연·구·원·기·고-1

정보검색 관리시스템 KRISTAL-2002

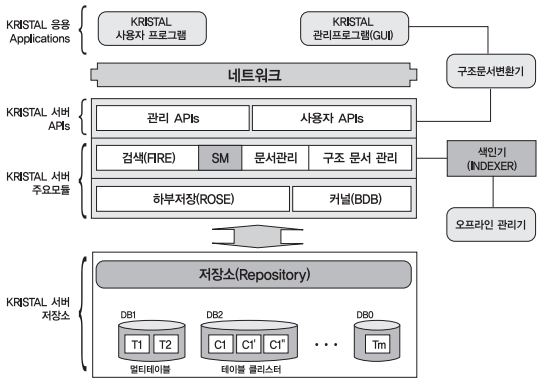
글 _ 류 범 종 · 정보시스템개발실 책임연구원 · ybj@kisti.re.kr
최 윤 수 · 정보시스템개발실 선임연구원 · armian@kisti.re.kr

1. 서 론

원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공의 여부를 결정 짓는 중요한 요소이다.

그러나 수많은 주제들에 대한 대용량의 정보로부터 한정된 시간 내에 원하는 정보를 찾는 것은 매우 어려운 일이다. 이러한 문제를 해결하기 위해 1960년대 초에 컴퓨터를 이용하여 원하는 정보를 찾도록 도와주는 "정보 검색"이라는 연구 분야가 확립되었으며, 지금까지 대용량의 문서를 효율적으로 적재 및 검색할 수 있는 방법에 대한 많은 연구들이 수행되어 왔다.

KISTI는 10여 년간의 연구개발 노력을 거쳐 정보검색엔진과 데이터베이스 관리 기능을 결합한 정보검색 관리시스템(IRMS : Information Retrieval & Management System) KRISTAL-2002 버전 1.0을 발표하였다. KRISTAL-2002는 평면 정보와 XML 구조정보 문서를 하나의 데이터베이스에서 저장, 관리하고 검색할 수 있도록 설계 및 개발되었다.



[그림 1] KRISTAL-2002의 전체 구조

[그림 1]은 정보검색 관리 시스템 KRISTAL-2002의 전체 구조를 도시한다. 본 논문에서는 KRISTAL-2002을 중심으로 시스템 개발 과정에서 고려한 사항들과 구현 내용을 각 구성 요소별로 나누어 설명한다.

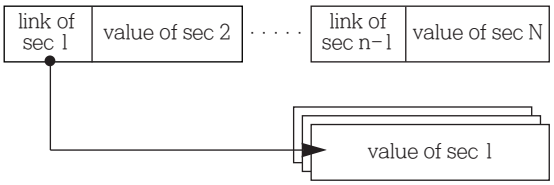
2. 저장엔진

저장엔진은 일반적인 서지데이터를 처리하기 위한 평면문서 관리기, 책자 형태의 XML 문서를 단편화하여 관리하는 구조문서 관리기와 저장된 문서들에 대한 빠른 접근을 지원하기 위한 역화일 접근 방식의 색인 관리기로 구성되어 있다.

2.1 평면문서 관리기

전통적인 정보 검색의 대상이 되는 문서는 제목, 초록, 본문, 저자 등과 같은 서지 항목들로 구성된다. 각각의 서지 항목들은 모든 문서에 걸쳐 정형화된 크기이거나, 서로 다른 크기를 갖는 특성을 지니며, 전체적으로 문서들은 요약문에서부터 책 한권의 크기에 이르기까지 매우 다양한 크기를 갖는다. 저장엔진의 평면문서 관리기는 이러한 가변적인 크기를

갖는 정형, 비정형 텍스트 문서를 저장 관리한다. 또한 내부적으로 UTF-8 형식으로 문서를 저장하므로, 다양한 기호 및 언어를 모두 처리하도록 구현되었다.

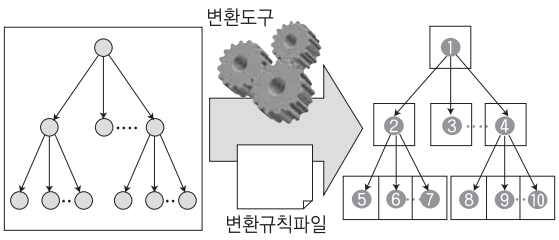


[그림 2] 평면문서 구조

KRISTAL-2002에서 문서는 섹션들의 집합으로 구성되며, [그림 2]는 평면 문서 관리기에서 다루는 문서의 논리적 저장 구조를 보여 준다. 문서를 구성하는 섹션들은 고정길이 문자열, 숫자 불리언 형식인 경우 실제 값이 저장되고, 가변길이 문자열인 경우 실제 저장소에 대한 링크를 갖는다. 문서에 대한 수정이 빈번한 경우, 레코드의 내부적인 불필요한 이동을 최소화하여 연산의 안정성과 신속성을 제공하도록 설계되었다.

2.2 구조문서 관리기

KRISTAL-2002가 취급하는 구조문서는 내부적으로 트리구조를 갖는 책자형식의 XML문서로 한정하고, 구조문서에 대한 계층적인 구조검색을 수행하기 위해 구조문서의 트리구조형태를 유지하면서 단편화하여 저장하는 방식을 제공한다. 이를 위해 구조문서의 단편화를 위한 단편화 기준과 관련 규칙들을 정의한 변환규칙파일을 제공한다. 변환도구는 변환규칙파일을 이용하여 구조문서를 KRISTAL-2002가 취급하는 평면문서로 변환한다. 단편화된 문서는 단편화되기 전의 부모, 형제간의 내부 링크에 대한 정보를 평면문서의 시스템 섹션에 저장한다.

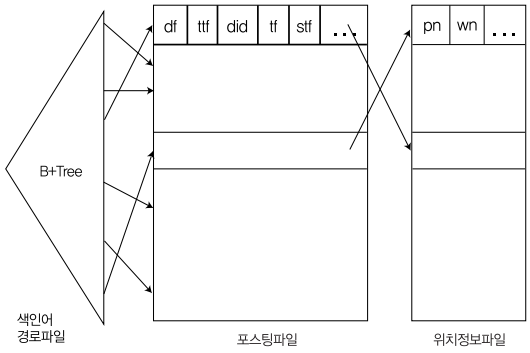


[그림 3] 구조문서 단편화 과정

[그림 3]은 구조문서를 단편화하는 과정을 보여준다. 구조문서 관리기는 각각의 단편화된 문서 단위의 삽입, 삭제, 수정, 이동 등의 연산 기능과 책자 단위 문서로의 복원기능을 제공하도록 설계되었다.

2.3 색인관리기

KRISTAL-2002의 색인관리기는 경로파일, 포스팅파일, 위치정보파일로 구성되며 [그림 4]와 같이 나타난다.



[그림 4] 색인 관리기 구성도

색인어 경로파일은 사용자의 질의에 나타난 단어를 효율적으로 탐색할 수 있도록 전체 문서파일에 출현한 색인어들을 B+트리 구조로 관리한다. 즉, 리프 노드에 색인어가 저장되며, 루트 노드와 내부 노드는 이들 색인어를 탐색하기 위한 접근 경로를 제공한다.

아울러 단말 노드는 색인어에 대응하는 포스팅파일 레코드에 대한 포인터를 색인어와 함께 유지한다.

포스팅파일은 색인어들이 출현한 문서들에 대한 정보를 저장하는 파일로서, 색인어마다 하나의 엔트리를 갖는다.

포스팅파일의 엔트리에는 색인어가 출현한 문서 식별자를 저장하는 것은 물론, 문서빈도수, 색인어 빈도수 등을 저장하여, 검색결과 우선순위를 계산할 때 이용한다. 위치정보파일은 사용자 질의의 단어간 근접도 연산을 지원하기 위해 문서 내에서의 색인어의 위치 정보를 저장한다. 이 위치정보는 문서 내에서의 단어 번호로써 표현되며, 색인어 추출 과정에서 색인어 추출 시스템에 의해 자동적으로 부여된다.

3. 검색엔진

과거 정보검색에 대한 수요가 일부 전문가 집단에 의해 주도 되었던 때와 달리, 여러 계층의 다양한 요구사항을 수용하기 위하여 다양한 방식의 정보 검색 모델을 제시하고 지원하기 위해 KRISTAL-2002는 검색모델로서 불리언(Boolean) 모델, 벡터(Vector) 모델과 벡터불리언(VectorBoolean) 모델을 제공한다.

불리언 모델은 논리곱(AND), 논리합(OR), 부정(NOT) 및 근접도(WITHIN/NEAR) 연산자를 기반으로 하여 섹션별로 정확한 일치 검색을 위해 사용하도록 설계되었으며, 전문가용 검색 모델로서 검색속도가 느리고 정확율(precision)은 낮지만, 정확한 재현율(recall)을 지원한다.

벡터 모델은 검색속도 향상을 위해 Pruning처리를 사용하고, 문서의 우선순위 연산을 수행하여 높은 정확율을 제공하므로 일반적인 웹 포탈 기반의 사용자를 위한 검색 모델로서 사용하기 위해 제공된다.
벡터 불리언 모델은 위의 두 모델의 장점을 상속하는 형태를 가진 모델로서, 불리언에서 지원하는 연산자(AND, OR, NOT, WITHIN/NEAR)와 벡터에서 제공되는 문서의 우선순위를 지원한다. 벡터 모델과의 차이점은 벡터 모델이 자동으로 Pruning처리를 하는데 비해, 벡터 불리언 모델은 사용자가 그 값을 지정하지 않으면 Pruning처리를 하지 않는다는 것이다.

KRISTAL-2002에서 사용되는 사용자 질의어는 국제 표준화 기구 중의 하나인 National Information Standards Organization(NISO)에서 1991년에 발표한 Z39.58-199x를 기반으로 한다. 이것은 온라인 정보 검색을 위한 사용자 명령을 표준화한 것으로서, Z39.58-199x의 FIND 명령어는 불리언 모델을 근간으로 하고 있다. 사용자 질의의 구문 형식은 Z39.58-199x의 FIND 명령어를 기초로 만들어졌으며, 다음과 같은 다양한 연산을 제공한다.

① 불리언 연산 : 질의를 구성하는 단어들 사이의 논리적인 관계를 명세할 수 있도록 AND, OR, NOT의 불리언 연산자를 지원한다.
② 근접도 연산 : 질의를 구성하는 단어들 사이의 문서 내에서의 상대적인 거리와 순서를 명시할 수 있도록 NEAR와 WITHIN 연산자를 지원한다.
③ 절단 연산 : 질의를 구성하는 단어들의 우측 절단을 지원한다. 이를 위해 사용자는 '*' 기호를 사용한다.
④ 결과 내 재검색 기능 : 시스템은 사용자가 입력한 질의들의 히스토리를 유지하며, 이를 접근하기 위한 SET 연산자를 지원한다.

4. 색인어 추출시스템

정보 검색에서 자동 색인은 문서의 내용을 대표할 수 있는 색인어를 추출하는 것을 말하며, 일반적으로 색인어 추출 방법은 정보 검색 시스템의 검색 효과에 중요한 영향을 미치는 것으로 알려져 있다. 현재 KRISTAL-2002는 데이터베이스에 대한 기본 섹션별 색인과, 기본 섹션들의 묶음인 통합 섹션별 색인을 지원하며, 색인어 추출 방식에 따라 섹션 단위, 음절단위, 어절단위, 형태소 단위의 색인 방법을 제공하고, 이와 더불어 한문의 경우 한자 한글 변화 옵션과 영문의 경우 스테밍 옵션을 제공한다.

섹션 단위의 색인은 섹션 값 전체를 하나의 색인어로 선정하는 방식으로, 섹션 값에 대한 완전 일치의 검색을 지원한다. 음절 단위의 색인은 한 음절 단위로 색인어를 생성하며, 한문의 경우 동형이음어, 이체자, 두음법칙 등을 적용한다. 어절 단위의 색인은 각 섹션에서 색인어로서의 가치가 없는 불용어를 제외한 모든 어절들을 원문에 나타난 형태 그대로 색인어로써 추출한다.

형태소 단위의 색인은 한글 문장에 대해 형태소 분

석을 수행하여 모든 어절들을 명사, 조사, 부사 등의 형태소 단위로 분리한 후 불용어들을 제거하고 색인어로서 의미가 있는 단순 명사들을 색인어로 추출한다. 복합명사의 경우 다양한 형태로 형태소 분석이 가능하므로, 각각의 경우에 대해 모두 개별적으로 색인어를 추출하여 재현율을 높일 수 있도록 구현하였다.

5. 데이터베이스 관리기

데이터베이스 관리기는 데이터베이스에 관련된 각종 관리 기능을 수행하는 응용 프로그램이다. 이러한 기능을 수행하기 위해서 관리할 데이터베이스, 하위 테이블스키마, 테이블 및 적재할 데이터에 대한 정보를 담고 있는 스키마 파일을 정의하고 해당 스키마 파일의 내용에 따라 사용자의 요청을 처리한다. 또한 세부 문서에 대한 처리는 제공하는 관리 API 및 사용자 API를 이용한다.

데이터베이스 관리기는 사용자가 편리하게 데이터베이스에 문서를 일괄 적재 및 추가할 수 있도록 설계 및 구현되었으며 다음과 같은 특징을 갖는다.

첫째, 전 세계 언어를 표현할 수 있는 유니코드로 문서 및 색인 정보를 저장함으로써, 이전에 표현하기 힘들었던 한자 및 고문자 등을 데이터베이스에 적재하고 검색 응용 프로그램에서 이용할 수 있게 되었다. 또한 적재할 문서의 인코딩 형식을 스키마 파일에 지정하면 한국어뿐만 아니라, 일본어, 중국어 등도 적재할 수 있어 국제화 된 소프트웨어의 특징을 가지고 있다.

둘째, 텍스트 파일과 압축 파일을 동시에 적재할 수 있다. 내부적으로 압축 파일(gzip)을 읽는 모듈을 사용하여 압축된 파일의 내용을 데이터베이스에 적재함으로써 원시 데이터의 크기가 매우 클 경우 저장 공간을 효율적으로 사용할 수 있는 장점이 있다.

셋째, 기본 키(primary key)를 제공한다. 적재할 문서가 여러 곳에서 가공되거나 작성자가 여러 명일 경우 문서가 중복되어서 생성될 수 있다. 문서 적재기는 중복된 문서가 데이터베이스에 적재되지 않도록 하는 기능을 가지고 있는데 이것이 기본 키다. 사용자가 스키마 파일에 기본 키로 사용할 하나 이상의 섹션을 지정하면 데이터베이스 관리기는 섹션을 조합하여 기본 키를 생성하고 기본 키를 체크해 중복되는 문서는 경고 메시지를 출력한 후 적재하지 않는다. 사용자가 기본 키를 지정하지 않을 경우에는 문서의 빠른 접근을 위해 데이터베이스 관리기가 기본 키를 내부적으로 생성하여 사용하고 있다.

6. 정보검색관리 통신 프로토콜

정보 검색 관리 시스템 KRSTAL-2002를 사용하기 위해서는 클라이언트와 서버 사이에 약속된 규약을 따르는 통신을 해야 한다. 이러한 통신 규약을 정보검색 관리 프로토콜이라고 한다.
KRISTAL-2002 프로토콜은 비세션 지향적인 프로토콜이다. 이는 서버가 접속 상태를 유지하지 않는다는 것을 의미하며, 따라서 한번의 연결에는 하나의 서비스만을 사용할 수 있다. 물론 시스템 설계자에 따라 TCP 계층에서 한번 연결한 후 여러 서비스를 요청할 수도 있다. 하지만 KRISTAL-2002 서버에서 일정 시간이 지나면, 연결되어 있고 작업 중인 상태라도 강제로 연결을 해제한다.

정보검색 관리 통신 프로토콜에서 제공하는 서비스

는 KRISTAL-2002 시스템과 클라이언트 사이의 다양한 정보 교환으로부터 이루어진다. 여기서 정보 교환은 XML 문서를 통해서 이루어지고, 양쪽의 시스템은 XML 문서에서 필요한 정보를 추출하여 적절한 작업을 수행하도록 한다.

객체에서 사용하는 파라미터 중 값이 문자열인 파라미터는 Base64 코드를 사용하여 엔코딩 된 메시지를 전송하고, 메시지 수신부에서는 다시 원래의 문자열로 Base64 디코딩을 수행한다.

메시지 구조	설 명
<Message>	〈Header〉와 〈Body〉로 구성됨
<Header>	
<Version> 1.0 </Version>	프로토콜의 버전을 표시
</Header>	
<Body>	
<Process>	서비스와 관련된 내용
Service	서비스명
</Process>	
<Object>	
Parameter	서비스에 관계된 파라미터들
Parameter	
</Object>	
</Body>	
</Message>	

[표 1] KRISTAL-2002 프로토콜 메시지 구조

7. 결 론

정보검색관리시스템 KRISTAL-2002는 대용량의 데이터베이스로부터 사용자의 질의를 만족하는 문서들의 검색을 지원한다. 이 시스템은 NISO에서 제정 발표한 질의 구문 형식을 구현함으로써 불리언 연산, 근접도 연산, 절단 연산 등의 다양한 질의 기능들을 제공하며, UTF-8을 기반으로 하므로 영어, 한글 외에도 다양한 언어로 작성된 문서에 대해서도 저장 및 검색을 지원한다.

저장엔진은 일반적인 문헌정보인 평면문서에 대한 저장 및 관리와 책자 형태의 XML문서인 구조문서에 대한 단편화 방법과 변환도구를 통한 저장 및 관리 방법을 제공하며, 검색엔진은 사용자의 요구사항을 만족하기 위해 다양한 검색 모델을 제시한다.

KRISTAL-2002는 현재 과학 기술 정보 서비스와 역사통합 검색 시스템 등을 위해 활용되고 있다. 이 시스템은 이외에도 전자 도서관, 고문서 편찬 시스템, 특허 정보, 인명 정보, 게시판 등과 같은 정보 서비스에 폭넓게 이용될 수 있을 것으로 기대된다.