글로벌 리포지터리 정보 수집모듈 개발 및 검증

A Development and Verification on a Harvest Module for Global Repository Information Data

김 선 태, 최 명 석
한국과학기술정보연구원

Suntae Kim, Myung-Seok Choi
Korea Institute of Science and Technology Information

요약

본 연구의 목적은 글로벌 리포지터리의 현황 정보를 수집하기 위한 모듈을 개발하고 이의 완전성을 검증하는데 있다. OpenDOAR와 ROAR 레지스트리에 등록되어 있는 리포지터리 정보를 수집하기 위해서 각각의 서비스에서 제공되는 API와 웹페이지에 대한 분석을 진행 하였다. 개발 모듈로는 대륙별 국가정보 수집모듈, 리포지터리 소프트웨어 정보 수집모듈, 리포지터리 주제 정보 수집모듈, 리포지터리 상세정보 수집모듈을 개발하였다. 수집모듈로 수집된 리포지터리 정보를 대상 정보원의 데이터와 비교하는 방법으로 모듈의 완전성을 검증하였다.

Ⅰ. 서론

1. 연구배경 및 목적

국내에서도 글로벌 리포지터리의 동향을 지속적으로 분석할 수 있는 플랫폼 구축이 필요하다. 플랫폼을 통해 제공되는 다양한 통계데이터는 학술커뮤니티를 지원하는 정부나 연구비 지원기관의 정책수립에 활용될 수 있을 것이다.

1.1 연구배경 및 목적

본 연구를 통해 개발된 리포지터리 정보 수집모듈은 향후에 글로벌 리포지터리 검색 플랫폼 구축 시 핵심엔진으로 사용될 수 있을 것이다.

1.2 연구방법 및 선행연구

수집모듈 검증은 OpenDOAR와 ROAR에서 제공하는 데이터와 모듈을 통해 수집된 데이터를 비교하는 방법을 사용하였다. 선행연구를 검토한 결과, 리포지터리에 대한 현황 조사를 위해서, OpenDOAR와 ROAR가 제공하는 메타데이터의 단순 분석 연구가 주를 이룬다. 또한 연구대상 리포지터리를 선정할 때 OpenDOAR와 ROAR가 핵심 정보원으로 활용되었음을 알 수 있다[1]. 하지만, 글로벌 리포지터리 정보를 수집하여 수집데이터를 활용하는 연구 사례는 존재하지 않는다. Sahu et al. (2013)도 OpenDOAR와 ROAR가 전 세계 리포지터리들의 최신 정보 제공 한다 주장하며 자신의 연구에 정보원으로 사용한바 있다. 수집모듈 검증은 OpenDOAR와 ROAR에서 제공하는 데이터와 모듈을 통해 수집된 데이터를 비교하는 방법을 사용하였다[3].

Ⅱ. 본론

2. 모듈개발

수집모듈 개발은 대한민국, 중국, 일본의 리포지터리 현황을 비교 분석하기 위해 Kim(2014)이 개발한 소스 코드를 수정 보완하는 방식으로 진행하였다[2]. 개발은 윈도우즈 기반의 Eclipse 환경에서 개발을 하였다. 자바 언어를 사용하였으며, 가상 머신은 호환성을 위해서 1.7 버전을 사용하였다. 데이터베이스는 범용적으로 사용되는 MySQL 오픈소스를 사용하였다.

2.1 대륙별 국가정보 수집 모듈

DOAR에서는 대륙별 국가정보를 별도로 수집하기 위한 API를 제공하고 있지 않다. 서비스 분석 결과 다음과 같이 'cContinent' 파라미터에 대륙의 이름을 파라미터로 전달하게 되면 해당 대륙에 속한 국가와 국가코드 정보를 획득 할 수 있다.

2.2 리포지터리 소프트웨어 정보 수집 모듈

ROAR에서는 리포지터리 소프트웨어와 코드 정보를 별도로 수집하기 위한 API를 제공하고 있지 않다. 서비스 분석 결과 'http://roar.eprints.org/cgi/schema' 문서의

'subject.eprint.software' 이름속성 값을 이용하여 ROAR에서 관리하는 소프트웨어와 소프트웨어코드 값을 획득할 수 있다.

2.3 리포지터리 주제 정보 수집 모듈

ROAR에서는 리포지터리 주제 정보와 코드 정보를 별도로 수집하기 위한 API를 제공하고 있지 않다. 서비스 분석 결과, 소프트웨어 정보를 수집했던 것과 동일하게 'http://roar.eprints.org/cgi/schema' 문서의 'subject.eprint.subjects' 이름속성 값을 이용하여 ROAR에서 관리하는 주제와 주제코드 값을 획득 하였다.

2.4 리포지터리 상세정보 수집 모듈

OpenDOAR와 ROAR 레지스트리를 대상으로 리포지터리정보 수집을 위한 핵심 모듈은 공통된 프로세스로 작동한다. SAX, DOM 구문을 분석하기 위한 표준화된 API를 사용하였다. SAX 구문 분석기 객체의 환경 설정 및 생성을 위해서 제공되는 'DocumentBuilderFactory' 추상클래스를 사용하였다. 객체 성성 시 시스템 프로퍼티로 지정된 팩토리 클래스 객체를 생성하여 사용하였다. 디폴트로 생성된 팩토리 객체를 이용하여 XML 문서로부터 DOM 트리 객체를 생성하는 'DocumentBuilder' 추상클래스 객체를 사용하여 XML 문서를 읽어서 리포지터리 정보를 수집하였다.

3. 리포지터리 정보 수집모듈 검증

OpenDOAR 데이터 수집 모듈을 검증하기 위해서, 국가별로 등록된 모든 리포지터리 정보를 수집하였다. 'AD'는 등록된 첫 번째 국가코드를 의미하며, 맨 하단의 "는 등록된 마지막 국가코드를 의미한다. DOAR 데이터 수집 모듈 작동 시 국가코드별 리포지터리 정보 획득에 사용되는 쿼리는 다음과 같다.

http://www.opendoar.org/api13.php?show=max&co=AD
http://www.opendoar.org/api13.php?show=max&co=AE
http://www.opendoar.org/api13.php?show=max&co=AF
〈중간 생략 ...〉
http://www.opendoar.org/api13.php?show=max&co=ZW

국가코드 'GT(과테말라)'와 'MC(모나코)'는 OpenDOAR 서버에서 보내준 XML 문서에 오류가 있어서 해당 국가의 리포지터리 정보는 수집되지 않았다. ROAR 데이터 수집 모듈 검증을 위해서는 ROAR에서 제공하는 'rawlist.xml' 파일을 사용하였다.

Result Grid | Filter Rows:

대륙명	리포지터리수	퍼센트
Africa	134	4.4210
South America	266	8.7760
Asia	605	19.9604
Australasia	66	2.1775
Caribbean	15	0.4949
Central America	16	0.5279
Europe	1352	44.6057
North America	574	18.9376
Oceania	3	0.0990

▶▶ 그림 6. 대륙별 리포지터리 분포 현황
(2016.2.18. 현재) - 수집모듈 제공

2016년 2월 18일 현재, ROAR에 등록된 리포지터리 개수는 3532개이다. 본 연구에서는 ROAR에서 제공하는 'rawlist.xml' 파일을 분석하여 데이터를 구축하였다. 구축된 데이터는 총 4171개의 레코드로 현재의 ROAR에서 제공되는 리포지터리 개수와 차이가 발생하였다. 본 연구를 통해 개발된 ROAR 데이터 수집모듈은 ROAR에 등록되었던 과거의 모든 리포지터리 정보를 수집하는데 의미가 크다. 하지만 수집된 데이터와 현재의 ROAR 데이터의 일관성을 검증할 수는 없었다.

Ⅲ. 결론

글로벌 리포지터리 정보를 수집하는 모듈을 개발하고 실제 수집된 데이터를 분석하여 모듈을 검증하였다. 수집정보원으로는 OpenDOAR과 ROAR를 사용하였다. 대륙별 리포지터 수량을 보여줌으로써 OpenDOAR에서 제공하는 데이터와 수집모듈에서 제공한 데이터가 일관성이 있음을 증명하였다. ROAR의 경우, 과거에 데이터를 제공하는 정보원을 사용하여 수집데이터의 일관성을 증명하지 못했다. 하지만 향후에 현재의 ROAR시스템에서 데이터를 제공하는 서비스를 분석하여 질의어 패턴을 제시하였다.

■ 참 고 문 헌 ■

[1] Bhat, Mohammad Hanief. 2010. "Interoperability of open access repositories in computer science and IT - and evaluation", Library Hi Tech. Vol. 28 No. 1. pp.107-118

[2] Kim, Suntae. Lee, Wongoo. 2014. "Global data repository status and analysis: based on Korea, China and Japan", Library Hi Tech, Vol. 32 Iss: 4, pp.706 - 722

[3] Sahu, Surendra Kumar. Arya, Satish Kumar. 2013. "Open access practices in India", Library Hi Tech News Number 4 pp.6-12