저자 식별을 위한 대용량 평가셋 구축

A Large-scale Test Set for Author Disambiguation

강인수*, 김평**, 이승우**, 정한민**, 류범종**
경성대학교 컴퓨터정보학부*, 한국과학기술정보연구원 정보기술연구실**

In-Su Kang(dbaisk@ks.ac.kr)*, Pyung Kim(pyung@kisti.re.kr)**,
Seungwoo Lee(swlee@kisti.re.kr)**, Hanmin Jung(jhm@kisti.re.kr)**,
Beom-Jong You(ybj@kisti.re.kr)**

요약

현재의 논문 중심적 학술정보 탐색의 한계에서 벗어나 저자 중심적 검색을 제공하기 위해서는 저자명이 갖는 동명이인의 문제가 해결되어야 한다. 그 해법으로 제시된 저자식별은 논문에 출현한 저자명 개체에 실세계 연구자에 대응하는 식별자를 부여하는 작업이다. 최근의 선도적 저자식별 연구들은 90%를 상회하는 식별 성능을 보고하고 있으나 실질적인 학술정보서비스에서 저자식별 기능이 탑재된 예는 거의 없다. 본 논문에서는 학술정보서비스에 보다 직접적으로 기여할 수 있는 광범위한 저자식별 연구를 위해 한국과학기술정보연구원에서 새롭게 구축한 대용량 저자식별 평가셋에 대해 기술한다. 평가셋은 DBLP 데이터에 출현한 고빈도 저자명들에 대해 웹 검색을 통한 수작업 식별 과정을 거쳐 만들어졌다. 현재 881개 저자명에 대해 수집된 41,673개의 저자명개체레코드로 구성되어 있으며 총 6,921명의 실세계 저자 식별자가 존재한다.

■ 중심어 : | 저자식별 | 저자식별 평가셋 | 평가셋 구축 절차 |

Abstract

To overcome article-oriented search functions and provide author-oriented ones, a namesake problem for author names should be solved. Author disambiguation, proposed as its solution, assigns identifiers of real individuals to author name entities. Although recent state-of-the-art approaches to author disambiguation have reported above 90% performance, there are few academic information services which adopt author-resolving functions. This paper describes a large-scale test set for author disambiguation which was created by KISTI to foster author resolution researches. The result of these researches can be applied to academic information systems and make better service. The test set was constructed from DBLP data through web searches and manual inspection. Currently it consists of 881 author names, 41,673 author name entities, and 6,921 person identifiers.

■ keyword : | Author Disambiguation | Test Set for Author Disambiguation | Test Set Construction |

I. 서론

저자식별(Author disambiguation)은 학술정보에 출현한 저자명에 대응하는 실세계 저자를 결정하는 것이다. 이를 통해 동일 표기의 저자명이 출현한 논문 집합을 논문의 실제 저자 중심으로 분류할 수 있으므로 저

접수번호 : #090929-004
접수일자 : 2009년 09월 29일
심사완료일 : 2009년 10월 14일
교신저자 : 김평, e-mail : pyung@kisti.re.kr

자명을 통한 학술정보 탐색의 정확률을 향상시킬 수 있다. 현재 저자식별 기술의 최고 성능은 통제된 실험 환경에서 90%(F1 지표)대 초반으로 보고되고 있으나[1], 실제적인 대용량 학술정보서비스에서 저자식별을 시도한 사례는 Scopus[1]를 제외하고는 거의 찾기 힘들다. 저자식별 기능이 학술정보서비스에서 보편화되기 위해서는 저자명 출현 학술정보의 실제 상황을 고려한 평가셋을 사용하여 광범위한 성능 평가가 전제되어야 할 것이다. 이러한 관점에서 현재까지 구축되어 활용되어 온 저자식별 평가셋들[2-5]은 대용량성과 출현 저자명의 다양성 측면에서 부족한 부분이 있어 저자식별을 위한 새로운 평가셋의 구축이 요구된다. 이 연구는 학술정보 출판의 공용어라 할 수 있는 영어를 대상으로 대용량으로 새롭게 구축한 저자식별 평가셋에 대한 소개를 다룬다.

저자식별 평가셋은 기본적으로 식별 대상이 되는 저자명(예: Michael Lay)과 그에 대응하는 실세계 저자 식별자(예: MLay-001)를 항목으로 포함하는 저자명개체레코드들의 모음으로 구성된다. 이 외에도 저자명개체레코드는 저자명이 출현한 문헌의 기본적인 서지정보(예: 논문제목, 저자명(들), 게재연도, 개재지 등)를 항목으로 포함한다. 예를 들어 세 저자명 M. Lay, J. Smith, D. Ullman에 대해 각 저자명이 출현한 100편씩의 논문서지항목을 수집하여 평가셋을 구축한 경우 저자명개체수와 동명저자명개체그룹[2]의 수는 각각 300과 3이 되며, 각 저자명개체 100개마다 매핑된 실세계 저자수가 5라면 저자중의성은 5가 될 것이다. 기존 영어 평가셋들의 경우 저자명개체수가 최대 8,442개[6], 개체그룹의 수는 최대 24개[3]이므로 저자명과 저자중의성의 다양성과 관련하여 실제 저자명 출현 학술정보의 대표 샘플로서 부족한 측면이 없지 않다.

기존 평가셋과 달리 이 논문에서 소개할 새로운 평가셋은 저자명개체레코드 41,673개, 동명저자명개체그룹 881개로 구성된 대용량이며, 저자명에 대해 식별된 실세계 저자가 총 6,921명으로, 하나의 저자명을 평균 12.7명 중 한 사람으로 식별해야 하는 복잡도를 내포하고 있다. 평가셋 구축의 대략은 다음과 같다. 먼저 저자명 수집을 위한 원천 논문집합으로 정확성, 공용성, 대용량성 등을 고려하여 2007년 당시 87만여편의 논문정보를 보유한 DBLP 데이터를 선정하였다. 다음으로 DBLP 전체 논문집합에 출현한 상위 1000개의 고빈도 저자명 집합을 정의하고 이들 각 저자명이 출현한 DBLP 논문들로부터 저자명개체레코드들을 생성하였다. 마지막으로 저자명개체에 대응하는 실세계 저자를 확인하기 위해 웹에 존재하는 개인출판논문리스트 웹페이지를 활용하였다.

논문의 구성은 2장에서 관련 연구를 기술하고, 3장에서 새로운 평가셋 구축 과정을 상술하며, 4장, 5장에서 각각 평가셋의 특성과 저자식별실험 및 성능에 대해 기술하고, 6장에서 결론을 맺는다.

II. 관련 연구

이 장에서는 신규 저자식별 평가셋 구축의 관련 연구로 기존 저자식별 연구에서 사용된 평가셋들의 특성을 기술한다.

표 1. 기존 저자식별 평가셋

Test set	# of Records	# of Persons	Ambiguity	Performance
psu-citeseer-9	3,028	447	49.7	93.6%
psu-citeseer-10	3,355	490	49.0	90.6%
psu-citeseer-14	8,442	480	34.3	84%
psu-pike-24	724	49	2.0	83.6%
umass-dblp-17	841	97	5.7	92.2%
umass-rexa-8	1,302	219	27.4	90.6%
umass-penn-7	1,588	93	13.3	35.5%
southampton-8	4,799	n/a	n/a	89.9%

[표 1]은 기존 저자식별 실험에서 활용되었던 평가셋들을 평가셋명(Test set), 저자명개체레코드 수(# of Records), 실세계 저자수(# of Persons), 저자중의성

1) Scopus(http://www.scopus.com/)는 논문에 출현한 정규화된 동일 표기 저자명들을 소속기관, 주제 분야, 공저자 등으로 구별한 다음 실세계 사람으로 추측되는 서로 다른 그룹으로 묶어 제시한다.

2) 저자식별의 대상이 되는 같은 이름의 저자명 개체들의 모음이며, 같은 이름의 저자명 개체가 출현한 논문들의 집합으로도 볼 수 있다. 이 논문에서 동명저자명개체집합으로도 기술한다.

(Ambiguity), 최대성능(Performance)의 항목들로 정리한 것이다. 평가셋명에서 마지막 부분 숫자는 평가셋에 포함된 서로 다른 저자명의 개수이며 동명저자명개체집합의 개수와 같다. 예를 들어 평가셋 psu-citeseer-14는 서로 다른 14개의 저자명에 대해 8,442개의 저자명 개체레코드를 포함하고 있으며 수작업으로 각 저자명의 실세계 저자를 확인한 결과 총 480명의 실세계 저자가 존재함을 의미한다. 이 경우 저자중의성 34.3은 실세계 저자수 480을 서로 다른 저자명 수 14로 나눈 것으로, 정확하게는 14개 각 저자명에 대해 저자중의성을 계산하여 평균을 구해야 할 것이다.

표에서 psu-citeseer-9/10/14는 미국 펜실베니아주립대학교의 CiteSeer 연구그룹에서 DBLP 데이터에 기초하여 만든 평가셋으로 서로 다른 논문에서 서로 다른 크기(9, 10, 14)의 저자명 수를 실험에 사용하였다. 표에 제시된 psu-citeseer-9/10/14의 수치들은 각각 Song[1], Huang[7], Pereira[6]의 논문에서 발췌한 것들이다. psu-pike-24는 미국 펜실베니아주립대학교의 PIKE 연구그룹에서 DBLP 데이터를 주로 하여 만든 평가셋이다[3][8]. 미국 메사츄세츠대학교에서 구축한 세 평가셋 umass-dblp-17, umass-rexa-8, umass-penn-7은 DBLP 데이터와 REXA 코퍼스를 기초로 구축한 것이다[4]. southampton-8은 영국 사우스앰튼대학교에서 AKTiveAuthor 시스템의 저자식별기능 평가를 위해 구축한 것이다[5].

기존 평가셋들의 경우 대상이 되는 저자명 수가 7~24로 많지 않다. 이로 인해 먼저 비영어권 연구자의 영어식 이름 표기를 포함한 다양한 저자명 표현의 문제를 다루는데 평가셋의 활용도가 제한될 수 있다. 또한 하나의 저자명은 평가셋 내에서 저자식별의 기본 단위가 되는 동명저자명개체집합에 대응되므로, 저자명 수의 적음은 개체집합 내에서 발생할 수 있는 다양한 저자 매핑 상황을 평가하는데 어려움이 있을 수 있다. 극단적으로 어떤 평가셋의 각 개체집합 내의 모든 저자명개체들이 단일 저자로만 사상된 경우를 가정해 보라.

psu-citeseer 평가셋은 저자중의성이 큰 저자명개체들로 구성되어 있으며 psu-pike 평가셋은 그 반대이다. 저자중의성과 저자식별 문제의 복잡성이 비례하는지는 확실치 않다. 그러나 크기 n인 개체집합의 저자중의성이 최고값 n인 경우 개별군집법[3]을 적용되면 간단하게 최고성능을 얻을 수 있을 것이다. 따라서 적어도 다양한 분포를 갖는 저자중의성이 내재되도록 평가셋을 구축할 필요가 있을 것이다. 이러한 관점에서 기존 평가셋들은 저자중의성이 발현되는 개체집합의 개수가 많지 않아 다양한 저자중의성을 갖는 실제 학술정보데이터의 상황을 반영하는데 어려움이 있다고 판단된다.

III. 평가셋 구축

새로운 저자식별 평가셋 구축 과정은 다음과 같다.

- 단계-1: 논문서지집합 결정
- 단계-2: 저자명집합 결정
- 단계-3: 저자명개체집합 생성
- 단계-4: 저자식별 정보 수집
- 단계-5: 저자식별자 부착
- 단계-6: 검증 및 단계-5 반복

[단계-1]은 평가셋 구축의 대상이 될 저자명이 출현한 논문의 서지레코드 집합을 결정하는 단계이다. 이를 위해 ArXiV, CiteSeer, CS BiBTeX, DBLP, NCSTRL 등의 기존 서지데이터베이스[4]를 비교 검토하였으며, 정확성, 공용성, 대용량성, 획득용이성, 대중성 등을 동시에 고려하여 DBLP 데이터를 논문서지집합의 모집단으로 선정하였다. DBLP는 2009년 현재 백이십만 편 이상의 전산학 분야 논문의 서지레코드를 수작업 구축하여 온라인 서비스하는 사이트이다[9]. 2007년 후반 원천 서지집합인 DBLP로부터 논문서지레코드를 다운로드 받아 87만여편에 달하는 논문서지집합(DBLP-Bib)을 확보하였다.

[단계-2]는 DBLP-Bib에 출현한 저자명 중 평가셋에 포함시킬 저자명집합을 선정하는 과정이다. 저자식별 문제를 탐구하는 관점에서 다양한 저자중의성을 갖는

3) 크기 n인 개체집합의 각 개체를 하나의 군집으로 하여 총 n개의 군집들로 만드는 군집법

4) DBLP와 CiteSeer의 비교는 Petricek(참고문헌 [10])를 참조

저자명이 평가셋에 포함되어야 할 것이다. 또한 같은 수준의 저자중의성을 갖는 다양한 크기의 동명저자명 개체집합들이 평가셋에 포함되어야 할 것이다. 그 이유는, 예를 들어 저자명 J. Smith의 저자중의성이 2라 하더라도 J. Smith가 출현한 논문(동명저자명개체집합)의 수가 2인 경우와 200인 경우의 저자식별문제의 어려움은 큰 차이가 있을 것이기 때문이다.

그러나 저자명에 대한 실세계 저자로의 대응이 확인되기 전에는 저자중의성을 알 수 없으므로 [단계-2]에서 전술한 두 가지 인자를 고려하기에는 어려움이 따른다. 이 문제를 다루기 위해, "저자명 출현 회수와 저자중의성은 비례할 가능성이 크다"는 가정에 기초하여 DBLP-Bib 내 저자명 출현 고빈도 순으로 상위 1000개 저자명을 추출하여 식별 대상 저자명집합(DBLP-NameSet)으로 선정하였다. 예를 들어 논문서지집합이 아래 3편의 논문으로 구성된 경우 저자명 출현 빈도 순으로 상위 2개의 저자명을 추출하면 J. Mitchell(3회)과 P. Lincoln(2회)이 저자명집합으로 선정될 것이다.

[논문서지집합 예]

- J. Mitchell. 1983. File Servers. AC, 221-259.
- P. Lincoln, J. Mitchell. 1991. Algorithmic Aspects of Type Inference with Subtypes. POPL, 293-304.
- P. Lincoln, J. Mitchell, A. Scedrov. 1996. Linear logic proof games and optimization. BSL, 322-338.

[단계-3]은 이전 단계에서 결정된 DBLP-NameSet를 구성하는 1000개 각 저자명의 출현 개체들을 DBLP-Bib로부터 수집하여 저자명개체집합(DBLP-NameEntitySet)을 생성하는 것이다. 예를 들어 전술한 예인 [논문서지집합 예]를 논문서지집합으로 보고 여기서 얻어진 저자명집합이 {J. Mitchell, P. Lincoln, A. Scedrov}라고 하면, 이에 대응하는 저자명개체집합은 다음과 같다.

[저자명개체집합 예]

- <J. Mitchell> J. Mitchell. 1983. File Servers. AC, 221-259.
- <J. Mitchell> P. Lincoln, J. Mitchell. 1991. Algorithmic Aspects of Type Inference with Subtypes. POPL, 293-304.
- <J. Mitchell> P. Lincoln, J. Mitchell, A. Scedrov. 1996. Linear logic proof games and optimization. BSL, 322-338.
- <P. Lincoln> P. Lincoln, J. Mitchell. 1991. Algorithmic Aspects of Type Inference with Subtypes. POPL, 293-304.
- <P. Lincoln> P. Lincoln, J. Mitchell, A. Scedrov. 1996. Linear logic proof games and optimization. BSL, 322-338.
- <A. Scedrov> P. Lincoln, J. Mitchell, A. Scedrov. 1996. Linear logic proof games and optimization. BSL, 322-338.

실제로 저자명개체집합은 동명저자명개체집합(들)의 모음으로 이루어진다. 위 예의 저자명개체집합은 저자명집합을 구성하는 3개 저자명 각각에 대한 동명저자명개체집합들의 모음인 것이다. 예를 들어 위에서 저자명 <P. Lincoln>에 해당하는 동명저자명개체집합은 P. Lincoln이 출현한 논문 두 편의 모음이다.

즉 [단계-3]은 DBLP-NameSet를 구성하는 1000개 저자명에 대응하는 1000개 동명저자명개체집합들로 이루어진 DBLP-NameEntitySet를 생성하는 것이다.

[단계-4]는 DBLP-NameEntitySet 내의 각 저자명개체에 대해 실세계 저자를 대응시키기 위한 정보를 수집하는 단계이다. 기존 평가셋 구축의 경우 각 저자명개체의 홈페이지 내 출판논문리스트페이지(Personal Publication List page, PPLpage)를 참조하거나 저자명개체가 출현한 논문의 원문에 기재된 전자메일주소로 확인 메일을 발송하는 방식 등을 통해 실세계 저자의 신원을 확인했다. 그러나, 이 연구에서 사용하는 DBLP 데이터의 경우 전자메일 획득을 위한 원문 확보가 쉽지 않고, 1000개 저자명의 실세계 저자들의 홈페이지를 수작업으로 검색하는 것 또한 시간/인력 집약적 작업이 되는 것을 피할 수 없다.

이 문제를 다루기 위해 저자의 출판논문정보가 기재된 웹페이지를 구글 웹 검색을 통해 자동 획득하고자 시도하였다. 먼저 기존 홈페이지 탐색 기법[11]에서 활용된 단서 용어들(curriculum vitae, cv, resume, homepage, publication)과 특정 저자명개체가 출현한 논문의 제목을 저자명과 함께 구글검색엔진의 다양한 검색 옵션(intitle:, allintitle:, site: 등)과 조합하여 웹검색을 시도하였다. 그러나 이 방법이 만들어 낼 수 있는 검색식의 조합이 적지 않아, 저자명개체집합으로부터 무작위 추출된 100개 저자명개체에 대해 정답셋을 만들고 이를 이용해 최적 검색식의 조합을 찾는 과정을 거쳤다. 그 결과 웹페이지의 제목 문자열에 저자명의 성(lastname)이 출현하면서 웹페이지의 본문에 저자개체의 논문제목이 같이 출현하는 웹페이지를 검색하는 구글검색식이 가장 좋은 성능을 보였다. 다음은 J. Mitchell에 대한 특정 저자명개체와 그 개체에 대한 PPLpage를 웹검색하기 위한 최적 구글검색식의 예를 보인 것이다.

- 저자명개체: <J. Mitchell> P. Lincoln, J. Mitchell. 1991. Algorithmic Aspects of Type Inference with Subtypes. POPL, 293-304.
- 구글검색식: intitle:Mitchell Algorithmic Aspects of Type Inference with Subtypes

전술한 구글검색식을 사용하여 [단계-3]에서 얻어진 DBLP-NameEntitySet 내의 각 저자명개체에 대해 구글웹검색을 수행하여 상위 20개의 검색결과를 자동 수집하였다.

[단계-5]는 이전 단계에서 수집된 저자명개체의 식별정보를 바탕으로 각 저자명개체에 식별자를 부여하는 단계이다. 이를 위해 먼저 DBLP-NameEntitySet 내의 동명저자명개체집합 단위로, 각 저자명개체에 대해 [단계-4]에서 수집된 20개 웹페이지 중 정답 PPLpage의 URL을 수작업으로 찾아 할당하고, 동일 URL이 부여된 저자명개체들에 동일 고유식별자(자연수)를 부착하는 절차를 거쳤다.

이 과정에서 구글검색결과에서 정답 PPLpage URL을 찾을 수 없는 저자명개체의 수가 적지 않았고 [단계-4]의 검색결과를 생성하지 못한 저자명개체도 다수 발견되었다. 그 결과 최초 1000개 DBLP-NameSet은 867로 줄어들었고, DBLP-NameEntitySet은 총 41,673개의 저자명개체를 포함하게 되었다.

[단계-6]에서는 [단계-5]의 식별자 부착 결과를 재확인하여 수정하고 필요할 경우 [단계-5]의 작업을 재수행하는 과정을 거친다. 이 단계의 주요 작업 사례로, 서로 다른 PPLpage URL이 할당되어 서로 다른 저자식별자가 부착된 저자명개체들이 재확인을 통해 동일 저자식별자로 병합되는 예가 있다. 구체적인 예로는 서로 다른 두 PPLpage URL들이 웹 서버의 부모-자식 디렉토리 위치에 존재하거나, 연구자의 소속 변경으로 인해 이전 소속기관과 현재 소속기관의 웹사이트에서 유사한 논문출판리스트들이 유지되고 있는 경우 등이 해당된다.

IV. 평가셋 특성

이 연구에서 구축한 저자식별 평가셋은 영어(English) 저자명을 대상으로 하는 저자식별(Author Disambiguation)을 위해 한국과학기술정보연구원(KISTI)에서 구축한 첫 번째(01) 평가셋(TestSet)이라는 의미에서 KISTI-AD-E-01-TestSet으로 명명하였으며 그 통계는 [표 2]와 같다.

표 2. KISTI-AD-E-01-TestSet 통계

항목	값
논문수	37,613
동명저자명개체그룹수	881
실세계 저자수	6,921
저자명개체수	총 116,564 중 41,673
저자중의성	12.7 (=6,921/881)
논문 당 평균 저자수	3.1 (=116,564/37,613)

[표 2]에 제시된 바와 같이 평가셋은 서로 다른 881개 영어저자명이 출현한 41,673개 저자명개체레코드(논문서지레코드)들로 이루어져 있다. 저자명개체레코드

41,673개를 서로 다른 논문의 수로 카운트하면 37,613편이고, 37,613편 내에 출현한 저자명개체의 수는 총 116,564이다. 다시 말하면 전체 37,613편의 논문집합에 출현한 총 116,564개의 저자명개체 중 41,673개의 저자명개체 각각에 대해 실세계 저자 6,921명 중 한 사람의 식별자를 부여한 것이며 평균 저자중의성은 12.7이다.

[그림 1]은 특정 크기의 동명저자명개체그룹의 수가 평가셋 내에서 분포하는 모습과, 총 881개 그룹 중 특정 크기 이하 그룹들이 차지하는 비율(점선)을 나타낸 것이다. 최소 크기 1에서 최대 크기 325까지의 881개 그룹들이, 그룹 크기가 증가할수록 같은 크기 그룹들의 개수는 점차 감소하는 양상을 보인다. 또한 크기 10, 30, 50, 100, 150, 200 이하 그룹들이 각각 전체의 약 21%, 49%, 67%, 87%, 96%, 99%를 차지하여 적은 크기 그룹들(크기 30이하 거의 50%)이 상대적으로 많으나 큰 크기 그룹들(크기 100이상 13%, 117개 그룹)도 적지 않았다. 이는 저자식별의 단위가 되는 동명저자명 개체 집단의 크기 분포가 획일적이지 않음을 보여준다.

[그림 2]는 특정 개수의 실세계 저자수를 갖는 동명저자명개체그룹의 개수를 나타낸 것이며, 점선은 총 881개 중 특정 수 이하 저자를 갖는 동명저자명개체그룹의 비율이다. 그림에서 X축인 저자수는 최소 1에서 최대 71까지이며 저자수가 적을수록 해당 동명저자명개체그룹의 수는 가파르게 증가하였다. 구체적으로 저자수 1, 3, 5, 10, 20, 30, 50 이하 그룹들이 각각 전체의 약 12%, 33%, 51%, 80%, 93%, 97%, 99%를 차지하였다. 즉 저자중의성이 낮은 그룹이 대부분이지만, 20인 이상의 다수 저자들 중 하나로 저자명개체들을 매핑해야 하는 그룹들도 69개(전체의 7.8%)로 적지 않은 개수임을 알 수 있다.

[그림 3]은 특정 수의 공동저자를 갖는 저자명개체의 수를 보인 것으로 점선은 총 41,673개 중 특정 수의 공동저자를 갖는 저자명개체의 비율이다. 저자식별에서 공동저자명은 전자메일주소, 소속 등과 함께 개인 저자 개별성이 큰 자질이므로 공동저자수의 분포를 살피는 것은 의미가 있다. 구체적으로 평가셋 내에서 공동저자수 0, 1, 2, 3, 4를 갖는 저자명개체들이 각각 전체의 8%, 30%, 30%, 18%, 8%를 차지하였다. 즉 공동저자수 2-4인 개체들이 전체의 80%정도를 차지한다. 공동저자수가 0인 저자명개체 3,349개(8%)에 대해서는 저자식별과정에서 공동저자명 자질을 활용할 수 없음을 의미한다.

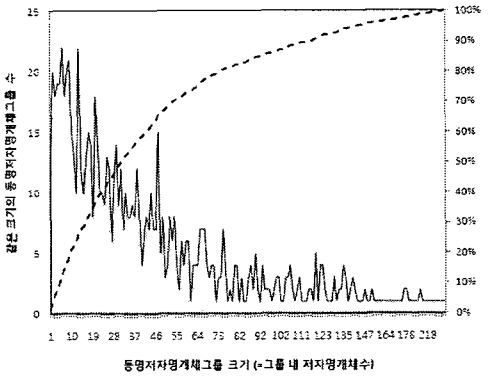

그림 1. 동명저자명개체그룹의 크기 분포

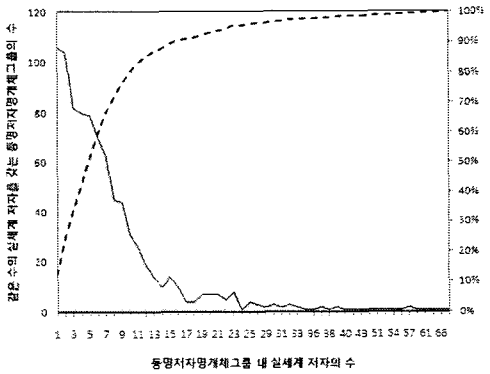

그림 2. 동명저자명 저자의 수 분포

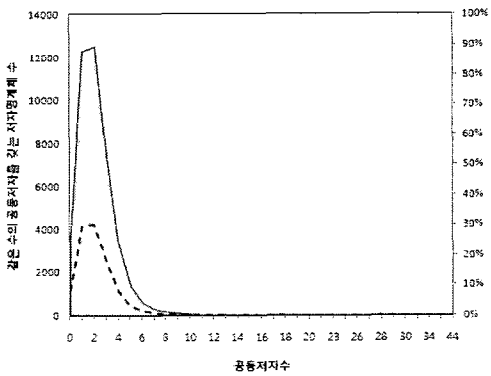

그림 3. 공동저자수 분포

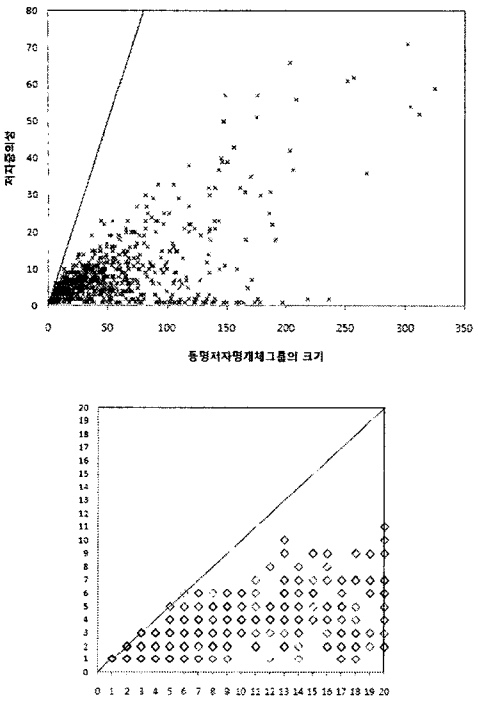

그림 4. 저자명개체수-저자수 분포

[그림 4]는 전체 881개 동명저자명개체그룹에 대해 그룹 내 저자명개체의 수와 실세계 저자의 수(저자중의성)를 2차원 좌표 상의 점(표식 x)으로 보인 것으로, 하단 그림은 상단 그림의 밀집 영역을 확대한 것이다. 예를 들어 동명저자명개체그룹 G={a, b, c}에서 저자명개체집합 {a, b}, {c}가 실세계 저자 P1, P2에 각각 대응될 경우 G는 [그림 4]에서 좌표 (3,2)에 하나의 점으로 표시된다. 그림에서 실선은 저자명개체수와 저자수가 같은 점들을 연결한 참고용 기준선(Y=X)이다. 기준선에 가까운 점일수록 저자군집의 수가 저자명개체수에 가까운 개체그룹임을 의미한다.

극단적으로 기준선에 위치한 군집들은 개별군집법(singleton clustering)으로 최상의 성능을 낼 수 있고, Y=1에 대응하는 선에 위치한 군집들은 단일군집법[5](single clustering)으로 최상의 성능을 만들 수 있다. 따라서 군집문제에 해당하는 저자식별의 경우 Y=X와 Y=1 사이의 영역에 고르게 분포하도록 평가셋이 구성되는 것이 좋을 것이다. 그렇게 함으로써 동일 크기 그룹들 내에서 저자수 분포가 다양할 것이고 동일 저자수를 갖는 그룹들 내에서 그룹 크기의 분포 또한 다양할 것이기 때문이다.

이러한 관점에서 [그림 4]를 고찰하면 현재의 평가셋은 적은 크기 그룹들의 경우 저자중의성의 고른 분포를 보인다. 그러나 크기 50에서 150이상 그룹들의 경우 그룹 크기의 약 1/3이상(예: 크기 150인 그룹의 경우 1/3 크기인 50이상의 저자중의성을 찾기 힘들다)의 저자중의성 분포는 거의 발견되지 않는다. 또한 크기 150이상 그룹들의 경우 그룹 크기의 중간 범위에 해당하는 일부 저자중의성 분포만이 발견된다. 이는 동명저자명개체그룹의 크기가 증가할수록 저자중의성의 분포 범위가 커지므로 충분한 수의 개체그룹들이 평가셋에 포함되어야 하지만 현재의 평가셋은 [그림 1]의 개체그룹 크기 분포에서 알 수 있듯이 개체그룹의 크기가 증가할수록 개체그룹의 개수는 감소하기 때문이다.

V. 저자식별성능

이 장에서는 평가셋 KISTI-AD-E-01-TestSet에 대한 저자식별성능을 제시한다. 먼저 평가셋에 포함된 총 881개 개체그룹 중 크기가 1인 14개의 개체그룹을 제외하여, 총 867개 그룹 41,659 저자명개체들을 실험대상으로 하였다. 이는 적어도 2개 이상의 저자명개체를 포함한 개체그룹을 저자식별의 실험대상으로 하기 위함이다. 저자식별을 위한 군집법으로는 단일링크법(Single-Linkage Agglomerative Hierarchical Clustering)을 사용하고 군집법이 요구하는 개체거리함수로는 코사인함수와 이진거리함수[12]를 적용하였다. 저자식별의 평가지표[13]로 재현율, 정확률, F1, OE(Over-clustering Error, 과다군집오류), UE(Under-clustering Error, 과소군집오류)를 사용한다.

[표 3]은 새로운 평가셋의 저자식별성능을 보인 것이다. 표에서 F, C, T, P, Y는 저자명개체레코드가 갖는 자질들이며 각각 저자명완전명(Fullname), 공동저자명(Coauthor names), 논문제목(Title), 게재지명(Publication title), 게

5) 크기 n인 개체집합을 하나의 단일 군집으로 만드는 군집법

재연도(Year)이다. 자질 CTPY는 공동저자명, 논문제목, 게재지명, 게재연도 문자열의 모음을 하나의 문헌으로 보고 문헌 간 유사도를 계산한 것이다. 논문제목(T) 자질이나 CTPY 자질이 사용된 경우 유사도 계산을 위해 TF/IDF 용어가중치에 기반한 코사인함수가 사용되었다. F+T 등에 사용된 기호 +는 +로 연결되는 두 자질의 자질값들의 합집합을 의미한다.

표 3. 저자식별 성능

자질	Rec.	Pre.	F1	OE	UE
F	.9797	.9352	.9569	.0273	.0082
C	.4205	.9645	.5856	.0062	.2330
T	.4989	.5010	.4999	.1998	.2015
P	.0679	.9152	.1263	.0025	.3748
Y	.0815	.3520	.1323	.0603	.3693
F+T	.9832	.9340	.9579	.0279	.0068
F+C	.9870	.9104	.9472	.0391	.0052
C+T	.7063	.8942	.7892	.0336	.1181
CTPY	.5355	.7018	.6075	.0915	.1868
F+CTPY	.9797	.9352	.9569	.0273	.0082

단일자질 성능으로 저자명완전명 자질이 가장 좋았고 다음으로 공동저자명, 논문제목, 게재연도, 게재지명 순이었다. 단일자질 중 상위 성능의 것들을 이중 결합한 경우 F+T의 조합이 가장 좋았다. 결합자질 CTPY는 개별 자질들의 성능은 능가하였으나, 저자명완전명(F)와의 결합(F+CTPY)에서는 잉여적으로 작용하였다.

VI. 결론

저자 중심의 학술정보서비스를 제공하기 위해서 저자식별 문제 해결은 반드시 필요하다. 기존 저자식별 평가셋들의 한계를 극복하고 다양한 저자식별 실험을 장려하기 위해서, 본 연구에서는 대용량성과 출현 저자명의 다양성이 보강된 새로운 저자식별 평가셋을 구축하였다. 새롭게 구축된 평가셋은 DBLP 데이터에 출현한 고빈도 저자명들에 대해 웹 검색을 통한 수작업 식별 과정을 거쳐 만들어졌으며, 현재 881개 저자명에 대해 수집된 41,673개의 저자명개체레코드를 대상으로 총 6,921명의 실세계 저자 식별자가 존재한다.

본 논문은 저자식별 평가셋의 구축 절차, 특성 및 저자식별 성능을 기술하였으며, 총 6단계 구축 절차 중 4단계에서 웹 검색을 통해 저자식별정보를 자동 수집한 것은 구축 과정의 가속화에 큰 기여를 하였다. 그러나 5단계 저자식별자 부착에서 웹 검색 결과를 수작업으로 검사하여 저자명 개체의 개인논문출판정보 페이지를 확인하는 작업은 오랜 시간이 소요되었다. 단계 5에서 정답 PPLpage가 상위에 배치되도록 검색 결과의 순위를 조정하는 시도는 사람의 확인 시간을 대폭 감소시킬 수 있을 것이다. 향후 단계 5에 대한 자동화 연구를 통해 반자동 저자식별 기법들을 탐구할 필요가 있을 것이다.

참 고 문 헌

[1] Y. Song, J. Huang, I. Councill, J. Li and C. L. Giles, "Efficient topic-based unsupervised name disambiguation," In Proceedings of the ACM IEEE Joint Conference on Digital Libraries (JCDL), 2007(6).

[2] H. Han, H. Zha, and C. L. Giles, "Name disambiguation in author citations using a k-way spectral clustering method," In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries(JCDL), pp.334-343, 2005(6).

[3] D. W. Lee, B. W. On, J. W. Kang, and S. H. Park, " Effective and scalable solutions for mixed and split citation problems in digital libraries," In Proceedings of the International Workshop on Information Quality in Information Systems(IQIS), pp.69-76, 2005(6).

[4] P. Kanani and A. McCallum, "Efficient strategies for improving partitioning-based author coreference by incorporating Web pages as graph nodes," In Proceedings of the 6th International Workshop on Information

Integration on the Web(IIWeb-07), 2007(7).

[5] D. M. McRae-Spencer and N. R. Shadbolt, "Also by the same author: AKTiveAuthor, a citation graph approach to name disambiguation," In Proceedings of ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp.53-54, 2006(6).

[6] D. A. Pereira, B. Ribeiro-Neto, N. Ziviani, A. H. F. Laender, M. A. Goncalves, and A. A. Ferreira, "Using web information for author name disambiguation," In Proceedings of ACM/IEEE-CS Joint Conference on Digital Libraries(JCDL), pp.49-58, 2009(6).

[7] J. Huang, S. Ertekin, and C. L. Giles, "Efficient name disambiguation for large scale databases," In Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD), pp.536-544. 2006(9).

[8] Y. F. Tan, M. Y. Kan, and D. W. Lee, "Search engine driven author disambiguation," In Proceedings of ACM/IEEE Joint Conference on Digital Libraries(JCDL), pp.314-315, 2006(6).

[9] M. Ley, "DBLP - some lessons learned," In Proceedings of International Conference on Very Large Data Bases(VLDB), 2009(8).

[10] V. Petricek, I. J. Cox, H. Han, I. G. Councill, and C. L. Giles, "A comparison of on-line computer science citation databases," In Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), 2005.

[11] O. Fatemieh, K. Manzoor, A. Jain, and A. Ramani, "Home Page Finder. University of Illinois at Urbana-Champaign," 2005.

[12] 강인수, "한글 저자명 군집화를 위한 계층적 기법 비교", 정보관리연구, 제40권, 제2호, pp.95-115, 2009.

[13] I. S. Kang, S. H. Na, S. W. Lee, H. M. Jung, P. Kim, W. K. Sung, and J. H. Lee, "On co-authorship for author disambiguation," Information Processing and Management, Vol.45, No.1, pp.84-97, 2009.

저 자 소 개

강 인 수(In-Su Kang) 정회원

- 1995년 : 경북대학교 컴퓨터공학과 학사
- 1999년 : 포항공과대학교 컴퓨터공학과 석사
- 2006년 : 포항공과대학교 컴퓨터공학과 박사
- 1995년 ~ 1997년 : (주)포스데이타
- 1999년 ~ 2001년 : 포항공과대학교 학술정보원
- 2006년 ~ 2008년 : 한국과학기술정보연구원 선임연구원
- 2008년 ~ 현재 : 경성대학교 컴퓨터정보학부 교수

<관심분야> : 시맨틱 웹, 정보검색, 자연어처리

김 평(Pyung Kim) 정회원

- 1997년 : 충남대학교 전산학과 학사
- 1999년 : 충남대학교 컴퓨터과학과 석사
- 2004년 : 충남대학교 컴퓨터과학과 박사
- 2000년 ~ 2003년 : ㈜엔퀘스트테크놀러지 개발실장
- 2003년 ~ 2004년 : 충남대학교 강사
- 2004년 ~ 현재 : 한국과학기술정보연구원 선임연구원
- 2008년 ~ 현재 : 한국정보교육학회 이사

<관심분야> : 시맨틱 웹 , 정보검색 , 자연어처리

이 승 우(Seungwoo Lee) 정회원

- 1997년 : 경북대학교 컴퓨터공학과 학사
- 1999년 : 포항공과대학교 컴퓨터공학과 석사
- 2005년 : 포항공과대학교 컴퓨터공학과 박사
- 1999년 ~ 2000년 : 포항공과대학교 정보통신연구소 연구원
- 2005년 ~ 2006년 : 대구가톨릭대학교 컴퓨터교육과 강의전담교원
- 2006년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 시맨틱 웹, 정보검색, 자연어처리

정 한 민(Hanmin Jung) 정회원

- 1992년 : 포항공과대학교 전자계산학과 학사
- 1994년 : 포항공과대학교 전자계산학과 석사
- 2003년 : 포항공과대학교 컴퓨터공학과 박사
- 1994년 ~ 2000년 : 한국전자통신연구원 선임연구원
- 2000년 ~ 2004년 : ㈜다이퀘스트 연구소장/기술이사
- 2004년 ~ 현재 : 한국과학기술정보연구원 책임연구원
- 2004년 ~ 현재 : 과학기술연합대학원대학교 겸임교수
- 2009년 ~ 현재 : 한국외국어대학교 언어연구소 초빙연구원

<관심분야> : 시맨틱 웹, 정보검색, 정보추출, 자연어처리, HCI

류 범 종(Beom-Jong You) 정회원

- 1984년 : 서강대학교 전자공학과 학사
- 2000년 : 충남대학교 문헌정보학과 석사
- 2004년 : 충남대학교 문헌정보학과 박사
- 1987년 ~ 1993년 : 시스템공학연구소실장
- 1990년 : IBM in Australia 연구원
- 1993년 ~ 2000년 : 연구개발정보센터 실장
- 2001년 ~ 현재 : 한국과학기술정보연구원 실장
- 2001년 ~ 2006년 : 충남대학교 문헌정보학과 강사
- 2007년 ~ 현재 : 충남대학교 문헌정보학과 겸임교수

<관심분야> : 자연어처리, 시맨틱 웹, 정보검색