분산 통합검색 프로토콜을 사용한 과학기술 종합정보시스템 구현

이민호, 정창후, 주원균, 서정현, 류범종
한국과학기술정보연구원
e-mail : cokeman@kisti.re.kr

Implementation of Science Technology Information System with Distributed Integration Searching Protocol

Min-Ho Lee, Chang-Hoo Jeong, Won-kyun Joo, Jerry Hyeon Seo, Beom-jong You
Dept. of Information System Development, KISTI

요 약

오늘날 과학기술 정보는 아주 다양한 곳에서 다양한 방식으로 생산· 유통된다. 그러므로 원하는 정보의 위치나 각 정보 소스의 사용법을 모르는 사용자들은 일일이 정보 소스를 찾아 들어가 검색을 수행하여야 하는 번거로움이 있다. 본 논문에서는 사용자들에게 여러 곳에 위치한 과학기술 정보들을 한번에 검색하여 제공할 수 있는 과학기술 종합정보 시스템을 구현한다. 이를 위하여 KISTI 와 ㈜엔퀘스트에서 공동 개발한 분산통합검색 프로토콜을 사용한다. 사용된 분산통합 검색 프로토콜은 XML 기반으로 정의되어 있으며, 원천서버의 위치와 질의 형식 등에 관한 메타데이터를 유지하고 있어 분산된 서버들을 쉽게 통합검색에 참여 시킬 수 있다.

1. 서론

인터넷의 급속한 발전으로 인해 정보가 기하급수적으로 증가하고 정보검색 엔진의 보급이 증대된 사실은 이제 많이 알려져 있는 사실이 되었다. 이렇게 많은 정보들이 여러 서버에서 서비스됨에 따라 증가되는 사용자들의 불편을 해소하기 위해 이들을 통합하여 검색하려는 시도가 많이 진행되었는데 이러한 시도들은 주로 정보 집합의 크기 변화가 거의 없는 정적 환경에서 연구되었다. 하지만, 정보 집합은 계속하여 새로운 정보가 추가되는 동적인 상황인 경우가 많으며 동적 분산환경에서 기존의 검색엔진이 갖는 한계를 극복하려는 노력이 진행되어야 한다. 과학기술분야의 정보들도 새로운 기술발전에 힘입어 동적으로 변경되며 다양한 여러 기관들로부터 방대하고, 끊임없이 생성되고 있다. 따라서, KISTI 에서는 과학기술 정보 이용자에게 편리함을 제공하기 위하여 과학기술 종합정보 서비스 시스템을 개발하게 되었다. 이러한 동적, 분산 환경에서의 분산통합검색을 위하여 KISTI 와 ㈜엔퀘스트에는 분산통합 검색 프로토콜을 공동 개발하였다.[1] 개발된 분산통합검색 프로토콜은 각 검색엔진들의 특성을 알지 못하더라도 원하는 결과를 얻을 수 있고, 검색 서버들을 쉽게 분산통합에 참여시킬 수 있도록 설계되었다.

2. 분산 통합검색 프로토콜

KISTI 와 ㈜엔퀘스트에서 공동 개발한 분산통합검색 프로토콜(이하 k-protocol)은 다양한 분산 통합 검색 방법중 기존에 널리 사용되고 있는 통신 프로토콜인 HTTP 를 이용하는 방법으로 일반적으로 많이 사용되고 있는 메타검색과 유사한 방법이다. 하지만 기존 메타검색이 다양한 웹 화면 결과를 파싱하여 결과를 통합하는 것과는 달리, 표준화된 질의· 결과 포맷(프로토콜)을 사용하며 이를 위해 모든 원천 시스템에는 각 시스템에 대한 질의 및 결과 변환기가 설치된다. 이를 통하여 분산통합 프로토콜을 다음과 같은 장점

들을 갖는다.

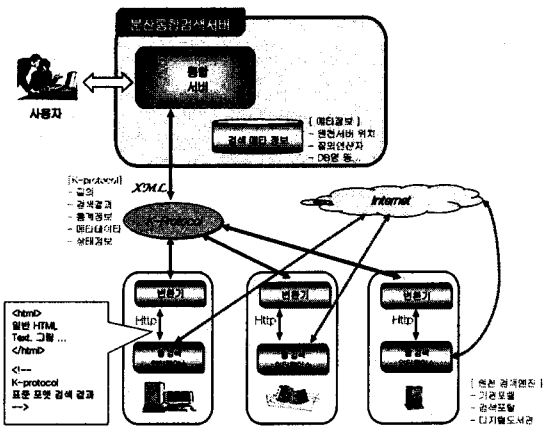
개발된 분산통합 프로토콜의 장점

- 원천 검색 시스템의 물리적 모델에 구현 받지 않는다.
- 웹을 사용한 개방적인 형태로서 쉽게 분산통합 검색에 참여시킬 수 있다.
- 기존 원천 검색 시스템의 검색 서비스에 영향을 주지 않는다.
- 프로토콜의 확장을 쉽게 할 수 있다.

k-protocol 에서는 일반적인 표준 개념 모델을 사용하고 각 원천 시스템에서는 각각의 시스템에 맞는 물리적 모델을 선택하도록 하여 원천 검색 시스템들은 자신의 정보 소스에 대한 스키마를 외부에 공개할 필요는 없이 분산통합 검색을 수행할 수 있도록 하였다. 하부 구조에서 제공하는 정보 시스템 (원천 검색 시스템)들은 정보를 웹 상으로 끌어 올림으로써, 입력 및 출력에 있어서의 복잡도를 감소시키며, DBMS · IRS 등의 물리적인 시스템 구성과는 관계없이 분산통합 검색에 쉽게 참여시킬 수 있다.

사용자로부터 원천 서버까지의 질의와 원천서버로부터 사용자까지의 결과는 k-protocol 에서 정의된 XML 형식으로 전달되므로, 프로토콜의 추가적인 확장이나 변경을 쉽게 해준다. 또한, 원천 서버의 웹 검색 결과도 이 형식으로 쉽게 변경되도록 요구된다. 하지만, 기존 웹 화면의 변경을 가하도록 요구되는 것은 아니고 주석 형태로 추가되는 것이기 때문에, 기존 원천 검색 시스템의 검색 서비스에는 영향을 주지 않는다.

검색 과정을 통하여 이를 더 자세히 설명하면, 통합 서버로의 사용자 질의는 k-protocol 에서 정의한 표준화된 형태의 질의로 만들어진다. 이 질의는 원천 서버의 위치와 DB 이름이 적혀있는 전역 메타 정보를 참고하여 각 원천 서버에 설치되는 질의-결과 변환기로 전해진다. 질의-결과 변환기는 원천 검색 서버의 웹 인터페이스에 맞는 질의 방법으로 사용자 질의를 변형하는데 변형에 필요한 웹 검색 인터페이스 URL, 검색 연산자, 검색 필드의 종류와 개수 등의 정보는 질의-결과 변환기 내부에 있는 지역 메타 데이터에 미리 저장되어 있다. 이렇게 질의-결과 변환기는 원천 검색 엔진에 질의를 하는 것이 아니라 이들 정보가 끌어올려진 웹 인터페이스에 질의를 함으로써, 원천 검색 시스템의 물리적 모델로부터 영향을 받지 않는다. 웹 인터페이스에서의 검색 결과는 검색 후 처리를 단순화 시키고, 서버의 부하를 줄이기 위해서 k-protocol 에서 정의한 표준화된 검색 결과 형태로 출력되어야 한다. 그러나, HTML 주석 형태로 추가적으로 지정되므로 기존 검색 화면에는 아무런 영향을 끼치지 않아 약간의 추가적인 작업만으로도 원천 서버가 쉽게 분산통합 검색에 참여할 수 있도록 한다.



[그림 1] k-protocol 모델

따라서, K-protocol 에서 제안하는 시스템 모델은 다양한 정보 소스인 원천검색 시스템, 이들 원천 검색 시스템에 맞는 질의로 표준 질의를 변환하여주거나 검색 결과를 표준 검색 결과로 바꾸어주는 질의-결과 변환기, 사용자에게 통합검색 인터페이스를 제공하여주고 각 질의-결과 변환기로부터 받아온 결과를 통합하여주는 통합 서버 세 부분으로 이루어져있다. (그림 1 참조)

3. 과학기술 종합정보 시스템 구축

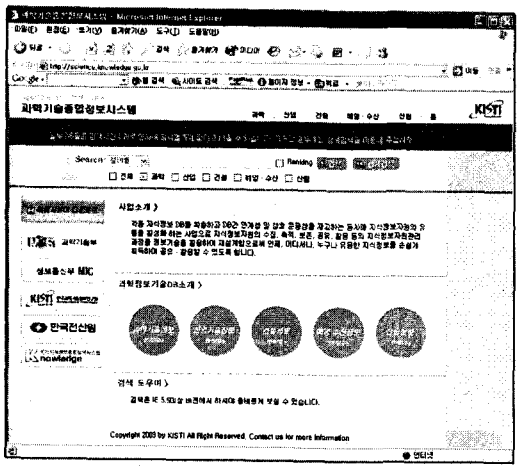
과학기술 종합정보 시스템은 과학, 산업, 건설, 해양· 수산, 산림 분야의 정보들을 종합적으로 검색 및 조회해 볼 수 있는 시스템으로 지식정보 연계 사업의 일환으로 구축되었다.[2] 현재 4 개 기관 총 25 개 DB 가 분산통합 검색 프로토콜로 연계되어 있다.

시스템 구축은 다음과 같은 절차로 진행되었다.[3]

A. 지원할 검색 방법 결정

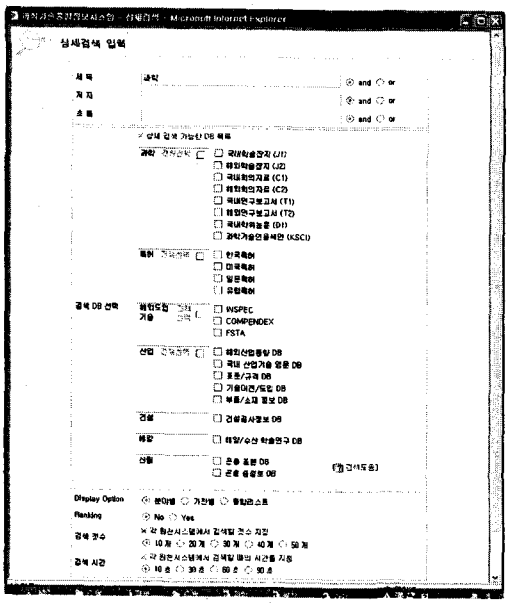
과학기술 종합정보 시스템은 간략검색과 상세 검색을 지원한다.

간략검색은 전체 필드를 대상으로 검색을 원하는 분야나 전체 DB 를 대상으로 검색하는 것이다. 상세검색은 DB 를 개별적으로 선택할 수 있으며, 필드를 제한하여 검색할 수도 있다. 또한, 검색 결과 화면을 분야별· DB 소장 기관별 또는 통합 리스트로 보여줄 수 있으며, 검색된 결과의 제목과 요약정보로부터 유사도를 추출하여 순위를 정할 수도 있다.



[그림 2] 간략 검색 화면

B. 각 DB들의 공통된 필드 추출

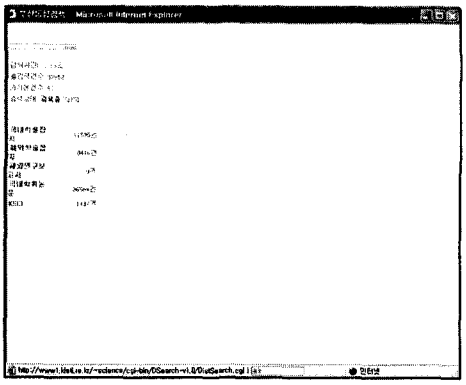


[그림 3] 상세 검색 화면

각 DB 마다 정의되어 있는 필드는 무척 다양하다. 통합검색에서는 공통된 필드만을 검색할 수 있으므로, 사용자에게 꼭 필요하면서 전체 DB 에서 공통된 필드를 잡는 일은 매우 중요하다. 25 개 DB 를 조사한 결과 제목-저자-초록 필드가 공통적으로 구성되어 있었다. 산림 DB 의 경우는 필드의 의미가 정확히 일치하지는 않아서 곤충명-제목, 채집자-저자로 유사한 의미를 찾아서 매핑시켰다.

C. 원천 서버의 검색 소요시간 측정

원천서버마다 검색에 소요되는 시간은 전부 다르다. 따라서, 원천 서버의 평균 검색 소요시간을 측정하여 적절한 타임 아웃값을 선택하는 일은 매우 중요하다. 25 개 DB 의 검색 시간을 측정하여 적절한 값인 30 초를 간략 검색 시간으로 잡아주었고, 상세 검색에서는 사용자가 선택할 수 있도록 하였다. 하지만 이렇게 다양한 시간을 선택 가능하도록 만들어도 문제가 될 수 있다. 최종적으로 검색이 완료된 후 결과를 사용자에게 보여줄 경우, 만약 일부 DB 가 검색에 소요되는 시간이 무척 오래 걸린다거나 서버가 다운되어 있을 경우 전체 결과가 화면에 보여지는 시간은 상당히 느려지므로 사용자의 분산통합검색 시스템에 기대하는 신뢰도는 떨어진다. 과학기술종합정보 시스템에서는 일부의 DB 만이라도 검색되어 결과를 받는 즉시 결과 건수와 DB 이름을 출력하면서 동적으로 화면을 갱신하도록 하여 사용자가 현재 검색이 잘 진행되고 있음을 느낄 수 있도록 하였다.



[그림 4] 동적 검색 화면

D. 원천 서버의 결과 화면 변경

k-protocol 표준 형식의 결과는 주석 처리되어 기존 HTML 결과 화면에 추가되어야 한다. 따라서 각 관계기관에 협조를 요청하여 원천 서버 결과 화면을 변경하였다. 추가되는 항목들은 다음과 같다.

① *DATABASE_NAME* : 검색 결과를 보내는 주체인 원천 데이터 집합의 이름
② QUERY : 질의어
③ NUM_OF_TOTAL_RESULT : 총 검색 결과 레코드의 수
④ NUM_OF_RECORD_PER_PAGE : 페이지당 보여줄 레코드의 수이다. 즉 총 검색 결과를 나누어 보여줄 때 한 페이지에 보여줄 레코드의 수
⑤ THIS_PAGE_START_RECORD : 현재 페이지

에서 보여줄 레코드의 시작 번호

⑥ THIS_PAGE_END_RECORD : 현 페이지에서 보여줄 레코드의 마지막 번호

⑦ FIRST_RESULTPAGE_URL : 첫 페이지 보기 화면의 URL

⑧ PREVIOUS_RESULTPAGE_URL : 전 페이지 보기 화면의 URL

⑨ NEXT_RESULTPAGE_URL : 다음 페이지 보기 화면의 URL

⑩ LAST_RESULTPAGE_URL : 마지막 페이지 보기 화면의 URL

⑪ BEGIN_RESULT_PAGE : 레코드 리스트의 시작을 표시

⑫ RECORD_TITLE : 레코드의 제목

⑬ RECORD_URL : 레코드의 URL

⑭ RECORD_ABSTRACT : 레코드의 요약 설명

⑮ END_RESULT_PAGE : 레코드 리스트의 끝을 표시

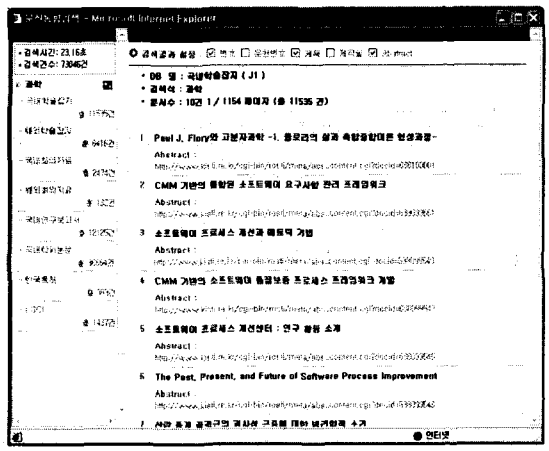
E. 메타데이터 작성

검색을 위해서 전역 메타데이터에 25 개 DB 를 담당하는 각 질의-결과 변환기의 위치와 DB 이름은 적어주었으며, k-protoco 표준 질의 형태의 사용자 질의를 원천 검색 인터페이스에 맞는 실질적인 질의로 변경하기 위해서는 지역 메타데이터에 웹 인터페이스의 정보를 기록하였다. 지역 메타데이터에 기록할 정보들은 각 기관의 웹담당자에게 협조를 구하여 얻었으며, 다음과 같은 내용들이다.

기관명, 웹 사이트 URL, DB 명, 로그인 url, 대표 id, 패스워드, 쿠드셋, 파라미터 전달방식, DB 분야, 사용되는 검색연산자, 간략검색 URL, 간략검색에 사용되는 파라미터들의 이름과 의미, 검색 결과화면의 페이지 이동 URL, 페이지 이동시 사용되는 파라미터 정보, 상세검색 URL 과 파라미터 정보, 사용언어 등을 기록한다.

F. 사용자 인증 문제

과학기술 DB 중 일부 DB 는 반드시 사용자 인증을 거쳐야 검색 및 결과보기를 할 수 있다. 따라서, 질의-결과 변환기에 이들 DB 의 사용자 인증을 할 수 있는 모듈을 추가하였다. 인증을 거쳐야 하는 DB 들은 cookie 를 통한 인증방식이기 때문에, 먼저 대표사용자 ID 와 패스워드를 질의-결과 변환기가 로그인을 수행하는 URL 에 넣어 인증을 하고 cookie 값을 얻어온다. 그 후 이 cookie 값과 사용자 질의를 같이 검색 URL 에 던져 검색을 수행한다.



[그림 5] 최종 결과 화면

4. 결론 및 향후 연구

인터넷의 발달로 사용자들은 다양한 정보를 쉽게 확보할 수 있게 되었으며, 정보 제공자 역시 인터넷으로 쉽게 정보를 출판할 수 있게 되었다. 하지만 이런 점 때문에 사용자들은 다수의 정보 소스를 찾아야만 하는 불편함도 생기게 되었으며, 분산통합검색을 통하여 이를 해결하려는 여러 시도가 있었다. 본 논문에서는 기존 HTTP 위에서 동작하는 분산통합 검색 프로토콜을 이용하여 쉽게 과학기술 정보를 통합 검색하여 사용자에게 제공하는 시스템을 구현하였다. 물론 원천 서버에 약간의 변경은 필요하나 이를 표준화함으로써 향후 기능의 확장이나 원천 서버의 사용자 인터페이스의 변경에도 쉽게 대처할 수 있다.

향후 과제로는 현재 사용자 인증과정에서 검색 시마다 인증을 거쳐 인증서버에 부하를 주는 문제가 있으며, 대표 ID 와 패스워드를 질의-변환기의 지역메타정보에 기록하므로 다수의 사용자가 각자의 고유한 ID 를 가지고 사용하기는 어렵다. 통합 서버측에서 사용자 ID 와 패스워드를 받아 Single Sign On 할 수 있는 기능이 더 개발되어야 할 것이다.

참고문헌

[1] 정창후, 이중현, 이현숙, 김평, 양명석, 맹성현, 서정현, 김현(2001). " 분산통합검색을 위한 시스템 개발 ", 2001 년도 한국정보과학회 봄 학술발표논문집(B) Vol.28 No.382-384

[2] 과학기술 종합정보 시스템, http://science.knowledge.go.kr

[3] KISTI, "KISTI 분산통합기 설치 및 DB 공유절차", Technical Note 2003-정보시스템연구실-2