한글 저자명 중의성 해소를 위한 기계학습기법의 적용

Application of Machine Learning Techniques for Resolving Korean Author Names

강인수(In-Su Kang)*

초 록

동일한 인명을 갖는 서로 다른 실세계 사람들이 존재하는 현실은 인터넷 세계에서 인명으로 표현된 개체의 신원을 식별해야 하는 문제를 발생시킨다. 상기의 문제가 학술정보 내의 저자명 개체로 제한된 경우를 저자식별이라 부른다. 저자식별은 식별 대상이 되는 저자명 개체 사이의 유사도 즉 저자유사도를 계산하는 단계와 이후 저자명 개체들을 군집화하는 단계로 이루어진다. 저자유사도는 공저자, 논문제목, 게재지정보 등의 저자식별자질들의 자질유사도로부터 계산되는데, 이를 위해 기존에 교사방법과 비교사방법들이 사용되었다. 저자식별된 학습샘플을 사용하는 교사방법은 비교사방법에 비해 다양한 저자식별자질들을 결합하는 최적의 저자유사도함수를 자동학습할 수 있다는 장점이 있다. 그러나, 기존 교사방법 연구에서는 SVM, MEM 등의 일부 기계학습기법만이 시도되었다. 이 논문은 다양한 기계학습기법들이 저자식별에 미치는 성능, 오류, 효율성을 비교하고, 공저자와 논문제목 자질에 대해 자질값 추출 및 자질 유사도 계산을 위한 여러 기법들의 비교분석을 제공한다.

ABSTRACT

In bibliographic data, the use of personal names to indicate authors makes it difficult to specify a particular author since there are numerous authors whose personal names are the same. Resolving same-name author instances into different individuals is called author resolution, which consists of two steps: calculating author similarities and then clustering same-name author instances into different person groups. Author similarities are computed from similarities of author-related bibliographic features such as coauthors, titles of papers, publication information, using supervised or unsupervised methods. Supervised approaches employ machine learning techniques to automatically learn the author similarity function from author-resolved training samples. So far, however, a few machine learning methods have been investigated for author resolution. This paper provides a comparative evaluation of a variety of recent high-performing machine learning techniques on author disambiguation, and compares several methods of processing author disambiguation features such as coauthors and titles of papers.

키워드: 저자 식별, 동명저자, 기계학습,
author disambiguation, same-name authors, machine learning techniques

---

* 경성대학교 컴퓨터정보학부(dbaisk@ks.ac.kr)

■ 논문접수일자 : 2008년 7월 2일 ■ 최초심사일자 : 2008년 7월 11일 ■ 게재확정일자 : 2008년 7월 23일
■ 情報管理學會誌, 25(3): 27-39, 2008. 〔DOI:10.3743/KOSIM.2008.25.3.027〕

1. 서 론

1.1 연구의 목적

학술 정보의 저자를 표현함에 있어, 이름의 사용은 학술 정보 검색의 불편을 야기시킨다. 예를 들어, '홍길동'이라는 이름의 저자가 저술한 논문을 찾기 위해, 검색어로 '홍길동'을 입력한 상황을 가정해 보자. 그 검색 결과에는, '홍길동'이라는 이름을 갖는 서로 다른 저자들(예: '서울대학교'의 '홍길동', '부산대학교'의 '홍길동', '한국통신'의 '홍길동' 등)의 논문들이 모두 출력될 것이고, 검색 사용자는 그 논문들 중에서 자신이 의도한 '홍길동'의 논문을 찾기 위해 많은 시간을 투자해야 할 것이다. 이름 검색이 가져오는 이러한 검색 정확률의 저하는 실세계에서 동일한 인명을 갖는 서로 다른 다수의 사람들이 존재할 수 있음에서 비롯된다.

상기의 동명 저자 중의성의 문제를 해결하기 위해서는 학술 정보에 출현하는 동일 이름들을 실세계에서의 서로 다른 개체들로 구분하여야 한다. 이를 동명 저자 중의성 해소(author disambiguation) 혹은 저자 식별(author resolution)이라고 부른다. 저자 식별은 보다 큰 문제인 개체 식별(entity resolution)을 저자 개체로 한정시킨 것으로, 대용량 문헌 정보의 시맨틱 지식화를 위한 핵심 기능으로 인식되면서 최근 들어 활발히 연구되고 있다(Alani et al. 2003; Aswani et al. 2006).

저자 식별은 동일 저자명이 출현한 논문들을, 동일인이 작성한 논문들을 같은 군집으로 묶는 방식으로, 군집화(clustering)하는 문제이다. 저자식별문제에 주로 사용된 응집형 군집법(agglomerative clustering)의 경우 군집화를 위해 먼저 군집 대상이 되는 개체 집합에 대해 임의의 두 개체쌍의 개체유사도가 계산되어야한다. 개체유사도는 저자 식별의 경우 두 저자 개체 사이의 유사도(이후 저자유사도)에 해당된다. 저자유사도 계산을 위한 기존 연구들은 저자 식별된 학습데이터의 사용 유무에 따라 교사 방법(supervised method)과 비교사 방법(unsupervised method)으로 나뉜다. 교사 방법은 저자 식별된 학습데이터로부터 기계학습기법을 통해 두 저자 개체가 동일인인지 아닌지를 판별하는 분류기(classifier)를 학습하고, 분류기의 출력을 저자유사도로 활용한다. 저자유사도 함수가 분류기의 형태로 학습되는 교사방법과 달리, 비교사 방법은 저자 식별된 학습데이터 없이 저자유사도 함수를 사람이 직접 정의하여 사용한다. 비교사 방법의 경우 일차결합 함수(Aswani et al. 2006)와 유클리디안 거리 함수(Song et al. 2007) 등이 저자유사도 함수로 시도되었다.

기존 교사 기반 방법들은 SVM(Huang et al. 2006; Yang et al. 2006), 최대엔트로피모델(Kanani & McCallum. 2007) 등의 기계학습기법에 한정되어 있으며 그 외 다양한 기계학습기법의 적용을 통한 평가는 미흡했다.

이 연구에서는 여섯 가지 최신 기계학습기법들에 기반한 저자 식별 성능을 비교하여 한글 저자식별문제에 적합한 기계학습기법 선택을 위한 평가 자료를 제시한다. 이를 위해 대용량 한글저자식별테스트셋을 사용하여 공저자, 논문제목 자질 등에 대해 다양한 자질값 추출기법과 자질유사도 계산 기법들에 대한 비교분석을 병행한다.

1.2 연구의 방법

한글 저자명 식별 문제에 기계학습 기법을 적용하고 평가하기 위해 먼저 저자 식별 문제를 저자유사도 계산과 저자 군집화의 두 단계로 구분하고, 저자유사도 계산을 위해 기계학습 기법을 적용한다. 기계학습 기법으로는 기법의 최신성과 성능 우수성 및 기계학습 기법의 다양성 등을 고려하여 SVM, 메모리 기반 학습, 신경망, 결정트리, 최대 엔트로피 모델, CRF 모델의 여섯 가지 방법을 선택하였다. 기계학습의 학습데이터로는 한국과학기술정보연구원에서 국내 학술대회발표논문들을 대상으로 개발한 한글저자식별 평가셋을 사용하였다. 여섯 가지 기계학습 기법의 비교를 위한 척도로는 저자식별 정확률과 재현율, 오류율, 기계학습모델의 학습시간 및 분류시간 등을 사용하였다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개한다. 3장에서는 저자유사도 계산을 위해 기계학습기법을 적용하는 방식을 기술한다. 4장과 5장에서는 다양한 기계학습기법을 적용한 저자 식별의 실험 방법과 그 결과를 기술하고, 6장에서 결론을 맺는다.

2. 관련연구

저자 식별은, 같은 이름에 해당하는 저자명 표현들을 실세계 사람 개체에 해당하는 군집들로 그룹화하는 것이다. 저자명을 실세계 개체로 구분하기 위한 자질로, 기존 연구에서는 공동저자명, 논문제목, 게재지명, 게재년도, 저자의 전자메일주소나 소속(Huang et al. 2006; Kanani & McCallum. 2007), 저자의 홈페이지 정보(Aswani et al. 2006; Yang et al. 2006) 등이 혼용되어 사용되었다. 이 중 전자메일주소, 소속, 그리고 홈페이지 정보의 사용을 위해서는 논문 원문파일이나 웹이라는 외부자원 및 그들로부터의 고성능 추출기의 존재가 전제되어야 하는 부담이 있다. 결국, 저자 식별 자질 중 항상 이용 가능한 보편적 자질은 기본서지 항목들인 공동저자명, 논문제목, 그리고 게재지정보이다. 이 연구에서는 논문의 기본서지 레코드 내의 항목들을 자질로 사용하여 한글 저자명의 식별 문제를 다룬다.

기존 한글 저자명 식별 연구들(이승우 외 2006; 강인수 2008; 강인수 외 2008)은 저자유사도 계산을 위해 일차결합 함수 형태의 비교사 기법을 시도하였으며 이 논문이 다루는 기계학습 기반의 교사 기법은 시도된 적이 없다.

한글의 경우 이름을 적는 단일의 방식이 사용되지만, 영어의 이름 표기는 다양하다. 예를 들어 Andrew S. Tannenbaum이라는 영어 이름은 A. S. Tannenbaum, Andrew Tannenbaum 외에도 Tannenbaum, A. S.로 쓰이기도 한다. 따라서 영어 저자명 식별의 경우는 같은 이름을 지칭하는 서로 다른 저자명 표현을 정규화시키는 저자명 매칭 혹은 이름 매칭(name matching) 단계가 추가로 요구된다. 이름 매칭은, 두 개의 이름 문자열을 입력으로 받아 edit-distance, Jaccard, soft TF/IDF 등의 유사 문자열 매칭 기법을 통해 이름 유사도를 계산하는 방식으로 동작한다(Bilenko et al. 2003). 이름 매칭은 전통적으로 레코드 링키지 분야에서 다루어져 왔으며 이의 역사 및 연구 동향에 대

해서는 Winkler의 최근 연구(Winkler. 2006)를 참조하라.

저자 식별은, 식별 대상을 저자 개체로 국한시킨 인명 식별의 특별한 경우에 해당한다. 인명 식별 연구는 웹 페이지에 출현하는 인명을 실세계 개체로 구분하는 문제를 주로 다루었다(Guha & Garg. 2004; Wan et al. 2005). 웹 인명 식별로 불리는 이 문제에서는 웹 페이지 내용으로부터 한 개인의 출생일, 출생지, 전자메일주소, 우편주소, 소속, 직업, 전화번호 등의 개인 신상 정보들을 주요 자질로 추출하여 사용하였다.

인명 식별은 식별 대상을 사람으로 국한시킨 일반적인 개체 식별의 특수한 경우이다. 인명 외의 개체 식별 대상으로는 논문의 참고문헌에 기술되는 인용 레코드가 주로 다루어졌다. 이는 인용 매칭(citation matching)(McCallum et al. 2000)으로 불리며 같은 논문을 지칭하는 서로 다른 인용(레코드)들을 같은 그룹으로 묶는 것을 의미한다. 인용 매칭을 통해 연구자나 저널의 피인용지수 계산이 자동으로 이루어질 수 있으며 인용 링크를 기반으로 하는 사회망(social network) 표현의 질적 향상에도 도움이 된다.

3. 기계학습 기반 저자 유사도 계산

저자유사도 함수를 기계학습기법으로 자동 학습하기 위해, 먼저 저자 식별된 동명 저자 개체 레코드 집합으로부터 학습 샘플들을 생성한다. 〈표 1〉은 동명저자 '홍길동'에 대한 저자 식별된 개체 레코드의 예이다. 표에서 논문 C1, C2의 두 홍길동은 동일인으로 식별ID p1이 부여되어 있고, 논문 C3, C4의 두 홍길동도 동일인으로 식별ID p2가 부여되어 있다.

〈표 1〉과 같은 저자 식별된 동명 저자 개체 레코드 집합으로부터 〈표 2〉와 같이 임의의 두 개체쌍에 대한 학습샘플들을 생성한다. 표의

〈표 1〉 저자 식별된 서지 레코드 예

C1	**홍길동(p1)**, 김철수, 이영희. (2008). "고속 추론을 위한 시맨틱 웹 ...", KISS, ...
C2	이영희, **홍길동(p1)**, 김철수. (2007). "시맨틱 웹 추론의 soundess ...", KISS, ...
C3	박만수, 정선희, **홍길동(p2)**. (2007). "DBMS 저장 구조 개선 ...", KIPS, ...
C4	**홍길동(p2)**, 박만수. (2008). "고속 벌크 저장이 가능한 DBMS ...", KIPS, ...

〈표 2〉 〈표 1〉로부터 얻어지는 학습샘플들

학습샘플	학습샘플이 추출된 개체쌍
2, 3, 1, 1, 1	〈C1, C2〉
0, 0, 0, 1, 0	〈C1, C3〉
0, 1, 0, 0, 0	〈C1, C4〉
0, 0, 0, 0, 0	〈C2, C3〉
0, 0, 0, 1, 0	〈C2, C4〉
1, 2, 1, 1, 1	〈C3, C4〉

각 학습샘플을 구성하는 5개 숫자 필드들은 차례로 “공유 공저자수”, “공유 제목단어수”, “게재지 일치 여부 (일치: 1, 불일치: 0)”, “두 출판년도 차의 절대값”, “개체식별자 일치 여부 (일치: 1, 불일치: 0)”를 의미한다. 예를 들어 개체쌍 〈C1, C2〉로부터 얻어진 첫 번째 학습샘플 〈2, 3, 1, 1, 1〉에서 첫 번째 필드값 2는 C1과 C2 사이에 공유되는 공저자가 두 명(‘김철수’, ‘이영희’)이라는 의미이다. 두 번째 필드값 3은 C1과 C2의 논문제목 단어들 중 겹치는 단어가 세 개(‘시맨틱’, ‘웹’, ‘추론’)임을 의미한다. 세 번째 필드값 1은 C1과 C2의 게재지명이 KISS로 일치함을 의미하며, 네 번째 필드값 1은 C1과 C2의 출판년도 차의 절대값이 1(=|2008-2007|)이라는 것을 의미한다. 마지막 필드값 1은 두 개체 레코드 C1과 C2에 출현한 ‘홍길동’들이 식별 ID가 p1으로 같은 동일인이라는 의미이다.

학습샘플의 다섯 개 필드 중 처음 네 개의 필드는 학습자질(feature)에 해당하며 논문의 기본 서지 항목들인 공저자, 논문제목, 게재지명, 출판년도 각각이 두 저자 개체 레코드 사이에 일치되는 정도를 수치화한 것이다. 마지막 다섯 번째 필드는 특정 학습샘플이 속하는 클래스(class)에 해당된다. 즉, 하나의 학습샘플은 두 동명 저자 개체 레코드가 동일인일 경우(클래스 레이블=1)와 그렇지 않을 경우(클래스 레이블=0)의 조건들은 네 개의 학습자질값으로 표현하고 있다. 학습샘플의 직관적 의미는 두 동명 저자 개체 사이에 (1) 공유되는 공동저자가 많을수록, (2) 공유되는 연구분야(논문제목 단어)가 많을수록, (3) 논문의 게재지가 같을수록, 그리고 (4) 논문 게재년도의 차가 크지 않을수록 그 두 저자 개체가 동일인을 지칭할 확률이 높다는 가정을 반영한 것이다.

기계학습 기법들은 표 2와 같은 학습샘플들을 입력으로 받아 두 저자 개체가 동일인인지 아닌지의 여부를 판별하는 개체식별 분류기(classifier)를 학습하게 된다. 이후 새로운 동명 저자 개체쌍에 대해 개체식별 분류기는 임의의 입력 저자 개체쌍 사이의 저자유사도를 1과 0의 값으로 출력한다. 따라서, 개체식별 분류기로부터 군집 대상이 되는 개체집합 내의 임의의 두 개체쌍에 대한 개체 유사도를 얻을 수 있고, 이후 군집화 단계에서 미리 계산된 개체 유사도를 사용하여 군집화를 수행하게 된다.

4. 실 험

저자 식별 성능 평가를 위해 한국과학기술정보연구원에서 저자 식별을 위해 구축한 평가셋[1]을 사용하였다. 이 평가셋은 1999년부터 2006년까지 개최된 국내 IT 관련 주요 9개 학회 29개 학술대회에 발표된 논문들을 대상으로 한 것으로, 논문에 출현한 각 저자 개체에 대해 논문 서지 항목들이 구축되어 있다. 이 연구의 실험을 위해 평가셋에서 추출한 저자 개체 레코드는 다음과 같은 필드로 구성된다.

- 저자명, 저자식별ID, 논문제목, 공동저자명(들), 게재지명,[2] 출판년도

1) 이 평가셋은 한국과학기술정보연구원을 통해 배포됩니다. (연락처: wksung@kisti.re.kr)
2) 게지지명은 전거통제를 하였다.

앞의 저자 개체 레코드에서 알 수 있듯이 개별 저자명에 대해 실세계의 서로 다른 저자에 대해 서로 다른 저자식별ID가 부여되어 있다. 이 평가셋은 실세계 저자 식별을 위해 먼저 동명 저자들에게 모두 다른 식별자를 부여한 다음 국가과학기술인력종합정보시스템,[3] 홈페이지 검색, 서지 메타데이터 등을 참조하여 동일 인임이 판명된 저자들을 하나의 식별자 아래로 병합하는 과정을 수작업으로 거친 것이다. 표 3은 실험 대상 논문 7,677편에 대한 통계치를 보여준다. 7,677편의 논문에 대해, 8,307명의 실세계 저자들이, 20,614개의 저자 개체와 5,164개의 동명 저자 그룹으로 출현하였다.

〈표 3〉의 테스트셋을 임의 삼등분하여 2/3를 학습집합으로 사용하여 개체식별 분류기를 학습하였고 나머지 1/3을 평가집합으로 사용하였으며, 평가 성능은 서로 다른 세 개의 평가집합에 대한 평균을 계산한 것이다.

하나의 학습샘플은 3장에서 기술한 것과 유사하게 다음 네 가지 자질을 갖도록 정의하였고, Tsim의 경우는 3장의 예제에서 단순화시켜 기술한 제목 유사도 계산 방식과 다름을 주의하라.

- Csim(Coauthor Similarity): 공저자 유사도
- Tsim(Title Similarity): 제목 유사도
- Psim(Publication Similarity): 게재지 유사도
- Ysim(Year Similarity): 연도 유사도

위에서 Csim은 두 동명 저자 개체와 관련된 공저자 집합들의 교집합의 크기 즉 공유 공저자수로 계산된다.

Tsim은 두 저자 개체의 논문제목들 사이의 유사도를 의미하는데 논문제목을 표현하는 방법과 유사도를 계산하는 기법의 차이에 따라 몇 가지 변형들이 가능하다. 논문제목의 표현 방법으로 논문제목으로부터 추출된 명사류 키워드들의 벡터로 표현하는 방법과, 논문제목에서 내재된 토픽을 추출하여 벡터로 표현하는 방법을 사용하였다. 명사류 키워드의 추출을 위해 논문제목 텍스트를 형태소 분석하고 품사 태깅을 거쳐[4] 보통명사, 고유명사, 미등록어 명사로 태깅된 용어들을 키워드로 추출하였다. 토픽 추출을 위해서는 최근 제안된 토픽 추출 기법인 LDA(Latent Dirichlet Allocation)(Blei et al., 2003)를 사용하여 고정된 토픽수 3, 5, 10, 15, 20, 30 각각에 대해 토픽을 추출하였다. LDA 학습을 위해 논문제목의 키워드 표현을 입력으로 사용하였고, LDA 학습 결과로부터 각 키워드가 고정된 수(예, 3개)의 토픽들 각각으로부터 생성되었을 확률(키워드-토픽확률)을 얻었다. 하나의 논문제목에서 추출된 키워드

〈표 3〉 테스트셋 통계

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

3) 〈http://www.hrst.or.kr/〉
4) 한글 형태소분석과 품사태깅을 위해 포항공과대학교 KLE연구실의 언어분석기를 사용하였다.

들에 대해 상기의 키워드-토픽확률들을 벡터로 표현한 것이 논문제목의 토픽 벡터 표현이다.

논문제목 벡터 사이의 유사도 계산을 위해 키워드 벡터 표현의 경우 Dice 계수와 tf-idf 기반 Cosine 유사도 함수를 각각 사용하였고, 토픽 벡터 표현의 경우 유클리디언 거리(Euclidean distance) 함수를 사용하였다.

Psim은 두 저자 개체 사이의 게재지명 유사도를 의미하는데 실험에 사용한 테스트셋 내의 게재지명이 이미 전거 표현되어 있어 게재지명의 일치 여부를 0과 1의 값으로 표현하였다. Ysim은 두 저자 개체 사이의 게재년도 인접도를 의미하며 이를 위해 두 게재년도의 차의 절대값을 사용하였다.

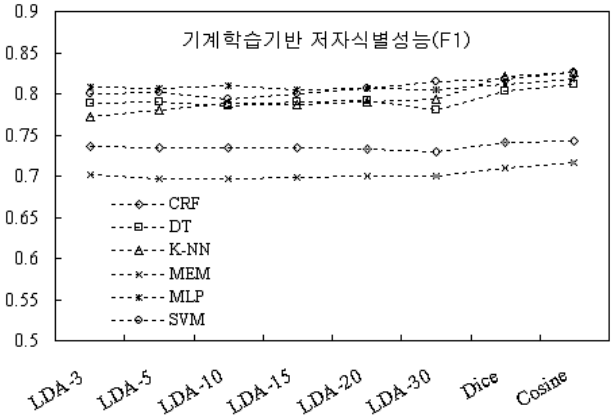
분류기 학습을 위한 기계학습 기법으로는 기법의 최신성과 여러 응용 분야에서의 성능 우수성을 고려하여 CRF(Conditional Random Field),[5] 결정트리(Decision Tree, DT), K-NN(K Nearest Neighbor),[6] MEM(Maximum Entropy Model),[7] 신경망(Neural Network, MLP: Mult-Layer Perceptron), SVM(Support Vector Machine)[8]의 여섯 가지 기계학습기법을 선택하였다. DT와 MLP는 Weka[9] 툴을 사용하였다. 기계학습기법을 통해 얻어진 개체식별 분류기의 출력은 군집화 알고리즘에서 요구하는 개체쌍 유사도(저자유사도)를 표현하기 위해 사용된다. 군집화 방법으로는 단일 링크 응집형 군집법(single-link agglomerative clustering)[10]을 사용하였다.

저자 식별 성능 평가를 위해 사용한 pairwise-F1 지표는 동일 정답 군집 내의 임의의 저자 개체쌍이 동일 시스템 (출력) 군집 내에서 발견되는 비율인 재현율(recall)과 동일 시스템 (출력) 군집 내의 임의의 저자 개체쌍이 동일 정답 군집 내에서 발견되는 비율인 정확률(precision)의 조화평균으로 계산된다. 군집 오류 평가를 위한 지표로 과다군집오류(overclustering error: OE)와 과소군집오류(underclustering error: UE) 지표를 사용하는데, OE는 전체 저자 개체쌍 중 정답에서는 서로 다른 군집에 속하나 시스템은 같은 군집으로 묶은 개체쌍의 비율이며, UE는 그 반대로 정답에서는 같은 군집이나 시스템은 같은 군집으로 묶지 않은 개체쌍의 비율이다(강인수 외 2008).

5. 실험 결과

〈그림 1〉은 서로 다른 기계학습기법들을 사용한 개체식별분류기의 분류결과에 기반한 저자식별의 성능을 비교한 것이다. 그림의 x축은 Tsim의 계산을 위한 여러 기법들을 나열한 것이다. 여섯 개 기계학습기법 중 저자식별 능력

5) 〈http://mallet.cs.umass.edu/〉
6) 〈http://ilk.uvt.nl/timbl/〉
7) 〈http://www.cs.ualberta.ca/~lindek/maxent.tgz〉
8) 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉
9) 〈http://www.cs.waikato.ac.nz/ml/weka/〉
10) single-linkage 군집법은 기존 저자식별 연구(Tan et al. 2006; Song et al. 2007)에서 complete-linkage나 group-linkage에 비해 대등하거나 우수한 것으로 보고되었다.



〈그림 1〉 기계학습기법들의 저자식별 성능 비교

은 SVM이 가장 좋았고 다음으로 K-NN, MLP, DT, CRF, MEM 순이었다. SVM은 RBF 커널함수를 사용했고, MLP는 2개의 히든 레이어 노드로 학습한 것이며, K- NN에서 K는 4[11]를 사용하였고, DT는 C4.5 알고리즘을 적용한 것이다.

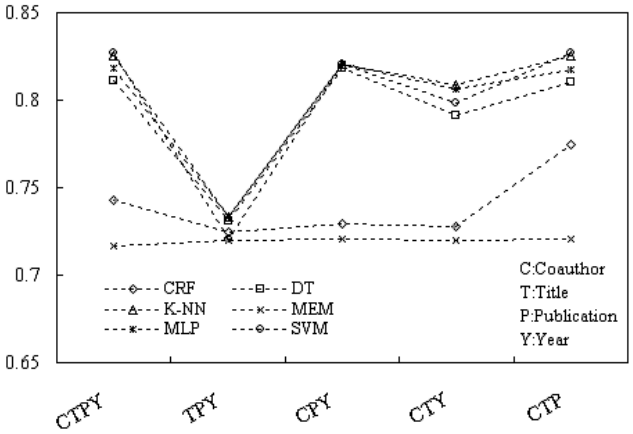
Tsim 계산을 위한 기법들 중에는, 키워드 벡터 표현을 사용한 Dice 계수와 Cosine 유사도가 LDA 토픽 기반의 유클리디안 거리 함수보다 저자식별에서 근소하게 좋은 성능을 보였으나 K-NN과 DT의 경우를 제외하고 큰 차이는 없었다. 이 결과는 텍스트 표현을 위한 차원 축소 기법으로 제안된 LDA의 효과를 보이는 것으로 해석될 수 있다. 즉, 테스트셋의 제목에 출현한 전체 표층 키워드 수 12,184개를 3~30개의 토픽으로 압축했음에도 불구하고 저자식별 성능의 큰 저하가 없었기 때문이다. 그러나 이는 현 단계에서 결론 짓기 힘든 면이 있으며 이후 단락에서 더 논의된다.

〈그림 2〉는 네가지 저자식별자질의 동명저자식별력을 비교하기 위해 네 가지 자질 중 하나의 자질을 사용하지 않은 각 경우의 성능을 기계학습기법 별로 비교한 것이다. 예를 들어, CTPY는 모든 자질을 사용한 경우이며, TPY는 공동저자자질(C)을 배제한 경우의 성능이다. 그림에서 자질배제로 인해 평균적 성능 저하가 심한 순으로 자질을 나열하면 공동저자(C), 게재지명(P), 게재년도(Y), 논문제목(T) 순이었다. 이는 저자식별력이 큰 순서로 볼 수 있다. 특히, 논문제목(T)과 게재년도(Y) 자질을 각각 배제한 CPY, CTP의 경우 CRF를 제외하고는 CTPY의 성능과 큰 차이가 없었다. 따라서 이 결과만으로는 논문제목자질이 저자식별에 도움이 되지 않는 것으로 판단될 수 있다

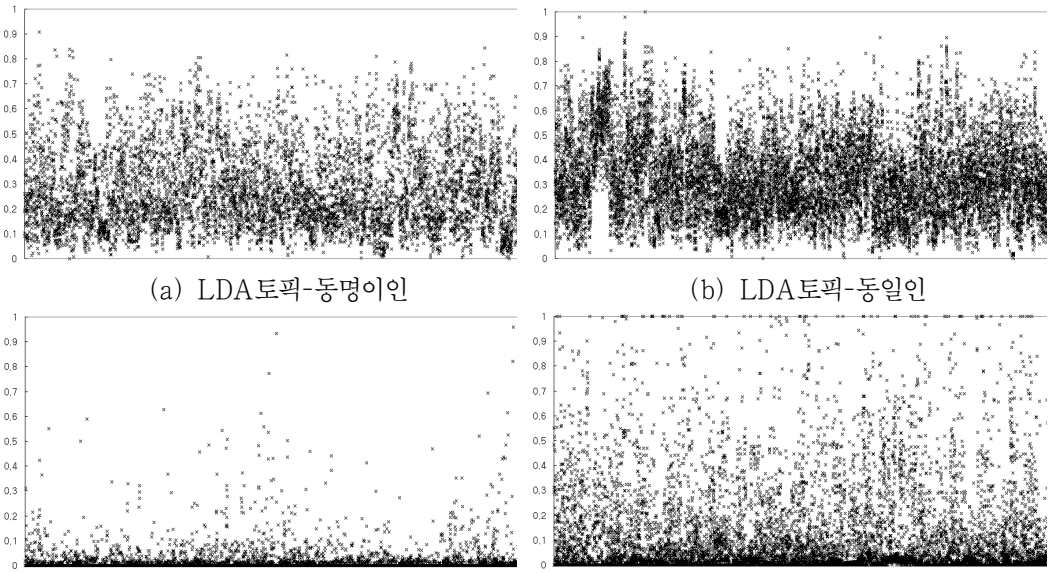
이의 분석을 위해 동명이인저자군집과 동일인저자군집 각각에 대해 저자개체쌍이 보이는 논문제목유사도 (Tsim) 분포를 표층키워드 표현과 LDA토픽 표현에 대해 각각 생성해 보았

11) TiMBL툴에서 옵션 "-a 0 -m 0 -w 1 -k 4"로 실험한 것이다. TiMBL은 tie 발생시 k를 1씩 증가시키는 방식 등의 tie-breaking 기법을 사용한다.

다(그림 3 참조). 그림에서 알 수 있듯이 표층 키워드에 기반한 Tsim의 분포((c)와 (d))는 동명이인군집과 동일인군집 사이에서 두드러진 차이를 보이는 반면, LDA토픽에 기반한 Tsim의 분포((a)와 (b))는 두 군집사이에 구별되는 특성을 찾기 힘들다. 이처럼 표층키워드 표현의 경우 LDA토픽 표현과 달리 동일인과 동명이인을 분별하는 특성을 보이고 있으므로, 〈그림 2〉의 결과만으로 논문제목이 저자식별에 긍정적 역할을 하지 못한다고 결론짓기는



〈그림 2〉 자질별 저자식별력 비교 (Cosine기반 Tsim사용, K-NN의 경우 CTPY는 k=4, 나머지는 k=2 사용)



(a) LDA토픽-동명이인 (b) LDA토픽-동일인

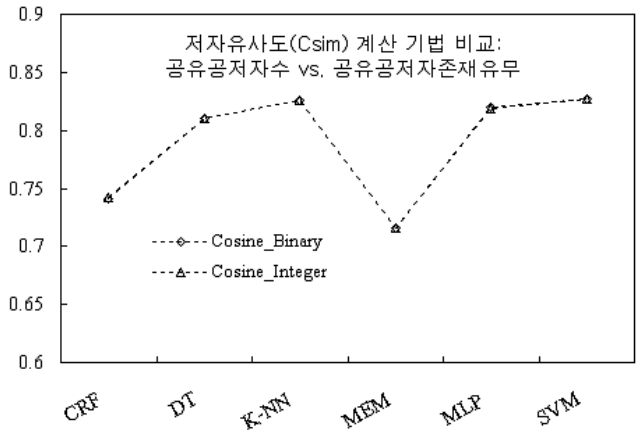
(c) 표층키워드-동명이인 (d) 표층키워드-동일인

〈그림 3〉 논문제목 유사도(Tsim) 분포(x축:저자개체쌍번호, y축:저자개체쌍의 Tsim)

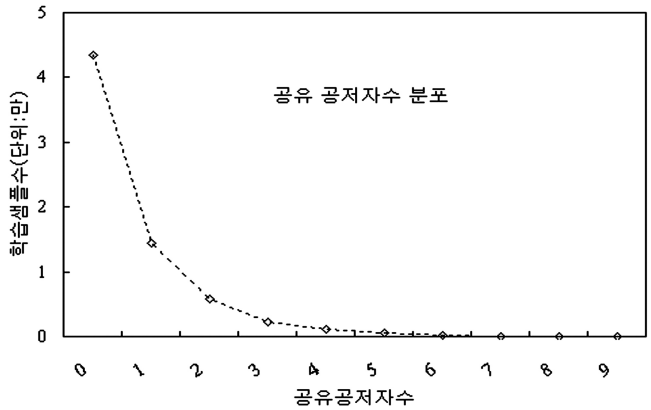
힘들며, 타자질의 식별력에 잉여적으로 기여한다고 보는 것이 적절할 것이다.

저자식별력이 가장 큰 공동저자자질(그림 2 참조)의 속성을 더 살펴보는 측면에서, 공유 공저자수의 많고 적음이 어떤 영향을 미치는지 여부를 살펴보기 위해, Csim을 공유 공저자수로 계산한 경우와 공유 공저자의 존재 유무로 계산한 경우를 비교해 보았다(그림 4 참조). 〈그림 4〉에서 전술한 Csim 계산의 두 방법들은 차례로 구분자 Integer와 Binary로 표시되었다. 그림에서 알 수 있듯이, 공저자수의 많고 적음은 저자 식별 측면에서 거의 차이가 없었는데 그 이유 중 하나로 실제 공유 공저자수가 두 명 이상인 경우가 극히 드물기 때문이라고 추정할 수 있다.

이의 분석을 위해 학습샘플에서 공유 공저자수의 분포를 살펴보았으며(그림 5 참조), 실제 한국의 학술대회논문 저자들의 대부분이 한 명의 공저자와만 논문을 작성하는 것은 아님을 알



〈그림 4〉 저자유사도 계산 기법 비교(공유공저자수 vs. 공유공저자존재유무)



〈그림 5〉 테스트셋 전체에서 공유 공저자수의 분포

수 있다. 따라서, 〈그림 4〉에서 Binary와 Integer 방식 Csim 계산이 거의 차이를 보이지 않는 것은, 한 동명 저자 개체 집단 내에서 동명 공저자를 갖는 저자 개체들이 실세계에서 서로 다른 사람들일 확률이 거의 희박하다는 것을 의미한다. 만약 그렇지 않다면 Integer 방식의 Csim을 사용한 저자 식별 방법이, 동명 공저자를 갖는지의 여부만을 계산하는 Binary 방식의 Csim을 사용한 저자 식별 방법이 해결하지 못하는 애매함을 해소하는데 기여했을 것이기 때문이다.

〈표 4〉는 여섯 가지 기계학습기법 각각에 대해 최고 성능(F1)을 보인 저자 식별 결과를 보여 준다. 실험에 사용된 기계학습기법 중에는 SVM과 K-NN이 차례로 최고 성능을 보였으며, 과소 및 과다 군집 오류 측면에서도 어느 한 쪽으로 치우치지 않는 균형 있는 결과를 보였다. 다른 기계학습 기법들은 과소군집오류에 비해 과다군집오류를 많이 발생시켰다.

〈표 4〉의 마지막 두 열은 각 기계학습기법에 대해 저자식별 분류기를 학습하고 분류하는 시간을 제시하고 있다. 표의 학습 및 분류 시간은 각각 학습샘플 전체와 평가샘플 전체에 대한 소요 시간이다. 오프라인으로 수행될 수 있는 분류기 학습에 소요되는 시간을 배제하고 온라인 저자식별을 위한 소요 시간인 분류시간 관점에서 기계학습 기법을 비교하면, DT와 MLP가 차례로 가장 빨랐고 이들은 저자식별성능에서도 SVM과 K-NN에 크게 뒤지지 않았다. K-NN은 분류해야할 테스트 샘플과 가장 가까운 K개 학습샘플이 속한 클래스 정보를 바탕으로 분류를 수행하므로, 테스트 샘플과 학습샘플 각각의 유사도를 비교하는 시간이 소요되어 일반적으로 가장 느린 분류 시간을 보이는 기법이다. 그러나, 이 실험에서는 SVM이 가장 느린 분류 시간을 보였다. SVM은 학습을 통해 얻어진 서포트벡터(support vector)들을 이용하여 테스트 샘플의 클래스를 결정하므로 서포트벡트의 개수가 분류 시간을 결정하는 요인이 된다. 이 실험 테스트셋의 경우 SVM을 통해 학습된 서포트벡터의 수는 전체 학습샘플수(약 45,386)의 50%정도를 차지할 만큼 많았으며 이로 인해 SVM의 분류 시간이 길어진 것으로 판단된다. 최근 이러한 SVM 분류시간 지연의 원인인 과다 서포트벡터의 수를 줄이는 기법이 연구되고 있다(Xia et al. 2005).

실험을 통해 한글 학술대회발표논문에 대한 기계학습기법 기반의 저자 식별 성능은 80%대 초반임을 확인하였다. 이 결과는 논문의 기본 서지 항목들인 저자, 논문제목, 게재지명, 게재년도만을 사용한 성능임을 감안할 때, 논문 원

〈표 4〉 기계학습기법 vs. 저자식별 최고 성능, 오류

학습기법	Tsim	F1	OE	UE	저자식별 분류기 (초)	
					학습시간	분류시간
CRF	Cosine	74.27%	24.16%	2.69%	94.79	1.49
DT	Cosine	81.08%	11.27%	6.82%	21.32	0.68
K-NN	Cosine	82.57%	7.71%	8.82%	8.02	4.70
MEM	Cosine	71.64%	27.99%	2.56%	76.22	4.91
MLP	Cosine	81.85%	10.09%	7.23%	740.26	1.82
SVM	Cosine	82.70%	8.02%	8.48%	142.82	39.20

문이나 웹 상의 개인홈페이지 정보의 존재를 배제한 상황에서 얻을 수 있는 저자식별의 보편적 성능임을 의미한다. 기계학습기법 중 저자식별력으로는 SVM과 K-NN이 최고 성능을 보였고 온라인 저자식별시간을 동시에 고려할 경우 DT와 MLP가 가장 좋았다.

6. 결 론

이 연구를 통해 다양한 최신/고성능 기계학습기법들을 저자식별문제에 적용하여 성능을 비교하였다. 실험을 통해, 논문의 원문 텍스트나 웹으로부터 획득되어야 하는 저자식별 자질들을 배제하고 논문의 기본 서지 항목들만을 사용하였을 경우, 한글 학술대회발표논문에 출현한 저자 식별의 실질적 최고 성능은 82.7%이었다. 기계학습기법들 중에는 SVM과 K-NN이 F1, 과소/과다군집오류의 세 가지 측면에서 실험에 사용한 테스트셋에 대해 가장 우수했고, 온라인 저자식별의 수행 속도면을 동시에 고려할 경우 결정트리와 신경망이 우수했다.

참 고 문 헌

강인수, 이승우, 정한민, 김평, 구희관, 이미경, 성원경, 박동인. 2008. 저자 식별을 위한 자질 비교. 『한국콘텐츠학회논문지』, 8(2): 41-47.

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

이승우, 정한민, 김평, 강인수, 성원경. 2006. 서지정보의 동명이인 구별을 위한 공저자 관계의 효용성 연구. 『한국컴퓨터종합학술대회 논문집』, pp.10-12.

Alani, H., Dasmahapatra, S., O'Hara, K., and Shadbolt, N. 2003. "Identifying communities of practice through ontology network analysis." *IEEE Intelligent Systems* 18(2): 18-25.

Aswani, N., Bontcheva, K., and Cunningham, H. 2006. "Mining information for instance unification." *Proceedings of ISWC-2006*, pp.329-342.

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P. and Fienberg, S. 2003. "Adaptive name matching in information integration." *IEEE Intelligent Systems*, 18(5): 16-23.

Blei, D., Ng, A., and Jordan, M. 2003. "Latent Dirichlet allocation." *Journal of Machine Learning Research*,3: 993-1022.

Guha, R., and Garg, A. 2004. "Disambiguating people in search." *Proceedings of WWW-2004*

Huang, J., Ertekin, S., and Giles, C.L. 2006. "Efficient name disambiguation for large scale databases." *Proceedings of*

PKDD-2006, pp.536-544.

Kanani, P., and McCallum, A. 2007. "Efficient strategies for improving partitioning-based author coreference by incorporating Web pages as graph nodes." *Proceedings of IIWeb-2007*.

McCallum, A., Nigam, K., Ungar, and L. H. 2000. "Efficient clustering of high-dimensional data sets with application to reference matching." *Proceedings of KDD-2007* pp.169-178.

Song, Y., Huang, J., Councill, I., Li, J., and Giles, C.L. 2007. "Efficient topic-based unsupervised name disambiguation." *Proceedings of JCDL-2007*.

Tan, Y.F., Kan, M.Y., and Lee, D.W. 2006. "Search engine driven author disambiguation." *Proceedings of JCDL-2006*, pp.314-315.

Yang, K.H., Jiang, J.Y., Lee, H.M., and 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.

Wan, X., Gao, J., Li, M., and Ding, B. 2005. "Person resolution in person search results: WebHawk." *Proceedings of CIKM-2005* pp.163-170.

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

Xia, X., Lyu, M., Lok, T., and Huang, G. 2005. "Methods of decreasing the number of support vectors via k-mean clustering." *LNCS*, 3644: 717-726.