PubMed 메타데이터 수집기 및 동적 RSS 채널 생성기 개발 연구

A Study on PubMed meta-data aggregator and dynamic RSS channel generator development

김선태, 김재훈, 한희준, 현미환, 이태석, 예용희, 여일연, 윤희준
한국과학기술정보연구원

Kim sun-tae, Han hee-jun, Kim jay-hoon, Yae yong-hee
Korea Institute of Science and Technology Information

요약

PubMed는 의학 분야의 독보적인 데이터베이스이다. 본 논문에서는 PubMed에서 제공하는 특정시점의 메타데이터 정보를 수집·구축하였으며, 이를 통해 웹2.0의 핵심 서비스인 RSS 채널을 동적으로 생성하는 서비스모델을 설계·구현한다. PubMed에서 제공하는 E-Utilities서비스(API)를 분석하였으며, 메타데이터 수집·구축기를 개발하였다. 또한 동적 RSS채널 생성기를 개발하였으며, RSS리더기를 이용한 활용방안을 제시하였다.

Abstract

PubMed stands alone among database related to biomedical literature. On this thesis, PubMed meta-data at some point is collected and built. The service model which serve the dynamic RSS channel, one of the core services in an web2.0 era, is designed and developed. E-Utilities service(API) served by PubMed are analyzed. The meta-data aggregator and builder are developed. Also the dynamic RSS channel generator is developed and the method how it could be used in the RSS reader is suggested.

1. 서 론

1.1 연구 목적

PubMed는 미국 국립 보건원 (National Institutes of Health, NIH) 산하, 미국립의학도서관 (United States National Library of Medicine, NLM)의 부서로써 국립 바이오테크놀로지 정보센터 (National Center for Biotechnology Information, NCBI)에서 서비스하고 있는 여러 데이터베이스 중 하나이다. 일반에게 무료로 공개하고 있는MEDLINE을 위한 검색도구로 사용되고 있다. 검색 대상으로는 1950년도부터 2007년도까지 약 5,000 종 저널, 1천 7백 9십만 건 이상의 의·생명과학 분야의 논문 정보이다.

국내의 대표적인 학술정보 서비스인 KISTI의 NDSL(National Digital Science Library)은 KISTI Service 2.0 프로젝트를 통해 PubMed의 논문 메타데이터를 확충 하였다. 과학기술분야에 집중되었던 데이터베이스가 의·생명분야까지 확대되면서 명실상부한 학술데이터베이스로서 한걸음 도약했다할 수 있다.

지속적인 PubMed 메타데이터를 수집하기 위해서는 특정 시점으로부터 부분적인 메타데이터 수집이 가능해야 하기 때문에 이에 대한 연구가 필요하며, KISTI에서 제공하는 RSS서비스는 의·생명과학 분야로 특화된 서비스는 제공하지 못하고 있으므로, 이를 지원할 수 있는 채널개발이 필요하다.

본 연구에서는 특정 시점으로부터 부분적인 메타데이터를 수집하는 방법을 분석하였으며, 이를 토대로 수집·구축기를 개발하였다. 또한 수집된 데이터를 대상으로 동적인 RSS 피드를 생성하는 채널생성기를 개발하였다.

1.2 연구의 방법 및 절차

PubMed에서 제공하고 있는 E-Utilities를 분석하여, 특정 시점으로부터 부분적인 메타데이터 수집이 가능한 방법을 조사하였다.

국내·외 RSS 개발 및 서비스 수준을 조사하였으며, PubMed메타데이터를 이용해 서비스 할 수 있는 RSS 엘리먼트를 선정하였다.

2. PubMed E-Utilities 분석

2.1 E-Utilities(API)분석

PubMed에서는 E-Utilities라는 메뉴를 통해서 활용할 수 있는 API에 대한 상세한 설명을 제공하고 있다. 제공되는 핵심 API로는 <표 1>과 같이 EInfo, ESearch, EPost, ESummary, EFetch, ELink, EGQuery, ESpell과 같이 일반적인 웹 쿼리 인터페이스로 제공된다. 또한 SOAP기반으로도 서비스를 제공하고 있다.

[표 1] PubMed 제공 주요API

API	설명
EInfo	필드인덱스 용어 카운트 및 DB링크제공
ESearch	**식별자(ID)리스트 검색**
EPost	식별자(ID)리스트 파일 업로드
ESummary	**식별자(ID)리스트로부터 요약문 검색**
EFetch	**식별자(ID)리스트로부터 순차검색**
ELink	관련 레코드 검색
EGQuery	Entrez데이터베이스 대상 검색건수제공
ESpell	스펠링 제안

모든 API의 파라미터는 순서가 무시되며, 부정확한 파라미터와 널(null)값은 무시된다. 파라미터로 '&tool'과 '&email'의 사용이 권고되는데 이것은 이용자가 전송한 쿼리를 식별하는데 사용된다. '&tool'의 값으로는 쿼리를 전송하는 시스템의 이름을, '&email'의 값으로 전송자의 이메일 주소를 넣는다.

PubMed에서는 히스토리 서버를 운영하는데, 이는 검색결과(식별자PMIDs)리스트를 임시로 보관하는 서버로서 정보검색에 용이함을 주기 위함이다. 히스토리 서버는 ESearch등의 반환 값 중 하나인 쿼리 키를 참조하는데 이때 사용하는 '#'은 '%23'으로 인코딩 되어야 한다. 각각의 API내용을 살펴보면 아래와 같다. EInfo는 특정데이터베이스의 상세정보(필드 및 링크정보)를 제공한다. 주어진 데이터베이스에 대한 정보를 제공하며, 데이터베이스의 인덱싱 필드 리스트와 다른 데이터베이스들과의 링크를 제공한다.

ESearch의 경우 'WebEnv' 파라미터만 제외하고 모두 소문자를 사용해야 하며, 구성되는 파라미터는 아래 <표 2>와 같다. ESearch는 실행한 결과 값(UIDs)을 히스토리 서버에 올릴 수 있다.

[표 2] ESearch 파라미터 활용방법

파라미터	활용 방법
term	검색어를 입력하며, 불리언 연산자 포함가능
field	검색필드 지정 (예: field=mesh)
reldate	오늘 날짜를 기준으로 Entrez Date값이 경과된 일자에 포함된 레코드 검색
mindate	검색결과를 날짜 필드의 값으로 제한할 경우 사용
maxdate	
datetype	날짜로 검색결과를 제한 할 경우 대상이 되는 날짜 필드명을 지정 함
retstart	디폴트값이 0으로 첫 번째 레코드의 순차번호를 말함
retmax	최대 검색건수 지정
retmode	검색결과 유형 예: retmode=xml
rettype	검색 유형 예: rettype=count rettype=uilist (default)
sort	Web Environment 와 함께 사용하여 ESummary and EFetch의 결과값 소팅
db	EInfo에서 제공하는 데이터베이스 중 하나의 이름
db	EInfo에서 제공하는 데이터베이스 중 하나의 이름
usehistory	이용자의 환경에 검색결과를 유지하는 것을 요구할 때 사용
WebEnv	ESearch 혹은 EPost 이후 XML결과 값 안에 리턴 되는 값으로서 ESearch 혹은 EPost을 호출시 매번 값이 변경된다. WebEnv가 사용되면, 검색히스토리숫자(History search numbers)가 ESummary URL에 포함될 수 있다.
query_key	검색히스토리숫자

EGQuery는 ESearch의 글로벌 버전으로서 모든 데이터베이스를 대상으로 동시에 검색을 진행한다.

EFetch는 UIDs를 입력 값으로 받고 정형화된 결과 값을 반환한다. 예를 들어 PubMed 데이터베이스로 부터 초록을 만들어 내고, 디스플레이 포맷을 위한 기능(함수)는 EFetch에 의해 수행되고, 연결된 레코드들의 기능은 ELink에 의해 수행된다.

ELink는 UIDs를 입력 값으로 받고 Entrez의 다른 데이터베이스와 관련된 UIDs 리스트를 결과 값으로 만들어 낸다.

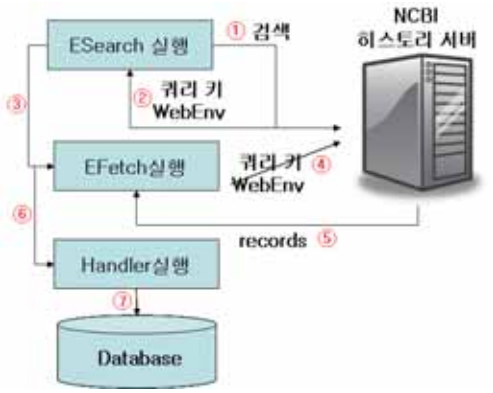
UIDs 리스트를 히스토리 서버에 저장해서, 다음 쿼리와 조합하거나 재조정하여 사용할 수 있는데, EPost는 히스토리 서버에 UIDs 리스트를 업로드 할 수 있는 기능을 제공한다. 쿼리 키(query key)와 WebEnv 값을 리턴 해 준다.

EPost나 ESearch의 결과 값인 쿼리 키와 WebEnv는 ESummary, EFetch, ELink의 UIDs 리스트로 대체될 수 있다. 이것은 대량의 레코드를 처리하는데 아주 유용한 기능이다.

본 논문에서는 ESearch와 EFetch를 이용한 데이터 수집방법을 사용하였다.

3. PubMed 메타데이터 수집·구축기 개발

E-Utilities API 중 특정 시점으로부터 부분적인 메타데이터 수집이 가능한 쿼리인 ESearch를 사용하여, 쿼리 키와 WebEnv값을 얻은 후 EFetch를 사용하여 관련 레코드를 수집하였다. 주요로직은 아래 <그림 4>와 같다. 그림에서 ④, ⑤, ⑥, ⑦프로세스는 핸들러에서 지속적으로 반복된다.



▶▶ 그림 4. PubMed 메타데이터 수집·구축 프로세스

ESearch는 현재 시점을 기점으로 특정 시점까지의 레코드를 검색한다. 사용한 기본URL은 아래와 같다.

• http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?

위 기본 URL에 ESearch에서 사용가능한 파라미터를 붙여 최종 완성된 URL은 아래와 같다.

- http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=&reldate=10&datetype=edat&retmax=100&usehistory=y&tool=KISTI_stkim&email=stkim@kisti.re.kr

파라미터 reldate의 값을 10으로 한 이유는, PubMed에 요청을 전송하는 시점에서 10일 이전부터 PubMed에 구축된 레코드를 검색하겠다는 것을 나타낸다. retmax는 100건씩 검색을 결과를 받겠다는 것을 의미하며, usehistory는 PubMed의 히스토리 서버를 사용하겠다는 것을 의미한다.

아래 XML은 ESearch에 대한 응답 내용중 중요한 엘리먼트만 나열한 것으로서, <Count>는 전체 검색건수, <RetMax>는 전달된 ID개수, <QueryKey>와 <WebEnv>는 히스토리서버를 사용할 수 있는 값, <IdList>는 검색된 레코드의 PMID리스트를 의미한다.

```
<Count>21948</Count>
<RetMax>100</RetMax>
 ...
<QueryKey>1</QueryKey>
<WebEnv>0NkFC7Na...DNkFC7</WebEnv>
<IdList>
        <Id>18421827</Id>h
        .....
        <Id>18421826</Id>
</IdList>
```

EFetch의 기본 URL은 아래와 같다.

- http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?

위 기본 URL에 EFetch에서 사용가능한 파라미터를 붙여 최종 완성된 URL은 아래와 같다.

- http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&WebEnv=0bmnU-rke3bp6gK6A0PAavIKURnC_m1tIq3NN_7sCF8d0tq31XJuQFwYOg@1EDF157E80C79AB0_0033SID&query_key=1&retstart=0&retmax=10&retmode=xml&tool=KISTI_stkim&email=stkim@kisti.re.kr

위 URL에선 주목할 것은 WebEnv와 query_key 파라미터이다. 이것은 히스토리 서버에서 사용하기 위한 킷값으로 ESearch 쿼리를 전송 후 전달받은 값을 그대로 사용해야 한다.

핸들러는 xerces를 활용한 SAX API를 사용하였다. 대량의 데이터를 파싱하는데, 메모리 사용문제 및 처리속도를 고려하여 DOM 보다는 SAX를 활용 하였다.

4. 동적 RSS 채널생성기 개발

RSS는 'Really Simple Syndication'의 약자로서 웹 콘텐츠를 배급/배포하기 위한 포맷이다. RSS는 XML로 표기된 데이터이며, 모든 RSS 파일은 반드시 W3C 에서 정의한 XML1.0 규격을 준수하여야 한다.[3] RSS는 현재 블로그, 뉴스, 기업정보, 사이트공지사항, 취업정보, 쇼핑 정보 등과 같이 주기적으로 자주 업데이트 되어지는 콘텐츠들에서 RSS를 이용해 손쉽게 한 곳에서 편하게 제공 중이며, 각 종 RSS reader기를 통해 피드를 받아 볼 수 있다.

본 논문에서는 구축된 PubMed 메타데이터를 대상으로, 전송된 RSS 채널생성 호출에 응답하여 피드를 생성하는 모듈을 개발하였다. 이것은 서비스를 위한 피드를 미리 생성하는 것이 아닌 최신 피드를 요청 시 생성하여 제공하기 위함이다.

4.1 RSS 2.0 엘리먼트 선정 및 채널생성기 개발

RSS 2.0의 두 번째 Depth 엘리먼트는 총 20개 인데, 이중에서 <표 3>과 같이 학술정보 서비스 있어 필요한 엘리먼트 8개만 선정하였다. 제공되는 피드의 목적이 PubMed 데이터베이스에 추가되는 신규 메타데이터에 대한 간략정보를 제공하고, 실제 상세정보 및 원문(full text)은 PubMed에서 제공한 식별자(PMID)를 이용해 연계하는 것이기 때문에, RSS 2.0에서 사용하는 모든 엘리먼트를 사용하지 않았다.

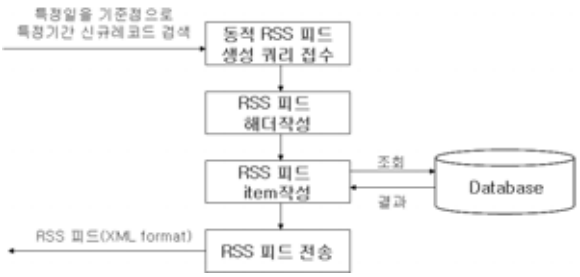
'channel - title'부터 'channel - generator'까지는 서비스 되는 피드에 대한 설명(제목, 제공자정보, 기술정보 등) 부분이며, 세 번째 계층부터는 제공되는 메타데이터 한건 한건에 대한 세부정보를 기술하는 부분이다.

'channel - item - description'엘리먼트에는 초록정보를 담았으며, 'channel - item - link'엘리먼트에는 PubMed 상세화면을 호출할 수 있는 URL과 PMID의 조합된 문자열을 값으로 사용하였다. 'channel - item - guid' 엘리먼트는 PMID를 값으로 갖는다. 'channel - item - pubDate'은 피드로 제공되는 메타데이터가 PubMed에 생성된 날짜를 값으로 갖는다.

[표 3] 선정된 RSS 엘리먼트 리스트('-' 은 Depth를 나타냄)

RSS 2.0 선정 엘리먼트
channel - title
channel - link
channel - description
channel - copyright
channel - managingEditor
channel - webMaster
channel - generator
channel - item - title
channel - item - link
channel - item - description
channel - item - author
channel - item - guid
channel - item - pubDate

채널생성기는 <그림 5>와 같이 전송된 쿼리를 이용하여 RSS 피드를 생성한다.



▶▶ 그림 5. 채널생성기 프로세스

5. 결 론

웹 환경에서 운영되는 다양한 학술정보서비스는 메타데이터 정보를 공개하고 수집하기 위해 다양한 방법을 사용한다. 대표적인 방식에는 OAI프로토콜을 통한 배포자(Provider)와 수집자(Harvester) 간 커뮤니케이션 방식, OpenAPI 쿼리 및 XML결과값 전송방법인 REST방식, 이 기종 서버 간 SOAP을 기반으로 쿼리 및 결과 값을 전송하는 방식 등이 있다.

본 연구에서는 REST방식을 사용하여 PubMed 데이터베이스를 대상으로 특정일을 기준점으로한 특정기간의 신규레코드를 수집하였다. NCBI의 히스토리 서버를 활용하였으며, ESearch와 EFetch API를 사용하여 PubMed 메타데이터 수집·구축기를 개발하였다.

동적 RSS 피드 생성기를 개발하여 실시간으로 요청되는 쿼리에 의해 피드를 생성할 수 있도록 구현 하였다. 동적 RSS 채널 생성기로부터 전달받은 피드를 XSL을 이용해 브라우저에서 확인한 화면은 아래 <그림 6>과 같다.

▶▶ 그림 6. XSL을 이용한 RSS 피드 검증 및 PubMed 연계

한국과학기술정보연구원(KISTI)에서는 2007년 8월부터 시작된 KISTI Service 2.0 프로젝트를 통해 PubMed메타데이터를 NDSL 학술 데이터베이스에 확충하였다. 본 연구를 통해 개발된 특정시점의 PubMed 메타데이터 수집기를 활용하여 지속적인 메타데이터 수집이 가능할 것이다. 또한 동적 RSS 채널 생성기를 통해 의·생명분야의 신착자료 피드를 제공할 수 있을 것이다.

▌참 고 문 헌▐

[1] Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils) [2008.1.14] http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/chapter_eutils.pdf

[2] Creating a Web Link to the Entrez Databases [2008.1.14] http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helplinks.chapter.linkshelp

[3] RSS Tech notes & History[2008.4.21] http://cyber.law.harvard.edu/rss/index.html

[4] NCBI소개[2008.4.21] http://file.blog.empas.com/attachfile_download.php?a=27677308&seq=10233309