시맨틱 웹 응용 서비스에서의 텍스트 처리 기술 적용

정한민 강인수 구희관 이승우 김평 성원경
한국과학기술정보연구원 정보시스템연구팀
jhm@kisti.re.kr

Use of Text Processing Technologies in a Semantic Web Application

Hanmin Jung In-Su Kang HeeKwan Koo Seungwoo Lee Pyung Kim Won-Kyung Sung
Information System Research Lab., KISTI

요 약

본 논문은 시맨틱 웹 응용 서비스를 구현함에 있어 필수적으로 요구되는 온톨로지 인스턴스 구축을 효율적으로 처리하는 데 있어 텍스트 처리 기술이 어떤 역할을 수행할 수 있는가를 OntoFrame-K®라는 시맨틱 웹 기반 정보 유통 체계에의 적용 사례를 통해 살펴본다. 본 논문에서 소개하는 텍스트 처리 기술은 개체 확인을 통한 개념 사례화, 주제·분야 할당을 통한 메타데이터 확장에, 그리고 인용 정보 추출 및 인용 관계 구축을 통한 객체 관계 속성 구축에 적용된다. 개체 확인에서는 메타데이터 비교 및 병합을 사용하였으며 이를 기반으로 한 수작업 구축을 통해 8,543명의 인력 URI를 확보하였다. 주제 및 분야 할당에서는 색인어와 분야분류명이 매핑된 시소러스 개념어의 매칭을 통해 색인어 별 TF (Term Frequency), 색인어와 매칭된 개념어 별 TF, 색인어와 매칭된 개념어 별 시소러스에서의 깊이, 색인어와 매칭된 개념어 별 개념 패싯, 색인어와 매칭된 각 개념어에 부착된 분야분류명 목록 등 할당을 위한 다양한 자질을 확보·적용하였다. 인용 정보 추출과 인용 관계 구축에서는 객체 URI와 인력 URI를 기반으로 하여 자동 추출된 인용 정보를 반영하는 방식으로 7,237개 문헌으로부터 총 135개의 인용 네트워크 그룹을 자동으로 확보하였다. 본 연구를 통해 제시된 텍스트 처리 기술의 활용 방안이 향후 시맨틱 웹 응용 서비스 및 인프라 구현에서 다각적으로 활용될 수 있기를 기대한다.

1. 서론

시맨틱 웹에 대한 관심의 증가는 온톨로지, 추론, 시맨틱 웹 서비스 등에 대한 기술 발전을 가져오고 있다. 특히, 2005년 이후에는 추론을 위한 규칙 기술과 시맨틱 웹 기반 질의 기술이 그 중요성을 더하게 되었다 [10]. 그렇지만, 아직까지는 대용량 데이터 처리, 시맨틱 웹 포털 구축 등에 있어서 그 발전이 성숙되지 않아 텍스트 처리 기술의 중요성이 크게 강조되고 있지 않은 상황이다.

시맨틱 웹 응용 서비스 개발을 위해서는 시간·인력 소모적인 콘텐츠의 의미화가 우선적으로 이루어져야 한다. 여기에서 우리가 텍스트 처리 기술에 주목하는 이유는 단순한 수작업을 획기적으로 감소시킴으로써 지속적이고 풍부한 의미 기반 콘텐츠를 확보할 수 있도록 하는 중요한 역할을 하기 때문이다. 콘텐츠의 의미화에서 텍스트 처리 기술은 주로 의미 태깅을 통해 어휘에 의미를 부여하고 어휘 간에 관계를 설정할 수 있게 하거나, Wrapper를 이용하여 메타데이터에 의미를 부여할 수 있도록 하는 정보 추출 관점에서의 지원 도구 역할을 수행하는 데 한정되어 있었다.

본 연구는 이러한 한정된 텍스트 처리 기술의 적용 범위를 좀더 넓힐 수 있도록 URI (Uniform Resource Identifier), 전문 용어, 시소러스와 같은 언어 자원을 적극 활용하고자 한다. 또한, 시맨틱 웹 기반 자원으로서의 온톨로지 인스턴스 구축을 위해 개체 해소, 주제·분야 할당, 인용 정보 추출 및 인용 관계 구축에 다각적으로 텍스트 처리 기술이 활용됨을 보임으로써 시맨틱 웹과 텍스트 처리 기술의 접목 가능성을 보이고자 한다.

2. 시맨틱 웹 기반 정보 유통 체계

OntoFrame-K®는 시맨틱 웹과 정보 자원 공유 기술의 심화 연구와 융합을 통해 방대한 양의 과학기술 정보 자원을 효율적으로 공유하고 유통시킬 수 있도록 하는 정보 유통 체계이다 [11]. 특히 URI 서버를 이용하여 온톨로지 인스턴스를 다루고, 과학기술 분야 시소러스 및 분야 분류 체계 등 언어 자원을 통합하여 관리하고, DBMS 기반 추론 서비스를 제공하는 등 다른 시맨틱 웹 포털에서 볼 수 없는 특징을 가진다. OntoFrame-K®가 제공하는 서비스는 크게 정보 유통 서비스와 추론 서비스로 나누어진다. 정보 유통 서비스는 중앙 지식 서버를 통한 성과 정보의 등록·검색을 포함하는 정보의 체계적인 생명 주기 관리를 담당한다. 추론 서비스는 추론 규칙과 URI 서버를 이용하여 지식을 확장하고 SPARQL (SPARQL Protocol And RDF Query Language) 기반 사용자 질의에 대해 사용자에게 연구자 네트워크, 연구자 정보, 연구 성과 맵, 통계 정보, 성과 정보, 기관 정보 등 다양한 서비스를 제공한다.

본 연구는 상기 정보 유통 체계 상에서의 핵심 지식인 온톨로지 인스턴스를 구축하는 과정에서 이용되는 여러 텍스트 처리 기술을 소개함으로써 시맨틱 웹 기술과 텍스트 처리 기술의 결합을 모색하고자 한다. 온톨로지 인스턴스를 구축하기 위해서 일반적으로 참조하는 자원은 메타데이터와 온톨로지 스키마가 있다. 본 연구에서는 온톨로지 인스턴스에 대한 정합성을 보장하기 위해 URI 서버를 추가로 참조한다. 온톨로지 인스턴스 구축은 메타데이터 내의 값을 온톨로지 스키마를 참조하여 개체 확인 (Identity Resolution)하는 개념 사례화 (Concept Instantiation)를 거쳐 URI 서버로부터 URI를 부여받고, 해당 개체에 대해 온톨로지 스키마에 정의된 데이터타입 속성 (Datatype Property)과 개체 관계 속성 (Object Property)을 구축하는 과정을 통해 이루어진다.

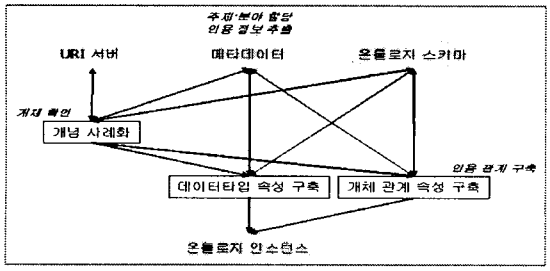

그림 1. 온톨로지 인스턴스 구축 과정 및 텍스트 처리 기술 적용 예

그림 1은 이러한 온톨로지 인스턴스 구축 과정에서 사용된 개체 확인, 주제·분야 할당, 인용 정보 추출 및 인용 관계 구축에의 텍스트 처리 기술 적용 예를 보여준다. 개체 확인은 메타데이터 내의 값이 가지는 애매성을 해소하는 데 도움을 주며, 주제·분야 할당은 메타데이터가 가지고 있지 않지만 데이터타입 속성으로서 요구되는 부가 정보를 획득할 수 있도록 하며, 인용 정보 추출 및 인용 관계 구축은 추론 서비스 중 연구자 네트워크에서 인용 관점의 연구자 간 관계를 제공할 수 있게 한다. 이 외에도 여러 부분에서 텍스트 처리 기술이 적용될 수 있으나, 본 논문에서는 이 세 가지 부분에 한정하여 설명하고자 한다.

3. 텍스트 처리 기술 적용

텍스트 처리 기술은 사람이 효율적으로 처리하지 못하는 부분에 적용할 때 좋은 결과를 가져오는데, 본 연구에서도 시간 소모적이며 오류 발생 가능성이 높은 부분에 그 적용 초점을 맞춘다. 예를 들어, 인력에 대한 개체 확인의 경우 수작업에만 의존하게 되는 경우 동명이인 해소에서 검색 범위 한정, 공저자 분석 오류 등으로 인해 부정확한 결과를 가져올 수 있다. 특정 성과에 대한 주제·분야 할당 역시 관리자나 등록자의 작업 환경, 판단 오류 등으로 인해 일관성 있는 결과를 낳기 힘들며, 특히 분야 분류 체계와 시소러스 등 언어 자원의 변경에 따른 신속한 대처가 어렵다. 효율적인 온톨로지 인스턴스 구축을 위한 텍스트 처리 기술의 적용 예를 다음에서 살펴보고자 한다.

3.1 개념 사례화 - 개체 확인

온톨로지 상의 개념 (Class; Concept)에 해당하는 메타데이터 내의 값을 사례화 (Instantiation)할 때 애매성을 해소하는 개체 확인 (Identity Resolution)이 필수적이다. 시맨틱 웹 구조에서 최하위 기반이 되는 URI는 개체가 유일한 존재임을 보장한다. 그렇지만, URI를 할당하는 과정에서 메타데이터 내의 개체를 표현한 문자열의 동형이의나 이형태로 인해 애매성이 생기기 때문에 개체 확인이 쉽지 않다. 특히, 인력 개체를 구축할 때 동명이인 해소가 필요함에도 활용할 수 있는 정보에 제약이 있는 경우가 흔하다. 예를 들어, 논문의 경우에 저자를 식별할 수 있는 정보로는 소속 정보, 이메일, 공저자 정보 외에는 이용할 수 있는 정보가 별로 없다 (이 중 소속 정보와 이메일은 필수로 요구되지도 않는다.). 이러한 정보가 메타데이터로 구축이 되어 있다고 할 지라도 수작업만으로 동명이인을 해소하는 것은 시간 소모적일 뿐만 아니라 유사한 대량의 데이터 처리로 인해 오류를 포함할 가능성이 커질 수밖에 없다.

텍스트 처리 기술은 특정 개체를 표현하는 메타데이터 내 비교 가능한 필드를 이용함으로써 구축 이

전의 전처리 과정이나 구축 이후의 검증 과정에서 도움을 줄 수 있다. 본 절에서는 문헌 내의 동명이인 해소를 포함한 인력 인스턴스 생성에서의 텍스트 처리 기술 적용 예를 보여준다.

문헌으로부터 직접적으로 구축할 수 있는 메타데이터는 제목, 출처, 권·호, 원문파일명, 저자, 소속 기관, 소속 부서, 이메일 등을 포함한다. 이 중 저자(인력) 클래스를 사례화할 때 이용할 수 있는 정보로는 소속 기관, 소속 부서, 이메일이 있다. 이 세 필드를 하나의 비교 대상으로 묶는다면 동일한 소속 정보와 이메일을 가지는 경우에 동명이인인지 동일 인물인지를 어느 정도 구분할 수 있을 것이다. 그렇지만, 특정 인력이 문헌을 저술한 후 소속을 변경한 경우나 소속 정보가 불완전한 경우에 인력 추적을 위한 부가 정보가 필요하다. 공저자 정보가 문헌 상에 나타난 인력의 개체 확인에 중요한 역할을 하는 이유가 여기에 있다. 비록 소속이 변경된 경우라 할 지라도 공저자 관계를 유지하는 경우가 작지 않으며, 소속 정보의 확보 정도가 서로 다른 문헌이 존재할지라도 공저자 확인을 통해서 어느 정도 개체 확인 문제를 해결할 수 있다. 본 연구에서는 다음의 절차를 이용하여 동명이인이 해소된 초벌의 인력 그룹을 제공하여 인력 정보 구축에 도움을 준다.

동명을 가진 문헌 각각을 하나의 인력 그룹으로 하여,
(1) 소속 정보와 이메일이 같은 동명을 묶어 하나의 인력 그룹으로 병합한다. <인력 간 병합>
(2) 공저자를 공유하는 서로 다른 인력 그룹이 존재하는 경우에 두 인력 그룹을 하나의 인력 그룹으로 병합한다. 이때 소속 정보와 이메일 쌍은 하나의 목록으로서 관리하여 비교 대상을 확대한다. <인력 그룹 간 병합>
(3) (1)과 (2) 과정을 더 이상의 인력 그룹이 병합되지 않을 때까지 반복한다 [3].

표 1. 소속 정보, 이메일, 공저자 관계를 이용한 인력 그룹 생성 (빗금 친 부분은 자동으로 부여된 값으로 동명에 대한 개체 확인 결과임)

인명	그룹 ID	공저자 관계	소속 기관	소속 부서	이메일
김종원	ID1	한상우; 이동훈; 김종원;	광주과학 기술원	정보통신 공학과	jongwon@kjist.ac.kr
김종원	ID2	강민수; 김종원; 이원철; 신요안;	숭실대학교	정보통신 전자공학부	
김종원	ID1	권영우; 김종원;	광주과학 기술원	정보통신 공학과 네트워크 미디어연구실	jongwon@kjist.ac.kr
김종원	ID1	홍기원; 김양근; 최덕재; 김종원; 박주원; 정중렬;	광주과학 기술원		jongwon@kjist.ac.kr
김종원	ID2	김종원; 이장욱; 이원철; 유명식; 신요안;	숭실대학교	정보통신 전자공학부	ocosjwt@amcs.ssu.ac.kr
김종원	ID3	이승주; 김종원;	광주과학 기술원	정보통신 공학과 네트워크 미디어연구실	jongwon@netmedia.gist.ac.kr
김종원	ID2	이장욱; 김종원; 신요안;	숭실대학교	정보통신 전자공학부	
김종원	ID1	이철호; 최정용; 권영우; 김종원; 신지태; 김재곤;	광주과학 기술원	정보통신 공학과 네트워크 미디어연구실	jongwon@gist.ac.kr
김종원	ID1	한상우; 김종원;	광주과학 기술원	정보통신 공학과 네트워크 미디어연구실	jongwon@nm.gist.ac.kr
김종원	ID1	황구연; 김재윤; 이동훈; 김종원; 이진영; 주성순;	광주과학 기술원	정보통신과	jonwon@netmedia.gist.ac.kr

표 2. 인력 그룹 정보

인명	그룹 ID	소속 기관	소속 부서	이메일
김종원	ID1	광주과학 기술원	정보통신공학과	jongwon@kjist.ac.kr
		광주과학 기술원	정보통신공학과 네트워크미디어 연구실	jongwon@kjist.ac.kr
		광주과학 기술원		jongwon@kjist.ac.kr
		광주과학 기술원	정보통신공학과 네트워크미디어 연구실	jongwon@gist.ac.kr
		광주과학 기술원	정보통신공학과 네트워크미디어 연구실	jongwon@nm.gist.ac.kr

		광주과학기술원	정보통신과	jonwon@netmedia.gist.ac.kr
김종원	ID2	숭실대학교	정보통신전자공학부	
		숭실대학교	정보통신전자공학부	ocosjwt@amcs.ssu.ac.kr
김종원	ID3	광주과학기술원	정보통신공학과 네트워크미디어 연구실	jongwon@netmedia.gist.ac.kr

표 1은 (1)과 (2)를 통해 처리된 인력 그룹을 보여준다. 표 2는 각 인력 그룹에 속한 소속 정보와 이메일 쌍을 보여준다. 이러한 자동화된 처리를 통해 수작업으로 판단하기에 부담이 되는 부분을 용이하게 해결함으로써 온톨로지 인스턴스 확장 및 개념 사례화의 효율성을 높일 수 있다. 표 2를 통해 10명의 '김종원' 이라는 저자가 동명이인을 해소하면 3명으로 줄어든다는 것을 알 수 있다. 물론 ID1과 ID3이 동일 인물이라는 것을 짐작할 수 있지만, 현재 병합을 위한 제약으로서 부분 매칭된 쌍은 인정하지 않기 때문에 별개의 인물로 간주가 된다. 그렇지만, 이는 구축자의 확인 작업을 통해 동일 인물로서 간주되어 '과학기술인력 종합정보시스템'에서 사용하는 인력 ID '6410138462'를 부여 받게 된다. 그룹 ID가 'ID1'인 경우 소속 정보와 이메일 조합 쌍이 6가지나 존재한다. 수작업에만 의존하는 경우 이를 하나로 인지하는 과정에서 일관성을 유지하기 어렵기 때문에 개체 확인 결과의 신뢰성이 떨어진다.

이와 같이 개념 사례화 과정에서 적용된 텍스트 처리 기술은 수작업을 간소화시킬 수 있는 주요한 수단으로서 작용하게 되며 지속적인 확장 및 검증에서도 중요한 역할을 한다는 것을 알 수 있다. 이러한 자동화된 처리 방법은 사례화 대상이 되는 메타데이터의 크기가 커질수록 더욱 효과적인데, 상기 예에서 보듯이 비교 대상이 되는 소속 정보와 이메일 쌍 집합이 점점 커지기 때문이다. 통계 기반 방법이 신뢰성을 가지기 위해 충분한 크기의 말뭉치를 필요로 하는 것처럼 개념 사례화 과정에서도 충분한 메타데이터가 수작업을 줄이는 데 도움이 된다.

3.2 메타데이터 확장 - 주제·분야 할당

메타데이터는 온톨로지를 이용한 개념 체계화 (Formalization)의 주요 대상이다. 원문의 경우 정확도 측면에서 의미 태깅 수준 이상의 체계화는 현실적으로 어려우며, 개념 단위를 어휘로 보지 않고 하나의 객체로 간주할 때는 잘 구축된 메타데이터가 더욱 필요하다. 그렇지만, 객체를 표현하는 메타데이터가 응용 서비스에서 요구하는 다양한 관계를 갖도록 처음부터 설계하지 않는 이상 부족한 관계가 생길 수 밖에 없다. 주제 및 분야 역시 이러한 관계 유형의 하나로 볼 수 있는데, 정보 양이 많을수록 체계적인 분류의 필요성은 증가하지만 이에 대처하는 것이 쉽지 않다. 시스템 관리자나 정보 관리자가 직접 문서 내 주제나 분야를 할당하는 경우에는 유지 보수에의 대처, 일관성 유지 측면에서 여러 문제점을 가지기 때문이다.

기존에 색인어를 이용하여 자동으로 주제를 선정하는 시도 [2] [8]와 시소러스 상의 관계나 범주 사전을 이용하여 문서를 분류하려는 시도 [5] [4]가 있었으나, 주제어 통제, 주제 및 분야 동시 할당, 일관성 있는 정보 갱신을 보장하기 힘들었다. 본 연구에서는 주제어 통제를 위해 시소러스 개념어를 사용하고, 분야 할당을 위해 시소러스 개념어를 매개체로 한 분야분류명-개념어 간 매핑 정보를 참조하는 방법을 도입한다 [7].

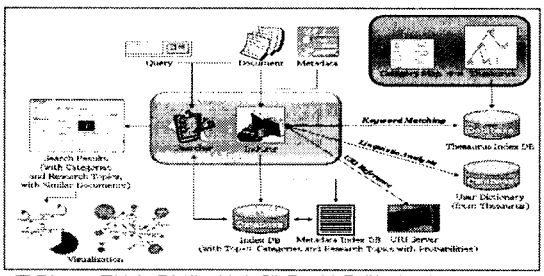

그림 2. 정보 검색 시스템을 이용한 주제 및 분야 할당 방안

먼저 원문을 대상으로 색인기를 이용하여 색인어를 추출한다. 추출된 색인어는 시소러스 개념어와 매칭된다. 주제어 통제를 위해 시소러스 개념어 중 과학기술 분야[1]를 대표한다고 보기 어려운 것을 불용어로서 처리하는데, 현재 본 연구에서는 표준국어대사전 [12]과 DF (Document Frequency)를 이용하여 그 대상을 선별한다[2].

상기 과정을 통해 얻어진 시소러스 개념어 목록과 매칭된 색인어 목록을 기본적으로 이용하여 해당 문서의 주제 및 분야를 결정하는 알고리즘을 적용한다. 사용하는 자질로는 색인어 별 TF (Term Frequency), 색인어와 매칭된 개념어 별 TF, 색인어와 매칭된 개념어 별 시소러스에서의 깊이, 색인어와 매칭된 개념어 별 개념 패싯 (Conceptual Facet), 색인어와 매칭된 각 개념어에 부착된 분야

[1] 본 연구는 과학기술 콘텐츠 서비스를 대상으로 하므로 시소러스 구축 범위도 과학기술 분야로 한정하였다.

[2] 불용어 목록은 본 연구를 통해 구축한 31,000여 개의 개념어 중 표준국어대사전과 매칭된 2,332개와 DF 상위 개념어 중 선별된 310개를 포함한다.

분류명 목록 등이 있다. 다음은 주제·분야 할당 알고리즘을 보여준다.

(1) 색인어 목록: 전체 문헌 집합 중 k번째 문헌 $D_k = \{t_{k1},...,t_{km}\}$ 는 m개의 색인어를 가진다. t_{ki} 는 D_k 에 나타난 i번째 색인어를 의미한다.
(2) 개념어 목록: 시소러스 개념어 집합 $S = \{s_1,...,s_p\}$ 는 전체 p개의 개념어를 가진다.
(3) 색인어 별 TF: $tf_{D_k}(t)$ 는 문헌 D_k 내에서 용어 t가 출현한 빈도수 (TF)이다.
(4) 색인어와 매칭된 개념어 별 TF: 각 개념어 s에 대해 문헌 D_k 내에서의 개념어 빈도수는 $tf_{D_k^S}(s)$ 와 같이 정의한다.
색인어와 매칭된 개념어 별 TF $tf_{D_k^S}(s)$ 는 색인어와 매칭된 개념어 빈도수이며, 이것은 한 문서 내에서 해당 개념어 s로 대응되는 색인어의 빈도수 합이다. 즉, $tf_{D_k^S}(s) = \sum_{\substack{s=f_{동의어그룹대표어}(t),\\ t\in D_k \cap S}} tf_{D_k}(t)$ 이며, 분야 별 TF $tf_{D_k^c}(c)$ 는 $tf_{D_k^c}(c) = \sum_{c\in Topic(s)} tf_{D_k^s}(s)$ 와 같다[3].
여기에서 주목할 내용은 시소러스 상에서의 동등관계('USE/UF 관계'와 'RT-동등관계')에 해당하는 개념어 (동등개념어 집합)는 대표개념어로 정규화된다는 것이다. 대표개념어는 일반적으로 USE를 사용하며, 일단 정해지면 동등개념어 그룹 내에서 USE가 변경이 되더라도 전체색인을 수행하기 전까지는 대표개념어를 유지해야 한다. 그렇지 않으면, 동등개념어 집합 내에서의 서로 다른 개념어가 주제로서 문서에 할당되는 혼란이 일어날 수 있다. $f_{동의어그룹대표어}(s)$ 는 개념어 s가 속한 동등개념어 집합 내의 대표개념어를 반환하는 함수이다. D_k^S 는 D_k 의 색인어 중 시소러스 개념어에 매칭되는 색인어를 그 색인어와 매칭된 동등개념어 집합의 대표개념어로 정규화한 색인어 집합으로 $D_k^S = \{f_{동의어그룹대표어}(t) \mid t \in D_k \cap S\}$ 에 해당한다.
(5) 색인어와 매칭된 개념어 별 시소러스에서의 깊이
(6) 색인어와 매칭된 개념어 별 개념패싯 (Conceptual Facet)
(7) 색인어와 매칭된 각 개념어에 부착된 분야분류명 목록

주제 순위화 수식: $tf_{D_k^c}(s)$

주제 가중치[4] 수식: $w(s) = \frac{tf_{D_k^c}(s)}{\sum_{s'\in Top3s} tf_{D_k^c}(s')}$

분야 순위화 수식: $tf_{D_k^c}(c)$

분야 가중치 수식: $w(c) = \frac{tf_{D_k^c}(c)}{\sum_{c'\in Top3c} tf_{D_k^c}(c')}$

텍스트 처리 기술은 상기에서 언급한 주제·분야 외에도 메타데이터나 말뭉치 등으로부터 개체를 표현하는 필드를 확보하는 과정에도 적용할 수 있다. 예를 들어, 문서 수집 및 정보 추출 기술을 이용하여 인력 개체와 연관된 웹 사이트나 홈페이지 등을 통해 해당 개체의 신상 정보, 저작 정보 등을 다양하게 수집·추출·가공할 수 있다. 온톨로지와 연관된 메타데이터를 풍부하게 하는 것은 결국 다양한 시맨틱 웹 응용 서비스를 가능하게 한다는 것을 의미한다.

3.3 객체 관계 속성 구축 - 인용 정보 추출 및 인용 관계 구축

정보 추출 (Information Extraction) 기술은 대용량 언어 자원을 처리하기 위해 필수적인 기술이다. 시맨틱 웹에서 서비스를 다양화하기 위해 필요한 관계를 다음과 같은 추론 규칙을 통해 유도 객체 관계 속성 (Derived Object Property)으로서 획득할 수도 있지만, 추론으로 획득할 수 없는 기본 객체 관계 속성 (Basic Object Property)은 별도의 방법으로 획득해야 한다. 그러한 방법으로서 정보 추출 기술이 있으며, 본 연구에서는 인용 네트워크 (Citation Network) 구축을 위한 인용 관계 획득을 통해 그 적용 예를 보이고자 한다.

[공저자 객체 관계 속성 ('*isCoCreatorOf*') 획득을 위한 추론 규칙]
(?x hasCreationformation ?y1) (?y1 hasCreator ?z1) (?x hasCreationformation ?y2) (?y2 hasCreator ?z2) notEqual(?z1, ?z2) → (?z1 isCoCreatorOf ?z2)[5]
기본 객체 관계 속성: 'hasCreationformation', 'hasCreator'
유도 객체 관계 속성: 'isCoCreatorOf'

인용 네트워크는 문헌 간의 인용 관계를 이용하는

[3] 분야 c는 해당 분야분류명에 대응하는 개념어들로부터 구해진다.

[4] 주제 가중치는 상위 3개 주제어에 부여되는 수치로 각 주제가 해당 문서를 대표할 확률값을 상대적으로 정의한 것이다.

[5] 변수 ?x는 문헌, ?y1과 ?y2는 저작 정보, ?z1과 ?z2는 저자를 의미하는 변수이다. 하나의 문헌에 대해 서로 다른 저자가 있을 때 이 두 저자를 공저자 ('*isCoCreatorOf*') 관계로 연결하는 추론 규칙이다.

것으로 문헌 간 인용 네트워크를 구축할 수도 있고, 저자 간 인용 네트워크를 구축할 수도 있다. 저자 간의 경우에도 1저자만 고려하는 방안도 있으며 복수 저자 (2저자 이상)를 고려하는 방안도 있다. 다만 저자 간 인용 네트워크의 경우 저자에 대한 정확한 개체 확인이 필수적이다. 본 연구에서는 문헌에 나타난, 개체 확인된 모든 저자를 대상으로 저자 간 인용 관계 및 네트워크를 구축하고자 한다.

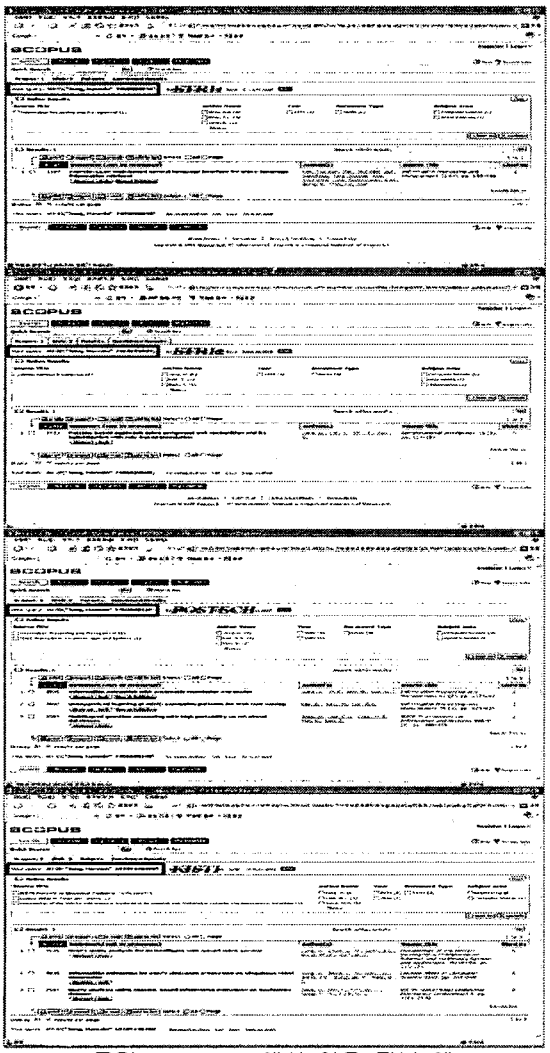

그림 3. SCOPUS에서 인용 정보 예

최근 서비스를 개시한 SCOPUS [13]의 경우 저자 ID를 이용하여 동명이인 문제를 어느 정도 해소하고 있으나, 저자 ID의 식별 요소로서 소속 정보와 이름만을 이용함으로써 동일 인물임에도 서로 다른 인력으로 보여지는 예를 쉽게 찾을 수 있다. 그림 3은 SCOPUS에서 '정한민'이라는 동일 인물이 4가지 다른 인력으로서 보여지는 예를 보여준다. 이 예에서는 해당 인력에 대한 소속 정보 변경을 고려하지 못하여 다른 인력으로 인식했는데 이러한 문제는 인력 정보를 다루는 데 있어서 흔히 발생할 수 있는 문제이다. 상기 예에서 '정한민'의 경우 다음과 같은 4가지 소속 정보로서 구별된다.

ETRI, Technology Lab, Taejon, South Korea
Electronics and Telecom. Res. Inst, Comp. and Software Technol. Lab., Taejeon, South Korea
Pohang Univ. of Sci. and Technology, Dept. of Comp. Sci. and Engineering, Pohang, South Korea
NULL

인력에 대한 URI 기반 정보 관리가 이루어지지 않는 이상 소속 변경에 따른 대응이 쉽지 않다. 본 연구에서는 URI 서버를 이용하여 OntoFrame-K®에서 다루는 온톨로지 상의 모든 개체[6]를 관리한다 [1]. URI 서버는 문헌 개체를 위한 객체 URI, 소속 기관을 위한 기관 URI, 소속 부서를 위한 부서 URI, 인력 정보를 위한 인력 URI 등 9개 개체 유형을 종합적으로 관리한다. 즉, 개체에 URI를 할당하고 해당 개체의 메타데이터를 관리하고 온톨로지 스키마에 기반하여 다른 개체와의 연결 (예. 인력 URI는 'hasDepartmentOfPerson'이라는 객체 관계 속성으로 부서 URI와 연결되며, 'hasOrganizationOfPerson'이라는 객체 관계 속성으로 기관 URI와 연결된다.)을 관리한다.

객체 관계 속성 구축을 위한 텍스트 처리 기술 적용에서 URI가 중요한 이유는 정보 추출된 결과를 개체를 표현하는 메타데이터로 매핑하고 객체 관계 속성과 연관된 필드 값의 경우에 대상 개체와의 비교를 URI 기반으로 함으로써 추출된 정보의 정확성을 보장할 수 있기 때문이다. 본 연구에서는 이러한 URI 서버를 기반으로 하여 인력 간 인용 관계 구축을 위해 문헌 내 참고 문헌으로부터 인용 정보를 추출하고, 각각에 해당하는 문헌 개체 (객체 URI에 대응)를 URI 서버에서 확인하고, URI 기반으로 문헌 간 인용 관계를 구축하고, 문헌 내 저자 중심으로 인용 관계를 변환한다.

문헌 내 인용 정보 추출은 참고 문헌 영역 인식, 인용 문헌 추출, 인용 정보 메타데이터 구축 과정으로 구성되는데, 정규 표현 (Regular Expression)을 이용하여 저자, 제목, 학회, 권·호, 연도 등을 추출한다. URI 기반의 인용 관계 구축은 다음 예와 같이 이루어진다 (문헌 1이 문헌 2와 문헌 3을 참고 문헌으로 가지는, 즉 인용하는 경우). ('7010186243', '7110110372')와 ('6810332620', '7110110372')는 인용이 2번씩 일어나므로 인용 강도를 더 크게 하여 인

[6] 온톨로지 상에서 자원 (Resource)로 표현된다.

용 네트워크를 구성한다.

[문헌 1]
객체 URI: 'KISTI.PCD.0000001'
저자 URI: '7010186243', '6810332620'
[문헌 2]
객체 URI: 'KISTI.PCD.0000100'
저자 URI: '0000000083', '7320003992', '7110110372'
[문헌 3]
객체 URI: 'KISTI.PCD.0000200'
저자 URI: '7110110372'
[문헌 기반 인용 관계 쌍]
('KISTI.PCD.0000001', 'KISTI.PCD.0000100')
('KISTI.PCD.0000001', 'KISTI.PCD.0000200')
[저자 기반 인용 관계 쌍]
('7010186243', '0000000083')
('7010186243', '7320003992')
('7010186243', '7110110372')
('7010186243', '7110110372')
('6810332620', '0000000083')
('6810332620', '7320003992')
('6810332620', '7110110372')
('6810332620', '7110110372')

공저자 네트워크 (공저자 관계를 중심으로 구성한 연구자 네트워크)는 문헌 메타데이터로부터 직접적으로 구축하는 것이 가능하지만, 인용 네트워크의 경우 수작업으로 인용 정보를 구축하는 것이 현실적으로 어렵기 때문에 정보 추출 기술이 필요한 것이고 여기에 URI 관리 기법을 추가하여 정확한 인용 네트워크가 구성될 수 있도록 한다. 인력의 소속 정보 변경 등에 대해서도 인력 URI를 관리함으로써 인력의 소속 정보나 기타 관련 정보의 변경에 영향 받지 않고 일관성 있게 인력 중심 네트워크를 구성할 수 있다. 이와 같이 텍스트 처리 기술과 시맨틱 웹 기반 기술인 URI를 결합하여 정교한 객체 관계 속성을 추가로 자동 획득할 수 있음으로 해서 시맨틱 웹 응용 서비스를 다양화할 수 있게 된다.

4. 서비스 구현 결과

그림 4는 개념 사례화를 통해 온톨로지 인스턴스로 구축된 인력 URI와 이를 관리하는 URI 서버를 보여준다 [1] [6]. 각 인력 정보는 개체 확인을 거친 결과이며 여기에 3.1절에서 언급한 동명이인 해소 기법이 사용되었다. 전체 8,543명의 인력이 등록되어 있는데, '과학기술인력 종합정보시스템'에 등록되어 있는 인력이 1,332명으로 약 15.6%에 해당한다. 구축된 인력 정보를 검증하기 위해서도 동명이인 해소 기법이 사용되었는데 이때 성능 평가 수단으로 Rand Index [9]를 사용하였다. 성능[7]은 정답 집합 (인력 정보 구축 결과)과 대상 집합 (자동 인력 정보 검증 결과) 간의 모든 인력 쌍에 대해 일치 비율을 측정하는 방식으로 구해지는 데 본 연구 결과에 적용한 결과 약 0.925이다. 이 실험 결과를 수작업으로 다시 검증한 결과 103명 (약 0.46%)이 잘못 구축된 것으로 판단되어 이를 수정하였다. 이와 같이 텍스트 처리 기술은 수작업 구축 이전에 미리 정제 (인력 그룹핑)하고 수작업 구축 이후에 오류를 검증하는 등 온톨로지 인스턴스 구축의 효율성을 높일 수 있게 하는 이중적 수단으로서 그 가치를 가진다.

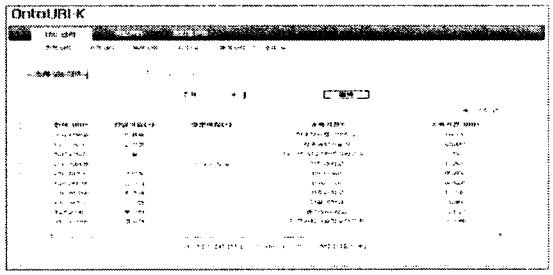

그림 4. URI 서버에서의 인력 정보 검색 결과

그림 5는 메타데이터 확장의 일환으로 구현된 주제·분야 할당의 적용 예를 보여준다. 입력된 문서로부터 색인어를 추출하고, 분야분류명이 매핑된 시소러스 개념어를 이용하여 주제와 분야를 할당한다. 현재 각 문서에 대해 상위 3개까지의 주제 및 분야를 할당한다. 정보가 많아질수록 일관성 있는 정보 분류가 필요하게 되며, 그림 하단과 같이 주제 및 분야 별로 유사 문서를 검색하거나 중복 과제를 검출하는 등의 다양한 응용에 적용될 수 있다. OntoFrame-K®에서는 추론 서비스의 각 세부 서비스의 범위를 제약하는 공통 제약 조건으로서 주제·분야를 활용한다. 즉, 특정한 주제나 분야에 해당하는 연구자를 검색하거나 (연구자 정보), 연구가 활발한 지역을 파악하거나 (연구 성과 맵), 성과물을 확인하거나 (성과 정보), 연구자 네트워크를 분석하거나 (공저자 네트워크, 인용 네트워크), 전문가를 추천하는 등 다양한 서비스를 위한 제약으로 사용된다.


그림 5. 주제·분야 할당 결과 예

[7] P = #Agreement/(#Agreement+#Disagreement)

그림 6은 문헌으로부터 자동 추출한 인용 정보를 이용하여 문헌 및 저자 인용 관계를 구축하고 시맨틱 웹 서비스를 통해 시각화한 연구자 네트워크를 보여준다. 상단의 공저자 네트워크는 추론 규칙에 의한 지식 확장을 통해 획득한 유도 객체 관계 속성에 의해 구현된 것이고, 하단의 인용 네트워크는 자동 정보 추출 기술을 이용하여 문헌으로부터 획득한 기본 객체 관계 속성에 의해 구현된 것이다. 전체 7,237개 문헌으로부터 해당 문헌 집합 내에서 인용이 이루어진 인용 정보만을 대상으로 하여 366쌍의 문헌 기반 인용 관계를 자동 구축하였다. 이를 저자 기반 인용 관계로 변환한 결과, 총 2,872쌍이 얻어졌는데, 이에 대해 인용 네트워크를 구성하여 135개의 인용 네트워크 그룹 (인용-피인용 관계로 연결된 그룹으로서 해당 그룹 내에서는 특정 인력에 대해 최소한 1번 이상 그룹 내 다른 인력과 인용-피인용 관계가 맺어진다.)을 획득하였다. 인용 네트워크를 구성하는 전체 저자 수 (URI 기준)는 827명이며, 가장 큰 인용 네트워크 그룹의 경우 12명의 구성원으로 이루어져있다.

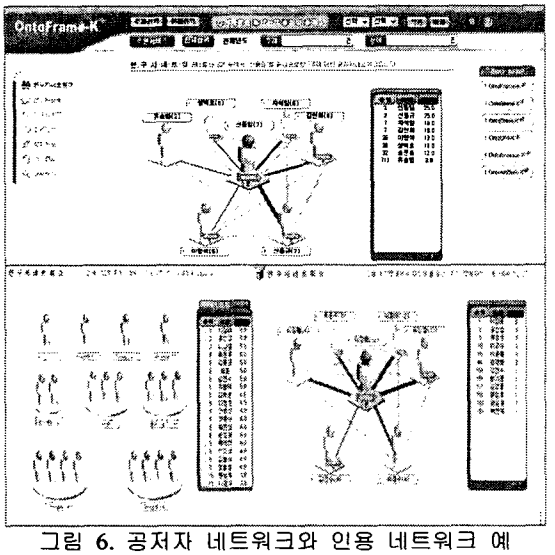

그림 6. 공저자 네트워크와 인용 네트워크 예

5. 결론

본 논문에서는 시맨틱 웹 기반 정보 유통 체계인 OntoFrame-K®에서의 기반 지식인 '국가 과학기술 R&D 기반 정보 온톨로지' 인스턴스 구축에서 텍스트 처리 기술이 어떻게 적용되었는 지를 개념 사례화, 메타데이터 확장, 객체 관계 속성 구축의 예를 통해 살펴보았다. 텍스트 처리 기술이 기존에 시맨틱 웹 기술과 밀접하게 다루어지지 못하고 별개의 기술로서 인식됨으로써 막대한 인적·물적 자원을 필요로 하는 온톨로지를 포함한 시맨틱 웹 자원 구축에 큰 도움을 주지 못하였다. 본 연구는 이러한 문제점을 개선하고자 여러 방향에서 텍스트 처리 기술 도입의 필요성을 본 논문을 통해 주장하였다. 특히 OntoFrame-K®를 실제 데이터, 텍스트 처리 기술, 시맨틱 웹 기술을 접목하여 구현함으로써 그 효용성을 실제적으로 보여주었다는 데 본 연구의 의의가 있다.

향후 메타데이터 수집과 가공에 보다 폭넓게 텍스트 처리 기술이 적용될 것으로 보이며, 본 연구 또한 이 부분에 많은 관심을 가지고 있다.

참고 문헌

[1] 구희관, 정한민, 강인수, 성원경, 이승준, 심빈구, "국가 과학기술 R&D 기반정보 온톨로지 구축을 위한 URI 관리 및 서비스 시스템 구현", 한국정보과학회 한국컴퓨터종합학술대회논문집, 2006.

[2] 안찬민, 박선, 박상호, 최범기, 이주홍, "분류 주제 자동 생성 및 동적분류체계 방법을 이용한 이메일 분류", 한국정보과학회 춘계학술대회논문집, 2004.

[3] 이승우, 정한민, 김평, 강인수, 성원경, "서지정보의 동명이인 구별을 위한 공저자 관계의 효용성 연구", 한국정보과학회 한국컴퓨터종합학술대회논문집, 2006.

[4] 이용배, 맹성현, "장르와 주제 범주간 용어 편차 정보를 이용한 디지털 문서의 장르기반 분류", 정보과학회논문지: 소프트웨어 및 응용 30(1), 2003.

[5] 이창범, 박혁로, "시소러스를 이용한 문서 자동 요약", 한국정보과학회 춘계학술대회논문집, 2001.

[6] 정한민, 강인수, 구희관, 이승우, 성원경, "URI 서버에 기반한 국가 R&D 기반정보 온톨로지 설계 및 구현", 정보관리연구 37(2), 2006.

[7] 정한민, 강인수, 성원경, "시소러스와 분야분류체계를 이용한 과학기술 문헌에의 주제 및 분야할당", 제7회 한국언어정보학회 하계학술대회논문집, 2006.

[8] 정호석, 임종태, 나혜숙, 민철호, "자동 문서 분류를 위한 분류 주제어의 자동 증식 방법", 한국정보과학회 추계학술대회논문집, 2000.

[9] 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", Proceedings of 13th International Conference on Knowledge Engineering and Knowledge Management, 2002.

[10] Tim Berners-Lee, "Putting the Web back in Semantic Web", http://www.w3.org/2005/Talks/1110-iswc-tbl/

[11] H. Jung "OntoFrame-K: Semantic Web-based Information Dissemination Platform", The 9th International Forum on Metadata Registry, 2006.

[12] http://www.korean.go.kr/000_new/50_dic_search.htm

[13] http//www.scopus.com/