□ 기술해설 □

정보검색 시스템 KRISTAL-II

연구개발정보센터 이준호*·안정수*

● 목 차 ●

1. 서 론
2. 커널
3. 저장 엔진
3.1 문서 관리기
3.2 색인 관리기
4. 카탈로그 관리기
5. 데이타베이스 관리기
6. 검색 엔진
6.1 사용자 질의
6.2 질의 연산
7. 색인어 추출 시스템
8. 웹 게이트웨이
9. 결 론

1. 서 론

원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공의 여부를 결정짓는 중요한 요소이다. 그러나 수많은 주제들에 대한 대용량의 정보로부터 한정된 시간내에 원하는 정보를 찾는 것은 매우 어려운 일이다. 이러한 문제를 해결하기 위해 1960년대 초에 컴퓨터를 이용하여 원하는 정보를 찾도록 도와주는 정보 검색이라는 연구 분야가 확립되었다.

지금까지 컴퓨터를 이용하여 대용량의 문서를 효율적으로 검색할 수 있는 정보 검색 시스템에 관한 많은 연구들이 수행되어 왔다. 정보 검색 시스템의 사용은 원하는 정보에 대한 접근을 용이하게 함으로써 다양한 분야의 정보들에 대한 수집 시간과 노력을 단축시킨다. 특히 관리할 정보의 양이 기하급수적으로 증가하고 있는 정보화 시대에서 효율적인 정보 검색 시스템에 대한 요구는 더욱 절실하다.

외국에서는 이미 1960년대 초에 일괄 처리 방식에 의한 MEDLARS 시스템이 구축된 이후 DIALOG, ORBIT, BRS, STAIRS 등과 같은 많은 정보 검색 시스템들이 상용화되어 현재까지 사용되고 있다. 그러나 이들 외국의 정보 검색 시스템은 주로 영어 문서들의 처리를 목적으로 개발되었기 때문에, 한글 문서들을 처리하는데 있어서 어려움을 지니고 있다.

한편, 정보 검색 분야는 많은 경우에 시스템의 사용 목적이나 환경에 따라 기능 확장이나 변경을 요구한다. 이러한 상황에서 외국 시스템을 그대로 도입하여 사용하는 것은 시스템의 기능 변경을 크게 제약하기 때문에, 우리로 하여금 원하는 기능을 외국 시스템에서 제공하기를 기다려야 하는 수동적인 입장으로 만든다. 따라서 이러한 문제점들을 근본적으로 해결하기 위해서는 정보 검색 시스템을 우리의 기술로 개발하려는 노력이 요구된다.

과학과 기술에 관련된 정보를 수집, 관리하고 이들 정보에 대한 검색 서비스를 제공하기 위해 설립된 연구개발정보센터는 1993년부터 4년여에 걸쳐 정보 검색 시스템을 개발하여 왔다[1, 2]. KRISTAL-Ⅱ라고 명명된 이 시스템은 부울 모델을 기반으로 하며, 영어 문서뿐만 아니라 한글 문서들의 검색을 지원한다. 그림 1은 정보 검색 시스템 KRISTAL-Ⅱ의 전체 구조를 도시한다. 본 논문에서는 KRISTAL-Ⅱ 버전 1.3을 중심으로 시스템 개발 과정에서 고려한 사항들과 구현 내용을 각 구성 요소

* 정회원

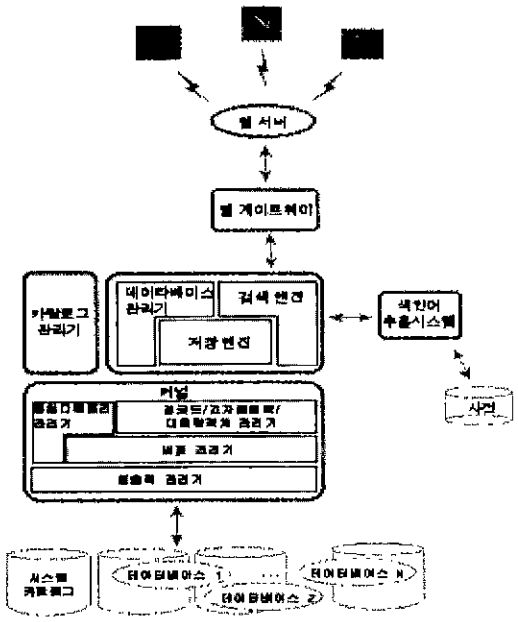

그림 1 KRISTAL-II의 전체 구조

별로 나누어 설명한다.

2. 커널

커널은 디스크에 데이타를 저장하거나 디스크에 저장된 데이타를 사용자의 메모리 영역으로 읽어들이는 작업을 제어한다. KRISTAL-Ⅱ 커널은 현재 입출력 관리기, 버퍼 관리기, 레코드 관리기, 긴 자료 항목 관리기를 위해 위스콘신 대학에서 개발한 저장 시스템 WISS를 상당 부분 이용하고 있다[5]. 이 저장 시스템은 효율적인 데이타의 입출력을 위해 유닉스 화일 서브 시스템을 우회하여 직접 디스크를 관리한다. 그리고 LRU 방식의 버퍼 교환 알고리즘을 사용하여 페이지 단위의 버퍼링을 수행하며, 또한 디스크에 생성된 모든 화일들의 속성 정보를 화일 디렉토리라는 자료 구조를 통해 저장 관리한다.

WISS에서 사용자는 레코드와 긴 자료 항목의 자료 구조를 통해 정보를 저장할 수 있다. 이들 자료 구조는 모두 일련의 바이트 스트림으로 간주되어 처리되며, 그것에 논리적인 의미는 부여되지 않는다. WISS에서 레코드는 하나의 페이지보다 작은 크기의 데이타를 유지하며, 긴 자료 항목은 하나의 페이지보다 큰 데이타를 저장 관리함으로써 레코드의 크기 제한을 보완한다. 예를 들면, 페이지 크기가 4K 바이트인 경우 최대 1.6M 바이트의 데이타 크기를 지원할 수 있다.

최근 정보 검색의 응용 범위가 이미지, 그래픽, 오디오 등과 같은 대용량 객체로 급속히 확산되고 있는 현실을 감안할 때, WISS에서의 레코드나 긴 자료 항목은 한계를 지니고 있다. 이러한 문제를 해결하기 위해 KRISTAL-Ⅱ 커널의 대용량 객체 관리기는 블롭(BLOB, Binary Large Object)이라는 개념을 사용하여 대용량 객체를 지원한다. 블롭은 가변적인 크기를 갖는 자료 구조로, 이론적으로 최대 2^{31} 바이트까지 지원할 수 있다. 대용량 객체 관리기는 사용자의 부분적인 갱신 요구에 대해 동적으로 대응할 수 있도록 EXODUS의 객체 관리 방식을 이용하여 블롭을 지원한다[4].

3. 저장 엔진

저장 엔진은 하위의 커널을 기반으로 비정형화된 텍스트 문서들을 저장하거나 이를 판독할 수 있는 기능을 지원한다. 그리고 저장된 텍스트 문서들에 대한 빠른 접근을 지원하기 위해 역화일 접근 방식의 색인 화일을 구현한다.

3.1 문서 관리기

일반적으로 정보 검색의 대상이 되는 문서는 제목, 초록, 본문, 저자 등과 같은 서브 항목들로 구성된다. 각각의 서브 항목들은 문서마다 서로 다른 크기를 갖는 특성을 지니며, 전체적으로 문서들도 요약문에서부터 책 한권의 크기에 이르기까지 매우 다양한 크기를 갖는다. 저장 엔진의 문서 관리기는 이러한 가변적인 크기를 갖는 비정형 텍스트 문서를 저장 관리한다.

KRISTAL-Ⅱ에서 문서는 섹션들의 집합으로 구성되며, 그림 2는 문서 관리기에서 다루는 문서의 논리적 저장 구조를 보여 준다. 문서를 구성하는 모든 섹션들의 수와 길이 정보가 문서의 헤더에 저장되며, 섹션들의 실제 내

그림 2 문서의 구조

용이 뒤이어 순서적으로 저장된다.

사용자는 레코드 식별자를 통해 화일내에 저장된 문서들을 접근할 수 있다. 그리고 문서내의 섹션은 섹션 번호를 통해 접근할 수 있다. 문서 관리기는 사용자에게 화일에 문서를 추가하는 연산, 문서나 섹션의 내용을 판독할 수 있는 연산, 그리고 화일내의 모든 문서들을 순차적으로 스캔할 수 있는 연산 등을 제공한다.

3.2 색인 관리기

그림 3에서 보듯이 KRISTAL-Ⅱ에서의 색인 화일은 색인어 경로 화일과 포스팅 화일로 구성된다. 색인어 경로 화일은 사용자의 질의에 나타난 단어를 효율적으로 탐색할 수 있도록 전체 문서 화일에 출현한 색인어들을 B+ 트리 구조로 관리한다. 즉, 리프 노드에 색인어가 저장되며, 루트 노드와 내부 노드는 이들 색인어를 탐색하기 위한 접근 경로를 제공한다. 아울러 리프 노드는 색인어에 대응하는 포스팅 화일 레코드에 대한 포인터를 색인어와 함께 유지한다.

포스팅 화일은 색인어들이 출현한 문서들에 대한 정보를 저장하는 화일로서, 색인어마다 하나의 엔트리를 갖는다. 대개의 경우 데이타베이스에서 색인어들의 출현 빈도는 각기 다르기 때문에 포스팅 화일의 엔트리는 앞절에서의

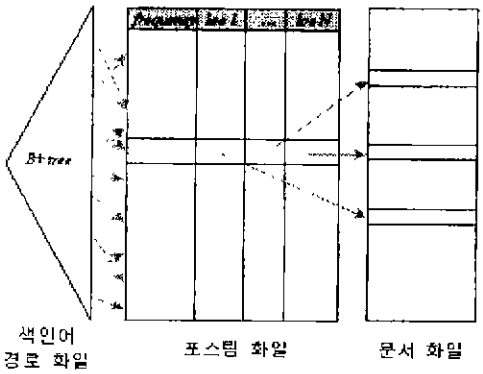

그림 3 색인 화일의 구조

문서와 같이 가변적인 길이를 갖는다. 따라서 색인 관리기는 문서 관리기와 같은 방식으로 포스팅 화일을 관리하고 있다.

포스팅 화일의 엔트리에는 색인어가 출현한 문서 식별자를 저장하는 것은 물론, 사용자 질의의 단어간 근접도 연산을 지원하기 위해 문서내에서 색인어의 위치 정보도 함께 저장한다. 이 위치 정보는 문서내에서의 단어 번호로써 표현되며, 색인어 추출 과정에서 색인어 추출 시스템에 의해 자동적으로 부여된다.

4. 카탈로그 관리기

카탈로그 관리기는 검색 엔진이나 응용 프로그램들이 데이타베이스를 보다 편리하게 접근할 수 있도록 데이타베이스 구조, 문서 구조, 색인 정보 등의 시스템 정보를 관리하며, 또한 이들에 접근할 수 있는 기능을 제공한다. 데이타베이스의 시스템 정보는 시스템 카탈로그에 저장되며, 시스템 카탈로그는 데이타베이스 그룹 카탈로그, 데이타베이스 카탈로그, 문서 화일 카탈로그, 섹션 카탈로그, 결합 섹션 카탈로그 등으로 구성된다.

KRISTAL-Ⅱ에서 시스템 관리자는 동일한 스키마 구조를 갖는 다중의 데이타베이스를 하나의 데이타베이스 그룹으로 정의할 수 있다. 이러한 기능은 데이타베이스 그룹 카탈로그와 데이타베이스 카탈로그에 의해 지원된다.

정보 검색 분야는 대용량의 데이타를 다룬다. 특히 정보의 양이 기하 급수적으로 증가하는 현실에서 데이타베이스는 더 이상 하나의 볼륨에 저장할 수 없는 상황을 초래하고 있다. 따라서 늘어나는 데이타베이스를 여러 개의 볼륨에 분산 저장하는 기능이 필수적이며, KRISTAL-Ⅱ에서는 문서 화일 카탈로그를 통하여 이를 지원하고 있다.

섹션 카탈로그와 결합 섹션 카탈로그는 데이타베이스의 문서를 구성하는 섹션들의 정보를 유지한다. 섹션 카탈로그는 데이타베이스 그룹의 기본적인 섹션들에 대한 섹션 이름, 색인 방식 등에 대한 정보를 저장하며, 결합 섹션 카탈로그는 여러 개의 기본 섹션들을 결합해 만들어진 가상 결합 섹션들에 대한 정보를 저

장한다.

5. 데이타베이스 관리기

데이타베이스 관리기는 시스템 관리자에 의해 작성된 데이타베이스 스키마 정보를 해석하여 카탈로그 관리기를 통해 이 정보를 시스템 카탈로그에 기록한다. 또한 저장 엔진을 이용하여 유닉스 화일로서 존재하는 원시 문서들을 데이타베이스에 적재한다. 적재 과정에서 데이타베이스에 저장되는 문서들은 색인어 추출 시스템을 통해 관리자가 결정한 색인 방식에 따라 섹션별로 자동적으로 색인된다. 데이타베이스 관리기는 이들 스키마 정보와 적재 정보를 기술하기 위한 별도의 언어를 정의하고 있으며, 이 언어를 실행하기 위한 인터프리터를 사용한다.

6. 검색 엔진

부울 모델에서 문서는 색인어들의 집합으로 표현되고, 질의는 색인어들을 부울 연산자 AND, OR, NOT으로 연결한 부울 수식이며, 검색되는 문서는 질의로서 주어진 부울 수식을 만족하는 문서들이다. KRISTAL-Ⅱ의 검색 엔진은 부울 모델을 지원하며, 카탈로그 관리기와 저장 엔진을 이용하여 사용자의 질의를 만족하는 문서들을 데이타베이스로부터 검색한다.

6.1 사용자 질의

국제 표준화 기구 중의 하나인 National Information Standards Organization(NISO)에서는 1991년에 Z39.58-199x를 발표하였다[6]. 이것은 온라인 정보 검색을 위한 사용자 명령을 표준화한 것으로서, Z39.58-199x의 FIND 명령어는 부울 모델을 근간으로 하고 있다. KRISTAL-Ⅱ에서 사용자 질의의 구문 형식은 Z39.58-199x의 FIND 명령어를 기초로 만들어졌으며, 다음과 같은 다양한 연산을 제공한다.

가. 검색 범위 지정 : 사용자 질의의 검색 범위를 한정시키기 위해 질의의 각 단어마다 검색 섹션을 지정할 수 있다. 만일 아무것도 지시하지 않으면, 시스템은 미리 내정된 섹션에 대해 검색을 수행한다.

나. 절단 연산 : 질의를 구성하는 단어들의 중간 절단과 우측 절단을 지원한다. 이를 위해 사용자는 *, ?의 기호를 사용할 수 있다.

다. 부울 연산 : 질의를 구성하는 단어들 사이의 논리적인 관계를 명세할 수 있도록 AND, OR, NOT의 부울 연산자를 지원한다.

라. 근접도 연산 : 질의를 구성하는 단어들 사이의 문서내에서의 상대적인 거리와 순서를 명시할 수 있도록 NEAR와 WITHIN 연산자를 지원한다.

마. 사용자 질의의 히스토리 기능 : 시스템은 사용자가 입력한 질의들의 히스토리를 유지하며, 이를 접근하기 위한 SET 연산자를 지원한다.

6.2 질의 연산

검색 엔진은 사용자의 질의를 만족하는 문서들을 데이타베이스로부터 검색하기 위해 질의 분석과 질의 연산의 두 단계를 수행한다. 즉, 사용자 질의 분석 과정은 어휘 분석, 구문 분석, 의미 분석의 단계를 거쳐 구문 파싱 트리를 생성하고, 질의 연산 과정은 구문 파싱 트리를 순회하면서 저장 엔진과 커널의 기능을 호출하여 사용자 질의를 만족하는 문서들을 검색한다.

한편, 어휘 분석 단계에서는 사용자의 질의를 구문 분석 단계에서 사용할 수 있도록 일련의 토큰들로 변환하여 전달하며, 구문 분석 단계에서는 사용자의 질의가 BNF로 표현된 문법 구문에 맞는가를 확인하고, BNF 구문에 맞는 구문 파싱 트리를 생성한다. 그리고 의미 분석 단계에서는 카탈로그 관리기를 이용해 질의에 나타난 섹션 이름들의 적절성을 확인한다.

7. 색인어 추출 시스템

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-Ⅱ는 데이타베이스에 대한 섹션별 색인을 지원하며, 색인어 추출 방식에 따라 섹션 단위, 어절 단위, 형태소 단위의 색인 방법을 제공한다.

섹션 단위의 색인은 섹션 값 전체를 하나의 색인어로 선정하는 방식으로, 섹션 값에 대한 완전 일치의 검색을 지원한다.어절 단위의 색인은 각 섹션에서 색인어로서 가치가 없는 불용어를 제외한 모든 어절들을 원문에 나타난 형태 그대로 색인어로써 추출한다.

형태소 단위의 색인은 한글 문장에 대해 형태소 분석을 수행하여 모든 어절들을 명사, 조사, 부사 등의 형태소 단위로 분리한후, 불용어들을 제거하고 색인어로서 의미가 있는 단순 명사들을 색인어로 추출한다. KRISTAL-Ⅱ의 색인어 추출 시스템은 한글 문장의 분석을 위해 연구개발정보센터에서 개발한 형태소 분석기를 이용하고 있다[3].

8. 웹 게이트웨이

KRISTAL-Ⅱ에서는 기존의 단순한 명령어 방식이나 메뉴 방식을 지양하고, 오늘날 우리에게 친숙한 웹 사용자 인터페이스를 제공한다. 이를 위해 웹 서버와 검색 엔진을 연결하기 위한 웹 게이트웨이를 구현하고 있다. 사용자는 웹 브라우저를 통해 원하는 정보에 대한 질의를 입력할 수 있으며, 입력된 질의는 웹 서버를 통해 웹 게이트웨이에 전달된다. 그리고 웹 게이트웨이는 검색 엔진에 접근하여 사용자의 질의를 만족하는 문서들을 검색한 후, 검색 결과를 다시 웹 서버를 통해 사용자의 웹 브라우저로 전달한다.

웹은 일반적인 클라이언트-서버 구조와는 달리 웹 서버가 웹 브라우저의 요구에 응답한 후에 웹 브라우저와의 연결을 즉시 종료한다는 특성을 갖고 있다. 따라서 이러한 상황에서 검색 엔진에 의해 검색된 결과는 자연히 유실되므로 이후의 연산에 사용될 수 없다는 문제점이 발생한다. 웹 게이트웨이는 이를 해결하기 위해 검색 엔진에서 생성된 검색 결과를 외부 화일로써 유지하며, 일정한 시간이 경과된 후에 그 화일을 제거하는 방식을 채택하고 있다.

9. 결 론

정보 검색 시스템 KRISTAL-Ⅱ는 대용량의 데이타베이스로부터 사용자의 질의를 만족하는 문서들의 검색을 지원한다. 이 시스템은 NISO에서 제정 발표한 질의 구문 형식을 구현함으로써 부울 연산, 근접도 연산, 절단 연산 등의 다양한 질의 기능들을 제공하며, 영어 문서뿐만 아니라 한글 문서에 대한 검색을 지원한다.

저장 엔진은 정보 검색의 주된 대상이 되는 가변 길이의 비정형 텍스트 문서들에 대한 저장과 접근 방법을 제공하며, 카탈로그 관리기는 서비스 제공자로 하여금 대용량의 데이타베이스를 디스크 용량의 한계를 넘어 분산 저장할 수 있도록 한다. 아울러 웹 게이트웨이는 사용자로 하여금 데이타베이스를 웹을 통하여 보다 친숙하게 접근할 수 있도록 도와준다.

KRISTAL-Ⅱ는 현재 과학 기술 정보 서비스와 신문 기사 검색 서비스 그리고 웹 문서 검색 서비스 등을 위해 활용되고 있다. 이 시스템은 이외에도 전자 신문, 전자 잡지, 문헌 정보, 법률 정보, 특허 정보, 인물 정보 서비스 등과 같은 정보 서비스에 폭넓게 이용될 수 있을 것으로 기대된다.

참고문헌

[1] 연구개발정보센터, "정보검색을 위한 효율적인 저장시스템 개발", 1996.

[2] 연구개발정보센터, "정보검색을 위한 고성능 검색시스템 개발", 1996.

[3] 연구개발정보센터, "정보서비스를 위한 언어처리기술", 1996.

[4] Carey J. Michael, David J. Dewitt, Joel E. Richardson, and J. Shekita, "Object and File Management in the EXODUS Extensible Database System", Proceedings of the 12th VLDB Conference, pp. 91-100, 1986.

[5] H.T. Chou, David J. Dewitt, Randy H. Katz, and Anthony C. Klug, "Design

and Implementation of the Wisconsin Storage System", Software Practice and Experience, Vol. 15, No. 10, pp.943-962, 1985.

[6] National Information Standards Organization, "Z39.58 Common Command Language for Online Interactive Information Retrieval", 1991.

이 준 호

1987.2 서울대학교 전자계산기공학과 학사
1989.2 한국과학기술원 전산학과 석사
1993.2 한국과학기술원 전산학과 박사
1993.2~1994.1 한국과학기술원 인공지능연구센터연구원
1994.1~현재 연구개발정보센터 선임연구원
1994.3~1995.3 코넬대학교 전산학과 방문연구원
1996.4~1996.6 메사츄세스대학교 전산학과 방문연구원
관심분야 : 정보검색, 정보시스템, 데이타베이스

안 정 수

1993.2 숭실대학교 전자계산학과 학사
1995.8 한국과학기술원 전산학과 석사
1995.9~현재 연구개발정보센터 연구원
관심분야 : 정보검색, 저장시스템, 데이타베이스