한글대장경 검색 시스템

이금석, 이용규, 홍영식

컴퓨터・멀티미디어 공학과, 전자불전연구소(EBTI)

한태식

선학과, 전자불전연구소(EBTI)

목 차

요 약

한글대장경의 전산화 및 검색 시스템 개발은 세계에 우리 문화의 우수성을 널리 알리고, 우리 문화유산에 대한 국민적 관심을 높이며 해당 분야에 대한 효과적인 연구를 지원하기 위해 반드시 필요하다. 한글대장경은 고려대장경의 번역본으로 주로 한글로 기록되어 있지만, 원문의 이해를 돕기 위해서 다양한 한자가 사용되고 있다. 이러한 한자들은 본자(本字)와 뜻은 같지만 모양이 틀린 이체자(異體字)및 오자(誤字)나 탈자(脫字)로 간주되는 파자(破字) 등을 포함하고 있으므로 입력이나 저장 뿐 아니라 검색 관점에서

도 여러 가지 문제점을 갖는다. 또한 인터넷이 널리 보급되면서 인터넷을 통한 문헌 검색이 증가하였지만 기존의 문헌 검색 시스템은 용어 검색, 색인 검색 등과 같은 단순한 검색만을 지원하기 때문에 사용자의 다양한 검색 요구를 충족시키기가 어려웠다.

본 연구에서는 한글대장경의 효과적인 입력과 저장을 위해 유니코드(unicode)를 사용하였고, 효율적인 검색을 위해 문서 구조를 XML로 정의하였으며, 결과내 검색과 히스토리 검색 등과 같은 다양한 검색 방법을 지원하는 웹 기반 한글 대장경 검색 시스템을 개발하였다.[1)]

Ⅰ. 서 론

문화유산의 전산화 및 전자도서관화는 세계적인 추세로 일찍이 미국과 중국어권 나라들을 중심으로 문헌 전산화에 대한 연구가 활발히 진행되어 왔으며 연구 목적뿐만 아니라 상업적 측면에서의 여러 산물들이 발표되었다. 또한 문화유산의 가치를 드높인다는 측면에서 다양한 결과물들이 속속 발표되고 있다. 이에 비해 우리나라의 문헌 전산화 및 웹에서의 검색 시스템 제공은 일부 연구소에서만 작업이 진행 중이다. 특히 웹에서의 한글대장경 검색 서비스 제공은 세계에 우리 문화의 우수성을 알리고, 우리 문화유산에 대한 국민적 관심을 높이며, 해당 분야에 대한 효과적인 연구를 지원하는데 꼭 필요하다.

현재 구현되어 사용되고 있는 문헌 검색 시스템은 대부분 해당 문헌을 찾는 서비스만을 제공하고 일부만이 문헌의 내용을 검색하는 서비스를 제공한다. 해당 분야를 연구하는 연구자들에게는 문헌을 찾는 것도 중요하지만 문헌의 원문 내용을 빠르고 정확하

1) 본 연구는 문화관광부와 동국대학교의 지원으로 수행되었음

게 검색할 수 있는 서비스가 필요하다. 문헌 원문 검색 서비스를 제공하는 시스템의 경우도 용어 검색과 색인 검색 등과 같은 단순한 검색만을 제공하기 때문에 사용자의 다양한 검색 요구를 충족시키기 어렵다. 따라서 본 연구에서는 결과내 검색, 히스토리 검색 등과 같은 다양한 검색을 지원하는 웹 기반 한글대장경 검색 시스템을 개발하였다.

본 연구는 2장에서 기존의 문헌 검색 시스템에 대하여 살펴보고, 3장에서는 한글대장경 검색시스템에서 제공하는 검색 방법의 이용에 대하여 설명한다. 4장에서는 다양한 검색 방법을 지원하는 한글대장경 검색 시스템의 구현에 대하여 기술하고, 5장에서는 이 논문의 결론과 향후 연구 과제를 살펴본다.

Ⅱ. 관련 연구

본 장에서는 기존의 고문헌 검색 시스템과 한글 문서 검색 시스템에서 제공하는 검색 방법과 특징을 살펴본다. [표 1]은 각 검색 시스템 별로 지원하는 검색 방법과 특징을 나타낸다.

[표 1] 검색 시스템 별 검색 방법 및 특징

이름	검색 방법 / 특징
	URL
고려대장경 웹 검색	고문헌 원문 검색, 문서내 일치 한자 검색, 색인 검색
	http://211.46.71.249/condsearch/
고려대장경 CD롬 검색	고문헌 원문 검색 지원, 복수개의 경전 선택 가능, 경전내 이동 · 페이지 이동 바로가기 제공

한국역사정보 통합 시스템	고문헌 원문 검색 지원, 검색어 입력결과가 분류별로 표현, 문자입력기 이용한 한자, 일어, 옛한글 등 입력 가능. 결과내 검색 지원
	http://www.koreanhistory.or.kr/index.html
서울대학교 규장각	한글, 한자 입력 가능. 확장연산, AND, OR 연산 지원
	http://147.46.103.67/3_1.htm
성균관대학교 존경각	검색어로 시작하거나 중간에 포함된 형태 선택하여 검색
	http://east.skku.ac.kr/asp/oldbook/oldbk.asp
엠파스	한글 원문 검색, 부정어 지원, 결과내 검색. 유의어 제공. 고급검색 지원
	http://www.empas.com
네이버	한글 원문 검색, 통합검색 및 디렉토리별, 웹문서, 이미지 등 다양한 분류에서의 검색결과
	http://www.naver.com

표 1에서 보는 바와 같이 한자를 포함하는 문헌의 원문 검색 서비스는 일부의 시스템들만이 제공하고, 그 중에서도 결과내 검색은 한글역사정보 통합시스템에서만 지원하고 있다. 엠파스의 경우 결과내 검색 외에 다양한 검색 기능을 제공하지만 한글 문서 검색만 가능하고 한자를 포함하는 문헌의 검색은 불가능하다. 특히 이전에 검색한 결과를 유지하여 검색에 사용하는 히스토리 검색은 대부분의 시스템에서 제공하지 못하고 있다.

Ⅲ. 한글대장경 검색 시스템의 이용

웹을 통하여 한글대장경 검색 시스템을 사용하기 위해서 사용자는 웹 브라우저를 이용하여 EBTI(Electronic Buddhist Text Institute), 전자불전연구소(http://ebti.dongguk.ac.kr)에 접속한다.

그림 1은 EBTI의 초기화면을 나타낸다.

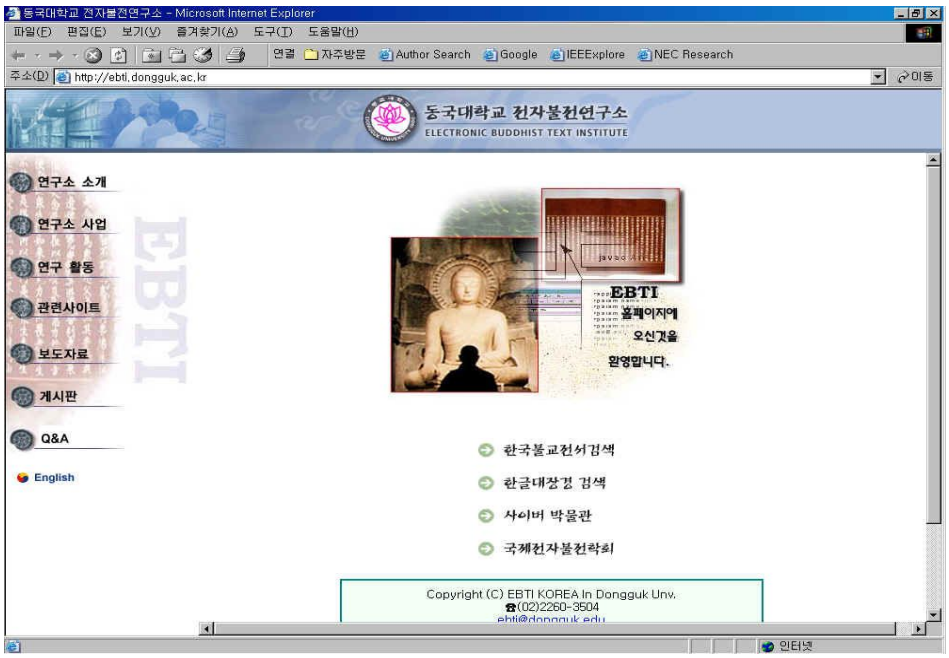

그림 1 EBTI 웹 사이트 초기화면

그림 1에서 '한글대장경 검색'을 선택하면 새로운 창에 한글대장경 검색 시스템이 나타난다. 그림 2는 한글대장경 검색 시스템의 초기화면을 나타낸다.

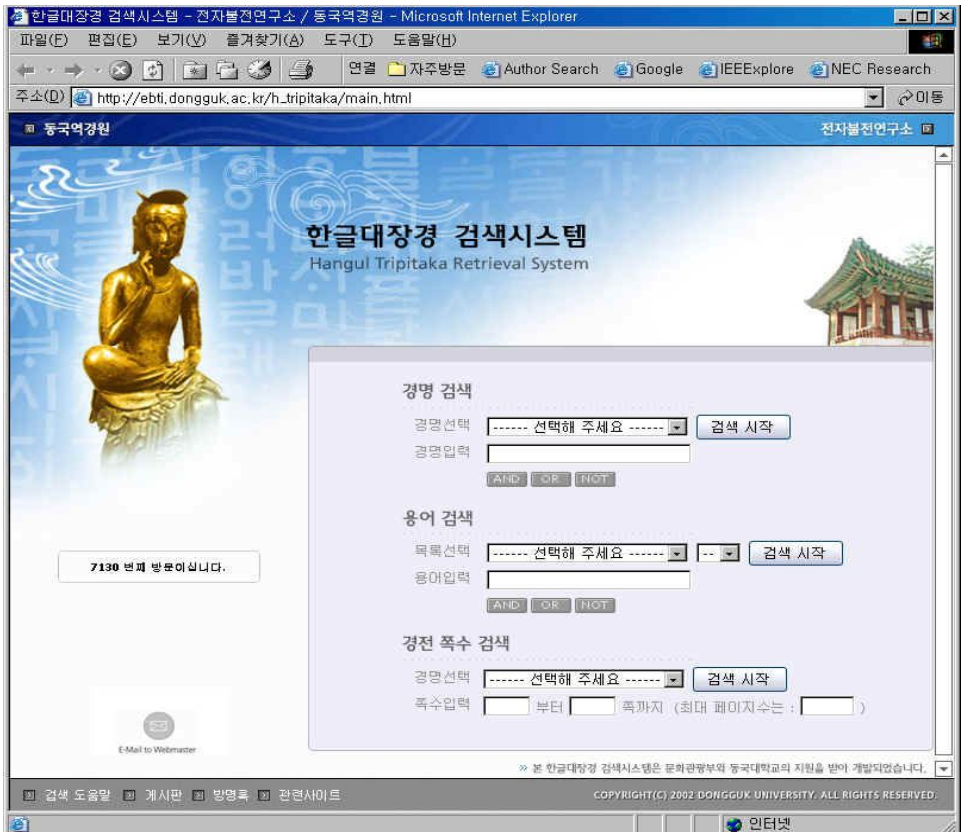

그림 2 한글대장경 웹 검색 인터페이스 초기화면

초기화면은 검색 메뉴와 다양한 링크로 구성된다. 상위 프레임은 전자불전연구소, 동국역경원 링크로 구성되고, 하위 프레임은 검색도움말, 게시판, 방명록, 관련사이트 링크로 구성된다. 중간 프레임은 경명 검색, 용어 검색, 쪽수 검색으로 구성되고, 각 검색 메뉴에서 조건을 선택하거나 입력하고, '검색 시작'을 클릭하면 검색 방법에 맞는 해당 검색 페이지로 이동한다.

한글대장경의 기본 검색 방법은 입력된 경명이나 용어의 한글 독음을 이용하여 검색하는 것이다. 또한 불리언 검색을 제공하여 사용자가 정확한 질의를 할 수 있도록 한다. 그림 3은 한글대장경 검색 시스템에서 제공하는 검색방법의 목록을 나타낸다.

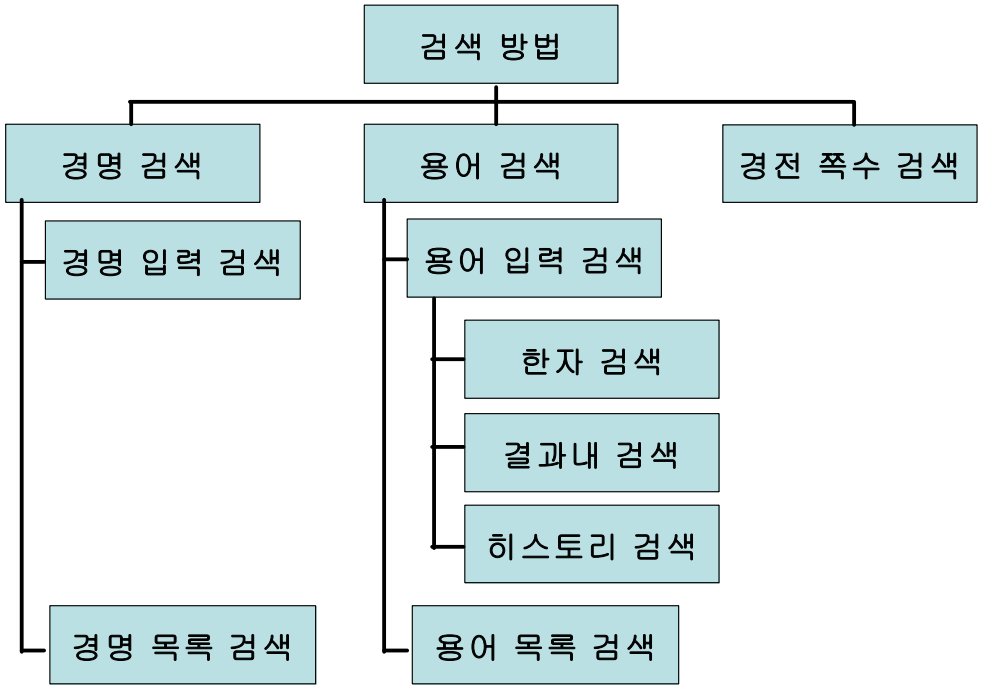

그림 3 제공하는 검색 방법

1. 경명 검색

한글대장경은 많은 경(經)으로 이루어져 있기 때문에 경단위의 검색은 한글대장경에서는 필수적인 기능이라 할 수 있다. 경의 이름을 입력하여 검색하는 경명 입력 검색이나 경명으로 구성된 목록에서 선택하여 검색하는 경명 목록 검색은 기존의 고문헌 검색 시스템에서 제공하고 있다. 경은 많은 제목으로 구성되어 있기 때문에 제목을 색인 목록으로 만들면 신속한 검색이 어렵다. 따라서 본 검색 시스템은 경을 구성하는 제목을 트리형태로 만들었다.

‘경명선택’에서 경명을 선택하면 해당 경의 제목 트리가 나타난다. 나타난 제목 트리의 항목을 선택하면 결과가 오른쪽 화면에 나타난다. 그림 4는 ‘경명선택’을 이용하여 경을 선택하고, 선택한 경의 제목 중 하나를 선택하여 검색한 화면이다.

그림 4 경명선택 후 제목 트리를 이용한 검색

'경명입력'에 경명을 입력하면 입력한 경명을 포함하는 경명을 제공한다. 그림 5는 '화엄'이란 단어를 포함하는 경전 목록을 검색한 화면이다.

그림 5 경명 입력으로 검색한 경전 목록

제공된 경명 중 검색하고자 하는 경명을 선택하면 선택한 경을 구성하는 제목이 트리형태로 나타난다.

2. 용어 검색

용어 검색은 등록된 용어를 목록에서 선택하거나 직접 한글 독음을 입력하여 해당 용어를 포함하는 페이지를 검색하는 방법이다. 한글대장경 검색 시스템은 현재 50,000개의 불교 용어를 이용하여 경전별로 용어의 위치를 색인화함으로써 빠른 검색이 가능하도록 하고, 또한 사용자의 다양한 검색 요구에 맞도록 결과내 검색, 히스토리 검색 및 한자 검색을 지원한다.

2.1 결과내 검색

검색 결과는 검색 범위에 따라 정확하고 신속하게 검색할 수 있다. 이미 검색한 결과와 찾고자하는 결과가 관련이 있는 경우, 검색한 결과 안에서의 재검색은 보다 빠르고 정확한 검색을 가능하게 한다.

‘결과내 검색’ 체크박스를 선택한 후 용어 입력에 결과내 검색을 하고자 하는 문장을 입력하면 이전 결과에 한하여 검색을 실행하게 된다. 사용자가 검색을 진행해 나감에 따라 점차적으로 사용자가 원하는 방향으로 검색의 범위를 좁혀갈 수 있어 빠르고 정확한 검색을 할 수 있다. 그림 6은 ‘대방’이란 용어를 검색한 결과에서 ‘보살’이란 용어를 결과내 검색한 화면이다.

그림 6 결과내 검색

2.2 한자 검색

독음을 이용한 용어 검색은 독음은 같지만 한자가 다른 용어가 존재하기 때문에 정확한 검색을 위해 한자를 이용한 검색 방법이 필요하다. ‘한자 검색’ 체크박스를 선택한 후 용어를 입력하면 한자와 한글독음이 함께 저장된 데이터베이스인 키워드 테이블에 접근하여 독음에 해당하는 한자를 검색한다. 검색 결과의 한자가 하나 이상이면 라디오 버튼을 제공하여 검색하고자 했던 정확한 한자를 선택하여 검색할 수 있도록 한다. 그림 7은 ‘대방’과 ‘사리’라는 한자를 검색한 화면이다.

그림 7 한자 검색 결과 화면

‘대방’이라는 독음의 한자는 1개가 존재하고, ‘사리’라는 독음의 한자가 3개가 존재하기 때문에 3개의 라디오 버튼이 화면에 나타나고, 사용자는 정확한 검색을 위해 해당 한자의 라디오 버튼을 선택한 후 ‘검색시작’을 눌러 검색하면 한글검색 보다 정확한 검색을 할 수 있다.

2.3 히스토리 검색

히스토리 검색은 검색한 결과를 유지하여 이용하는 방법으로 웹 브라우저의 히스토리 기능과 동일하게 동작한다. 즉, ‘이전’ 또는 ‘다음’ 버튼을 이용하여 이미 검색한 결과를 신속하게 제공한다. 이 검색 방법은 단독으로 사용할 수도 있지만 결과내 검색과 연동하여 사용할 수도 있다.

3. 경전 쪽수 검색

한글대장경 검색 시스템은 기본 검색 결과로 10쪽을 사용하고 있지만 사용자가 10쪽 이상의 검색 결과를 원하거나 특정 쪽으로 이동하고자 할 경우 경전 쪽수 검색을 사용할 수 있다.

검색을 위해 사용자는 경을 먼저 선택하는데 사용자는 선택한 경이 몇 쪽으로 구성되어 있는지 알 수 없기 때문에 잘못된 쪽수를 입력할 수 있다. 따라서 본 시스템에서는 사용자가 경을 선택하면 경이 몇 쪽으로 구성되어 있는지를 알려주어 오류를 막는다.

경을 선택한 후 시작 쪽수와 마지막 쪽수를 입력하고 ‘검색시작’을 누르면 입력한 쪽이 경의 범위에 속하는 지를 확인하고, 경명과 쪽수를 이용하여 검색한다. 그림 8은 ‘대방광불화엄경 40권’의 1쪽부터 20쪽까지를 검색한 화면이다.

그림 8 경전 쪽수 검색

한글대장경 검색 시스템은 반복적인 검색에서도 현재 검색한 결과가 한글대장경의 어느 부분에 속하는지 쉽게 알 수 있도록 위치 정보와 부가 정보를 제공한다. 그림 9는 검색 결과의 위치정보와 부가 정보를 나타낸다.

그림 9 검색 결과의 위치 정보와 부가 정보

검색한 결과는 '대방광불화엄경 40권본'의 제1권, 소제목 '1. 이경을 말씀한 곳과 들은 이들'에 포함되고, 고려대장경의 K.1262(36-1), 신수대장경의 T.293(10-661)의 해당 표시 부분과 연관이 있다.

Ⅳ. 한글대장경 검색 시스템 구현

그림 10은 본 연구에서 개발한 웹 기반 한글대장경 검색 시스템의 구성을 나타낸다.

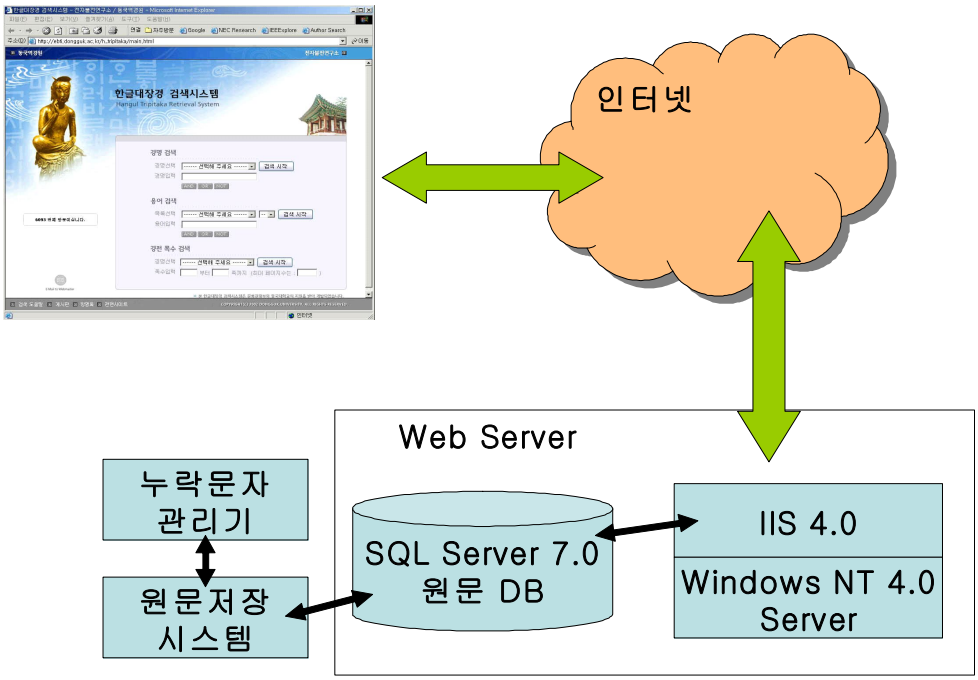

그림 10 시스템 구성도

사용자는 웹 브라우저를 이용하여 웹 서버에 접속하고, 사용자의 요청은 IIS를 통해 ASP로 구현된 검색 인터페이스에 전달된

다. 검색 인터페이스는 구축된 데이터베이스와 연동하여 결과를 사용자에게 반환한다.

한글대장경 검색 시스템의 개발 및 운영 환경은 다음과 같다.

- 운영체제 : Microsoft Windows NT 4.0
- 데이터베이스 : Microsoft SQL Server 7.0
- 웹 서버 : Microsoft Internet Information Server 4.0
- 개발 언어 : ASP, Javascript, Visual Basic 6.0, Java
- 클라이언트 환경 : Internet Explorer 5.0 이상

웹에서 한글대장경 검색 서비스를 제공하기 위해서는 입력 문제 처리, 데이터베이스 구축 등의 작업이 선행되어야 한다.

1. 입력 처리

한글대장경은 주로 한글로 기록되어 있지만 다양한 한자를 병행하여 사용한다. 이러한 한자는 이체자, 파자 등을 포함하고 있고, 이를 제공하기 위해 자체적인 폰트를 제작하거나 임의로 코드를 부여하는 방법이 사용되었지만 새로운 폰트를 설치해야 하기 때문에 인터넷에서 웹 서비스로 고문헌을 제공하는데 문제가 된다[1,2]. 따라서 본 연구에서는 대략 20,902자 정도의 다양한 한자를 지원하는 유니코드(Unicode)[7]를 사용하고, 유니코드 체계에서 지원하는 않는 문자, 즉 누락문자(Missing Character)를 처리하는 누락 문자 관리기(Missing Character Manager)를 사용하여 한자 입력 문제를 처리하였다. 누락문자 처리가 완료된 문서는 XML 태그를 이용하여 문서의 구조를 정의하였다.

현재 입력, 교정 및 태깅 작업을 수행하여 서비스되고 있는 한글 대장경의 목록은 총 30권으로 [표 2]와 같다.

[표 2] 한글대장경 검색 시스템에서 서비스되고 있는 경전 목록

장아함경 1권	출요경 1권
중아함경 3권	법구경 1권
별역잡아함경 1권	60화엄경 3권
대루탄경 1권	80화엄경 3권
불반니원경 1권	40화엄경 1권
아라한구덕경 1권	구사론 2권
비화경 1권	중론 1권
찬집백연경 1권	선문염송 5권
현우경 1권	법원주림 2권

2. 데이터베이스 구축

XML로 마크업된 문서를 저장하기 위해서는 파일 시스템, 관계형 데이터베이스, 전문 데이터베이스, 객체지향 데이터베이스 등이 사용될 수 있으나 본 연구에서는 일반적이고, 높은 안전성과 신뢰성, 그리고 호환성을 제공하는 관계형 데이터베이스인 마이크로소프트사의 SQL Server 7.0을 사용하였고, Java와 JDBC를 이용하여 개발하였다.

한글대장경의 데이터베이스 구축은 용어 및 원문 저장 단계, 인덱스 구축 단계로 이루어진다.

용어 및 원문 저장 단계에서는 용어 사전 파일로부터 용어를 추출하여 용어 테이블에 저장한다. keyword 테이블에는 한글 독음과 한자 유니코드가 저장된다. 원문 저장은 유니코드 편집기에서 작성된 유니코드 원문을 그대로 저장한다. 원문 파일을 라인별로 읽어 저장하면서 페이지 태그를 검사하여 페이지 당 라인수와 ncontinue 등의 부가 정보를 생성한다.

인덱스 구축 단계에서는 keyword 테이블을 사용해서 원문을

순차 검색한다. 용어가 나타나면 페이지와 라인 번호를 keyword_index 테이블에 저장하고, 제목이 나타나면 tag_jmok_table, tag_jmok_table, tag_kyung_table 테이블을 저장한다.

3. 웹 검색 인터페이스 구현

3.1 제목 트리 생성 및 저장

제목을 저장해 놓은 데이터베이스에 접근하여, Javascript와 CSS(Cascading Style Sheet)로 작성된 트리 인터페이스를 생성하고 저장한다. 사용자의 요구에 따라 각 경전별로 생성된 트리 객체로부터 제목 트리를 생성하여 화면에 나타낸다.

3.2 히스토리 검색

히스토리 검색을 하기 위해서는 사용자의 세션(Session)별로 검색 결과를 유지해야 한다. 이를 위해 쿠키(Cookie)를 이용할 수 있지만, 쿠키는 파일 형태로 사용자 컴퓨터에 저장되기 때문에 세션 단위로 처리하기 어렵다. 한 사용자는 다수의 브라우저를 이용하여 검색할 수 있기 때문에 세션 단위로 히스토리 검색을 지원해야 한다.

본 연구에서 개발한 검색 인터페이스는 세션 객체와 Scripting 객체의 Dictionary 객체[8]를 이용하여 히스토리 검색을 구현하였다. 세션 객체는 세션 동안만 유지된 후 제거된다. 세션 시작 후 몇 번째 검색인지를 나타나는 번호인 Session("current")는 사용자가 검색할 때 마다 증가하고, 컬렉션(Collection)의 Key값으로 검색 결과와 함께 Dictionary 객체형의 세션 변수에 저장된다. Session("current")가 0보다 큰 경우 사용자가 '이전' 버튼을 선택하면 Session("current")의 값을 감소시키고 세션 변수의

Dictionary 객체의 값을 이용하여 이전 검색 결과를 생성한다.

3.3 결과내 검색

개발한 검색 인터페이스의 결과내 검색은 MS-SQL 서버에서 제공하는 전역 임시테이블(Global Temptable)[9]을 이용하여 구현하였다. 전역 임시테이블은 세션 동안만 유효한 테이블로 결과를 임시로 저장하는데 적합하다.

3.4 한자 검색

데이터베이스 구축단계에서 한자와 한글 독음으로 구성된 테이블을 데이터베이스에 저장한다. 사용자가 질의한 한글 독음을 데이터베이스에 질의하고, 결과가 다수개인 경우 라디오 버튼을 이용하여 선택할 수 있게 하였다. 한자 검색의 경우 한자 유니코드를 용어 색인 테이블에 질의하여 정확한 검색 결과를 생성한다.

V. 결론 및 향후연구

한글대장경에 대한 연구 활성화와 대중화를 위해서는 한글대장경의 전산화와 검색 시스템 개발이 필수적이다. 그러나 기존의 문헌 검색 시스템으로는 한글대장경을 효율적으로 검색하기 어렵고, 사용자의 다양한 검색 방법 요구를 충족하기 어려웠다.

본 연구에서는 한글대장경의 효과적인 입력과 저장을 위해 유니코드를 사용하였고, 효율적인 검색을 위해 문서 구조를 XML로 정의하였으며, 사용자의 다양한 검색 방법 요구를 충족시키기 위해 기본적인 검색 외에 결과내 검색, 히스토리 검색 등을 지원하는 웹 기반 한글 대장경 검색 시스템을 개발하였다. 또한 인터넷을 사용할 수 없는 환경에서 사용할 수 있도록 웹과 동일하게 동작하는 CD-ROM 시스템으로도 개발하였다.

향후연구 과제로는 검색 결과와 제목 트리를 동기화하는 문제를 해결하기 위해 트리를 동적으로 생성하는 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] 유응구, 김정숙, 이용규, 이금석, 홍영식, “유니코드를 기반으로한 한자 입력 시스템 개발,” 한국정보과학회, ‘98봄 학술발표논문집, Vol 25, No 1, 1998.

[2] 김태규, 유병인, 한인, 이용규, 이금석, 홍영식, “ 유니코드 한자 지원 문법 지시적 SGML 편집기의 설계 및 구현,” 한국정보과학회, ‘98봄 학술발표논문집, Vol 25, No 1, 1998.

[3] 주신탁, 설승진, 이용규, 이금석, 홍영식, “ 유니코드와 SGML을 이용한 한국 고문헌 데이터베이스 구축,” 한국정보과학회, ‘98봄 학술발표논문집, Vol 25, No 1, 1998.

[4] 조은정, 신훈철, 이용규, 이금석, 홍영식, “웹에서의 한국 고문헌 검색 시스템,” 한국정보처리학회, ‘98 봄 학술발표논문 CD, 1998.

[5] 이금석 외, “웹에서의 한국 고문헌 관리 및 검색기술 개발,” 정보통신부, ‘97 초고속 정보통신 응용기술개발사업 최종 연구결과 보고서, 7, 1998.

[6] 이용규, 홍영식, 이금석, 김정숙, 한인, 설승진, 신훈철, “한국 고문헌 데이터베이스,” 동국대학교, 동국논총 제 三十七, 12월, 1998.

[7] Unicode Homepage, http://www.unicode.org

[8] C.Ullman외 6인, “Beginning Active Server Pages 3.0,” WROX, pp359-410, pp461-513, 2000.

[9] Microsoft SQL Server2000 Online Help