□ 기술해설 □

바다-III : 전문 정보 검색 기능을 내장한 멀티미디어 데이타베이스 시스템

한국전자통신연구소 이미영* · 전성택* · 허대영** · 김명준**

● 목 차 ●

1. 개 요

컴퓨터 기술과 통신망 기술의 발전으로 멀티미디어 시대가 도래함에 따라 기존의 정형 자료에 비하여 상대적으로 대용량인 전문(Full-Text), 정지 화상(Image), 음성(Audio) 자료, 동화상(Video) 자료 등의 다양한 미디어 자료가 복합되어 있는 멀티미디어 정보가 대량으로 발생되고, 기존 정보들도 멀티미디어화 되고 있다. 이에 따라 멀티미디어 응용 시스템(멀티미디어 전자 우편, 문서 작성 시스템, 이미지 관리 시스템...)들과 전자 신문, 홈 쇼핑 등과 같은 새로운 멀티미디어 정보 서비스가 확산되고 있고, 문헌 정보, 인물 정보, 법률 정보 등의 정보 검색 시스템에서도 기존의 전문 위주의 정보 검색 서비스에서 멀티미디어 정보 검색 서비스로 급격히 변화되고 있다.

멀티미디어 정보는 그 특성상 자료의 저장, 관리 및 검색 방법이 기존의 정형 자료와는 다르다. 자료가 크고 복잡하므로 모형화하는 방법 및 저장하는 방법이 달라지고, 특히 검색 방법도 기존 데이타베이스 시스템의 검색 방법인 정확히 일치하는 자료를 검색하는 것과는 달리 유사성에 의한 검색을 요구한다.

그러므로 새로운 응용 분야의 요구를 충족시키기 위해서는 On-Line Transaction Processing(OLTP) 또는 경영정보 시스템 같은 분야의 정형 데이타베이스를 관리하던 기존 데이타베이스 관리 시스템과 전문 정보 검색 시스템, 이미지 검색 시스템 및 Geometric Information System(GIS) 시스템과 같은 비정형 자료 관리를 지원하던 정보 검색 시스템의 통합의 필요성이 제기되고 있다[1]. 즉, 기존 Data-Base Management System(DBMS) 기능(트랜잭션 처리, 동시성 제어, 회복, 질의, 자료의 일치성 보장)에 멀티미디어 정보 모형화 기능, 멀티미디어 자료에 대한 내용 기반 검색 기능, 대용량 자료 저장 및 관리 기능 그리고 실시간 처리 기능 등을 지원할 수 있는 멀티미디어 데이타베이스 시스템이 필요하다[1].

이에 따라 DBMS 분야에서는 멀티미디어를 지원하기 위해 멀티미디어 자료를 모형화하는데 용이한 객체지향 데이타베이스 시스템의 개발 또는 기존 관계 데이타베이스 시스템의 자료 모델을 확장하고 있고, 대용량 자료를 관리하기 위해 Binary Large Object(BLOB) 관리

*정 회 원
**종신회원

기능을 추가하고 있으며 정보 검색 시스템의 검색 기술을 채용하고 있다. 그러나 현재 나와 있는 멀티미디어 DBMS들은 멀티미디어 DBMS로서 갖추어야 할 많은 기능 중 부분적인 기능만을 지원하고 있으며, 아직까지 필드에서 검증되어 있지 못한 상태이고, 필요 기술들의 많은 부분들이 아직 성숙되지 못한 상태이다.

한국전자통신연구소에서는 초고속통신망 시대의 멀티미디어 정보 서비스를 지원하기 위해 전문 정보 검색 기능을 내장한 멀티미디어 데이타베이스 시스템인 바다-Ⅲ을 4개년(94. - 97.)에 걸쳐 개발하고 있다.

본 논문에서는 바다-Ⅲ 멀티미디어 데이타베이스 시스템에 대해 서술하기로 한다. 2장에서 바다-Ⅲ을 개발하면서 고려되었던 사항들에 대해 서술하고, 3장에서는 바다-Ⅲ의 구조 및 바다-Ⅲ 엔진의 주요 특징들인 객체지향 자료 모델, 전문 정보 검색 기능이 통합된 검색 모델, 자동 색인 방법, 그리고 멀티미디어 자료 관리 등에 대해 기술하고, 마지막으로 결론을 맺기로 한다.

2. 바다-Ⅲ 멀티미디어 데이타베이스 시스템 개발시 고려 사항

바다-Ⅲ을 설계함에 있어서 멀티미디어 DBMS로서 지원해야 할 기능 중 주요 고려 사항으로, 바다-Ⅲ이 지원하는 다양한 단일 미디어들로 구성되는 복합 멀티미디어 정보의 효율적인 모형화를 위한 자료 모델(관계형 자료 모델 또는 객체지향 자료 모델)의 선정, 멀티미디어 자료중 내용 기반 검색 방법을 제공하는 전문 검색 모델 및 이를 위한 시스템 구조에 대한 문제를 들 수 있다.

2.1 자료 모델

멀티미디어 정보 서비스를 하기 위해서는 복합 멀티미디어 정보를 갖는 멀티미디어 문서를 표현하는 방법이 있어야 한다. SGML(Standard Generalized Markup Language)[2]은 다양한 문서의 논리적 구조를 표현할 수 있는 범용 마크업 언어로 멀티미디어 문서를 표현하는 수단으로 널리 이용되고 있다. 예를 들어 현재 World Wide Web(WWW)에서 통용되고 있는 HTML(HyperText Markup Language)는 SGML로 특정 DTD(Document Type Definition)를 정의한 것이다. 바다-Ⅲ에서는 SGML이 멀티미디어 정보를 표현하는데 가장 적합한 방법이라는 인식하에 SGML 문서를 모형화하는데 적합한 자료 모델을 선택하기로 한다.

SGML은 범용 마크업 언어로 DTD를 정의하여 문서의 논리적 구조를 표현할 수 있게 한다. DTD는 문서의 논리적 구조를 이루는 구성요소(Element)들과 이들간의 관계를 정의한 것이다. 구성요소들은 또 다른 구성요소들을 내포할 수 있으므로 구성요소간에 트리 구조를 형성한다. 예를 들어 책이란 문서를 정의한다면 DTD에는 book이란 구성요소가 있고, book은 title, author 구성요소와 여러 개의 chapter 구성요소로 이루어지고, chapter는 title 구성요소와 section 구성요소를 포함하고, section 구성요소는 paragraph 구성요소들을 포함한다고 정의할 수 있다.

특정 DTD 형태를 갖는 각 문서들은 문서내에 DTD에 정의된 구성요소들을 마크업 태그로 하여 문서를 기술한다. 즉, 마크업 태그들은 문서의 내용내에서 어떤 구성요소들이 나타났는지와 어떤 순서로 나타났는지를 가르킨다. SGML 문서 트리의 중간 노드들은 구성요소들의 구성 정보를 갖고 있고 트리의 마지막 노드에 실제 정보가 들어 있다.

이와 같은 논리적 구조를 갖는 SGML 문서의 모형화는 객체지향 자료 모델이 가장 적합하다. DTD에 정의되어 있는 각 구성요소 타입은 객체지향 개념의 클래스로 대응될 수 있고, 마크업 태그들과 실제 내용들로 구성된 각각의 문서는 구성요소별로 나뉘어 각 대응 클래스의 객체들로 표현될 수 있다. 중간 노드들의 구성요소에 대응되는 클래스들은 여러 객체들로 구성된 복합 객체로 표현된다. 그러므로 바다-Ⅲ에서는 이와 같은 개념을 지원하는 객체지향 자료 모델을 지원하기로 한다.

2.2 전문 자료에 대한 검색 모델

멀티미디어 자료에 대한 내용 기반 검색 기

법은 전문, 정지 화상, 기하학적 도형 등에서 많은 발전을 이루고 있으나, 아직까지도 대부분의 시스템에서 전문을 제외한 멀티미디어 자료에 대한 내용 기반 검색을 위해서는 간접 방식을 택하고 있다. 즉, 객체에 대하여 묘사한 전문 정보를 이용하여 검색을 실시하는 방식을 제공하고 있다.

전문 정보 검색의 중요성은 현재 정보로서 가장 많이 이용되며, 효율적인 정보 검색 기술이 제공되고 있고, 다른 멀티미디어 자료에 대한 내용 기반 검색을 간접적으로 지원할 수 있다는 점이다. 또한 앞으로 멀티미디어 정보 검색에 있어서 가장 중요한 자연어 정보 검색 기술의 기반 기술로 활용될 수 있다는 것이다.

이러한 전문 정보 검색은 전문에 대한 전체 비교가 아니라 전문을 대표하는 색인어를 추출하고, 이를 이용한 색인어 검색을 실시하여 적합한 문서를 제공한다. 전문 정보 검색에서는 얻고자 하는 정보일 가능성이 있는 정보를 놓치지 않으면서도 빠르게 검색해 주어야 하며, 일반적으로 재현율(Recall)과 정확율(Precision)에 의해 검색의 효율성을 판단한다[3].

전문 정보 검색은 사용자 요구의 부정확성, 불확실성을 고려하여 색인어에 가중치를 부여한 검색, 이전 결과를 이용하는 검색, 검색 결과들에 순위를 제공하는 것, 시소러스를 이용하여 관련성이 있는 것을 검색해 주는 기능 등이 지원되고 있다.

전문 검색 모델은 불리안 모델, 확장 불리안 모델, 벡터 공간 모델 등 다양한 모델이 존재한다. 불리안 모델에 의한 검색은 효과적인 검색과 비교적 쉬운 질의어로 널리 사용되고 있으나, 질의의 불확실성에 대한 고려가 부족하며, 색인어에 가중치 부여 및 검색 결과들에 순서를 부여하는 기능이 없다는 단점이 있다[3]. 벡터 공간 모델은 문서에 가중치를 줄 수 있지만, 구문 처리나 동의어 처리가 가능하지 않다. 확장 불리안 모델은 불리안 연산자를 이용하면서도 랭킹이 가능하도록 한 것으로 전문 정보 검색 모델로 널리 채택되고 있다.

전문 검색에서는 효율적인 접근을 위해 전문의 색인어에 대한 인덱스를 제공한다. 인덱스는 지원되는 검색 기능에 따라 역화일, 시그너쳐 등이 사용된다. 일반적으로 시그너쳐는 저장 공간을 적게 사용하는 반면 가중치를 고려한 검색을 할 수 없고, 역화일은 저장 공간은 많이 사용하나, 가중치를 고려한 검색이 가능하고, 검색 속도도 빠르다.

바다-III에서는 불리안 모델의 데이타베이스 검색과 확장 불리안 모델의 전문 정보 검색을 통합하여, SGML 문서에 대한 검색을 지원하기로 한다. 예를 들어 위에 언급한 책이란 문서가 데이타베이스내에 모형화되어 들어 있다면 "book중에서 chapter의 title에 객체지향 자료 모델이 들어 있고, section에 계승에 대한 설명이 들어 있는 문서를 찾아 주세요"라는 질의를 할 수 있고. 이것은 chapter, title, section이 book 이란 문서와 갖는 복합 관계에 따라 데이타베이스 질의를 구성하고, section처럼 전문 검색이 필요한 것에 대해서는 전문 정보 검색을 실시하는 통합 질의로 지원한다.

2.3 전문 정보 검색 기능의 통합 방안

데이타베이스 시스템 기능과 전문 정보 검색 기능을 동시에 제공하려는 시도는 여러 곳에서 오랫동안 진행되어 왔다. 다양한 방법 등이 제시되었는데 이를 분류하면 밀결합 방식(Tight-ly-Coupled Integration, 내부 통합)과 느슨한 결합 방식(Loosely-Coupled Integration, 외부 통합)으로 나눌 수 있다[4, 5]. 밀결합 방식은 두 기능이 하나의 시스템으로 통합되어 같은 데이타베이스내에서 관리하는 것이고, 느슨한 결합 방식은 두 시스템이 독립적으로 존재하고 자료도 필요에 따라 중복 관리하며, 이 두 시스템을 이어주는 특정 모듈이 있어 이에 의해 통제되는 방식이다.

밀결합 방식을 채택한 시스템으로는 관계 데이타베이스 시스템에 중첩 릴레이션을 추가한 시스템도 있고, 기존의 관계 데이타베이스 시스템에 사용자가 정의한 인덱스를 생성할 수 있도록 확장하여 지원한 시스템이나, ORION에 전문 정보 검색 기능을 추가한 것 등이 있다[5].

느슨한 결합 방식은 통제의 주체가 누구냐에 따라 3가지 방식으로 구현될 수 있다[4]. 첫번째 방법은 데이타베이스 시스템과 정보 검색

시스템을 동급으로 취급하고 이 두 시스템을 통제하는 제어 모듈을 제공하는 것으로 INQUERY와 IRIS를 통합한 COINS가[6] 이에 해당한다. 이 방법은 제어 모듈의 능력에 따라 시스템의 기능이 달라지고, 두 시스템의 기능을 충분히 활용하기 위해서는 제어 모듈의 규모가 상당히 커지게 된다. 두번째 방법은 정보 검색 시스템이 통제의 주도권을 잡고 데이타베이스 시스템이 필요할 때 이를 활용하여 수행하는 것으로, 이 방법은 데이타베이스 시스템의 기능을 충분히 활용하지 못하는 단점을 갖는다. 세번째 방법은 데이타베이스 시스템이 주도권을 가지며, 정보 검색 시스템은 데이타베이스 시스템의 통제하에 있는 것으로, 느슨한 결합 방식중에 가장 좋은 방법이며, TextMachine과 OpenODB을 통합한 것[5]과 INQUERY와 VODAK를 통합한[4] 예가 이에 해당한다.

느슨한 결합 방식은 기존의 데이타베이스 시스템과 정보 검색 시스템을 적절히 활용하여 적은 노력으로 두 가지 기능을 제공할 수 있고, 어떤 특정 시스템에 얽매이지 않고 자유롭게 시스템을 선택하여 구축할 수 있다는 장점이 있으나, 자료의 중복으로 인한 자료들간의 일치성 유지 문제에 어려움이 있고, 두 시스템에서 수행되는 트랜잭션을 관리해야 하는 부담이 있다.

밀결합 방식은 하나의 시스템내에서 모든 정보를 총체적으로 관리하므로 트랜잭션 처리가 용이하고, 백업, 회복 등 버전 관리가 용이하다는 장점이 있다.

바다-III에서는 전문 정보 검색이 멀티미디어 자료 검색을 위한 가장 기반이 되는 검색 기법이라는 인식하에 시스템내에 통합하여 총체적으로 관리할 수 있는 밀결합 방식으로 전문 정보 검색을 지원한다.

3. 바다-III 엔진

본 장에서는 바다-III이 2장에서 고려한 사항을 중심으로 바다-III의 구조 및 자료 모델, 검색 모델, 멀티미디어 자료 관리 방법에 대해서 설명한다.

3.1 시스템 구조

그림 1은 바다-III의 시스템 구성도로서 각 블록들의 기능은 아래와 같다.

- 바다-III/MIDAS
  MIDAS(Multiuser Index-based Data Access System) 블록은 디스크 관리, 트랜잭션 관리, 회복 관리 등의 기능을 수행하는 바다-III의 최하위 저장 시스템[7]으로 일반적인 DBMS 인덱스 관리 외에 정보 검색용 인덱스 관리 기능을 수행한다.
- 바다-III/OK
  OK(Object-oriented Kernel) 블록은 객체지향 자료 모델을 제공하며 클라이언트/서버 구조를 지원한다[8, 9]. 또한 정보 검색 기능을 내장하고 있는 블록이다.
- 바다-III/C++
  바다-III의 사용자 인터페이스인 C++ 바인딩을 제공하는 블록으로 C++ 클래스 라이브러리 형태로 제공된다.
- 바다-III/VM
  VM(Visual Master)은 바다-III의 데이타베이스 관리자가 사용할 데이타베이스 관리 도구로서 데이타베이스를 효율적으로 관리하기 위한 Graphical User Interface(GUI) 형태의 사용자 도구로 스키마 브라우징 및 편집, 객체 브라우징 및 편집, 객체 적재 및 하적 기능 등을 제공한다[10].
- 바다-III/Webgateway
  WWW를 통하여 바다-III의 데이타 서비스를 제공하는 블록이다.

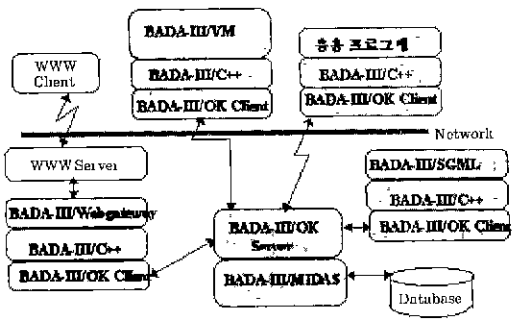

그림 1 바다-III 시스템 구성

● 바다-III/SGML
SGML 문서를 관리하는 블록으로 SGML 문법의 오류를 Parsing 하고 대량의 SGML 문서를 데이타베이스화 하는 기능을 제공한다.

위의 블록 중 바다-III 엔진은 객체지향 자료 모델을 지원하는 바다-III/OK와 자료 저장 관리기인 바다-III/MIDAS로 구성되며, 다음과 같은 특징을 갖고 있다.

- 객체지향 자료 모델 지원
- 전문 정보 검색 지원
- 멀티미디어 자료의 저장 관리 지원
- 동시성 제어 및 회복 기능 지원

바다-III 엔진 구조를 상세히 살펴보면 그림 2와 같이 전문 정보 검색 기능이 데이타베이스 시스템내에 통합된 밀결합 방식을 채택하고 있다.

바다-III/OK의 스키마 관리기, 객체 관리기,

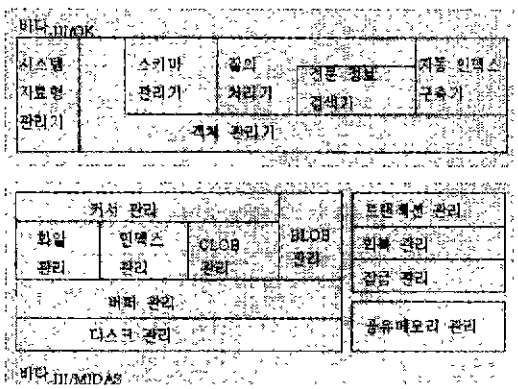

그림 2 바다-III 엔진 구조

시스템 자료형 관리기는 객체지향 자료 모델을 지원하고, 질의 처리기는 전체 검색을 통제하며, 데이타베이스 검색 기능을 수행하고, 전문 정보 검색기는 색인어를 이용한 정보 검색을 수행한다. 자동 인덱스 구축기는 문서로부터 색인어를 추출하여 정보 검색 인덱스를 생성한다.

바다-III/MIDAS는 실제적으로 디스크내에 모든 자료(객체, 인덱스)를 저장, 관리하는 업무를 담당하며, 다중 사용자에게 자료의 일치성을 유지시켜 주기 위해 잠금 기능, 복구 기능 등을 지원한다. 또한 자료의 빠른 검색을 위해 B+ 트리와 정보 검색 인덱스인 역화일 인덱스를 지원하고, 대용량 자료와 전문 자료에 대한 저장 관리를 지원한다.

3.2 바다-III 객체지향 자료 모델

바다-III은 멀티미디어 정보 서비스를 고려하여 다음과 같은 객체지향 자료 모델을 지원한다[11, 12, 13, 14].

● 객체
독립적으로 존재하는 모든 것은 객체의 대상이 되고, 객체의 상태가 변경된다고 객체가 달라지지 않으며, 모든 객체는 객체 식별자를 갖는다. 객체 식별자는 유일하며 변경 불가능한 것으로 객체의 소멸과 함께 없어지며 다시 사용되지 않는다. 객체는 상태를 변경할 수 있는 변경 가능 객체와 상태를 변경할 수 없는 변경 불가능 객체(즉, 값)가 있다.

● 복합 객체
객체가 값만으로 응집(aggregation) 관계를 가지면 단순 객체이고, 값이 아닌 다른 객체와 응집 관계를 가질 때 이를 복합 객체라 한다.

● 클래스
클래스는 타입과, 같은 타입을 갖는 객체들의 모임인 클래스 익스텐트의 의미를 갖는다.
타입은 비슷한 특성을 갖는 객체들을 추상화한 것으로, 객체의 정적인 특성을 나타내는 어트리뷰트와 객체의 동적인 특성을 나타내는 메소드로 표현된다. 어트리뷰트는 이름과 도메인으로 구성되고, 어트리뷰트 이름은 클래스내에서 유일하여야 한다. 메소드는 메소드 인터페이스와 메소드의 구현으로 구성되며, 메소드 인터페이스는 메소드 수행후 복귀값의 자료형과 메소드의 서명으로 구성된다. 메소드의 서명은 메소드명과 메소드 호출시 주어질 인자들의 자료형으로 구성되며, 이는 클래스내에서 유일하여야 한다.
기본적으로 객체의 상태는 캡슐화되어 있고, 캡슐화 수준으로 private, public을 지원한다.

- 계승
  계승은 단일 계승을 지원하고, 타입과 메소드의 구현 내용을 계승한다. 타입 계승은 어트리뷰트와 메소드 인터페이스를 모두 계승하는 것이고, 메소드의 구현 계승은 서브 타입에 속하는 객체가 슈퍼 타입에 정의된 메소드의 구현을 계승받아 사용할 수 있게 하는 것이다. 슈퍼 타입으로부터 계승받은 메소드의 구현은 서브 타입에서 재정의할 수 있고, 계승받은 어트리뷰트 이름이나 어트리뷰트 타입도 재정의할 수 있다.
- 시스템 제공 타입
  BLOB, Character Large Object (CLOB), COLLECTION 타입을 제공한다. BLOB은 대량의 비트 정보를 위한 타입이고, CLOB은 정보 검색이 허용되는 전문을 위한 타입이다. COLLECTION은 동질의 원소들로 구성된 집합 객체을 위한 것이다. BLOB은 바이트 위치에 따라 읽고, 쓰고, 삭제하는 연산 등을 제공하고, CLOB은 전문의 의미를 고려해 문장, 단어 등의 단위로 읽고, 쓸 수 있는 연산 등을 제공한다. COLLECTION은 집합 연산인 합집합, 교집합, 차집합 이외에도 원소의 첨가, 삭제, 검색 등의 연산을 제공한다.

3.3 검색 모델

바다-III은 기존의 데이타베이스 검색 방법인 불리안 모델과 정보 검색 모델중 가중치를 고려한 확장 불리안 모델을 통합하여 다음과 같은 모델을 제시한다[15].

- 프레디키트 연산자
  - 비교 연산자 : EQ, NE, GT, GE, LT, LE, CONTAINS
    전문에 적용된 EQ 연산자는 색인어 검색에 가중치를 고려하여 참일 확률이 어느 정도 이상이냐에 의해 판별한다.
  - 부분 매치 연산자 : LIKE, NLIKE
  - 근접성 정보 검색 연산자 : ADJ, NEAR, WITHINPARA
    전문 자료에만 사용할 수 있는 연산자로 색인어들간의 상대 위치, 혹은 한 단락내 같이 존재하는지 여부 등 근접성에 의한 검색을 지원한다.
- 불리안 연산자 : AND, OR, NOT
  기본적으로 단순 불리안 모델을 택하여 AND이면 참일 확률의 곱, OR이면 양 피연산자가 참일 확률중 더 큰 값, NOT이면 참일 확률을 1에서 뺀 값으로 계산하여 기준치 이상의 값만 검색한다. 정보 검색이 아닌 경우 프레디키트의 결과는 참일 확률이 1이다. 그러나 도메인이 같은 전문에 대한 검색이면서 불리안 연산자가 이용되면, 불리안 연산자의 의미가 확장 불리안 모델의 의미가 된다.
- 순위에 따른 결과값의 검색
  참일 확률이 높은 결과값부터 검색한다.
- 이전 검색 결과를 이용한 검색
  이전 검색 결과를 이용하여 그에 대한 재검색을 허용하므로서 사용자가 검색의 범위를 점차 줄여 나가며 원하는 정보를 얻을 수 있도록 한다.

통합 검색 모델을 지원하는 통합 질의 인터페이스의 기본 구문은 ODMG93[13]의 OQL (Object Query Language) 양식을 따르고, 정보 검색 질의 사양은 ISO8777 CCL(Common Command Language)[16]의 Find 구문을 참조하여, 다음과 같이 검색 결과를 얻는 방법에 따라 3가지 인터페이스를 제공한다.

- 하나의 객체 검색 인터페이스
  int oql(OM_REF<T> &result, const char *query)
- 검색 결과를 집합으로 얻는 인터페이스
  int oql(OM_LIST<T> &result, const char *query)
- 검색 결과를 커서를 통해 브라우징하는 인터페이스
  int oql(OM_ITERATOR<T> &result, const char *query)

각 인터페이스에서 query 스트링에 의해 정의할 수 있는 질의문은 다음과 같은 기본 형태를 갖는다.

```
SELECT project_list
FROM from_clause
```

WHERE search_condition

project_list는 추출할 어트리뷰트의 리스트 혹은 객체 자체를 명시할 수 있으며, from_clause는 검색할 클래스로 하나의 클래스 혹은 첫번째 명시한 클래스와 복합 관계가 있는 클래스만 명시할 수 있다. search_condition은 위 검색 모델에서 언급한 내용을 담고 있는 불리안 표현식이다.

3.4 자동 색인

자동 색인은 문서로부터 그 문서를 대표할 수 있는 중심 단어들을 자동으로 찾아내어 인덱스를 구축하는 것을 말한다.

일반적으로 전문 정보 검색은 구축된 인덱스를 이용하여 사용자가 요구한 자료를 빠르게 접근한다. 그러므로 전문 정보 검색을 하기 위해서는 전문으로부터 색인어를 추출하여 인덱스를 구축하여야 하고, 전문의 첨가, 삭제 등이 이루어지면 이를 인덱스에 반영하여 전문과 인덱스간의 일치성을 유지해 주어야 한다. 그러나 일반적으로 자동 색인 과정은 시간과 비용이 많이 드는 작업이므로 오프라인으로 수행시키며 문서들과 인덱스간의 불일치성을 어느 정도 감수한다. 바다-Ⅲ은 온라인 구축이 의미있는 응용 분야를 위해 색인어에 대한 가중치 부여 방식을 오프라인과는 다른 방식으로 한 온라인 색인 구축 인터페이스를 제공한다.

색인어 추출 방법은 한국어에 적용하기 용이하고, 비교적 구현하기 쉬운 형태소 분석 방법[17]을 이용하여 자동 색인을 한다.

바다-Ⅲ 자동 색인은 색인 대상 문서를 분석하는 입력 단계, 분석된 입력 문서에서 색인어를 추출하는 단계, 추출된 색인어와 문서와의 관련도를 나타내는 가중치 부여 단계, 마지막으로 추출된 색인어로 정보 검색 인덱스를 구축하는 단계로 나뉜다. 입력 단계에서는 형태소 분석의 기본 단위인 어절로 분리하고, 색인어로 사용할 필요가 없는 어절들을 제외시키는 전처리 작업을 실시하고, 단어들에 대한 위치 정보를 계산한다. 색인어 추출 단계에서는 명사 사전, 조사 사전, 불용어 사전을 이용하여 전처리가 끝난 단어들에 대해 최장일치법을 적용하여 색인어 후보를 생성하고 불용어를 제거한다. 가중치 부여 단계에서는 색인어가 적은 수의 문서에 나타나고 한 문서내에서는 여러 번 출현할 수록 가중치가 높게 나오도록, 색인어의 문서내 출현 빈도와 역문서 빈도의 곱에 의해 가중치를 구한다. 인덱스 구축 단계에서는 색인어 정보(색인어, 가중치, 문서내 출현 빈도, 위치 정보)를 디스크에 저장하여, 이후 검색시 이용한다.

3.5 멀티미디어 자료 관리

멀티미디어 자료는 자료의 크기가 큰 만큼, 자료의 저장 및 효율적인 접근 방법, 자료의 조작 등에 있어 특별한 고려를 하여야 하며, 미디어의 타입에 따라 사용되는 형태 및 자료의 규모가 다르므로 이에 대한 고려가 이루어져야 한다. 예를 들어 전문 자료는 정지 화상, 음성 등의 다른 멀티미디어 자료들에 비해 자료의 크기가 비교적 작은 중규모이면서 그 크기가 매우 가변적이라는 특성을 갖고 있다. 그러므로 바다-Ⅲ에서는 전문을 제외한 모든 멀티미디어 자료 관리에 기반이 되는 BLOB과 전문의 특성을 고려한 CLOB으로 구분하여 지원한다.

- BLOB 관리

일반적으로 크기가 큰 자료를 효율적으로 처리하기 위해서는 페이지 단위의 저장 공간 할당 정책을 사용한다. BLOB은 대용량의 자료로 구성되므로 페이지 단위로 할당하고, 페이지의 리스트 형태로 저장 관리한다. 페이지의 리스트로 관리하는 방식은 BLOB 자료가 부분 자료의 접근보다는 주로 전체 자료를 대상으로 읽고 쓰는 형태가 많은 것을 고려하여 이에 적합하도록 설계한 것이다.

- CLOB 관리

전문은 요약문과 같이 매우 작은 크기에서부터 책 한권의 크기까지 매우 가변적인 크기를 가지므로 BLOB과 같이 페이지 방식으로 할당하면 작은 크기에 대해서는 저장 공간의 낭비가 발생한다. 그러므로 자료의 크기가 작은 경우에는 기존의 일반 자료를 저장하는 방식으로 관리하고, 자료가 한 페이지를 넘는 경우에만 2단계 디렉토리 구조로 관리한다[18]. 2단계 디렉토리 구조는 CLOB처럼 자료의 크기가 방

대하지 않고, 자료의 일부에 대한 접근 및 변경 연산이 빈번한 환경에 적합한 방식으로 슬라이스라는 데이타 세그먼트들과 이 세그먼트에 대한 디렉토리 구조로 되어 있다.

또한 효율적인 접근을 위해 역화일 인덱스를 제공한다. 역화일 인덱스는 단어 색인 화일과 포스팅 화일로 구성되며, 단어 색인 화일은 사용자의 질의 단어를 효율적으로 탐색하기 위하여 전체 문서에서 출현한 색인어를 B+ 트리를 이용하여 구성한다. 포스팅 화일은 색인어와 문서간의 밀접도, 문서내에서 색인어가 출현한 위치 정보 등으로 이루어진다.

4. 결 론

바다-III은 멀티미디어 문서 모델링에 적합한 객체지향 자료 모델을 제공하고, 한글 전문 자료에 대한 내용 기반 정보 검색을 제공하는 멀티미디어 DBMS이다. 그 밖의 정지 화상, 음성, 동영상 등의 방대한 멀티미디어 자료의 저장 및 관리를 위하여 BLOB 자료 형태를 지원하고 있으며, 16 테라 바이트의 방대한 양의 멀티미디어 자료를 효율적으로 저장 관리할 수 있다.

바다-III은 전자 신문, 전자 잡지 등 네트워크를 통한 멀티미디어 정보 서비스와 문헌 정보, 법률 정보 및 인물 정보 검색 등 멀티미디어 정보 검색이 필요한 응용 분야에 효율적으로 이용될 수 있으며, 특히 정형 데이타베이스 검색과 비정형 데이타베이스 검색을 함께 다루어야 하는 응용 시스템 개발시 DBMS와 정보 검색 시스템을 함께 사용해야 하는 경제적인 낭비 및 비효율성을 제거할 수 있을 것으로 판단된다.

현재 바다-III의 개발 상태는 1차 버전이 구현 완료되어 이 시스템을 이용하여 WWW에서 4 기가 바이트 크기의 HTML 문서의 시범 서비스를 구축하고 있으며, 기존의 다양한 전문 정보 검색 시스템을 바다-III을 이용 재구축하는 시범 서비스들을 개발중이다. 97년말을 목표로 개발중인 바다-III의 최종 버전은 시범 서비스 개발과 운용을 통해 도출된 요구 사항을 반영시켜 보강할 것이다.

그리고 WWW 환경에서의 JAVA 지원의 중요성을 고려하여 JAVA 인터페이스를 제공할 예정이다. 97년 이후에는 정지 화상, 기하학적 도형 등 다양한 미디어에 대한 내용 기반 정보 검색의 지원 및 자연어 검색을 할 수 있도록 시스템을 확장할 예정이다.

참고문헌

[1] Setrag Khoshafian, A. Brad Baker. Multi-Media and Imaging Databases, Morgan Kaufmann Publishers, Inc., 1996.

[2] "Information Processing - Text and Office Systems - Standardized Generalized Mark-up Language(SGML)," ISO 8879-1986 (E), International Organization for Standardization, 1986.

[3] William B. Frakes, Richardo Baeza-Yates, Information Retrieval Data Structures & Algorithms, PTR Prentice Hall, 1992.

[4] Marc Volz, Karl Alberer, Klemens Bohm, "A Flexible Approach to Combine IR Semantics and Database Technologe and its Application to Structured Document Handling," GMD Technical Report No. 891, 1995.

[5] Tak W. Yan, Jurgen Annevelink, "Integrating a Structured-Text Retrieval System with an Object-Oriented Database System," Proceedings of the 20th VLDB Conference, pp.740-749, 1994.

[6] W. Bruce Croft, Lisa A. Smith, "A Loosely-Coupled Integration of a Text Retrieval System and an Object-Oriented Database System," 15th Ann. Intl SIGIR92. pp.223-232, 1992.

[7] 이진수, 박순영, 채미옥, 김준, 허대영, "MIDAS-II의 설계 및 구현", 한국정보과학회 93 가을학술발표논문집, 제20권 2호, pp. 183-196, 1993.

[8] 채미옥, 이미영, 김평철, 전성택, "OMEGA 시스템의 구조 설계", 한국정보처리학회, 95 춘계 학술발표논문집, 제2권 1호, pp.224-227, 1995.

[9] Mi-Ok Chae, Ki-Hyung Hong, Mi-

Young Lee, June Kim, Ok-Ja Cho, Sungtaeg Jun, Young-Kyun Kim. "Design of the Object Kernel of BADA-III : An Object-Oriented Database Management System for Multimedia Data Services." International Workshop on Network and System Management 1995, pp.143-152, 1995.

[10] 김완석, 이용현, 정광철, 전성택, "바다-III/VM의 설계", 한국정보처리학회, 96 춘계 학술발표논문집, 제3권 1호, pp.660-665, 1996.

[11] Atkinson, M., "The Object-Oriented Database System Manifesto," In Proc. 2nd Intl. Conf. on Deductive and Object-Oriented Databases, pp.40-56, 1989.

[12] Won Kim, Introduction to Object-Oriented Databases, The MIT Press, 1990.

[13] Cattell, R.G.G., The Object Database Standard : ODMG-93, Morgan Kaufmann, 1993.

[14] Bjarne Stroustrup, The C++ Programming Language, Addison-Wesley Publishing Company, 1991.

[15] 채미옥, 김준, 홍기형, "문서 서비스를 위한 객체 지향 질의 처리 API의 설계 및 구현", 한국정보처리학회, 96 춘계학술발표논문집, 제3권 1호, pp.632-635, 1996.

[16] "Information and Documentation - Commands for Interactive Text Searching," ISO 8777, International Organization for Standardization, 1993.

[17] 최기선, "한국어 정보검색", 정보과학회지, 제12권 제8호, pp.24-32, 1994.

[18] 이정기, 안성현, 김강석, 김연중, 장재우, 허대영, "MIDAS-III에 기반한 정보검색 하부구조의 설계", 한국정보과학회 95 가을 학술발표논문집(A), 제22권 2호, pp.259-262, 1995.

이 미 영

1981.2 서울대학교 식품영양학과(계산통계학 부전공) 학사
1983.2 서울대학교 계산통계학과 석사
1983.2~85.5 한국전기통신연구소 연구원
1988.3~현재 한국전자통신연구소 컴퓨터연구단 선임연구원
관심분야 : 데이타베이스 시스템, 객체지향 시스템, 분산 시스템

전 성 택

1977.2 서울대학교 공업교육과(전기공학사)
1977~80 동양정밀 주식회사 중앙연구소
1980~81 대화기기 주식회사
1982~83 University of Detroit (Master in Computer Science)
1983~92 University of Michigan(PhD in Computer Engineering)
1993~현재 한국전자통신연구소 컴퓨터연구단 선임연구원
관심분야 : 멀티미디어 DBMS, Real-time 시스템, Computer Graphics

허 대 영

1982.2 숭실대학교 전산학과 학사
1991.12 정보처리기술사
1982.3~현재 한국전자통신연구소 책임연구원, 데이타베이스연구실장
관심분야 : 데이타베이스 시스템, 객체지향 시스템, 소프트웨어 개발 방법

김 명 준

1978.2 서울대학교 계산통계학과 학사
1980.2 한국과학기술원 전산학과 석사
1986.5 프랑스 낭시(Nancy) 제1대학교 응용 수학 및 전산학과 박사
1980.2~81.6 아주대학교 종합연구소 연구원
1981.10~86.5 프랑스 낭시 전산학 연구수(CRIN) 연구원
1986.7~현재 한국전자통신연구소 책임연구원, S/W연구부장
관심분야 : 데이타베이스 시스템, 데이타베이스 설계, 트랜잭션 처리, 분산 시스템, 소프트웨어 개발 방법