논문 원문을 이용한 동명 저자 자동 군집화

Automatic Clustering of Same-Name Authors Using Full-text of Articles

강인수, 정한민, 이승우, 김평, 구희관, 이미경, 구남앙, 성원경
한국과학기술정보연구원, NTIS 사업단

Kang In-Su, Jung Han-Min, Lee Seung-Woo, Kim Pyung, Goo Hee-Kwan, Lee Mi-Kyung, Goo Nam-Ang, Sung Won-Kyung
NTIS Division, Korea Institute of Science and Technology Information

요약

대용량 과학 기술 문헌의 탐색 및 검색에 있어서 저자, 저자 소속 기관, 게재지 등에 대해 고유 식별자에 기반한 표현의 필요성이 증가하고 있다. 특히, 과학 기술 문헌의 저자가 단순히 이름으로 표현될 경우, 동일명을 가진 서로 다른 저자들에 대한 구분은 사용자의 검색 부담을 가중시키게 된다. 이러한 동명이인의 문제를 해결하기 위한 기존의 접근법들은 공저자 정보, 논문 제목 등의 서지정보에 의존하는 공통점을 지닌다. 그러나, 기존의 방법들은 공저자가 없거나 논문 제목 간의 공통 어휘가 발견되지 않을 경우 어려움을 겪게 된다. 본 연구에서는, 동명저자 문제 해소를 위한 기존의 접근법을 보완하기 위해, 동명저자들의 논문 원문의 내용에 기반한 문서 군집화 방법을 사용한다. 국내 학술대회 발표 논문집을 대상으로 한 실험에서 제안한 방법이 기존의 서지정보에 기반한 해법의 단점을 보완할 수 있다는 가능성을 보였다.

Abstract

Bibliographic information retrieval systems require bibliographic data such as authors, organizations, source of publication to be uniquely identified using keys. In particular, when authors are represented simply as their names, users bear the burden of manually discriminating different users of the same name. Previous approaches to resolving the problem of same-name authors rely on bibliographic data such as co-author information, titles of articles, etc. However, these methods cannot handle the case of single author articles, or the case when articles do not have common terms in their titles. To complement the previous methods, this study introduces a classification-based approach using similarity between full-text of articles. Experiments using recent domestic proceedings showed that the proposed method has the potential to supplement the previous meta-data based approaches.

I. 서 론

웹의 보편화는 기존에 종이의 형태로 접근했던 논문 정보(예: 저자, 제목, 출처 등 서지정보와 원문 그리고 인용정보 등)를 전자화된 형태로 이용할 수 있도록 변화시켰다. 이후, 전자화된 논문 정보의 양은 점점 증가하고 있으며, 이에 따라 다양한 논문 검색 사이트들이 생겨났다. 그러나, 현재의 문자 기반의 논문 정보 표현 방식에 있어서는, 연구자나 발명자가 원하는 논문 정보를 찾는 데 많은 시간과 노력을 요구하게 된다. 예를 들어, 논문 정보 중 저자는 이름으로 표현되는데, 그 기술 방식이 영어의 경우 통일되어 있지 않으며(예: John R. Smith vs. Jonathan Richard Smith), 동명 저자의 중의성(예: Harvard Univ.의 John R. Smith와 MIT의 John R. Smith)이 해소되어 있지 않음으로 인해, 저자명으로 논문을 검색할 경우 재현율과 정확률을 떨어뜨리게 된다. 또한, 이러한 저자 표현의 문제는 대용량 논문들로부터 구축된 저자 인용망의 질적 수준을 떨어뜨리게 된다.

본 연구에서는 전술한 저자명 표현의 문제 중 동명저자의 중의성을 해소하는 방법을 다룬다. 동명 저자 중의성 해소를 위한 기존 방법들은[Bhattacharya and Getoor, 2005; Han et al., 2004, 2005a; Han et al., 2005b; Lee et al., 2005; On et al., 2005; 이승우 et al., 2006], 중의성 해소의 자질 측면에서, 개별 동명 저자가 작성한 논문의 공동저자(들), 제목, 게재지명 등을 사용하고 있다. 이러한 자질을 사용하는 기본 가정은, "한 연구자는 일정 기간 동안, 특정한 몇 명의 연구자와 공동 연구를 수행하며, 연구 분야가 크게 바뀌지 않으며, 제한된 몇 개의 학술대회나 학술지에 논문을 투고하는 경향이 있다"는 것이다.

그러나, 동명 저자의 논문집합에서 논문간 공유되는 공저자가 없을 수 있으며, 한 저자의 연구 분야를 기술하는 데 있어서 논문 제목만으로는 부족할 수 있으며, 동일 게재지에 투고

하는 동명 연구자는 적지 않을 수 있다. 본 연구에서는 서지 정보로부터 획득되는, 동명 저자의 자질 타입의 한계를 극복하기 위해 논문의 원문 텍스트를 자질로 활용하고자 한다. 특히, 이 연구에서는 논문의 원문 텍스트 자질이 군집화(clustering) 기반 동명 저자 중의성 해소 방법에 효과가 있는지를 실험적으로 보이는 데 중점을 둘 것이다.

논문의 구성은 다음과 같다. 먼저 2장에서 관련 연구를 기술한다. 3장에서는 원문 텍스트 자질을 활용한 동명 저자 중의성 해소를 위한 군집화 방법과 그 효과에 대해 실험적 평가 절차를 중심으로 기술한다. 마지막으로 4장에서 결론을 맺는다.

II. 관련 연구

개체 집단과 개별 개체를 지칭하는 표현을 구분하는 문제는 데이터베이스 분야에서 레코드 링키지(recordd linkage)란 이름 아래 연구되고 있다[Newcombe et al., 1959; Elmagarmid et al., 2006]. 이 문제를 세분하면, 동일 개체에 대한 서로 다른 표현을 찾는 동일개체/복수표현의 문제와, 동일 표현을 갖는 서로 다른 개체를 국분하는 복수개체/동일표현의 문제로 나누어 볼 수 있다. 동일개체/복수표현의 문제는, 데이터베이스 분야에서 여러 DB를 통합할 때 빈번히 발생하는 중복 레코드 제거와 관련된 것이며, 레코드 링키지 연구의 주요 토픽이다. 이 분야의 연구는, 주로 사람이나 주소 개체를 지칭하는 영문 이름(예: John R. Smith vs. Jonathan Richard Smith)이나 영문 주소 표현(예: 17 East Fifth Street vs. 17 E5th) 등의 중복 제거에 적용됨으로 인해, 유사 표현을 찾는 방법은 주로 두 표현간의 스트링 비교 기법(예: edit distance[Levenshtein, 1965], Jaro[Jaro, 1976], cosine similarity[Cohen, 1998])에 의존한다.

복수개체/동일표현의 문제는, 최근 원문을 포함한 대용량 서지 정보가 가용화되면서, 인용망의 자동 구축 측면에서 논문 저자명의 동명이인 해소를 위해 새롭게 다루어지고 있다. 전술한 것처럼 동일개체/복수표현의 문제는 표현상의 형태적 변이를 다루기 위해 표현 내적인 자질(비교 대상이 되는 두 스트링 자체)을 적극적으로 활용할 수 있지만, 복수개체/동일표현의 문제는 개체 표현이 동일하므로 표현 외적인 자질간 유사성을 사용하여 해결해야 한다. 서지 정보의 경우 동명 저자 해소를 위한 외적 자질로는 공동 저자명, 논문 제목, 게재지명 등이 사용된다[Bhattacharya and Getoor, 2005; Han et al., 2004, 2005a; Han et al., 2005b; Lee et al., 2005; On et al., 2005; 이승우 et al., 2006]. 기존 연구에서 사용된 상기의 외적 자질과 달리, 본 연구에서는 논문 원문에서 추출한 텍스트를 동명이인 해소 문제의 자질로 사용하고자 한다.

외국인 저자에 대한 동명이인 해소에 있어서는, 동일 저자명 표기의 다양한 변이형들로 인해, 동일개체/복수표현과 복수개체/동일표현의 두 문제가 복합적으로 발생한다. 한글 저자명의 경우는 이름 표기에 일관성이 있어 동일개체/복수표현의 문제가 없는 듯 보이나, "홍길동"과 "Gil-Dong Hong" 혹은 "Hong, G.D."을 매치해야 하는 상황을 고려하면 유사표현의 그룹화 문제를 무시할 수 없을 것이다. 따라서, 동일 저자명 해소 문제를 다루기 위해서는 개체 표현의 내적 자질과 외적 자질을 동시에 고려하는 것이 일반적이다. 그러나, 본 연구에서는 동명 저자의 중의성 해소 문제를 한글로 제한하여 다룰 것이므로, 내적 자질은 사용되지 않으며, 외적 자질 중에서도 새롭게 제안하는 원문 텍스트 자질의 효용성만을 평가하고자 한다.

동명 저자 해소를 위한 기존 연구는 크게 군집(clustering) 기반 방식[Bhattacharya and Getoor, 2005; Han et al., 2005b; 이승우 et al., 2006]과 분류(classification) 기반 접근법[Han et al., 2004, 2005a; Lee et al., 2005; On et al., 2005]으로 나눌 수 있다. 군집화 방식은 학습데이터가 필요 없지만, 분류 기반 방법은 분류 모델을 만들기 위해 학습데이터가 요구된다는 단점이 있다. 또한, 원론적으로 분류 기반 방식은 분류 대상이 되는 목표 클래스가 미리 정의되어 있을 때 적용된다. 따라서, 동명 저자 해소의 경우처럼 동일 저자명을 갖는 개체들이 실제 몇 명의 사람(클래스)에 대응될 지 미리 정할 수 없는 상황에서 분류적 접근법을 적용하는 것은 무리가 있다 하겠다. 본 연구에서는, 동명 저자 해소 문제에 보다 바람직한 접근법인 군집 기반 방식을 사용한다.

III. 원문을 이용한 동명 저자 군집화

이 장에서는 원문의 텍스트 자질이 동명 저자 중의성 해소에 미치는 영향을 실험적 평가 절차를 중심으로 기술한다. 먼저, 3.1절에서는 실험 집합의 구축 과정을 설명하고, 3.2절에서는 원문을 이용한 동명저자 군집화를 CLUTO 군집화 툴킷을 활용하여 수행하는 방법을 기술한다. 3.3절에서는 평가방법을 설명하고, 3.4절에서는 실험 결과를 논한다.

3.1 실험 집합

실험집합 구축을 위해, 원문 입수가 용이한 최근(2002~2006년) 국내 학술대회 발표 논문집 CD[1]로부터 논문 원문파일을 획득하여, 먼저 개별 논문의 서지정보를 구축하였다. 서지정보는 '논문제목', '게재지정보(게재지명, 권/호, 년도)', '저자정보

1) CD를 입수한 학회는 한국정보과학회, 한국정보처리학회, 한국통신학회, 대한전자공학회, 한국멀티미디어학회, 한국HCI학회 등이다.

(저자명, 소속기관명, 전자메일주소)', '원문파일패스'로 구성하였다. 다음으로, 동명 저자 해소 실험을 위한 정답셋을 구축하기 위해, 개별 동명 저자의 공동저자 리스트, 논문 작성 연도 및 그 당시의 소속기관 그리고 전자메일주소, 과학기술인력검색사이트[2]를 참조하여 수작업으로 서지정보에 출현한 동명 저자들의 중의성을 해소하고 고유한 식별자를 부여하였다.

표 1에 보인 것처럼, 실험을 위해 구축한 정답셋에는, 전술한 학술대회 발표 논문집 CD 원문파일으로부터 텍스트 추출이 가능한 논문 8,418 편과, 동명 저자 해소의 평가 목적에 맞게 적어도 2편 이상의 논문을 작성한 저자 4,483명으로 구성되어 있다.

[표 1] 실험 집합 (원문으로부터의 텍스트추출 가능 논문 8,418편 대상)

동명저자 출현회수	저자수	비율
2	1,799	40.13%
3	887	19.79%
4	552	12.31%
5	335	7.47%
6	229	5.11%
7	143	3.19%
8	140	3.12%
9	87	1.94%
10회 이상	311	6.94%
계	4,483	

3.2 실험 방법

원문의 텍스트 자질을 사용한 동명 저자들의 군집화를 수행하기 위해, 최근 개발된 군집화 툴킷인 CLUTO[3]를 활용하였다. CLUTO를 통해 군집화 알고리즘의 세 가지 패러다임인 partitional, agglomerative, graph-partitioning 알고리즘들과, 4가지 클러스터 간 유사도 함수들(similarity function), 그리고 7가지의 서로 다른 군집화 최적화 함수(criterion function for optimization)들을 선택하여 적용할 수 있다. 본 연구에서는 전술한 CLUTO의 다양한 알고리즘, 유사도 함수 및 최적화 함수들을 동명저자 군집화에 적용해 본 다음, 최적의 옵션으로 graph-partitioning 알고리즘과, 코사인(cosine) 유사도 함수, 그리고 i2 최적화함수를 선정하였다. 또한, CLUTO는 자질값을 계산하는 다양한 함수들을 제공하는데, 이들 중 maxtf와 idf를 결합하여 사용하였다. 이상의 CLUTO 파라미터들을 정리하면 다음과 같다.

%vcluster -clmethod=graph -crfun=i2 -sim=cosine -rowmodel=maxtf -colmodel=idf -clustfile=OUTPUT.txt INPUT.mat CLUSTER_SIZE

위에서, vcluster는 CLUTO에서 제공하는 군집화 실행파일이며, OUTPUT.txt는 군집결과가 저장되는 파일이고, INPUT.mat는 행렬형태로 표현된 군집대상 입력파일이고, CLUSTER_SIZE는 최종 군집의 개수를 지정하는 정수값이다.

군집화 입력 파일인 INPUT.mat는 다음과 같이 만들어 진다. 먼저, 텍스터 필터를 이용하여 논문 원문으로부터 원문 텍스트를 추출한 다음, 형태소 분석기[4]를 통해 원문 텍스트로부터 명사에 해당하는 용어를 그 빈도수와 함께 추출한다. 다음으로, 전체 논문의 원문에서 추출된 전체 용어 집합을 벡터로 구성하고, 개별 논문을 그것의 출현 용어의 빈도수를 벡터요소의 값으로 하는 하나의 벡터로 만든다. 이렇게 만들어진 개별 논문의 용어 벡터들을 결합하여 하나의 행렬로 구성한 것이 INPUT.mat이며, 행의 수는 전체 논문 수에 해당하는 8,418이고, 열의 수는 전체 용어 수에 해당하는 28,338이었다.

CLUTO를 통해 INPUT.mat를 군집화하면 동명 저자의 군집이 아닌 논문의 군집이 만들어지는데, 이 논문의 군집으로부터 저자의 군집을 만드는 것은 간단하다. 즉, 논문과 그것의 저자가 서지정보(3.1절 참조)를 통해 연결되어 있으므로, 군집 A에 속하는 논문들의 저자들을 모아서 군집A'으로 묶어버리면 되기 때문이다. 이런 식으로 논문의 군집에 대응되는 저자의 군집을 하나씩 만들어 나가면 되는 것이다.

3.3 평가 방법

군집화의 성능 평가 수단으로 Rand Index[Rand, 1971]를 사용하였다. Rand Index는, 군집 대상이 되는 임의의 두 개체쌍에 대해, 그 개체쌍이, 시스템이 출력한 군집들과 기준 정답내의 군집들에서 묶이고 떨어지는 여부가 같은지를 판단하여 군집화 성능을 계산한다. 즉, 임의의 개체쌍 (a,b)에 대해, 그것이 시스템의 군집화 결과와 기준 정답에서 공히 같은 군집에 속해 있거나 공히 다른 군집에 속해 있다면 (a,b)의 군집은 정답과 일치한다고 보고, 그 외의 두 경우는 정답과 불일치한다고 판단하는 것이다. 그리하여, 군집화 성능(P)는 아래와 같이 전체 개체쌍의 개수에 대한 일치 개수의 비율로써 계산되며, 군집화의 정확도를 측정한다.

P = number of agreement / (number of agreement + number of disagreement)

2) 국가과학기술인력 종합정보시스템(http://www.hrst.or.kr/)을 활용하였다.

3) http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview/

4) 다이퀘스트(http://www.diquest.com/) 형태소분석기를 사용하였다.

식에서 number of agreement와 number of disagreement는 개체쌍을 시스템의 군집 결과와 기준 정답에서 비교했을 때 일치하는 수와 불일치하는 수이다.

3.4 실험 결과

CLUTO를 이용한 동명 저자 해소를 위한 군집화에서, 군집의 개수(3.2절의 CLUSTER_SIZE)를 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 50 등의 값으로 바꾸어가며 실험했을 때, 군집의 개수를 3으로 지정했을 때의 성능이 가장 좋았다. 표 2는 군집 개수가 3일 때의 동명 저자 군집화 성능을 보여준다. 표 2에서 동종군집 정확률은 기준 정답에서 같은 군집에 속해 있는 개체쌍의 시스템에 의한 군집 정확률이며, 이종군집 정확률은 기준 정답에서 다른 군집에 속해 있는 개체쌍의 시스템에 의한 군집 정확률이다.

[표 2] 동명 저자 군집화 성능 (군집수=3)

동명저자 출현회수	군집 정확률	동종군집 정확률	이종군집 정확률
2	78.40	86.19	13.81
3	75.33	81.06	18.94
4	75.85	78.23	21.77
5	73.58	78.78	21.22
6	75.51	74.71	25.29
7	77.41	80.30	19.70
8	74.62	75.73	24.27
9	73.43	73.46	26.54
10회 이상	75.26	78.74	21.26
계	75.49	78.58	21.42
baseline	70.05	100	0

표 2에서 baseline은 CLUSTER_SIZE를 1로 하여 전체를 하나의 군집으로 구성했을 때의 성능이다. 70%대에 이르는 baseline의 성능으로 추정해 보면, 실험 집합의 동명 저자 중의성은 그리 높지 않다고 볼 수 있다. baseline에 비해 원문 텍스트 자질을 이용한 동명 저자 해소 방법은 군집 정확률 측면에서 전체적으로 7.8%의 성능 향상을 보였고, 동명저자 출현회수에 무관하게 baseline의 성능을 뛰어 넘고 있음을 알 수 있다. 이것은 동명 저자가 작성한 논문의 내용이 저자의 신원을 결정하는 자질로 효과가 있음을 보이는 것이다.

표 2에서는 동명 저자 출현 회수가 커질수록 군집화 성능이 조금씩 떨어지는 경향을 보이는 것을 알 수 있다. 이것은, 동명 저자 출현 회수에 무관하게 (다소 적은) 동일한 군집수(=3)를 적용했기 때문으로 추정된다. 만약, 동명 저자 출현 회수가 클수록 동명저자의 중의성이 높다는 가정을 세울 수 있다면, 향후 동명 저자 출현 회수에 따라 서로 다른 군집수를 적용하여 군집화를 수행하여 성능 향상을 기대해 볼 수 있을 것으로 판단된다.

IV. 결 론

본 연구에서는 논문의 저자 표현에서 발생하는 동명 저자의 중의성 해소 문제를 다루기 위해, 중의성 해소를 위한 자질 타입으로 개별 동명 저자의 논문 원문 텍스트를 사용할 것을 제안하였다. 본 연구의 실험집합에서, 제안한 원문 텍스트 자질은 그 자질만으로 동명 저자 중의성 해소에 효과가 있음을 보였다. 향후, 원문 텍스트 자질이 다른 자질들과 결합되었을 때, 동명 저자의 중의성을 해소하는 데 도움이 되는 지 살펴볼 필요가 있을 것이다.

▌참 고 문 헌▐

[1] Bhattacharya, I., and Getoor, L. (2004). "Iterative record linkage for cleaning and integration", *Proceedings of SIGMOD-2004 Workshop on Research Issues in Data Mining and Knowledge Discovery*, Paris:France, Jun. 13, 2004, pp.11-18.

[2] Cohen, W.W. (1998). "Integration of heterogeneous databases without common domains using queries based on textual similarity", *Proceedings of ACM SIGMOD International Conference on Management of Data*, Jun. 2-4, 1998, Seattle:Washington, pp.201-212.

[3] Elmagarmid, A.I., Panagiotis G.V., and Vassilios, V. (2006). "Duplicate Record Detection: A Survey", *Information Systems Working Papers Series: CeDER-06-05*, Stern School of Business, New York University.

[4] Han, H., Giles, C.L., Zha, H., Li, C., and Tsioutsiouliklis, K. (2004). "Two supervised learning approaches for name disambiguation in author citations", *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*, Tuscon:AZ, Jun. 7-11, 2004, pp.296-305.

[5] Han, H., Xu, W., Zha, H., and Giles, C.L. (2005a). "A hierarchical naive Bayes mixture model for name disambiguation in author citations", *Proceedings of the 2005 ACM Symposium on Applied Computing*, Santa Fe:New Mexico, Mar. 13-17, 2005, pp.1065-1069.

[6] Han, H., Zha, H., and Giles, C.L. (2005b). "Name disambiguation in author citations using a K-way spectral clustering method", *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*, Denver:CA, Jun. 7-11, 2005, pp.334-343.

[7] Jaro, M.A. (1976). "Unimatch: a record linkage system: user's manual", *Technical report*, U.S. Bureau of the Census, Washington, D.C.

[8] Lee, D.W., On, B.W., Kang, J.W., and Park, S.H. (2005). "Effective and scalable solutions for mixed and split citation problems in digital libraries", *Proceedings of SIGMOD-2005 Workshop on Information Quality in*

Information Systems, Baltimore:Maryland, Jun. 17, 2005, pp.69-76.

[9] Levenshtein, V.I. (1965). "Binary codes capable of correcting deletions. Insertions and reversals", *Doklady Akademii Nauk SSSR*, 163(4):845-848. (in Russian), English translation: Soviet Physics Doklady, 10(8):707-710, 1966.

[10] Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959). "Automatic linkage of vital records", *Science*, 130(3381):954-959.

[11] On, B.W., Lee, D.W., Kang, J.W., and Mitra, P. (2005). "Comparative study of name disambiguation problem using a scalable blocking-based framework", *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*, Denver:CA, Jun. 7-11, 2005, pp.344-353.

[12] Rand, M. (1971). "Objective Criteria for the Evaluation of Clustering Methods", *Journal of the American Statistical Association*, 66:846-850.

[13] 이승우, 정한민, 김평, 강인수, 성원경 (2006). "서지정보의 동명이인 구별을 위한 공저자 관계의 효용성 연구", *한국컴퓨터종합학술대회 발표 논문집*, 용평:강원도, 2006년 6월 21-23일.