을 분석하여 같은 값의 다른 표현을 서로 연결하는 작업을 진행한 뒤, 더 이상 새로운 패턴이 발견되지 않을 때까지 같은 과정을 반복하였다.

4.1.1 기관명 및 부서명 정규화

기관명과 부서명 데이터의 특징을 분석하기 위해 데이터 셋의 모든 기관명과 부서명을 추출하여 오름차순으로 정렬한 후 패턴을 분석하였다. 추출된 기관명은 535개이고, 부서명은 856개이다. 정규화 작업은 크게 눈으로 식별 가능한 데이터를 웹 검색을 통해 확인한 뒤 정규화 리스트를 작성하는 수작업 과정과 알고리즘을 사용한 자동화 방식으로 진행되었다.

수작업 과정을 통해 기관을 대학과 대학이 아닌 기관으로 분류하고, 대학 이름이 영어 또는 영문 약자로 표기된 경우 대학 홈페이지를 참조하여 한국 표기로 변경하였다. 홈페이지에서 추가로 발견된 이형 표기들도 모두 리스트에 추가하였다. 또한 '국립보건연구원'과 '국립보건원'과 같이 유사한 이름을 가진 기관에 대해 기관 홈페이지를 참조하여 같은 기관임이 확인되면 두 기관명을 연결하였다. 수작업으로 정규화 작업을 진행한 기관명은 51개이며, 부서명의 경우 대학 또는 기관 홈페이지에 명확하게 기재되어 있지 않은 경우가 많아 확인이 어려워 띄어쓰기와 'lab', 'Lab', '랩'을 서로 다르게 표기한 경우나 띄어쓰기만 다른 경우 등 같은 부서임이 확실한 5건에 대해서만 정규화 작업을 진행하였다.

알고리즘을 활용하여 자동으로 진행한 정규화 작업은 다음과 같이 구성되어 있다.

- ㈜와 괄호 및 괄호 안의 데이터 삭제
- 키보드에 없는 ' ` ' 등의 기호 삭제
- 앞뒤 스페이스 삭제
- ' / ' 뒤에 있는 문자열 삭제

추가로 기관명의 경우는 'OO대학' 및 'OO대'를 모두 'OO대학교'로, 'uni'를 'university'로 통일하고, 영어 이름인 경우 'the' 모두 삭제하였다. 부서명의 경우 'dept', 'department', 'division', 'school', 'of', 'and'를 모두 삭제하였다. 알고리즘을 통해 정규화 한 기관명의 개수는 244개이고, 부서명은 110개이다. 정규화를 마친 후 기관명은 535개에서 419개로, 부서명은 856개에서 836개로 감소하였다.

4.1.2 영어 공저자명 정규화

한국어 이름의 경우 표기법이나 표기 순서가 같아 정규화를 하지 않아도 된다는 장점이 있으며, 만약 오타가 입력된 경우 없는 글자로 표기된 것이 아닌 이상 오타임을 식별할 수 없으

므로, 저자 이름에 대한 정규화 작업은 영어 공저자명 자질에 대해서만 진행하였다. 공저자명 자질은 저자명리스트 필드에 세미콜론(;)으로 구분되어 입력되어 있으며, 이를 모두 추출하여 가나다순으로 나열하였다. 영어로 표기된 공저자명은 모두 39개였으며, 데이터양이 적어 눈으로 확인한 결과 영어와 한국어 표기가 섞여서 기재된 레코드 2건을 원문 및 저자식별자 확인을 통해 영문 표기로 수정하였고, 이름 뒤에 '?'가 삽입된 경우 레코드 1건을 수정하여 총 3건을 정규화 하였다. 그 결과 공저자를 포함한 고유 저자명은 9,372개에서 9,369개로 감소되었다.

4.1.3 논문제목 추출 및 정규화

데이터 셋에서 박은정, 조성준(2014)이 한국어 자연어 처리를 위해 개발한 파이썬 패키지인 코엔엘파이(KoNLPY)를 사용하여 논문제목 자질값에서 명사를 추출하였다. 추출 결과 총 14,868개의 명사가 추출되었으나, 영어 단어를 조사와 함께 삭제하거나 영어와 숫자가 붙어 있는 '802.11i'와 같은 문자열을 모두 삭제하는 등 숫자와 영어 단어를 추출하는 것에 미흡한 점이 발견되었다. 따라서 숫자와 영어 단어는 따로 추출하고 중복 방지를 위해 한국어 명사 리스트에서 숫자를 모두 제거하였다. 또한 따로 추출한 영어 단어에서 단어가 하나의 알파벳인 경우 단어가 아니라고 간주하고 삭제하였다. 그 결과, 추출된 단어 수는 11,304개로 감소되었다.

4.1.4 email 주소 오류 수정 및 정규화

본 단락에서는 email 주소의 문자열 오류를 자동으로 식별하는 알고리즘 방식과 email 주소 자질과 저자명 자질의 관계를 통해 오류 패턴을 분석하는 휴리스틱 방식으로 email 주소가 가진 문제점이 무엇인지 찾고, 그 해결 방안을 모색해보고자 한다. 또한 같은 기관의 여러 도메인 주소를 정규화 하여 같은 기관 소속의 한 명의 개인이 사용자 ID는 같으나 여러 기관 도메인 주소를 사용하는 경우를 연결함으로써 같은 저자가 여러 email 주소를 사용하는 경우를 식별하였다.

1) email 주소 오류 수정 알고리즘

데이터 셋의 e-mail 주소 자질값의 오류를 분석한 결과, 문자열 앞뒤에 공백이 입력된 경우 또는 물음표나 콤마와 같은 문장부호의 삽입 오류와 같이 구두법이 잘못 된 경우 등의 9가지 오류 패턴이 다음과 같이 발견되었다.

- 문자열 앞 또는 뒤에 공백이 입력된 경우: 16개
- 온점이 아닌 콤마가 입력된 경우: 7개
- email 주소의 사용자 ID 또는 도메인 주소가 입력되어 있지 않은 경우: 2개
- @이가 없는 경우: 1개
- 도메인 주소에 1단계 도메인(최상위 도메인) 또는 2단계 도메인이 없는 경우: 1개
- @뒤에 온점(.) 또는 콤마(,)가 있는 경우: 4개
- 온점(.)이 두 개 연속 찍힌 경우: 1개
- 도메인에 온점(.)을 제외한 특수 문장부호(')가 입력된 경우: 1개
- 도메인 주소의 철자가 잘못 된 경우: 78개

구두법 오류가 발견된 email 주소는 모두 111개이며, 발견한 패턴 9개를 알고리즘으로 구축하여 데이터 셋에서 오류 email 주소를 자동으로 검색하였다. 검색된 오류 email 주소는 원문 검색과 구글 검색, 기관 홈페이지 검색 등을 이용하여 수작업으로 수정하였다.

2) email 주소 오류 수정 휴리스틱

본 단락에서는 '같은 email 주소를 사용하지만 저자명이 다른 레코드'를 오류로 가정하고 같은 email 주소이나 서로 다른 저자명 또는 저자식별자가 기재되어 있는 경우를 찾아보았다. 같은 email 주소를 사용하지만 저자명이 다른 email 주소의 수는 66개이며, 레코드 수는 303개이다. 303개의 레코드와 원문을 비교해본 결과 같은 논문의 모든 공저자에게 같은 email 주소가 기재된 경우와 원문 자체에 email 주소가 잘못 기재되어 있는 경우가 발견되었다.

공저자에게 모두 같은 email 주소가 부여된 경우는 공저자들 간에 같은 email 주소를 가질 확률은 거의 없기 때문에 같은 email 주소를 가진 서로 다른 공저자 중 반드시 한 명만이 기재된 email 주소의 확실한 실제 저자이다. 이와 같은 오류를 찾기 위해 데이터 셋에서 같은 email 주소와 같은 논문제목, 다른 저자명인 데이터를 검색한 결과, email 주소는 40건, 총 레코드 수는 164건이 검색되었고, 검색된 데이터에서 email 주소의 본래 주인을 찾기 위해 다음과 같은 과정을 진행하였다.

원문을 검색하여 email 주소 추출 과정의 오류인지 원문 자체의 오류인지 알아보고 원문 추출 과정의 오류인 경우 원문에 맞게 수정한다. 원문 추출 과정에서 발생한 기재 오류인 경우, email 주소의 사용자 ID와 저자명 이니셜이 일치하거나 유사한 레코드를 찾는다. 사용자 ID와 저자명 이니셜이 일치하는 경우 같은 저자식별자를 사용하는 다른 레코드(같은 저자의 다른 저작)의 원문을 검색하여 email 주소를 확인한다. 사용자 ID로 email 주소의 주인을 식별할 수 없는 경우 email 주소를 구글에 검색하고 저자 홈페이지 또는 같은 저자의 다른 저작을 찾아 email 주소를 확인한다. email 주소의 실제 주인을 확인되면, 같은 논문의 다른 공저자에게 기재된 email 주소를 모두 삭제하고, 실제 주인을 찾지 못한 경우 해당 논문 관련 레코드의 email 주소 자질을 모두 삭제[1] 한다. 그 결과 51개의 레코드에서 email 주소를 삭제하였고, 같은 email 주소를 사용하지만 저자명이 다른 email 주소의 수는 66개에서 36개로, 같은 email 주소를 사용하지만 저자명이 다른 레코드의 수는 303개에서 139개로 감소되었다.

한 논문의 공저자에게 동일 email 주소가 부여된 오류 이외에 동일 email 주소이나 다른 저자명을 가진 36건의 경우에 대해 위의 과정을 다시 반복 수행한다. 139개 레코드의 email 주소와 저자명을 모두 확인한 결과 확실하지 않은 email 주소를 가진 레코드 46개에서 email 주소를 삭제하였고, 같은 email 주소와 같은 논문제목, 다른 저자명에서 삭제한 51개와 함께 총 97개의 레코드에서 email 주소가 삭제되었으며 이로 인해 같은 email 주소를 사용하지만 저자명이 다른 경우인 레코드는 모두

1) 단, email 주소를 수정하거나 삭제할 때, 데이터 셋의 전자메일주소 필드의 데이터를 직접 수정하지 않고 새로운 필드를 추가하여 수정한 email 주소를 입력하였다.

제거되었다.

3) email 도메인 주소 정규화

정규화란 다르게 표기된 하나의 개체를 식별하여 연결하는 과정을 의미하나, 본 연구에서 진행하는 email 주소 자질의 정규화는 한 사람이 다양한 email 서비스 기관의 email 주소를 사용하는 경우를 제외하고, 같은 기관의 도메인 주소이나 부서나 연구팀에 따라 또는 시간이 흐름에 따라 기관 도메인 주소가 변경되었기 때문에 도메인 주소를 다르게 사용하는 경우만 그 대상으로 하였다. 즉, 같은 기관 소속의 한 명의 개인이 사용자 ID는 같으나 부서나 연구 팀에 따라 한 기관 내의 하위 도메인 주소를 사용하는 경우나 기관 도메인 주소 자체가 변경된 경우를 연결함으로써 같은 저자가 여러 email 주소를 사용하는 경우를 식별하도록 하였다. 같은 기관 email 주소의 도메인 부분을 정규화 하는 이유는 사용자 ID에서는 오타를 식별할 수 없는 것과 달리 도메인의 구두법 오류나 철자 오류 등은 식별해낼 수 있으며 기관 도메인은 도메인 주소가 변경되었더라도 기존의 기관 도메인 주소를 확인할 수 있는 반면, 다양한 email 서비스 기관의 email 주소를 사용하는 경우는 데이터 셋의 정보만으로는 확실하게 식별할 수 있는 방법이 없고, 구글 검색 또는 저자 홈페이지를 확인하더라도 누락되는 부분이 많기 때문이다.

도메인 주소 정규화를 위해 추출한 고유 도메인 주소를 오름차순으로 정렬하여 패턴을 분석하고, 레벤슈타인 거리를 사용하여 클러스터링 하였다. 이를 통해 스펠링이 유사한 도메인끼리 군집화 한 결과 'hanyang.ac.kr'의 오타인 'hanynag.ac.kr', 'sogang.ac.kr'의 오타인 'goang.ac.kr' 등의 도메인 오류 31건을 발견하고 121개의 레코드를 수정하였다. 같은 기관의 여러 도메인 주소를 찾기 위해 추출한 도메인을 기관명으로 묶어 158개의 기관에 대한 도메인 리스트를 생성하였다. 생성한 기관별 도메인 리스트 중에서 'empal.com', 'dreamwiz.com'과 같이 기관명과 관련 없는 도메인 주소를 제외한 도메인 주소를 구글에 검색하여 같은 기관의 도메인 주소가 맞는지 확인하는 작업을 진행하였다. 그 결과 30개 기관이 2개 이상의 도메인 주소를 사용하는 것이 발견되어 정규화 리스트를 작성하였고, 레코드 489개에 대한 정규화 작업을 진행하였다. 정규화 과정에서 오류를 수정하고 정규화 작업을 진행한 결과 고유한 도메인 주소의 개수는 1,164개에서 1,055개로 감소하였고 고유 email 주소 또한 7,575개에서 7,423개로 감소하였다.

자질 수정 및 정규화 과정을 통해 변경된 데이터의 변화를 요약하면, 고유한 자질의 개수 변화는 〈표 4〉와 같고, 자질값이 변경된 레코드의 개수는 〈표 5〉와 같다.

〈표 4〉 데이터 클리닝 후 고유한 자질의 개수 변화

	기관명	부서명	영어 공저자명	논문제목	도메인
수정 전	535개	856개	9,372개	14,868개	1,164개
수정 후	419개	836개	9,369개	11,304개	1,055개

〈표 5〉 자질값이 변경된 레코드의 개수

	기관명	부서명	email 주소
총 개수	17,957개	15,226개	15,384개
변경된 개수	244개(1.4%)	110개(0.7%)	707개(4.6%)

4.2 저자식별자 수정

한국과학기술정보연구원에서 제공 받은 데이터 셋의 저자식별자 필드는 연구자가 임의로 실세계 저자라고 식별된 레코드에 부여한 고유한 식별자이다. 즉 저자식별자 필드는 저자 개개인을 식별한 연구 결과로 본 연구의 저자명 식별 결과와 비교될 정답 데이터이다. 그러나 일부 오류가 발견되어 다음과 같은 방식으로 오류를 수정[2] 하였다. 오류 발견을 위해 먼저 저자식별자가 잘못 부여되었을 가능성이 있는 경우를 다음 두 가지로 구분하고 해당 레코드를 검색하였다.

① 같은 email 주소이나 저자식별자가 2개 이상 부여된 경우
② 같은 email 주소-저자명-기관명이나 저자식별자가 2개 이상 부여된 경우
③ ①의 검색 결과에서 나타난 email 주소를 제외한 ②의 경우 재검색

실제로 다른 저자라면 같은 email 주소를 가질 수 없으며, 설령 드물게 email 주소가 재활용된 경우가 있다 하더라도 같은 시기에 같은 email 주소를 사용할 가능성은 거의 없기 때문에 위와 같이 email 주소를 기준으로 레코드를 검색하였다. 그 결과 ①의 경우는 107건이며 레코드 수는 221개이고, ②는 62건이며 레코드 수는 128개, ③의 경우는 45건이고 레코드 수는 45개가 검색되었다. ②와 ③의 레코드 수를 합하면 ①의 레코드 수 221개보다 3개가 적은데, 그 이유는 같은 email 주소에 저자식별자가 3개 이상 부여되어 있어서, ②의 경우와 ③의 경우 모두 해당하는 레코드가 ③에서 ②에서 검색된 email 주소를 삭제하는 과정에서 해당 email 주소가 삭제되었기 때문이다. 검색되지 못한 3개의 레코드에 대해서는 따로 검증 과정을 진행하였다.

검색된 레코드가 실제 오류 값인지 확인하기 위해 다음과 같은 과정으로 식별 작업을 진행하였다. 동일한 email 주소-저자명-기관명의 레코드들에 저자식별자가 2개 이상 부여된 경우와 동일한 email 주소의 레코드들에 저자식별자가 2개 이상 부여된 경우 각 건에 대한 모든 레코드의 논문 원문을 검색하여 논문의 첫 번째 페이지에서 저자명, 기관명, email 주소를 확인하고 원문 추출 과정에서 오류가 발생한 경우에는 각 자질 값을 수정하고 같은 저자인지 식별한다.

원문 첫 페이지 정보로 식별되지 못한 레코드에 대해서는 한국연구자정보(KRI) 사이트에서 해당 레코드의 저자명과 기관명을 검색하여 '논문실적'과 '학술활동' 페이지를 비교하고 같은 저

2) 단, 실제 데이터 셋의 저자식별자 필드를 직접 수정하지 않고, 새로운 필드를 추가하여 저자식별자를 수정하였다.

자로 판단되면 저자식별자 수정하였다. 한국연구자정보에 정보가 없는 경우에는 구글에 저자명과 기관명을 검색하여 저자 홈페이지가 있는 경우 저자 홈페이지의 논문 리스트를 확인하고 기관 홈페이지의 인물 소개란을 참고하여 해당 저자의 정보를 확인하였고, 같은 저자로 식별되는 경우 저자식별자를 수정하였다. 저자식별자를 수정한 레코드는 155개이며 고유 저자식별자 수는 8,304개에서 8,230개로 감소하였다. 원문 첫 페이지를 확인할 수 없거나 한국연구자정보, 저자 홈페이지 등을 확인할 수 없어 저자식별자가 정확하게 부여된 것이 맞는지 확인 할 수 없는 데이터 15건에 대해서는 email 주소를 삭제하여 같은 email 주소에 2개 이상의 저자식별자가 부여된 경우를 모두 제거하였다.

그러나 이 방법을 통해 저자식별자의 오류를 100% 발견해서 수정하였다고 할 수 없으므로 원 데이터 셋의 저자 식별자를 연구의 기본 정답 데이터로 사용하였고, 클린 저자식별자는 원래의 저자식별자를 정답 데이터로 사용한 저자명 식별 실험 평가 자료와 비교하기 위한 데이터로만 사용하였다.

5. 저자명 식별 실험

본 연구에서는 'email 주소는 고유하므로, 같은 email 주소를 사용하는 저자는 동일인이다.'는 연구 가정을 설정하고, email 주소 자질을 1차 자질로 활용한 룰 베이스 기반 저자명 식별 방식이 저자명 식별에 효과적인지 알아보기 위해 룰 베이스 방식과 머신러닝을 결합한 저자명 식별 실험과 머신러닝만 활용한 저자명 식별 실험을 비교하였다. 룰 베이스 기반 저자명 식별 방식은 고유한 자질인 email 주소[3]가 같은 레코드를 같은 저자로 식별한 후 Single-Link 클러스터링을 실행하는 것을 말하는데 email로 식별된 레코드들의 pairwise 유사도를 1로 지정한 후 Single-Link 클러스터링을 실행함으로써 email의 고유식별성을 보존함과 더불어 군집단의 seed 데이터 역할을 하는 장점이 있다.

머신러닝만 활용한 방식은 모든 자질에 대해 자카드 계수 기법으로 유사도를 계산하고 자질 간 유사도의 임계치가 0.7 이상인 레코드를 Single-Link 클러스터링으로 군집화 하는 방식을 말한다.

룰 베이스 방식의 실험에서 사용된 email 주소 자질은 email 주소의 정확도가 저자명 식별 성능에 미치는 영향을 알아보기 위해, 원래의 email 주소와 클린 email 주소를 각각 따로 사용한 실험을 진행하였다. 원래의 email 주소 자질로 1차식별 된 레코드는 15,091개이고 클린 email 주소 자질로 1차식별 된 레코드는 14,894개이며, email 주소로 식별되지 못한 나머지 레코드는 Single-Link 클러스터링을 수행하였다. 룰 베이스 방식에서의 유사도 계산은 email 주소가 같으면 유사도를 1로 계산하여 같은 저자로 식별하고, 그렇지 않은 경우에는 학술지명, 기관명, 부서명, email 주소, 중복을 제외한 공저자 수, 중복을 제외한 단어 수마다 분모를 1씩 증가시키고, 같은 자질이 있는 경우 또는 공유 공저자 수와 공유 단어 수만큼 분자를 1씩 증가시키는 방식을 사용하여 유사도를 계산하였다.

3) 수정된 email 주소를 사용함

즉 학술지명, 기관명, 부서명, email 주소 자질은 자질 개수 당 분모를 1씩 증가시키고 같은 자질이 있는 경우 분자를 1씩 증가시킨다. 공저자명은 중복을 제외한 공저자 수를 분모로 공유 공저자 수를 분자로, 논문제목의 단어는 중복을 제외한 단어의 수를 분모로 공유 단어 수를 분자로 계산하였다. 자질 간 유사도의 임계치는 머신러닝과 마찬가지로 0.7이며, Single-Link 클러스터링으로 군집화 하였다.

룰 베이스 방식과 머신러닝 방식에서 진행한 클러스터의 개수는 단일 자질 클러스터부터 최대 6개의 자질을 결합한 클러스터까지 총 18개이다. 저자명 식별 실험에 사용한 자질은 '공저자명(a)', '학술지명(c)', '도메인(d)', 'email 주소(e)', '기관명 및 부서명(i)', '논문제목(t)'이며 클러스터의 결합은 'a', 'e', 'i', 't', 'ae', 'ai', 'at', 'ei', 'et', 'it', 'aei', 'aet', 'ait', 'eit', 'aeit', 'aceit', 'acdit', 'acdeit'이다. 클러스터에 사용한 자질 간의 결합은 선행연구를 참고하여 저자명 식별에 효과적이라고 생각되는 것을 선정하였다. 예를 들어, '학술지명과 발행연도' 자질의 경우에 한 명의 저자가 여러 학술지에 논문을 기고할 수 있고, 20,396개의 데이터를 23개의 학술지명과 발행연도 자질값으로 식별하는 것은 의미가 없다고 판단하여 '학술지명과 발행연도' 자질의 단일 클러스터는 실험 제외하였다.

저자명 식별 실험에서는 데이터 셋에서 하나의 저자명에 하나의 레코드만 있는, 즉 동명이인 문제가 발생하지 않는 레코드 379개를 제외한 20,017개의 레코드만 사용하였고, 그 결과 고유 저자 수는 4,785명이며 고유 저자식별자 수는 8,001개가 되었다. 이는 동명이인 문제가 발생하지 않는 데이터를 실험에 포함하면 평가 성능이 실제 성능보다 높게 나올 가능성이 있기 때문이다.

6. 실험 결과

실험 결과 분석에 사용한 평가 지표는 저자명 식별 연구에서 일반적으로 사용하는 정확률(precision), 재현율(recall), F1 지표와 전체 경우의 수에서 정확히 예측한 비율인 정확도(Accuracy), 매크로 평균(Macro-Averaging), 마이크로 평균(Micro-Averaging) 및 컨퓨전 매트릭스(Confusion Matrix)의 구성항목인 True Positive(TP), False Positive(FP), False Negative(FN), True Negative(TN)이다.

정확도와 컨퓨전 매트릭스 방식은 일반적으로 머신러닝의 분류 방식에 대한 식별 성능을 평가하기에 더 적합한 방식이다. 본 연구에서는 클러스터링 결과를 평가하기 위해 정확률과 재현율, F1 지표 및 정확도에 pair-wise 방식을 적용하여 모든 레코드를 두 개의 쌍으로 연결한 뒤 각각의 쌍이 맞게 연결되었는지 여부를 정답 데이터와 비교하고 그 결과를 계산하였다. 매크로 평균과 마이크로 평균은 cluster-wise accuracy 방식을 적용하여 클러스터의 쌍을 비교하였다.

본 연구에서 시행한 저자명 식별 실험 6가지에 대한 F1 지표와 정확도의 최고 성능을 비교하면 〈표 6〉과 같다.

F1 지표를 기준으로 가장 좋은 성능을 보인 자질의 조합은 공저자-기관명 및 부서명-논문제목을 결합한 ait이고, 클린 데이터 셋과 클린 email 주소를 사용한 룰 베이스 방식의 실험의

〈표 6〉 실험별 최고 성능 비교

실험 방법	F1 지표	정확도
원 데이터 셋 - 룰: 원래의 email 주소	0.8769(ait)	0.8444(ait)
클린 데이터 셋 - 룰: 원래의 email 주소	0.8785(ait)	0.8449(ait)
원 데이터 셋 - 룰: 클린 email 주소	0.8775(ait)	0.8443(ait)
클린 데이터 셋 - 룰: 클린 email 주소	0.8792(ait)	0.8449(ait)
원 데이터 셋 - 머신러닝	0.7704(ait)	0.8324(a)
클린 데이터 셋 - 머신러닝	0.7823(ait)	0.8324(a)

식별 성능이 87.92%로 가장 높은 성능을 보였다. 이는 데이터 셋의 오류를 수정하고 정규화 작업을 한 데이터 셋과 email 주소를 사용하는 것이 데이터 클리닝 작업을 거치지 않은 원 데이터를 바로 사용하는 것보다 효과적임을 보여준다.

또한 머신러닝 방식에서 최고 성능을 보인 클린 데이터 셋에 머신러닝 방식의 사용한 실험 결과가 룰 베이스 방식의 최고 성능보다 약 10% 가량 낮게 나온 것은 email 주소가 같은 저자명을 먼저 식별하고 나머지 데이터를 클러스터링 하는 룰 베이스 방식이 모든 데이터를 대상으로 클러스터링 한 일반적인 머신러닝 방식보다 효과적임을 나타낸다.

한편, 데이터 클리닝 작업을 마친 클린 데이터 셋에 원래의 email 주소를 사용한 룰 베이스 방식의 최고 성능이 87.85%로 원 데이터 셋에 클린 email 주소를 사용한 최고 성능이 높게 나타난 것은 email 주소에 대한 데이터 클리닝 작업을 시행할 때 설령 정확한 email 주소였더라도 분명하지 않은 97개의 레코드와 저자식별자 수정 과정에서 확인되지 않은 저자식별자의 email 주소 15개를 삭제함으로써 양성 예측치인 True Positive와 False Positive가 줄어들었기 때문으로 보인다. 그러나 이 과정을 통해 음성 예측치인 False Negative와 True Negative는 향상되어 정확하지 않은 클린 email 주소가 서로 다른 저자를 같은 저자로 판단하는 오류를 줄여줌을 알 수 있다.

7. 결 론

저자명 식별 문제는 한 명의 저자가 여러 가지 방식으로 표기되거나 필명 등의 이유로 여러 이름을 가지는 다명일인 문제와 서로 다른 저자가 같은 이름을 갖는 동명이인 문제로 구분되며 저자의 연구 업적이나 인용지수 등을 평가하고자 할 때, 특정 분야의 전문가를 찾거나 특정 저자의 연구 목록을 검색하고 싶을 때, 그리고 학술 정보 서비스의 원활한 운영을 위해 반드시 해결해야 할 문제이다. 그러나 이때까지 발행된 논문부터 지금 이 순간에도 새롭게 출판되는 엄청난 양의 논문 등 학술 논문의 폭발적인 증가와 학술지마다 제공하는 서지정보 요소와 표기 방법이 달라 저자명 식별 자질을 추출할 때 많은 비용이 발생하는 문제는 저자명 식별 문제 해결을 어렵게 하고 있다.

본 연구에서는 머신러닝 방법을 사용한 저자명 식별 문제에서 발생하는 한계점을 보완하기

위해 데이터의 오류 및 이형 문제를 처리하고 임의의 가정을 설정하는 인간의 중재가 도움이 되는지 알아보고자 하였다. 먼저 머신러닝은 데이터 자체에 오류가 있을 경우 스스로 식별하지 못하며 같은 데이터라도 표기법이 다른 경우 같은 데이터로 식별하지 못한다는 점에 주목하여 저자명 식별을 위한 데이터 셋의 데이터 클리닝 방법 제시하고 그 효용성을 평가하였다. 또한 저자마다 고유한 값을 가지는 email 주소 자질이 저자명 식별을 위한 자질 중 가장 우수한 식별 성능을 가질 것이라는 가정을 바탕으로 한 룰 베이스 방식의 저자명 식별 실험을 진행하고 그 실험 결과를 머신러닝만 사용한 실험 결과와 비교하여 email 주소 자질이 저자명 식별 문제 해결에 실제로 도움이 되는지 알아보았다. 마지막으로 룰 베이스 방식에서 email 주소의 정확도가 높을수록 저자명 식별 성능이 올라가는지를 평가하기 위해 원래의 email 주소 자질과 오류 수정 및 정규화 작업을 마친 클린 email 주소 자질을 원 데이터 셋과 클린 데이터 셋 각각 적용하여 실험하고 그 결과를 비교하였다.

저자명 식별 실험은 연구 목적 세 가지를 해결하기 위해 원 데이터 셋과 클린 데이터 셋 각각에 대하여 원래의 email 주소와 클린 email 주소를 각각 적용한 4가지의 룰 베이스 방식 실험과 Single-Link 클러스터링을 적용한 2가지의 머신러닝 방식 실험을 진행하고 총 6가지 방식의 식별 결과를 비교하였다. 룰 베이스 방식과 머신러닝 방식에 사용한 자질은 '공저자명(a)', '학술지명(c)', '도메인(d)', 'email 주소(e)', '기관명 및 부서명(i)', '논문제목(t)'이며, 6가지 자질들의 여러 조합 중에 저자명 식별에 가장 효과적이라고 판단한 18개의 클러스터를 선정하여 실험을 진행하였다.

머신러닝의 성능 평가에서 가장 많이 사용하는 F1 지표를 기준으로 6가지 실험을 비교한 결과 최고 성능을 발휘하는 자질은 공저자-기관명 및 부서명-논문제목의 결합한 클러스터로 모두 같았다. 최고 성능은 클린 데이터 셋-클린 email 주소-룰 베이스 방식을 사용한 실험의 결과인 87.92%이며, 클린 데이터 셋-원래의 email 주소-룰 베이스 방식이 87.85%, 원 데이터 셋-클린 email 주소-룰 베이스 방식이 87.75%, 원 데이터 셋-원래의 email 주소-룰 베이스 방식이 87.69%, 클린 데이터 셋-머신러닝 방식이 78.28%, 원 데이터 셋-머신러닝 방식이 77.04% 순으로 식별 성능이 높게 나타났다. 머신러닝 방식의 최고 성능은 78.28%로 룰 베이스 방식의 최고 성능인 87.92% 보다 약 9.6% 낮으며, 룰 베이스 방식의 최저 성능인 87.69%보다도 낮게 나타났다. 이는 email 주소 자질이 같은 레코드를 같은 저자로 판단하여 먼저 식별하는 것이 저자명 식별에 효과적임을 알 수 있는 지표이다

또한 클린 데이터 셋-클린 email 주소-룰 베이스 방식을 사용한 실험이 가장 우수한 성능을 보인 것은 데이터 셋과 email 주소에 대한 데이터 클리닝 작업이 저자명 식별 성능 향상에 도움이 된다는 것을 증명한다. 한편, 데이터 셋-클린 email 주소-룰 베이스 방식의 최고 성능이 87.75%인데 반해 클린 데이터 셋-원래의 email 주소-룰 베이스 방식이 87.85%로 더 높게 나타나 email 주소 자질 하나만 클리닝 작업을 하는 것보다 email 주소에 오류가 있더라도 데이터 셋 전체에 대한 데이터 클리닝 작업을 하는 것이 더 효과적인 것으로 나타났다. 이는 email 주소에 대한 데이터 클리닝 작업에서 정

확한 email 주소가 입력되었다 하더라도 정확한 email이라는 증거가 없는 데이터를 모두 삭제하여 레코드 개수가 부족해졌기 때문으로 보인다. 그러나 클린 email 주소 자질을 사용한 실험은 같은 저자를 다른 저자로 식별하는 False Negative와 다른 저자를 바르게 식별한 True Negative인 음성 예측치 성능이 클린 데이터셋에 원래의 email 주소를 사용한 경우보다 높게 나타났다. 이를 통해 정확한 email 주소는 서로 다른 저자를 같은 저자로 식별하는 오류를 줄여준다는 것을 알 수 있다.

정보의 양이 폭발적으로 증가하고 사람의 힘만으로는 모든 데이터를 처리할 수 없는 현실에서 저자명 식별 문제 또한 결국은 머신러닝을 활용한 방식으로 해결해야 한다. 그러나 머신러닝은 데이터 자체를 스스로 학습하여 문제를 해결해 나가는 방식으로 데이터 자체에 오류가 있거나 학습할 데이터의 양이 부족하면 결과가 좋을 수 없다. 그리고 데이터에는 항상 오류가 발생한다. 저자명 식별을 위한 데이터셋은 원문 텍스트에서 저자의 소속기관 정보 또는 email 주소 자질을 추출해야 하지만 학술지마다 서로 다른 양식을 사용하여 자질을 추출해내는데 많은 어려움이 따르고 이로 인해 오류가 발생한다. 추출 성능이 좋아졌다 하더라도 오타 등의 이유로 원문 자체에 오류가 있을 수 있기 때문이다.

본 연구에서는 이러한 머신러닝의 단점을 보완하고 성능을 향상시키기 위해 인간의 패턴인식과 지식기반의 추론능력을 활용하여 머신러닝이 스스로 할 수 없는 데이터의 오류수정과 정규화작업을 수행하는 휴리스틱과 전거통제 방법을 도출하였다. 다시 말해 이 연구의 목적은 데이터의 오류를 일일이 수작업으로 수정하는 것이 아니라 데이터의 오류와 변형들의 패턴을 분석하여 데이터 수정과 정규화를 최대한 자동화시킬 수 있는 방법을 조사하는 것이다. 또한 모든 레코드를 유사도 계산 기법으로 계산하고 임계치 이상의 레코드를 서로 군집화할 뿐 어떤 자질이 가장 효과적인지, 어떤 자질에 가중치를 두는 것이 좋은지는 스스로 판단할 수 없는 머신러닝 방식을 보완하기 위해, 본 연구에서는 정규화한 email 주소로 식별한 저자들의 데이터를 Single-Link 클러스터링에 도입하는 (유사도 = 1) 저자명 식별방법을 제안하였다.

실험 결과 데이터 클리닝 작업과 가정을 설정하고 문제를 해결하는 룰 베이스 기법을 활용한 결과가 머신러닝 결과 보다 좋게 나타났다. 그러나 본 연구에서 진행한 정규화 작업은 아주 적은 수의 레코드에만 적용되어 그 효과가 미미하였다. 또한 데이터 셋에 사용된 논문에 IT 분야로 한정되어 있어 기관명 및 부서명, 논문제목의 다양성이 부족한 점도 아쉬운 부분이다. 다양한 분야의 대용량 데이터 셋을 사용하여 보다 정교하고 많은 양의 정규화 작업을 진행한다면 그 효과가 커질 것으로 예상한다. 따라서 후속 연구를 통해 다양한 주제의 최근 논문을 사용한 대용량 데이터 셋을 구축하고 여러 분야의 논문에 대한 저자명 식별 실험에서 발생하는 패턴을 분석하는 연구가 필요해 보인다. 추가로 기관명 또는 부서명이나 도메인에 대한 정규화 파일을 누적하고 공유함으로써 정규화 리스트를 일종의 전거통제 파일처럼 구축한다면 저자명 식별의 결과를 더욱 좋게 할 수 있을 것으로 예상된다.

기계와 인간의 장점들을 결합하여 저자명 식별의 향상을 시도한 본 연구에서는 동명이인 식별을 위한 데이터 수정과 정규화에 초점을 두고 일인다명 문제의 해결을 기계학습에 의존했지만, 추후연구에서는 전거통제같은 방법으로 일인다명 저자 식별의 결과를 향상시킬 수 있는지를 알아보고자 한다.

참 고 문 헌

강인수. 2008a. 저자식별을 위한 전자메일의 추출 및 활용.『한국콘텐츠학회논문지』, 8(6): 261-268.

강인수. 2008b. 한글 저자명 중의성 해소를 위한 기계학습기법의 적용.『정보관리학회지』, 25(3): 27-39.

강인수. 2009. 한글 저자명 군집화를 위한 계층적 기법 비교.『정보관리연구』, 40(2): 95-115.

강인수. 2011a. 동시인용정보를 이용한 동명이인 저자의 중의성 해소.『정보관리연구』, 42(3): 167-186.

강인수. 2011b. 저자 식별에 기반한 저자 그래프 생성.『정보관리연구』, 42(1): 47-62.

강인수 외. 2006. 논문 원문을 이용한 동명 저자 자동 군집화.『한국콘텐츠학회 종합학술대회 논문집』, 4(2): 652-656.

강인수 외. 2009. 저자 식별을 위한 대용량 평가셋 구축.『한국콘텐츠학회논문지』, 9(11): 455-464.

김일환, 이도길. 2015. 저자 판별을 위한 전산 문체론 - 초기 현대소설을 대상으로.『국어국문학』, 170(0): 207-239.

김하진, 정효정, 송민. 2014. 토픽모델링을 통한 저자명 식별 성능 비교.『한국정보관리학회 학술대회 논문집』, 149-152.

신동욱. 2009.『사회망을 이용한 서지정보의 저자명 명확화 기법』. 석사학위논문. 한양대학교 컴퓨터공학과.

이미화. 2014. 전거제어를 위한 국제표준이름식별자(ISNI의 활용가능성에 관한 연구.『정보관리학회지』, 31(3): 133-151.

이승우 외. 2006. 서지정보의 동명이인 구별을 위한 공저자 관계의 효용성 연구.『한국정보과학회 학술발표 논문집』, 10-12.

정한민, 이승우, 강인수, 성원경. 2006. 과학기술 문헌으로부터의 URI 기반 인력정보 구축.『한국콘텐츠학회논문지』, 6(9): 152-163.

조재인. 2013. ORCID 기반의 학술 연구 결과물 저자명 식별 시스템 구축 방안에 관한 연구.『한국비블리아학회지』, 24(1): 45-62.

Smalheiser, Neil R. and Vetle I. Torvik. 2009. "Author Name Disambiguation." *Annual Review*

of Information Science and Technology, 43(1): 1-43.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Cho, Jane. 2013. "A Study on the Construction Methods for Author. Identification System of Research Outcome based on ORCID." *Journal of the Korean Biblia Society for Library and Information Science*, 24(1): 45-62.

Jung, Han-Min, Seung-Woo Lee, In-Su Kang, and Won-Kyung Sung. 2006. "The Construction of URI-Based Human Resource Information from Science and Technology Papers." *The Journal of the Korea Contents Association*, 6(9): 152-163.

Kang, In-Su. 2008a. "Email Extraction and Utilization for Author Disambiguation." *The Journal of the Korea Contents Association*, 8(6): 261-268.

Kang, In-Su. 2008b. "Application of Machine Learning Techniques for Resolving Korean Author Names." *Journal of the Korean Society for Information Management*, 25(3): 27-39.

Kang, In-Su. 2009. "Exploration of Hierarchical Techniques for Clustering Korean Author Names." *Journal of Information Science Theory and Practice*, 40(2): 95-115.

Kang, In-Su. 2011a. "Disambiguation of Author Names Using Co-citation." *Journal of Information Science Theory and Practice*, 42(3): 167-186.

Kang, In-Su. 2011b. "Author Graph Generation Based on Author Disambiguation." *Journal of Information Science Theory and Practice*, 42(1): 47-62.

Kang, In-Su et al. 2006. "Automatic Clustering of Same-Name Authors Using Full-text of Articles." *Proceeding of The Journal of the Korea Contents Association*, 4(2): 652-656.

Kang, In-Su et al. 2009. "A Largescale Test Set for Author Disambiguation." *The Journal of the Korea Contents Association*, 9(11): 455-464.

Kim, Ha-Jin, Hyo-jung Jung, and Min Song. 2014. "A Comparison of Author Name Disambiguation Performance through Topic Modeling." *Proceedings of the 21th Korean Society for Information Management 2014*, 149-152.

Kim, Il-hwan and Do-Gil Lee. 2015. "Computational Stylistics for Authorship Attribution - based on Early Modern Korean Novels." *The Korean Language and Literature*, 170(0): 207-239.

Lee, Mi-hwa. 2014. "A Study on the Applicability of ISNI for Authority Control." *Journal of*

the Korean Society for Information Management, 31(3): 133-151.

Lee, Seung-woo et al. 2006. "A Research on the Effectiveness of Co-authorship for Identity Resolution in Bibliography." *Korea Computer Congress 2006*, 10-12.

Shin, Dong-Wook. 2009. *Name Disambiguation Using Social Networks on Bibliographic Data*. M.A. Thesis. HanYang University.