대용량 멀티미디어 문서를 위한 정보검색 시스템

진두석, 최윤수, 안성수
한국과학기술정보연구원 정보시스템연구실

Information Retrieval System for Very Large Multimedia Docuement

Du-Seok Jin, Yun-Su Choi, Sung-Soo Ahn
Dept. of giis, Korea Institute of Science and Technology Information

요 약

인터넷의 급속한 보급과 함께 멀티미디어 문서의 사용에 대한 사용자의 요구가 증가하고 이에 따라 멀티미디어 문서 정보 검색에 관련된 연구들이 국내외적으로 활발하게 진행되고 있다. 멀티미디어 문서는, 데이터의 양이 방대할 뿐 아니라 데이터가 비정형화되어 있기 때문에 분석이 복잡하며 또한 효율적으로 저장, 검색하기가 매우 어렵다. 그러므로 이를 위해서는 적절한 멀티미디어 자료 저장 구조를 지닌 정보 검색 시스템이 절실히 요구된다. 따라서 본 논문에서는 대용량 멀티미디어 문서에 적합한 저장 구조를 가진 정보검색 시스템을 제안한다.

1. 서론

최근 멀티미디어 문서의 증가에 따라 멀티미디어 문서 정보 검색에 관한 연구가 활발히 진행되고 있다. 비정형화된 대용량 멀티미디어 문서를 위한 효율적인 정보검색 시스템은 첫째, 멀티미디어 문서 중에서 텍스트 정보에 해당하는 정보를 안정적이고 빠르게 검색하기 위한 색인저장 시스템이 필요하다. 둘째, 멀티미디어 문서에 포함된 멀티미디어 객체의 정보를 효율적으로 저장 및 관리할 수 있는 멀티미디어 객체 저장 구조가 필요하다. 따라서 본 논문에서는 빠르고 안정적인 색인구조와 대용량 객체저장 시스템을 이용한 멀티미디어 문서 정보검색 시스템을 소개한다.

2. 멀티미디어 문서 정보검색시스템

본 논문에서 사용한 멀티미디어 문서 정보검색시스템의 전체적인 구조는 [그림1]과 같다. 멀티미디어 문서 정보 검색 시스템은 크게 3가지 모듈로 구성된다. 첫째, 멀티미디어 문서를 저장시스템에 유효한 포맷 즉, 텍스트정보 와 멀티미디어 정보로 변환하거나 이전에 저장된 멀티미디어 문서를 검색하여 편집하는 작업을 위한 문서관리시스템과 둘째, 텍스트 데이터와 멀티미디어 객체를 저장 및 색인하는 저장시스템, 그리고 검색질의를 분석하여 다양하고 최적화된 검색을 처리하는 검색시스템으로 구성된다. 3가지 모듈의 특징은 다음과 같다.

문서관리시스템은 비정형화된 멀티미디어 문서를 저장시스템에 적합한 반정형문서로 변환 하여 저장시스템에 저장하고, 검색결과를 브라우징하는 역할을 수행하며, 아래와 같은 기능을 포함한다.

- 데이터베이스 생성 및 관리 기능
- 데이터베이스 벌크적재 및 백업 기능
- 멀티미디어 문서 오류검사
- 멀티미디어 문서 편집 및 브라우징 기능
- 멀티미디어 문서 실시간 삽입, 삭제, 수정 기능

저장시스템의 특징은 대용량의 멀티미디어 문서에 빠른 적재능력이 있으며, 데이터베이스의 압축기능을 사용하여 저장 공간을 줄일 수 있다. 또한 트랜잭션 처리를 통한 안정적인 멀티미디어 문서의 삽입, 삭제, 수정을 보장하며 클라이언트와 서버는 소켓으로 연결되어 있어 클라이언트의 요구사항을 실시간으로 처리하여 검색결과에 즉시 반영한다. 텍스트 정보로부터 색인어들을 추출하는 색인기는 한글, 영문, 숫자, 한자 형태소 분석을 위한 다양한 형태의 색인타입을 지원

하며 어절분석속도 향상을 위한 최적의 알고리즘을 사용하여 구현하였고, 보다 정확한 검색을 위한 특정 분야 전문사전을 이용한다.

검색시스템은 전문화된 검색문법을 제공하며 다양한 검색모델을 지원하고 검색결과의 랭킹이 가능하다. 뿐만 아니라, 빠른 검색성능을 위해서 멀티 쓰레드를 이용한 분산검색을 수행하며, 셋 관리기를 사용하여 기존의 검색된 결과에 대한 결과 내 재검색이 가능하다.

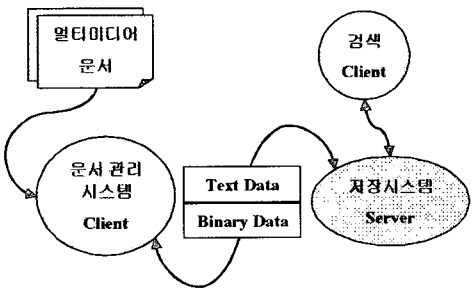

그림 1 멀티미디어 정보 검색 시스템의 구조

2.1 멀티미디어 문서 저장 구조

본 논문에서 제안하는 멀티미디어 문서 정보검색시스템은 비정형화된 멀티미디어 문서를 텍스트 정보와 멀티미디어 객체를 분리하여 서로 다른 레코드에 저장한다. 이는 검색이나 문서 수정시에 사이즈가 큰 멀티미디어 객체와 텍스트 정보가 같은 레코드에 저장되어있으면 텍스트의 내용을 접근하기위해서 멀티미디어 객체의 내용도 디스크로부터 읽어야 하는 부담이 줄이기 위해서이다. [그림2]와 같이 텍스트 레코드에는 다수의 텍스트 필드와 멀티미디어 객체의 포인터를 가지고 있고, 멀티미디어 레코드에는 멀티미디어 객체에 대한 설명과 내용이 저장된다.

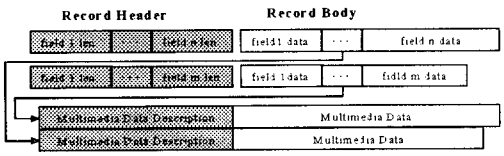

그림 2 멀티미디어 문서 저장 구조

멀티미디어 문서중 텍스트 데이터에 대한 다양한 검색모델을 지원하기 위해서 역파일 구조를 이용하여 색인어를 저장한다. 포스팅 파일에는 문서번호와 색인어 빈도수, 문단번호, 단어번호를 저장한다. [그림3]은 B+tree를 이용한 역파일 구조를 보여준다.

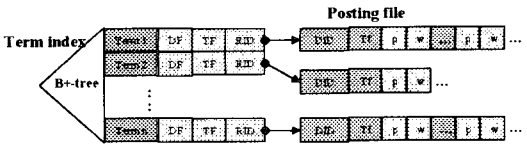

그림 3 색인어 저장 구조

2.2 정보검색시스템 서버

멀티미디어 문서 정보검색시스템 서버는 저장엔진, 데이터관리기, 검색기, 색인기, 셋관리기, 작업관리기로 구성된다. 각 모듈은 소켓이나 파이프 또는 직접 연결되어 상호간에 데이터를 전송한다. 소켓으로 연결된 모듈은 서로 다른 시스템에 위치할 수 있으며 파이프나 직접 연결된 모듈 간에는 공유메모리를 사용하여 필요한 정보를 교환한다. 각 모듈에 대한 설명은 다음과 같다.

첫째, 저장엔진(Storage Engine)은 멀티미디어 데이터베이스의 카탈로그정보를 관리하며, 텍스트 데이터에서 색인기를 사용하여 추출된 색인정보를 역파일을 이용한 저장구조에 저장한다. 그리고 멀티미디어 객체는 [그림2]의 문서 저장 구조에 맞게 변형하여 저장시스템에 저장한다.

둘째, 작업 관리기(Job scheduler)는 클라이언트 모듈로부터 검색 질의 또는 데이터베이스 관리에 관한 요구를 입력받아서 각 하부 모듈에 전달한다. 검색질의는 검색기(Fire)를 통해 처리하고, 멀티미디어 문서 관리는 데이터관리기(Data Manager)를 통해 처리한다. 서버의 구동시 작업관리기는 환경설정 파일에 지정된 개수만큼 프로세스를 생성하여 프로세스 개수만큼의 동시 클라이언트 요구를 처리할 수 있다. 또한, 멀티미디어 문서관리를 처리하는 중에는 검색질의에 대하여 잠금(Locking)기능을 수행한다.

셋째, 검색기(Fire)는 작업 관리기로부터 받은 검색질의를 처리하는 부분으로 질의어에 대해서 검색기는 Z39.58의 FIND 명령어를 바탕으로 하여 질의어를 확장한다. 확장된 문법은 불리언 연산자 AND(&), OR(|), 근접도 연산자 NEAR(/N), WITHIN(/W), 관계연산자 =, <, =<, >=, > 등으로 질의어에 대한 최적화를 한다. 또한 Boolean, Vector 검색 모델을 지원하고 있다. Boolean 모델은 부울 질의를 만족하는 문서들을 검색 처리한다. Vector 모델은 부울 질의를 만족하는 문서들을 검색하고, 검색된 문서 결과를 문서 우선순위 결정 알고리즘에 의하여 결정된 문서를 가중치에 따라 랭킹한다. 그리고 검색 속도를 빠르게

하기위해서 멀티쓰래드를 이용한 분산검색을 수행한다. 즉 검색할 데이터베이스가 N개일 경우 N개의 쓰래드가 생성되어 각각 검색을 수행한 후 통합한다.

넷째, 색인기(Indexer)는 문서의 삽입과 질의처리에 필요한 다양한 색인타입의 형태소 분석을 위해 자동 한글 형태소 분석기를 사용한다. 자동 한글 형태소 분석기는 어절 분석 속도를 높이기 위하여 품사 사전의 구조화와 탐색 방법에 대한 다양한 접근 방법의 평가를 통해 최적의 알고리즘을 사용하였고, 시스템 구조를 디자인함에 있어서 모듈화된 하부 시스템의 유기적이고 효율적인 결합에 중점을 두고 개발하였다. 그리고 대부분의 형태소 분석 시스템에서 적용하고 있는 재귀적 복합명사 분석을 탈피하여 빈번한 재귀적 호출에 따른 시스템 부하를 줄이고 확장성을 도모하였다. 또한 다층적 수사 패턴 인식에 기반한 수사 형태소 분석 시스템을 개발하여 형태소 분석 시스템과 결합하였다.

다섯째, 데이터관리기(Data Manager)는 안정적인 온라인 멀티미디어 문서 관리를 목적으로 설계되었다. 작업관리기로부터 전달된 작업을 데이터관리기는 트랜잭션 처리를 통한 실시간 문서 갱신 작업을 수행한다. 따라서 에러가 발생할 경우 회복(Recovery)기능을 이용하여 복원 할수 있다.

마지막으로, 셋 관리기(Set Manager)는 검색 결과의 저장 및 관리를 담당하는 부분으로서, 사용자가 클라이언트에 검색 결과를 저장하지 않고서도 자신의 검색 결과를 탐색 및 관리 할 수 있도록 하고, 캐쉬(cache) 기능을 제공함으로써 검색기의 빠른 검색 수행을 보조한다. [그림4]는 멀티미디어 문서 정보검색시스템의 서버 구조를 보여준다.

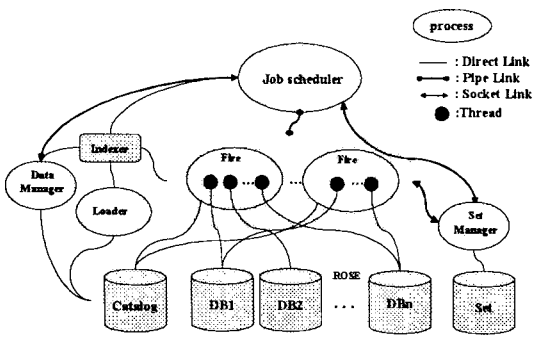

그림 4 정보검색시스템 서버 구조

2.3 정보검색시스템 클라이언트

정보검색 시스템의 클라이언트는 서버와 소켓통신으로 정보를 주고 받는다. 문서 검색, 관리에 관한 모든 요청은 서버의 작업관리기에 전달되어 서버의 작업처리 정책에 따라 수행되며, 초기 데이터베이스 생성 및 벌크적재, 백업관리등은 서버의 적재기(Loader)를 통해 처리된다. [그림5]는 클라이언트API를 이용한 문서관리시스템을 보여준다.

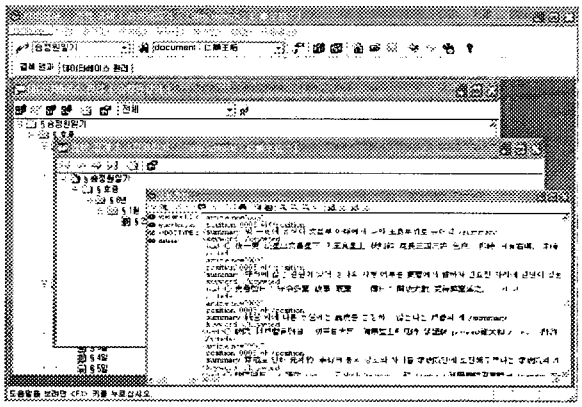

그림 5 문서관리 시스템(클라이언트)

3. 구현 및 성능

본 논문에서 구현한 멀티미디어 문서 정보검색시스템을 사용하여 웹사이트를 구현하고 성능을 테스트하였다. 대용량의 멀티미디어 문서를 테스트 하기위해서 약 3000만건(86GB)의 웹 문서와 약 1800만건(40GB)의 유전자 데이터 문서를 저장하고 검색 및 온라인 문서 관리를 테스트 하였다[2][3]. 테스트 결과 50시간정도의 빠른 적재시간과 안정적인 온라인 문서 관리를 수행하였다. 뿐만아니라, 평균 100만건 이상 검색되는 질의에 대해서 3초 이내의 검색 성능을 보인다. [표1]은 1800만건(40GB)의 유전자 데이터에 대한 검색성능을 나타낸다. 또한 유전자 데이터베이스는 저장 공간을 줄이기 위해서 문서내용과 색인정보를 압축하여 저장하였다. 따라서 압축하지 않은 데이터베이스의 경우는 [표1]에 나타난 시간보다 더 빠른 검색 성능을 보인다.

표 1 GenBank DB 검색 속도

Query	문서 수	검색 시간
clone	13,584,311	35.07(s)
human	6,364,105	15.13(s)
epidermal	4,525	0.422(s)
epidermal or growth	190,965	1.128(s)
NADH or oxydase or subunit	199,372	1.026(s)

4. 결론

본 논문에서는 비정형화된 대용량 멀티미디어 문서를 위한 정보검색 시스템을 구현하였다. 제안하는 멀티미디어 정보검색 시스템은 멀티미디어 문서 중에서 텍스트 정보에 해당하는 정보를 안정적이고 빠르게 검색하기 위한 색인저장 시스템과 멀티미디어 문서에 포함된 멀티미디어 객체 정보를 효율적으로 저장 및 관리할 수 있는 멀티미디어 객체 저장시스템을 구현하였다. 그리고 본 논문에서 제시한 멀티미디어 문서 정보검색을 이용하여 대용량 멀티미디어 정보검색에 효과적으로 적용 할 수 있음을 보였다.

[참고문헌]

[1] KRISTAL2000, http://ace.kisti.re.kr/~k2demo
[2] GenBank, http://blast.kisti.re.kr/genbank
[3] WebData, http://ace.kisti.re.kr/~k2web