가용자원을 활용한 해외학술정보 데이터베이스제작방법에 관한 연구

노 경 란, 권 오 진
한국과학기술정보연구원 정보콘텐트개발실

A Study on Construction Method of Foreign Scientific Database by Utilizing Available Information Resources

Noh Kyung Ran*, Kwon Oh Jin
KISTI, S&T Development Dept.
E-mail : infor@kisti.re.kr

요 약

전통적인 해외학술정보 데이터베이스 제작시스템은 자료의 수집에서부터 DB로딩에 이르기까지 여러 문제들을 내포하고 있다. 본 논문에서는 분산되어 있는 가용자원들로부터 에이전트 기술을 이용하여 필요로 하는 메타데이터를 효율적으로 수집하고 활용하는 데이터베이스 제작모델을 설계하고자 한다. 본 논문은 해외학술정보 데이터베이스를 구성하는 수록정보에 대해 기술하고, 이들 수록정보와 연계된 가용자원들에 대해 살펴본다. 분산되어 있는 가용자원으로부터 메타데이터 구성에 필요한 기본요소를 수집하고, 이 메타데이터에 가공정보를 추가하고, 메타데이터의 품질을 검증하는 과정에 대해 기술한다.

Abstract

There are many problems in conventional database construction of foreign scientific journal from material aquisition to DB loading. This paper's purpose is to design database construction model which utilizes available information resources scattered several locations and uses agents technology to gather essential metadata efficiently. This paper describes component informations of foreign scientific database and related available resources. And it describes a process of DB construction that include metadata gathering method, automatic metadata classification method, and metadata quality monitoring method.

I. 서론

오늘날은 변화의 내용과 방향을 파악하지 못한 채, 과거의 경험이나 지식에 의존한 데이터베이스 제작방법으로는 더 이상 살아남기 힘든 시대가 되었다. 하루에도 무수히 많은 종류의 정보가 대량으로 유입되고 있지만 이를 체계적이고 일관되게 관리할 수 있는 데이터베이스의 미비로 효과적으로 정보를 활용하고 있지 못하고 있는 실정이다. 통합된 데이터베이스 체계를 이용하여 필요로 하는 데이터를 신속하게 수집하여 관리활용하기 위해서는 무엇보다도 필요로 하는 데이터를 수집하

는데 어떠한 가용자원을 활용할 수 있는지 파악하고, 데이터베이스를 설계, 구축하는 것이 필요하다.

본 논문의 목적은 전통적인 해외학술정보 데이터베이스 제작방식이 기존 여러 데이터베이스들간 연계성을 고려하지 않고 폐쇄적이고 독립적으로 제작됨에 따라 발생한 여러 문제들을 해결하기 위해, 공개된 환경에서 가용자원을 활용하여 데이터베이스를 제작하는 방법을 기술하고자 한다.

본 논문에서는 해외 학술정보 데이터베이스 구축시 필요한 구성요소들에 대해 살펴보고, 가용자원을 최대한 활용한 학술정보 데이터베이스 제작방법에 관해 설명한다. 공개된 환경에서 메타데이터 생성에 필요한 기본 정보를 추출하며, 추출된 정보를 필터링하고 변환함으로써 기본 메타데이터를 작성한다. 이 데이터에 부가가치를 부여하는 가공정보를 추가한 후 일련의 검증과정을 거쳐 양질의 데이터베이스를 구축한다.

II. 해외학술정보 데이터베이스의 구성

해외 학술정보 DB는 학술지 수록기사에 대한 메타데이터로 구성된다. 메타데이터의 구성요소는 학술지에 대한 기본정보, 학술지 수록기사에 대한 기본정보와 가공정보로 구성된다. 각 구성요소별 수록정보는 다음과 같다.

표1. 해외학술정보 데이터베이스의 수록요소

학술지 기본정보
학술지명, ISSN, 발행국, 자료유형, 학술지 분류정보
학술지수록기사 기본정보
기사명, 저자명, 수록권호, 페이지, 발행일자, 수록언어
학술지수록기사 가공정보
한글기사명, 기사분류정보, 색인어, 초록, 자료소장정보 등

해외학술정보 DB를 구축할 때 사용할 수 있는 가용자원으로는 학술지목록 DB, 학술지체크인 DB, 학술지목차DB, 그리고 보유중인 기타 CD-ROM 자원과 네트워크를 이용해 구할 수 있는 공개정보들이 있다.

학술지목록DB를 활용하여 학술정보DB 제작에 필요한 학술지기본정보를 작성한다. 학술지 목록DB는 학술지에 대한 종정보를 수록하고 있으며, 학술지 변경정보를 관리한다. 학술지체크인 정보를 활용하여 해외 학술지 수록기사에 대한 기본정보를 작성한다. 해외 학술지의 각 권호마다 부착된 SISAC 바코드를 스캐닝함으로써 학술지 목록DB와 연동하여 자동으로 자료 입수처리가 이루어진다. 이 SISAC 바코드를 SICI코드로 변환하여 학술지 수록기사를 자동 추출하는데 사용한다. 학술지 체크인정보에는 학술지명, ISSN, 권호정보, 발행일정보, 소장정보 등이 수록되어 있어, 해외학술정보 DB제작에 필요한 항목만 추출하여 재사용이 가능하다. 또한 학술지 수록기사의 기본정보를 작성하기 위해 보유중인 학술지 목차DB, CD-ROM뿐만 아니라, 웹을 통해 얻을 수 있는 여러 공개소스로부터 데이터를 추출, 수집한다. 해외학술지 수록기사의 가공정보를 작성하기 위해 보유중인 학술지 목록DB뿐만 아니라, 웹에 공개된 여러 정보자원을 활용할 수 있다.

이미 보유중이거나 이용할 수 있는 가용자원들을 연계 활용하여 해외학술정보 데이터베이스를 제작했을 때 장점은 다음과 같다.

① 데이터의 정확성, 완전성, 최신성, 연속성 향상
② DB제작 대상이 되는 자료입수부터 DB제작까지 소요되는 제작시간 단축
③ 과거의 시공간적 제약조건을 극복하고, 공개된 환경에서 인적자원, 정보자원, 시스템자원 공유
④ 양질의 콘텐츠를 적은 비용으로 구축

III. 가용자원을 활용한 데이터베이스 제작

1. 메타데이터 수집기

1.1 메타데이터 자동추출

학술정보 데이터베이스의 대상이 되는 원자료는 책자형, CD-ROM, 웹으로 되어있다. 책자형 자료의 경우 해당 페이지를 스캐닝한 후 문장인식을 통해 필요한 1차 데이터를 추출한다.

학술정보 데이터베이스의 구축대상저널이 입수처리되면 SISAC 코드를 이용하여 보유중인 학술지목차 DB이나 CD-ROM으로부터, 또는 웹상에서 이 저널에 대한 기본 메타데이터를 추출한다. 정보수집 에이전트는 미리 정의된 정보수집프로파일을 바탕으로 여러 곳에 흩어져 있는 이형질의 정보소스로부터 필요한 메타데이터를 자율적이고 지속적으로 수집한다. 에이전트는 DB개발자를 대신하여 DB제작에 필요한 작업을 자동으로 처리해 주는 프로그램이라 할 수 있다. 따라서 DB 개발자는 원하는 정보를 얻기 위해 직접 질의어를 작성, 수정하여 데이터를 수집하는 등의 작업을 할 필요가 없다.

인터넷에 존재하는 메타데이터는 대부분 하이퍼링크를 이용하여 다른 정보사이트와 연결되어 있는데, 인덱싱을 위해서는 하나의 문서에서 출발하여 그 문서내에 있는 여러 링크를 어떠한 순서로 검색할지 결정하여야 한다[1].

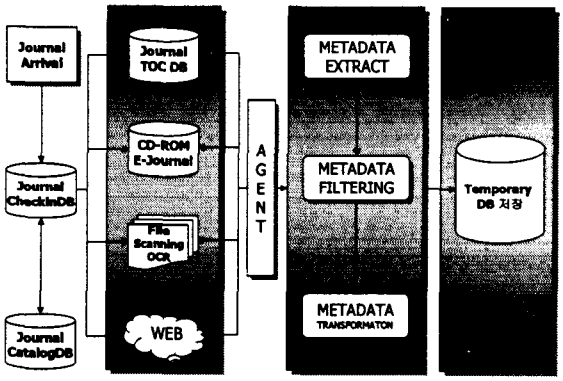

그림 1. 메타데이터 수집과정

1.2 메타데이터 필터링

메타데이터 필터링은 끊임 없이 유입되는 정보 중에서 필요한 것이 무엇이고 필요없는 것이 무엇인지를 판단하여 필요하지 않은 것은 무시한다. 필요한 정보가 무엇인지 알기 위해 정보수집 프로파일이 중요한 역할을 한다. 에이전트가 가용가능한 자원으로부터 메타데이터를 가져올 때는 프로파일과 비교하여 필요한 데이터만 걸러서 저장한다. DB 개발자는 이 필터링 과정을 거친 결과를 보고 실제로 원하는 것인지를 판단하게 되고, 피드백을 거쳐서 정보수집 프로파일을 재구성하게 된다. 데이터 필터링을 거쳐 기구축된 데이터와 중복된 메타데이터를 차단한다.

1.3 메타데이터 변환

수집한 문서를 분석하여 미리 정의된 형태에 따라 불필요한 부분은 제거하고 필요한 부분만 추출하여 임시DB에 저장한다.

웹사이트와 CD-ROM로부터 추출한 메타데이터는 다수의 정보소스로부터 수집되어 다양한 포맷을 가지고 있기 때문에 학술정보 DB의 데이터 포맷에 맞추어 변환하여 DB에 저장한다. 웹사이트와 CD-ROM 자료는 그 형태가 매우 다양하므로, 에이전트는 변환테이블(conversion table)을 이용한다. 이 변환테이블은 추출한 소스데이터의 특정값을 지정한 값으로 바꾸어 임시DB에 저장한다.

2. 메타데이터 에디터

에이전트를 통해 수집된 정보는 학술정보 데이터베이스 형식에 적합하게 임시데이터베이스에 저장된다. 정보입력기는 정보입력기는 입력, 수정, 삽입, 저장, 조회, 삭제 기능을 갖추고 있으며, 에이전트에서 생성한 데이터를 검증하는 역할을 수행한다.

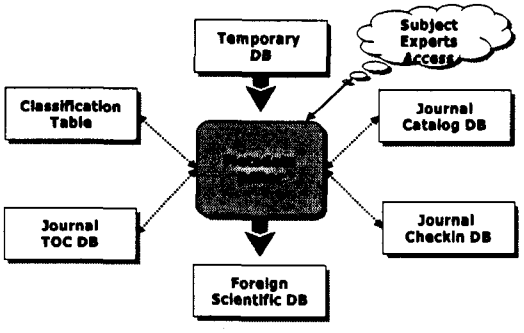

그림 2. 메타데이터에디터와 가용자원간의 관계

3. 메타데이터 자동분류

메타데이터 자동분류기는 미리 정의되어 있는 범주를 메타데이터의 내용에 근거하여 컴퓨터가 자동으로 범주를 할당하는 작업을 수행한다.

과학기술정보의 폭증으로 다루어야 할 정보량이 증가하였다. 전통적인 메타데이터 분류방식은 수

작업으로 수행되므로, 정보생성속도에 비해 정보가공 속도가 지나치게 뒤쳐져 정보순환주기에 적응하지 못할 뿐만 아니라 정보가공에 시간과 인력 등 경제적으로 많은 비용을 요구하였으며[2], 분류작업자에 따라 일관성이 결여되기도 하였다.

이와같은 비효율성을 감소시키고 경제성을 도모하기 위해 메타데이터 자동분류가 요구되고 있다.

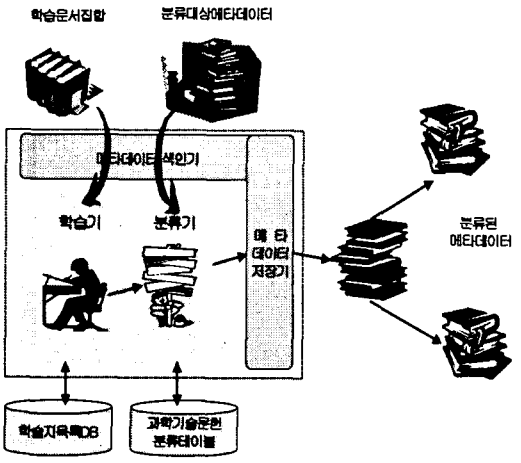

그림3. 메타데이터 자동분류과정

학술정보 메타데이터 자동분류를 위해서 매칭테이블을 이용하며, 그 절차는 다음과 같다.

① 이미 학술지목록 DB에 수록되어 있는 종단위 분류정보에 의해 메타데이터가 수록되어 있는 학술지단위 범주화 작업을 수행한다.
② 학술정보 DB에 이미 구축되어 있는 메타데이터를 이용하여 학습분류기를 만든다.
③ 학술정보 데이터베이스에 구축대상인 메타데이터 정보로부터 자질을 추출한다.
④ 전처리과정을 거친 메타데이터에 나타난 여러 용어중에서 자질을 선정한다.
⑤ 자동분류기를 이용해 메타데이터에 해당되는 분류정보를 할당한다.
⑥ 주제전문가는 자동분류된 데이터가 적합한 범주로 분류되었는지 검증한 후 데이터베이스에 저장한다.
⑦ 메타데이터 분류기는 구축되는 학술정보의 분류정보를 학습함으로써 재구성된다.

4. 메타데이터 검증

에이전트를 통해 수집된 메타데이터의 가공이 완료되면 교열검증단계를 거쳐 데이터 가공시 발생하게 되는 논리적, 물리적 오류를 찾아 정정한다. 데이터 검증방법은 인간의 중재여부에 따라 정적 검증방법과 동적 검증방법으로 구분된다. 정적 검증방법은 인적 요소의 중재없이 기계적으로 데이터의 형태적 무결성을 측정하며, 동적 검증방법은 이미 구축되어 있는 데이터를 이용하여 기계적 오류가능성을 판단한 후 인간의 지적분석 작업을 요구한다. 가용자원을 활용하여 기본 메타데이터를 추출하기 때문에 메타데이터의 오류는 현저히 감소하게 되고, 데이터의 품질은 향상된다.

메타데이터의 교열검증인력은 웹에서 공개선정된 고급주제전문가들로 구성된다. 주제전문가들은 시간과 공간에 구애받지 않고 웹을 통해 메타데이터 검증을 수행한다. 검증과정을 거친 데이터는 데이터베이스로 구축된다[3].

IV. 결론

해외학술정보 데이터베이스 구축방법은 기존의 노동집약적 수작업 제작방식에서 탈피하여 에이전트를 이용하여 다양한 소스로부터 데이터를 수집하고 정해진 포맷에 맞추어 가공되고, 변환저장되는 개방형 제작방식으로 전환되고 있다. 가용자원을 최대한 활용하여 데이터베이스를 제작함으로써 데이터 품질의 정확성, 완전성, 최신성, 연속성을 향상시킬 수 있다. 또한 데이터 제작의 신속성을 기할 수 있고, 제작에 소요되는 경비를 감소할 수 있다.

참 고 문 헌

[1] 최중민 "인터넷 정보추출 에이전트", 정보과학회지, 제18권, 제5호, pp. 48-53, 2000.
[2] 임희석 "자동문서분류기의 개발동향 및 구축", 한국인터넷정보학회, 제3권 제3호, pp.48-56, 2002.
[3] 노경란, 권오진, 유현종 외 "실시간 서지데이터베이스 평가방법에 관한 연구", 한국콘텐츠학회 논문지, 제2권 제4호, pp.76-84, 2002.