학술지정보서비스를 위한 해외학술정보 가공모델 연구

이석형*, 김한기*, 윤희준*, 윤화묵*
*한국과학기술정보연구원 정보유통기술개발실
e-mail:skyi@kisti.re.kr

A Study of Data Processing Model for Overseas S&T Information System

Seok-Hyoung Lee*, H.K. Kim*, H.J. Yoon* H.M. Yoon*
*Dept of Information Technology, KISTI

요 약

한국과학기술정보연구원(KISTI)은 국내 과학기술자의 연구를 돕기 위해 국내외의 여러 기관으로부터 정기적, 또는 비정기적으로 발간되는 학술지, 연구보고서, 특허 및 각종 동향, 분석 정보를 수집하여 서비스하고 있다. 국내에서 생산되는 정보는 그 발생양이 방대하지 않고 일정 수준 이상의 서지정보를 담고 있으며 원문정보까지 제공이 가능하기 때문에 정보의 질이 비교적 높은 편이나, 해외에서 수집되는 학술 정보는 데이터 처리량이 방대하기 때문에 처리시간이 오래 걸릴 뿐만 아니라, 기본적인 서지 정보만을 담고 있어 원문 서비스나 초록 및 주제 분류 등의 부가적인 서비스를 위해서는 추가적인 데이터 가공이 필수적이다. 따라서 본 논문에서는 데이터 처리 속도와 이용자 중심의 해외학술정보의 원문 및 부가 서비스 제공 등을 고려한 데이터 가공 방법에 대해 연구한다.

1. 서론

인터넷의 발달로 인해 사용자들은 웹을 이용하여 많은 정보를 습득할 수 있게 되었다. 또한 정보 수요가 증가함에 따라 다양하고 양질의 정보를 제공하기 위한 많은 방법이 연구되고 활용되고 있다. [1]

그 중에 과학기술정보는 한 국가의 과학기술발전의 중요한 밑거름이 되기 때문에 이에 대한 효율적인 정보의 제공이 절실한 실정이다[2]. 이에 한국과학기술정보연구원에서는 과학기술 종합정보시스템을 개발하여 국내 과학기술자에게 과학기술정보를 서비스 하였고, 2004년 11월 원내 모든 DB를 대상으로 하는 통합검색서비스를 제공하고 있다. 과학기술 통합검색서비스는 국내외에서 발간된 학술잡지, 회의자료, 연구보고서, 학위논문, 특허기술, 분석 및 동향정보 등의 서지정보를 KRISTAL-2002[3] 기반의 검색 데이터베이스를 구축하고, 사용자가 이를 검색하고 필요한 정보에 대해 원문 신청까지 이루어질 수 있도록 하는 서비스이다[4].

그런데, 현재 서비스되고 있는 분야 중 해외과학기술정보는 국내 산학연의 연구개발 및 기술혁신에 있어 의존도가 매우 높고, 선진국의 기술보호주의의 심화와 핵심기술정보의 대외유출규제가 강화되고 있어 물리적으로 접근하기 어려운 자료, 또는 지속적으로 접근하기 어려운 자료에 대해서는 데이터베이스 구축을 통한 자료 제공이 필수적이다[5]. 현재 한국과학기술정보연구원에서 수집하는 해외학술정보는 약 12,000여종의 3,000,000여건의 기사건수인데, 이를 효과적으로 처리하여 사용자에게 양질의 서비스를 제공하기 위한 방법이 필요하다. 따라서 본 논문에서는 데이터 처리 속도와 이용자 중심의 해외학술정보의 원문 및 부가 서비스 제공 등을 고려한 데이터 가공 방법에 대해 연구한다.

2. 해외학술정보 구성

해외학술정보서비스를 위해 수집하는 서지정보는 크게 6종이다. 6종을 수집하는 이유는 각 정보를 생산하는 기관별로 데이터의 성격과 특징이 상이하기 때문에, 이용자에게 다양한 정보를 제공할 수 있도록 하기 위함이다.

2.1 ADONIS

1986년 7월 미국 과학기술 분야의 10개 잡지 출판사 연합으로 설립되었으며 480여종의 최신학술지에 수록된 논문의 전문을 CD_ROM으로 축적하고 있으며 각국의 대표적인 도서관이나 정보 서비스 기관에 유료로 배포하였다. 기본적인 서지정보 이외에 표준주제 분류와 초록정보를 담고 있다. 가공대상정보건수는 1992년부터 2002년까지 생산된 1,076종 1,300,000여건의 논문이다.

2.2 EBSCO

해외에서 발간되는 과학기술문헌에 대한 목차데이터로서 EBSCO 사로부터 저작권을 확보한 정보이다. 이 데이터의 특징은 출간된 지 얼마 안 된 정보를 가장 빠르게 이용자에게 알릴 수 있도록 기본적인 서지정보만을 제공한다. 따라서 이 데이터는 해외학술지 정보 가공의 근간이 된다고 할 수 있다. 가공대상 정보건수는 24,000여종 17,000,000여건이다.

2.3 SwetScan

약 16,000여종의 종정보와 TOC 및 921종의 초록 데이터를 ASCII, SGML등 14개 포맷으로 제공하고 있으며 SwetsScan에 저장되는 출판물의 목차 및 초록 정보에 대해 출판사로부터 구입한 소유권 및 운영권을 Swets사가 보유하고 있다. 가공 대상 건수는 1993년도부터 현재까지 16,000여종 20,000,000여건이다.

2.4 JTOC

일본에서 발행되는 과학기술 학/협회지를 중심으로 구축한 목차정보로 한국과학기술정보연구원에서 2002년부터 자체제작하고 있다. 초록정보는 존재하지 않으며, 가장 기본적인 서지정보만을 담고 있다. 현재까지 700여종 200,000여건의 데이터가 구축되어 있다.

2.5 CrossRef

CrossRef는 2001년 1월부터 PILA(Publisher International Linking Association)에 의해 운영되는 단체로 비영리로서 독립적인 단체이며, 350여개 학술저널 출판사, 많은 제휴업체등 많은 단체들이 참여하고 있다. CrossRef에서 제공하는 기사정보에는 DOI 정보를 포함하고 있어 이용자 원문서비스를 위한 필수 수집대상 기관이며, 가공대상건수는 약 9,000 여종 12,000,000여건의 데이터이다.

2.6 BIST

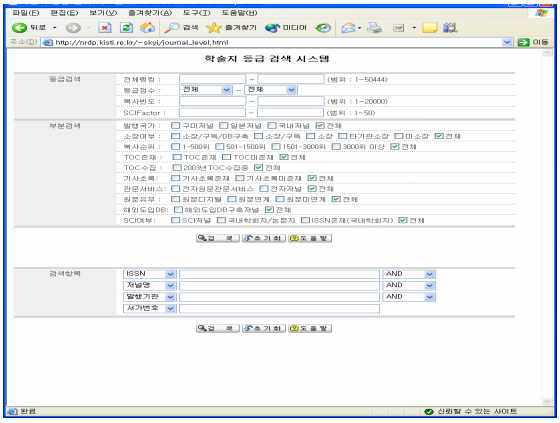
과학기술문헌 데이터베이스(BIST)는 KISTI가 국내외에서 발간되는 과학기술분야 정기간행물에 수록된 기사를 정보가치와 시사성을 기준으로 기사를 엄선하고 이들을 한글로 번역한 다음, 분류와 색인등의 가공과정을 거쳐 DB로 제작하였다. 가공대상 건수는 6,000여종 2,800,000 여건의 기사를 제작하였다.

3. 해외학술정보 가공 및 DB 구축 모델

2장에서 살펴본 바와 같이 해외학술정보 서비스를 제공하는데 여러 기관으로부터 정보를 수집하는 이유는 사용자에게 다양하고 양질의 정보를 제공하기 위함이다. 각각의 정보는 상호 보완적인 내용을 담고 있기 때문에 이종의 메타 정보간 중복데이터를 체크하여 양질의 정보를 생산해내기 위한 정보가공 모델이 필요하다. 본 장에서는 이런 다양한 기관에서 수집하는 여러 정보를 효율적으로 처리할 수 있는 정보가공 모델을 제시하고, 실제 적용하고 있는 예를 설명한다. 해외학술정보 가공을 위한 워크 플로우는 크게 4단계로 이루어진다. 첫째는 정보 가공 대상 선정이고, 둘째는 입수 기관별 우선순위 선정 및 대상 기사 추출, 셋째는 우선 서비스 항목 적용 및 이종의 메타 정보간 중복데이터 체크, 마지막으로 데이터 구축 이다.

3.1 정보가공 대상 선정

6개 기관에서 수집하는 모든 정보를 이용자에게 서비스하는 것은 자료의 양이 방대하고 가비지 데이터가 많기 때문에, 양질의 데이터베이스를 서비스하기 위해 우선 과학기술핵심저널을 선정하여 대상 종에 해당하는 기사만을 구축하였다. 과학기술핵심저널 선정은 불필요한 자원수집과 사용자에게 꼭 필요한 정보만을 제공한다는 측면에서 중요하다.

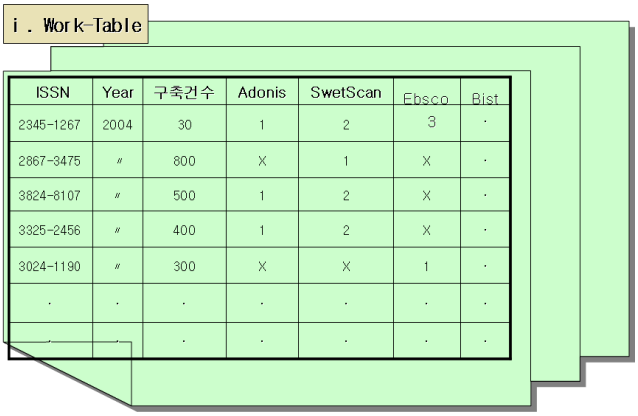


(그림1) 학술지 등급 검색시스템

과학기술핵심저널 선정기준은 SCI 등재저널, BLDSC 추진저널, 복사빈도, 최근 10여년간 수집여부 등을 종합적으로 고려하여 등급화 하였다. (그림 1)은 자료관리시스템에 등록된 약 5만여 종을 대상으로 등급을 부여하여 검색할 수 있도록 개발된 학술지 등급 검색시스템을 보인 것이다. 이를 바탕으로 약 11,000종의 과학기술핵심저널을 선정하였고, 이 이외에 수집된 종에 해당하는 기사들은 서비스 대상에서 제외하였다. 이로 인해 약 20,000,000여건의 불필요한 기사정보가 제거되어 데이터 가공 속도 감소와 이용 편의성 증대의 효과를 가져오게 되었다.

3.2 입수기관별 우선순위 선정 및 대상 기사 추출

입수기관별로 데이터 품질은 큰 차이가 있다. EBSCO에서 제공하는 TOC목차 속보는 서비스 대상 종수는 많지만, 가장 기본적인 기사정보만 수록되어 있으며, BIST와 같이 제작된 기사정보는 서비스 대상 종수는 한정되어 있으나, 주제 분류, 초록 등의 부가 정보를 제공한다. 따라서, 입수기관별 우선순위를 결정하여 질이 높은 정보는 선순위로 가공 및 구축될 수 있도록 한다. 이를 위해 각 종별로 입수기관별 우선순위를 기록할 수 있는 work-table 구성이 필수적이다. (그림 2)는 3.1절에서 선정된 11,000여 대상 종을 가지고 구성한 work-table 예시이다. 데이터의 품질과 자료량 등을 기준으로 1순위 입수기관을 Adonis, 2순위를 SwetScan, 3순위를 EBSCO로 선정하였고 CrossRef, BIST, JTOC를 4순위 항목 추가 입수기관으로 선정하여 가공하기로 하였다.

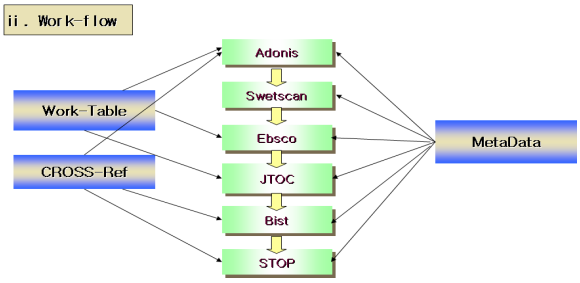


ISSN	Year	구축건수	Adonis	SwetScan	Ebsco	Bist
2345-1267	2004	30	1	2	3	·
2867-3475	〃	800	X	1	X	·
3824-8107	〃	500	1	2	X	·
3325-2456	〃	400	1	2	X	·
3024-1190	〃	300	X	X	1	·
.	.	.	.	.	.	.
.	.	.	.	.	.	.

(그림 2) Work-Table 구성도

3.3 입수기관별 우선 서비스 항목 선정 및 메타정보 중복 체크

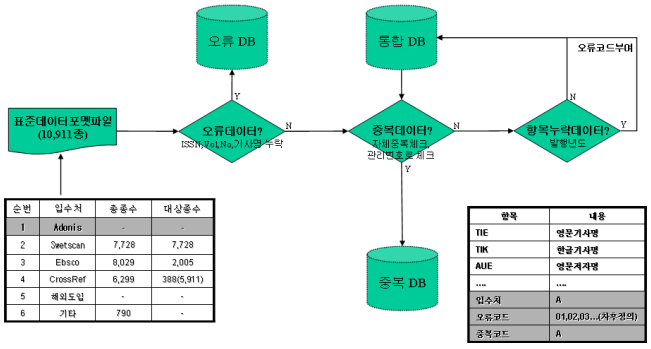
주제 분류나 초록, 원문정보를 식별할 수 있는 DOI(Document Object Identifier), 기사 식별자인 SICI등은 별도의 가공 절차를 거쳐 해당 입수기관에서 포함하는 고유의 내용이다. 이러한 입수기관별로 우선 서비스 항목을 선정하여 하나의 기사에 대해 위와 같은 고급 정보를 한꺼번에 보여줄 수 있도록 하는 것이 중요하다. 또한 이를 바탕으로 입수기관별로 대상 종에 해당하는 기사를 추출하는데, 각 기관별로 제공 포맷이 상이하기 때문에 하나의 표준 포멧으로 가공하여 차후에 재가공시 입수파일을 직접 가공하는 것보다 표준데이터 포맷으로 가공하여 가공의 효율성을 극대화하는 것이 중요하다. (그림 3)은 work-table을 바탕으로 입수기관별 우선순위를 두어 데이터를 처리하는 흐름도를 보인 것이다.



(그림 3) 우선순위별 데이터 처리 흐름도

3.4 데이터 구축

3.1~3.3 단계를 거친 데이터 정제작업 후에 데이터베이스를 구성한다. 데이터 구축은 크게 두 부분으로 나뉘는데 하나는 일반적인 레코드 삽입이며 다른 하나는 중복레코드의 항목보완이다.

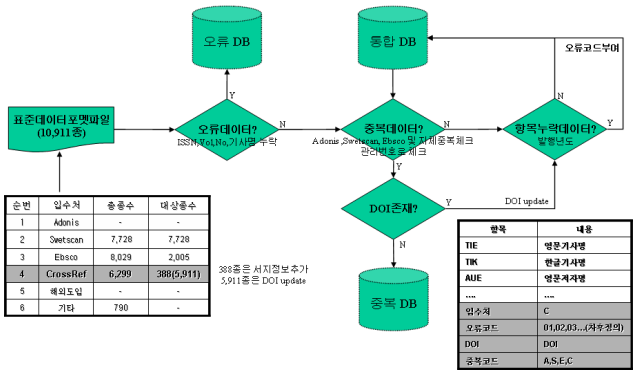


순번	입수처	총종수	대상종수
1	Adonis	-	-
2	Swetscan	7,728	7,728
3	Ebsco	8,029	2,005
4	CrossRef	6,299	388(5,911)
5	해외도입	-	-
6	기타	790	-

항목	내용
TIE	영문기사명
TIK	한글기사명
AUE	영문저자명
....	....
입수처	A
오류코드	01,02,03...(자유정의)
중복코드	A

(그림 4) Adonis 데이터 구축 흐름도

일반적인 레코드 삽입은 중복 체크 후 동일한 정보가 없는 경우 새로운 레코드를 생성하는 것이고 중복 정보는 중복테이블로 적재하여 관리하거나 제거하는 작업이다. 적용 모델에서 이에 해당하는 입수정보는 Adonis, SwetScan, EBSCO에서 제공하는 정보로 주로 TOC 목차 속보를 제공하는 것들은 위의 절차를 따른다. (그림 4)는 Adonis데이터의 DB 구축 과정을 나타낸 것이다.



순번	입수처	총종수	대상종수
1	Adonis	-	-
2	Swetscan	7,728	7,728
3	Ebsco	8,029	2,005
4	CrossRef	6,299	388(5,911)
5	해외도입	-	-
6	기타	790	-

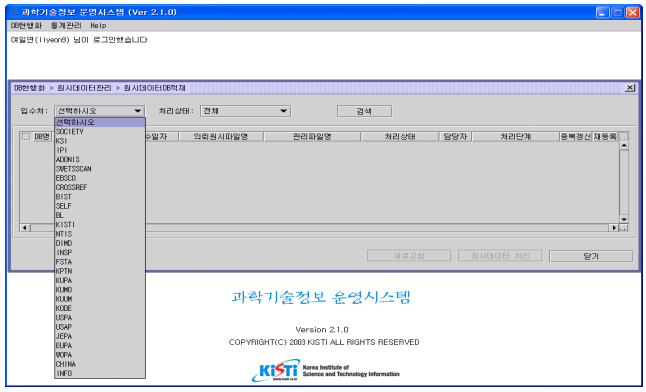
항목	내용
TIE	영문기사명
TIK	한글기사명
AUE	영문저자명
....	....
입수처	C
오류코드	01,02,03...(자유정의)
DOI	DOI
중복코드	A,S,E,C

(그림 5) CrossRef 데이터 구축 흐름도

중복레코드의 항목 보완은 3.3절에서 정해진 우선 서비스 항목을 바탕으로 중복 체크 후 동일한 정보가 있는 경우 대상 필드를 업데이트하는 하거나, 동일한 정보가 없는 경우 새로운 레코드를 삽입하는 경우이다. (그림 5)는 CrossRef 데이터의 DB 구축과정을 나타낸 것이다.

3.5 응용시스템

앞에서 설명된 정보가공모델을 기반으로, 정보처리를 담당하는 관리자를 위한 해외학술정보 운영시스템을 개발하였다. (그림 6)은 해외학술정보 운영시스템 실행화면이다. 본 시스템은 데이터 처리시에 발생되는 모든 트랜잭션에 대한 관리가 가능하며, 원스톱으로 db가공이 이루어질 수 있도록 설계되었다. 또한 각종 오류데이터 및 중복데이터를 손쉽게 처리할 수 있도록 개발 되었다.



(그림 6) 해외학술정보 운영시스템

4. 결론

효율적인 학술지 정보 서비스를 위한 정보가공 모델 연구가 활발히 진행되고 있는 가운데, 본 논문에서는 학술지 정보 가공 및 처리를 위해 필요한 여러 요소 및 절차에 대해서 설명하였다. 본 가공 모델을 적용하여 학술지 가공을 수행하였을 경우 데이터 처리 속도와 검색 속도가 절반 이하로 감소되고, 또한 이용자 만족도는 크게 증가됨을 알 수 있었다. 추가로, 현재 표준포멧파일 및 XML 기반의 기사정보 제작을 일원화하여 모든 학술지 정보를 표준 XML화 하여 관리하고, 기사정보에 대한 식별자 부여를 위한 DOI 서비스 및 권, 호, 페이지 정보 패턴 분석시스템을 개발하여 본 가공모델을 보완할 예정이다.

참고문헌

[1] Seok-Hyoung Lee, N.G.Kang, H.M. Yoon, Y.H. Yae, H. Kim, "Implementation of the XML based Science and Technology Information System using KRISTAL", The 7th IASTED International Conference , IMSA 2003 , 2003.08

[2] 한국과학기술정보연구원, "과학기술정보유통체제 구축" 국무총리실 공공기술연구회, 2002.12

[3] 주원균,정창후,이민호, "KRISTAL-2002를위한 JAVA 사용자 API의 설계 및 구현", 정보과학회 가을 학술발표논문집 Vol.31,No.2, p433-435, 2003

[4] 한국과학기술정보연구원, "과학기술종합정보 서비스시스템 개발 및 운영", 2004

[5] Beom-Jong You, E.B. Lee, A Study on the Construction of Meta Database and Management Systemof the Foreign Technical Information, Journal of the Korean Society for Library and Information Science, volume 36 number 3, 2002.02