서지정보의 동명이인 구별을 위한 공저자 관계의 효용성 연구

이승우[0] 정한민 김평 강인수 성원경
한국과학기술정보연구원 NTIS사업단
{swlee[0], jhm, pyung, dbaisk, wksung}@kisti.re.kr

A Research on the Effectiveness of Co-authorship for Identity Resolution in Bibliography

Seungwoo Lee[0], Hanmin Jung, Pyung Kim, In-Su Kang, Won-Kyung Sung
NTIS Division, Korea Institute of Science and Technology Information

요 약

동명이인 구별 문제는 최근 시맨틱 웹과 온톨로지에 대한 관심이 높아지면서 이슈로 부각되고 있다. 논문의 서지정보를 온톨로지로 구축하기 위해서는 이름이 같은 저자들이 동명이인인지 여부를 판단하는 것이 중요하다. 일반 문서에서와는 달리, 서지정보에서는 소속과 Email를 유용한 정보로 사용할 수 있으나 그것만으로는 충분치 못하며, 이를 보완하기 위한 한 방법으로 공저자 관계를 이용하는 것이 유용함을 살펴본다.

1. 서론

이름은 고유한 개체를 구별하기 위한 것이지만, 실세계에서는 둘 이상의 서로 다른 개체가 동일한 이름을 갖는 경우를 흔히 접할 수 있다. 인명이 그 대표적인 예라고 할 수 있다. 예를 들어, 싸이월드[1] 의 회원찾기에서 '*김영희*'라는 1980년생 여자를 찾으면 무려 213명이나 나타난다. 이들은 모두 이름은 같지만 서로 다른 고유한 개체를 지칭하는 동명이인인 셈이다.

동명이인 구별 문제는 최근 시맨틱 웹과 온톨로지에 대한 관심이 높아지면서 이슈로 부각되고 있다. 시맨틱 웹에서는 모든 개체는 식별자(Unique Resource Identifier; URI)로 구별되어야 하기 때문이다. 즉, 213명의 '*김영희*'를 시맨틱 웹에서 다루기 위해서는 하나하나를 구별할 수 있어야 한다. 실세계에서는 주민등록번호와 같은 식별자로 구분할 수 있지만, 보통의 문서에서는 그러한 식별자 없이 단순히 이름만으로 언급되므로 서로 다른 두 문서에 나타나는 '*김영희*'라는 이름이 동일한 사람을 가리키는지를 판단하기 위해서는 온전히 문맥정보에 의지해야 하므로 쉽지 않은 작업이다. 이에 비해 논문의 서지정보에 나타나는 저자 이름에 대한 동명이인 구별은 상대적으로 쉬운 작업이라 할 수 있다. 논문의 서지정보는 메타정보를 가리키는 것으로, 제목을 비롯하여 저자명, 저자의 소속, Email, 출처, 연도 등을 포함한다. 여기서, 소속과 Email 등은 같은 이름의 저자를 식별하는데 필요한 양질의 문맥정보로서 잘 정리되어 제공될 수 있기 때문이다.

논문을 비롯하여 보고서, 특허 등의 연구 성과물과 인력에 대한 정보는 최근 국가과학기술정보유통 서비스를 위한 온톨로지 구축의 대상으로 인식되고 있다[1]. 온톨로지로 구축된 인력 및 연구 성과물 정보는 개별 연구자나 기관, 지역의 연구 동향 파악과 공저자관계를 이용한 연구자 네트워크 파악, 연구 분야별 전문가 추천 등의 고급 응용 서비스에 다양하게 활용될 수 있다. 연구 성과물의 서지정보는 바로 이러한 온톨로지를 구축하기 위한 첫 단계라 할 수 있다. 따라서, 서지정보의 저자 이름에 대한 동명이인 여부를 판단하는 작업은 반드시 필요하며 동시에 높은 정확도를 요구한다.

앞서 언급하였듯이, 서지정보에서의 동명이인 구별은 보통의 문서의 경우에 비해 상대적으로 쉬운 것은 사실이다. 이는 서지정보에 저자의 소속과 Email을 포함하기 때문이다. 그러나, 3장에 기술한 바와 같이, 이러한 정보가 저자의 동명이인 문제를 완전히 해결하지는 못한다. 따라서, 이를 보완하기 위한 한 방법으로 본 논문에서는 공저자 관계의 효용성을 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 동명이인 구별에 관한 기존 연구를 살펴보고, 3장에서는 서지정보에서의 동명이인 구별 방법, 특히 공저자 관계의 이용에 대해 기술한다. 4장에서는 실험을 통해 공저자 관계의 효용성을 확인하고, 5장에서 결론을 맺는다.

2. 관련 연구

한 문서 내에서의 인명에 대한 참조 해결에 대해서는 이미 많은 연구가 있었지만, 여러 문서에 나타나는 동일 이름에 대한 참조 해결에 대한 연구는 최근에 두드러지게 나타나고 있다. Mann과 Yarowsky[2]는 부트스트래핑을 통해 문서에서 자동으로 추출한 인물의 출생지, 생년월일, 직업 등의 사적인 기록들을 자질로 사용하는 비교사 방식의 클러스터링 기법을 소개하였다. 이 방법은 동명이인을 두 클러스터로만 구분할 수 있다. Fleischman과 Hovy[3]는 동명이인일 확률을 학습하는 Maximum Entropy 모델에 기반한 클러스터링 기법을 제시하였다. 최근에 Bekkerman과 McCallum, Malin은 사회망(Social Network)에서의 동명이인 구별 문제를 다루었다.

[1] http://cyworld.nate.com

Bekkerman과 McCallum[4]은 서로 관련된 인물들의 이름이 매치되는 웹 페이지에서 실제 그 인물들을 찾아내기 위해 링크 구조와 A/CDC 클러스터링 기법을 고안하였다. Malin[5]은 인터넷 영화DB의 동명이인을 구별하기 위해 이름에 기반한 관계 망의 유사성을 이용하였다.

이 연구들은 텍스트에서의 동명이인 문제를 다룬 것인데 비해, Alani외[6]는 시맨틱 웹을 위한 온톨로지의 통합 과정에서 발생할 수 있는 동명이인 문제를 다루고 있다. 서로 다른 두 온톨로지에 나타나는 두 인물이 동일인지 여부를 결정하기 위해 각 인물의 CoP (Communities of Practice)를 비교한다. CoP의 유사성이 임계치 이상이면 동일인으로 간주한다. 본 논문에서 다루는 서지정보의 공저자 관계는 CoP의 한 예라고 볼 수 있다는 점에서 유사성을 갖는다.

<표 1> 학술대회별로 서지정보 입력된 논문 수

한국정보과학회 (KISS)	2002	춘계(S)	555
		추계(F)	757
	2003	춘계(S)	774
		추계(F)	870
	2004	춘계(S)	682
대한전자공학회 (IEEK)	2003	하계(S)	665
	2004	하계(S)	419
	2005	추계(F)	290
한국HCI학회 (HCI) (년 1회 개최됨)	2003	-	241
	2004	-	312
	2005	-	326
	2006	-	369
한국정보처리학회 (KIPS)	2004	추계(F)	484
	2005	춘계(S)	431
합계	-	-	7,175

3. 서지정보에서의 동명이인 구별

논문의 서지정보는 제목, 저자명, 소속, Email, 출처, 연도 등을 포함하는 메타데이터를 말한다. 이 서지정보를 온톨로지로 구축하기 위해서는 서지정보에 포함된 각 개체를 유일하게 식별할 수 있어야 한다. 여기서 가장 문제가 되는 것이 저자명에 대한 동명이인 구별이다. 즉, 서로 다른 서지정보에 같은 이름의 저자가 나타날 때, 그것이 동일인을 가리키는지 여부를 판단하는 것이 중요하다.

본 연구에서는 국가과학기술 R&D 기반정보 온톨로지를 구축하기 위해, 그 첫 단계로서 IT분야의 국내 학술대회의 논문들을 대상으로 서지정보를 수작업으로 입력하였다. 학술대회별로 입력된 논문 수는 <<표 1>과 같다. 입력된 7,175건의 논문에서 저자이름의 출현 회수는 23,105회로, 한 논문은 평균 3.22명의 공저자를 가졌다. 이중, 동일 이름의 저자가 출현한 최대 회수는 55회로 나타났는데, 온톨로지로 구축하기 위해서는 이러한 동일이름의 출현이 실제 동일인을 가리키는지 아닌지를 판단해야 한다.

일반 문서에서와는 달리, 논문의 서지정보에서는 각 저자의 소속과 Email에 대한 정보를 담고 있어서, 비교적 구분이 쉬울 수 있다. 그러나, 실제에서는 이 정보만으로 온전히 구분하기 어려운 경우가 흔히 발생한다.

첫 번째로 논문에 소속이나 Email 정보가 기재되지 않은 경우가 그러하다. 양식이 엄격한 논문지와는 달리, 학술대회의 논문인 경우, 비록 지정된 양식에는 소속과 Email을 기재하도록 되어 있다고 하더라도 저자가 이를 따르지 않는 경우가 종종 있다. 저자가 다수인 경우, 각 저자별로 소속과 Email을 명확히 구분하기 어려운 경우도 이에 해당한다. 실제로 <<표 1>에서 입력된 서지정보에서 소속이 기재되지 않은 저자는 27회, Email이 기재되지 않은 저자는 5,562회, 소속과 Email 모두 기재되지 않은 경우도 16회나 출현하였다.

두 번째로 동일 저자의 소속과 Email 표기가 일관되지 않은 경우이다. 소속을 표기하는데 있어서, 기관만을 표기하거나 기관과 부서를 함께 표기한 경우, 부서를 '*XYZ학과*' 혹은 '*XYZ학부*', '*XYZ연구실*'과 같이 다르게 표현한 경우를 흔히 볼 수 있다. 덩치가 큰 기관에서는 한 기관 내에 동명이인이 흔히 있을 수 있다는 점을 감안할 때, 저자의 기관명만이 일치한다고 해서 동일인으로 판단하기에는 위험요소가 있다. 예를 들어, "*연세대학교 컴퓨터과학과*"의 '*조성배*'와 "*연세대학교 대학원 인지과학협동과정*"의 '*조성배*'가 동일인이라고 단정할 수는 없다. Email의 경우도 마찬가지로, 한 사람이 둘 이상의 Email주소를 사용하는 경우를 흔히 볼 수 있다. 기관에서 제공하는 Email과 부서 혹은 연구실에서 제공하는 Email, 웹 포털사이트에서 제공하는 Email 등, 복수 개의 Email을 혼용할 때, 그 Email이 가리키는 사람이 동일인임을 알기 어렵다. 예를 들어, '*sbcho@cs.yonsei.ac.kr*', '*sbcho@sclab.yonsei.ac.kr*', '*sbcho@candy.yonsei.ac.kr*', '*sbcho@csai.yonsei.ac.kr*'와 같이 다른 Email을 갖는 여러 '*조성배*'가 동일인이라고 단정하기는 어렵다.

세 번째로 시간에 따라 소속과 Email이 바뀌는 경우이다. 저자가 소속을 옮기는 경우는 흔히 발생하며, 이 때, Email도 함께 바뀐다. 기관 내에서 부서를 옮기거나 부서명 자체가 변경되는 경우도 이에 해당한다. 이 경우, 논문의 발행연도를 참조하는 것이 도움이 될 수 있다.

이와 같이, 소속과 Email은 동명이인을 구별하는데 필요한 중요한 정보를 제공하지만, 그것만으로는 부족한 경우가 흔히 발생한다. 이런 경우에, 추가적으로 이용할 수 있는 정보가 바로 공저자 관계이다.

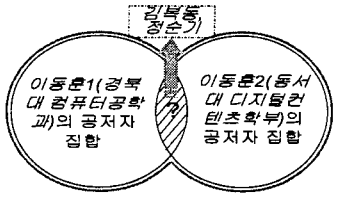
3.1 공저자 관계 이용

소속이나 Email과 마찬가지로, 공저자 관계는 동명이인을 구별하는데 중요한 정보로 활용될 수 있다. 논문의 공동 저자들은 서로간에 학술적으로 밀접한 관계를 맺고 있으며, 공저자로 들어간 논문의 수가 많을수록 그 관계의 정도는 더 밀접하다고 할 수 있다. 따라서, 두 논문에 대해, 각각 두 명씩의 저자의 이름이 서로 같을 때 그들이 동일인일 가능성은, 한 명씩의 저자 이름이 서로 같을 때 그들이 동일인일 가능성에 비해 훨씬 높다. 또한 공저자 관계가 더 많이 공유될수록 동일인일 가능성은 더욱 높아진다.

공저자 사이의 학술적 관계는 어느 정도 지속되는 것이 보통이다. 심지어, 저자가 소속을 바꾼 경우 – 예를 들면, 학위 과정을 마치고 같은 분야에서 학술활동을 계속하는 경우 – 에도 학술적 관계를 계속 유지하는 것을 흔히 볼 수 있다. 이 말은 저자의 소속과 Email이 같지 않더라도 공저자 관계를 통해 그 저자가 동일인임을 알아내는 것이 가능하다는 것을 뜻한다.

입력된 서지정보에서 보면, 한 논문은 최대 17명, 평균

3.22명의 저자를 가지는 것으로 나타났는데, 이는 평균적으로 두 명을 공저자 관계에서 비교해 볼 수 있음을 말한다. 이미 동일인으로 판명되었으면 각각의 공저자 집합을 하나로 합침으로써 비교할 수 있는 공저자 집합을 늘려 나갈 수 있다. ((그림 1)은 이 과정의 한 예를 그림으로 설명하고 있다.



(그림 1) '*이동훈1*'의 공저자와 '*이동훈2*'의 공저자가 공유될 때 둘은 동일인으로 볼 수 있으며 공저자 집합을 늘려 나간다.

4. 실험

동명이인 구별을 위한 공저자 관계의 효용성을 확인하기 위해 <<표 1>의 서지정보 중에서 4회 이상 출현하는 저자이름 11,322개에 대한 서지정보를 사용하였다. 여기에는 1,522개의 서로 다른 이름이 들어있다.

같은 이름의 저자가 동일인인지에 대한 기준정답(ground truth)을 만들기 위해 우선 소속과 Email이 모두 일치[2]하거나 공저자를 공유하는 경우를 클러스터링하였다. 이를 기초로 하여 국가과학기술인력 종합정보시스템[3]에서 저자의 이름을 검색하여 여기에 등록된 정보를 바탕으로 동명이인을 구분하였다. 이렇게 하여 4,670개의 이름에 대해 확인할 수 있었다. 나머지에 대해서는 소속과 Email, 공저자 관계를 통해 명백히 확인할 수 있는 경우에만 동일인으로 간주하였다. 이렇게 동일인 간주를 엄격히 함으로써 동일인을 서로 다른 사람으로 잘못 구분하는 경우는 있지만 동명이인을 동일인으로 잘못 판단하는 경우는 없도록 하였다. 이 기준정답은 대상 서지정보에 대해서 동명이인을 구분하는 정확한 성능을 평가하는 목적에는 부적합하지만, 공저자 관계를 이용한 동명이인 구분의 효용성만을 판단하는 상대적인 비교 자료로 사용하기에는 충분하다. 즉, 공저자 관계는 같은 이름의 저자들을 동일인으로 그룹을 지어주는 역할을 하는 것이지, 동일인이 아님을 판단하는 기준은 아니기 때문이다.

$$P = \frac{\#agreement}{\#agreement + \#disagreement} \quad (1)$$

성능 평가 수단으로 *Rand Index*[7]를 사용하였다. *Rand Index*는 클러스터링된 각 객체의 모든 쌍에 대해, 기준정답과 비교하여 일치하는 수(*#agreement*)와 불일치하는 수(*#disagreement*)로부터 계산된다. 각 객체의 쌍이 기준정답과 시스템의 결과 둘 다에서 같은 클러스터에 있거나 둘 다에서 다른 클러스터에 있으면 일치로 계산하고 그 외의 경우에는 불일치로 계산한다. 그리하여 성능(*P*)는 식(1)과 같이 전체에 대한 일치의 비율로써 계산되며 클러스터링의 정확도를 측정한다.

<<표 2>는 공저자 관계 이용 여부에 따른 성능비교를 보여준다. 이 표에 따르면, 공저자를 추가로 이용한 경우 동일인으로 그룹짓지 못한 불일치 쌍의 수(*#under-clustering*)가 소속과 Email만을 사용한 경우에 비해 13%로 월등히 줄어듦을 알 수 있다. 반면 동일인으로 그룹지은 불일치 쌍의 수(*#over-clustering*)는 2.18배 늘어나는데 그쳤다. 이 때의 *#over-clustering*은 기준 정답에서는 동일인임이 명백하지 않아 그룹짓지 않은 것을 공저자 관계를 통해 그룹지은 경우가 대부분으로 이것이 오류라고 단정할 수는 없다. 특히, ((그림 1)에 보여진 바와 같이 공저자 관계를 이용함으로써 소속과 Email이 전혀 일치하지 않는 '*이동훈1*'과 '*이동훈2*'를 동일인으로 그룹지을 수 있었다.

<표 2> 공저자 관계 이용 여부에 따른 성능 비교

	Base[4]	소속+Email (1)	(1)+공저자	공저자 only
#agreement	40,021	37,369	52,093	40,568
#disagreement	14,849	17,501	2,777	14,302
P	**0.73**	**0.68**	**0.95**	**0.74**
#over-clustering	14,849	234	511	160
#under-clustering	0	17,267	2,266	14,142

5. 결론

논문의 서지정보에 나타나는 저자의 동명이인 구분을 위해서는 소속과 Email만으로는 충분치 못하며, 이를 보완하기 위한 한 방법으로 공저자 관계를 이용하는 것이 유용함을 살펴보았다. 그 외에도 분야분류체계에서의 논문의 분야를 서로 비교하거나 시소러스를 통한 제목 키워드의 매칭 유사도를 이용하는 것도 고려해 볼 수 있다. 이는 동일 저자는 비슷하거나 관련된 분야의 논문을 작성할 것이라는 가정에 기반한다.

참고문헌

[1] 이미경, 정한민, 성원경, "*지식 기반 정보 유통 플랫폼 개발*", 제10회 한국과학기술정보인프라 워크숍, 2005.

[2] G.S. Mann and D. Yarowsky, "*Unsupervised Personal Name Disambiguation*", In Proceedings of the 7th Conference on Computational Natural Language Learning, 2003, Canada.

[3] M. B. Fleischman and E. Hovy, "*Multi-Document Person Name Resolution*", In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Reference Resolution Workshop, 2004, Barcelona, Spain.

[4] R. Bekkerman, A. McCallum, "Disambiguating Web Appearances of People in a Social Network", WWW2005, pp. 463-470, 2005, Japan.

[5] B. Malin, "*Unsupervised Name Disambiguation via Social Network Similarity*", In Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security, in conjunction with the SIAM International Conference on Data Mining, pp. 93-102, 2005, Newport Beach, CA.

[6] H. Alani, S. Dasmahapatra, N. Gibbins, H. Glaser, S. Harris, Y. Kalfoglou, K. O'Hara and N. Shadbolt, "*Managing Reference: Ensuring Referential Integrity of Ontologies for the Semantic Web*", In Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02), pp. 317-334, 2002, Sigenza, Spain.

[7] M. Rand, "*Objective Criteria for the Evaluation of Clustering Methods*", Journal of the American Statistical Association, Vol. 66, pp. 846-850, 1971.

[2] 두 소속이 서로 포함관계에 있을 때 – 예를 들어, "*연세대학교 컴퓨터과학과*"와 "*연세대학교 컴퓨터과학과 생체인식연구센터*" – 소속이 일치하는 것으로 간주하였다. 시스템의 결과에서 소속 명칭의 정규화는 현재 처리되지 않았다.

[3] http://www.hrst.or.kr/

[4] 같은 이름의 저자를 모두 동일인으로 간주한 경우. '소속+Email'이 Base보다도 낮은 것은 소속 명칭의 정규화가 처리되지 않은 영향으로 보인다.