한국불교전서 검색 시스템 개발

구현우*, 김영희*, 박미화*, 이재수**, 신병삼**,
이용규***, 이금석***, 홍영식***, 한보광****

목 차

요 약

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.

이러한 불교경전에 대한 편찬 사업은 활발히 진행되고 있는 것에 비하여 불교전적의 전산화 작업은 현재 미비한 상태이나, 세계적으로는 이 분야에 대한 연구가 활발히 진행 중에 있다.

* 동국대학교 컴퓨터공학과
** 동국대학교 불교학과, 전자불전·문화재콘텐츠연구소 전임연구원
*** 동국대학교 컴퓨터공학과 교수
**** 동국대학교 선학과 교수, 전자불전·문화재콘텐츠연구소 소장

이에 발맞춰 본 연구는 한국불교전서 전산화 7차 사업으로 한국불교전서 제 13책과 14책을 전산화하여 전 세계에서 활발하게 사용되고 있는 인터넷을 통하여 검색할 수 있도록 하는 것에 목적을 두고 있다.

이에 한 걸음 더 나아가 다양한 환경에서 한국불교전서를 검색 가능케 하기 위한 검색 시스템을 CD로 제작하여 보급한다.

이러한 연구 목적을 달성하기 위해 크게 3가지 기술이 필요하다. 한국불교전적을 컴퓨터에 입력하고 이를 편집하여, 데이터베이스에 저장하고, 데이터베이스에 저장된 내용들을 웹뿐만 아니라 인터넷이 되지 않는 컴퓨터에서도 검색할 수 있는 인터페이스와 검색 기술이 필요하다

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 7차 사업으로 한국불교전서 제 13책과 14책을 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.

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 문

헌을 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].

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 전

산화만을 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.

EBTI에서 논의되고 있는 기술들은 본 연구를 수행하는데도 필요한 기술로서 첫째로 한자를 컴퓨터에 입력할 수 있는 입력방법 및 입력된 한자들을 편집할 수 있는 편집 시스템의 개발이 필요하다.

둘째로 이를 통해 웹 상에서 한자를 검색할 수 있는 시스템이 필요하다.

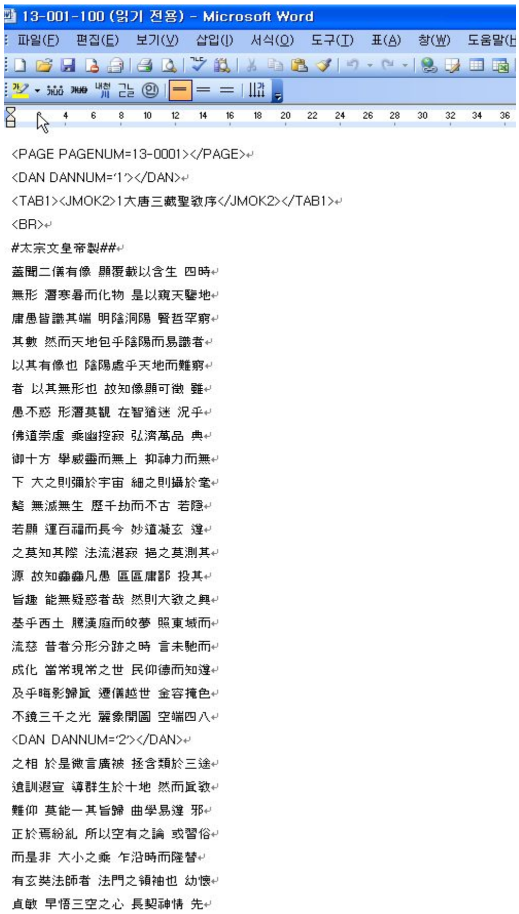
2. 한국불교전서 전산화 7차 사업

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 먼저 한국불교전서의 입력 및 교정 방법에 대해 설명한다.

2.1. 한국불교전서의 입력 및 교정

한국불교전서의 입력은 전문 입력기관인 (주)솔트웍스에 입력용역을 주어서 입력하였다. 기존의 연구원들이 직접 스캔을 통해 교정하여, 입력하였을 때보다 보다 입력의 정확성을 기할 수 있게 되었다. 다음의 〔그림 1〕은 한국불교전서 13책의 1쪽이다.

그림 1. 한국불교전서 13책 1면 사진



13-001-100 (읽기 전용) - Microsoft Word

<PAGE PAGENUM=13-0001></PAGE>
<DAN DANNUM='1'></DAN>
<TAB1><JMOK2>1大唐三藏聖敎序</JMOK2></TAB1>

#太宗文皇帝製##
蓋聞二儀有像 顯覆載以含生 四時
無形 潛寒暑而化物 是以窺天鑒地
庸愚皆識其端 明陰洞陽 賢哲罕窮
其數 然而天地包乎陰陽而易識者
以其有像也 陰陽處乎天地而難窮
者 以其無形也 故知像顯可徵 雖
愚不惑 形潛莫覩 在智猶迷 況乎
佛道崇虛 乘幽控寂 弘濟萬品 典
御十方 擧威靈而無上 抑神力而無
下 大之則彌於宇宙 細之則攝於毫
釐 無滅無生 歷千劫而不古 若隱
若顯 運百福而長今 妙道凝玄 遵
之莫知其際 法流湛寂 挹之莫測其
源 故知蠢蠢凡愚 區區庸鄙 投其
旨趣 能無疑惑者哉 然則大敎之興
基乎西土 騰漢庭而皎夢 照東域而
流慈 昔者分形分跡之時 言未馳而
成化 當常現常之世 民仰德而知遵
及乎晦影歸眞 遷儀越世 金容掩色
不鏡三千之光 麗象開圖 空端四八
<DAN DANNUM='2'></DAN>
之相 於是微言廣被 拯含類於三途
遺訓遐宣 導群生於十地 然而眞敎
難仰 莫能一其旨歸 曲學易遵 邪
正於焉紛糺 所以空有之論 或習俗
而是非 大小之乘 乍沿時而隆替
有玄奘法師者 法門之領袖也 幼懷
貞敏 早悟三空之心 長契神情 先

그림 2. 입력된 원문 텍스트

〔그림 1〕의 한국불교전서 13책 1면의 텍스트를 MS-Word 2003을 통해 입력된 문서는 〔그림 2〕와 같다.

2.2. 교정과 색인

이번 전자불전연구소에서 수행한 한국불교전서 전산화 7차 사업에서는 한국불교전서 제13책(보유편 3)과 제14책(보유편 4)의 전체를 입력하고 교정하였다.

지금까지 한국불교전서 총 14권 가운데 전 14책을 입력하고 교정, 태그를 달아 전산화함으로써, 한국불교전서의 신라시대편과 고려시대편, 조선시

대편과 보유편 4권의 전산화를 완성하였다. 이는 한국불교전서 전체 분량에 해당한다.

또한 이번 7차 사업인 한국불교전서 제13책(보유편 3)과 제14책 (보유편 4)에서는 주로 원문 텍스트 위주로 구성되었다.

이번 사업에서 입력·교정한 저술은 제13책과 제14책은 모두 瑜伽論記 1종, 20권이다. 입력한 저술들의 목록은 다음과 같다.

제 13 책 (보유편 3)

제 14 책 (보유편 4)

한국불교전서 13책(1102쪽)과 14책(962쪽)의 분량은 전체 2,064쪽이며, 글자수로는 약 206만자를 교정하였다. 입력팀의 3명이 분량을 나누어 5차례에 걸쳐 교정작업을 하였으며, 교정과정은 다음과 같다.

1) 제1차 교정 : 처음 입력된 것을 틀린 자가 없는지 한글자 한글자 원본과 대조하면서 확인하고, 누락문자로 표시된 & 기호가 정확한 누락문자인지 이체자 인지를 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 다시 입력팀에서 태그와 문단구성, 누락문자, 이체자 등 전체적으로 오류는 없는지 다시 한번 최후 교정 작업함.
5) 제5차 교정 : 작업이 모두 끝난 원문이 웹 상에서 제대로 구현되는지를 다시 한 번 살펴본 후 최종 교정 작업을 마친다.

이러한 입력과 5번의 교정 등 여섯 단계를 거쳐 한국불교전서 제13책, 제14책의 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
兎	兎	乌	烏	财	財
骚	騷	攒	攢	畧	略
烟	煙	鬪	鬪	覓, 覓	覓

색인작업은 다음과 같이 진행되었다.

현재 불교사전에 입력되어 있는 약 50,000단어를 모두 색인어로 등록하여 불교학 용어를 거의 망라하고 있다. 향후에는 불교단어뿐만 아니라 선어록에 많이 등장되는 선어와 인명, 지명 등을 추가 등록하여 보다 많은 색인어로 사용자가 보다 편리하게 이용할 수 있도록 개선할 예정이다.

2.3. 태깅

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

우선 태깅 작업은 다음과 같다.

1) 제목 태깅 - 원제목을 나타내준다.

<JMOK1>瑜伽論記</JMOK1>

2) 소제목 태깅 - 원제목에 딸린 소제목을 나타낸다.
<JMOK2>大唐三藏聖教序</JMOK2>

3) 쪽수 태깅 - 쪽수를 알려준다.
<PAGE PAGENUM='13-5'></PAGE>

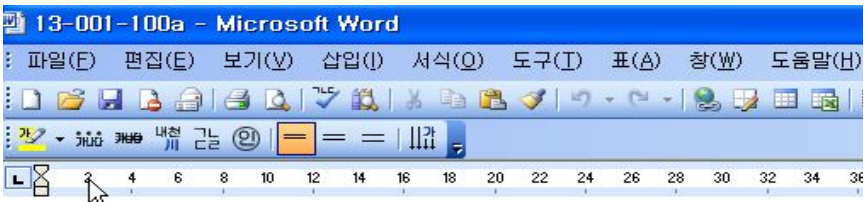
4) 단락을 표시해 주는 태깅 - 1, 2, 3단을 나타내준다.
<DAN DANNUM='1'></DAN>
<DAN DANNUM='2'></DAN>
<DAN DANNUM='3'></DAN>

5) 이미지 태깅 - 이미지로 처리해야 할 부분이다.
<IMAGE 10-111-1-1></IMAGE>

6) 주석 태깅 - 주석임을 나타내준다.
<COMMENT>
{1}〔持神〕作〔特〕{甲}. {2}〔座〕作〔坐〕{甲}. 甲本註曰〔坐〕作〔座〕{甲}. {3}〔漸〕作〔流〕{甲}. {4}〔式〕作〔求〕{甲}. {5}甲本註曰〔時〕作〔持〕{乙}.
</COMMENT>

7) 탭 태깅 - 들여쓰기를 책과 같이 해준다.
1칸 들여쓰기: <TAB1></TAB1>
2칸 들여쓰기: <TAB2></TAB2>
3칸 들여쓰기: <TAB3></TAB3>

실제 태깅 작업이 진행된 원문은 다음의 〔그림 3〕과 같다.



<PAGE PAGENUM='13-1'></PAGE>↵
<DAN DANNUM='1'></DAN>↵
<JMOK1>瑜伽論記</JMOK1>↵
<JMOK2>大唐三藏聖敎序</JMOK2>↵
<TAB1>1大唐三藏聖敎序</TAB1>↵
↵
<TAB9>太宗文皇帝製**</TAB9>**↵
蓋聞二儀有像 顯覆載以含生 四時↵
無形 潛寒暑而化物 是以窺天鑒地↵
庸愚皆識其端 明陰洞陽 賢哲罕窮↵
其數 然而天地包乎陰陽而易識者↵
以其有像也 陰陽處乎天地而難窮↵
者 以其無形也 故知像顯可徵 雖↵
愚不惑 形潛莫覩 在智猶迷 況乎↵
佛道崇虛 乘幽控寂 弘濟萬品 典↵
御十方 擧威靈而無上 抑神力而無↵
下 大之則彌於宇宙 細之則攝於毫↵
鼇 無滅無生 歷千劫而不古 若隱↵
若顯 運百福而長今 妙道凝玄 遵↵
之莫知其際 法流湛寂 挹之莫測其↵
源 故知蠢蠢凡愚 區區庸鄙 投其↵
旨趣 能無疑惑者哉 然則大敎之興↵
基乎西土 騰漢庭而皎夢 照東域而↵
流慈 昔者分形分跡之時 言未馳而↵
成化 當常現常之世 民仰德而知遵↵
及乎晦影歸眞 遷儀越世 金容掩色↵
不鏡三千之光 麗象開圖 空端四八↵
<DAN DANNUM='2'></DAN>↵
之相 於是微言廣被 拯含類於三途↵
遒訓遐宣 導群生於十地 然而眞敎↵
難仰 莫能一其旨歸 曲學易遵 邪↵
正於焉紛紜 所以空有之論 或習俗↵
而是非 大小之乘 乍沿時而隆替↵
有玄奘法師者 法門之領袖也 幼懷↵

그림 3. 태깅 작업을 한 원문

3. 데이터베이스 저장

본 절에서는 한국불교전서를 데이터베이스로 저장하고 관리하기 위한 기술을 설명한다. 한국불교전서를 데이터베이스로 저장하기 위해서는 원문을 구별해주는 각 태그들의 유효성을 검증하는 작업과 원문으로부터 제목, 원문 내용, 키워드를 추출하여 유니코드로 변환하고 해당 테이블에 값을 저장하는 작업이 필요하다. 또한 데이터베이스 저장과 관리를 위한 저장 시스템의 구현이 요구된다. 이를 위해, 본 절에서는 한국불교전서 데이터베이스 구축 단계와 데이터베이스 저장 시스템의 구성과 기능, 데이터베이스 구조에 대해 자세히 설명한다.

3.1 한국불교전서 데이터베이스 구축 단계

한국불교전서는 제목, 원문, 주석 등으로 구성되어 있고, 각각에 해당되는 내용은 태그로 구별하여 데이터베이스를 구축한다. 원문에 나타나는 한문은 기존 문자 셋으로 표현하는데 한계가 있어서 유니코드로 변환하여 저장한다.

한국불교전서 데이터베이스를 구축하기 위해 먼저, 원문을 구별해주는 각 태그들의 유효성을 검증하는 작업을 진행한다. 유효성 검증 작업 후 원문으로부터 제목, 원문 내용, 키워드를 추출하여 유니코드로 변환한 후 그 값을 해당 테이블에 저장한다.

3.1.1. 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 여는 태그(〈...〉)와 닫는 태그(〈/...〉)가 쌍으로 구성되어야 한다. 태그가 잘못 작성되면 잘못된 데이터가 데이터베이스에 들어갈 수 있으므로 유효성 검증 작업이 필요하다. 원문의

유효성 검증 작업은 다음과 같은 순서로 이루어진다.

① "*.txt"로 저장된 파일들을 "*.xml"로 확장자 명을 바꾼다.

② xml 문서를 웹 브라우저에서 읽어 들인다.

③ 웹 브라우저에 에러 메시지가 나타나지 않으면 유효한 문서이고, 에러 메시지가 나타나면 해당되는 내용을 찾아 원문을 수정한다.

3.1.2. 키워드 추출 및 저장

키워드 추출 및 저장 단계에서는 원문 내에서 키워드로 지정된 단어를 찾아 그 위치와 단어를 keyword_index 테이블에 저장한다. 미리 지정된 키워드 목록 정보는 ekeyword, hkeyword 테이블에 저장되어 있으며, 각각 키워드의 유니코드 값, 한글 값을 저장하고 있다. 키워드에 관련된 테이블의 자세한 설명은 2.3절에서 살펴본다. 키워드 추출 및 저장 작업은 다음과 같은 순서로 이루어진다.

① ekeyword 테이블로부터 키워드의 목록을 해쉬(hash) 테이블 자료구조 형태로 구축한다.

② 원문을 한 라인씩 읽으면서 ①의 해쉬 테이블에 있는 키워드들을 찾는다.

③ 키워드가 있으면 해당 단어들을 keyword_index 테이블에 저장한다. 최종적으로 keyword_index 테이블에는 각 권에 대한 유일한 키워드 목록이 저장된다.

3.1.3. 원문 저장

원문 저장 단계에서는 유니코드 편집기에서 작성된 유니코드 원문을 테이블에 저장한다. 원문 파일을 줄(line) 단위로 읽어 edocdata 테이블에 저장한다. 또한 페이지 태그와 단 태그를 검사하여 페이지 당 라인 수와 단 번호 등의 부가 정보를 생성한다. 부가 정보는 원문에 대한 인덱스 역할을 한다. 원문 저장에 사용되는 edocdata 테이블의 속성은 2, 3절에서 설명한다.

3.2 데이터베이스 원문 저장 시스템 구성과 기능

〔그림 4〕는 데이터베이스 원문 저장시스템의 모듈 구조도를 나타낸 그림이다. 주화면 처리 모듈에서 DB 관리 모듈을 호출하면, DB 관리 모듈에서는 파일 관리 모듈을 호출하여 저장할 원문 텍스트 파일을 읽어온다. 읽어온 원문 텍스트 파일은 파서 모듈에서 태그를 분리하여 원문 내용, 페이지, 단, 라인 정보를 추출한다. 추출한 키워드 및 원문 내용은 유니코드 변환 모듈을 거쳐 edocdata, tag_page_table, tag_jmok_table 테이블 등에 각각 저장한다.

키워드 인덱스 생성 모듈은 해시 테이블을 이용하여 ekeyword, hkeyword, tag_jmok_table 테이블에서 키워드 목록을 가져와 keyword_index 테이블에 저장한다. 이 중에서 유일 키워드 목록은 Idx_keyword_index 테이블에 저장한다.

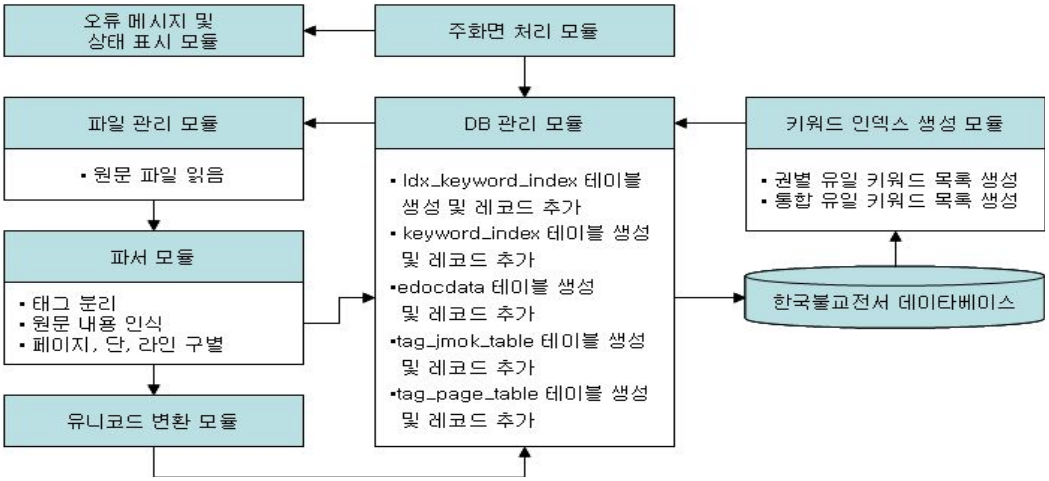
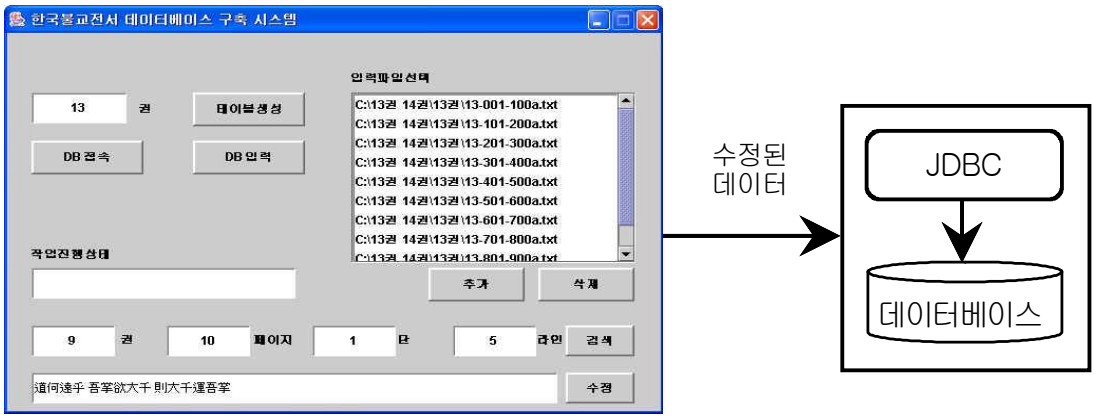


그림 4. 데이터베이스 원문 저장시스템 모듈 구조도

〔그림 5〕는 데이터베이스 저장 시스템의 주 화면을 나타낸다. 원문 저장 프로그램을 이용하여 원문 데이터를 입력하면 이를 데이터베이스에 저장한다. 이때, 프로그램과 데이터베이스간의 연동을 위하여 JDBC를 사용하고,

DBMS로는 마이크로소프트사의 SQL 2000을 사용하였다.



원문 저장 프로그램

그림 5. 데이터베이스 원문 저장 시스템의 주 화면

데이터베이스 원문 저장 프로그램의 주요 기능인 원문 저장 기능, 원문 검색 기능, 원문 수정 기능들에 대해서 자세히 살펴보면 다음과 같다.

3.2.1. 원문 저장 기능

원문 저장은 〔그림 6〕과 같이 새로 저장할 원문의 권을 입력하고 "테이블 생성" 버튼을 눌러 테이블을 생성한다. 이미 테이블이 생성되어 있다면 에러 메시지를 출력한다. "입력 파일 선택"란에 "추가" 버튼을 이용하여 해당 권에 관련된 원문 파일을 순서대로 추가한다. "삭제" 버튼을 이용하여 추가된 항목을 삭제할 수 있다. 이때

그림 6. 원문 저장 화면

권에 해당하는 모든 파일을 목록에 추가해야 정확한 데이터가 생성된다. 그 후 "DB 입력" 버튼을 눌러 원문을 데이터베이스에 저장한다. 작업의 진행 상황은 "작업진행상태"란에서 볼 수 있다. 원문이 유효한 문서가 아닌 경우 처리 과정 중에 에러가 발생하고 에러 메시지를 출력하게 된다. 이러한 일련의 과정을 거쳐 정상적으로 처리되면 차례로 keyword, keyword_index, edocdata 테이블이 생성된다.

3.2.2. 원문 검색 기능

원문 검색은 〔그림 7〕과 같이 원문 저장 프로그램에 의해 입력된 데이터를 검색한다. 먼저 프로그램의 하단에 검색하고자 하는 책의 권, 페이지, 단락, 라인 정보를 입력한다. 그런 다음 "검색" 버튼을 누르면 해당되는 내용의 결과가 아래쪽 글상자에 보인다. 입력된 정보의 범위가 잘못된 경우 에러 메시지가 표시된다. 원문 저장 프로그램의 이러한 검색 기능을 이용하면 원문에서 검색하고자 하는 내용을 찾을 수 있을 뿐만 아니라, 데이터의 오류 부분을 쉽게 찾아낼 수도 있다.

그림 7. 원문 검색 화면

3.2.3. 원문 수정 기능

원문 수정은 〔그림 8〕과 같이 오류가 있는 원문 부분을 찾아 수정한다. 원문 수정 기능은 원문의 양이 많아져서 원문 전체를 전부 입력할 때 생기는 시간과 노력을 줄이기 위해 필요하다. 수정하고자 하는 권, 페이지, 단의 정보를 입력하고 수정할 라인 수를 입력한 후에 "검색" 버튼을 누른다. 그러

면 아래쪽에 있는 글상자에 해당 내용이 출력된다. 오류가 있는 부분을 수정한 후에 오른쪽 아래에 있는 "수정" 버튼을 누르면 해당부분이 수정된다.

그림 8. 원문 수정 화면

3.3 데이터베이스 테이블 구성

본 절에서는 한국 불교전서 데이터베이스를 구성하는 테이블들을 자세히 기술한다.

3.3.1. hkeyword 테이블

hkeyword 테이블은 키워드에 대한 한글 독음과 획수를 포함하고 있다. 〔그림 9〕는 hkeyword 테이블의 구조와 입력된 값을 나타낸다.

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.

'hkeyword' 테이블의 데이터('budhaj_...

hkeynum	hkeyword	stroke
100	가루라	9
101	가루라법	9
102	가루바다	9
103	가루라염	9
104	가루염	9
105	가루오타이	9
106	가루오타이경	5
107	가루자	5
108	가류타이	9
109	가류파다	9
110	가를다	12

그림 9. hkeyword 테이블의 구조와 데이타

3.3.2. ekeyword 테이블

ekeyword 테이블은 키워드에 대한 유니코드 값과 획수를 저장한다. 〔그림 10〕은 ekeyword 테이블의 레코드 구조와 입력된 값을 나타낸다.

① 테이블 명 : ekeyword

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

'ekeyword' 테이블의 데이터('budhaj_...

ekeynum	ekeyword	stroke
50	EF67857F	9
51	4C6B857F	14
52	E68F857F426C85	9
53	E68F857F426C85	9
54	E68F857FE99C51	9
55	E68F857FE99CB	9
56	E554857F857F	10
57	EF67857F857F	9
58	A052857F857F	5
59	E68F857F857F	9
60	4C6B8F90857F	14

그림 10. ekeyword 테이블의 구조와 데이터

③ 필드의 역할

• ekeynum : 키워드에 대한 유일키를 저장한다.

• ekeyword : 키워드 한자에 대한 유니코드 값을 저장한다.

• stroke : 키워드의 첫 단어 획수를 저장한다.

3.3.3. keyword_index 테이블

keyword_index 테이블은 키워드에 대한 인덱스 정보를 저장한다. 〔그림 11〕은 keyword_index 테이블의 레코드 구조와 입력된 값을 나타낸다.

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, hkeword 테이블의 hkeynum와 ekeyword 테이블의 ekeynum의 값이 일치한다.
- ekeyword : 각 키워드의 유니코드를 저장한다.
- 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.

'keyword_index1' 테이블의 데이터('budhaj_06v01' 데이터베이스, 'EBTI7' 컴퓨터)

uid	keynum	pagenum	dannum	linenum	nbooknum
100	21396	1	1	11	1
101	41120	1	1	11	1
102	41155	1	1	11	1
103	49083	1	1	11	1
104	49136	1	1	11	1
105	55481	1	1	11	1
106	56355	1	1	11	1
107	56887	1	1	11	1
108	57975	1	1	11	1
109	6516	1	1	12	1
110	24387	1	1	12	1
111	27715	1	1	12	1
112	27773	1	1	12	1
113	31058	1	1	12	1
114	41049	1	1	12	1
115	45463	1	1	12	1

그림 11. keyword_index 테이블의 구조와 입력 데이터

3.3.4. edocdata 테이블

edocdata 테이블은 원문의 내용과 부가 정보를 저장한다. 〔그림 12〕는 edocdata 테이블의 레코드 구조와 입력된 값을 나타낸다.

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.

'edocdata1' 테이블의 데이터('budhaj_06v01' 데이터베이스, 'EBTI7' 컴퓨터)

nlinenum	sdocdata	npagenum	ndannum	ndanline	nbooknum
100	7075200045657C5	2	2	5	1
101	F88A5B4F0C545F	2	2	6	1
102	8C4E2E7A200000	2	2	7	1
103	8C4E05801653AE	2	2	8	1
104	E68F8259864F20(	2	2	9	1
105	10625B4F2000D6:	2	2	10	1
106	4E6567F9559006;	2	2	11	1
107	4C88F16D2C82E!	2	2	12	1
108	2C7B8C4EA88FF	2	2	13	1
109	8C5FA88F7A662E	2	2	14	1
110	C0897A662000F1	2	2	15	1
111	4C88F16D200021	2	2	16	1
112	2171FD8040624C	2	2	17	1
113	2759C154F06620(	2	2	18	1
114	4C88F16D2C82E!	2	2	19	1
115	2171F8762000767	2	2	20	1

그림 12. edocdata 테이블의 구조와 입력 데이터

3.3.5. tag_jmok_table 테이블

tag_jmok_table 테이블은 제목에 대한 정보를 저장한다. 〔그림 13〕은 tag_jmok_table 테이블의 레코드 구조와 입력된 값을 나타낸다.

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.

'tag_jmok_table' 테이블의 데이터('budhaj_06v01' 데이터베이스, 'EBTI7' 컴퓨터)

tag_num	jmok	npagenum	ndannum	nlinenum	nbooknum	ndanline	nlevel	endpage
100	5C743D4F41F918	674	1	49002	2	1	1	700
101	5C743D4F2B5E3(	700	3	50951	2	1	1	710
102	5C743D4F41F918	711	1	51707	2	1	1	733
103	5C743D4F41F918	733	3	53366	2	1	1	749
104	5C743D4F41F918	749	3	54525	2	1	1	762
105	5C743D4F41F918	763	1	55500	2	1	1	776
106	5C743D4F41F918	776	3	56507	2	1	1	796
107	5C743D4F41F918	797	1	57998	2	1	1	810
108	5C743D4F41F918	810	3	58985	2	1	1	837
109	5C743D4F41F918	837	2	60949	2	1	1	846
110	5C743D4FD68A1(	1	1	1	3	1	1	13
111	5C743D4F41F918	14	1	912	3	1	1	35
112	5C743D4F41F918	36	1	2505	3	1	1	52
113	5C743D4F41F918	52	2	3667	3	1	1	68
114	5C743D4F41F918	69	1	4907	3	1	1	88
115	5C743D4F41F918	88	2	6385	3	1	1	100

그림 13. tag_jmok_table 테이블의 구조와 입력 데이터

3.3.6. tag_page_table 테이블

tag_page_table 테이블은 제목 검색 편의를 위해 제목에 대한 요약 정보를 저장한다. 〔그림 14〕는 tag_page_table 테이블의 레코드 속성과 입력된 값을 나타낸다.

'tag_page_table' 테이블의 데이터('bu...

tag_num	nlinenum	nbooknum
100	49002	2
101	50951	2
102	51707	2
103	53366	2
104	54525	2
105	55500	2
106	56507	2
107	57998	2
108	58985	2
109	60949	2
110	1	3

그림 14. tag_page_table 테이블의 구조와 데이터

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. tag_jmok_table의 tag_num을 참조한다.
- nlinenum : 제목이 있는 곳의 edocdata 테이블의 nlinenum을 저장한다.
- nbooknum : 현재 권 번호를 저장한다.

3.3.7. tag_hjmok_list 테이블

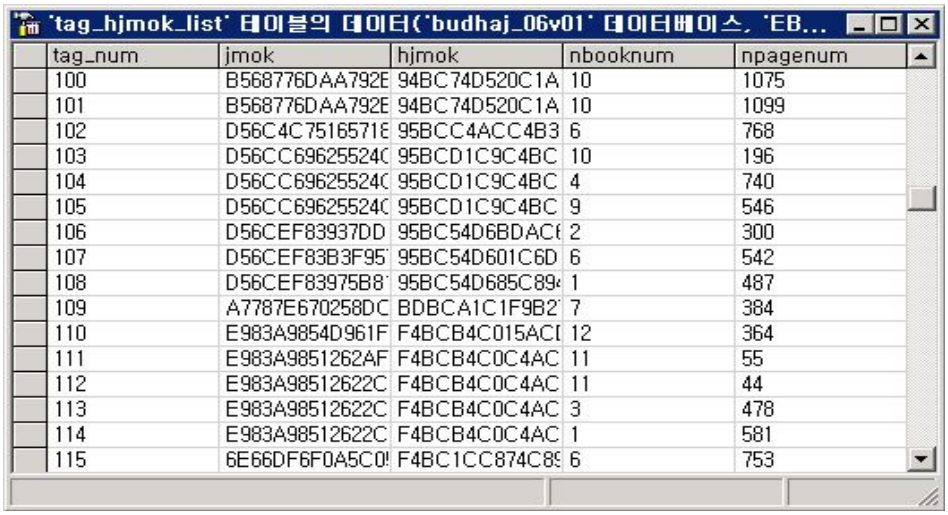
tag_hjmok_list 테이블은 제목 리스트를 한자와 한글로 저장한다. 〔그림 15〕는 tag_hjmok_list 테이블의 레코드 구조와 입력된 값을 나타낸다.

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.


'tag_hjmok_list' 테이블의 데이터('budhaj_06v01' 데이터베이스, 'EB...

tag_num	jmok	hjmok	nbooknum	npagenum
100	B568776DAA792E	94BC74D520C1A	10	1075
101	B568776DAA792E	94BC74D520C1A	10	1099
102	D56C4C7516571E	95BCC4ACC4B3	6	768
103	D56CC69625524C	95BCD1C9C4BC	10	196
104	D56CC69625524C	95BCD1C9C4BC	4	740
105	D56CC69625524C	95BCD1C9C4BC	9	546
106	D56CEF83937DD	95BC54D6BDACE	2	300
107	D56CEF83B3F95	95BC54D601C6D	6	542
108	D56CEF83975B8	95BC54D685C89	1	487
109	A7787E670258DC	BDBCA1C1F9B2	7	384
110	E983A9854D961F	F4BCB4C015ACE	12	364
111	E983A9851262AF	F4BCB4C0C4AC	11	55
112	E983A98512622C	F4BCB4C0C4AC	11	44
113	E983A98512622C	F4BCB4C0C4AC	3	478
114	E983A98512622C	F4BCB4C0C4AC	1	581
115	6E66DF6F0A5C0	F4BC1CC874C8	6	753

그림 15. tag_hjmok_list 테이블 구조와 입력 데이터

3.3.8. Idx_keyword_index 테이블

Idx_keyword_index 테이블은 키워드 검색 속도 향상을 위해 모든 권의 키워드 인덱스 테이블을 통합한 유일한 키워드 인덱스 목록을 저장한다. 통합 키워드 인덱스 목록은 한국불교전서의 각 권별 키워드 인덱스 테이블인 keyword_index 테이블을 이용하여 생성한다. 〔그림 16〕은 Idx_keyword_index 테이블의 레코드 구조와 입력된 값을 나타낸다.

① 테이블 명 : Idx_keyword_index

② 테이블의 역할 : 모든 권의 키워드 인덱스 테이블을 통합한 유일 키워드 목록 정보를 저장한다. 키워드가 발견된 원문의 페이지, 단에 대한 정보를 저장한다.

③ 필드의 역할

- uid : Idx_keyword_index 테이블의 유일키를 저장한다.
- keynum : 키워드에 대한 유일키를 저장하며, hkeword 테이블의 hkeynum와 ekeyword 테이블의 ekeynum의 값을 참조한다.
- 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.

'Idx_keyword_index2' 테이블의 데이터('budhaj_06v01' 데이터베이스, 'EBTI7' 컴퓨터)

uid	keynum	pagenum	dannum	linenum	nbooknum
100	21396	1	1	11	1
101	41120	1	1	11	1
102	41155	1	1	11	1
103	49083	1	1	11	1
104	49136	1	1	11	1
105	55481	1	1	11	1
106	56355	1	1	11	1
107	56887	1	1	11	1
108	57975	1	1	11	1
109	6516	1	1	12	1
110	24387	1	1	12	1
111	27715	1	1	12	1
112	27773	1	1	12	1
113	31058	1	1	12	1
114	41049	1	1	12	1
115	45463	1	1	12	1

그림 16. Idx_keyword_index 테이블 구조와 입력 데이터

4. 검색 시스템 설계 및 구현

4.1. 개요

한국불교전서는 인터넷과 웹 브라우저를 이용한 검색 서비스가 제공되고 있다. 그리고 이에 한 걸음 더 나아가 인터넷이 지원되지 않는 환경에서도 손쉽게 한국불교전서 문헌을 검색 및 열람이 가능하도록 하기 위해 CD-ROM으로 검색 시스템을 제작할 필요가 있다. 따라서 7차 사업에서는 오프라인 검색 서비스를 지원하기 위한 한국 불교 전서 검색 시스템의 설계 및 구

현을 진행하였다. 이러한 검색 시스템은 인터넷이 지원되지 않는 PC에 설치하여 사용할 수 있으며 한국불교전서의 방대한 량의 데이터 보관 기능을 포함하기도 한다.

오프라인 검색 시스템의 모든 검색 방법 및 사용법을 웹-기반 검색 서비스와 동일하게 개발하였으며, 윈도우즈에서 자주 사용되는 인터페이스인 도움말 구조를 사용하여 사용자에게 쉽고, 일관성 있는 사용법을 제공하고 있다.

4.2. 검색 시스템의 구조

〔그림 17〕은 검색 시스템의 전체 구조도를 보여준다. 검색 시스템은 크게 데이터베이스 검색 엔진, HTML 문서 생성기 그리고 HTML 문서를 사용자에게 보여줄 HTML 브라우저로 구성된다. 검색 시스템에서 사용되는 불교 전서 데이터베이스는 웹에서 사용되는 데이터베이스를 가져와 검색 시스템에 필요한 요소만 추출하여 사용한다. 검색 시스템의 전체 구조를 자세히 살펴보면 먼저 사용자가 검색을 명령하면 데이터베이스 검색 엔진에서 해당 데이터베이스에서 원문을 가져오기 위해 입력 받은 사용자 명령을 SQL Query 문으로 변경하고 데이터베이스에서 해당 원문을 가지고 온다. 데이터베이스에서 검색된 원본 데이터는 HTML 생성기에 입력 자료로 사용된다. 데이터베이스에 저장되어 있는 원본 데이터는 유니코드 형식의 텍스트로 되어 있어 웹 브라우저에서 바로 사용할 수가 없다. 따라서 HTML 생성기는 입력 받은 원본 데이터를 웹 브라

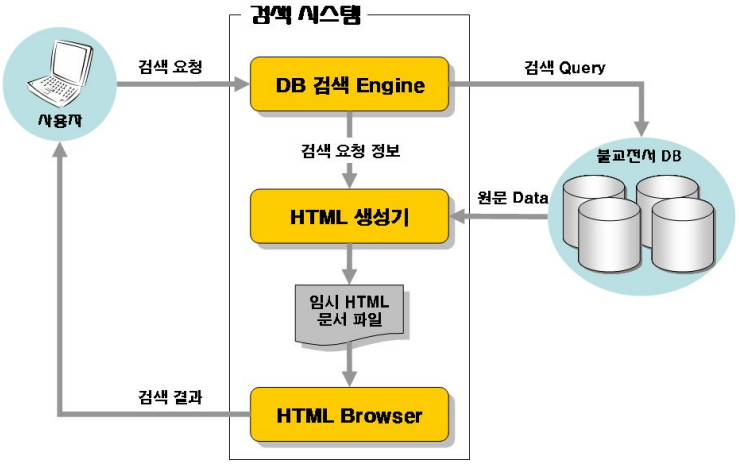


그림 17. 검색 시스템의 구조도

우저에서 볼 수 있는 형태의 HTML 문서로 변화하여 HTML 브라우저에게 전달한다. HTML 생성기는 원본 데이터 시작과 끝에 몇 가지 정보를 추가하고 편리한 검색 서비스를 제공하기위한 페이징 기법 및 이미지 파일 처리하여 임시 HTML 문서를 생성한다. 이렇게 생성된 HTML 문서는 HTML 브라우저에 전달되어 사용자가 검색을 원하는 원본 데이터를 화면에 출력한다. HTML 브라우저는 인터넷을 이용한 검색 서비스와 동일한 화면 구조를 가지고 있어 인터넷을 이용하여 불교 전서를 검색한 경험이 있는 사용자에게 편이성을 제공할 수 있다.

4.3. 검색 시스템의 검색 방법에 따른 세부 구조

검색 시스템은 크게 네 가지 검색 방법을 사용자에게 지원한다. 웹을 통한 검색과 같이 키워드에 의한 검색, 원문 쪽수를 이용한 검색, 원문 제목을 이용한 검색 그리고 키워드의 한자 획수를 이용한 검색이다. 〔그림 18〕은 각 검색 방법의 작동 흐름을 보여준다.

첫 번째 검색방법인 키워드 검색은 사용자가 검색할 권 또는 전체를 선택하고 검색하고 싶은 용어를 입력하면 불교전서 데이터베이스 중 키워드 테이블에서 입력된 용어가 지정된 키워드인지를 검사한다. 만약 지정된 키워드가 아니면 경고 메시지를 출력하고 지정된 키워드이면 한글로 입력된 용어와 음이 같은 키워드들을 가지고 키워드 HTML 문서를 작성한다. 키워드 HTML 문서는 용어와 음이 같은 한자 키워드를 사용자에게 보여주기 위해 작성된 문서이다. 작성된 키워드 HTML 문서에서 사용자가 찾고자 하는 하나의 키워드를 선택하면 키워드 색인 테이블에서 선택된 키워드의 ID를 가지는 원문 페이지들을 검색한다.. 검색된 원문 페이지들의 페이지 번호와 키워드가 들어 있는 하나의 문장을 사용자에게 보여주기 위한 키워드 검색 결과 HTML 문서를 작성한다. 키워드 검색 결과 HTML 문서에 나타나는 정보는 찾고자 하는 키워드가 포함되어 있는 원문 페이지의 요약이다. 키워드 검색 결과 HTML 문서에서 하나의 원문 페이지를 선택하면 원문 테이블에 선택된 원문 페이지의 ID가 입력되고 검색을 하게 된다. 검색된 결과 원문

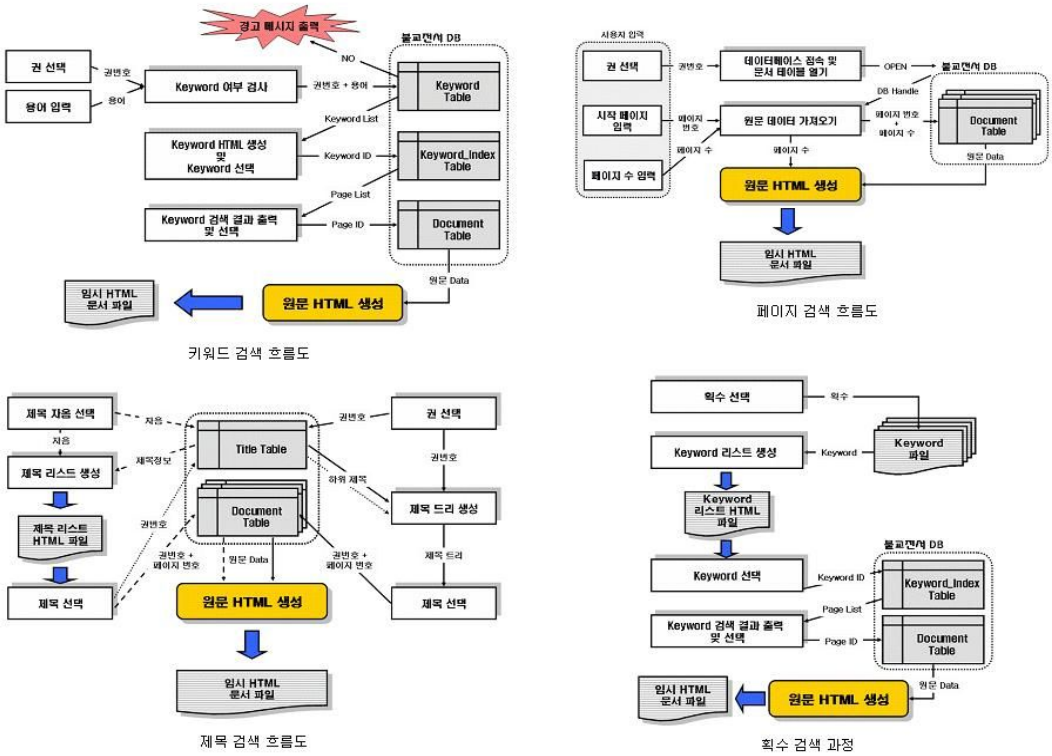


그림 18. 검색 흐름도

은 HTML 생성기의 입력이 되고 최종적으로 원문을 출력하게 된다. 〔그림 19〕는 키워드 검색의 사용자 인터페이스와 사용자의 검색 작업 순서를 보여주고 〔그림 20〕은 키워드 검색의 결과를 보여준다.

그림 19. 키워드 검색 인터페이스

그림 20. 키워드 검색 결과

두 번째, 검색 방법인 페이지 검색은 사용자가 검색할 권을 선택하면 데이터베이스 검색 엔진은 불교전서 데이터베이스에서 입력된 권의 문서 테이블을 열고 해당 테이블에 대한 제어권을 가지고 온다. 사용자가 권에서의 시작 페이지를 입력하고 한 화면에 보여줄 페이지의 수를 입력하면 시작 페이지 번호를 이용하여 제어권을 넘겨받은 테이블에서 해당 원문 데이터를 추출하고 원문 HTML 생성 작업을 하여 최종적으로 원문을 출력한다. 〔그림 21〕은 페이지 검색의 사용자 인터페이스와 사용자의 검색 작업 순서 및 검색의 결과를 보여준다.

세 번째 검색 방법인 제목 검색은 크게 두 가지 검색 방법을 제공한다. 첫 번째는 한글 자음에 해당하는 제목의 리스트에서 검색을 원하는 제목을 선택하는 방법이다. 두 번째 검색 방법은 권을 선택하고 해당하는 권의 제목을 트리 형태로 표현하여 해당하는 제목을 선택하여 원문을 출력하는 방법이다. 두 번째 방법과 같이 첫 번째 검색 방법에서도 선택된 제목이 포함되어 있는 권의 하위 제목을 트리 형태로 보여준다.

그림 21. 페이지 검색 인터페이스 및 결과

두 방법의 동작 과정을 자세히 살펴보면 다음과 같다. 첫 번째 방법을 살펴보면 먼저, 찾고자 하는 제목의 첫 번째 글자의 한글 자음을 선택하면 해당하는 자음을 첫 번째 글자의 자음으로 가지고 있는 제목 리스트를 보여주는 창이 생성된다. 이때 제목 리스트를 보여주기 위해 데이터베이스에 있는 제목 테이블에 접근하여 정보를 가져온다. 제목 리스트에서 찾고자 하는 제목을 클릭하면 현재 제목 리스트를 보여주는 창이 사라지고 메인 화면에 선택된 제목을 지닌 권의 하위 제목 트리와 검색결과를 보여주게 된다. 두 번

째 방법은 먼저 찾고자 하는 권을 선택하고 검색을 시작하면 해당하는 권의 하위 제목 트리가 생성되고 제목 트리에서 원하는 제목을 선택하면 결과가 화면에 출력되는 동작 과정을 보인다.

그림 22. 제목 리스트 검색 인터페이스

그림 23. 제목 검색 결과 화면

첫 번째 방법으로 생성된 제목 트리를 이용하여 두 번째 방법에서 결과를 출력한 과정을 그대로 진행 할 수 있다. 〔그림 22〕와 〔그림 23〕은 첫 번째 제목 검색 방법의 사용자 인터페이스와 사용자의 검색 작업 순서 및 검색의 결과를 보여주고 〔그림 24〕는 두 번째 제목 검색 방법의 사용자 인터페이스와 사용자의 검색 작업 순서 및 검색의 결과를 보여준다.

마지막 검색 방법인 획수 검색은 키워드의 한자 표기에서 첫 번째 글자의 한자 획수를 이용한 검색으로 키워드들을 첫 글자의 획수로 분류하고 사용자가 획수를 선택하면 해당하는 키워드를 화면에 출력한다. 그리고 출력된 키워드를 사용자가 선택하는 방법을 취한다. 키워드 선택 이후의 동작은 앞서 본 키워드 검색 방법의 흐름과 같다. 획

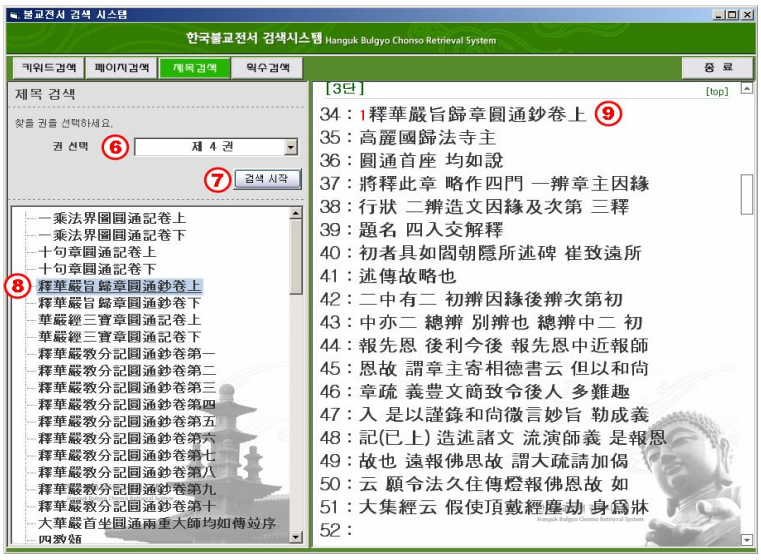


그림 24. 제목 리스트를 이용하지 않는 제목 검색 인터페이스와 결과

수를 선택하여 키워드 리스트를 화면에 출력할 때 데이터베이스를 접근하지 않고 미리 획수별 분류된 키워드 파일에서 키워드 데이터를 불러 키워드 리스트를 작성한다. 데이터베이스를 직접 접근하지 않는 이유는 원문에서 사용되고 있는 키워드의 전체 수가 58,000개 이상이기 때문에 데이터베이스에 직접 접근하여 해당 데이터를 불러오는 시간이 오래 걸리는 문제가 있기 때문에 미리 획수에 따른 키워드를 분류하고 하나의 파일을 생성하여 획수 검색에 사용한다.

〔그림 25〕는 획수 검색의 사용자 인터페이스와 사용자의 검색 작업 순서를 보여주고 〔그림 26〕은 획수 검색의 결과를 보여준다.

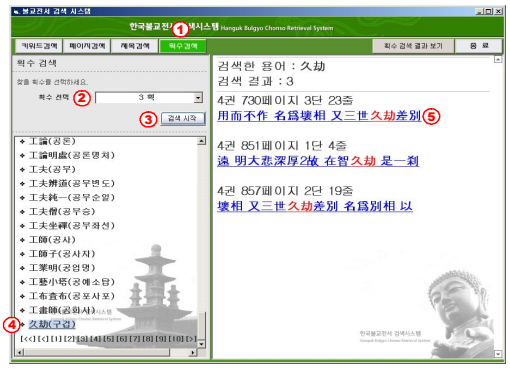


그림 25. 획수 검색 인터페이스

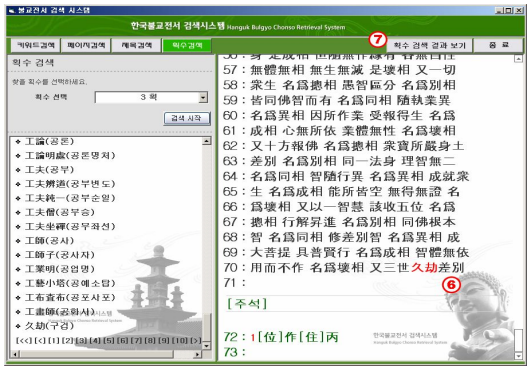


그림 26. 획수 검색 결과

4.4. 원문 HTML 생성

데이터베이스의 실제 한국 불전 원문 데이터의 모든 내용은 "edocdata" 테이블에 저장되어 있다. 원문 HTML 생성기는 SQL_SERVER로부터 데이터를 불러 HTML형식에 맞게 변환 후 임시 HTML 파일로 변환시킨다.

〔그림 27〕은 원문 HTML 생성기의 동작 흐름을 보여준다. 각 검색 방법에서 전달 받은 검색 정보 중 권 번호와 한 화면에 표시할 페이지 수를 이용하여 다음 페이지 이동, 이전 페이지 이동, 첫 페이지 이동, 마지막 페이지 이동을 위한 Tag를 생성한다. 이러한 작업 Paging 기법이라 부른다. 검색 정보에서 권 번호 와 페이지 번호를 이용하여 제목 과 원문을 추출하고

각각의 데이터를 웹에서 사용자 볼 수 있는 형태의 데이터로 변환한다. 원문은 단과 주석을 다른 형태로 표시하기 위한 추가적인 처리과정을 거친다. 만약 키워드 검색이나 획수 검색을 통해 키워드가 존재하면 키워드를 화면에서 다른 색으로 표시하기 위한 키워드 처리 단계를 거치게 된다.

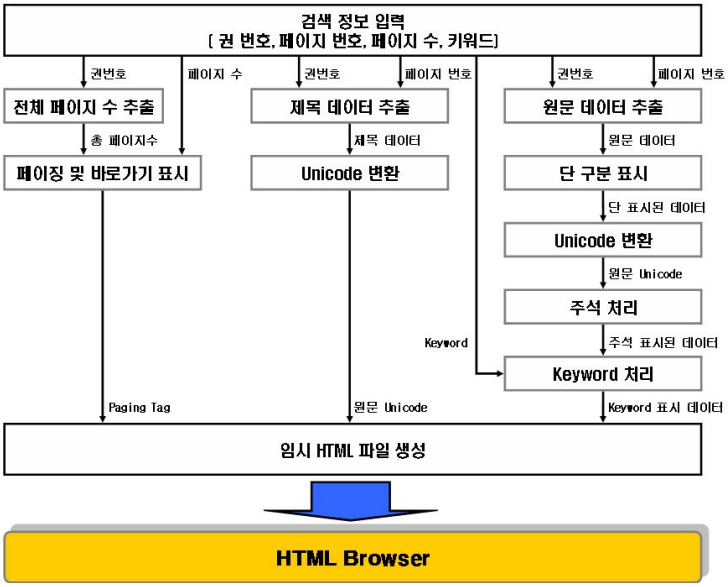


그림 27. HTML 생성기의 동작 흐름

다음은 HTML 생성기의 Visual Basic 코드 중 중요한 몇몇 부분을 보여주고 있다.

```
Function MakeHtml(ByVal iBookNum As Integer, ByVal iPageNum As Integer,
                                ByVal icount As Integer, ByVal strKey As String) As String
    Dim strHtml, temp, strSub, strSubTitle As String
    Dim i, j, iDanNum, idantemp As Integer
    Dim totPage As Integer '총 페이지 수
    Dim m, n, x As Integer '페이징 사용 변수
            ......
'********************임시 HTML 파일 생성 시작************************************
    strHtml = "<html><link rel=stylesheet href=../HTML/default.css type=text/css>"
    strHtml = strHtml & "<body bgcolor='#FFFFFF' text='#333' leftmargin='8'
              topmargin='0' marginwidth='0' "
            ......
'********************총 페이지 추출*****************************************************
    AdoData.RecordSource = "select MAX([npagenum]) as totpage from edocdata" &
                                    iBookNum
    AdoData.Refresh
    totPage = CInt(AdoData.Recordset.Fields("totpage"))
            ......
'********************페이징 표시 및 바로 가기 처리 부분 ************************
    '첫 패이지로 이동
```

```
    strHtml = strHtml & "<a href=?" & CStr(iBookNum) & "?1?" & CStr(nPageNum) &
              ">[<<]</a>"
    m = Int(iPageNum / 10)
    n = iPageNum Mod 10
    If (n = 0) Then
        m = m - 1
    End If

    '10단위 페이징
    For x = 1 To 10
        If ((m * 10 + x) <= totPage) Then
        If iPageNum <= (m * 10 + x) And (m * 10 + x) <= (iPageNum + icount - 1)
        Then
            strHtml = strHtml & "<font size=2><a href=?" & CStr(iBookNum) & "?" & m
                      * 10 + x & "?" & CStr(nPageNum) & ">[" & m * 10 + x &
                      "]</a></font>"
        Else
            strHtml = strHtml & "<a href=?" & CStr(iBookNum) & "?" & m * 10 + x &
                      "?" & CStr(nPageNum) & ">[" & m * 10 + x & "]</a>"
        End If
        End If
    Next x

    '마지막 페이지로 이동
    strHtml = strHtml & "<a href=?" & CStr(iBookNum) & "?" & totPage & "?" &
              CStr(nPageNum) & ">[>>]</a>"
            .....
'*******************원문 데이터 와 제목 데이터 추출*****************************
    For j = 0 To icount - 1
        AdoData.RecordSource = "select * from edocdata" & iBookNum & " where
                               npagenum = " & CStr(iPageNum) & " order by
                               nlinenum"
        AdoData.Refresh
        adoData2.RecordSource = "select * from tag_jmok_table where nbooknum = " &
                                iBookNum & " and " & CStr(iPageNum + j) & " >=
                                npagenum and " & CStr(iPageNum + j) & " <=
                                endpage"
        adoData2.Refresh
        strSubTitle = adoData2.Recordset.Fields("jmok")
        strSubTitle = ToUnicode(CStr(strSubTitle))
        adoData2.Recordset.Close
           .....
        For i = 1 To AdoData.Recordset.RecordCount
```

```
'****************단 구분 표시 ****************************************************
            idantemp = AdoData.Recordset.Fields("ndannum")
            If idantemp <> iDanNum Then
                iDanNum = idantemp
            End If
            temp = Trim(CStr(AdoData.Recordset.Fields("sdocdata")))
            strSub = ToUnicode(CStr(temp))
'***************  주석 처리 ****************************************************
            strSub = Replace(strSub, "<COMMENT>", "<NOTE>[주석]<img src='../image/
                     dot_gray1.gif'>
            strSub = Replace(strSub, "</COMMENT>", "</NOTE>")
            strSub = Replace(strSub, "<com>", "<com><font style='font-family: 새굴림,
                     Verdana; font-size: 15pt; color: red'>")
            strSub = Replace(strSub, "</com>", "</font></com>")
            AdoData.Recordset.MoveNext
        Next i
        AdoData.Recordset.Close
    Next j
    strHtml = strHtml + "</body></html>"
'****************Keyword 처리*************************************************
    If strKey <> "" Then
       strHtml = Replace(strHtml, strKey, "<font color = red>" & strKey & "</font>")
    End If
'******************임시 HTML 파일 저장 및 화면 출력*****************************
    Open App.Path & "/HTML/Pagehtml.htm" For Output As #1
    Print #1, strHtml
    Close #1
    webMain.Navigate App.Path & "/HTML/Pagehtml.htm"
End Function
```

5. 웹 검색 인터페이스의 구현

한국불교전서 웹 검색 시스템의 검색 결과 화면은 경전과 동일한 구조인 제목, 원문, 주석 등으로 구성하였고, 원문과 동일한 형태로 들여쓰기 기능을 제공함으로서 사용자에게 학술적인 참고 자료로서 가치가 있도록 하였다. 또한 사용자 입장에서 쉽고 편리한 방법으로 검색 할 수 있도록 키워드 검색, 페이지 검색, 제목 검색, 그리고 획수 검색에 이르는 다양한 검색 서비스를 제공하였다. 사용자가 보다 편리하게 사용할 수 있도록 개선한 기능을

살펴보면 다음과 같다.

- 키워드 검색 결과 한 화면에 10개씩 보여주기
- 여러 키워드 검색
- 전권 대상으로 한 키워드 검색 속도 개선
- 한자의 음이 동일한 한글 키워드 검색 오류 수정

5.1 웹 검색시스템의 주요 기능

웹 검색시스템의 주요 기능에 해당하는 인터페이스는 [그림 22]에 나타나있다. 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.

5.2 개선한 기능

사용자에게 유용하고 편리한 기능을 제공하고 완성도 높은 웹 검색시스템을 위하여 일부 기능을 개선하였다. 그 기능을 살펴보면 크게 4가지로 첫째, 키워드 검색결과 한 화면에 10개씩 보여주는 기능, 둘째, 여러 키워드 검색가능, 셋째, 전권 대상으로 한 키워드 검색 속도 개선, 넷째, 키워드 검색에서 한자의 음이 동일한 한글 키워드 검색 문제 해결 등이 있다. 각 기능에 해당하는 상세한 내용은 다음과 같다.

5.2.1. 키워드 검색 결과 한 화면에 10개씩 보여주는 기능

이전 사업에서는 전체 키워드 검색 또는 개별 키워드를 검색하면, 키워드 검색 결과 목록이 한 화면에 모두 나타난다. 이것의 문제점은 검색한 후 결과 목록이 많으면, 사용자는 우측에 있는 이동박스를 움직여서 원하는 항목을 찾

는 번거로움이 있다. 그래서 본 사업에서는 한 화면에 10페이지씩 10개의 목록을 보여줌으로써 기존의 문제점을 해결하였다. [그림 28]은 개선한 키워드 검색 인터페이스 화면으로 '보살'이라는 키워드를 검색하였을 때 한 화면에 10페이지씩 10개의 목록으로 표현한 것이다.

그림 28. 한 화면에 10개씩 보여주기

5.2.2. 여러 키워드 검색 기능

여러 개의 키워드를 검색하는 기능이다. 예를 들어, 용어 입력란에 '사리 보살 불교 검색 대반야'를 포함한 공백으로 구분되는 용어를 검색 하였을 때 검색 결과는 각 키워드를 포함하고 있는 내용으로 나타나고 [그림 29]는 그 결과 화면을 나타낸 것이다.

그림 29. 여러 키워드 검색

5.2.3. 전권 대상으로 한 키워드 검색 속도 개선

전체 권을 대상으로 키워드 검색을 하였을 때, 개별 권을 대상으로 키워드 검색을 한 것보다 검색 속도가 현저히 저하되고, 또한 전체 경에 포함된 용어의 개수가 많을 때 검색시간 초과 오류가 발생한다. 전체 권을 대상으로 하였을 때 나타나는 문제점을 다음과 같이 개선하였다. 먼저, 전체 권을 대상으로 하는 통합 용어 테이블에 인덱스를 생성하였다. 다음, 검색한 용어를 테이블에서 모두 가져오는 것이 아닌, 한 화면에 필요한 10페이지씩 10개 목록만큼만 메모리로 가져와서 화면에 나타내는 방식으로 검색 속도를 개선하였다. 좀 더 구체적으로 설명하면, 지금까지 번역된 총 12권 중에서 가장 많은 용어의 개수는 '사리' 와 '보살'이다. 이 키워드를 검색 하였을 때 키워드의 개수는 총 11,835 개이고 평균 검색시간은 1초미만으로 나타났다.

5.2.4. 한자의 음이 동일한 한글 키워드 검색 오류 수정

기존 키워드 검색에서 한자의 음이 동일한 한글 키워드 검색에 문제 가 있다. 예를 들어, '사리' 음을 가진 한자 검색 하였을 때, '사리'에 해당하는 한자는 6개로 '闍利 事理 舍利 師利 師利 捨離' 이다. 검색된 결과에서 사용자가 찾고자 하는 키워드 한자(한글)를 선택한 후 '검색시작' 버튼을 누르면, 한자키워드는 항상 마지막에 위치한 용어에 해당하는 본문내용을 우측 화면에 나타내 준다. 좀 더 구체적으로 설명을 하면 〔그림 30〕에서 '사리'라는

그림 30. '사리' 검색 결과

키워드를 검색하면 좌측화면에 한글 '사리' 음을 가진 한자 6개가 나타난다. 검색된 6개중 하나를 선택한다고 해도 언제나 제일 마지막 키워드 '捨離 (사리)' 만 우측화면에 본문 내용으로 나타난다. 그래서 용어 검색 후 한자 키워드의 목록이 여러 개일 경우 사용자가 원하는 키워드를 선택할 수 있고, 해당하는 본문 내용을 우측 화면에서 확인 할 수 있다.

5.3 웹 검색 인터페이스의 구현

본 검색시스템의 전체 구성도가 〔그림 31〕에 나타나있다. 사용자가 메인

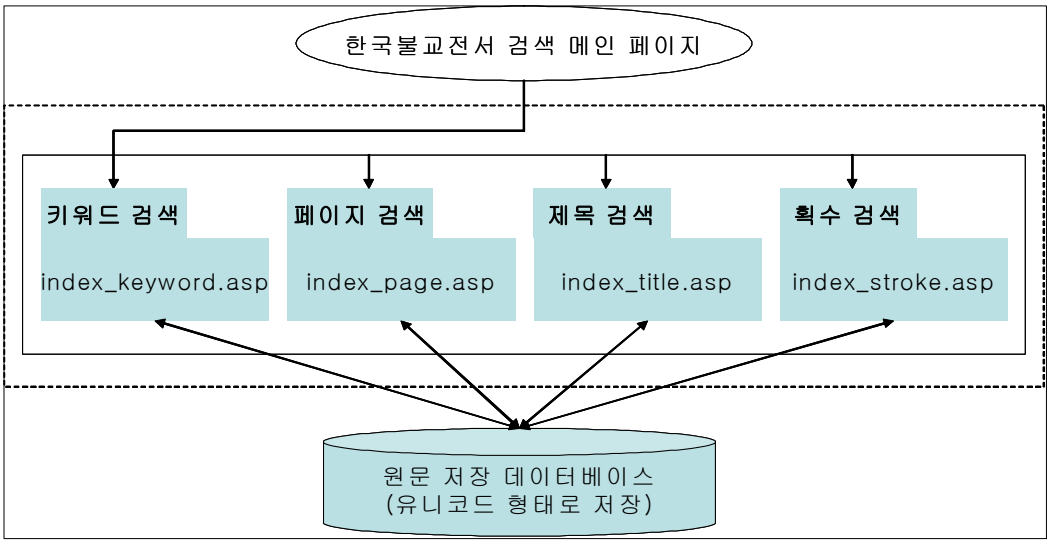


그림 31. 웹 검색 인터페이스의 전체 구성도

화면에서 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.

또한 데이터베이스에 저장할 때 경전의 내용은 유니코드 형태로 변환해서 저장한다. 그러므로 유니코드와 텍스트간의 변환 기능이 필요하고, 〔그림 32〕은 유니코드와 일반 텍스트 간 변환을 담당하는 함수와 변환 과정을 나타

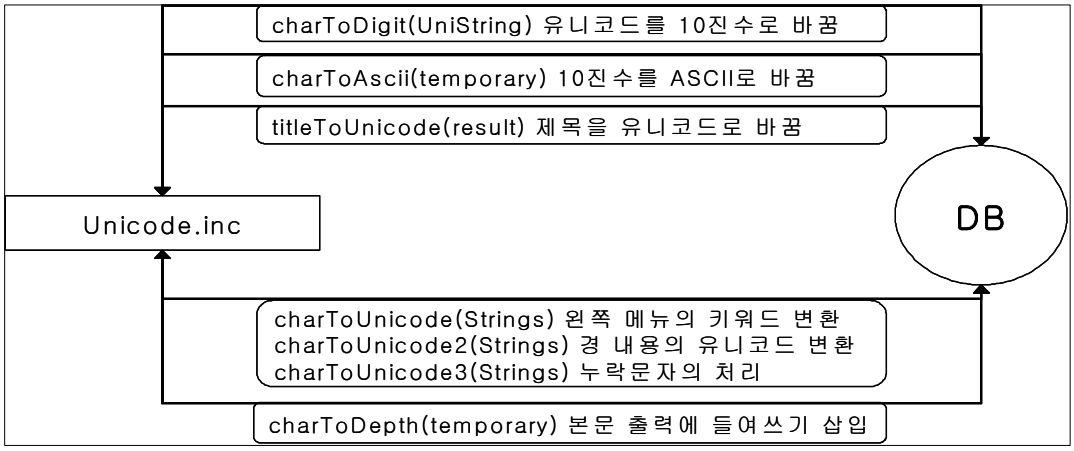


그림 32. 유니코드와 텍스트 간 변환 함수

낸 그림이다.

6. 결론 및 향후 연구 방향

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 제13책과 제14책을 전산화 하여 이를 연구하는 연구자들이나 열람을 원하는 사람들에게 도움이 될 뿐만 아니라 우리의 귀중한 문화유산을 전 세계에 널리 알릴 수 있다.

본 연구를 수행하는데 필요한 기술로 가장 먼저 한자를 컴퓨터에 입력할 수 있는 입력 방법 및 유니코드 상에 없는 고문헌 상의 문자를 처리할 수 있는 시스템의 개발이 필요하다. 이러한 기술을 개발하기 위해서는 유니코

드 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 제13책과 제14책의 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 제13책과, 제14책에 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 Databa ses,"
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 전자불전 연구소 제2회 세미나, 2000.
〔27〕 한태식, "불교학 연구에 있어서 한국불교전서의 위상," '00 동국대학교 전자불 전연구소 제2회 세미나, 2000.
〔28〕 허인섭, "Report on the Digital Tripitaka Koreana 2001," '01 동국대학 교 개교 95주년 기념 세계전자불전학회 학술대회, 2001.
〔29〕 현득창, 임광택, 이수연, "SGML 기본 파서를 이용한 SGML문서 편집기의 구 현," 한국정보과학회, 정보과학회 눈문지, Vol 25, No. 1, 1998.
〔30〕 홍영식, "한국 불교전서 데이터베이스에서 누락문자 검색," '01 동국대학교 개 교 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.kr/~jwjung/seminar/hangul-i18n/ko-code.html.
〔44〕 종림스님, "팔만대장경 전산화 추진경과와 이후 계획," 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