한국불교전서 웹 검색 시스템 개발

김영희*, 장은실*, 구현우*, 박영희**,
이용규***, 이금석***, 홍영식***, 한보광****

목 차

요 약

현재까지 한국불교전적 및 한국 고문헌에 대한 전산화는 매우 미미한 실정이며, 웹에서의 한국불교전적을 포함한 한국 고문헌에 대한 검색 시스템은 전례를 찾아보기 힘들다. 그러나 문화유산의 전산화 및 전자도서관화는 세계적인 추세이며, 고문헌의 전산화 및 웹에서의 검색 서비스 제공은 해당 분야의 연구자들에게는 시급히 필요한 과제이다. 또한 본교는 불교학을 중심으로 한 한국학과 컴퓨터 정보통신 두 분야를 특성화의 큰 축으로 하고 있으며, 불교자료의 전산화야말로 본교의 특성화 방향

*동국대학교 컴퓨터공학과,
**동국대학교 선학과, 전자불전연구소 전임연구원,
***동국대학교 컴퓨터·멀티미디어공학과 교수,
****동국대학교 선학과 교수, 전자불전연구소 소장

인 “불교학과 정보통신 기술”의 연계에 가장 적합한 프로그램이라 할 수 있다.

따라서 본 연구의 목적은 우리의 문화유산인 한국불교전서 중의 11책과 12책을 컴퓨터에 입력하여 데이터베이스에 저장하고, 이를 효율적으로 검색할 수 있는 검색엔진을 개발하여, 현재 전 세계적으로 범용화 되고 있는 인터넷의 웹을 통해 검색할 수 있도록 하는 것이다.

이렇게 본 연구를 수행함으로써 고문헌의 전산화를 활성화할 수 있으며, 전자도서관을 이용한 문화유산의 관리를 촉진할 수 있을 뿐만 아니라, 우리나라 문화유산에 대한 종합적인 전산화 기술을 개발할 수 있다.

1. 서 론

본 연구는 한국불교전서 전산화 6차 사업으로 한국불교전서 제 11책과 12책을 전산화하여 전 세계에서 활발하게 사용되고 있는 인터넷을 통하여 검색할 수 있도록 하는 것이다. 한국불교전적은 우리나라에 불교가 전래된 이래 삼국시대부터 우리나라의 선조들이 남긴 옛 문헌들을 발굴 수집하여 출간한 한국불교전서와 해인사 고려대장경을 동국대학교의 역경원에서 한글로 번역한 한글 대장경을 위시한 한국 불교 문헌을 총칭한다.

그 중 한국불교전서는 동국대학교가 30여 년 동안의 오랜 시간과 막대한 예산을 투입하여 한국에 불교가 전래된 이래 삼국시대부터 구한말에 이르기까지의 불교계의 고승대덕, 명현학자 등 우리의 선조들이 남긴 옛 문헌들을 낱낱이 발굴 수집하여 전 13책으로 출간한 한국 고전 학술자료의 대총서이다. 이 전서는 신라의 원측이 저술한 ‘반야심경찬’으로부터 구한말의 서진하가 쓴 ‘선문재정록’에 이르기까지 석학 고승 1백 71인이 남긴 2백 88종의 옛 문헌을 그대로 활자화하여 수록한 것인데 대교본을 포함하면 552부 1506권 21편에 이른다. 또한 한국불교전서는 고려의 대각국사 속장경 이후 한국불교의 모든 전적을 집대성 한 것으로 우리나라 불교사상의 흐름을 일목요연하게 파악할 수 있을 뿐만 아니라, 정신문화·역사·철학 등 여러 분야에 걸쳐 효율적인 연구 자료이다. 이의 발간은 우리나라의 역사를 주도해 온

불교사의 심층을 조명하고, 불교 사상과 아울러 한국의 전통사상을 정리한 것이라 할 수 있다[32].

그리고 우리의 귀중한 문화유산인 고려대장경의 한글화 작업이 동국대학교 역경원에서 진행 중이며, 총목록, 목록색인 및 해제, 내용 색인 그리고 저역자 색인 등과 번역된 한글 대장경을 2000년 6월까지 총 316여권을 발간하였다[32].

이러한 불교경전에 대한 활발한 편찬 사업에 비하여 불교전적의 전산화 작업은 현재 미비한 상태이나, 세계적으로는 이 분야에 대한 연구가 활발히 진행 중에 있으며 이러한 디지털 경전에 대한 정보교환을 위한 국제 회의도 개최되고 있다[34]. 가상공간에서 불법을 펴고, 법신불 비로자나 부처님을 모시는 방법을 논의하는 국제회의인 전자불전회의(Electronic Buddhist Text Initiative, EBTI)는 '전자 경전을 만드는 것을 발의(發意)한다'는 회의다. 불교정보화와 관련 유일한 국제회의인 EBTI는 인터넷에서 제공되고 있는 다양한 불교정보를 서로 자유롭게 이용할 수 있는 호환성을 키우는데 그 중요성이 있다. 이 회의를 통해 각국에서 진행하는 불교정보화에 대해 현재까지 진행된 상황 등 여러 가지 의견을 교환한다.

2000년 12월 16일부터 22일까지 미국의 버클리대학에서 20여 개국의 70여명의 학자가 참석한 가운데 개최된 EBTI와 2001년 5월 25일부터 26일까지 본교에서 10여 개국의 70여명의 학자가 참석한 가운데 개최된 EBTI에서 한문・빨리어・산스크리트어・티벳어 경전에 관한 전산화, 사전, 문헌정보 등이 논의되었고, 불교문화를 데이터베이스로 저장하는 다양한 프로젝트들도 소개되었다. 뿐만 아니라 전자사전 개발 및 원전 보전 계획도 활발한 것으로 밝혀졌다. 그러나 한문 경전 전산화에서 우리나라에서만 고려대장경이 연구될 뿐, 일본・대만・중국・미국에서는 대정신수대장경에 대한 전산화만을 활발히 추구하고 있는 것으로 나타났다. 또한 EBTI에서는 불경 전산화를 위해 필요한 여러 가지 기술과 표준화도 논의되고 있다. 논의되고 있는 기술 중에서 가장 큰 쟁점이 되고 있는 것은 '컴퓨터에서 읽혀지지 않는 한자' 즉 '누락(missing)' 한자를 구현하는 방법이다.

이렇게 EBTI에서 논의되고 있는 기술들은 본 연구를 수행하는데도 필요

한 기술로서 첫째로 한자를 컴퓨터에 입력할 수 있는 입력방법 및 입력된 한자들을 편집할 수 있는 편집 시스템의 개발이 필요하다.

2. 한국불교전서 전산화 6차 사업

본 연구에서는 한국불교전적을 전산화하여 인터넷을 통해 손쉽게 검색할 수 있도록 하는데 그 목적이 있다. 이러한 연구 목적을 달성하기 위해 크게 3가지 기술이 필요하다. 한국불교전적을 컴퓨터에 입력하고 이를 편집하여, 데이터베이스에 저장하고, 데이터베이스에 저장된 내용들을 웹에서 검색할 수 있는 인터페이스와 검색 기술이 필요하다. 이들을 위해 본 연구에서 개발한 기술 내용 및 사용방법에 관해 먼저 1절에서는 한국불교전서의 입력 및 교정 방법에 대해 기술하고, 2절에서는 유니코드에 없는 글자를 처리하는 방법, 3절에서는 데이터베이스에 저장하는 방법을 살펴보고, 4절에서는 웹 검색 인터페이스 기술 및 검색방법에 대해 설명한다.

2.1. 한국불교전서의 입력 및 교정

이번 전자불전연구소에서 수행한 한국불교전서 전산화 6차 사업에서는 한국불교전서 제11책(보유편 1)과 제12책(보유편 2)의 전체를 입력하고 교정하였다. 아울러 지난 3차 사업에 시행한 제5책의 오류가 많아서 이번 사업에서 교정을 완료하였다.

지금까지 한국불교전서 총 14권 가운데 1책에서부터 12책 까지를 입력하고 교정, 태그를 달아 전산화함으로써, 한국불교전서의 신라시대편과 고려시대편, 조선시대편과 보유편 2권의 전산화를 완성하였다. 이는 한국불교전서 전체 분량의 약 90%에 해당하는 분량이다.

또한 이번 6차 사업인 한국불교전서 제11책 (보유편 1)과 제12책 (보유편 2)의 특징은 도표, 이미지, 영인본, 옛한글과 진언이 등장하였다.

이번 사업에서 입력·교정한 저술은 제11책 총 26종과 제12책 총 34종 모두 60종이다. 입력한 저술들의 목록은 다음과 같다.

제 11 책 (보유편 1)

新羅時代篇

高麗時代篇

제 12 책 (보유편 2)

起信本末五重(一卷)
禪敎摠判門(一卷)
參禪念佛文(一篇)
東國僧尼錄(一卷)

한국불교전서 11책(871쪽)과 12책(875쪽)의 분량은 전체 1,746쪽이며, 글자수로는 약 196만자, 재교정한 5책(924쪽)은 99만 8천자이며, 총 295만 8천여자를 교정하였다. 입력팀의 3명이 분량을 나누어 5차례에 걸쳐 교정작업을 하였으며, 교정과정은 다음과 같다.

1) 제1차 교정: 처음 입력된 것을 틀린 자가 없는지 한 글자 한 글자 원본과 대조하면서 확인하고, 누락문자로 표시된 '&' 기호가 정확한 누락문자인지 이체자인지를 옥편과 대조해보며, 또한 옥편에 있더라도 반드시 MS WORD 문서에서 인식되는 한문인지를 확인해 본 후 최종 누락문자로 처리하고, 한국불교전서 원문을 대조해가며 세심하게 문장에 정확도를 기하였다.
2) 제2차 교정: 1차 교정본에 제목, 소제목, 쪽수, 단락수, 주석, 들여쓰기 등의 태그 처리를 하면서 다시 한 번 교정함.
3) 제3차 교정: 태그가 다 끝난 후 다시 한 번 전체적으로 오류가 없는지 빠진 누락문자와 태그는 잘 처리되었는지에 대해 교정함.
4) 제4차 교정: 태그 처리가 끝나 누락문자를 담당하는 제2팀에서 누락문자 처리가 끝나면 입력팀에서 태그와 문단구성, 누락문자, 이체자 등 전체적으로 오류는 없는지 다시 한 번 최후 교정 작업함.
5) 제5차 교정: 작업이 모두 끝난 원문이 웹상에서 제대로 구현되는지를 다시 한 번 살펴본 후 최종 교정 작업을 마친다.

이러한 입력과 5번의 교정 등 여섯 단계를 거쳐 한국불교전서 제11책, 제12책의 전산화가 이루어졌다. 교정의 원칙은 아래와 같다.

1) 최대한 원문과 동일하게 한다.
2) 원문에 충실하되 古字는 異體字로 대체하고 누락문자는 이미지화 한다.
3) 교감은 하지 않는다.

위의 교정원칙에 근거하여 古字는 다음과 같은 異體字로 교정하였다.

[표 1] 고자의 이체자 교정의 예

원문	교정한자	원문	교정한자	원문	교정한자
纏	纏	餙	飾	總	總
惣	惣	龕, 龍	龕	纸	紙
纪	紀	悗	悅	虛	虛
顔	顏	函	函	髮	髮
兔	兎	乌	烏	财	財
骚	騷	攒	攢	畧	略
烟	煙	鬬	鬪	覓, 覎	覓

2.1.1 태깅

1차 교정 작업이 끝나고 나면 태깅을 시작한다. 태깅 작업은 문서를 웹상에 띄우기 위한 작업으로 매우 정확해야 하고 중요하다.

우선 태깅 작업은 다음과 같다.

제목 태깅 - 원제목을 나타내준다.

<JMOK1>高麗國新雕大藏校正別錄</JMOK1>

소제목 태깅 - 원제목에 딸린 소제목을 나타낸다.

<TAB2><JMOK2>高麗國新雕大藏校正別錄卷第一</JMOK2></TAB2>

쪽수 태깅 - 쪽수를 알려준다.

<PAGE PAGENUM='11'-60></PAGE>

단락을 표시해 주는 태깅 - 1, 2, 3단을 나타내준다.

<DAN DANNUM='1'></DAN>
<DAN DANNUM='2'></DAN>
<DAN DANNUM='3'></DAN>

이미지 태깅 - 이미지로 처리해야 할 부분이다.

<IMAGE 10-111-1-1></IMAGE>

주석 태깅 - 주석임을 나타내준다.

<COMMENT>
{1}此文(漢文及諺解)無有{甲}. {2}[誠]當作[識]{編}.
</COMMENT>

탭 태깅 - 들여쓰기를 책과 같이 해준다.

1칸 들여쓰기: <TAB1></TAB1>
2칸 들여쓰기: <TAB2></TAB2>
3칸 들여쓰기: <TAB3></TAB3>

2.1.2 이미지

제 6차 사업인 제11책과 제12책에 나오는 도표나 그림은 이미지 처리를 위하여 스캔 작업을 하였다. 이미지가 제11책에는 85개, 제12책에는 11개가 등장하였다.

이미지는 본문글자와 구별을 하기 위하여 색상을 진한 청색으로 처리하였다.

2.2. 유니코드에서 누락된 문자의 처리

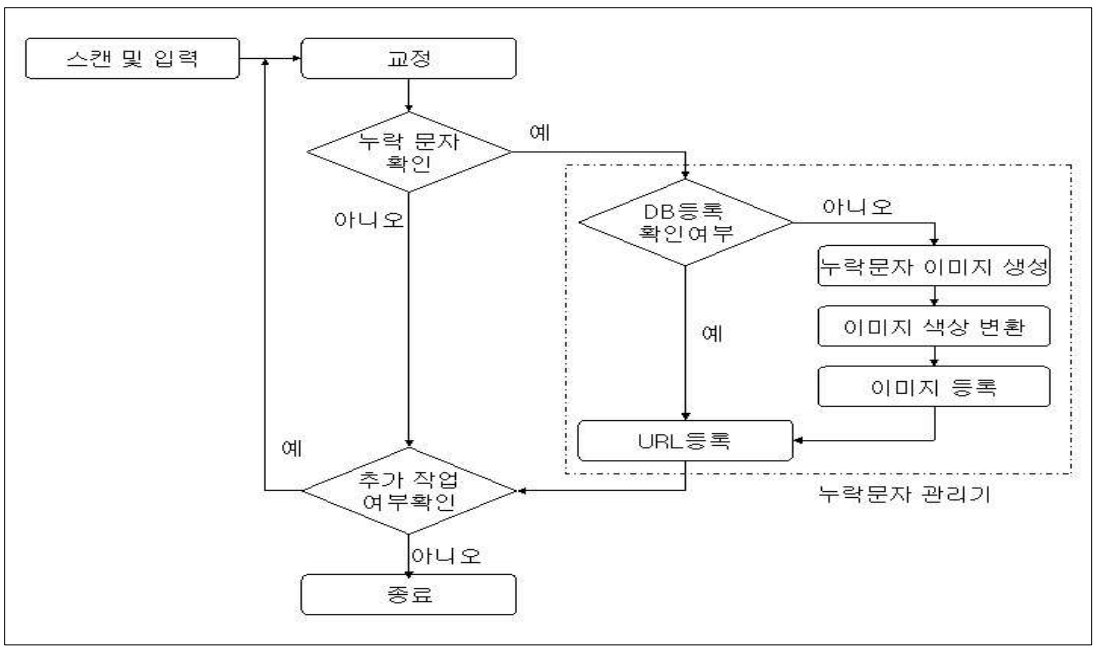
누락문자란 현재 윈도우즈 운영체제 및 인터넷 환경에서 사용가능한 한자에 포함되지 않는 문자를 뜻한다. 한글 윈도우즈에서 채택하고 있는 KSC-5601 한글 체계상에서 한자는 대략 4,888자 정도 지원이 되고 있으며

유니코드를 사용할 경우에는 대략 20,902자 정도의 한자가 지원되고 있다. 그러나 한자로 집필된 불교 고문헌의 경우 KSC-5601 한글 체계나 유니코드 체계에서 지원하지 않는 문자들이 존재하고 있으며 이를 누락문자(Missing Character)라 칭한다.

입력 과정을 거쳐야 하는 한국불교전서의 분량이 방대하기 때문에 가능한 입력 도구의 간소화와 편리화를 필요로 하다. 실제 입력 작업은 스캐너와 OCR을 이용한 방식을 사용하지만 누락문자는 수작업을 통해 문서상에 정해진 태그의 형태로 삽입되어야 한다. 따라서 누락문자 자체를 입력하는 과정이 대단히 번거롭고 시간을 많이 소요하게 되는 작업이 된다.

유니코드에 나와 있지 않은 문자인 누락문자를 입력하기 위해서는 누락문자를 GIF 형식의 폰트 이미지 파일로 만들고 누락문자 DB에 등록한다. 그리고 한국 불교 전서를 인터넷을 통해 검색 시 다른 유니코드 문자들과 함께 등록된 누락문자를 웹 브라우저 상에서 보여 준다.

다음은 누락문자를 입력하는 작업을 나타내고 있다.



[그림 1] 한국 불교 전서 입력과정

[그림 1]을 살펴보면 누락문자 이미지를 생성하기 위한 준비 작업으로 원

문 스캔 및 입력 그리고 교정작업이 있다. 원문 스캔 및 입력 작업은 한국불교 전서 원문을 스캐닝하여 이미지 파일(BMP 파일)로 저장하고 이미지 파일을 글눈 2001 프로그램을 사용하여 한자들을 인식하여 텍스트 파일로 변환한다. 이 때 인식되지 않은 한자들은 텍스트 파일에 직접 입력하고 누락문자가 있다면 특별 기호로 표시하게 된다. 이 후 텍스트 파일을 원문과 비교하여 잘못 입력된 내용이 없는지 검토하고 잘못 입력된 내용이 있다면 교정한다. 해당하는 누락문자의 위치(책, 페이지, 단락, 라인)를 문서화한다.

이렇게 준비된 문서들을 이용하여 교정 작업 도중 누락문자가 발견되면 누락문자 검색 프로그램을 사용하여 이미 발견된 누락문자인지 검색한다. 만약 이미 발견된 누락문자라면 검색 프로그램을 이용하여 누락문자에 해당하는 Tag를 삽입한다. 여기서 Tag는 누락문자 이미지가 저장되어 있는 주소를 나타내고 URL 주소로 표현된다. 그러나 저장되어있는 누락문자 중에서 해당 누락문자를 찾지 못하면 누락문자를 이미지 파일로 만들고 누락문자 검색 프로그램에 등록한 후 이에 해당하는 Tag를 삽입한다.

이러한 누락문자 입력 과정을 단순화하고 실제 누락문자를 간단한 방법으로 문서상에 이미지 태그 형태로 삽입할 수 있는 누락문자 관리를 위한 인터페이스를 지난 과제에서 개발하였고, 본 과제에서는 누락문자 관리를 위한 기존에 사용하던 데이터베이스의 수정과 이미 제작된 누락문자 이미지를 데이터베이스화 하는 작업을 진행하였다. 이미지를 데이터베이스화 함으로써 누락문자 이미지의 관리를 용이하게 하고 시스템의 공간 절약을 꾀할 수 있다.

지난 과제 수행에 있어 필요한 누락문자 관리기의 요구 조건은 다음과 같았다.

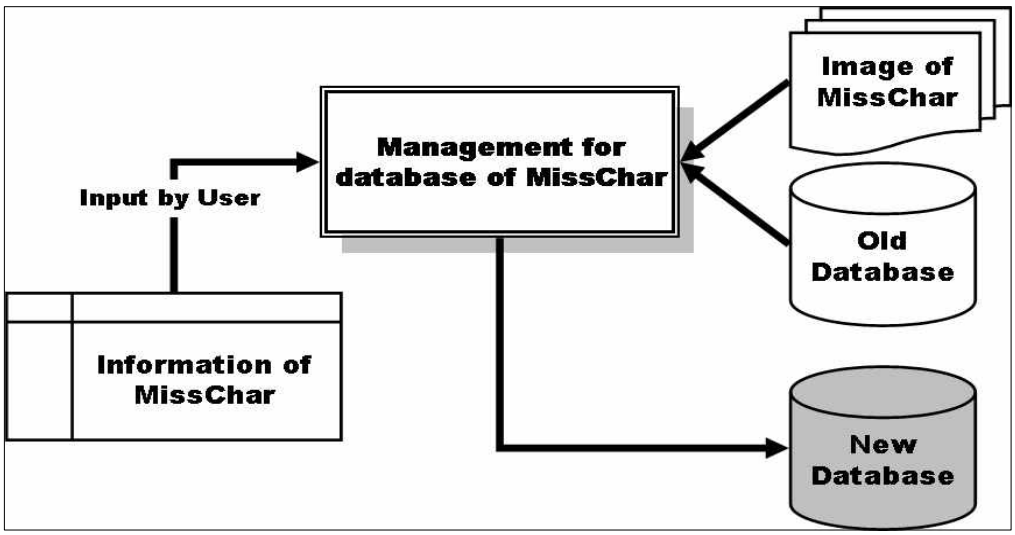
- 편리한 누락문자의 등록
- 등록된 누락문자에 대한 빠른 문서상의 입력
- 웹 문서에서 누락문자사용의 무 제약성 제공

본 과제에서는 추가적인 요구사항을 누락문자 관리기에 반영하였다. 추가적인 요구사항은 다음과 같다.

- 개선된 누락문자의 등록 기능 제공
- 등록된 누락문자의 다양한 검색 기능 제공

한국 불교 전서의 입력 작업은 매우 많은 원문의 분량으로 인한 많은 시간이 소요 된다. 그러므로 입력과정에서 발견되는 누락문자를 손쉽고 빠르게 등록시킬 수 있는 기능은 필수적이라 할 수 있다. 누락문자 관리를 위해서는 여러 한자 이미지를 체계적인 정렬 방법을 통해 제시해야 하며, 사용자는 이러한 한자 중에 자신이 원하는 누락문자의 이미지를 효과적인 방법으로 검색할 수 있어야 한다. 즉, 등록된 문자에 대한 효과적인 검색을 통하여 간편하게 원문 상에 등록된 문자에 대응하는 태그를 입력할 수 있어야 할 것이다. 이를 위해 문자 관리기는 현재 저장하고 있는 등록된 문자의 목록을 효과적으로 게시하며 게시된 문자를 여러 조작 없이, 예를 들어 1회의 마우스 클릭 등의 작업을 통해 원문 상 지정된 위치에 태그 정보를 삽입할 수 있다.

효율적인 누락문자 관리 프로그램의 사용을 위해 기존의 데이터베이스 확장과 이미지 파일로 저장되어 있는 누락문자 이미지의 데이터베이스화를 위한 프로그램이 필요하다. 이를 “누락문자 데이터베이스 관리기”라 부르기로 한다.



[그림 2] 누락문자 데이터베이스 관리기 구조도

[그림 2]는 누락문자 데이터베이스 관리기의 구조도를 보여준다. 누락문자 데이터베이스 관리기에서 기존 데이터베이스에서 총획수별 누락문자 데이터를 임시 테이블로 작성하고 한 레코드씩 로드하면서 해당하는 이미지를 불러온다. 디스 플레이된 누락문자를 관리자가 확인한 후 입력 팀에서 전달받은 누락문자에 대한 정보를 추가 입력하고 저장함으로써 새로운 데이터베이스를 완성한다.

[그림 3] 누락문자 데이터베이스 관리기 초기 인터페이스

[그림 3]은 누락문자 데이터베이스 관리기의 초기 인터페이스를 보여준다. 기존 데이터베이스에 포함되지 않은 정보를 추가하기 위한 여러 개의 구성요소로 되어 있다. 정보 입력을 위해서는 부수 선택을 위한 콤보 박스, 부수를 보여주기 위한 레이블, 그리고 음을 입력하기 위한 텍스트 박스로 구성된다. 기존의 데이터베이스와 이미지를 불러 오기 위한 이미지박스와 텍스트박스 그리고 버튼으로 구성되고 새로운 데이터베이스에 저장을 위한 버튼으로 구성되어져 있다. 마지막 구성요소는 유니코드 확장에 따라 누락문자

중 유니코드 표현이 가능한 문자를 찾아내기 위한 이후에 추가 구현될 부분으로 구성된다.

[그림 4] 데이터베이스 관리기 동작 화면

[그림 4]는 누락문자 데이터베이스 관리기 동작 과정을 보여준다. 먼저 누락문자의 총획수를 입력하고 “OLD DB Load" 버튼을 클릭하면 총획수에 대하여 기존 데이터베이스에서 레코드를 로드하고 첫 번째 레코드에 해당하는 이미지를 불러와 이미지 박스에 보여준다. 불린 이미지를 보고 해당 부수를 선택하고 음이 존재하면 음을 입력한 후 "SAVE" 버튼을 클릭하여 누락문자를 데이터베이스에 저장한다. 다음은 ”SAVE"버튼을 클릭하였을 때의 데이터베이스에 데이터를 저장하는 코드를 나타낸다.

```
Private Sub Save_Click()
    Dim adoControl As New ADODB.Connection
    Dim strCon, url, strProgPath As String
    Dim SQL As String
    Dim r As Long
    Dim arraybyte() As Byte
```

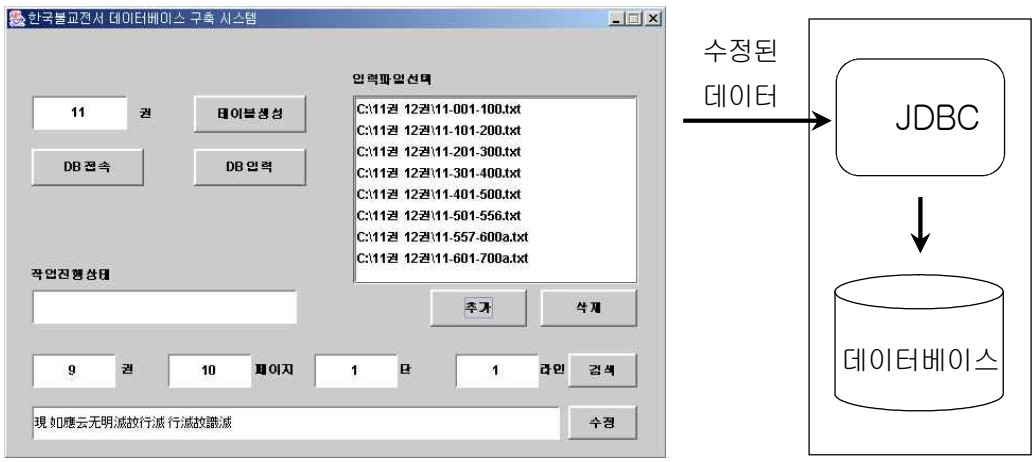
```
    '문자 저장
    SavePicture ImgChar.Picture, App.Path & "\temp.bmp"

    '누락문자를 임시화일에 저장하고 arraybyte 배열로 binary로 읽어들인다.
    ReDim arraybyte(FileLen(App.Path & "\temp.bmp"))

    Open App.Path & "\temp.bmp" For Binary As #2
    Get #2, , arraybyte
    Close #2

    '데이터베이스에 저장할 URL 정보 생성
     url = "<IMG SRC=""http://ebti.dongguk.ac.kr/images/k"
                          + Trim(labCharID.Caption) + ".gif"">"

     '데이터베이스의 테이블을 열고 해당 정보들 입력
     Adodc1.Recordset.Open "DBmisschar",
     "Provider=SQLOLEDB.1;Password=3338;Persist Security Info=True;
     User ID=misschar;Initial Catalog=misschar;
     Data Source=ALGO", adOpenDynamic, adLockOptimistic, adCmdTable

     Adodc1.Recordset.AddNew
     Adodc1.Recordset.Fields!charID = Trim(labCharID) '누락문자 ID 저장
     Adodc1.Recordset.Fields!busuID = Trim(labBusuID) '부수 ID 저장
     Adodc1.Recordset.Fields!busuNum = Trim(combBusu1.Text) '부수의 획수 저
      장
     Adodc1.Recordset.Fields!charNum = Trim(txtStrokeNum.Text) '누락문자 총
      획수 저장
     Adodc1.Recordset.Fields!Pron = Trim(txtPronReg.Text)  '누락문자의 음 저
     장
     Adodc1.Recordset.Fields!url = url       '누락문자 Tag URL 저장
     Adodc1.Recordset.Fields!Image = arraybyte  '누락문자 Image 저장
     Adodc1.Recordset.Fields!Size = Trim(CStr(FileLen(CStr(App.Path +
                                    "\temp.bmp"))))   '누락문자 Image 크기 저
     장
     Adodc1.Recordset.Update
     Adodc1.Recordset.Close
End Sub
```

2.3. 데이터베이스 저장

한국불교전서는 제목, 원문, 주석 등으로 구성되어 있고, 각각에 해당되는 내용은 태그로 구별하기 때문에 이를 이용하여 데이터베이스를 구축할 수가 있다. 이때 원문에 나타나는 한자는 기존 문자 셋으로 표현하는데 한계가 있어서 유니코드로 변환하여 저장한다.

2.3.1 데이터베이스 원문 저장 시스템 구현

[그림 5]는 데이터베이스 저장 시스템의 구조도를 나타낸 그림이다. 즉, 원문 저장 프로그램을 이용하여 원문 데이터를 입력하면 이를 데이터베이스에 저장한다. 이때, 프로그램과 데이터베이스간의 연동을 위하여 JDBC를 사용하고, DBMS로는 마이크로소프트사의 SQL 7.0을 사용하였다.



[그림 5] 데이터베이스 원문 저장 시스템 구조도

데이터베이스 원문 저장 프로그램의 주요 기능은 원문 저장, 원문 검색, 원문 수정 기능이며, 원문 저장 기능은 원문을 저장하는 기능으로 구성되어 있다. 그리고 원문 검색 기능은 원문 저장 프로그램의 입력된 데이터에 대한 검색 기능이며, 원문에서 검색하고자 하는 내용을 찾을 수 있을 뿐만 아

니라 데이터의 오류 부분을 쉽게 찾아낼 수 있다.

마지막으로 원문 수정 기능은 원문 검색 기능을 이용하여 오류가 있는 원문 부분을 찾아낸 후에 수정하는 기능으로 원문의 양이 많아져서 일부 오류가 있는 원문 전체를 전부 입력할 때 발생하는 시간과 노력을 줄이는 효과를 보인다.

2.3.2 데이터베이스 구축 절차

먼저, 원문을 데이터베이스에 저장하기 위해서는 원문을 구별해주는 각 태그들의 유효성을 검증하는 작업이 필요하다. 이러한 유효성 검증 작업을 마친 다음에는 원문으로부터 제목, 원문 내용, 키워드를 추출하여 유니코드로 변환한 후 그 값을 각 테이블에 저장한다. 제목은 tag_jmok_table 테이블에 저장하고, 원문은 edocdata 테이블에 권별로 저장하는데 이번 사업에서는 11책과 12책이 추가로 저장되었으며, 사용자 편의를 위한 제목 리스트 팝업창의 서비스 제공을 위하여 tag_hjmok_list 테이블에 대표 제목 리스트를 저장하였다. 이러한 한국불교전서의 데이터베이스 구축 단계를 정리하면 다음과 같으며 자세한 내용은 본문을 통해서 살펴보도록 한다.

① 태그가 삽입된 원문의 유효성 검증 작업
② 키워드 추출 및 저장
③ 원문 저장

2.3.2.1 태그가 삽입된 원문의 유효성 검증 작업

텍스트 파일로 변환된 원문에 페이지, 제목, 단락, 들여 쓰기, 주석 등을 구별하기 위하여 각각 <PAGE>, <JMOK>, <DAN>, <TAB>, <COMMENT>라는 태그들을 삽입한다. 이러한 태그들은 쌍으로 여는 태그(<...>)와 닫는 태그(</...>)로 구성되어야 하며, 만일 그렇지 않으면 잘못된 데이터가 데이터베이스에 저장될 수 있으므로 반드시 확인이 필요한 작업이다. 이러한 태그들의 검증 작업은 대략 다음과 같은 순서로 이루어진다.

① “*.txt”로 저장된 파일들을 “*.xml”로 확장자 명을 바꾼다.
② xml 문서를 웹 브라우저에서 읽어 들인다.
③ 웹 브라우저에 아무런 에러 메시지가 나타나지 않으면 유효한 문서이고, 에러 메시지가 나타나면 해당되는 내용을 찾아 원문을 수정한다.
④ 모든 태그들이 여는 태그와 닫는 태그로 이루어져야만 유효한 문서를 생성할 수 있다.
⑤ 최종적으로 이렇게 생성된 유효한 문서를 데이터베이스 구축에 사용한다.

2.3.2.2 키워드 추출 및 저장

키워드 추출 및 저장 단계에서 이루어지는 작업은 원문의 저장 단계에서 동시에 처리되는데, 원문 내에서 키워드로 지정된 단어를 찾아 그 위치와 단어를 keyword_index 테이블에 저장하는 작업을 한다. 키워드에 관한 정보는 ekeyword, hkeyword, stroke 테이블에 저장되어 있으며, 각각 키워드의 유니코드 값, 한글 값, 획수 값을 저장하고 있다. 지정된 키워드에 관한 테이블의 자세한 설명은 2.3.3절에서 살펴본다. 이러한 키워드 추출 및 저장 작업은 대략 다음과 같은 순서로 이루어진다.

① ekeyword 테이블로부터 키워드의 목록을 해쉬(hash) 테이블 자료구조 형태로 구축한다.
② 원문에서 한 라인씩 입력받은 문자열 내에서 ①의 해쉬 테이블을 이용하여 키워드들을 찾는다.
③ 키워드의 해당하는 단어들을 keyword_index 테이블에 저장한다.
④ 중복된 키워드를 제거하여 권마다 유일한 키워드 목록만을 따로 keyword 테이블에 저장한다.

2.3.2.3 원문 저장

원문 저장은 유니코드 편집기에서 작성된 유니코드 원문을 그대로 테이블

에 저장하는 단계이다. 원문 파일을 라인(line) 단위로 읽어 원문의 내용을 저장하면서 페이지 태그와 단 태그를 검사하여 페이지 당 라인 수와 단 번호 등의 부가 정보를 생성한다. 이러한 부가 정보는 원문에 대한 인덱스 역할을 한다. 원문 저장에 사용되는 edocdata 테이블의 속성은 다음과 같으며, 테이블에 대한 자세한 내용은 2.3.3절에서 설명한다.

테이블 명: edocdata		
칼럼명	데이타형(길이)	비고
nlinenum	integer	라인 수
sdocdata	nchar(800)	유니코드 원문저장
npagenum	integer	페이지 번호
npageline	integer	페이지에서의 라인 번호
ndannum	integer	단 번호

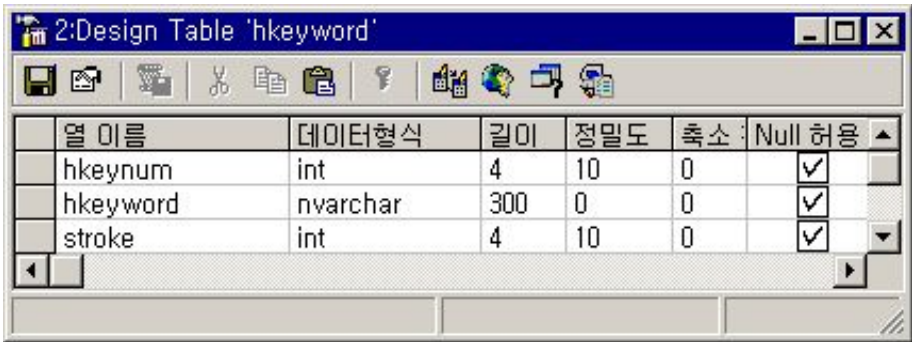
2.3.3 주요 테이블의 세부 내용

본 절에서는 앞에서 일부 다루었던 테이블들에 대해 좀 더 자세히 살펴보도록 한다.

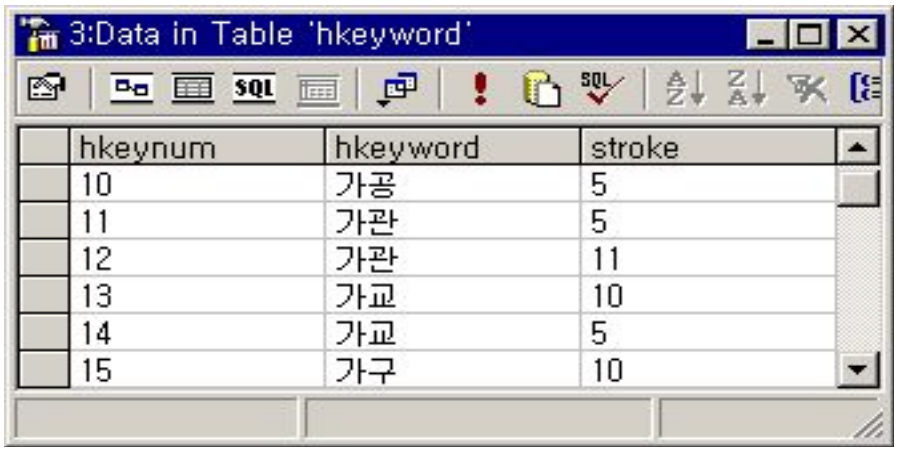
(1) hkeyword 테이블

[그림 6]과 [그림 7]은 hkeyword 테이블의 레코드 속성과 입력된 값을 나타낸 그림이다. hkeyword 테이블은 주로 키워드에 대한 한글 독음과 획수를 저장하고 있으며 자세한 설명은 다음과 같다.

① 테이블 명: hkeyword
② 테이블의 역할: 사용자로부터 한글 키워드를 입력받아, Select 문을 사용하여 해당 키워드의 hkeynum을 얻어오는데 사용한다.
③ 필드의 역할
- hkeynum: 각 키워드에 대한 유일키를 저장한다.
- hkeyword: 키워드에 대한 독음을 저장한다.
- stroke: 키워드의 첫 단어 획수를 저장한다.



[그림 6] hkeyword 테이블의 레코드 속성



[그림 7] hkeyword 테이블에 입력된 데이터

(2) ekeyword 테이블

[그림 8]과 [그림 9]는 ekeyword 테이블의 레코드 속성과 입력된 값을 나타낸 그림이다. ekeyword 테이블은 키워드에 대한 유니코드 값과 획수를 저장하며 자세한 설명은 다음과 같다.

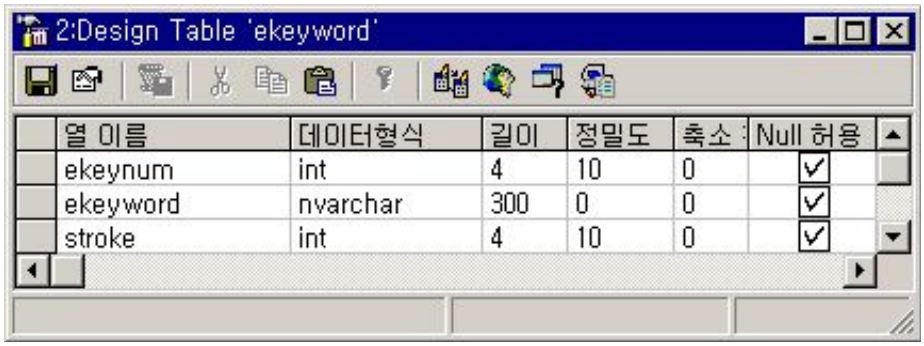
① 테이블 명: ekeyword

② 테이블의 역할: hkeyword 테이블에 저장된 독음에 대한 한자를 유니코드 형태로 저장한다.

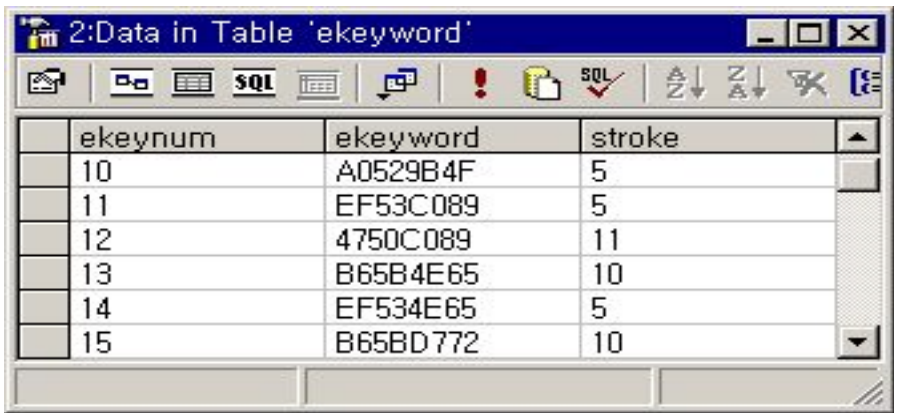
③ 필드의 역할

- ekeynum: 키워드에 대한 유일키를 저장한다.

- ekeyword: 키워드 한자에 대한 유니코드 값을 저장한다.
- stroke: 키워드의 첫 단어 획수를 저장한다.



[그림 8] ekeyword 테이블의 레코드 속성



[그림 9] ekeyword 테이블에 입력된 데이터

(3) keyword_index 테이블

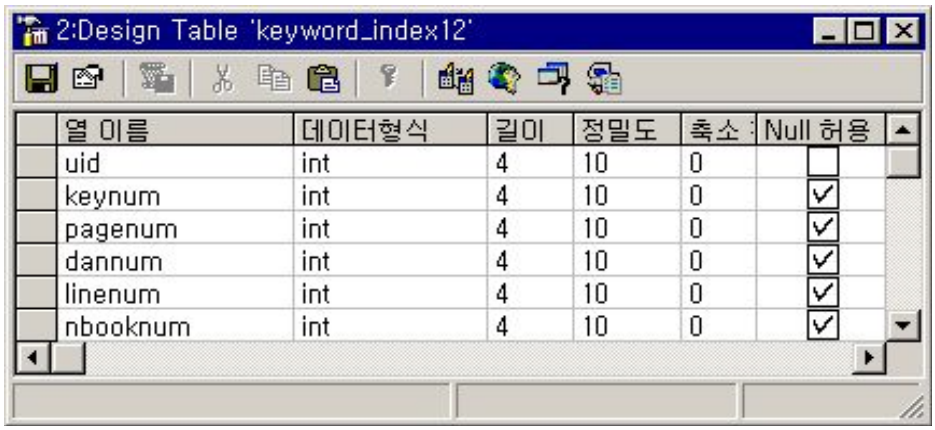
[그림 10]과 [그림 11]은 keyword_index 테이블의 레코드 속성과 입력된 값을 나타낸 그림이다. keyword_index 테이블은 키워드에 대한 인덱스 역할을 하며 자세한 설명은 다음과 같다.

① 테이블 명: keyword_index

② 테이블의 역할: 각 권별로 키워드 인덱스 테이블을 유지한다. 키워드가 발견된 원문의 페이지, 단, 라인에 대한 정보를 저장한다.

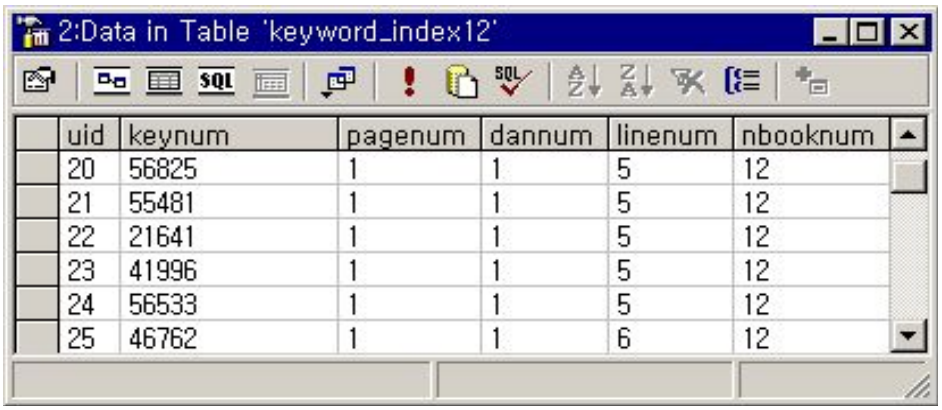
③ 필드의 역할

- uid: keyword_index 테이블의 유일키를 저장한다.
- keynum: 키워드에 대한 유일키를 저장하며, hkeyword 테이블의 hkeynum와 ekeyword 테이블의 ekeynum의 값이 일치한다.
- pagenum: 키워드가 발견된 곳의 페이지 번호를 저장한다.
- dannum: 키워드가 발견된 곳의 단 번호를 저장한다.
- linenum: 키워드가 발견된 곳의 라인번호를 저장한다.
- nbooknum: 현재의 권 번호를 저장한다.



열 이름	데이터형식	길이	정밀도	축소	Null 허용
uid	int	4	10	0	☐
keynum	int	4	10	0	☑
pagenum	int	4	10	0	☑
dannum	int	4	10	0	☑
linenum	int	4	10	0	☑
nbooknum	int	4	10	0	☑

[그림 10] keyword_index 테이블의 레코드 속성



uid	keynum	pagenum	dannum	linenum	nbooknum
20	56825	1	1	5	12
21	55481	1	1	5	12
22	21641	1	1	5	12
23	41996	1	1	5	12
24	56533	1	1	5	12
25	46762	1	1	6	12

[그림 11] keyword_index 테이블에 입력된 데이터

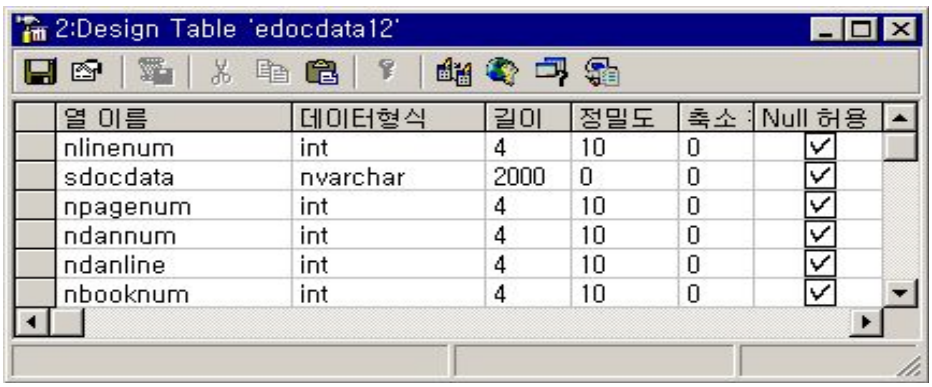
(4) edocdata 테이블

[그림 12]와 [그림 13]은 edocdata 테이블의 레코드 속성과 입력된 값을 나타낸 그림이다. edocdata 테이블은 주로 원문의 내용과 이에 대한 정보를

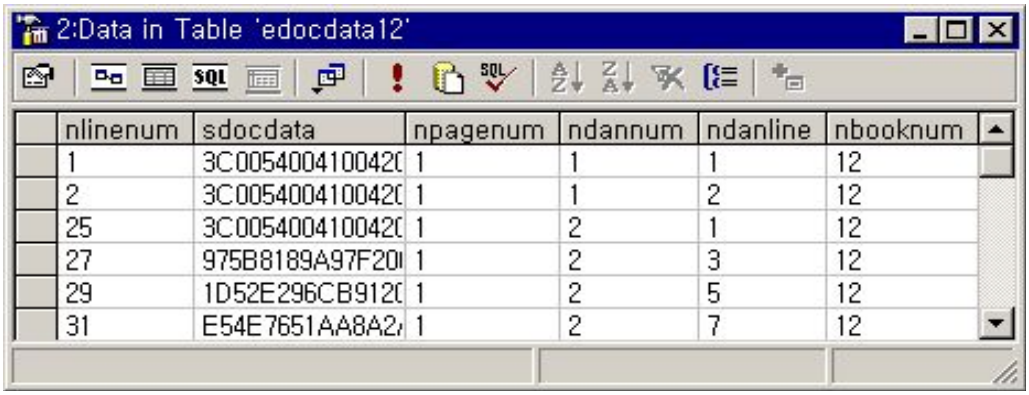
저장하고 있으며 자세한 설명은 다음과 같다.

① 테이블 명: edocdata
② 테이블의 역할: 각 권별로 원문을 저장한다.
③ 필드의 역할

- nlinenum: 원문에 대한 유일키를 저장한다.
- sdocdata: 원문을 유니코드 형태로 저장한다.
- npagenum: 페이지 번호를 저장한다.
- ndannum: 단 번호를 저장한다.
- ndanline: 단의 라인번호를 저장한다.
- nbooknum: 현재 권 번호를 저장한다.



[그림 12] edocdata 테이블의 레코드 속성



[그림 13] edocdata 테이블에 입력된 데이터

⑸ tag_jmok_table 테이블

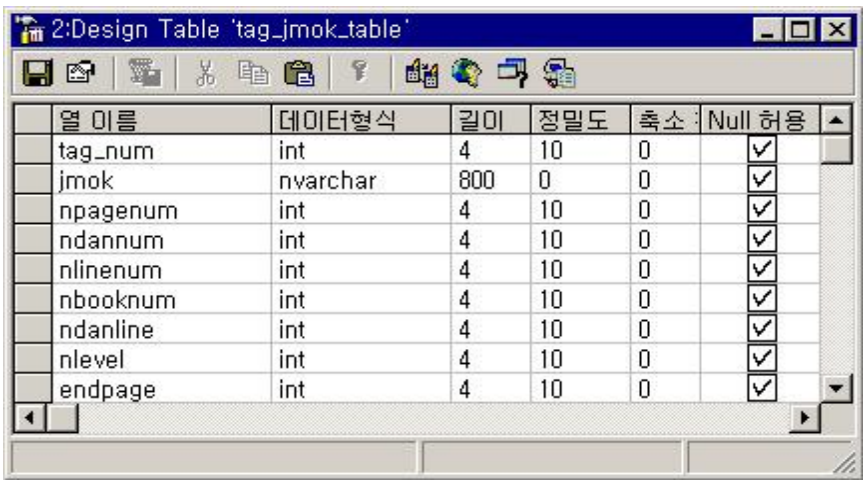
[그림 14]와 [그림 15]는 tag_jmok_table 테이블의 레코드 속성과 입력된 값을 나타낸 그림이다. tag_jmok_table 테이블은 주로 제목에 대한 정보를 저장하고 있으며 자세한 설명은 다음과 같다.

① 테이블 명: tag_jmok_table

② 테이블의 역할: 원문에서 제목이 나타나는 곳의 정보를 저장한다.

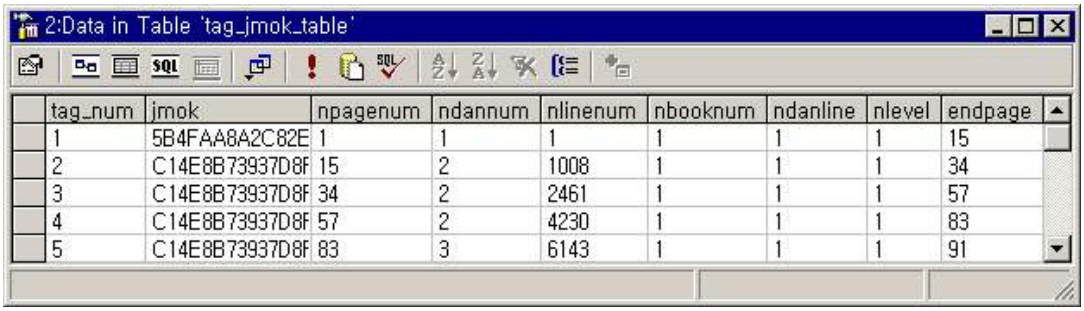
③ 필드의 역할

- tag_num: 각 제목에 대한 유일키를 저장한다.
- jmok: <JMOK> 태그가 나타난 곳의 제목을 유니코드 형태로 저장한다.
- npagenum: 제목 태그의 페이지 번호를 저장한다.
- ndannum: 단 번호를 저장한다.
- nlinenum: 라인번호를 저장한다.
- nbooknum: 현재 권 번호를 저장한다.
- ndanline: 단의 라인번호를 저장한다.
- nlevel: 제목의 레벨을 저장한다.
- endpage: 제목에 해당하는 내용이 끝나는 페이지 번호를 저장한다.



열 이름	데이터형식	길이	정밀도	축소	Null 허용
tag_num	int	4	10	0	☑
jmok	nvarchar	800	0	0	☑
npagenum	int	4	10	0	☑
ndannum	int	4	10	0	☑
nlinenum	int	4	10	0	☑
nbooknum	int	4	10	0	☑
ndanline	int	4	10	0	☑
nlevel	int	4	10	0	☑
endpage	int	4	10	0	☑

[그림 14] tag_jmok_table 테이블의 레코드 속성



tag_num	jmok	npagenum	ndannum	nlinenum	nbooknum	ndanline	nlevel	endpage
1	5B4FAA8A2C82E	1	1	1	1	1	1	15
2	C14E8B73937D8F	15	2	1008	1	1	1	34
3	C14E8B73937D8F	34	2	2461	1	1	1	57
4	C14E8B73937D8F	57	2	4230	1	1	1	83
5	C14E8B73937D8F	83	3	6143	1	1	1	91

[그림 15] tag_jmok_table 테이블에 입력된 데이터

(6) tag_page_table 테이블

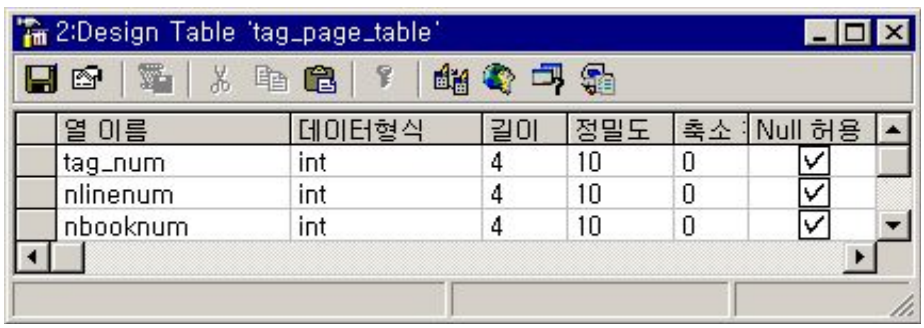
[그림 16]과 [그림 17]은 tag_page_table 테이블의 레코드 속성과 입력된 값을 나타낸 그림이다. tag_page_table 테이블은 tag_jmok_table에 저장된 일련번호를 가지고 페이지 번호만 따로 추출하여 저장하고 있으며 자세한 설명은 다음과 같다.

① 테이블 명: tag_page_table

② 테이블의 역할: 제목 테이블에 대한 요약정보를 가지고 있다.

③ 필드의 역할

- tag_num: 각 제목에 대한 유일키를 저장한다.
- nlinenum: 제목이 있는 곳의 edocdata 테이블의 nlinenum을 저장한다.
- nbooknum: 현재 권 번호를 저장한다.



열 이름	데이터형식	길이	정밀도	축소	Null 허용
tag_num	int	4	10	0	☑
nlinenum	int	4	10	0	☑
nbooknum	int	4	10	0	☑

[그림 16] tag_page_table 테이블의 레코드 속성

tag_num	nlinenum	nbooknum
1	1	1
2	1008	1
3	2461	1
4	4230	1
5	6143	1

[그림 17] tag_page_table 테이블에 입력된 데이터

(7) tag_hjmok_list 테이블

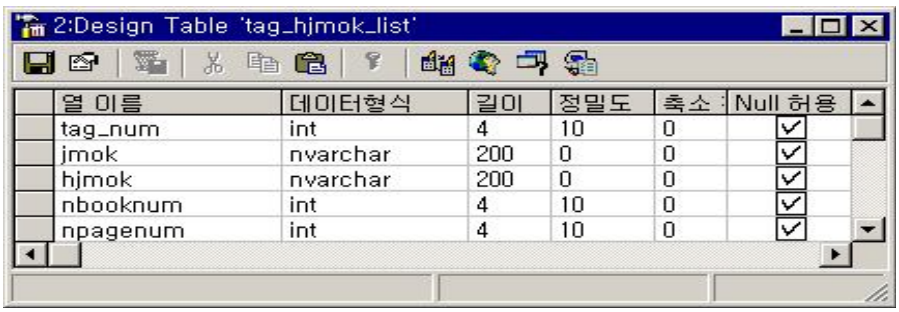
[그림 18]과 [그림 19]는 tag_hjmok_list 테이블의 레코드 속성과 입력된 값을 나타낸 그림이다. tag_hjmok_list 테이블은 제목 리스트가 한자와 한글로 저장되어 있으며 자세한 설명은 다음과 같다.

① 테이블 명: tag_hjmok_list

② 테이블의 역할: 제목 리스트 정보를 가지고 있다.

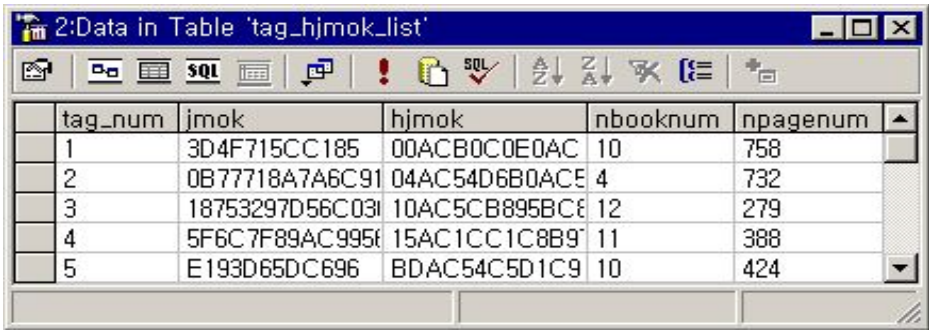
③ 필드의 역할

- tag_num: 각 제목에 대한 유일키를 저장한다.
- jmok: 한자 제목을 유니코드로 변환하여 저장한다.
- hjmok: 한글 제목을 유니코드로 변환하여 저장한다.
- nbooknum: 현재 권 번호를 저장한다.
- npagenum: 제목 태그의 페이지 번호를 저장한다.



열 이름	데이터형식	길이	정밀도	축소	Null 허용
tag_num	int	4	10	0	☑
jmok	nvarchar	200	0	0	☑
hjmok	nvarchar	200	0	0	☑
nbooknum	int	4	10	0	☑
npagenum	int	4	10	0	☑

[그림 18] tag_hjmok_list 테이블의 레코드 속성



tag_num	jmok	hjmok	nbooknum	npagenum
1	3D4F715CC185	00ACB0C0E0AC	10	758
2	0B77718A7A6C91	04AC54D6B0AC5	4	732
3	18753297D56C03I	10AC5CB895BC8	12	279
4	5F6C7F89AC995(	15AC1CC1C8B9	11	388
5	E193D65DC696	BDAC54C5D1C9	10	424

[그림 19] tag_hjmok_list 테이블에 입력된 데이터

(8) idx_keyword_index 테이블

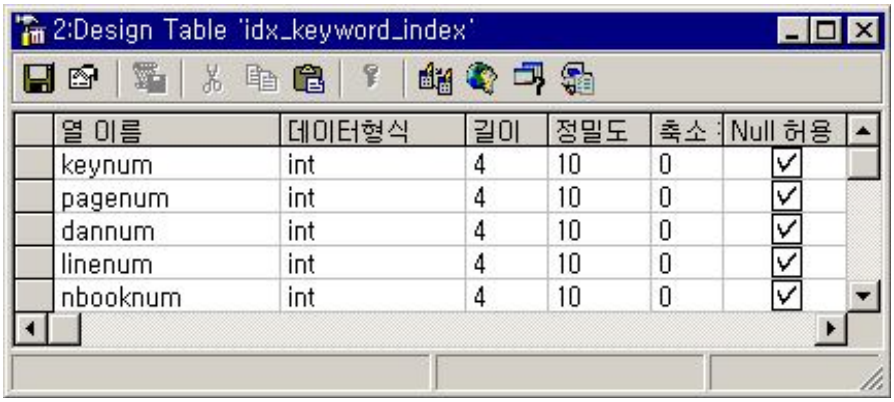
[그림 20]과 [그림 21]은 idx_keyword_index 테이블의 레코드 속성과 입력된 값을 나타낸 그림이다. idx_keyword_index 테이블은 keyword_index1 테이블부터 keyword_index12 테이블을 통합한 테이블로써 용어 검색을 빠르게 하기 위하여 필요한 테이블이며, 자세한 설명은 다음과 같다.

① 테이블 명: idx_keyword_index

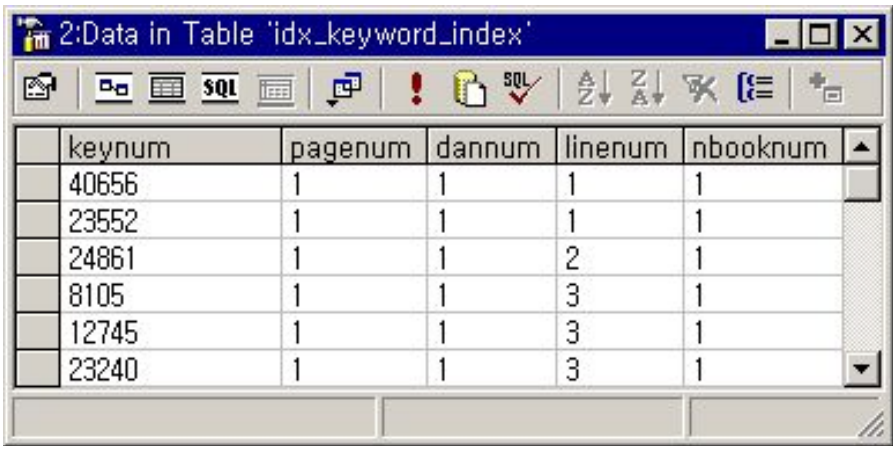
② 테이블의 역할: 모든 권의 키워드 인덱스 테이블을 통합한다. 키워드가 발견된 원문의 페이지, 단, 라인에 대한 정보를 저장한다.

③ 필드의 역할

- keynum: 키워드에 대한 유일키를 저장하며, hkeyword 테이블의 hkeynum와 ekeyword 테이블의 ekeynum의 값이 일치한다.
- pagenum: 키워드가 발견된 곳의 페이지 번호를 저장한다.
- dannum: 키워드가 발견된 곳의 단 번호를 저장한다.
- linenum: 키워드가 발견된 곳의 라인 번호를 저장한다.
- nbooknum: 현재의 권 번호를 저장한다.



[그림 20] idx_keyword_index 테이블의 레코드 속성



[그림 21] idx_keyword_index 테이블에 입력된 데이터

2.4. 웹 검색 인터페이스의 구현

한국불교전서가 제목, 원문, 주석 등으로 구성되어있기 때문에 사용자에게 검색 결과로 보여주는 화면도 경전과 동일하게 구성하였다. 사용자 입장에서 쉽고 편리한 방법으로 검색 할 수 있도록 키워드 검색, 페이지 검색, 제목 검색, 그리고 획수 검색에 이르는 다양한 검색 서비스를 제공하였다. 또한 원문과 동일한 형태로 들여쓰기 기능을 제공함으로서 사용자에게 학술적인 참고 자료로서 가치가 있도록 하였다.

이번 사업에서 사용자 요구사항에 맞게 개선한 기능을 살펴보면 다음과 같다.

- 전체 경 대상 용어 검색 가능
- 한 화면에 10페이지씩 10개 항목 보여주는 기능
- 키워드 검색에서 한자 키워드 검색기능
- 제목 검색을 위한 목록창(Pop-up) 보여주기
- 본문 내용을 3페이지씩 보여주기
- 주석을 나타내는 색인의 크기 줄이기

본 사업에서 진행한 내용은 본문을 통해서 자세히 살펴보도록 한다.

2.4.1 웹 검색시스템의 주요 기능

한국불교전서 웹 검색 인터페이스는 사용자에게 보다 편리한 검색을 위해서 여러 가지 검색 방법을 제공하고 있다. 주요 검색 방법으로는 키워드 검색, 페이지 검색, 제목 검색, 그리고 획수 검색으로 구성되어 있다. [그림 22]는 웹 검색시스템의 주요 기능의 인터페이스를 보여준다.

키워드 검색은 경전의 키워드를 이용해서 검색을 하며, 다양한 검색 조건을 처리할 수 있는 기능을 제공한다. 페이지 검색은 경전을 검색할 때 찾으려는 페이지를 직접 입력해서 검색하는 방법이며 제목 검색은 경전의 각 '권'에 포함되어 있는 제목을 이용하여 검색하는 방법이다. 마지막으로 획수 검색은 한자의 획수를 이용해서 검색하는 방법이다.

2.4.2 개선한 기능

사용자에게 유용하고 편리한 기능을 제공하고 완성도 높은 웹 검색시스템을 위하여 기능 개선을 하였다. 크게 네 가지로, 첫째로 전체 경 대상 용어 검색을 할 수 있도록 하였고, 이때 한 화면에 10페이지씩 10개 항목을 보여주도록 화면 인터페이스 기능을 개선하였다. 둘째로 키워드 검색에서 한자 키워드 검색이 가능하도록 하였고, 셋째로 제목 검색을 위한 목록창(Pop-up) 기능을 제공하였고, 마지막으로 한 화면에 3페이지씩 보여주는 기능 등을 개선하였다. 각 기능에 대한 내용을 좀 더 상세하게 설명하면 다음과 같다.

(1) 전체 경 대상 용어 검색 및 화면 인터페이스 변경

이전 사업에서 전체 경 대상으로 용어를 검색하면, 그 결과 목록이 한 화면에 모두 나타난다. 이것의 문제점은 검색한 후 결과 목록이 많으면, 사용자는 이동 박스를 움직여서 원하는 항목을 찾아야 하는 번거로움이 있다. 본 사업에서, 한 화면에 10페이지씩 10개의 목록을 보여줌으로써 기존의 문제점을 해결하였다. [그림 23]은 개선한 용어 검색 인터페이스 화면으로, 한 화면에 10페이지씩 10개의 목록을 표현한 것이다.

가. 키워드 검색 화면 나. 페이지 검색 화면

다. 제목 검색 라. 획수 검색 화면

[그림 22] 웹 검색 시스템의 주요 기능 인터페이스

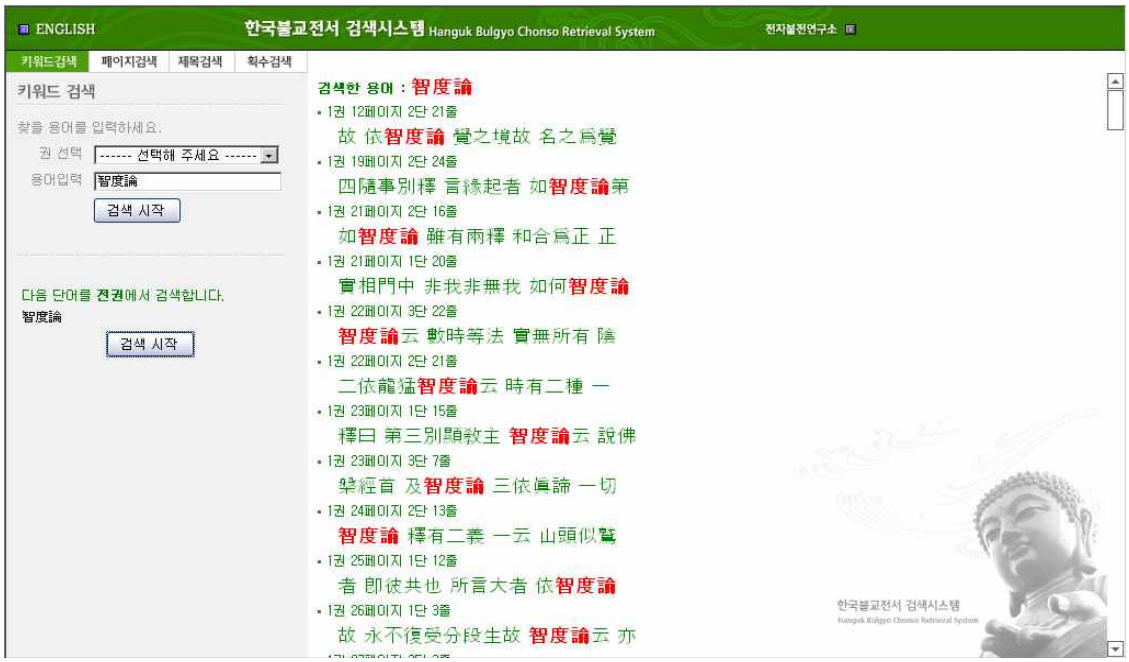
[그림 23] 한 화면에 10페이지씩 10개의 목록을 표현

(2) 한자 키워드 검색 기능

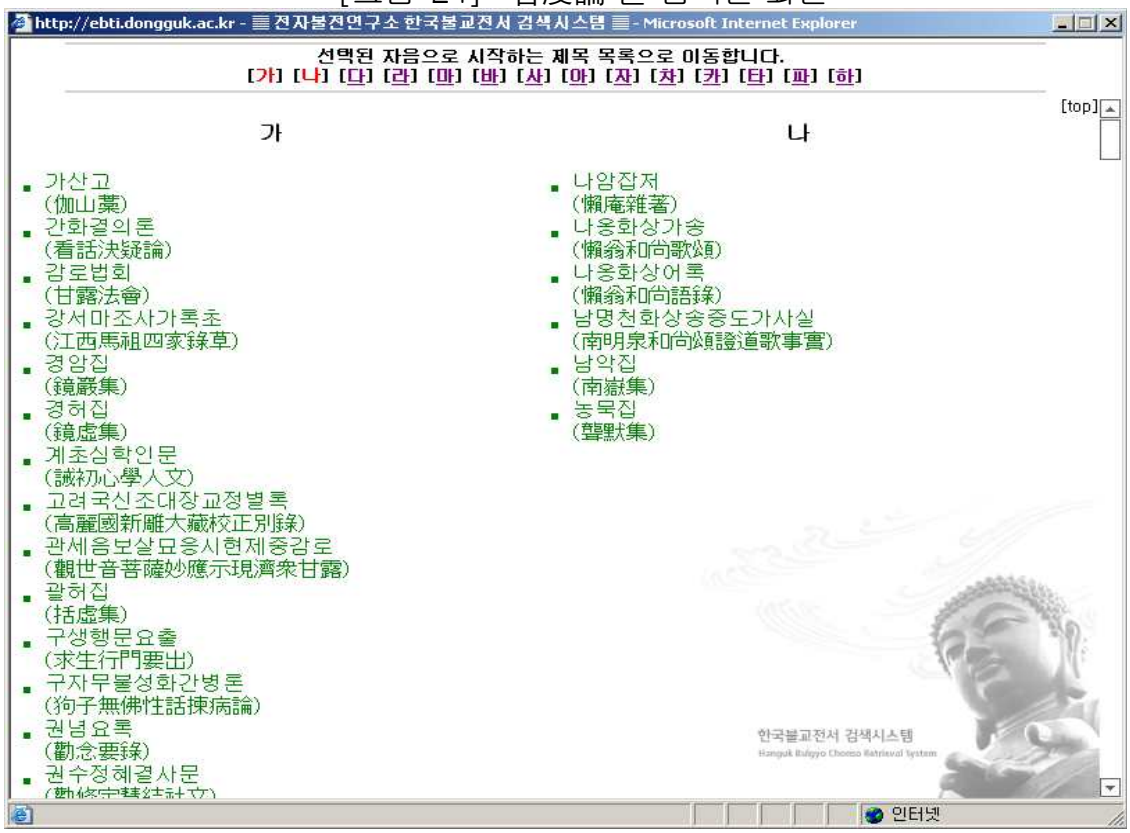
기존 한자 검색의 경우 한자를 용어입력란에 입력하는 것이 아닌 한글 독음을 이용해서 미리 등록되어 있는 한글 독음에 해당하는 한자들을 목록으로 나타내었다. 제6차 사업에서 개선한 기능은 한자를 용어입력란에 직접 키 입력해서 검색하도록 하였다. 이때 사용가능한 한자의 자수는 윈도우 2000 자전에서 제공해주는 기본 자수이다. [그림 24]는 용어입력란에 '智度論'을 입력한 후 검색한 화면이다.

(3) 팝업 창을 열어 제목 목록 보여주기

제목에 대한 한글 독음 목록을 권 순서가 아닌 자음 순서로 정렬하여 팝업 창을 통해서 보여주고, 이 제목을 선택하면 그 내용을 본 화면 우측 틀에 나타내 주는 기능을 추가하였다. [그림 25]는 제목 목록 팝업 창 화면이다.



[그림 24] '智度論'을 검색한 화면



[그림 25] 제목 목록 팝업 창

기존의 제목 검색은 찾고자 하는 제목이 어떤 권에 있는지 알아야만 검색할 수 있다는 단점이 있었으나 한글 제목 목록을 보여 줌으로써 사용자로 하여금 손쉽게 검색할 수 있도록 보완하였다. 상단의 자음을 누르면 해당 자음으로 시작하는 제목 목록으로 이동하며, 원하는 제목을 선택하면 본 화면 우측 틀에는 내용의 첫째 단을 보여주고 좌측 틀에는 세부 제목 목록을 보여준다.

(4) 3페이지씩 보여주기 기능

기존에 제공한 서비스는 한 화면에 한 페이지의 원문을 제공하였기 때문에 다음 페이지를 보기 위한 사용자에게는 다소 번거로움이 따랐다. 이 기능은 사용자가 한 화면에 나타나는 페이지의 수를 선택할 수 있는 기능으로 1페이지에서 3페이지까지 설정할 수 있도록 하였다.

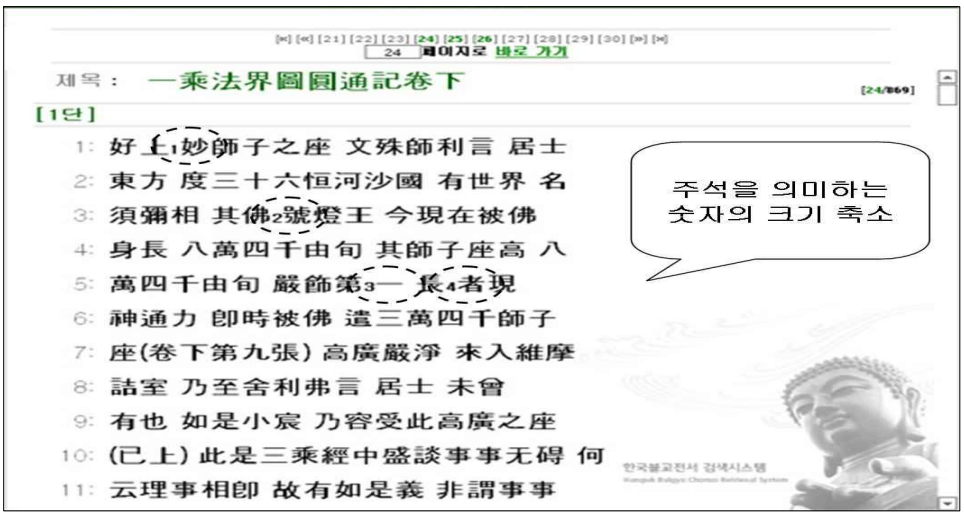
“페이지 검색”의 경우 기본 값은 3페이지로 설정되어 있고, 사용자가 설정한 값은 다음에 변경할 때까지 다른 페이지로 이동하여도 페이지 정보가 유지된다. 키워드 검색, 제목 검색, 그리고 획수 검색은 항상 3페이지씩 보여주도록 고정하였다. [그림 26]은 페이지 검색에서 3페이지를 보여주는 화면이다.

[그림 26] 3페이지씩 보여주기

(5) 주석을 나타내는 색인의 크기 줄이기

경의 원문은 본문 내용을 나타내는 아라비아 숫자와 주석을 나타내는 아라비아 숫자가 있다. 기존에 제공한 서비스는 주석과 본문의 내용을 나타내는 숫자의 글자 크기가 같았기 때문에 원문 내용과 주석을 쉽게 구분할 수 없었다. 개선한 기능은 주석의 아라비아 숫자 크기를 원문 내용보다 작은 글자 크기로 사용하였기에 원문 내용과 주석을 쉽게 구분할 수 있도록 하였다.

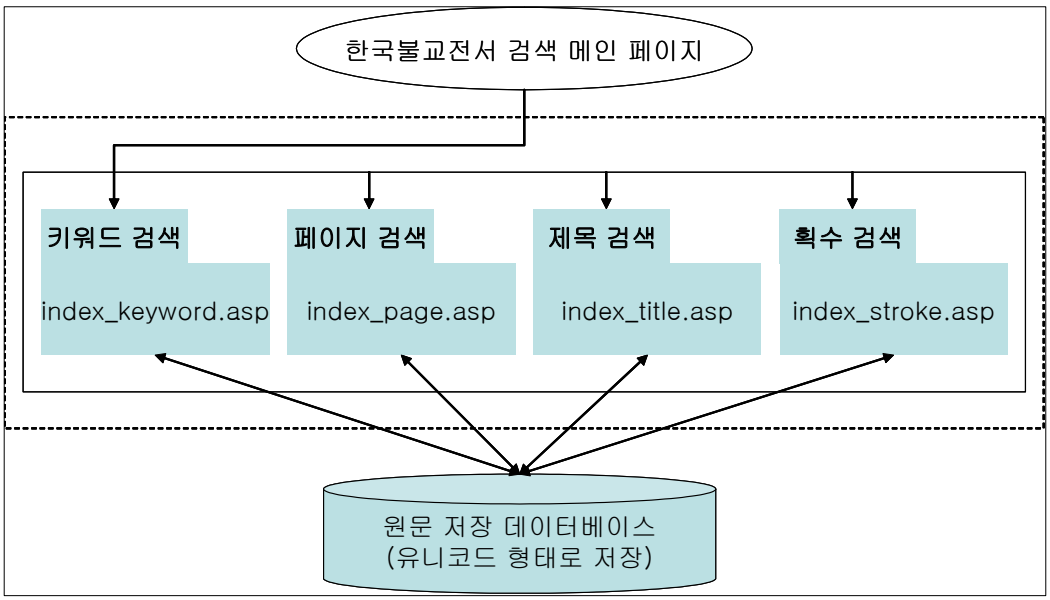
[그림 27]은 주석을 나타내는 숫자를 줄인 것이다.



[그림 27] 주석을 나타내는 숫자의 크기 줄이기

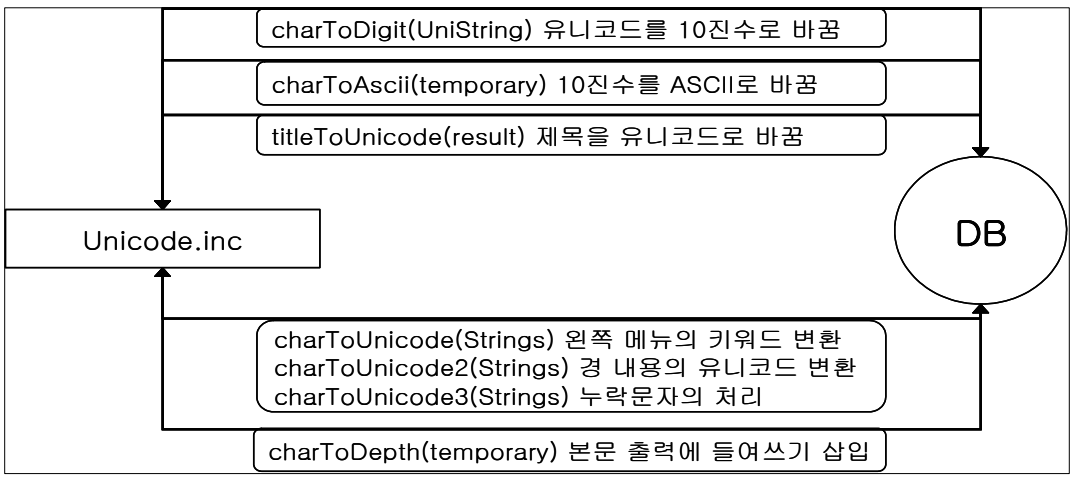
2.4.3 웹 검색 인터페이스의 구현

[그림 28]은 본 검색시스템의 전체 구성도를 나타낸 그림이다. 사용자가 메인화면에서 검색시작 단추를 누르면 기본적으로 키워드 검색 페이지로 이동한다. 페이지 검색, 제목 검색, 그리고 획수 검색으로 이동하기를 원할 때는 해당 탭을 누르면 된다. 사용자가 원하는 검색 결과를 얻는 과정은 해당 검색 페이지에서 검색을 요청하면, 본 검색시스템이 사용자의 검색 요청을 질의문으로 변경한다. 그 다음 원문 저장 데이터베이스에 질의하며, 해당 결과를 사용자에게 보여준다.



[그림 28] 웹 검색 인터페이스의 전체 구성도

또한 데이터베이스에 저장할 때 경전의 내용은 유니코드 형태로 변환해서 저장한다. 그러므로 유니코드와 텍스트간의 변환 기능이 필요하고, [그림 29]는 유니코드와 일반 텍스트 간 변환을 담당하는 함수와 변환 과정을 나타낸 그림이다.



[그림 29] 유니코드와 텍스트간 변환 함수

3. 결론 및 향후 연구과제

본교는 불교학을 중심으로 한 한국학과 컴퓨터 정보통신 두 분야를 특성화의 큰 축으로 하고 있으며, 불교자료의 전산화야 말로 본교의 특성화 방향인 “불교학과 정보통신 기술”의 연계에 가장 적합한 프로그램이라 할 수 있다. 따라서 본 연구에서는 한국불교전적 중 한국불교전서의 일부를 전산화하여 본교의 특성화 사업에 부응하고자 하였다.

현재 우리나라에는 귀중한 불교 문헌들을 포함하여 많은 한문 고문헌들이 있으나 이들에 대한 전산화 작업은 아주 미미한 실정이다. 특히 한국불교 및 한문 고문헌에 대한 연구를 하거나, 필요에 의해 한문 고문헌들을 열람하고 싶을 때 귀중한 자료들이 여러 도서관에 분산되어 있어 손쉽게 이용할 수 없다. 따라서 본 연구를 수행하면 한국불교전서 제11책과 제12책을 전산화 하여 이를 연구하는 연구자들이나 열람을 원하는 사람들에게 도움이 될 뿐만 아니라 우리의 귀중한 문화유산을 전 세계에 널리 알릴 수 있다.

한국불교전적은 우리나라에 불교가 전래된 이래 삼국시대부터 우리나라의 선조들이 남긴 옛 문헌들을 발굴 수집하여 출간한 한국불교전서와 해인사 고려대장경을 동국대학교의 역경원에서 한글로 번역한 한글 대장경을 위시한 한국의 불교 문헌을 총칭한다. 이러한 불교 경전에 대한 활발한 편찬 사업에 비해 전산화 작업이 현재는 미비한 상태이나 전 세계적으로 연구되고 있고, 이러한 디지털 경전에 대한 정보 교환을 위해 국제 회의인 EBTI가 개최되고 있다. 불교정보화와 관한 유일한 국제회의인 EBTI는 인터넷에 제공되고 있는 다양한 불교정보를 서로 자유롭게 이용할 수 있는 호환성을 키우는데 그 중요성이 있다. 이 회의를 통해 각국에서 진행하는 불교정보화에 대해 현재까지 진행된 상황 등 여러 가지 기술에 대한 의견을 교환한다.

이렇게 EBTI에서 논의되고 있는 기술들은 본 연구를 수행하는데도 필요한 기술로 가장 먼저 한자를 컴퓨터에 입력할 수 있는 입력 방법 및 유니코드 상에 없는 고문헌 상의 문자를 처리할 수 있는 시스템의 개발이 필요하

다. 이러한 기술을 개발하기 위해서는 유니코드 등 세계 여러 나라의 각국 언어에 대한 코드 체계 및 방대한 량의 한문 폰트의 확보가 시급하다. 따라서 본 연구소에서는 일본 모직교의 9만 여자의 폰트와 이에 대한 이미지를 이용하여 문헌 중 유니코드 상에 없는 문자를 처리할 수 있는 문자 관리 시스템을 개발하였다. 개발된 문자 관리 시스템을 통해 한국불교전서 제11책과 제12책의 내용을 입력하고, 입력된 내용들을 3번씩 교정 작업을 하여 원문과 다른 글자가 입력되거나 원문에 있는 내용이 빠진 경우들을 없애고 최대한 원문에 가깝게 컴퓨터에 입력하였다.

그리고 본 연구를 수행하는데 두 번째 필요한 기술은 입력된 한국불교전적 원문 내용들을 의미있는 단위로 분할하여 데이터베이스에 저장하는 것이다. 또한 저장된 데이터베이스에서 사용자가 질의를 하면 그 질의에 대해 효율적으로 검색할 수 있는 검색 기술 및 한국불교전적 색인파일 작성 기술이 필요하다. 따라서 크게 다음의 4단계로 데이터베이스 구축을 하였다. 가장 먼저 원문에서 키워드를 추출하여 테이블로 저장하는 동시에 인덱스를 구축하기 위한 파일을 생성하는 단계, 다음은 원문 저장 할 때 XML 태그들을 유지하고 원문의 라인을 유지하면서 저장하는 단계, 그리고 인덱스 구축 및 문서 구조 추출 순서로 이루어진다.

마지막으로 필요한 기술은 데이터베이스에 저장되어 있는 내용들을 검색하기 위하여 전 세계에서 사용하고 있는 인터넷의 웹을 통해 검색할 수 있는 웹 인터페이스와 인터넷이 되지 않는 환경에서도 사용자가 질의를 입력하면 이들을 검색할 수 있는 CD-ROM을 통한 검색 방법이 필요하다. 이에 본 연구에서는 위에서 언급한 3가지 기술들과 사용 방법들을 개발하였다.

본 연구에서 개발된 한국 불교전서 제11책과, 제12책에 대해 웹을 통해 검색하고자 한다면 다음의 URL을 이용하면 된다. URL은 http://ebti.dongguk.ac.kr/ 이다. 향후 연구 과제는 현재 불교사전에 입력되어 있는 약 50,000단어를 모두 색인어로 등록하여 불교학 용어를 거의 망라하고 있다. 그러나 불교어뿐만 아니라 선어록에 많이 등장되는 선어와 인명, 지명 등을 추가 등록하여 보다 많은 색인어로 사용자가 편리하게 이용할 수 있도록 개선할 예정이다. 더불어 본 연구에서 개발된 유니코드에서 누락된

문자 처리 시스템에서 좀 더 효율적이고 빠르며 체계적인 문자의 관리를 위해 기능을 수정 보완해야 한다. 그리고 데이터베이스에 저장된 내용을 검색할 때 원문 전체에 대한 전문 검색 방법도 가능하도록 해야 한다.

참고문헌

[1] Aming Tu, “중국 전자 불전 협회(CBETA)의 전자 『大正新脩大藏經』,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.

[2] Fred Coulson, “전기(傳記)-저서(著書) 목록 검색 데이터베이스로 링크된 텍스트 이미지를 위한 TBRC와 그 모델들,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.

[3] John Lehman, “탈자(脫字) 문제 처리를 위한 프로젝트,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.

[4] Robert Chilton, “아시아 고전 입력 프로젝트 (ACIP): 과거, 현재 그리고 미래,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.

[5] Eric Johnson, The Text Encoding Initiative, Text Technology, 1995.

[6] ISO 8879, Standard Generalized Markup Language, 2nd Edition, 1986.

[7] ISO/IEC 10646-1, “Information Technology - Universal Multiple-Octet Coded Character Set(UCS) - Part I: Architecture and Basic Multilingual Plane,” 1993.

[8] The Unicode Consortium, The Unicode Standard, Version 2.0, Addison Wesley, 1996.

[9] The Unicode Standard, Microsoft Developer’s Network, 1997.

[10] Unicode Enabling, Microsoft Developer's Network, 1997.

[11] Unicode Support in Win32, Microsoft Developer's Network, 1997.

[12] CJK Codes-Unicode/ISO-10646 Unicied “Ideographs,” http://www.mit.edu:8001/afs/athena.mit...r/a/k/akbar/www/Unicode-ideographs.html.

[13] Christian Wittern, “Chinese character codes: an update,” http://www.iijnet. or.jp/iriz/irizhtml/multling/codes/htm.

[14] EditTime, http://www.timelux.lu, TimeLUX.

[15] How to View Chinese/Japanese/Korean HTML with Netscape Communication on US version of Windows 95 or NT,

http://people.netscape.com/ftang/communicatorfont.html.
[16] “Installing Bitstrean Cyberbit Version 1.1,”
http://www.bitstrean.com/cyberbit.html.
[17] “Notes on CJK Character Codes and Encodings,”
http://www.ifcss.org/ftp-pub/software/info/cjk-codes.
[18] Panorama, http://www.softquad.com, Softquad.
[19] Public Unicode Font,
ftp://www.ifcss.org/ftp-pub/software/fonts /unicode.
[20] True Type and Unicode,
http://truetype.demon.co.uk:80/unicode.htm.
[21] Urs App, “A Look at the Korean Tripitaka Input Project,”
http://www.iijnet.or.jp/iriz/irizhtml/ebit/samsung.htm.
[22] Urs App, “Guidlines for the Creation of Large Chinese Text Databases,”
http://www.iijnet.or.jp/iriz/irizhtml/maketext/guideline. html.
[23] Urs App, “The Importance of Markup,”
http://www.iijnet.or.jp/iriz/irizhtml/maketext/foguang.html.
[24] 대장경학술용어연구회, “대정신수 대장경소인,” 제1권, 대장경학술용어연구회, 1975.
[25] 송석구, “전자불전과 미래불교의 향방,” ’01 동국대학교 개교 95주년 기념 세계 전자불전학회 학술대회, 2001.
[26] 이금석, “한국불교전서 전산화에서의 누락문자관리,” ’00 동국대학교 전자불전연구소 제2회 세미나, 2000.
[27] 한태식, “불교학 연구에 있어서 한국불교전서의 위상,” ’00 동국대학교 전자불전연구소 제2회 세미나, 2000.
[28] 허인섭, “Report on the Digital Tripitaka Koreana 2001,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
[29] 현득창, 임광택, 이수연, “SGML 기본 파서를 이용한 SGML문서 편집기의 구현,” 한국정보과학회, 정보과학회 눈문지, Vol 25, No. 1, 1998.
[30] 홍영식, “한국 불교전서 데이터베이스에서 누락문자 검색,” ’01 동국대학교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
[31] 김숙자, SGML의 모든 것, 성안당, 1997.
[32] 동국대학교 출판부 발행, 한국불교전서 제9-10권 조선시대편, 1979.
[33] 장회창, 현득창, 이수연, SGML 가이드, 사이버출판사, 1997.
[34] 한국불교신문, 시방시계 1월 27일자, 현대불교신문사, 1999.
[35] 황기태 역, 어드밴스 윈도우 NT, 도서출판 대림, 1995.

[36] 강석진, “팔만사천대장경 전산화를 위한 제언, 한자위주 문헌의 워드프로세서 데이터베이스, 탁상출판 시스템 개발을 위해,”
http://members.iWorld.net/hederein/menu22/Kang.html.
[37] 김응철, “고려장경 및 한자정보전산화에 관련한 문제제기,”
http://members.iWorld.net/hederein/menu22/Kim.html.
[38] 노용균, “불전 전산화와 SGML,”
http://members.iWorld.net/hederein/menu22/Dogam42.html.
[39] 심재룡, “정보화 사회와 불교 전산화,”
http://members.iWorld.net/hederein/menu22/Dogam32.html.
[40] 이규갑, “고려대장경 전산화에 있어서 이체자의 처리 문제,”
http://members.iWorld.net/hederein/menu22/Yi.html.
[41] 인터넷으로 만나는 불교,
http://members.iWorld.net/hederein/menu23/Pogyu121.html.
[42] 정주원, “ISO/IEC-10646 Universal Multiple-Object Coded Character Set (UCS)에 대해서,”
http://simac.kaist.ac.kr/~jwjung/seminar/hangul-i18n/iso10646.html.
[43] 정주원, “한글 코드에 대하여,”
http://simac.kaist.ac.l [illegible] ko-code.html.
[44] 종림스님, “팔만대장경 전산
http://members.iWorld.net/hederein/menu22/
[45] 혜묵스님, “세계의 불교자료 전산화 계획과 고려대장경 전산화를 위한 몇 가지 문제들,” http://members.iWorld.net/hederein/menu22/Hye. html.

키워드(Keyword)

한국불교전서, 한국불교전서 검색 시스템, 한국불교전서 전산화, 유니코드, 누락문자

Korea Bulgyo Chonso, Korea Bulgyo Chonso Retrieval System, Korea Bulgyo Chonso Digitalization, Unicode, Missing Character