온톨로지 구축 지원을 위한 과학기술 문헌으로부터의 인력정보 구축

Construction of Human Resources from Science & Technology Publications for Supporting of Ontology Construction

정한민, 이승우, 강인수, 성원경
한국과학기술정보연구원 NTIS 사업단

Jung Han-Min, Lee Seung-Woo, Kang In-Su, Sung Won-Kyung
KISTI

요약

시멘틱 웹의 발전은 정보의 규격화, 의미화를 통한 지식을 기본으로 요구하며, 온톨로지는 이러한 지식 표현을 위해 필수적으로 사용되는 도구이다. 온톨로지 상에서 Individual들은 식별체계인 URI(Universal Resource Identifier)를 이용하여 유일하게 지칭될 수 있어야 한다. 그렇지만, 현실에서 식별체계를 사용하는 경우를 발견하기가 힘들며, 특히 논문과 같은 과학기술 문헌은 그 적용 대상에서 제외되어 왔다. 이러한 이유로 인해 과학기술 문헌상의 인력정보를 식별체계 기반으로 구축하고자 하는 시도가 미약한 실정이었다. 이에 본 논문은 과학기술 문헌을 포함한 국가 과학기술 R&D 기반정보 온톨로지 구축에서 중심이 되는 인력정보를 URI 기반으로 구축하는 방법을 기술한다. 공저자 관계, 이메일, 출판 연도, 소속기관 등을 이용하여 동명이인 문제를 해소하고, 각 저자 그룹 별 URI 부여를 위해 국가과학기술인력 종합정보시스템을 활용한 사례를 소개한다. 이러한 과정을 통해 7,175건의 논문들로부터 획득한 인력정보는 연구자 네트워크 분석, 성과 통계 등 다양한 시멘틱 웹 응용 분야들에 기초적으로 활용될 것으로 기대한다.

Abstract

The development of Semantic Web basically requires knowledge induced from the formalization and semantization of information, and thus ontology should be introduced as a knowledgization tool. URI(Universal Resource Identifier) is an indispensible scheme to uniquely indicate individuals on ontology. However, it is difficult to find the use cases of URI in real data including science & technology publications. This paper describes the method to construct human resources based on URI which are the critical components on national R&D reference information ontology. We use co-authors, e-mails, publication date, and affiliation for discriminating authors with the same strings. HRST is referred to acquire human resource URIs. We expect the human resources would be adopted to outcome analysis applications such as researcher network analysis and outcome statistics.

I. 서 론

시멘틱 웹의 발전과 함께 온톨로지 구축의 중요성이 점점 강조되고 있는 현실에서, 온톨로지의 기반이 되는 식별체계인 URI(Universal Resource Identifier)가 필수적으로 요구되고 있다. 그렇지만, 이러한 중요성에도 불구하고 실험 데이터가 아닌 실제 데이터로부터 정보를 추출하여 정보 간 충돌을 해소하고 URI를 부여하고자 하는 시도가 외국의 몇몇 사례들을 제외하고는 찾아보기 힘든 형편이다[2]. [1]에서와 같이 인력정보에 대해 식별자 기반으로 서비스를 제공하는 곳이 있기는 하나 해당정보는 실 데이터로부터 가공된 것이 아닌 사용자가 직접 등록한 정보에 불과하다. 실 데이터로부터 발생할 수 있는 다양한 문제들을 해결할 수 있는 방안을 제시하지 않고서는 지속적인 정보 확장이 어려울 수밖에 없다.

기존에 논문으로부터 저자정보를 포함한 서지정보를 자동 추출하여 구축하는 연구가 있었다[5]. 그렇지만, 저자정보의 의미적 구축이 아닌 문자 인식 기반 문자열 추출에 초점이 맞추어졌을 뿐이다. 다만, 식별체계 구축 시스템 간의 연계에 대한 연구가 [3][4]를 통해 조금씩 이루어지고 있는데, 본 연구에서 과학기술 문헌에 부여하는 ID를 KOI(Knowledge Object Identifier) 기반으로 구축함으로써 이러한 연구 결과를 부분적으로 활용하고 있다.

II. 과학기술 문헌 서지정보 작성 및 가공

과학기술 문헌은 보고서, 지적재산권과 더불어 국가 과학기술 R&D 기반정보 중 성과정보를 구성하는 주요 요소이다. 본 장에서는 과학기술 인력정보 구축을 위한 기본 자료로서의 서지정보를 작성하고 가공하는 방법을 설명한다.

1. 과학기술 문헌 서지정보 작성

과학기술 문헌 서지정보는 제목, 저자정보(주저자, 소속기관, 소속부서, 이메일 등), 출처정보(문헌유형, 주최기관, 발행기관, 문헌명칭 등), 발행연도로 구성된다. 각 문헌마다 저자의 창작 특성이 다르기 때문에 일관성 있는 서지정보 확보를 위한 입력 지침이 필요하다. 예를 들어, 한글 저자명은 붙여 써야한다거나 한글 소속기관명과 소속부서명은 따로 구분하되 각각은 붙여 써야한다는 것 등이 있다. 특히, 학술대회 논문집에서는 소속정보를 기재하지 않거나, 이메일을 기재하지 않는 경우가 종종 발생한다. 소속정보나 이메일을 기재하더라도 공저자를 포함한 전체 저자들과 매칭되지 않는 경우도 발생한다. 이러한 경우 위첨자를 이용하거나 이메일의 이니셜을 분석하는 휴리스틱들을 사용하고 있으나 그 처리에 한계가 있을 수밖에 없다. 소속정보의 경우에도 "한국과학기술원", "과학기술원", "KAIST"와 같이 다양한 형태들이 존재할 수 있는데, 기관 URI를 이용하여 해결할 필요가 있다.

2. 과학기술 문헌 서지정보 가공

과학기술 문헌 서지정보 입력 후 효율적인 인력정보 구축을 위해 서지정보를 가공할 필요가 있다. 즉, 같은 저자명을 빈도순으로 정렬함으로써 구축 우선순위 결정과 상호 비교를 통한 용이한 저자 URI 해소가 가능해진다. 본 논문은 이러한 서지정보 가공을 형식 검증, 포맷 변환, 정보 정렬 과정으로 나눈다. 형식 검증에서는 공백 문자, 줄 바꿈 기호 등의 불필요한 문자들을 제거하고 불완전한 저자정보를 재확인한다. 포맷 변환은 출처 URI, 인력 URI, 기관 URI, 부서 URI 등을 용이하게 구축할 수 있도록 하는 정지 작업으로, 공저자 분석을 위해 공저자 목록 필드를 추가한다. 정보 정렬은 한정된 인적/물적 자원을 활용하여 효율적으로 동명이인 문제를 해소하고자 구축 목적에 맞도록 특정 필드 중심으로 정렬하는 작업이다.

```
ID, 논문 제목, 출처, 파일명,
파일 타입, 총 저자 수, 저자 순위, 주저자,
소속기관, 소속부서, 이메일, 공저자 목록
```

```
KISTI1.PCD.0003883                  // ID (KOI 생성 규칙 활용)
지능형로봇 환경에서의 질의처리
2005HCI                             // 출처
docsrc\논문(원문및텍스트)\2005HCI\원문\00219
pdf
5
1
정한민                              // 주저자
한국과학기술정보연구원  // 소속기관
차세대정보시스템연구실  // 소속부서
jhm@kisti.re.kr
정한민;선충녕;손주찬;성원경;박동인;
```

본 논문에서는 동일한 저자명이 많은 순으로 정렬함으로써 인력정보 구축의 효율성을 극대화하고자 한다. 다음은 가공된 서지정보가 가지는 필드들을 나열한 결과와 그 예이다. 이 중 URI 부여 대상은 ID, 출처, 저자명, 소속기관, 소속부서이다.

III. 저자 그룹 생성 및 인력 URI 부여

동명이인 문제 해소 후 부여하는 인력 URI는 주민등록번호와 같이 유일한 값이어야 한다. 본 연구에서는 신규로 인력 URI를 설계하는 대신에 이미 실제 시스템에서 활용되고 있는 식별자를 활용하고자, 국가과학기술인력 종합정보시스템에서 사용하고 있는 인력 ID를 인력 URI로서 사용한다. 또한, 해당 URI를 획득할 수 없는 미등록 인력에 대해서는 동일한 인력 ID 생성 규칙을 활용하여 가상 인력 ID를 생성하여 인력 URI로 사용한다.

1. 국가과학기술인력 종합정보시스템

국가과학기술인력 종합정보시스템 [1]은 국내의 여러 기관에 분산되어 구축 운영 중인 인력 DB를 연계하여 통합 메타 DB를 구축하고, 통합 검색 서비스 및 각종 현황정보 서비스를 제공하는 시스템이다. 현재 한국과학기술정보연구원(KISTI), 한국과학재단(KOSEF)을 포함하여 총 24개 기관이 참여하고 있다. 2006년 4월 현재 대학교, 연구소, 산업체 인력 등을 포함하여 약 323,000명이 등록되어 있다. 국가과학기술인력 종합정보시스템에서는 10자리로 구성된 고유 식별체계[1]를 인력 ID로 사용하고 있으며, 인명, 소장처 및 식별체계를 결합하여 해당 인력에 대한 상세정보에 바로 접근할 수 있도록 되어 있다. 예를 들어, 인명이 "정한민", 소장처가 "KISTI", 식별번호가 "7010186243"인 경우에 "http://hrst.or.kr/hrst/viewDetailFrameSet.jsp?koi=7010186243&korggubun=KISTI&kname=정한민"이라는 상세정보 URL을 자동 생성할 수 있다는 특징을 가진다. 본 연구에서는 이미 등록된 인력에 대해서는 국가과학기술인력 종합정보시스템의 식별번호를 그대로 URI화[2]하고, 미등록 인력에 대해서는 10자리를 유지하되 가상의 할당 영역에 해당할 수 있도록 "000"으로 시작하는 일련번호를 부여한다 (예를 들어, 첫 번째 미등록 인력에는 "0000000001"을 식별번호로 부여하므로, URI는 "http://www.kisti.re.kr/isrl#PER:0000000001"이 된다.).

1) 첫 두 자리는 연도를 의미하며, 세 번째 자리는 성별을 의미한다.
2) 인력 URI는 Namespace + Prefix + 식별번호의 조합으로 구성한다. Namespace로는 "http://www.kisti.re.kr/isrl#"를, URI 유형을 의미하는 Prefix로는 "PER:"을 사용한다.

2. 인력 URI 획득 및 생성

인력 URI 획득 및 생성 프로세스는 다음 프로세스와 같다. 가공된 과학기술 문헌 서지정보를 이용하여 저자 그룹들을 생성하고 각 그룹에 대해 국가과학기술인력 종합정보시스템을 참조하여 인력 URI를 부여하거나 신규 인력 URI를 생성하여 할당한다.

1. 공저자 분석 및 그룹화: 동명이인 문제를 해소하고자 하는 인력에 대해 공저자를 공유하는 인력들을 동일 인물로 간주하여 그룹화한다.
2. 이메일 분석 및 인력 그룹 병합: 두 인력 그룹이 같은 이메일을 하나 이상 공유하고 있는 경우 두 그룹을 하나로 병합한다.
3. '소속+연도' 분석 및 인력 그룹 병합: 두 인력 그룹이 동일한 연도에 같은 소속 (소속기관 AND 소속부서)을 공유하고 있는 경우 두 그룹을 하나로 병합한다. 단, 소속정보가 불완전한 경우에는 같은 소속이 아닌 것으로 간주한다.
4. 인력 그룹 병합 및 인력 URI 획득: 국가과학기술인력 종합정보시스템을 검색하여 경력정보 확인 등을 통해 소속 변경 등으로 두 인력 그룹이 동일하다고 판단하는 경우에는 해당 인력 그룹들을 병합하고, 각 인력 그룹에 대해 등록 인력 ID를 인력 URI로서 할당한다.
5. 인력 URI 생성: 국가과학기술인력 종합정보시스템에 등록되지 않은 인력이라고 판단하는 경우에 신규 인력 URI를 생성하여 할당한다.

다음은 이러한 프로세스에 의해 인력 URI를 해소한 결과의 예를 보여준다. "조성배"는 세 인력 그룹으로 구분되며, 두 그룹은 국가과학기술인력 종합정보시스템에 의해 인력 URI를 할당받고, 나머지 하나는 신규 인력 URI를 할당받는다.

ID	출처	저자명	인력 ID	소속	이메일	공저자 목록
KISTI1.PCD.0005989	2003KISSS[3]	조성배	6510145983	연세대학교 컴퓨터과학과	sbcho@cs.yonsei.ac.kr	한상준; 조성배;
KISTI1.PCD.0005998	2003KISSS	조성배	6510145983	연세대학교 컴퓨터과학과	sbcho@csai.yonsei.ac.kr	박찬호; 조성배;
KISTI1.PCD.0006014	2003KISSS	조성배	7510237363	한국전기연구원	sbcho@keri.re.kr	하현석; 황민태; 조성배; 이재조;
KISTI1.PCD.0000349	2002KISSF	조성배	6510145983	연세대학교 컴퓨터과학과	sbcho@cs.yonsei.ac.kr	한상준; 조성배;
KISTI1.PCD.0006161	2003KISSS	조성배	0000000007	순천향대학교 정보기술공학부	hopi@dkpower.com	김동균; 전병찬; 조성배; 이상정;
KISTI1.PCD.0007149	2003KISSS	조성배	0000000007	순천향대학교 정보기술공학부	hopi@dkpower.com	송재훈; 조성배; 이상정;
KISTI1.PCD.0007032	2003KISSS	조성배	6510145983	연세대학교 컴퓨터과학과	sbcho@cs.yonsei.ac.kr	민현정; 김경중; 조성배;

3) 2003KISSS: 2003년 한국정보과학회 춘계학술대회, 2002KISSF: 2002년 한국정보과학회 추계학술대회

3. 온톨로지 Property 활용

상기 프로세스에 의해 인력정보 구축 작업을 수행할 때 완전하게 동명이인 문제가 해소되지 않는 경우가 발생할 수 있다. 예를 들어, 하나의 인력이 소속을 변경하여 논문을 작성한 경우에는 공저자, 이메일, 소속+연도 모두에 대해 불일치할 수 있다. 또한, 국가과학기술인력 종합정보시스템에도 해당 인력의 경력이 제대로 업데이트되지 않았을 수 있다. 이러한 경우, 다른 인력으로 간주할 수밖에 없는데, 추후 경력정보가 업데이트가 되거나, 인력 홈페이지 등의 다른 루트를 통해 동명이인 문제가 해소되는 상황을 가정해야 한다.

본 연구에서는 이러한 가능성을 현 시점에서 해소하려고 하는 대신에 OWL(Web Ontology Language)기반의 온톨로지 Property를 이용하여 해소되는 시점에 처리할 수 있도록 한다. OWL Lite에서 제공하는 Equality Property들 중에서 'sameAs'는 서로 다른 이름을 가진 Individual들이 동일한 Resource를 지칭하고 있다는 의미를 부여한다. 이 Property를 이용하면 서로 다른 URI들을 가지는 Individual들을 연결할 수 있으며 비록 두 인력이 서로 다른 URI들을 가지더라도 추론 과정에서 동일 인력으로 처리하는 것이 가능하다.

IV. 구축 결과

인력정보 구축을 위해 수집/구축한 과학기술 문헌 종류 및 크기는 표 1과 같다. 구축 크기가 수집 크기와 다른 것은 중복 논문들이 존재하고, 원본을 얻을 수 없는 논문들을 배제했기 때문이다.

[표 1] 과학기술 문헌 구축 결과

학회명	연도	학술대회명	크기 (구축/수집)
한국정보과학회	2002	춘계학술대회	555/591
		추계학술대회	757/768
	2003	춘계학술대회	774/794
		추계학술대회	870/874
	2004	춘계학술대회	682/729
대한전자공학회	2003	하계학술대회	665/721
	2004	하계학술대회	419/421
	2005	추계학술대회	290/316
한국 HCI 학회	2003	학술대회	241/253
	2004	학술대회	312/318
	2005	학술대회	326/346
	2006	학술대회	369/383
한국정보처리학회	2004	추계학술대회	484/485
	2005	춘계학술대회	431/432
총합			7175/7431

중복을 포함하여 전체 저자 출현 회수는 23,105건 (논문 당 평균 공저자수는 약 3.22명)이며, 한 논문의 최대 공저자 수는 17명이다. 동명이인을 포함하여 동일 이름 저자의 최대 출현 회수는 55회이며, 4,903명은 1회만 출현한 저자들이다. 동명이인 문제 해소에 중요한 단서인 소속과 이메일을 분석해보면, 소속이 기재되지 않은 경우가 27건이며, 이메일이 기재되지 않은 경우가 5,562건이다. 즉, 이들에 대해서는 좀 더 신중한 인력정보 구축 및 검증 과정이 필요하다는 것을 알 수 있다. 본 구축에서는 동명이인 문제 해소를 위해 2번 이상 성과물에서 나타난 인력들을 대상으로 하였다. 3.2장에서 정의한 인력 URI 획득 및 생성 프로세스를 통해 구축한 결과는 다음과 같다.

내용	건수
성과물 수 (= 동명이인 수)	18,199
Unique 성과물 수	6,826
동일 인명 수	4,369
Unique 인력 수	6,977
국가과학기술인력 종합정보시스템 등록 인력 수 (Unique 인력 수 - 신규 URI 할당 인력 수)	1,176

V. 결 론

본 연구는 실제 과학기술 문헌을 대상으로 온톨로지에서 기본적으로 요구되는 URI 기반의 인력정보를 구축하는 방안을 설명하였다. 이러한 URI 기반 인력정보는 의미 기반 정보검색, 추론 서비스와 같이 문자열만으로는 해결할 수 없는 고급 응용 서비스 분야에서 중요한 기반정보로서의 역할을 할 것으로 기대한다. 앞으로 개인 홈페이지를 포함한 웹 페이지 등을 참조하여 보다 정교한 인력정보 구축이 이루어질 수 있도록 할 예정이다.

▌참 고 문 헌▌

[1] http://www.hrst.or.kr/hrst/index.jsp
[2] http://www.aktors.org/technologies/csaktivespace/
[3] 신동구, 김재수, 윤정모, 권이남, 전성진, 정택영, 식별체계 간 연계시스템 구축에 관한 연구, 한국정보과학회 추계학술대회, 2005.
[4] 이상환, 신동구, 김재수, 최진영, 정택영, 식별체계 기반의 과학기술분야 전자원문 연계시스템 설계 및 구현, 한국정보과학회 춘계학술대회, 2004.
[5] 장대근, 지수영, 오원근, 김의정, 논문 첫 페이지 영상의 논리적 구조형성을 위한 제목, 저자, 요약 영역의 추출, 한국정보처리학회 추계학술대회, 1998.