과학기술문헌 데이터베이스의 검색효율 향상을 위한 색인 보완 방안

A Study on Adding Index Terms for improving the retrieval efficiency of the STI database

김 병 규, 김 태 중, 강 무 영, 류 범 종
한국과학기술정보연구원

Kim Byung-kyu, Kim Tae-jung, Kang Mu-yeong, You Beom-jong
The Korea Institute of Science and Technology Information(KISTI)

요약

KISTI는 국내에서 발간되는 과학기술 학술논문을 가공해서 데이터베이스로 구축, 제공하고 있으며 그 규모는 2010년에 100만건을 넘어서고 있다. 규모가 늘어남에 따라 체계적인 주제분류 등 검색의 효율화를 위한 부가적인 가공이 필요하다. 전통적으로 정보를 가공하는 방법으로 초록화, 분류, 색인, 추록화 등을 혼용하여 사용하고 있다. 이 가운데 색인과 분류는 특히 정보 검색에 유용한 도구로 활용되고 있다. 이 논문에서는 기 구축된 과학기술문헌 데이터베이스에 분류 코드와 색인어를 부여하여 검색효율을 향상시키기 위한 방안을 제안한다.

Abstract

KISTI collects the scientific and technical articles published in Korea and builds the Korean STI database for scientists. The number of papers exceeds one million. To improve the search efficiency of the database additional processing is required. Abstracting, classification, indexing and extracting is a traditional processing method adding value to information. Indexing and classification are useful tool to assist efficient retrieval. In this paper, authors propose a method to improve information retrieval efficiency by assigning classification code and index terms to records of Korean STI database.

I. 서론

1. 국내 발행 과학기술 논문 현황

우리나라의 과학기술 논문은 미국, 유럽 등의 학술 논문이 상업 출판사에 의해 발간되는 경우와 달리 대체적으로 학회가 과총(한국과학기술단체총연합회)과 한국연구재단(전 학술진흥재단) 등의 지원을 받아 직접 발행하는 형식으로 발간되고 있다. 2010년 한국과학기술정보연구원(KISTI)에서 조사한 결과에 따르면 과학기술분야의 789개 학회에서 국영문으로 1,089개의 학회지 및 논문지가 발간되고 있으며 2010년말 현재 478개 학회의 706종 논문지 및 학회지를 데이터베이스로 구축하여 NDSL(www.ndsl.kr)을 통해 온라인으로 제공되고 있다[1]. 〈표 1〉은 최근 3개년도에 분야별 발간된 논문 현황을 보여주고 있다. KISTI는 2001년에 산업기술정보원(KINITI)와 연구개발정보센터(KORDIC)의 통합에 의해 새롭게 설립된 기관으로 주요 취급분야는 과학기술이며 각 기관은 독자적인 정보처리 기준(분류체계, 작업지침 등)을 채택하여 사용하였다. 통합후 산업기술정보원에서 사용하던 기준을 중심으로 추진하였으나 2007년 한국과학기술원의 NDSL을 통합하면서 국내 문헌의 분류 등을 적시에 수행하지 못하고 있어 국내 문헌의 검색 효율성을 높이기 위한 지원이 요청되어 오고 있는 상황이다.

표 1. NDSL을 통해 제공되는 최근 3년간의 논문 분포

학문분야	종수	2008	2009	2010	총계
이학	91	4,592	5,170	3,535	13,297
공학	220	14,159	15,312	10,573	40,044
의약학	111	5,437	5,446	3,265	14,148
농수해	29	1,407	1,505	899	3,811
예체	14	538	489	307	1,334
인문사회	108	5,319	5,434	3,155	13,908
기타	6	334	331	209	874
계	579	31,786	33,687	21,943	87,416

2. 정보 가공 방법 점검 및 조정

초록작성, 주제 분류, 색인, 추록이 대표적인 정보를 가공하는 방법으로 최근에는 초록의 경우에는 저자 초록을 사용하는 추세이며 추록은 요약의 형태로 자동 초록 등으로 활용되고 있다. 분류와 색인은 정보검색에서 재현율과 정확율을 높여주는 유용한 도구로 사용되고 있다. KISTI는 독자적인 분류체계(과학기술문헌분류체계-BIST 분류표)를 사용하였으나, 표 1에서와 같이 인문사회과학 등의 분야도 다수 포함되어 있어 2010년에 발표된 '2008년 재편 국가과학기술표준분류체계'로 변경하기로 하였다[2][3].

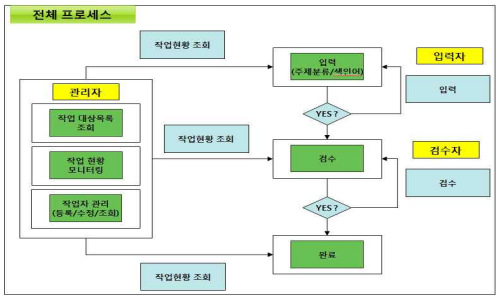
색인은 저자가 논문을 제출할 때 기입한 키워드와 논문의 제목과 초록으로부터 자동 추출한 색인어를 그대로

사용하고 있으며 여기에 분야별 전문가에 의한 색인어를 추가할 계획이다. 색인 작업은 평균적으로 5개 내외의 색인어를 추가하게 되며 최소한 1개는 '과학기술 용어 시소러스'에 포함된 용어를 사용하도록 규정하여 기존 시소러스의 갱신에 활용하는 한편 검색의 효율을 높이는 도구로 작용할 것으로 예상된다[4].

Ⅱ. 본론

1. 분류 및 색인 시스템 구성

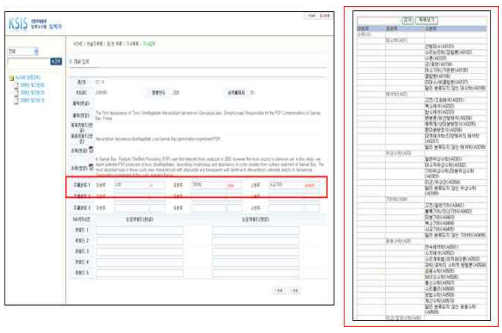
학술논문에 대한 분류 및 색인의 정보가공을 지원하기 위하여 KISTI 주제분류 입력시스템을 개발하였다[5]. 입력시스템은 입력/검수/관리의 등급으로 구성되며 시스템 사용자는 등급에 따른 권한과 업무프로세스에 따라 작업을 수행할 수 있다. 시스템의 전체 프로세스는 그림 1과 같다.



▶▶ 그림 1. KISTI 주제분류 입력시스템 DB구축 프로세스

2. 시스템 기능 및 사용자 인터페이스

사용자의 역할에 맞는 업무를 지원하기 위하여 사용자 등급별로 적합한 기능 및 인터페이스를 이 시스템에 적용 및 구현하였다. 먼저 입력자를 위한 사용자 화면은 그림 2와 같으며 지원 기능은 크게 세 가지이다. 첫째는 학술논문에 대한 메타 및 원문정보의 제공이며 둘째는 시스템 상에 정보화된 '과학기술표준분류체계' 코드표를 적용하여 주제분류를 원클릭으로 자동입력(주제분류 항목 및 코드검색 및 적용)할 수 있는 기능이다. 마지막으로 '과학기술 용어 시소러스'를 사전화하여 사용자가 색인정보입력 시 검색용어 자동완성 및 매칭된 한-영 용어의 자동입력기능이다. 논문당 주제분류의 개수는 최대 3개까지, 한.영 색인어 추출은 각각 5개씩으로 구성하고 입력자의 잘못된 입력에 대한 예외처리(문자열길이/NULL 체크 등)기능을 시스템에 반영하였다.



▶▶ 그림 2. 입력 Level 시스템 화면

이 시스템은 입력 후 검수기능을 제공한다. 검수자는 그림 3의 사용자 화면과 같이 개별 논문들에 대한 입력작업 완료여부를 확인할 수 있으며 해당 논문의 가공정보(분류 및 색인정보)의 적합여부를 판별하여 검수완료 또는 입력된 정보를 수정 할 수 있다. 따라서 입력 후 검수과정을 통해 분류 및 색인정보의 품질을 개선할 수 있고 입력정보와 검수정보의 비교를 통해 작업방식을 제고할 수 있다.

▶▶ 그림 3. 검수 Level 시스템 화면

마지막으로 시스템 관리자의 업무수행 지원을 위해 작업자 등록 및 관리 기능과 작업 현황 모니터링(작업자별/일자별/학술지별) 등의 기능을 제공한다. 향후 정보가공 대상목록(학술지/권호/논문)의 조회 및 목표대비 작업진행률 등의 다양한 통계정보를 제공할 예정이다.

Ⅲ. 결론

KISTI는 국내에서 발간되는 과학기술 학술논문을 가공해서 대규모의 데이터베이스로 구축하여 서비스하고 있다. 검색 효율성의 향상에 있어 논문에 대한 주제분류 및 색인어 추출은 필수적인 작업이나 쉽지 않은 작업과정과 환경으로 인하여 해당 정보의 가공에 어려움이 있어왔다. 이에 KISTI는 효과적이고 원활한 분류 및 색인정보 가공 환경을 제공할 수 있는 주제분류 입력시스템을 개발하였다. 이를 위해 시스템 사용자 역할별로 업무프로세스를 설계하고 역할별 기능 및 사용자 인터페이스를 구현하였다. 본 시스템의 활용을 통해 과학기술 학술논문에 대한 빠르고 정확한 분류 및 색인작업이 가능할 것으로 기대된다. 향후 시스템의 지속적인 업그레이드를 통해 성능을 개량하고 구축 정보의 서비스 적용 방안에 대해서도 연구를 수행할 예정이다.

▌참 고 문 헌▐

[1] 국가과학기술정보센터 http://ndsl.kr
[2] 국내 학술정보 가공 지침서 2010, KISTI
[3] 2008년 과학기술표준분류체계, KISTEP
[4] 과학기술 용어 시소러스, KISTI
[5] KISTI 전문주제분류 입력시스템
http://ksis.kisti.re.kr