시맨틱 웹 온톨로지에서의 OWL sameAs 적용
(Applying OWL SameAs to an Ontology in the Semantic Web)

강 인 수 † 정 한 민 † 이 승 우 †
(Insu Kang) (Hanmin Jung) (Seungwoo Lee)
김 평 † 이 미 경 † 성 원 경 ††
(Pyung Kim) (Mikyung Lee) (Wonkyung Sung)

요 약 차세대 시맨틱 기술 적용의 비전인 시맨틱 웹의 성공을 위해서는 그 기반 지식이 되는 온톨로지의 생산, 공유 및 연계가 전제되어야 함과 동시에 온톨로지를 구성하는 개체들의 정합성이 보장되어야 한다. 온톨로지 개체 정합성 측면에서, 두 개의 서로 다른 온톨로지 개체가 실세계에서 동일한 개체임을 표현하기 위한 방편으로 OWL에서는 sameAs가 사용될 수 있다. 그러나, 최근까지의 시맨틱 웹 응용 연구에 있어서 sameAs 적용과 관련된 문제점과 고려사항들이 충분히 고찰되지 못했다. 이 연구는 과학기술 연구분야 온톨로지에서의 sameAs 적용 사례를 소개함으로써 sameAs와 관련된 이해의 폭을 공유하고자 한다.

키워드 : 시맨틱 웹, 온톨로지 개체, sameAs

Abstract The ontology is the underlying knowledge base to create the semantic web. Prerequisites for the success of the semantic web include widespread uses/sharing/merging of ontologies. In addition, it is very crucial to secure the integrity of ontology instances such as instance-identifying/referring integrity, and attribute domain constraints. In terms of ensuring instance-identifying integrity, OWL provides owl:sameAs property which is used to connect two separate ontology instances in order to represent that the two instances are the same. Recent semantic web works, however, have not sufficiently investigated the issues one may face in applying owl:sameAs to real semantic web applications. This study introduces our experiences of sameAs in developing a semantic web service framework for a research domain.

Key words : Semantic web, ontology individuals, sameAs

1. 서 론

인간 중심적이고 구문적 표현을 사용하는 현재의 웹을 개선하기 위해 제안된 시맨틱 웹에서는 온톨로지에 기반하여 기계 중심적이며 의미적 표현 방식을 추구한다[1]. 온톨로지는, 특정 분야에 있어서 관련된 다수가 공유하는 개념 체계를 형식화해 놓은 것으로[2], 시맨틱 웹의 구현을 위한 핵심적인 지식베이스이며, 온톨로지에 정의된 개념과 개념 간 관계들은 지식화 대상이 되는 실세계 개체와 개체 간 관계의 태깅에 사용된다. 온톨로지에 기반하여 단위 지식과 지식 간의 관계에 의미적 태그를 부착함으로써, 시맨틱 웹에서는 지식 추출, 지식 통합 및 내재된 지식의 추론 등이 기계적 자동 처리로 가능케 된다.

시맨틱 웹의 성공을 위해서는 그 기반 지식이 되는 온톨로지의 왕성한 생산 및 공유가 전제되어야 함과 동시에 온톨로지 내 개체들의 정합성이 보장되어야 한다. 온톨로지 개체 정합성이란, 온톨로지 개체들이 모델링 대상이 되는 실세계의 부분을 투명하게 반영하고 있는 정도를 의미하며, 개체 식별 정합성, 개체 참조 정합성, 개체 속성값의 도메인 정합성으로 세분될 수 있다. 개체 식별 정합성은 온톨로지 내에 존재하는 개체들의 집합이 모델링 대상이 되는 실세계의 개체 집

† 정 회 원 : 한국과학기술정보연구원 정보시스템연구팀
dbaisk@kisti.re.kr
jhm@kisti.re.kr
swlee@kisti.re.kr
pyung@kisti.re.kr
jerryis@kisti.re.kr
†† 비 회 원 : 한국과학기술정보연구원 정보시스템연구팀
wksung@kisti.re.kr
논문접수 : 2006년 12월 4일
심사완료 : 2007년 2월 9일

합과 일치하는 정도를 의미하며, 이의 확보를 위해 대표적으로 실세계 단일 개체의 중복 표현이나 실세계 복수 개체들의 병합 표현 등을 통제할 수 있어야 한다. 개체 참조 정합성은 온톨로지 내 개체 간에 설정된 관계(예: '사람' 개체와 '기관' 개체 간에는 '소속되다'의 관계가 존재할 수 있다)가 올바른 정도를 의미한다. 이는 관계형 데이타베이스에서의 참조 무결성처럼 존재하지 않는 기관이나 사람 간에 소속 관계가 설정되지 않도록 제어할 수 있어야 함을 가리킨다. 개체 속성값의 도메인 정합성은, 사람의 나이가 양의 정수의 값을 가져야 한다거나 연도 속성값(예: 개인의 출생년도, 논문의 출판년도 등)이 부적절한 미래 연도를 갖지 않도록 하는 것과 같은, 해당 도메인 내에서 개체의 부적절한 속성값의 리스트나 범위를 통제할 수 있어야 함을 의미한다. 위의 세 가지 온톨로지 정합성 중에서, 이 논문에서는 가장 우선적으로 다루어져야 할 개체 식별 정합성에 집중할 것이다.

온톨로지 개체 식별 정합성 측면에서, W3C[1]의 OWL에서는[3], 두 개의 서로 다른 온톨로지 개체가 실세계에서 동일한 개체임을 표현하기 위한 방편으로, owl:sameAs를 사용할 수 있다. 그러나, sameAs를 실제 시맨틱 웹 응용에 적용함에 있어서는 여러 가지 고려해야 할 이슈들이 발생한다. 예를 들면, 개체 집단의 같음과 다름에 대한 기본 전제로, OWL에서는 "sameAs나 differentFrom 관계가 명시되지 않은 개체들이 존재한다"고 가정하고 있으나, 이는 같음과 다름을 판명할 수 없는 개체들을 추론 과정에서 어떻게 처리해야 하는가에 대한 의문에 직면하게 만든다. 또한, sameAs 해제 기능 지원의 필요성, sameAs 관계를 적용할 두 개체가 동일 클래스에 속한다는 제약을 가할 것이냐의 여부, sameAs 관계들로 연결된 sameAs 그룹에 대한 해석 및 sameAs 대표 개체의 표현, 그리고 sameAs 그룹 단위의 지식 처리 방안 등의 이슈에 대한 고려가 필요하다.

그러나, 최근까지의 시맨틱 웹 응용 연구에 있어서 전술한 sameAs 적용과 관련된 고려사항들이 충분히 고찰되지 못했다. 이 연구에서는 현재 한국과학기술정보연구원에서 과학기술 연구분야 온톨로지 및 그 응용 시스템을 개발하는 과정에서 개체 간 정합성 보장을 위해 sameAs를 적용한 경험을 공유함으로써, 시맨틱 웹 관련 연구자 및 개발자들 사이에 sameAs와 관련된 이해의 폭을 넓히고자 한다.

2. sameAs 개요 및 관련 연구

2.1 OWL sameAs 개요

W3C(World Wide Web Consortium)에서 정의한 온톨로지 기술을 위한 표준 언어인 OWL(Web Ontology Language)에서는 온톨로지 개체(Individual) 간 동일성(identity)을 표현하는 수단으로 owl:sameAs라는 property를 제공하고 있다[3]. "온톨로지(들) 내의 두 개체 A와 B가 동일한 실세계 개체를 의미한다"라는 하나의 sameAs 문장은 RDF(Resource Description Framework) 트리플 형태로 다음과 같이 표현될 수 있다.

A owl:sameAs B

위에서 A와 B는 동일한 온톨로지 내의 서로 다른 두 개체이거나 서로 다른 온톨로지 내의 두 개체일 수 있다. 각각 서로 다른 온톨로지에 속한 두 개체 간에 sameAs 관계를 설정하는 것은, 서로 다른 두 온톨로지의 개체들을 상호 연계하여 온톨로지를 통합하는 용도로 활용될 수 있다. 동일한 하나의 온톨로지 내의 서로 다른 두 개체 간에 sameAs 관계를 맺는 것은, 실세계 개체에 대한 신원이 그 신원 확인 시점마다 다를 수 있다는 전제하에 이루어지는 것으로, 온톨로지 개체 집단의 실세계 정합성을 보장하기 위해 사용된다.

동일 온톨로지 내에서 sameAs 관계를 설정할 필요가 있는 대표적인 개체 타입의 예는 "사람"이다. 만약, 온톨로지에 개체로 등록할 사람에 대한 속성 값으로 주민등록번호를 사용하기로 하고 등록할 사람에 대해 항상 그의 주민등록번호가 준비되어 있다면 사람에 대해 sameAs를 적용할 필요는 없을 것이다. 그러나, "논문의 저자"처럼 고유한 식별자가 준비되어 있지 않은 사람을 그것의 클래스 중 하나로 모델링해야 하는 온톨로지의 수도 적지 않을 것이다. 고유한 식별자가 미리 준비되어 있지 않은 개체들은 그들 간의 동일성에 대한 판단을 개체가 갖는 속성값에 의존할 수 밖에 없는데 이 경우 속성값만으로 개체 간의 같음과 다름을 판단하기 어려운 때가 많다. 예를 들어, "논문의 저자"라는 개체는 서지정보(bibliographic data)와 논문 원문으로부터 "이름", "전자메일주소", "소속기관" 등의 개체 속성값을 추출할 수 있으나, "이름"은 고유한 속성값이 아니고 "전자메일주소"는 원문에 항상 표현되는 것이 아니며 동일 기관 내의 동명이인의 출현도 배제할 수 없으므로, 실세계의 동일 저자가 온톨로지 내에서 서로 다른 동명이인으로 표현되는 것을 통제할 수단이 부족하게 된다. 따라서, (사람의 주민등록번호와 같은 식별자가 아닌) 비식별자 속성만을 갖는 클래스를 모델링하고 있는 온톨로지에서는 개체 간 신원 확인 정보의 부족으로 인해 온톨로지 개체의 무결성을 보장하는 측면에서 sameAs의 사용을 피할 수 없는 것이다.

1) http://www.w3.org/

2.2 sameAs 관련 연구

시맨틱 웹 환경에서 서로 다른 온톨로지를 사용하는 에이전트나 서비스들 사이의 상호 작용을 가능케 하기 위한 선결 요건 중 하나인 온톨로지 간 매핑 및 정렬 연구들에서, sameAs는 개체 간 명시적 동일성을 판단하는 자질로 사용되고 있다[4,5]. owl:sameAs 외에 개체 동일성 판단을 위해 사용될 수 있는 OWL 장치로는 owl:FunctionalProperty가 있다. 이는 다음과 같은 규칙을 통해, 한 개체 a가 또 다른 두 개체 b, c와 owl:FunctionalProperty에 속하는 한 property p1을 통해 연결되어 있다면 두 개체 b와 c는 동일 개체라는 사실을 유추해 내는 것을 가능케 한다[6,7].

- (p1 rdf:type owl:FunctionalProperty) (a p1 b) (a p1 c) -> (b owl:sameAs c)
- (p1 rdf:type owl:FunctionalProperty) (a1 p1 b) (a2 p1 c) (a1 owl:sameAs a2) -> (b owl:sameAs c)

sameAs 관계가 명시되지 않은 개체 집단 내에서 개체 간 sameAs 관계의 후보를 제시할 수 있는 자동화된 기법에 대한 연구들은[8-11], 주로 논문의 서지정보나 인용정보에 출현하는 저자 개체들의 동명이인 해소 문제를 다루고 있다. 즉, 논문 서지정보나 인용정보에 나타난 동일 이름(예: "홍길동") 저자들에 대해 그들의 실세계에서의 동일성 여부를 판단하는 것이다. 이들 연구들은 동일 저자명의 서로 다른 저자들을 클러스터링하는 데 적용되며, 이후 동일 클러스터 내의 저자 개체들은 모두 sameAs 관계를 설정함으로써 온톨로지 인스턴스의 개체 식별 정합성을 높이는 데 활용될 수 있다.

아쉽게도 최근까지의 시맨틱 웹 연구에서 전술한 내용을 제외하고는 sameAs와 관련된 연구는 거의 찾을 수 없다. 그 이유는, sameAs의 주용도가 서로 다른 온톨로지 간 매핑이나 병합이라는 점과 현재 시맨틱 웹 커뮤니티 내에서 온톨로지의 생산 및 공유 정도가 그리 크지 않은 점을 감안할 때 sameAs 적용 연구는 아직 시기상조이기 때문인지도 모른다.

3. sameAs 적용 시 고려 사항

온톨로지에 sameAs 관계를 적용함에 있어서 개체 집단을 바라보는 일반적인 기본 전제로 "sameAs 관계가 명시되지 않은 개체들은 모두 서로 다르다"를 고려해 볼 수 있다.[2] 그러나, 이것은 W3C의 OWL 권고문서[3]에 기술된 내용으로부터 유추될 수 있는 바와 같지 않다. W3C에서는 "sameAs나 differentFrom 관계가 명시되지 않은 개체들이 존재한다"고 가정한다. 이것은 현실 세계에 존재하는 개체들에 대한 사람들의 인식을 그대로 표현하는 가정이며, 바로 공용 시맨틱 웹(public or open-world semantic web)이 추구하는 바이기도 하다. 다시 말하면, 한 사람 혹은 집단이 같음과 다름에 대한 판단을 내릴 수 있는 개체들은 현실 세계에 존재하는 전체 개체의 크기를 감안할 때 무시할 수 있을 정도로 적은 부분일 것이므로, 공용 시맨틱 웹 측면에서 W3C의 가정은 적절하다. 하지만, 같지도 않고 다르지도 않은 개체들이 존재하는 온톨로지에 대해 질의를 가한다고 할 때, 그것이 가능한 것인지에 대한 의문이 제기될 뿐 아니라, 가능하다 하더라도 그 결과를 해석하는 몫은 고스란히 사용자에게 떠넘겨 질 것이다. 따라서, 이처럼 현재로서는 비전으로의 역할만을 갖는 공용 시맨틱 웹의 반대편에는 현실적으로 구현 가능한 전용 시맨틱 웹(corporate or closed-word semantic web)의 세계가 존재하며, 전용 시맨틱 웹의 관점에서 개체 집단에 대한 바람직한 가정은 앞서 언급한 것처럼 "sameAs 관계가 명시되지 않은 개체들은 모두 서로 다르다"일 것이다.

sameAs를 적용함에 있어 제기되는 두 번째 논의는, sameAs 해제 기능 지원의 필요성이다. 앞서 언급한 것처럼 비식별자 속성만을 갖는 클래스의 개체 집단에 sameAs를 적용한다는 것은 개체 간 동일성의 판단을 기계적으로 수행하기 어려울 수 있다는 것을 의미한다. 기계적 sameAs 판단이 불가하다는 것은, 온톨로지 개체 관리자가 sameAs 관계를 맺어야 할 것이므로, 수작업으로 인한 sameAs 설정의 오류가 발생할 수 있으며 그러한 오류의 수정을 위해 기존에 잘못 설정된 sameAs 관계를 해제할 필요가 발생할 수 있다는 것을 의미한다. 클래스에 대한 식별자 속성(예: "직원" 클래스에 대한 식별자 속성 "직번")이 존재할 경우 그 동일 클래스에 속하는 개체 간에는 sameAs 설정 자체가 필요 없을 것이고, 식별자 속성을 갖는 서로 다른 클래스에 속하는 두 개체 간에 sameAs를 맺는 경우에도 식별자는 이미 두 개체의 신원을 판단하는 역할을 상실한 것이 되므로 비식별자 속성만으로 개체의 동일성을 판단하는 경우와 같아진다. 따라서, sameAs를 설정해야 하는 경우라면, 설정 오류의 가능성을 동반하므로, sameAs 해제 기능이 필수적으로 요구된다.

sameAs 적용과 관련된 세 번째 논의는 sameAs 관계를 적용할 두 개체가 동일 클래스에 속한다는 제약을 가할 것이냐의 여부이다. OWL의 sameAs 명세에서는 sameAs 대상이 되는 두 개체에 대해 그들이 소속된 클래스에 대한 어떠한 제약도 찾아볼 수 없다. 즉, sameAs 관계는 클래스의 일치 여부에 상관 없이, 임의의 두 클래스의 임의의 두 개체 간에 맺을 수 있다는

2) 혹은, sameAs 대신 differentFrom 관계를 사용하기로 하고, 개체 집단에 대해 "differentFrom 관계가 명시되지 않은 개체들은 모두 서로 같다"라고 가정할 수도 있겠으나 이것은 현실적이지 못한 부분이 많다.

애기이다. sameAs 설정에 있어서 클래스 제약의 결여는, 공용 시맨틱 웹 환경에서 빈번히 발생될 것으로 예상되는 온톨로지 간 병합이나 링킹을 광범위하게 지원하기 위한 적절한 수단이 된다. 온톨로지 병합이나 링킹이 아니더라도, 단일 온톨로지 내에서 서로 다른 클래스 소속의 개체 간 sameAs 설정의 필요성이 발생할 수 있다. 예를 들어, "고객" 클래스와 "직원" 클래스를 포함하고 있는 한 온톨로지에서, 직원이면서 동시에 고객인 사람을 표현하기 위해, "고객" 클래스의 개체(http://abc.org/def#CUS_010369)와 "직원" 클래스의 개체 (http://abc.org/def#EMP_000123)를 다음과 같이 sameAs 관계로 연결하는 경우가 발생할 수 있다.

http://abc.org/def#CUS_010369 owl:sameAs http://abc.org/def#EMP_000123

그러나, 서로 다른 클래스 소속의 개체 간 sameAs 연결에 대한 제약을 가하지 않음으로 인해, 작업자의 오류나 부주의로 인한 부적절한 개체 간 sameAs 설정이 발생될 수도 있다. 이는 전술한 sameAs 해제 기능의 필요성과도 관련되지만, 혹자는 sameAs 관계 설정을 동일 클래스에 속하는 두 개체로만 제한하는 방식을 취할 수도 있다. 예를 들어, 어떤 온톨로지 내에 정의된 서로 다른 클래스들 간에 sameAs 설정의 허용이 불가하도록 스키마가 정의된 온톨로지의 경우, sameAs 설정을 동일 클래스 소속의 개체들로만 제약하는 것이 부주의한 sameAs 설정으로 인해 발생될 온톨로지 개체 집단 정합성의 결함을 최소화할 수 있는 한 방편이 될 수 있다.

sameAs 적용에 있어서 네 번째로 고려할 것은, sameAs 관계들로 연결된 sameAs 그룹에 대한 해석 및 대표 개체의 표현과 관련이 있다. sameAs 관계는 기본적으로 대칭성[3](symmetry)과 이행성[4](transitivity)을 갖는다. 이는 두 개 이상의 sameAs 관계로 연결된 세 개 이상의 개체들이 모두 상호간 sameAs 관계가 성립된다는 것을 의미한다. 예를 들어, A1-sameAs-A2, A2-sameAs-A3, A3-sameAs-A4의 네 개의 sameAs 관계들에 sameAs의 이행성을 적용하면 추가적으로 A1-sameAs-A3, A1-sameAs-A4, A2-sameAs-A4를 얻을 수 있다. 이처럼 sameAs 관계들로 연결된 그래프(sameAs 관계를 링크로, sameAs 대상 각 개체를 노드로 고려)에서 노드에 해당하는 개체들의 집합을 sameAs 그룹으로 정의할 때, sameAs 그룹을 사용자에게 제시하는 측면에서 sameAs 그룹의 대표 개체 정의의 필요성이 제기될 수 있다. 예를 들어, "사람"에 대한 다양한 클래스들(동일 온톨로지 내에 존재할 필요는 없음)이 존재하고 클래스별로 고유의 URI식별체계가 운용되고 있을 때, sameAs 그룹으로 통합된 "사람" 개체들에 대한 정보 제시 측면에서 대표가 되는 URI 식별체계가 선정될 수 있을 것이다. 즉, A기관의 소속 인력에 대한 "A기관직번체계"와 특허청에서 출원인에게 부여하는 "출원인코드"가 별도로 존재하고, A사와 특허청 온톨로지 내의 "사람" 개체들이 sameAs 관계로 연결되어 있다고 할 때, A기관에서는 sameAs 그룹의 대표 개체로 A기관의 ("A기관식별체계"를 갖는) 소속 인력이 전면에 출력되기를 원할 것이다.

sameAs 그룹의 대표 개체 지정의 필요성에 동의한다면, sameAs 그룹의 대표를 효율적으로 탐색하기 위한 sameAs 그룹의 표현 방식을 논할 필요가 있다. 그림 1은 sameAs 그룹에서의 대표 개체 표현의 한 예를 보이고 있다. 그림에서 좌측 다섯 개의 sameAs 관계 설정 트리플들로부터 sameAs 그룹에 해당하는 그래프(그림 1의 (a))를 얻을 수 있으며, 여기서 sameAs 관계 설정에 있어서의 대표 개체는 sameAs 트리플의 SUBJECT에 해당하는 개체로 가정하였다. 예를 들면, D sameAs F에서의 대표 개체는 SUBJECT에 해당하는 D이다.

그러나, 그림 1의 (a)와 같은 sameAs 그룹의 표현에 있어서는, sameAs 그룹 내의 임의의 한 개체에 대한 대표 개체를 탐색하는 데, 노드의 수를 n으로 가정했을 때, 최대 O(n)의 시간이 소요된다. 이 대표 개체 탐색 시간을 상수배의 시간으로 줄이기 위해서는, 그림 1의 (b)의 점선 링크(isrl:standForSameAsGroupOf)와 같이, sameAs 그룹 내의 다른 모든 개체들이 sameAs 그룹의 대표 개체를 가리키도록 하면 된다. 그림 1의 (b)에서 owl:sameAs 관계에 해당하는 실선 링크를 제거하는 것은 그림 1의 (a)의 원 sameAs 관계 설정의 구조를 상실하게 되는 위험을 초래할 수 있다. 그림 1의 (b)에서 원 sameAs 관계 설정의 구조를 유지해야 하는 이유는, 부적절하게 설정된 sameAs 관계 설정을 해제했을 때 이전 sameAs 관계 설정의 구조를 복원할 수 있게 하기 위함이다. 예를 들어, 그림 1에서 마지막 sameAs 관계 설정에 해당하는 A sameAs C가 오류로 판명되어 sameAs 관계 해제를 수행한다고 할 때, 그림 1의 (a)에서는 노드 C에서 노드 A로의 링크만 삭제하면 되는 것이므로 A sameAs C 설정 이전의 상태인 두 개의 sameAs 그룹 {A, B}, {C, D, E, F}를 복원해 낼 수 있는 것이다.

마지막으로, sameAs 그룹 단위의 지식 처리 방안을 논할 필요가 있다. 좀 더 풀어 설명하면, sameAs 그룹의 지식을 특정 sameAs 그룹에 소속된 개별 개체들이 갖는 속성(데이타타입속성 및 객체관계속성) 정보들이

3) A-sameAs-B는 B-sameAs-A를 내포한다
4) A-sameAs-B와 B-sameAs-C는 A-sameAs-C를 내포한다.

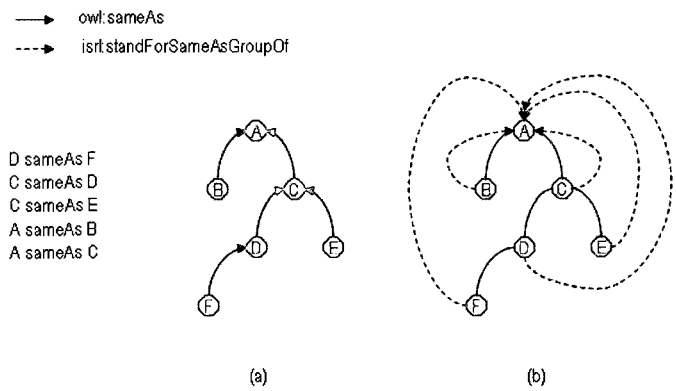

그림 1 sameAs 그룹의 대표 개체의 표현

병합된 것이라고 정의할 때, 병합에 소요될 계산의 복잡성을 어떻게 다룰 것이냐의 문제이다. 이에 대해서는, 4장에서 실제 예를 통해 설명할 것이다.

서론에서 언급한 것처럼, 이 장에서 기술한 sameAs 관련 고려 사항들은 개체 식별 정합성 측면에서 owl:sameAs 적용과 관련된 것들이다. sameAs 적용과 관련하여 개체 참조 정합성 측면에서도 여러 고려사항들이 발생될 수 있다. 예를 들어, 한 온톨로지 내의 두 개체 A2와 B3 간에 R1의 관계가 존재하는 상황에서, owl:sameAs 관계 설정들을 통해 신규로 두 sameAs 그룹 {A1, A2, A3}, {B1, B2, B3, B4}이 생성된 경우, 이제 R1 관계는 앞의 두 그룹의 임의의 두 개체 간에 존재하게 되는 것이므로, 그러한 R1 관계(들)을 두 sameAs 그룹 간에 어떤 방식으로 맺고 끊을 것인지를 고려할 필요가 있을 것이다. 개체 속성값의 도메인 정합성 측면에서도 owl:sameAs로 인해 곤란한 상황이 발생될 수 있다. 예를 들어, sameAs를 맺을 두 사람 개체 A와 B의 현 재직기관 속성의 값들이 속성값 입력 시점의 차이로 인해 서로 다를 수 있는데, 이 경우 두 재직기관 속성값들을 구별하는 방식을 취하여 별도로 유지해야 할 지 하나의 통일된 속성값을 선별하여 표현해야 할 지의 선택을 해야 할 것이다. 이처럼 이 논문에서 다루는 개체 식별 정합성 측면에서의 온톨로지 개체 간 sameAs 적용을, 개체 참조 정합성 및 개체 속성값 도메인 정합성들을 동시에 고려하는 상황으로 확장할 경우, 심도 깊게 고려되어야 할 여러 복잡한 이슈들이 추가로 발생될 것으로 예상된다.

4. 연구분야 온톨로지에서의 sameAs 적용

4.1 연구분야 온톨로지 소개

이 절에서는 연구분야 온톨로지의 하나로 한국과학기술정보연구원에서 개발한 "국가 과학기술 연구개발 기반정보 온톨로지"(이하 "기반정보 온톨로지")를 소개한다. 이 절에서 온톨로지 스키마를 구성하는 클래스는 **볼드체**로 표시하였다. 기반정보 온톨로지는, 연구의 주체들과 연구주체가 창출해 낸 연구성과물을 중심으로 설계되었다. 연구주체로는 **연구자**와 **연구기관**을, 연구성과물로는 **논문, 연구보고서, 지적재산권**을 각각 포함하고 있다. 또한, 기반정보 온톨로지는 연구주체와 연구성과물을 바라보는 다양한 관점들을 모델링하고 있는데, 현재로는 지역과 토픽의 개념을 포함하고 있다. 지역은, 연구성과물이 창출된 지역 정보를 제시하기 위해 활용된다. 토픽은, 연구성과물의 (원문)내용을 압축 표현할 수 있는 주제/분야 분류를 제시할 뿐 아니라, 연구주체 및 지역의 주요 연구 주제/분야를 표현하기 위해 도입되었다.

개별 연구성과물의 토픽 정보는 해당 연구성과물의 원문으로부터 텍스트추출/형태소분석/용어추출/토픽할당 과정을 거쳐 자동 획득되며, 이렇게 얻어진 연구성과물의 토픽 정보는 해당 연구성과물의 저자들과 저자의 소속 기관(현재 소속 기관이 아닌 연구성과물 창작 당시의 소속 기관)의 토픽 정보로 수집되고, 이후 기관의 토픽 정보는 기관이 위치한 지역의 토픽 정보로 모아지게 된다. 지역과 토픽 외에도, 기반정보 온톨로지는, 연구성과물과 관련하여 논문의 **게재지**(학술대회발표논문집과 학술지 등)와 연구보고서의 모체가 되는 **과제**를 포함하고 있다.

기반정보 온톨로지는 현재, 21개 클래스, 64개 속성관계(46개의 데이타타입속성, 18개의 객체관계속성)로 구성되어 있으며, 프로티지 온톨로지 편집 도구(Protégé 3.1.1)를 사용하여 W3C의 OWL DL(Description Logic) 기반으로 작성되어 있다.

현재 기반정보 온톨로지의 개체 데이타는, 2002년부터 2006년 전반기까지 국내 IT분야의 주요 학술대회[5] 발

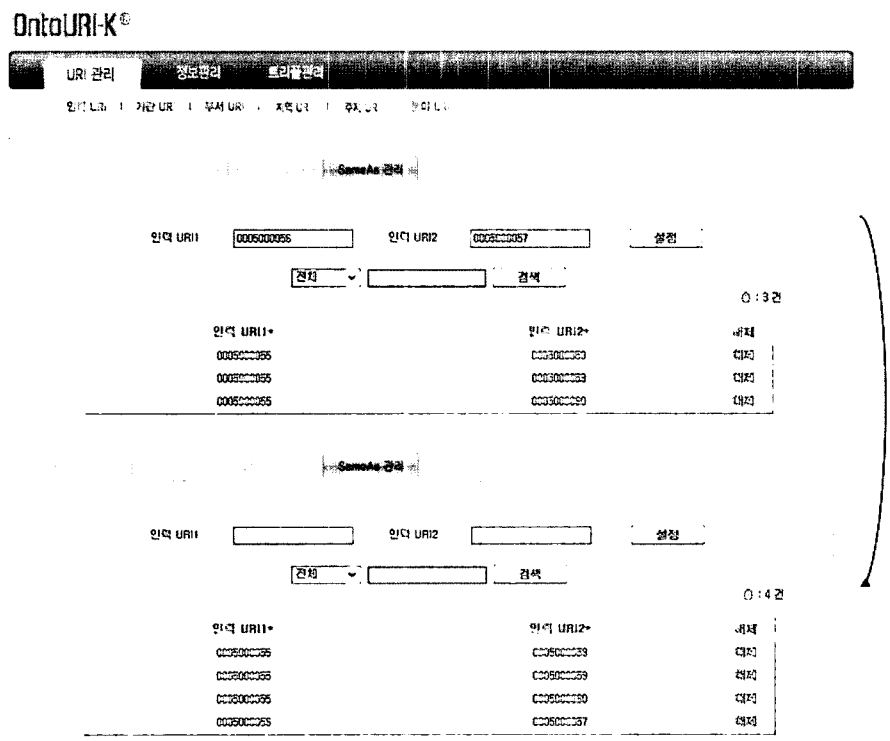

그림 2 sameAs 관계 관리기

표 논문들과 1999년부터 2006년 전반기까지의 한국과학기술정보연구원 내부 인력의 연구성과 데이타(논문, 연구고서, 특허, 과제 정보 등)를 대상으로 획득되었으며, 온톨로지 개체 지식의 크기를 가늠하는 단위인 RDF 트리플 기준으로, 추론 규칙 적용 이전에 781,296건, 추론 규칙 적용 이후에 1,459,047건이 구축되어 있다.

4.2 sameAs 관계 관리기

그림 2는 연구자 간 sameAs 관계를 설정하고 해제하는 기능을 담당하는 인력 sameAs 관계 관리기의 화면 인터페이스를 보여준다. 임의의 인력 개체들의 URI들을 입력으로 받아 그들 간에 sameAs 관계를 설정하거나, 기존에 설정된 sameAs 관계들을 검색/해제하는 기능을 포함하고 있다. 그림은 "서희달"이라는 이름의 두 인력의 URI를 입력으로 받아 sameAs 관계를 설정하는 예를 보인다. 그림 2의 상단에서 설정한 sameAs 관계는 온톨로지 개체 관계의 지식으로 다음과 같은 sameAs 트리플의 형태로 변환되어 저장된다. 이 연구에서는, sameAs 트리플이 표현하는 sameAs 관계에서의 대표 개체를, 트리플의 SUBJECT쪽으로 가정하였다.

http://www.kisti.re.kr/isrl#PER_0005000056 sameAs http://www.kisti.re.kr/isrl#PER_0005000057

sameAs 관계 관리기는, sameAs 관계 설정/해제 시, sameAs 그룹의 대표 개체를 표현하기 위해 도입된 isrl:standForSameAsGroupOf 관계의 적절한 변경도 담당한다.

4.3 sameAs 그룹의 지식 처리

sameAs 적용에 있어서, sameAs 그룹의 지식 획득을 위해 sameAs 그룹 내 소속 개체들의 속성 정보가 병합되는 단계에서는, 계산의 복잡성이 야기될 수 있다. 예를 들어, 기반정보 온톨로지에서, 각 연구자는 주요 연구 주제/분야 분류를 갖는데, 이는 해당 연구자의 연구성과물에 부여된 주제/분야 분류정보로부터 계산된 것이다. 즉, 아래 표 1에서처럼 "정한민[1]"은 그의 네 편의 논문들에 할당된 주제 분류 정보들이 통합되어 자신의 주제 분류 정보를 할당받는다.

만약, 정한민[1]과 또 다른 정한민[2]가 sameAs 관계로 묶인다면, sameAs 그룹 차원의 인력의 주제(Top 3)는 정한민[1]과 정한민[2]의 각 연구성과물의 주제 정보들을 참조하여 재계산되어야 할 것이다. 이러한 계산을 질의 시점에 처리하게 되면 만족스러운 실시간 응답 속도를 보장받기 어려울 수 있다. 이 연구의 sameAs 적용에서는, sameAs 그룹의 지식 중 높은 계산 복잡도가 요구되는 속성 정보(예: 표 1의 Top3 주제)를, sameAs 설정/해

5) 한국정보과학회, 한국정보처리학회, 전자공학회, 통신학회 등

표 1 연구성과물의 주제 분류 정보의 연구자로의 전이

논문	저자	주제 (Top 3)
논문 1	**정한민**[1], 성원경, 박동인	정보추출, 자연어질의, 시맨틱웹
논문 2	이병희, **정한민**[1], 성원경	정보추출, 시맨틱웹, 메타데이타
논문 3	**정한민**[1], 이승우, 성원경	정보추출, 메타데이타, 시맨틱웹
논문 4	**정한민**[1], 성원경, 김평	정보추출, 코퍼스, 용어생명주기
	정한민[1]	정보추출, 시맨틱웹, 메타데이타

제 시점에 미리 계산해 두는 방식을 취하였다. 좀 더 구체적으로, sameAs 관계의 설정/해제로 인해, sameAs 그룹의 지식 중 미리 계산되어야 할 (계산 복잡도가 높은) 속성 값들은 해당 sameAs 그룹의 대표 개체의 속성 값에 반영된다.

4.4 sameAs 후처리

이 연구에서, sameAs 후처리는 단순히, 사용자 질의의 결과에 나타난 개체 중 sameAs 관계를 갖는 개체가 발견될 경우, 해당 개체의 URI를 그 개체가 소속된 sameAs 그룹의 대표 개체의 URI로 교체하는 과정이다. 이는 사용자 질의에 대한 SPARQL 질의의 조건부에서, 질의의 결과로 생성될 각 개체(?x)에 대해, sameAs 트리플을 참조하여 그 개체의 대표 개체(?y)를 얻어 오도록 하는 다음과 같은 조건을 추가함으로써 간단히 수행될 수 있다.

OPTIONAL(?y isrl:standForSameAsGroupOf ?x)

isrl:standForSameAsGroupOf은 그림 1에서 설명한 것처럼, 하나의 sameAs 그룹 내에서 대표 개체를 표현하기 위해 본 연구에서 도입한 객체관계속성(object property)이다.

예를 들어 한국과학기술정보연구원에 소속된 연구자들을 가져오는 질의에 대한 원 SPARLQL 질의와 sameAs를 고려한 SPARQL 질의는 다음과 같다.

```
SELECT   ?x
WHERE    ?x isrl:hasInstitutionOfPerson ?z
         ?z isrl:nameOfInstitution "한국과학기술정보연구원"^^xsd:string
```

↓

```
SELECT   ?y
WHERE    ?x isrl:hasInstitutionOfPerson ?z
         ?z isrl:nameOfInstitution "한국과학기술정보연구원"^^xsd:string
         OPTIONAL(?y isrl:standForSameAsGroupOf ?x)
```

위와 같이 sameAs 처리를 SPARQL 질의에서 간단히 처리할 수 있는 것은, 하나의 sameAs 그룹 내에서 대표 개체를 제외한 다른 모든 개체가 해당 sameAs 그룹의 대표 개체를 가리키도록 하는 isrl:standForSameAsGroupOf 트리플을 sameAs 관계 관리기에서 미리 만들어 두었기 때문이다.

그림 3은 sameAs 관계 설정 전과 후의 공저 관점의 연구자 네트워크의 한 예를 보여 준다. 공저 관점의 연

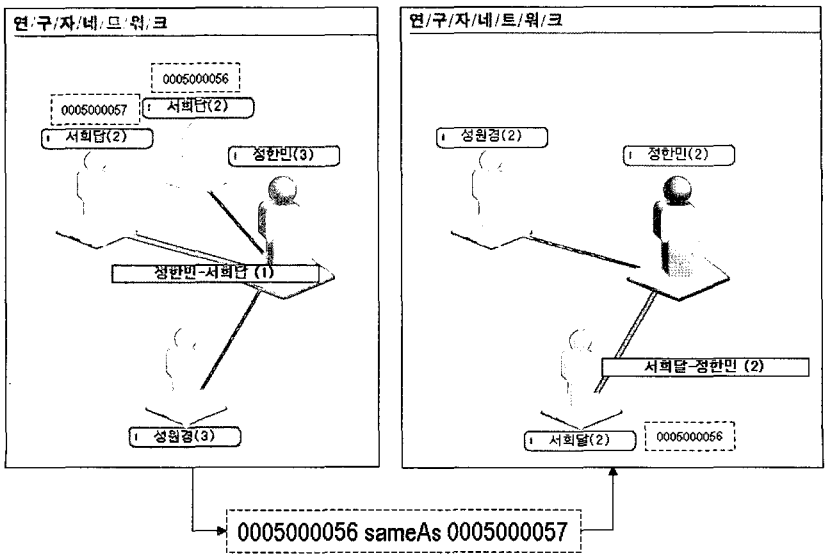

그림 3 sameAs 관계 설정으로 인한 연구자네트워크의 변화

구자 네트워크에서 노드는 URI로 구분되는 연구자를 의미하며, 노드 간 링크는 링크로 연결된 두 연구자가 공동 연구의 이력이 있음을 의미한다. 그림 3에서는 노드 간 링크의 굵기로 공동 연구의 강도를 표현하였고, 노드에 표시된 연구자 이름 뒤의 괄호 안의 숫자(예: 정한민(3))는 서로 다른 공동 연구자의 수를 의미한다. 그림 3의 좌측은 sameAs 설정 전에 두 명의 동명이인 "서희달"이 "정한민"과 공동 연구를 수행한 이력이 있음을 나타내고 있으며, 두 명의 동명이인 "서희달"의 URI 값(네임스페이스는 생략하고 식별자만 보임)들은 점선박스로 표시되어 있다. 또한, "0005000057"에 해당하는 "서희달"과 "정한민"간의 링크에 툴팁박스로 표시된 것은 두 사람이 공동 연구로 창출한 연구성과물의 건수이다. 그림 3의 우측은 sameAs 설정 이후 두 명의 동명이인 "서희달"이 하나의 노드로 병합된 모습을 보여 준다. 전술한 것처럼, 이 연구에서는, sameAs 트리플에서 SUBJECT쪽을 대표 개체로 가정했으므로, 그림 3의 우측 "서희달"의 URI는 sameAs 그룹의 대표 개체의 URI인 "0005000056"에 해당한다.

5. 결 론

본 연구에서는, 모델링 대상이 되는 실세계의 한 부분에 대한 전자적 표상으로써의 온톨로지 개체 집단의 정합성을 보장하는 측면에서, 개체 간 sameAs 관계 설정 장치 사용과 관련된 여러 이슈들을 제기하고 사례 중심의 현실적이고 경험적인 대안들을 제시하였다. 그러나, 온톨로지 내 임의의 두 개체쌍에 대해 암묵적으로 differentFrom을 가정하는 것은 개체 집단에 대한 통제가 어느 정도 가능한 전용 시맨틱웹 환경에서나 가능한 선택인 지도 모른다. 즉, 이 논문에서 제기된 여러 이슈들에 대해, 개체 간 같음과 다름에 대한 판단이 부재한 상황을 가정하는 공용 시맨틱웹 환경에서는 본 논문의 저자들이 택한 결정들과 다른 대안들이 취해질 수 있음을 잊지 말아야 한다.

또한, 논문에서는 특정 연구분야 도메인에 대해 실제 시맨틱 웹 온톨로지를 구축하고 관련 추론서비스를 위한 질의 처리를 수행하는 과정에, 개체 간 sameAs 관계가 적용되는 실제 사례를 소개하였다. sameAs 관계 관리기, sameAs 후처리기 등의 구현 사례들은 타 온톨로지 기반 응용 시스템의 개발에서 sameAs를 적용하는 경우의 선행 모델이 될 수 있을 것이다. 그러나, 본 연구의 sameAs 적용 경험은 특정 연구분야 온톨로지에 한정된 것일 수 있으며 타 도메인 온톨로지들에 공히 적용되지 않을 수도 있다. 따라서, 향후 다양한 도메인의 온톨로지에 sameAs를 적용하여 논문에서 제시한 대안들을 검증하고 일반화시킬 필요가 있다.

마지막으로, 이 논문에서는 온톨로지 정합성 보장을 위한 sameAs 적용 이슈 및 대안들을 개체 식별 정합성 측면으로 제한하여 고찰하였다. 3장 마지막 부분에 언급한 것처럼 sameAs 적용을 개체 참조 정합성과 개체 속성값 도메인 정합성 보장까지 동시에 만족시키는 상황으로 확장할 경우 여러 복잡한 이슈들이 추가로 발생하게 되므로, 향후 이 부분에 대한 추가 연구가 요구된다.

참 고 문 헌

[1] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). "The Semantic Web," *Scientific American* 279(5):34-43, 2001.

[2] Gruber, T.R. (1993). "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition* 5(2):199-220, 1993.

[3] Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., and Patel-Schneider, P.F. (2004). "OWL Web Ontology Language Reference," http://www.w3.org/TR/owl-ref/, 2004.

[4] Ehrig, M., and Staab, S. (2004). "QOM - Quick Ontology Mapping," *Proceedings of the Third International Semantic Web Conference*, Hiroshima: Japan, Nov. 7-11, pp.683-697, 2004.

[5] Gahleitner, E., and Wöß, W. (2004). "Enabling Distribution and Reuse of Ontology Mapping Information for Semantically Enriched Communication Services," *Proceedings of the 15th International Workshop on Database and Expert Systems Applications*, Zaragoza:Spain, Aug. 30 - Sep. 3, pp.116-121, 2004.

[6] Das, S., Chong, E.I., Eadon, G., and Srinivasan, J. (2004). "Supporting Ontology-based Semantic Matching in RDBMS," *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, Toronto:Canada, Aug. 31 - Sep. 3, pp.1054-1065, 2004.

[7] Bicer, V., Laleci, G.B., Dogac, A., and Kabak, Y. (2005). "Providing Semantic Interoperability in the Healthcare Domain through Ontology Mapping," *Proceedings of eChallenges-2005*, Ljubljana:Slovenia, Oct. 19-21, 2005.

[8] Alani, H., Dasmahapatra, S., Gibbins, N., Glaser, H., Harris, S., Kalfoglou, Y., O'Hara, K., Shadbolt, N. (2002). "Managing Reference: Ensuring Referential Integrity of Ontologies for the Semantic Web," *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management*, Siguenza:Spain, Oct. 1-4, pp.317-334, 2002.

[9] Han, H., Giles, C.L., Zha, H., Li, C., Tsioutsiouliklis, K. (2004). "Two Supervised Learning Approaches for Name Disambiguation in Author Citations," *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*, AZ:USA, Jun. 7-11, pp.296-305,

2004.
[10] Aswani, N., Bontcheva, K., Cunningham, H. (2006). "Mining Information for Instance Unification," *Proceedings of the 5th International Semantic Web Conference*, GA:USA, Nov. 5-9, pp.329-342, 2006.
[11] McRae-Spencer, D.M., Shadbolt, N.R. (2006). "Also by the Same Author: AKTiveAuthor, a Citation Graph Approach to Name Disambiguation," *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*, NC:USA, Jun. 11-15, pp.53-54, 2006.

강 인 수
1988년~1995년 경북대학교 컴퓨터공학과(학사). 1997년~1999년 포항공과대학교 컴퓨터공학과(석사). 2001년~2006년 포항공과대학교 컴퓨터공학과(박사). 1995년~1997년 (주)포스데이타 DBA. 1999년~2001년 포항공과대학교 학술정보원 연구원. 2006년~현재 한국과학기술정보연구원 초청연구원. 관심분야는 자연어처리, 시맨틱웹, 정보검색

정 한 민
1988년~1992년 포항공과대학교 전자계산학과(학사). 1992년~1994년 포항공과대학교 전자계산학과(석사). 2000년~2003년 포항공과대학교 컴퓨터공학과(박사) 1994년~2000년 한국전자통신연구원 선임연구원. 2000년~2004년 ㈜다이퀘스트 연구소장/기술이사. 2004년~현재 한국과학기술정보연구원 선임연구원. 2005년~현재 과학기술연합대학원대학교 겸임교수. 관심분야는 자연어처리, 시멘틱 웹, 정보 추출, 정보검색

이 승 우
1997년 2월 경북대학교 컴퓨터공학과(공학사). 1999년 2월 포항공과대학교 컴퓨터공학과(공학석사). 1999년~2000년 포항공과대학교 정보통신연구소 연구원. 2005년 8월 포항공과대학교 컴퓨터공학과(공학박사). 2005년~2006년 대구가톨릭대학교 컴퓨터교육과 강의전담교원. 2006년~현재 한국과학기술정보연구원 선임연구원. 관심분야는 자연어처리, 시맨틱 웹, 정보검색

김 평
1997년 충남대학교 전산학과(학사). 1999년 충남대학교 컴퓨터과학과(석사). 2004년 충남대학교 컴퓨터과학과(박사). 2000년~2003년 (주)엔퀘스트테크롤러지 개발실장. 2004년~현재 한국과학기술정보연구원 선임연구원. 관심분야는 정보검색, 텍스트 마이닝, 시맨틱 웹

이 미 경
1999년 대구대학교 전자계산학과(학사) 2002년 경북대학교 컴퓨터공학과(석사) 2002년~2005년 한국전자통신연구원 지능형로봇연구단 지식및추론연구팀 연구원. 2005년~현재 한국과학기술정보연구원 정보기술개발단 정보서비스연구팀 연구원. 관심분야는 시맨틱웹, 온톨로지 추론 시스템

성 원 경
1987년 2월 연세대학교 불어불문학과(학사). 1989년 2월 연세대학교 불어불문학과(석사). 1996년 12월 프랑스 파리7대학교 언어학과(박사). 1997년~1998년 한국전자통신연구원 Post-doc. 1998년~2001년 L&H Korea(주) 책임연구원. 2001년~2003년 (주)보이스텍 연구개발본부장/상무이사. 2004년~현재 한국과학기술정보연구원 정보시스템연구팀장/책임연구원. 2004년~현재 과학기술연합대학원대학교 겸임교수 관심분야는 자연어처리, 시맨틱 웹