KRISTAL-2000 에 기반한 문서 관리 시스템

이병희*, 정창후*, 전성진*, 강무영*, 서정현*
*한국과학기술정보연구원 정보시스템연구실
e-mail: {bhlee, chjeong, sjjhun, kmy, jerry}@kisti.re.kr

A Document Management System Based on KRISTAL-2000

Byeong-Hee Lee*, Chang-Hoo Jeong*, Sung-Jin Jhun*, Moo-Young Kang*, Jeon-Hyun Seo*
*Dept. of Information Systems, Korea Institute of Science and Technology Information

요 약

본 연구에서는 별도의 상용 DBMS 를 사용하지 않고, 국내에서 개발된 정보 검색/관리 엔진인 KRISTAL-2000 을 가지고 문서 관리 시스템을 설계하고 구축한다. KRISTAL-2000 이 국내의 언어 사용 환경을 고려하여 다양한 형태의 색인 기능을 제공하므로 검색어도 자유롭게 사용할 수 있는 장점이 있으며, 데이터 양이 대용량인 상황에서도 고속 검색이 가능함을 확인할 수 있었다. 현재는 데이터 입력을 원문을 보고 제목이나 담당자 등의 정보를 수동으로 하고 있는 실정이나, 향후에는 다양한 원문 파일에서 여러 정보를 자동으로 획득할 수 있도록 하는 문서 필터링 기능과 책이나 보고서 등의 원문을 문서 요약하여 이들을 색인할 수 있는 연구가 필요하다.

1. 서 론

현대 정보화 시대를 사는 많은 사람들은 사람의 관리가 불가능 할 정도로 많은 정보가 쏟아지는 일상 생활 속에서 살고 있다. 이를 위해 수 많은 정보를 체계적으로 저장, 관리, 검색하는 시스템이 필요하게 되는데, 이러한 시스템을 정보 검색 시스템이라고 한다[1].

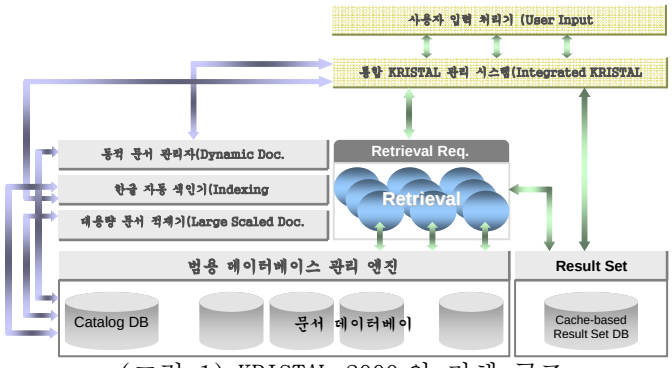
한국과학기술정보연구원(KISTI)은 1993 년부터 외국에서 도입된 정보 검색 시스템의 장점은 최대한 반영하고, 문제점인 한글 자료 처리의 한계, 시스템 기능의 확장이나 변경의 한계, 고가의 도입 비용 등을 해결하기 위하여 KRISTAL(Korea Retrieval Information of Science and Technology Access Line)이라는 정보 검색 관리 시스템을 개발하여 배포해 오고 있다.

2000 년부터 시작된 KRISTAL-2000 은 차세대 심층 전문 정보 검색 및 관리를 위한 시스템으로, 이제까지의 텍스트 정보 검색 관리뿐만 아니라, 구조 문서 및 멀티미디어 정보의 저장, 관리 및 검색을 지원하고 있다.

본 연구에서는 이렇게 개발된 정보 검색 관리 시스템인 KRISTAL-2000을 이용하여 부서 내에서 자주 오가는 문서들을 관리 및 검색하기 위한 문서 관리 시스템을 설계하고 구축한다.

2. KRISTAL-2000의 개요

KRISTAL-2000 은 국내 기술로 개발된 정보 검색 시스템과 DBMS 의 기능이 조화된 검색/관리 엔진으로 고속/대용량의 정보 검색을 지원하며, 필수적인 DBMS 관리 기능이 탑재되어 있고, 여러 API 를 이용하여 사용자의 선호도에 맞는 관리 도구를 구성할 수 있다. KRISTAL-2000 의 전체적인 구조를 보면 다음 그림 1 과 같다.



(그림 1) KRISTAL-2000 의 전체 구조

색인은 시스템의 검색 속도를 높일 뿐만 아니라 검색 효과에도 큰 영향을 미친다. 이를 위해 KRISTAL-2000 에서는 국내의 언어 환경을 고려하여 한글과 한자, 영문에 대해 여러 가지의 색인 방식을

지원하며, 데이터베이스 설계자는 문서의 기본 섹션 및 가상 섹션 마다 이들 색인 방식 중의 하나를 적용할 수 있다.

이러한 KRISTAL-2000을 가지고 현재까지 적용된 정보 검색 및 관리 분야에는 게시판, 웹 포털 서비스, 구조 문서[2], 문헌 정보 서비스[3], 멀티미디어 정보 서비스, 유전자 정보 서비스, 사용자 관리 시스템 등이 있으며, 이 외에도 여러 활용 분야로 전파되고 있다.

3. 문서 관리 시스템을 위한 스키마

KRISTAL-2000 을 이용하여 응용 프로그램을 작성하기 위해서는 먼저 적용할 분야의 스키마를 작성해야 한다. 문서 관리 시스템은 부서 내에서 자주 오가는 공문, 세미나/교육/회의, 업무 보고를 지원하는 응용 프로그램으로 이 스키마 작성은 KRISTAL-2000 매뉴얼[4,5]을 참조하여 작성하였다.

다음 표 1,2,3 은 각각의 필드명, 내용 타입(content type), 레이블, 그리고 색인타입을 보여 준다. 내용 타입 중에서 'FILE_CONTENT' 는 각각의 원문 문서를 저장하는 이진 데이터를 가리키며 이들 테이블을 합쳐 문서 관리 시스템 스키마(dms.schema)를 작성하였다.

(표 1) 공문 스키마

필드명	CONTENT_TYPE	LABEL:설명	색인타입
FORM	TEXT: 선택형	문서종류	INDEX_BY_TOKEN
TITLE	TEXT	제목	INDEX_BY_MA
DATE	TEXT: 선택형	날짜	INDEX_BY_TOKEN
NAME	TEXT	담당자성명	INDEX_BY_TOKEN
AFF_NAME	TEXT	담당자소속	INDEX_BY_MA
AFF_TEL	TEXT	담당자전화	INDEX_BY_TOKEN
AFF_EMAIL	TEXT	담당자이메일	INDEX_BY_TOKEN
DOC_NUM	TEXT	문서번호	INDEX_BY_MA
DOC_TYPE	TEXT: 선택형	공문형태	INDEX_AS_IS
FILE_NAME	TEXT	첨부파일명	INDEX_BY_MA
FILE_CONTENT	BLOB	첨부파일내용	DO_NOT_INDEX

(표 2) 세미나/교육/회의 스키마

필드명	CONTENT_TYPE	LABEL:설명	색인타입
FORM	TEXT: 선택형	문서종류	INDEX_BY_TOKEN
TITLE	TEXT	제목	INDEX_BY_MA
DATE	TEXT	날짜	INDEX_BY_TOKEN
NAME	TEXT	담당자성명	INDEX_BY_TOKEN
LOC	TEXT	개최장소	INDEX_BY_MA
ATTEND_TYPE	TEXT: 선택형	개최 구분	INDEX_AS_IS
ATTENDEE	TEXT	참석자	INDEX_BY_TOKEN
FILE_NAME	TEXT	첨부파일명	INDEX_BY_MA
FILE_CONTENT	BLOB	첨부파일내용	DO_NOT_INDEX

(표 3) 업무 보고 스키마

필드명	CONTENT_TYPE	LABEL:설명	색인타입
FORM	TEXT: 선택형	문서종류	INDEX_BY_TOKEN
TITLE	TEXT	제목	INDEX_BY_MA
DATE	TEXT	날짜	INDEX_BY_TOKEN
REPORT_TYPE	TEXT: 선택형	업무보고구분	INDEX_AS_IS
FILE_NAME	TEXT	첨부파일명	INDEX_BY_MA
FILE_CONTENT	BLOB	첨부파일내용	DO_NOT_INDEX

이렇게 작성된 스키마 테이블을 가지고 스키마 작성 문법에 맞게 스키마를 작성한다. 기본키(PRIMARY KEY)는 제목과 날짜로 하여 중복을 피할 수 있게 선언하였다.

```
/%
 % dms.schema
 %
 %/
// SCHEMA_ENCODING="EUC-KR"
KRISTAL_VERSION="2000"
KRISTAL_DIRECTORY='/raid1/k2demo/kristal'
DATABASE_DIRECTORY='/raid1/k2demo/kvolume/dms'
DATABASE_GROUP_NAME="DMS"

CREATE_SCHEMA
{
        STOPWORD_DEFINITION
        {
                (1) STOPWORD=swords1

        FILE='/raid1/k2demo/kdict/stopword/swords'
        }

        SECTION_DEFINITION
        {
                (1)     LABEL="문서종류"
                        SECTION_NAME=FORM
                        CONTENT_TYPE=TEXT

        INDEX_TYPE="INDEX_BY_TOKEN",
                (2)     LABEL="제목"
                        SECTION_NAME=TITLE
                        CONTENT_TYPE=TEXT
                        INDEX_TYPE="INDEX_BY_MA",
        . . .
                (15)    LABEL="첨부파일내용"
                        SECTION_NAME=FILE_CONTENT
                        CONTENT_TYPE=BLOB
                        INDEX_TYPE="DO_NOT_INDEX"
        }
        PRIMARY_KEY_DEFINITION
        {
                PRIMARY_SECTIONS=(TITLE, DATE)
        }
};
// 데이터베이스 생성
CREATE_DATABASE
{
        // 데이터베이스 이름과 크기를 정의
        (1)     DATABASE_NAME=DMS1
                DATABASE_SIZE=6000
};
DEFINE_DOCUMENT_STRUCTURE
{
  STRUCTURE_DEFINITION = ALL_DOCUMENTS
  {
                // 문서시작 태그 정의
                (1)     TAG_NAME="@dms_tag"
                        TAG_TYPE="SINGLE"
                        // SINGLE or PAIR
```

```
                                ACTION=DISCARD
                                // COPY or DISCARD
                                NEW_DOCUMENT_FLAG=TRUE,
                                // 섹션 태그 및 액션 정의
                        (2)     TAG_NAME="#FORM="
                                ACTION=COPY
                                SECTION_NAME=FORM,
                                . . .
                        (16)TAG_NAME="#FILE_CONTENT="
                                ACTION=COPY
                                SECTION_NAME=FILE_CONTENT
        }
};
// 문서 그룹 정의
DEFINE_DOCUMENT_GROUP
{
        (1)     dms1=('/raid1/k2demo/kdata/dms/list.01')
                        ENCODING=EUC-KR
};
// 문서 적재
LOAD_DATABASE
{
        (1)     FROM=dms1
                TO=DMS1
                WITH=ALL_DOCUMENTS
};
END
```

(그림 2) DMS 스키마 파일의 예

이렇게 작성된 스키마 파일을 가지고 'Loader'를 이용하여 스키마 파일의 문법을 검사한 후, 준비된 데이터 입력 리스트를 가지고 문서 파일과 인덱스로 구성되는 데이터베이스를 만들었다. 데이터 입력은 웹에서 온라인으로 추가할 수 있도록 설계되었다. 다음 그림 3 은 데이터 입력 리스트를 보이고 있다.

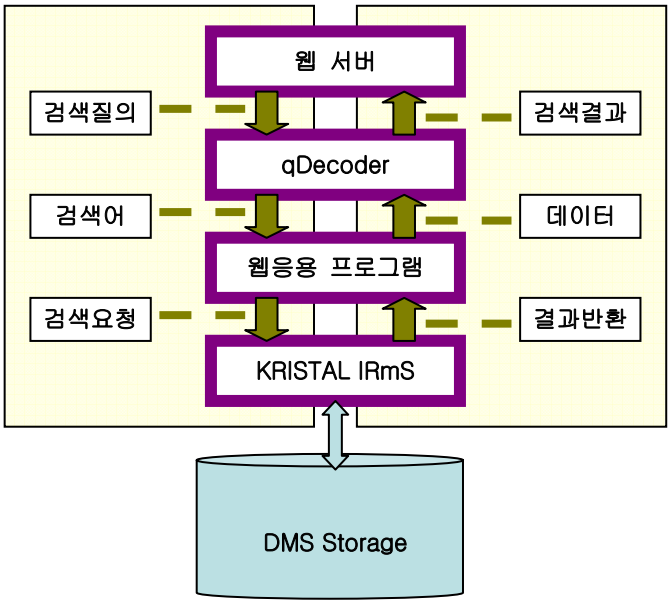
```
@dms_tag
#FORM=공문
#TITLE=수식을 포함한 XML 문서생성기 및 뷰어시스템 개발
        위탁연구계약 용역이행 검사조서
#DATE=2001-12-27
#NAME=이병희
#AFF_NAME=한국과학기술정보연구원 정보시스템연구실
#AFF_TEL=828-5106
#AFF_EMAIL=bhlee@kisti.re.kr
#DOC_NUM=정시 800-487
#DOC_TYPE=원내발송
#FILE_NAME=2002 위탁연구최종검수조서(정회경).hwp
#FILE_CONTENT=/raid1/k2demo/kdata/dms/2002 위탁연구최종검수
                조서(정회경).hwp
. . .
```

(그림 3) DMS 데이터 입력 리스트의 예

4. 문서 관리 시스템의 구현 및 결과

본 시스템은 Linux 환경하에서 C++언어를 이용하여 CGI 형태로 개발되었다. 다음 그림 4 는 웹 서버와 자료 저장소간의 데이터 교환을 보여 준다.

본 시스템은 클라이언트/서버 구조로 실행된다. 이를 위해 먼저 서버를 기동시킨 후, 웹 브라우저에서 클라이언트로 검색을 요청하는 형태로 실행한다. 서버를 기동하기 위해서는 환경 설정 파일(dms.conf)을 인자로 넘겨 주어야 하는데 다음은 본 시스템의 환경 설정이다.



(그림 4) 웹 서버와 자료 저장소의 데이터 교환

```
# K2000 검색 서버의 IP 주소와 포트
kristald 127.0.0.1          8800

# SetManager 서버의 IP 주소와 포트
kristalsmd          127.0.0.1          8801

# DataManager 서버의 IP 주소와 포트
kristaldmd          127.0.0.1          8802

# KRISTAL 이 설치된 디렉토리
KRISTAL_DIRECTORY /raid1/k2000/K2000

# 데이터베이스 위치
DATABASE_DIRECTORY /raid1/k2demo/kvolume/dms

# 데이터베이스 로그의 위치
DATABASE_LOG_DIRECTORY /raid1/k2demo/kconfig/log-dms

# 데이터베이스 그룹 이름
DATABASE_GROUP_NAME DMS

# K2000 검색 최소 서버 개수 (동시 접속자수)
MIN_SEARCH_SERVERS 3

# K2000 검색 최대 서버 개수 (동시 접속자수)
MAX_SEARCH_SERVERS 5

# 최대 검색 결과의 개수 (0 일 경우 무제한)
MAX_RESULT_SIZE                    0

# Shared Memory 의 크기
    KRISTAL_SHARED_MEMORY          0
```

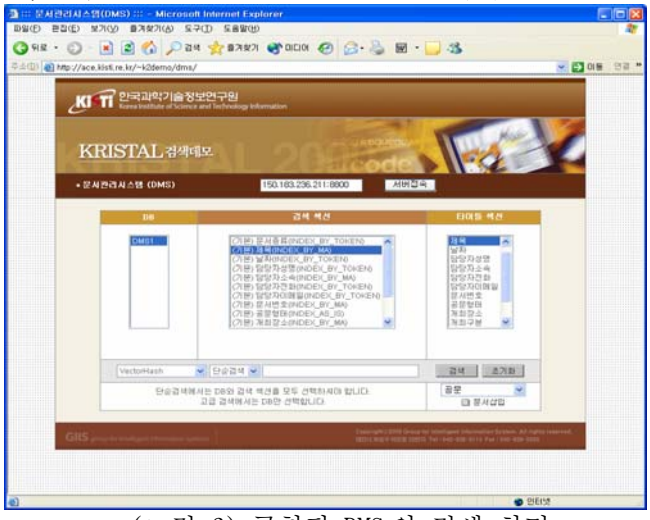
(그림 5) DMS 의 환경 설정 파일

이와 같은 환경 설정하에 서버를 기동하게 되는데 다음은 서버 기동 순서이다.

- 먼저 검색 엔진 데몬을 실행(kristald –f dms.conf)시킨다.
- 검색 실행 결과를 관리하는 결과 셋 데몬을 실행(kristalsmd –f dms.conf)시킨다.

- 온라인으로 문서를 삽입, 수정, 삭제하는 경우에는 데이터 관리 데몬을 실행(kristaldmd –f dms.conf)시킨다.

다음은 본 연구에서 구현된 여러 화면들로 그림 6 은 구현된 DMS 의 검색 화면이며, 그림 7 은 검색어를 입력하여 검색을 하였을 때 결과 출력 화면이며, 그림 8 은 웹에서의 온라인 문서 삽입 화면이다.



(그림 6) 구현된 DMS 의 검색 화면

(그림 7) 구현된 DMS 의 결과 출력 화면

(그림 8) 구현된 DMS 의 문서 삽입 화면

본 연구에서 구현된 문서 관리 시스템은 두 가지 입력 방법을 지원한다. 하나는 데이터를 모아서 한꺼번에(bulk) 입력하는 것이며, 다른 하나는 웹에서 온라인으로 문서를 삽입, 삭제, 수정하는 것으로 트랜잭션 처리를 통하여 안정적 온라인 문서를 관리할 수 있었다.

또한 구현된 문서 관리 시스템은 높은 재현률과 멀티 스레드를 이용하여 분산검색을 할 수 있으며, 결과셋을 관리할 수 있는 장점이 있었다. 좀더 다양한 형태의 문서를 관리하기 위해서는 스키마와 CGI 부분의 약간의 수정으로 실행시킬 수 있으며 상용 DBMS 를 사용하지 않고 정보 검색 및 관리를 할 수 있을 것으로 보인다.

5. 결 론

본 연구에서는 별도의 상용 DBMS 를 사용하지 않고, 국내에서 개발된 정보 검색/관리 엔진인 KRISTAL-2000 을 가지고 문서 관리 시스템을 설계하고 구축해 보았다. KRISTAL-2000 이 국내의 언어 사용 환경을 고려하여 다양한 형태의 색인 기능을 제공하므로 검색어도 자유롭게 사용할 수 있는 장점이 있다. 또한 데이터 양이 대용량인 상황에서도 고속 검색이 가능함을 확인할 수 있었다.

현재의 문서 관리 시스템에서는 데이터 입력을 원문을 보고 제목이나 담당자 등의 정보를 수동으로 하고 있는 실정이나, 향후에는 다양한 원문 파일에서 여러 정보를 자동으로 획득할 수 있도록 하는 문서 필터링 기능과 책이나 보고서 등의 원문을 문서 요약[6]하여 이들을 색인할 수 있는 연구가 필요하다.

참 고 문 헌

[1] 강현규, 박세영, “특집 정보검색,” 한국정보처리학회 정보처리학회지, pp.37-47, 1998.

[2] 진두석, 최윤수, 안성수, “XML기반 고문서 편찬 관리 시스템,” 한국정보처리학회 추계학술발표 논문집, 제9권 제2호, pp.1693-1696, 2002.

[3] 이석형, 양명석, 전유천, 최영수, “국내 과학기술 정보 통합 DB 구축 및 서비스 시스템 개발,” 제7회 한국과학기술정보인프라 워크샵 학술발표 논문집, pp.398-408, 2002.

[4] 한국과학기술정보연구원, 정보검색관리시스템 KRISTAL-2000 Programmer’s Manual for C++ User, 한국과학기술정보연구원 정보시스템연구실, 2002.

[5] 한국과학기술정보연구원, 정보검색관리시스템 KRISTAL-2000사용자 매뉴얼, 한국과학기술정보연구원 정보시스템연구실, 2002.

[6] 장동현, 맹성현, “자동 요약 시스템,” 한국정보과학회 정보과학회지, pp.42-29, 1997.