KRISTAL-2002를 이용한 대용량의 이기종 과학기술 데이터베이스 구축시스템 설계 및 구현

김명일° 신수미
한국과학기술정보연구원
{mikim°, sumi}@kisti.re.kr

A Design and Implementation of a Construction System for Huge and Heterogeneous Scientific Database Using the KRISTAL-2002

Myungil Kim° Sumi Shin
Korea Institute of Science and Technology Information (KISTI)

요 약

한국과학기술정보연구원은 국가 과학기술 지식정보인프라의 중심기관이자 과학기술종합정보센터로서 논문, 특허, 연구보고서, 사실정보, 생물다양성정보 등을 비롯한 다양한 종류의 과학기술관련 데이터베이스를 구축, 수집 및 서비스 하고 있다. 그러나 이러한 정보들은 약 40여개의 웹사이트에 분산되어 서비스되고 있어, 하나의 서비스 접점에서 다양한 정보를 쉽고 편리하게 얻고자 하는 이용자들의 요구를 만족시키지 못하고 있다. 따라서 본 논문에서는 분산 서비스되고 있는 대용량의 이기종 과학기술관련 데이터베이스를 각 데이터베이스의 특성을 고려하여 효과적으로 통합검색 될 수 있도록 하는 통합검색 데이터베이스 구축시스템을 설계한다. 본 논문의 통합검색시스템은 KISTI에서 자체 개발한 정보검색관리시스템인 KRISTAL-2002를 사용한다.

1. 서 론

한국과학기술정보연구원(Korea Institute of Science and Technology Information, KISTI)은 국가 과학기술 지식정보인프라 구축의 중추기관으로서, 과학·기술 및 이와 관련된 산업정보를 수집·분석·관리하고 이러한 정보를 유통하기 위한 시스템을 개발하여 국가 과학기술진흥에 이바지 하고 있다. 현재 KISTI에서는 분석, 동향, 논문, 연구보고서, 특허, 인력 등의 약 4,700여 만건의 과학기술관련 데이터베이스에 대한 통합검색 서비스를 제공하는 yesKISTI.net[1]을 구축·운영하고 있다. 그러나 플라즈마 물성, 무기결정 구조 등의 실험데이터를 포함하는 사실정보, 다양한 생물종에 대한 정보를 제공하는 생물다양성 정보, 각 웹사이트에서 제공하는 HTML 형태의 문서나 게시판·자료실 등의 게시물, 세미나 동영상 자료 등 KISTI에서 생산 및 서비스하는 모든 정보를 포함하고 있지 않다.

따라서 본 논문에서는 이러한 다양한 종류의 정보에 대해 각 정보의 특성에 맞는 검색기능을 제공하여 보다 효과적으로 정보검색을 수행할 수 있는 시스템을 설계하기 위한 전 단계로서 각 정보를 적절히 그룹화하여 통합 데이터베이스를 구성하는 방안을 제안한다. 이때, 각 데이터베이스를 통합 구축하기 위해 KISTI에서 자체개발하여 보급중인 정보검색관리시스템인 KRISTAL(Korea Retrieval Information of Science and Technology Access Line) 2002[2][3]를 사용한다.

본 논문의 구성은 다음과 같다. 2장에서는 통합검색 대상인 KISTI의 과학기술관련 데이터베이스 및 KRISTAL-2002 시스템을 소개하고, 3장에서는 KRISTAL-2002를 기반으로 대용량의 이기종 과학기술 데이터베이스를 대상으로 한 통합 DB 구축시스템을 제안한다. 마지막으로 4장에서는 결론과 향후 연구방향에 대해 논의한다.

표 1 다양한 종류의 과학기술 데이터베이스

(2005.9.13일 기준)

종류	내용	건수
분석	분석보고서	15,723
동향	국내·외 동향	97,707
논문	국내·외 학술지, 학술회의, 학위논문 등	30,426,585
특허	한국, 일본, 유럽 등의 공개특허, 등록특허 등	16,380,535
연구보고서	국내외 연구보고서	163,379
인력	과학기술관련 인력정보	293,729
세미나동영상	세미나 동영상자료	23,318
사실정보	플라즈마 물성, 무기결정구조, 신약 등의 실험데이터 및 분자구조 등	-
생물다양성정보	지렁이, 담수어 등의 13개 종에 대한 정보	7,933
기타정보	웹사이트의 HTML 형태의 문서, 게시판/자료실의 게시물 등	-

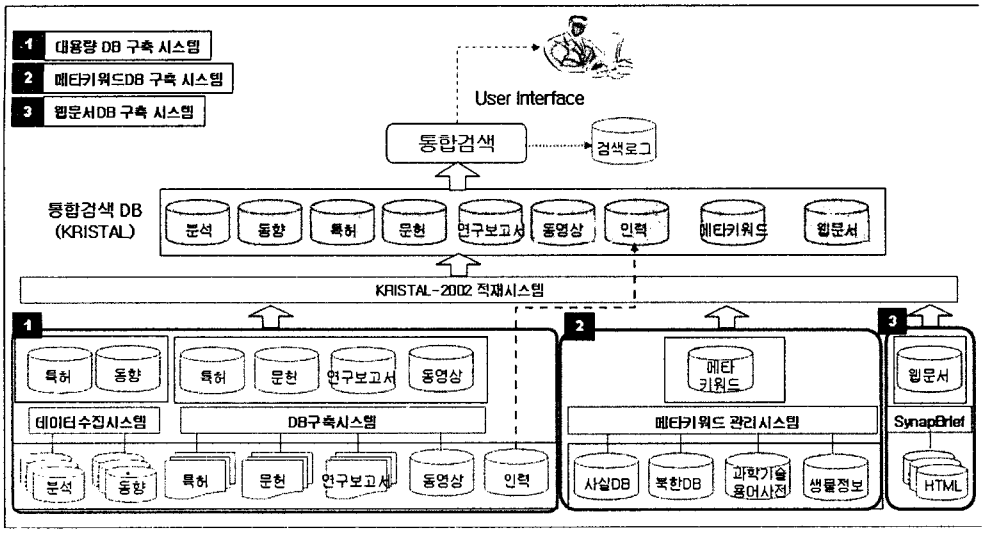

그림 1 대용량의 이기종 과학기술 데이터베이스 구축시스템

2. 관련연구

본 장에서는 통합검색 대상인 다양한 종류의 과학기술 데이터베이스를 소개하고, 이러한 정보를 효과적으로 구축하기 위해 사용되는 KRISTAL-2002를 소개한다.

2.1 과학기술 데이터베이스

KISTI에서 제공중인 과학기술 관련 데이터베이스는 표 1과 같이 그 종류가 매우 다양하고, 구축된 데이터의 양이 매우 방대하다. 표 1에 나열된 데이터베이스는 현재 yesKISTI.net을 비롯한 약 40여개의 홈페이지 분산되어 서비스되고 있으며, 이는 다양한 종류의 정보를 쉽고 편리하게 검색하고자 하는 이용자의 요구에 부합되지 않는다.

2.2 KRISTAL-2002

KRISTAL-2002는 KISTI에서 개발한 정보검색 관리시스템으로, UNIX 기반에서 데이터에 대한 저장, 관리 및 검색에 대한 처리를 효율적으로 처리할 수 있는 시스템이다. KRISTAL-2002의 특징은 아래와 같다.

- 유니코드 기반으로 데이터를 저장 및 관리하므로, 다국어 및 한글고어, 중국어 등을 쉽게 처리할 수 있다.
- 하나의 데몬으로 여러 개의 데이터베이스를 관리할 수 있어 시스템 메모리를 효과적으로 사용한다.
- 상용 DBMS에서 지원하는 다중 테이블 구조를 기반으로 설계되어 있다.
- 정형 데이터 형식과 반정형 데이터 형식 및 대용량 데이터를 지원하여 다양한 데이터를 취급할 수 있다.

위와 같은 특징을 가지는 KRISTAL-2002를 사용하여 정확하고 다양한 검색이 가능한 통합검색 시스템을 설계하고자 하며, 이를 위해 각 데이터베이스를 적절히 그룹화 하여 통합 데이터베이스를 구축한다.

3. 대용량의 이기종 과학기술 데이터베이스 구축시스템 설계 및 구현

대용량의 이기종 과학기술 데이터베이스를 구축하기 위한 시스템은 그림 1과 같이 구성되어 있으며, 각각의 구성 요소에 대한 설명은 아래와 같다.

- 대용량 데이터 구축시스템 : 논문, 특허, 연구보고서 등 그 양이 방대하고, Oracle, MySQL, MS-SQL 등 다양한 종류의 데이터베이스를 기반으로 구축되어 있는 정보들을 KRISTAL 적재 시스템을 이용하여 통합검색 데이터베이스를 구축한다.
- 메타키워드 구축 시스템 : 통합검색이 불가능한 사실정보나 전문정보에 속하는 생물다양성정보, 북한과학기술DB 등을 검색하기 위한 시스템으로 각 DB를 적절히 표현할 수 있는 키워드들로 구성된 메타키워드 DB를 대상으로 검색을 수행한다.
- 웹문서 데이터 구축 시스템 : 한민족과학기술자네트워크(KOSEN)[4]를 비롯한 KISTI의 약 40여개의 웹사이트에서 서비스중인 데이터베이스 형

태로 구축되어 있지 않은 HTML 문서나 게시판 또는 자료실의 게시물들을 콘텐트 수집 시스템인 SynapBrief[5]를 이용하여 수집하고 수집된 콘텐트를 MySQL을 이용하여 DB화 한다.

3.1 대용량 데이터 구축시스템

대용량 데이터 구축시스템은 논문, 특허, 연구보고서 등과 같이 그 양이 매우 방대한 데이터를 대상으로 통합 데이터베이스를 구축하는 시스템으로서, 텍스트, Oracle, MySQL 등 그 형태가 매우 다양한 이기종의 데이터베이스를 KRISTAL로 일원화하여 최종 검색대상이 되는 통합 데이터베이스를 구축한다. 이때, 각 데이터베이스간 참조 링킹이 가능하도록 하기 위해 논문간 인용 및 피인용 정보를 가지고 있는 참조논문 데이터[6]를 활용하고, 논문과 인력, 논문과 종정보, 특허와 인력 등의 참조검색(reference search)이 가능하도록 구현된다.

3.2 메타키워드 구축 시스템

플라즈마 물성, 신약 개발을 위한 화합물, 무기결정구조 등 실험값이나 분자구조에 대한 데이터인 사실정보 데이터베이스[7]나 전문정보의 성격을 가지는 생물종에 대한 특성을 표현하는 생물다양성 정보[8], 북한과학기술 관련 DB[9] 등은 데이터의 특성상 논문, 특허 등과 같은 DB와 동일한 방식으로 구축되어 검색하는 것은 적절치 않다. 따라서 이와 같은 정보들은 각 데이터베이스의 특성을 적절히 표현할 수 있는 다수의 키워드를 선정하고 이러한 키워드로만 구성된 메타키워드 데이터베이스를 구축한다. 즉, 사용자가 입력한 검색질의어에 대해 해당 키워드를 검색한 경우에만 그 결과를 노출하도록 하는 방식을 채택한다.

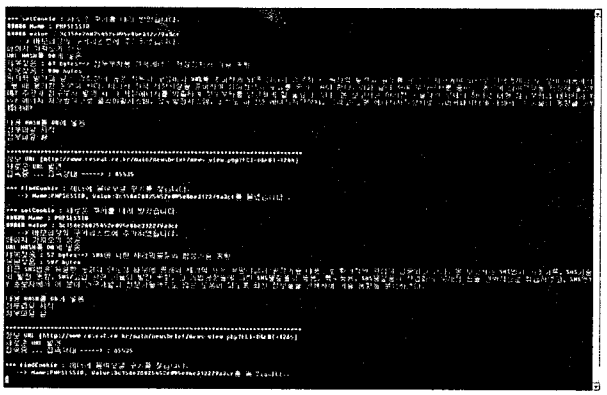

그림 2 분산되어 있는 분석/동향 데이터 수집

3.3 웹문서 데이터 구축시스템

웹문서 데이터 수집 시스템은 KISTI의 약 40여개 웹 사이트에 분산되어 있는 데이터베이스 형태가 아닌 HTML 문서나 게시판 또는 자료실의 게시물들을 그림 2와 같은 HTML 기반의 콘텐트 수집 시스템을 이용하여 일정 주기로 수집하고 MySQL을 이용하여 DB화 한다. 이때, 수집된 데이터들이 통합검색 되도록 하기위해 일정한 규칙을 갖는 KISTI 통합검색 고유 관리번호를 부여하고, pdf, hwp 등의 파일타입을 갖는 전자원문이 포함된 데이터에 대해서는 KISTI의 지식정보 식별체계인 KOI(Knowledge Object Identifier)[10] 번호를 부여하여 관리한다. 또한, 데이터를 수집할 때마다 해당 웹사이트의 데이터와 이미 수집된 데이터와의 비교를 통해 동기화 작업을 수행한다. 이러한 데이터 수집 방식은 웹페이지에 표현된 정보이외의 데이터는 수집할 수 없는 단점이 있지만, 웹페이지에 표현되지 않은 정보는 대부분 관리를 위한 데이터이므로 수집되지 않아도 무방하다.

4. 결 론

본 논문에서는 KISTI가 생산·구축·보급하는 다양한 종류의 방대한 과학기술관련 데이터를 통합검색이 가능하도록 하기위해 KISTI에서 자체 개발한 정보검색 관리시스템인 KRISTAL-2002를 기반으로 통합검색 데이터베이스를 설계하고 구현하였다. 이때, 각 데이터베이스의 특성을 고려하여 데이터베이스 구축시스템을 방대한 데이터를 위한 대용량 데이터 구축시스템, 통합DB를 구성하기 어려운 데이터를 위한 메타키워드 구축시스템, HTML 문서 등을 위한 웹문서 구축시스템으로 분류하여 구성하였다. 향후에는 기 구축된 통합검색 데이터베이스를 보다 편리하고 신속하게 검색될 수 있도록 하는 과학기술 종합 서비스 시스템을 구현해야 할 것이다.

참고문헌

[1] 과학기술정보 통합검색 서비스 [On-line]. Available: http://www.yeskisti.net

[2] GIIS(Group for Intelligent Information System) [On-line]. Available: http://giis.kisti.re.kr

[3] 한국과학기술정보연구원, "KRISTAL-2002 사용자 매뉴얼," 2004.

[4] 한민족과학기술자네트워크(KOSEN) [On-line]. Available: http://www.kosen21.org

[5] 사이냅소프트 [On-line]. Available: http://www.synapsoft.co.kr

[6] 한국과학기술인용 색인서비스 [On-line]. Available: http://ksci.kisti.re.kr

[7] 사실정보(Fact Database) [On-line]. Available: http://fact.kisti.re.kr

[8] 생물 자원정보 네트워크 센터 [On-line]. Available: http://biodiv.kisti.re.kr

[9] 북한과학기술네트워크(NK테크) [On-line]. Available: http://www.nktech.net

[10] KOI(Knowledge Object Identifier) [On-line]. Available: http://koi.kisti.re.kr