한글대장경 전산화 3차 사업의 현황

노진홍*, 구현우*, 유응구*, 박성은*, 박영희**,
이용규*, 이금석*, 홍영식*, 한보광**

*동국대학교 컴퓨터공학과
**동국대학교 선학과

요 약

본 연구는 한글대장경 전산화 3차 사업으로 한글대장경 30권 분량을 전산화하여 검색 시스템을 구축하는데 목적이 있다. 고려대장경의 우리말 번역본인 한글대장경을 개역하고 전산화함으로써 고문헌을 입력하여 데이터베이스에 저장하고 인터넷을 통해 검색할 수 있다. 한글대장경 고문헌은 확장한자와 누락문자 및 특수문자 등을 포함하고 있으므로 효과적인 입력과 저장을 위해 본 연구에서는 유니코드(Unicode)를 사용하며 유니코드로 표현하지 못하는 문자는 이미지 폰트로 만들었다. 그리고 고문헌의 문서구조 표현과 효율적인 검색을 위해 XML을 적용하여 웹상에서 실제 고문서와 같은 내용을 검색할 수 있다. 또한 효율적이고 편리한 검색 방법을 제공하는 검색엔진을 개발하여 유니코드로 저장된 고문헌은 인터넷을 통해 전세계에서 접근할 수 있다.

본 연구에서 구현된 검색 시스템은 윈도 2000 서버에서 운영되는 마이크로소프트사의 SQL Server와 IIS(Internet Information Server)를 사용하였다.

I. 서론

불법이 인도에서 전래되어 인류의 정신문화를 이끌어 온지도 어언 3000여 년이 지났다. 불교의 가르침은 보통 사람들이 구사하는 언어를 통해 전해져왔는데, 초기에는 부처님으로부터 신성한 가르침을 직접 듣는 것이 가능하였고, 입에서 입으로 구전되어 왔다.

부처님의 입멸 후 그러한 가르침의 전통은 인도에서 결집(結集)을 통해 문자화되어 보다 많은 인류를 깨달음의 길로 이끄는 지침이 되었다. 부처님의 가르침은 동아시아의 거의 모든 국가에 전해졌고, 그 국민들에게 안심낙도의 삶을 제시하였다.

이후 각 나라의 전법승들은 부처님의 가르침을 그 나라의 언어로 전하여 널리 일체중생을 이롭게 하는 역경사업에 진력하였다. 이는 국가의 지원을 받는 경우도 있었고, 전법승만의 불타는 신념에 의한 개인적인 사업인 경우도 있었다. 마침내 전법의 발길이 닿은 국가들에서는 불전을 자국어로 번역하여 편찬・유포하게 되었다.

우리나라에 불교를 전해준 중국에서는 한문(漢文)불경이 편찬・유포된 것이다. 우리나라에서는 중국의 불경을 전해 받아 국가와 국민의 정신적 지주로 삼아왔다. 이는 역사에서도 확인되는 바이다. 세계문화유산으로 등록된 고려대장경은 몽고의 침입으로 국가가 위기에 처했을 시기에 부처님의 가르침으로 국가의 안녕과 백성의 평안을 기원하기 위해 전 국가적으로 역량을 결집한 우리의 문화유산인 것이다.

조선시대에 이르러서는 불경을 민간에 널리 유포시키기 위하여 한글로 된 불경이 제작되기 시작하였다. 이는 지식인만의 불교에서 일체중생을 위한 불교로의 전환을 의미하게 된다. 조선 말기에서부터 가속화된 불경의 한글화는 일제의 강점기에 민족의 정신을 일깨우는 작업으로 진행되어 오늘에 이르게 되었다.

동국대학교의 역경원 설립과 함께 본격화되기 시작한 한글대장경

사업은 현대문명의 발달에 발맞추어 새롭게 전산화의 길을 모색하고 있다. 이는 한글대장경을 디지털화하여 인터넷을 통해 전 세계의 인류에게 제공함으로써 시간과 장소를 초월하여 불법의 진리를 홍보하는 것이며, 또한 우리나라의 뛰어난 정신문화를 전 세계에 알리는 새로운 전법활동이라고 할 수 있다.

II. 한글대장경 전산화 3차 사업

본 연구에서는 한글대장경을 전산화하여 전 세계에서 손쉽게 검색할 수 있도록 하는데 그 목적이 있다. 이러한 연구 목적을 달성하기 위해 크게 3가지 기술이 필요하다. 한글대장경을 컴퓨터에 입력하고 이를 편집하여, 데이터베이스에 저장 기술, 데이터베이스에 저장된 내용들을 웹에서 검색할 수 있는 인터페이스와 검색 기술, 누락문자, 진언 그리고 도표의 처리 기술이 필요하다. 이들을 위해 본 연구에서 개발한 기술 내용 및 3차 사업에서 시행되어 발전된 기술 내용에 관해 먼저 한글대장경의 입력 및 교정 방법에 대해 기술하고, 2차 사업보다 개선된 데이터베이스 저장 방법을 살펴본 후 검색 인터페이스의 개선사항과 누락문자, 진언 그리고 도표의 처리 기술에 대해 설명한다.

1. 한글대장경의 입력 및 교정

한글대장경의 입력은 (주)동국전산에 외주를 주어 입력하고, 3차에 걸쳐서 엄밀한 교정 작업을 수행하였다.

2003년도 한글대장경 전산화 제3차 사업에서 입력교정한 대장경의 목록은 총 68경, 791권으로 다음과 같다. (※ K번호는 고려대장경의 경전고유번호임.)

K.1 대반야바라밀다경 (301~600권)
K.5 마하반야초경 5권
K.6 도행반야경 10권
K.7 소품반야바라밀경 10권
K.26 무량수경 2권
K.57 대승대집지장십륜경 10권
K.60 보살염불삼매경 5권
K.61 허공잉보살경 2권
K.62 허공장보살경 1권
K.64 관허공장보살경 1권
K.65 대방등대집경보살염불삼매분 10권
K.66 대방등대집경현호분 5권
K.70 아차말보살경 7권
K.72 대애경 8권
K.73 대집비유왕경 2권
K.74 보녀소문경 4권
K.75 자재왕보살경 2권
K.76 분신왕문경 2권
K.77 무언동자경 2권
K.78 보성다라니경 10권
K.117 정법화경 10권
K.191 관무량수불경 1권
K.192 아미타경 1권
K.398 승가타경 4권
K.426 대불정여래밀인수증요의제보살만행수릉엄경 10권
K.427 대비로자나성불신변가지경 7권
K.429 금강정유가중략출염송경 4권
K.570 유가사지론 (24~48권)
K.571 현양성교론 20권

K.573 현양성교론송 1권
K.654 반니원경 2권
K.656 시가라월육방예경 1권
K.718 선생자경 1권
K.747 칠불부모성자경 1권
K.896 사분율 60권
K.945 아비달마법온족론 12권
K.946 아비달마집이문족론 20권
k.956 아비달마순정리론 (1~8권)
K.985 승가나찰소집경 3권
K.1050 경율이상 (1~50권)
K.1080 홍명집 14권
K.1119 법집요송경 4권
K.1177 비바시불경 2권
K.1179 대삼마야경 1권
K.1182 칠불경 1권
K.1247 인선경 1권
K.1248 신불공덕경 1권
K.1252 제석소문경 1권
K.1268 금강정경유가수습비로자나삼마지법 1권
K.1274 금강정일체여래진실섭대승현증대교왕경 3권
K.1289 금강정경유가십팔회지귀 1권
K.1290 보리장소설일자정륜왕경 5권
K.1335 금강정유가호마의궤 1권
K.1336 도부다라니목 1권
K.1339 대집대허공장보살소문경 8권
K.1372 금륜왕불정요략염송법 1권
K.1412 대집회정법경 5권
K.1418 일체여래금강삼업최상비밀대교왕경 7권

K.1429 대집법문경 2권
K.1453 대견고바라문연기경 2권
K.1463 니구타범지경 2권
K.1464 백의금당이바라문연기경 3권
K.1466 현증삼매대교왕경 30권
K.1481 해의보살소문정인법문경 18권
K.1499 종경록 (1~24권)
K.1502 법계도기총수록 2권
K.1503 조당집 (1~10권)
석화엄교분기원통초 (1~2권)

입력과 교정을 마친 한글대장경은 전자불전연구소에서 페이지・대제목・소제목・각주・색인어에 대하여 각각 태그(tag) 작업을 수행하였다.

1) 페이지 태그작업 : 페이지를 검색하여 해당 원문을 보여준다.
2) 제목 태그작업 : 경전의 대제목과 소제목을 검색할 수 있으며, 경전의 제목을 통하여 해당 원문을 확인할 수 있게 한다.
3) 각주 태그작업 : 한글대장경의 각주에 나타나 있는 원문을 확인할 수 있도록 한다.

2. 데이터베이스 저장

한글대장경의 원문은 제목, 원문 내용, 주석 등으로 구성되어있고, 각각의 해당되는 내용은 <JMOK>, <PAGE>, <COMMENT>와 같은 태그들로 구별하기 때문에 이를 이용하여 데이터베이스를 구축할 수가 있다. 이때 원문에 나타나는 한문은 기존 문자 셋으로 표현하는데 한계가 있어서 유니코드로 변환하여 저장한다.

먼저, 원문을 데이터베이스에 저장하기 위해서는 각 부분을 구별해

주는 태그들이 유효한지 검증하는 작업이 필요하며, 이를 위해 원문을 XML 파일로 저장하여 태그의 유효성을 검증한다. 유효성 검증 작업을 마친 후에는 원문으로부터 제목, 원문 내용, 키워드를 추출하여 유니코드로 변환한 후 그 값을 각 해당 테이블에 저장한다. 이러한 한글대장경의 데이터베이스 구축 단계를 간략히 정리하면 다음과 같으며 자세한 내용은 본문을 통해 설명한다.

- 1단계 : 태그가 삽입된 원문의 유효성 검증 작업
- 2단계 : 파일 처리
- 3단계 : 인덱스 추출
- 4단계 : 키워드 및 원문 저장

2.1 태그가 삽입된 원문의 유효성 검증 작업

텍스트 파일로 변환된 원문에는 페이지, 제목, 주석 등을 구별하기 위하여 각각 <JMOK>, <PAGE>, <COMMENT>라는 태그들을 삽입한다. 이러한 태그들은 여는 태그(<...>)와 닫는 태그(</...>)가 쌍으로 구성되어야하며, 만일 그렇지 않으면 잘못된 데이터가 데이터베이스에 입력될 수 있으므로 반드시 확인 작업을 거쳐야 한다. 이러한 태그들의 검증 작업은 다음과 같은 순서로 이루어진다.

① “*.txt”로 저장된 원문 파일들을 “*.xml”로 확장명을 바꾼다.
② 웹 브라우저에서 해당 XML 문서를 불러들인다.
③ 웹 브라우저에 에러 메시지가 나타나지 않으면 유효한 문서이고, 그렇지 않으면 오류가 발생한 부분을 찾아 원문 내용을 수정해야한다.
④ 모든 태그들은 여는 태그와 닫는 태그가 쌍으로 이루어져야만 유효한 문서를 생성할 수 있다.
⑤ 최종적으로 이렇게 생성된 유효한 문서를 데이터베이스 구축에 사용한다.

2.2 원문 저장 프로그램의 구현

[그림 1]은 데이터베이스를 구축하기 위한 저장 프로그램의 구현 화면이다. 즉, 왼쪽 중간에 있는 화면의 텍스트 필드에 각 경 이름을 넣어 '저장' 버튼을 누르면 해당 원문이 저장된다. 또한 저장된 내용을 검색하거나 수정할 수 있는 기능을 제공하는데, 검색하고자 하는 경 이름과 페이지, 단 정보를 넣은 후에 각각 '검색'과 '수정' 버튼을 누르면 해당 내용이 검색된다. 그러면 오른쪽 화면에 그 결과를 보여준다.

[그림 1] 원문 저장 프로그램의 화면

2.3 원문 파일에서 인덱스 구축

키워드 테이블을 사용해서 원문을 검색하며, 순차검색 방법을 사용하여 검색한다.

(1) 키워드 인덱스 구축

키워드 인덱스를 구축하기 위해서는 키워드 파일에서 순차적으로 키워드를 읽고 각각의 키워드를 원문과 비교하는 방법을 적용했으며, 자세한 과정은 다음과 같다.

① “edocdata”와 “keyword” 테이블을 사용한다.
② 각각의 유니코드로 저장된 테이블의 처음 시작 코드를 읽는다.
③ 키워드 테이블에서 추출한 하나의 키워드와 “edocdata” 테이블의 전체 내용과 비교한다. “keyword” 테이블에서 선택한 키워드와 같은 값을 발견하면, 일련번호, 키워드번호, 키워드, 페이지 정보를 저장한다. 그 다음 원문 파일을 비교한다.
④ “eodcdata” 테이블 수만큼 비교한 다음에 다음 키워드를 읽어 앞의 과정을 반복 수행한다. “keyword” 테이블 전체를 비교했을 때, 프로그램을 종료한다.

(2) 제목 인덱스 구축

원문을 검색할 때 <JMOK>과 </JMOK> 키워드를 검색하여 해당되는 부분이 발견되면 제목을 “tag_jmok_table” 테이블에 저장하고, 제목이 위치한 위치 정보를 각각 “book”, “page”, “line”에 저장한다.

<JMOK1> ... </JMOK1>
<JMOK2> ... </JMOK2>
<JMOK3> ... </JMOK3>
<JMOK4 SEARCH=‘TRUE’> ... </JMOK4>

위에서 보는바와 같이 제목 태그는 트리 구조의 형태를 갖는다. <JMOK 다음에 나타나는 숫자인 1, 2, 3, 4는 제목의 레벨을 나타내고, <JMOK4> 태그의 속성인 “SERCH=‘TRUE’”는 경제목을 의미하

며 그 값은 “tag_kyung_table”에 저장한다. “tag_page_table”에는 “tag_jmok_table”에 저장된 일련번호를 가지고 페이지 번호만 따로 추출하여 저장한다.

2.4 키워드 및 원문 저장

키워드 저장은 키워드가 저장된 텍스트 파일로부터 키워드를 추출하여 키워드 테이블을 구축하는 방법을 이용한다. 키워드 파일에는 한글과 한자 키워드가 저장되어있으며, 실제 “keyword” 테이블에는 한글과 한자 키워드에 대한 유니코드 값이 저장된다.

원문 저장은 유니코드 편집기에서 작성된 유니코드 원문을 그대로 테이블에 저장한다. 원문 파일을 라인별로 읽어 저장하면서 페이지 태그를 검사하여 각 페이지 해당 라인(line)수와 “ncontinue” 등의 부가 정보를 생성한다. 원문을 저장할 때 원문에서 한 라인이 레코드의 저장크기를 초과할 경우에 100자 단위로 나눠서 저장하고, “ncontinue”에 이에 대한 부가 정보를 저장한다.

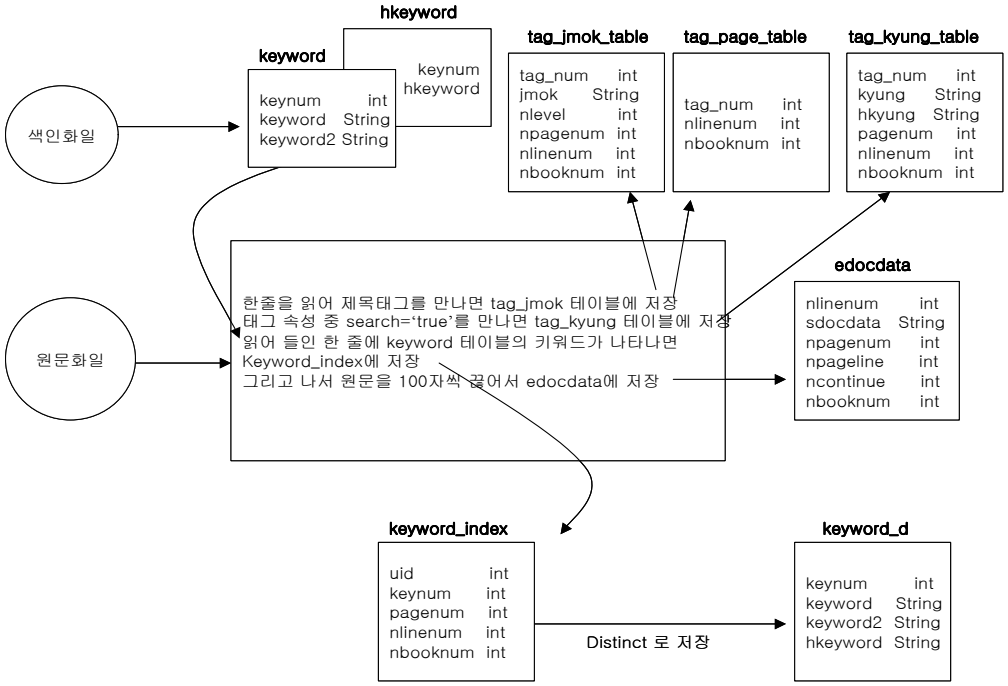
2.5 테이블 생성 방법 및 구조 설명

데이터베이스 구축을 위해서 마이크로소프트 SQL Server 7.0을 사용하며, 본 절에서는 테이블 생성 방법과 주요 테이블 구조에 대한 설명을 한다.

(1) 테이블 생성 방법

[그림 2]는 색인 파일과 원문 파일 이용하여 각각의 테이블을 생성하는 방법을 나타낸 그림이다. 먼저 색인 파일을 읽어서 “keyword”와 “hkeyword” 테이블을 만든다. 그런 다음 원문에서 한 줄을 읽어 제목 태그가 나타나면 각 해당 정보를 “tag_jmok_table”,

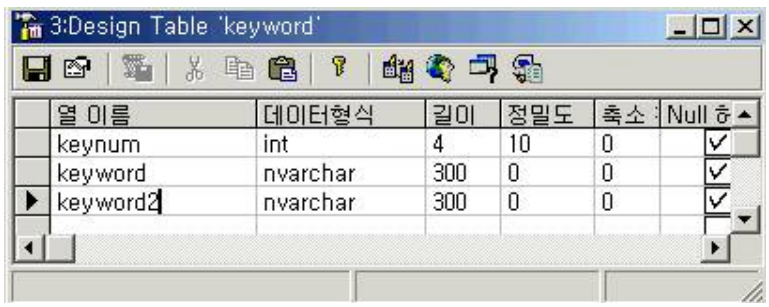
“tag_page_table”, “tag_kyung_table” 테이블에 저장한다. 그리고 원문을 읽어가면서 해당 키워드가 존재하면 그 키워드는 “keyword_index” 테이블에 저장한다.



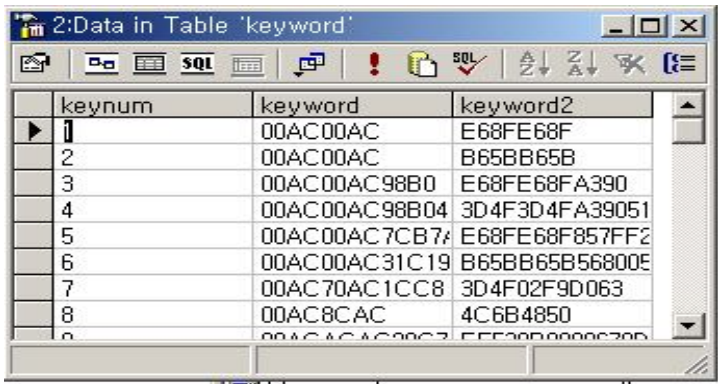
[그림 2] 테이블 생성 방법

(2) 테이블 구조

■ keyword 테이블



[그림 3] keyword 테이블의 레코드 형식



[그림 4] keyword 테이블의 내용

① 테이블 명 : keyword

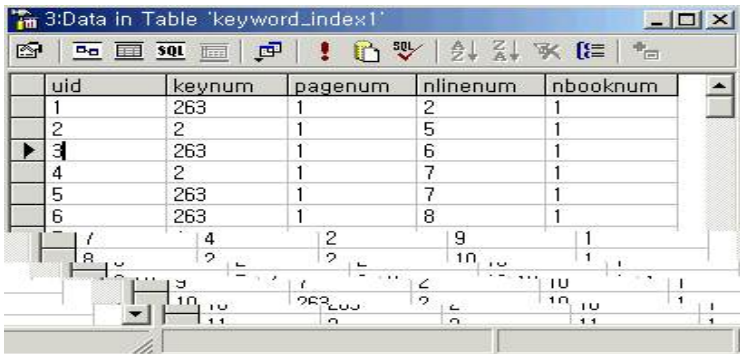
② 테이블의 역할 : 키워드 사전으로부터 입력받은 키워드를 저장한다.

③ 필드의 역할

- keynum : 각 키워드에 대한 유일키를 저장한다.
- keyword : 한글 키워드에 대한 유니코드 값을 저장한다.
- keyword2 : 한자 키워드에 대한 유니코드 값을 저장한다.

■ keyword_index 테이블

[그림 5] keyword_index 테이블의 레코드 형식



uid	keynum	pagenum	nlinenum	nbooknum
1	263	1	2	1
2	2	1	5	1
3	263	1	6	1
4	2	1	7	1
5	263	1	7	1
6	263	1	8	1

[그림 6] keyword_index 테이블의 내용

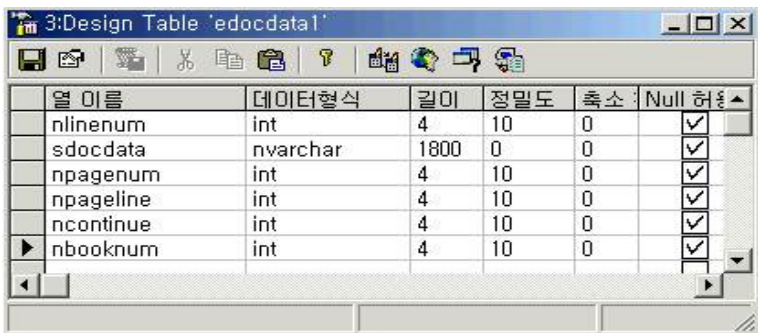
① 테이블 명 : keyword_index

② 테이블의 역할 : 각 권별로 키워드 인덱스 테이블을 유지한다. 키워드가 발견된 원문의 권, 페이지, 라인 정보를 저장한다.

③ 필드의 역할

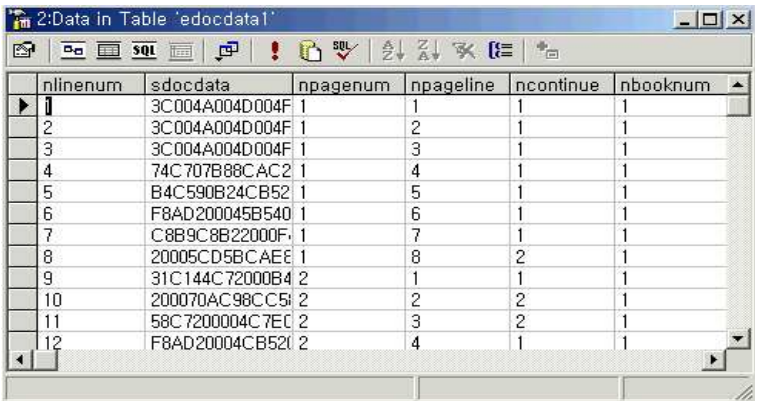
- uid : keyword_index 테이블의 유일키를 저장한다.
- keynum : “keyword” 테이블의 “keynum”과 일치한 값을 저장한다.
- pagenum : 키워드가 발견된 곳의 페이지 번호를 저장한다.
- linenum : 키워드가 발견된 곳의 라인 번호를 저장한다.
- nbooknum : 현재의 권 번호를 저장한다.

■ edocdata 테이블



열 이름	데이터형식	길이	정밀도	축소	Null 허용
nlinenum	int	4	10	0	☑
sdocdata	nvarchar	1800	0	0	☑
npagenum	int	4	10	0	☑
npageline	int	4	10	0	☑
ncontinue	int	4	10	0	☑
nbooknum	int	4	10	0	☑

[그림 7] edocdata 테이블의 레코드 형식



nlinenum	sdocdata	npagenum	npageline	ncontinue	nbooknum
1	3C004A004D004F	1	1	1	1
2	3C004A004D004F	1	2	1	1
3	3C004A004D004F	1	3	1	1
4	74C707B88CAC2	1	4	1	1
5	B4C590B24CB52	1	5	1	1
6	F8AD200045B540	1	6	1	1
7	C8B9C8B22000F	1	7	1	1
8	20005CD5BCAEE	1	8	2	1
9	31C144C72000B4	2	1	1	1
10	200070AC98CC5	2	2	2	1
11	58C7200004C7E	2	3	2	1
12	F8AD20004CB52	2	4	1	1

[그림 8] edocdata 테이블의 내용

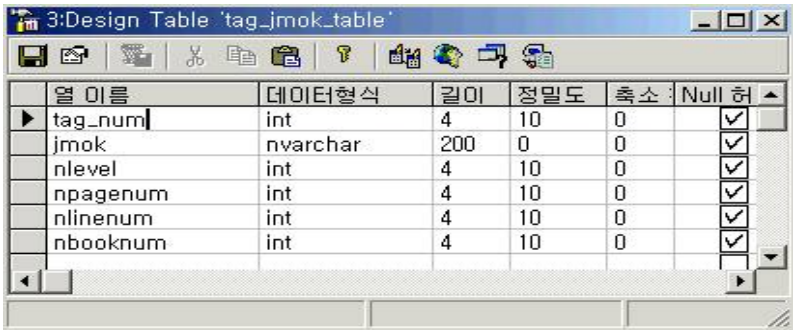
① 테이블 명 : edocdata

② 테이블의 역할 : 각 권별로 원문을 저장한다.

③ 필드의 역할

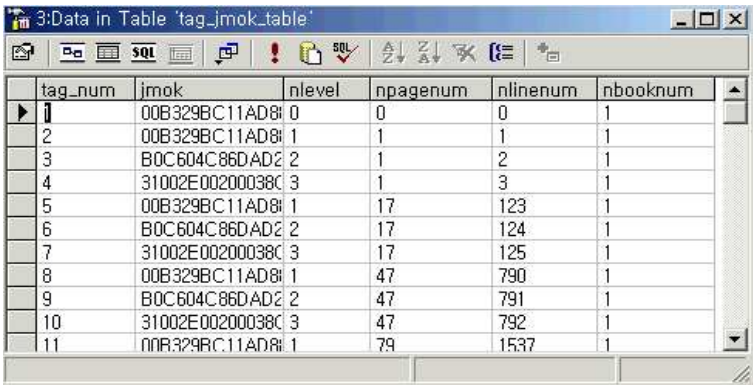
- nlinenum : 원문에 대한 유일키를 저장한다.
- sdocdata : 원문을 유니코드 형태로 저장한다.
- npagenum : 페이지 번호를 저장한다.
- npageline : 페이지 라인을 저장한다.
- ncontinue : 한 라인이 1800자를 초과할 경우 값을 증가한다.
- nbooknum : 현재 권 번호를 저장한다.

■ tag_jmok_table 테이블



열 이름	데이터형식	길이	정밀도	축소	Null 허
tag_num	int	4	10	0	☑
jmok	nvarchar	200	0	0	☑
nlevel	int	4	10	0	☑
npagenum	int	4	10	0	☑
nlinenum	int	4	10	0	☑
nbooknum	int	4	10	0	☑

[그림 9] tag_jmok_table 테이블의 레코드 형식



tag_num	jmok	nlevel	npagenum	nlinenum	nbooknum
1	00B329BC11AD8	0	0	0	1
2	00B329BC11AD8	1	1	1	1
3	B0C604C86DAD2	2	1	2	1
4	31002E00200038(	3	1	3	1
5	00B329BC11AD8	1	17	123	1
6	B0C604C86DAD2	2	17	124	1
7	31002E00200038(	3	17	125	1
8	00B329BC11AD8	1	47	790	1
9	B0C604C86DAD2	2	47	791	1
10	31002E00200038(	3	47	792	1
11	00B329BC11AD8	1	79	1537	1

[그림 10] tag_jmok_table 테이블의 내용

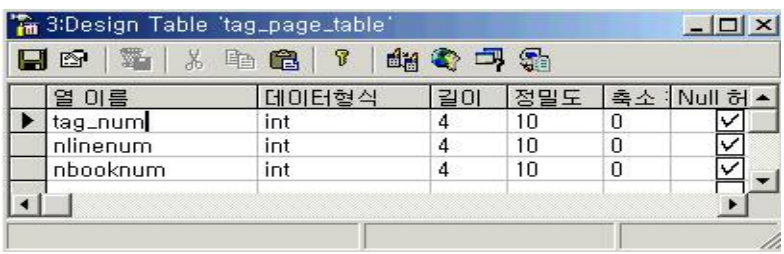
① 테이블 명 : tag_jmok_table

② 테이블의 역할 : 원문에서 제목과 제목이 나타나는 곳의 위치 정보를 저장한다.

③ 필드의 역할

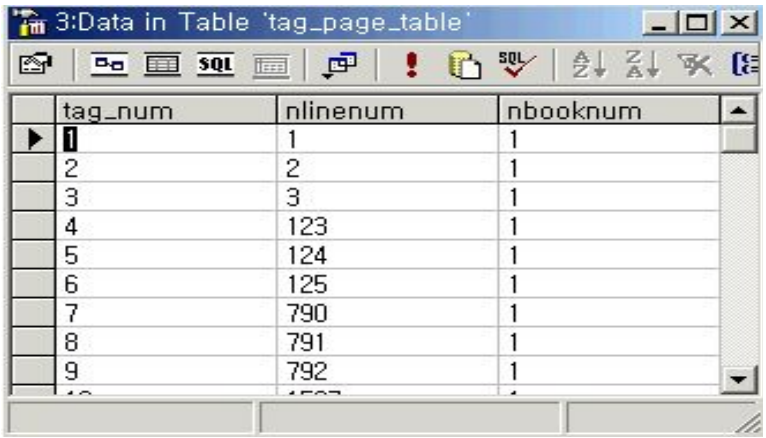
- tag_num : 각 제목에 대한 유일키를 저장한다.
- jmok : <JMOK> 태그가 나타난 곳의 제목을 유니코드 형태로 저장한다.
- nlevel : 제목의 레벨을 저장한다.
- npagenum : 제목 태그의 페이지 번호를 저장한다.
- nlinenum : 라인 번호를 저장한다.
- nbooknum : 현재 권 번호를 저장한다.

■ tag_page_table 테이블



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

[그림 11] tag_page_table 테이블의 레코드 형식



tag_num	nlinenum	nbooknum
1	1	1
2	2	1
3	3	1
4	123	1
5	124	1
6	125	1
7	790	1
8	791	1
9	792	1

[그림 12] tag_page_table 테이블의 내용

① 테이블 명 : tag_page_table

② 테이블의 역할 : 제목 테이블에 대한 요약 정보를 가지고 있다.

③ 필드의 역할

- tag_num : 각 제목에 대한 유일키를 저장한다.
- nlinenum : 제목의 위치 정보인 “edocdata” 테이블의 “nlinenum”의 값을 저장한다.
- nbooknum : 현재 권 번호를 저장한다.

■ tag_kyung_table 테이블

열 이름	데이터형식	길이	정밀도	축소	Null 허
tag_num	int	4	10	0	☑
kyung	nvarchar	200	0	0	☑
hkyung	nvarchar	200	0	0	☑
npagenum	int	4	10	0	☑
nlinenum	int	4	10	0	☑
nbooknum	int	4	10	0	☑

[그림 13] tag_kyung_table 테이블의 레코드 형식


3:Data in Table 'tag_kyung_table'

tag_num	kyung	hkyung	npagenum	nlinenum	nbooknum
1	00B329B(	대방광불화	1	1	1
2	00B329B(	대방광불화	17	123	1
3	00B329B(	대방광불화	47	790	1
4	00B329B(	대방광불화	79	1537	1
5	00B329B(	대방광불화	113	2312	1
6	00B329B(	대방광불화	142	3035	1
7	00B329B(	대방광불화	181	3845	1

[그림 14] tag_kyung_table 테이블의 내용

① 테이블 명 : tag_kyung_table

② 테이블의 역할 : 경 제목에 대한 정보를 가지고 있다.

③ 필드의 역할

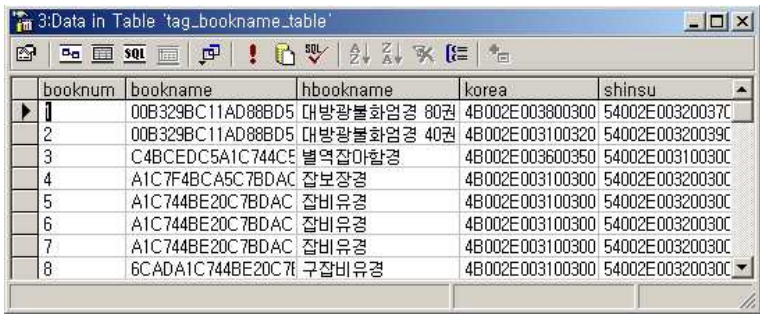
- tag_num : 각 제목에 대한 유일키를 저장한다.
- kyung : 경 제목에 대한 유니코드 값을 저장한다.
- hkyung : 유니코드에 대한 한글 경 제목을 저장한다.
- npagenum : 경 제목이 나타난 곳의 페이지 번호를 저장한다.
- nlinenum : 제목의 위치 정보인 “edocdata” 테이블의 “nlinenum”의 값을 저장한다.
- nbooknum : 현재 권 번호를 저장한다.

■ tag_bookname_table 테이블


2:Design Table 'tag_bookname_table'

열 이름	데이터형식	길이	정밀도	축소	Null 허용
booknum	int	4	10	0	☑
bookname	nvarchar	800	0	0	☑
hbookname	nvarchar	800	0	0	☑
korea	nvarchar	800	0	0	☑
shinsu	nvarchar	800	0	0	☑

[그림 15] tag_bookname_table 테이블의 레코드 형식



booknum	bookname	hbookname	korea	shinsu
1	00B329BC11AD88BD5	대방광불화엄경 80권	4B002E003800300	54002E00320037(
2	00B329BC11AD88BD5	대방광불화엄경 40권	4B002E003100320	54002E00320039(
3	C4BCEDC5A1C744C5	별역잡아함경	4B002E003600350	54002E00310030(
4	A1C7F4BCA5C7BDAC	잡보장경	4B002E003100300	54002E00320030(
5	A1C744BE20C7BDAC	잡비유경	4B002E003100300	54002E00320030(
6	A1C744BE20C7BDAC	잡비유경	4B002E003100300	54002E00320030(
7	A1C744BE20C7BDAC	잡비유경	4B002E003100300	54002E00320030(
8	6CADA1C744BE20C7(	구잡비유경	4B002E003100300	54002E00320030(

[그림 16] tag_bookname_table 테이블의 내용

① 테이블 명 : tag_bookname_table

② 테이블의 역할 : 책이 고려대장경과 신수대장경의 어느 부분에 해당되는지 나타낸다.

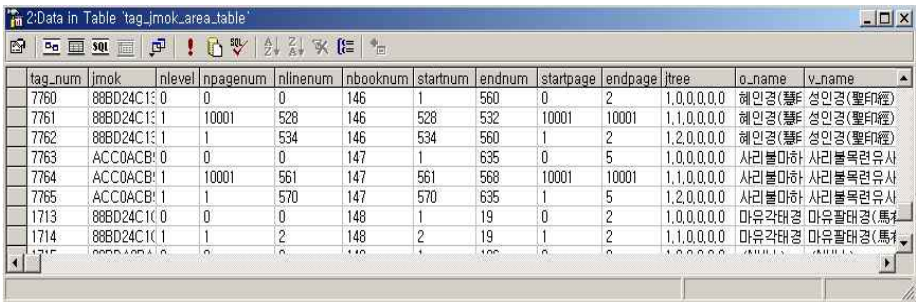
③ 필드의 역할

- book_num : 각 책에 대한 유일키를 저장한다.
- bookname : 책의 제목에 대한 유니코드 값을 저장한다.
- hbookname : 책의 제목의 한글 독음 값을 저장한다.
- korea : 고려대장경의 해당부분을 나타낸다.
- shinsu : 신수대장경의 해당부분을 나타낸다.

■ tag_jmok_area_table

열 이름	데이터형식	길이	정밀도	축소	Null 허용	기
tag_num	int	4	10	0	☑	
jmok	nvarchar	800	0	0	☑	
nlevel	int	4	10	0	☑	
npagenum	int	4	10	0	☑	
nlinenum	int	4	10	0	☑	
nbooknum	int	4	10	0	☑	
startnum	int	4	10	0	☑	
endnum	int	4	10	0	☑	
startpage	int	4	10	0	☑	
endpage	int	4	10	0	☑	
jtree	varchar	15	0	0	☑	
o_name	nvarchar	800	0	0	☑	
v_name	nvarchar	800	0	0	☑	

[그림 17] tag_jmok_area_table 테이블의 레코드 형식



tag_num	jmok	nlevel	npagenum	nlinenum	nbooknum	startnum	endnum	startpage	endpage	jtree	o_name	v_name
7760	88BD24C13	0	0	0	146	1	560	0	2	1,0,0,0,0,0	혜인경(慧E	성인경(聖印經)
7761	88BD24C13	1	10001	528	146	528	532	10001	10001	1,1,0,0,0,0	혜인경(慧E	성인경(聖印經)
7762	88BD24C13	1	1	534	146	534	560	1	2	1,2,0,0,0,0	혜인경(慧E	성인경(聖印經)
7763	ACC0ACB	0	0	0	147	1	635	0	5	1,0,0,0,0,0	사리불마하	사리불목련유사
7764	ACC0ACB	1	10001	561	147	561	568	10001	10001	1,1,0,0,0,0	사리불마하	사리불목련유사
7765	ACC0ACB	1	1	570	147	570	635	1	5	1,2,0,0,0,0	사리불마하	사리불목련유사
1713	88BD24C1(	0	0	0	148	1	19	0	2	1,0,0,0,0,0	마유각태경	마유팔태경(馬
1714	88BD24C1(	1	1	2	148	2	19	1	2	1,1,0,0,0,0	마유각태경	마유팔태경(馬

[그림 18] tag_jmok_area_table 테이블의 내용

① 테이블 명 : tag_jmok_area_table

② 테이블의 역할 : 각 레벨에 해당하는 제목들의 위치 정보를 저장한다.

③ 필드의 역할

- tag_num : 각 제목에 대한 유일키 값을 저장한다.
- jmok : 책의 제목에 대한 유니코드 값을 저장한다.
- nlevel : 각 제목의 레벨 정보를 저장한다.
- npagenum : 각 제목이 위치한 페이지 정보를 저장한다.
- nlinenum : 페이지 안에서 각 제목이 위치한 라인 정보를 저장한다.
- nbooknum : 현재 권의 번호를 저장한다.
- startnum : 각 제목에 해당되는 원문 내용이 “edocdata” 테이블에서 시작되는 정보를 저장한다.
- endnum : 각 제목에 해당되는 원문 내용이 “edocdata” 테이블에서 끝나는 정보를 저장한다.
- startpage : 각 제목이 원문에서 시작되는 페이지에 대한 정보를 저장한다.
- endpage : 각 제목이 원문에서 끝나는 페이지에 대한 정보를 저장한다.
- jtree : 트리 생성시 계층적 구조 정보를 저장한다.
- o_name : 해당 경에 대한 이경명의 정보를 저장한다.
- v_name : 해당 경에 대한 약경명의 정보를 저장한다.

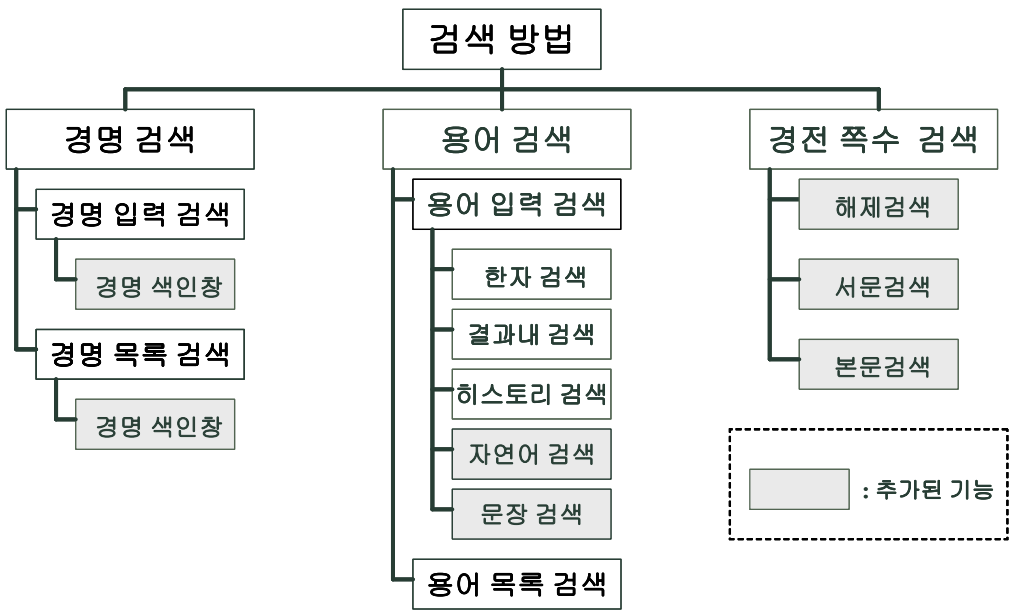
3. 웹 검색시스템

1) 웹 검색시스템의 주요 기능

한글대장경 웹 검색시스템은 사용자가 웹을 통하여 한글대장경을 효율적이고 편리하게 검색할 수 있도록 다양한 검색 방법을 제공하고 있다. 또한 사용자의 편의를 위하여 검색 도움말 및 게시판을 제공하고 있다.

학술적인 참고 자료로서 의미가 있도록 한글대장경 원문과 동일하게 검색 결과에 들여쓰기를 적용하였고, 검색 결과가 한글대장경의 어느 부분에 속하는지 쉽게 알 수 있도록 위치 정보를 제공하고 있으며, 연관된 고려대장경과 신수대장경의 정보를 제공하고 있다.

[그림 19]는 한글대장경 웹 검색시스템에서 제공하는 검색 방법들의 목록을 나타낸다.



[그림 19] 한글대장경 웹 검색시스템에서 제공하는 검색 방법

본 3차 사업에서는 [그림 19]에서 보는바와 같이 기존의 검색 기능에 자연어 검색과 문장 검색 기능을 추가하였고, 경명 선택 기능과 경명 쪽수 검색 기능을 강화하였다.

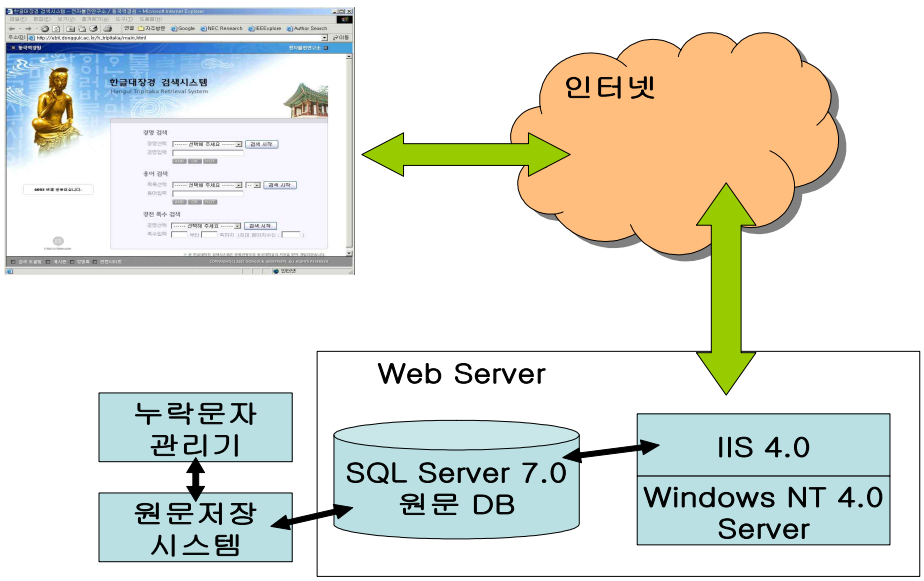
[표 1]은 한글 대장경의 웹 검색시스템에서 제공하는 검색 방법과 그 기능을 나타낸 것이다.

[표 1] 웹 검색시스템에서 제공하는 검색 방법 및 그 기능

검색 방법	기능
경명 검색	경명 입력창에 검색을 원하는 경명을 직접 입력하여 검색하는 방법과 본 검색시스템에서 제공하는 경명 목록에서 원하는 경명을 선택하여 검색하는 방법 등, 두 가지 방법을 제공한다. 본 3차 사업에서는 경명 색인창을 이용하여 본경명(本經名) 뿐 아니라 이경명(異經名), 약경명(略經名)을 이용한 검색을 지원한다.
용어 검색	경명의 목록으로부터 경명을 선택하고, 선택한 경에 대하여 사용자가 용어를 직접 입력하여 검색하는 방법과 제공되는 용어 목록에서 원하는 용어를 선택하여 검색하는 방법 등을 제공한다. 또한 한자 검색, 결과내 검색, 히스토리 검색을 제공한다. 본 3차 사업에서는 유니코드를 기반으로 자연어 검색과 문장 검색을 제공하여 사용자가 정확한 용어를 알지 못하는 경우도 검색이 가능하다.
쪽수 검색	경을 선택하고, 검색을 원하는 페이지의 시작 쪽수와 끝 쪽수를 입력하여 검색한다. 본 3차 사업에서는 선택한 경이 본문뿐 아니라 해제와 서문을 포함하는 경우도 원하는 페이지를 검색할 수 있는 기능을 제공한다.

2) 웹 검색시스템의 구성

사용자가 웹 브라우저를 이용하여 한글대장경 웹 서버에 접속하면 사용자의 요청은 IIS(Internet Information Server)를 통해 ASP로 구현된 웹 검색시스템에 전달된다. 웹 검색시스템은 사용자의 요청을 질의문으로 변경하고, 구축된 데이터베이스에 질의하여 결과를 사용자에게 반환한다. [그림 20]은 한글대장경 웹 검색시스템의 구성을 나타낸다.



[그림 20] 한글대장경 웹 검색시스템 구성도

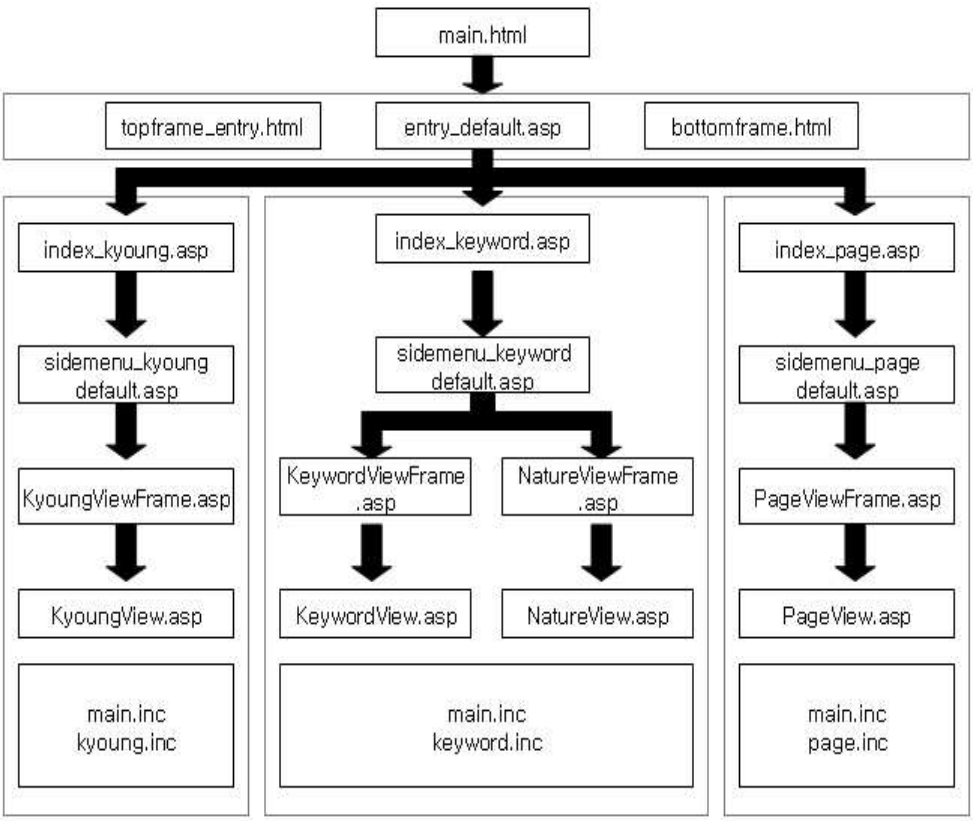
[그림 20]과 같이 개발한 웹 검색시스템은 검색 인터페이스를 구성하는 모듈과 구축된 원문 데이터베이스와 연동하여 질의를 처리하는 모듈로 구성된다.

3) 웹 검색시스템의 구현

한글대장경 웹 검색시스템의 구현 및 운영 환경은 다음과 같다.

- 운영체제 : Microsoft Windows 2000 Server
- 데이터베이스 : Microsoft SQL Server 7.0
- 웹 서버 : Microsoft Internet Information Server 5.0
- 개발 언어 : ASP, Javascript, Visual Basic 6.0, Java
- 클라이언트 환경 : Internet Explorer 5.0 이상

[그림 21]은 개발한 웹 검색시스템의 전체적인 제어 흐름을 나타낸다.



[그림 21] 웹 검색시스템의 제어 흐름도

한글대장경 웹 검색시스템을 접속하면 'main.html'이 호출되는데 'main.html'은 세 개의 프레임으로 구성된다. 상위 프레임은 전자불전

연구소와 동국역경원 링크로 구성된 ‘topframe-entry.html’이고, 하위 프레임은 검색도움말, 게시판, 방명록, 관련사이트 링크로 구성된 ‘bottomframe.html’이다. 중간 프레임에는 경명 검색, 용어 검색, 쪽수 검색 인터페이스로 구성된 ‘entry_default.asp’가 나타난다.

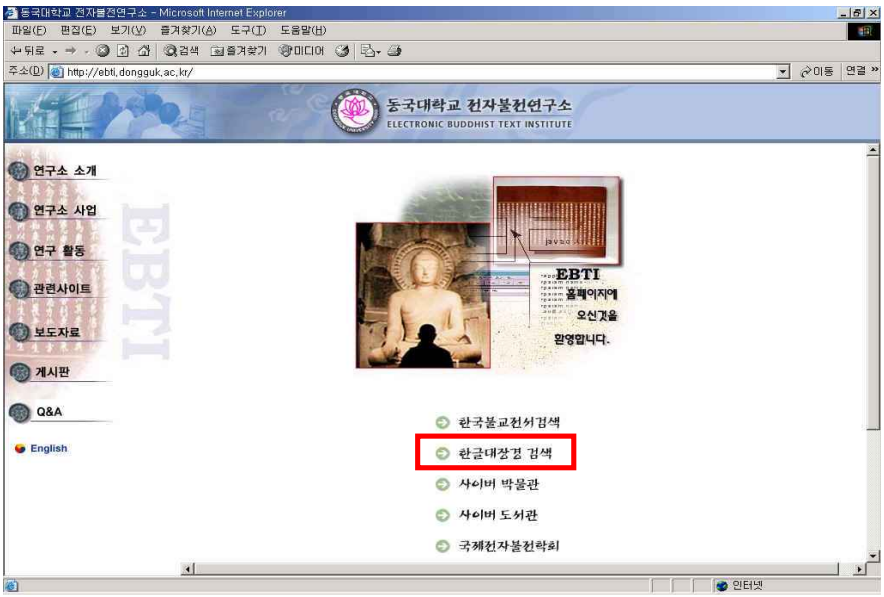
‘index_kyoung.asp’와 ‘sidemenu_kyoungdefault.asp’는 경명 검색 인터페이스를 생성한다. 경명 검색을 하면 ‘KyoungView.asp’가 원문 데이터베이스와 연동하여 경명 검색 결과를 생성한다. ‘kyoung.inc’는 경명 검색과 연관된 공통 기능을 제공한다.

용어 검색을 ‘index_keyword.asp’와 ‘sidemenu_keyworddefault.asp’는 용어 검색 인터페이스를 생성한다. 용어 검색을 하면 ‘KeywordView.asp’는 원문 데이터베이스와 연동하여 단순 용어 검색, 결과내 검색, 히스토리 검색 결과를 생성하며, ‘NatureView.asp’는 원문 데이터베이스와 연동하여 자연어 검색과 문장 검색 결과를 생성한다. ‘keyword.inc’는 용어 검색과 연관된 공통 기능을 제공한다. ‘index_page.asp’와 ‘sidemenu_pagedefault.asp’는 경전 쪽수 검 인터페이스를 생성한다. 경전 쪽수 검색을 하면 ‘PageView.asp’가 원문 데이터베이스와 연동하여 경명 검색 결과를 생성한다. ‘page.inc’는 경명 검색과 연관된 공통 기능을 제공한다.

4) 검색 방법

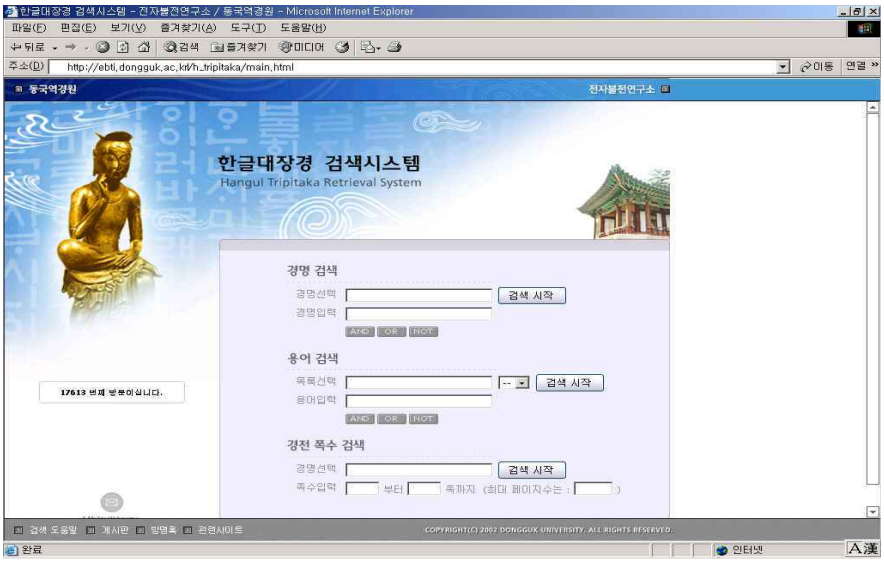
한글대장경 웹 검색시스템을 사용하기 위해서 웹 브라우저를 이용하여 동국대학교 전자불전연구소(http://ebti.dongguk.ac.kr)에 접속한다.

[그림 22]는 동국대학교 전자불전연구소의 초기화면을 나타낸다.



[그림 22] 동국대학교 전자불전연구소의 초기화면

[그림 22]에서 '한글대장경 검색'을 선택하면 새로운 창에 한글대장경 웹 검색시스템이 나타난다. [그림 23]은 한글대장경 웹 검색 시스템의 초기화면을 나타낸다.



[그림 23] 한글대장경 웹 검색시스템 초기화면

웹 검색시스템의 초기화면은 검색 메뉴와 다양한 링크들로 구성되어 있다. 상위 프레임은 전자불전연구소, 동국역경원 링크로 구성되고, 하위 프레임은 검색도움말, 게시판, 방명록, 관련사이트 링크들로 구성되어 있다. 중간 프레임은 경명 검색, 용어 검색, 쪽수 검색으로 구성되어 있고, 각 검색 메뉴에서 조건을 선택하거나 입력하고, '검색시작'을 클릭하면 검색 방법에 맞는 해당 검색 페이지로 이동한다.

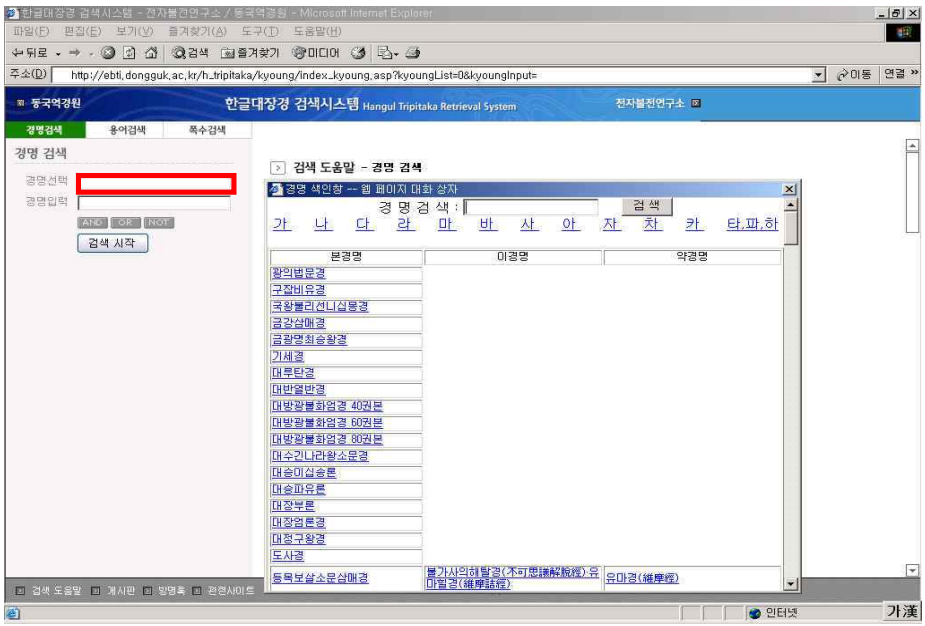
한글대장경 웹 검색시스템의 기본 검색 방법은 입력된 경명이나 용어의 한글 독음을 이용하여 검색하는 것이다. 또한 불리언 검색 기능을 제공하여 사용자가 정확한 질의를 할 수 있도록 한다.

5) 경명 검색

한글대장경은 수많은 경(經)으로 이루어져 있기 때문에 경단위의 검색은 한글대장경에서는 필수적인 기능이라 할 수 있다. 하나의 경은 많은 제목으로 구성되어 있기 때문에 제목을 색인 목록으로 만들면 신속한 검색이 어렵다. 따라서 본 검색시스템에서는 경을 구성하는 제목을 트리형태로 만들어 제공하고 있다. '경명선택' 상자나 '경명입력' 상자를 통해 경명을 선택한 후 '검색시작' 버튼을 누르면 해당 경의 제목 트리가 그 아래에 나타난다. 나타난 제목 트리의 항목을 선택하면 해당 검색 결과가 오른쪽 화면에 나타난다.

경에 따라서는 본경명(本經名)뿐 아니라 이경명(異經名)과 약경명(略經名)이 존재하기 때문에 본 검색시스템에서는 이를 위해 본경명, 이경명 및 약경명의 목록을 경명 색인창으로 구성하고, 경명을 검색하고 선택할 수 있는 기능을 추가하였다.

웹 검색시스템 초기화면이나 '경명검색' 화면에서 '경명선택' 상자를 선택하면 데이터베이스에 저장되어있는 경들의 본경명, 이경명 및 약경명으로 구성된 경명 목록이 본경명의 오름차순으로 경명 색인창을 통해 나타낸다. [그림 24]는 '경명선택' 상자를 선택하였을 때 나타나는 본경명, 이경명, 약경명으로 구성된 경명 색인창을 나타낸다.

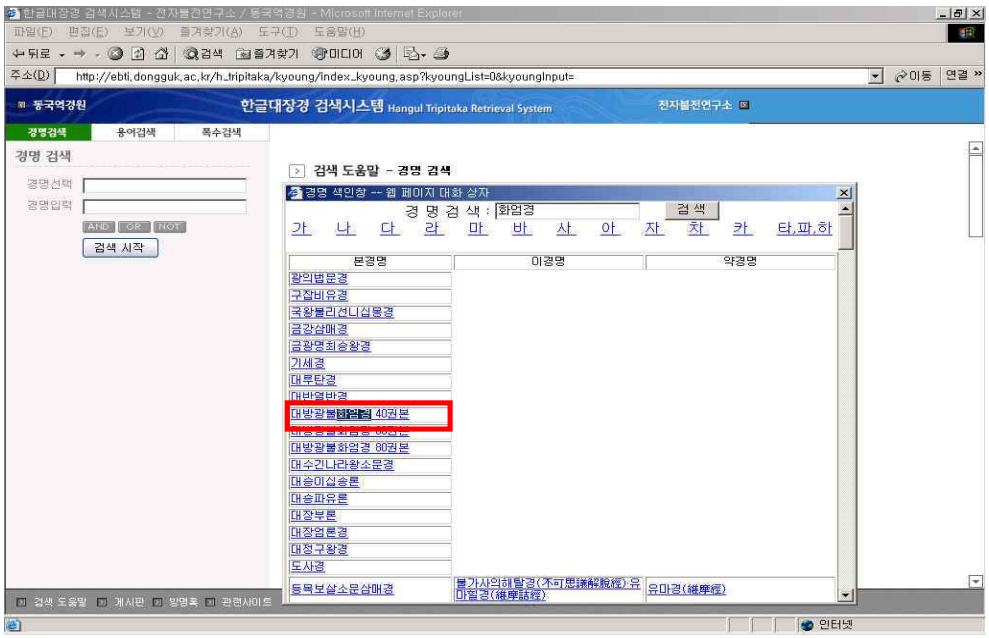


[그림 24] 본경명, 이경명, 약경명으로 구성된 경명 색인창

데이터베이스에 저장된 경의 수가 증가함에 따라 경명 색인창을 구성하는 경명 목록이 증가하게 되어 순차적인 접근으로는 원하는 경명을 찾기가 어렵다. 따라서 본 검색시스템에서는 경명 색인창에 경명 검색 기능을 추가하였다. 경명 색인창의 상단에 있는 입력창에 경명을 입력하고, '검색' 버튼을 누르면 입력한 경명을 포함하는 경명을 찾아 표시해준다.

사용자는 경명 색인창 상단의 가, 나, 다 색인을 통해 특정 철자로 시작하는 본경명으로 빠르게 이동할 수 있으며, 이경명이나 약경명을 검색할 수도 있다. 이경명이나 약경명을 선택하는 경우 '경명선택' 상자에는 해당 본경명이 입력되도록 처리하여 모든 경명이 식별 가능하도록 하였다.

[그림 25]는 '화엄경'이란 단어를 포함하는 경명을 검색한 화면이다.



[그림 25] '화엄경'이란 단어를 이용한 경명 검색

6) 용어 검색

용어 검색은 등록된 용어를 목록에서 선택하거나 직접 한글 독음을 입력하여 해당 용어를 포함하는 페이지를 검색하는 방법이다. 현재 사용중인 한글대장경 웹 검색시스템은 50,000여개의 불교 용어를 이용하여 경전별로 용어의 위치를 색인화함으로써 빠른 검색이 가능하도록 하고, 또한 사용자의 다양한 검색 요구에 맞도록 '한자 검색', '결과내 검색', '히스토리 검색', '자연어 검색'과 '문장 검색'을 지원한다. 용어 검색은 크게 '입력 검색'과 '목록 검색'으로 나눌 수 있다.

'입력 검색'은 검색시스템 초기화면이나 용어 검색 화면에서 경을 선택하고, 용어입력란에 검색할 용어를 입력하여 검색하는 방법이다.

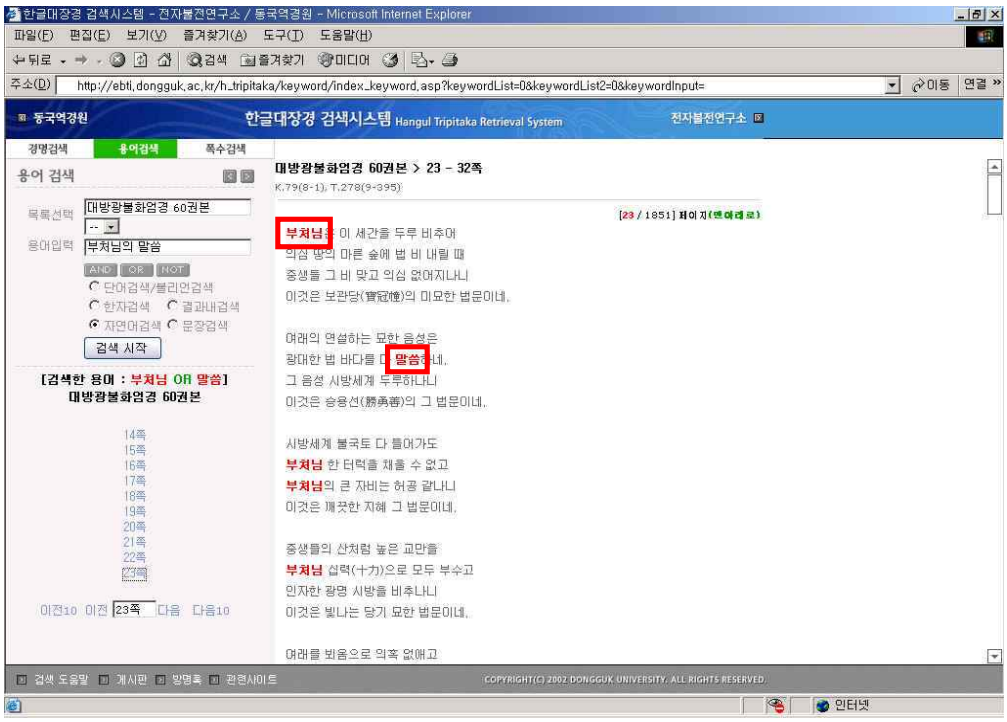
또 다른 용어 검색 방법 중의 하나는 '목록 검색'이다. '목록 검색'은 검색시스템 초기화면이나 용어 검색 화면에서 하나의 경을 선택하고, '목록 선택'에서 'ㄱ'부터 'ㅎ' 중에 하나를 선택하여 검색된 용어 목록 중 한 가지를 선택하여 검색하는 방법이다.

용어 하나만 입력하여 검색하거나 목록을 이용해서 선택하여 검색하는 단순한 방법 이외에도 '결과내 검색', '한자 검색', '히스토리 검색' 등을 제공한다. 본 3차 사업을 통해 '자연어 검색'과 '문장 검색' 방법까지 제공하고 있다.

자연어 검색은 문장 검색과 달리 입력된 문장을 형태소 분석을 통해 검색 가능한 용어를 추출하고 경 단위로 입력한 단어를 유니코드로 변경하고 유니코드의 패턴매칭을 통해 추출된 용어들을 검색한다.

'자연어 검색' 라디오 버튼을 선택한 후 '용어입력' 상자에 자연어 검색을 하고자 하는 문장을 입력하고 검색을 실행하면 형태소 분석을 통해 입력한 문장에서 용어를 추출하게 된다. 다시 검색을 실행하면 추출된 용어들을 이용해 검색한다.

[그림 26]은 '대방광불화엄경 60권본'에서 '부처님의 말씀'라는 문장을 입력하여 '자연어 검색'을 실행한 화면이다.

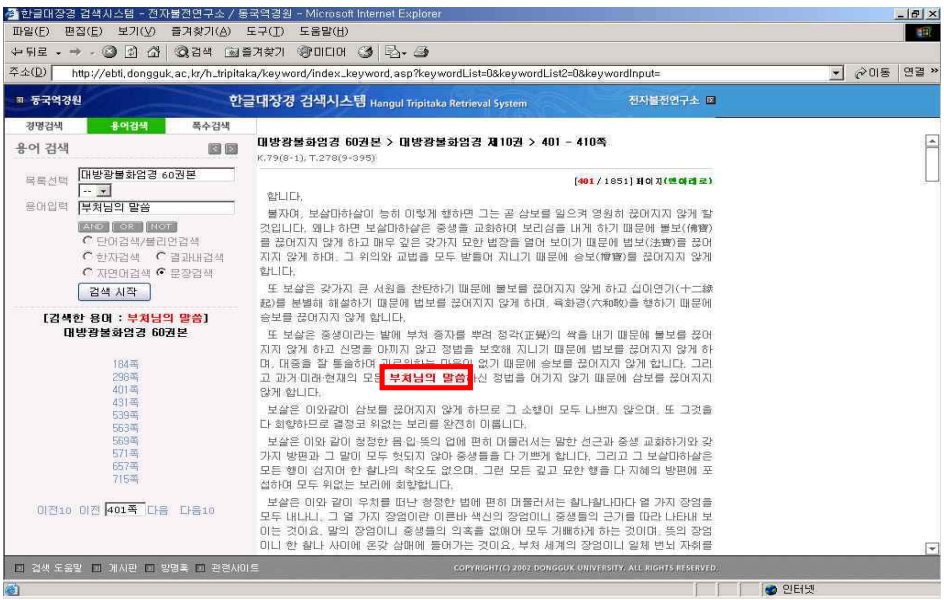


[그림 26] '부처님의 말씀'을 이용한 문장 검색 결과 화면

기존의 용어 검색 방법에서는 용어 선택이나 용어 입력에 의한 검색만이 가능하였지만 문장 검색에서는 입력한 문장 전체가 포함된 페이지를 검색하는 것이 가능하다.

‘문장검색’ 라디오 버튼을 선택한 후 ‘용어입력’ 상자에 문장 검색을 하고자 하는 문장을 입력하면 경 단위로 입력한 문장 전체를 포함하는 페이지를 검색한다.

[그림 27]은 ‘대방광불화엄경 60권본’에서 ‘부처님의 말씀’이라는 문장을 입력하여 ‘문장 검색’을 실행한 결과 화면으로 오른쪽 페이지에 ‘부처님의 말씀’이란 문장 전체가 포함된 것을 볼 수 있다.



[그림 27] ‘부처님의 말씀’을 이용한 문장 검색 결과 화면

7) 경전 쪽수 검색

한글대장경 웹 검색시스템에서는 사용자가 10쪽 이상의 검색 결과를 원하거나 특정 쪽으로 이동할 수 있도록 경전 쪽수 검색을 제공하고 있다. 또한 사용자가 경을 선택하면 경의 최대 페이지 수 정보를 제공하여 검색 페이지 입력의 오류를 방지하고 있다.

본 3차 사업에서는 사용자가 선택한 경에 따라 본문, 해제, 서문의 존재 유무를 라디오 버튼의 활성화를 통해 사용자에게 알려주는 기능을 추가였다. 또한 활성화된 본문, 해제, 서문 중 원하는 쪽수 검색을 선택하면 해당 경전의 최대 페이지 수를 알려줌으로써 사용자에게 보다 자세하고 편리한 쪽수 검색이 가능하도록 경전 쪽수 검색 기능을 강화하였다.

검색시스템의 첫 페이지나 쪽수 검색 화면에서 경을 선택하고 본문, 해제, 서문으로 나누어져 있는 경전의 종류 중 하나를 선택한 다음, 시작 쪽수와 마지막 쪽수를 입력하고 '검색 시작'을 누르면 입력한 쪽수가 경의 범위에 속하는지를 확인하고, 경명과 쪽수를 이용하여 검색한다.

본문 이외에 해제와 서문이 존재하는 경의 경우, 경전을 선택하면 해제와 서문에 해당하는 라디오 버튼이 활성화되고, 선택한 경전의 최대 페이지수를 나타내줌으로써 사용자에게 편리한 쪽수 검색을 제공한다. [그림 28]은 '대방광불화엄경 60권본' 해제의 2쪽부터 5쪽까지를 검색한 화면이다.

[그림 28] '대방광불화엄경 60권본'의 해제 쪽수 검색 결과

4. 유니코드에서 누락된 문자 및 진언 처리

누락 문자란 현재 윈도우즈 운영체제 및 인터넷 환경에서 사용 가능한 한자에 포함되지 않는 문자를 뜻한다. 한글 윈도우즈에서 채택하고 있는 KSC-5601 한글 체계상에서 한자는 대략 4,888자 정도 지원되고 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. 이러한 누락 문자가 존재하는 이유는 다음과 같이 볼 수 있다.

- 고문헌이 집필될 당시의 한자들이 유니코드 내에 포함되어 있지 않은 경우
- 고문헌의 기록 과정에서 오자 입력으로 인한 실제 존재하지 않는 글자인 경우

입력 과정을 거쳐야 하는 한글 대장경 원문의 분량이 방대하기 때문에 가능한 입력 도구의 간소화와 편리화가 필요로 하다. 특히, 누락 문자는 수작업을 통해 문서상에 정해진 태그의 형태로 삽입되어야 한다. 따라서 누락 문자 자체를 입력하는 과정이 대단히 번거롭고 시간을 많이 소요하게 되는 작업이 된다.

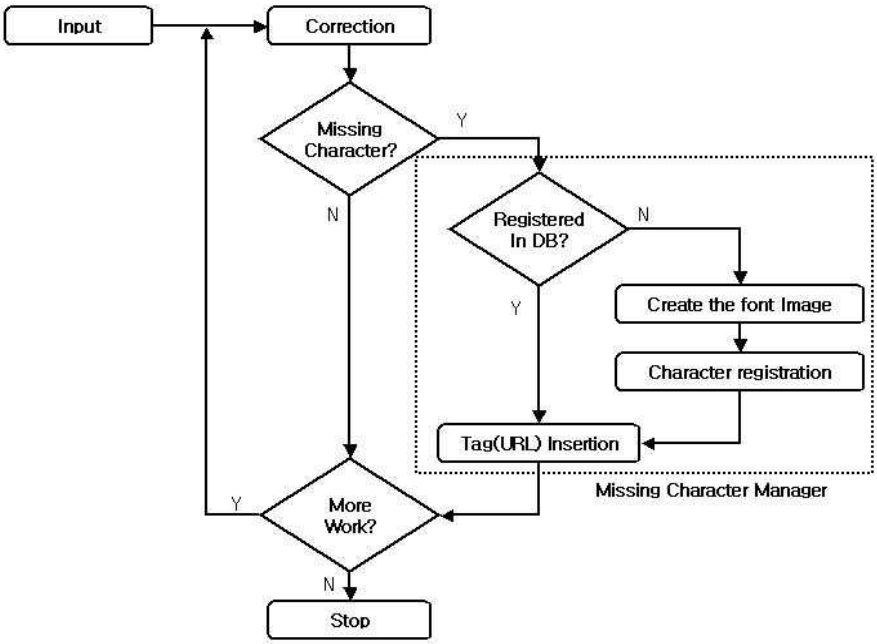
이러한 누락 문자 입력 과정을 단순화하고 실제 누락 문자를 간단한 방법으로 문서상에 이미지 태그 형태로 삽입할 수 있는 누락 문자 관리기를 개발하였다.

한국 고문헌 상에 나타난 누락 문자는 그 자체로 중요한 의미를 갖는다. 이것은 후에 한국 고문헌을 위한 폰트 체계를 정비하는데 있어 도움이 될 뿐만 아니라 한글 대장경 원문 상에 나타난 누락 문자의 발생 빈도 등을 한눈에 알아볼 수 있게 해준다. 따라서 누락 문자 관리기는 문자의 등록 및 이미지 태그 입력 기능뿐만 아니라 각각의

누락 문자에 대한 통계 자료를 제공하는 기능도 포함하여야 한다.

4.1 누락 문자 관리

유니코드에 나와 있지 않은 문자인 누락문자를 입력하기 위해서는 누락문자를 GIF 형식의 폰트 이미지 파일로 만들고 누락 문자 DB에 등록한다. 그리고 한국 불교 전서를 인터넷을 통해 검색 시 다른 유니코드 문자들과 함께 등록된 누락 문자를 웹 브라우저 상에서 보여준다. 다음 [그림 29]는 누락 문자를 입력하는 작업을 나타내고 있다.



[그림 29] 한글대장경 입력과정

(1) 원문 입력

한국 대장경 원문을 외주를 통해 직접 한글 97 프로그램을 이용하여 직접 입력하고, 이 때 입력이 불가능한 한자나 특수 기호는 특별 기호로 표시하게 된다.

(2) 교정작업

텍스트 파일을 원문과 비교하여 잘못 입력된 내용이 없는지 검토하고 잘못 입력된 내용이 있다면 교정한다. 해당하는 누락 문자의 위치(책, 페이지, 단락, 라인)를 문서화한다.

(3) 누락 문자 입력

교정 작업 도중 누락 문자가 발견되면 누락 문자 검색 프로그램을 사용하여 이미 발견된 누락 문자인지 검색한다. 만약 이미 발견된 누락 문자라면 검색 프로그램을 이용하여 누락문자에 해당하는 Tag를 삽입한다. 여기서 Tag는 누락 문자 이미지가 저장되어 있는 주소를 나타내고 URL 주소로 표현된다. 그러나 저장되어있는 누락문자 중에서 해당 누락 문자를 찾지 못하면 누락 문자를 이미지 파일로 만들고 누락 문자 검색 프로그램에 등록한 후 이에 해당하는 Tag를 삽입한다. 교정 작업 중 다시 누락 문자가 발견된다면 위의 작업을 반복한다.

4.2 누락 문자 관리기

본 과제의 수행에 있어 필요한 누락 문자 관리기의 요구 조건은 다음과 같다.

- 편리한 누락 문자의 등록
- 등록된 누락문자에 대한 빠른 문서상의 입력
- 문서상에 나타난 누락문자의 체계적인 관리 및 통계자료의 제공
- 웹 문서에서 누락 문자사용의 무 제약성 제공

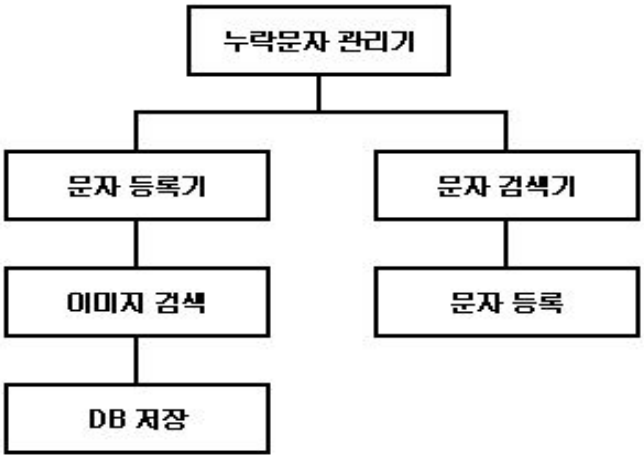
한글 대장경 원문의 입력 작업은 매우 많은 원문의 분량으로 인한 많은 시간이 소요된다. 이런 상황에서 입력과정에서 발견되는 누락문자를 손쉽고 빠르게 등록시킬 수 있는 기능은 필수적이라 할 수 있다. 누락 문자의 등록을 위해서 누락 문자 관리기는 여러 한자 이미

지를 체계적인 정렬 방법을 통해 제시하여야 하며, 사용자는 이러한 한자 중에 자신이 찾고자 하는 이미지를 효과적인 방법으로 검색해 낼 수 있어야 한다. 또한 찾고자 하는 누락 문자 이미지가 없는 경우 손쉬운 방법으로 누락 문자 이미지를 작성할 수 있어야 한다.

또한 문자 관리기는 등록된 문자에 대해 효과적으로 검색이 가능하여 간편하게 원문 상에 등록된 문자에 대응하는 태그를 입력할 수 있어야 한다. 문자 관리기는 현재 저장하고 있는 등록된 문자의 목록을 효과적으로 게시해 주어야 하며 게시된 문자를 여러 조작 없이, 예를 들어 1회의 마우스 클릭 등의 작업을 통해 원문 상 지정된 위치에 태그 정보를 삽입할 수 있어야 할 것이다.

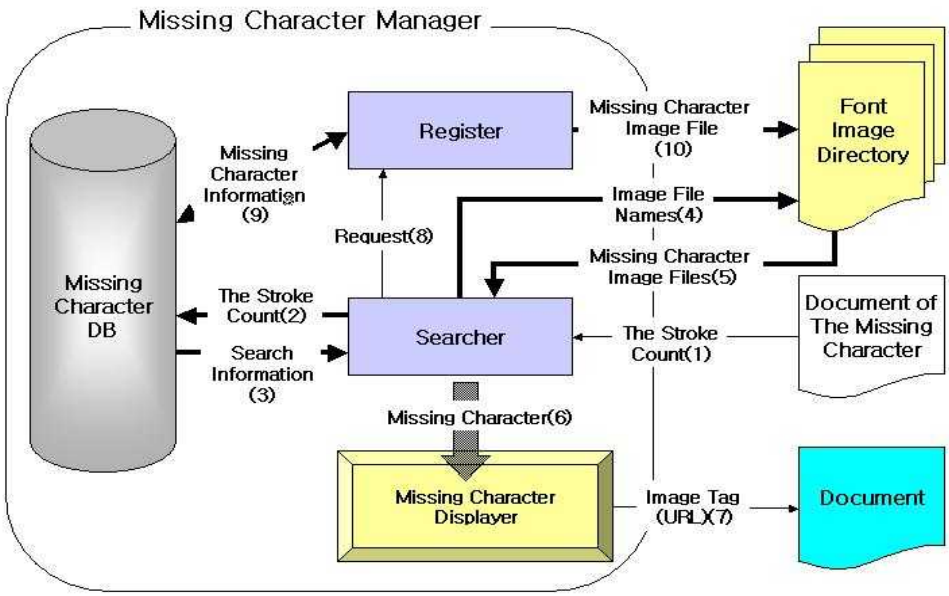
전산화된 한글 대장경은 인터넷을 통해 사용자에게 서비스를 제공하게 될 것이다. 이에 따라 다양한 형태(현재는 색상의 변환)의 문자가 제공되어야 할 것이다. 기본적으로 인터넷에서 사용되고 있는 문자의 색상은 링크가 걸리지 않을 경우 검정색이 링크가 연결되어 있을 경우 파란색에 밑줄이 그리고 링크를 거친 후에는 보라색의 밑줄의 형태로 문자가 제공되어진다.

유니코드로 제공되어지는 문자는 HTML의 Tag 문을 사용하여 이러한 기능을 모두 제공하지만 이번 작업에서 제작된 누락 문자 이미지는 그렇지 못하다 따라서 웹상에서 누락 문자의 표현에 제약성을 없애기 위해서는 이미지의 색상 변환이 필요로 하다. 쉽게 하나의 검정색을 가진 이미지 파일을 하나하나 색상을 필요시 바꿀 수 있지만 이러한 작업은 많은 시간이 필요로 한다. 이에 이미지 색상 변환도 손쉽게 할 수 있는 프로그램이 필요로 하다.



[그림 30] 누락 문자 관리기의 구조

[그림 30]과 같이 누락 문자 관리기는 문자의 등록, 검색, 문자의 색상 변환 등 몇 가지 기능별 구조를 갖는다. 문자 등록기 상에서는 표현되지 않는 문자들의 이미지 추출 및 내부적인 코드 부여, DB에 저장 등 기능을 가지며, 문자 검색기는 현재 등록된 문자의 검색 및 검색된 문자를 원문 상에 삽입하는 기능을 갖는다.



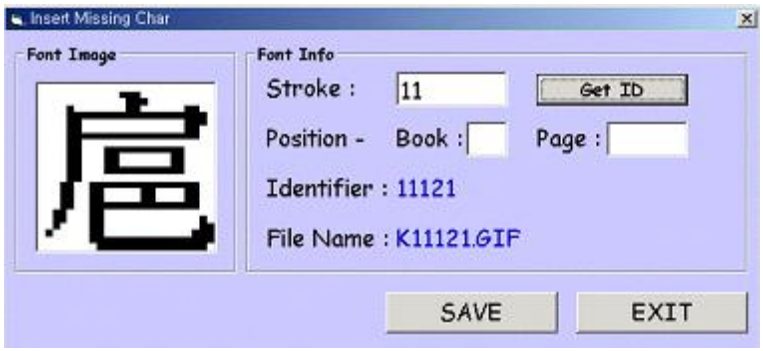
[그림 31] 누락문자 관리기의 동작 과정

[그림 31]은 누락문자 관리기의 전체적인 동작과정을 설명하고 있다. 굵은 실선은 네트워크로 연결되어 있음을 의미하고 각각의 숫자는 동작 순서를 의미한다. 먼저 원문에 누락문자가 존재하는 것을 발견하면 해당 누락문자의 위치 정보를 직접 손으로 작성한다. 작성된 누락문자에 관한 문서를 이용하여 누락문자를 이미지로 생성한다.

실질적인 누락문자 관리기에서 가장 먼저 시작되는 과정은 누락문자의 위치정보를 가진 문서에서 각각의 누락문자 총획수를 검색기에 전달한다. 검색기는 네트워크로 연결된 Missing Character DB에서 해당 총획수를 가지는 문자의 정보를 가져온다. 이때 가지고 오는 정보는 누락문자 이미지의 태그 정보(URL)와 해당 누락문자 이미지 파일 이름으로 구성된다. 누락문자 이미지 파일이름을 이용하여 네트워크로 연결된 누락문자 이미지 파일이 저장되어 있는 디렉토리에서 해당 파일을 가져와서 태그 정보과 이미지 파일을 병합하여 사용자에게 이미지 파일을 보여준다.

사용자는 보여진 이미지 파일 중 해당 문자가 있는지를 검사하고 만약 해당 문자가 있으면 원문에 해당 누락문자의 URL을 입력한다. 그렇지 않다면 검색기에서 등록기로 누락문자 등록을 요청한다. 문자 등록기는 누락문자에 대한 이미지와 누락문자 정보를 생성하여 각각 누락문자 디렉토리와 누락문자 데이터베이스에 저장하고 문자 검색기에 저장되었음을 알려준다. 응답을 받은 문자 검색기는 누락문자 검색 과정을 통해 해당 누락문자의 태그 정보(URL)을 원문에 입력한다.

문자 등록기는 한글대장경 원문 입력 시 나타나는 누락 문자를 문자 관리기에 저장하는 기능을 지니고 있다. 한글대장경 전산화에서는 누락문자를 이미지 형태로 관리하고 있으며 본문 상에 삽입하기 위해서는 HTML의 이미지 태그 정보를 사용하게 된다.



[그림 32] 문자의 등록

[그림 32]는 누락 문자를 데이터베이스에 등록하는 화면을 보여주고 있다. 사용자가 입력해야 할 정보는 문자의 총 획수(Stroke)와 문자가 처음 나타난 곳(Book, Page)에 대한 정보이다. 누락 문자에 대한 총 획수를 입력하고, <Get ID> 버튼을 누르게 되면 등록할 문자의 고유한 코드(Identifier)를 자동 생성해주며 실제 저장에 사용될 파일 이름(File Name)을 표시해 준다. 이와 같은 상태에서 <SAVE> 버튼을 누르면 누락 문자 이미지 파일과 부가적인 정보가 데이터베이스에 저장된다.

정보가 저장되는 데이터베이스의 테이블은 IDpool 테이블과 CharMap 테이블이다. IDpool 테이블은 총획수와 해당 총획수를 가지는 누락문자의 개수를 표현하는 필드 CharNum과 ID로 구성되어 있다. 그리고 CharMap 테이블은 누락문자 이미지 파일 이름으로 사용할 CharID 필드, 총획수를 나타내는 CharNum 필드, 해당 누락문자의 이미지 태그인 URL을 저장하는 URL 필드로 구성된다. 누락문자 파일 명을 결정하는 체계는 다음과 같다.

K[총획수][등록된 순서]

등록된 누락 문자 이미지 파일은 필요에 따라 다양한 곳에서 이미지 태그 형태로 삽입이 가능하다.

누락문자 입력 방법은 누락 문자 관리기의 검색 기능을 이용하여 가능하다. 입력하고자 하는 문자를 찾아내고 해당 글자를 더블 클릭하는 작업을 통하여 수행된다.

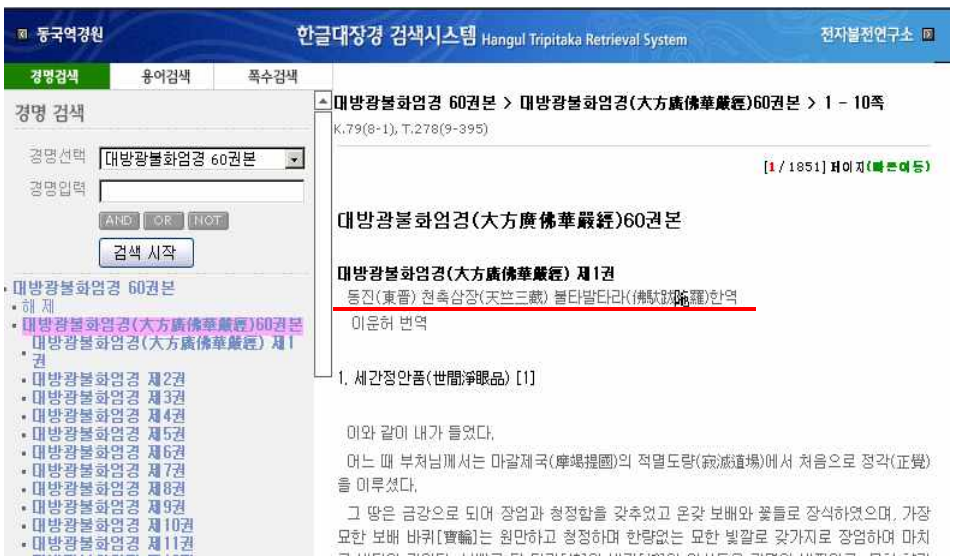
[그림 33] 누락문자 검색기

[그림 33]은 문자 등록기에 의해 데이터베이스에 저장된 누락 문자를 검색한 화면이다. 검색 방법은 등록 시 사용한 누락 문자의 총획수를 통해서 이루어진다. 찾고자 하는 누락 문자의 총획수를 입력하고 <Search> 버튼을 클릭하면 데이터베이스에 저장되어 있는 누락 문자 이미지 파일 중에서 입력된 총 획수에 대한 정보를 가지고 있는(파일 명을 결정하는 체계에 의해) 누락 문자 이미지 파일을 게시해 주며 찾고자 하는 누락 문자를 더블 클릭하게 되면 그림에서 보는 바와 같이 메시지 상자가 생기고, 누락 문자에 대한 이미지 태그 정보를 구하게 되며 메시지 박스 내의 확인 버튼을 누르면 이미지 태그 정보가 한글 2002에 입력 중인 원문에 입력하여 준다.

이와 같은 과정은 클립보드를 이용한 자료의 이동 방법을 통하여 구현되었다. 클립보드로 누락 문자를 복사한 후 다른 애플리케이션에 복사하기 위해서는 다음과 같은 작업을 한다.

1. 현재 스크린에 실행중인 모든 윈도우들을 검사
2. 이 중 윈도우 캡션 이름이 “txt”인 윈도우를 검사하고 찾은 윈도우를 화면 상단(On Top)으로 불러오고 활성화시킨다.
3. 복사하기(Ctrl-V) 버튼을 누른다.

위의 3가지 작업을 거치면 원문을 입력 중인 한글 2002가 활성화되고 누락 문자 검색기를 실행하기 직전의 커서가 있던 지점 이후로 누락 문자 이미지 파일에 대한 정보를 가진 이미지 태그 문이 복사가 된다.



[그림 34] 누락문자 이미지 사용 예

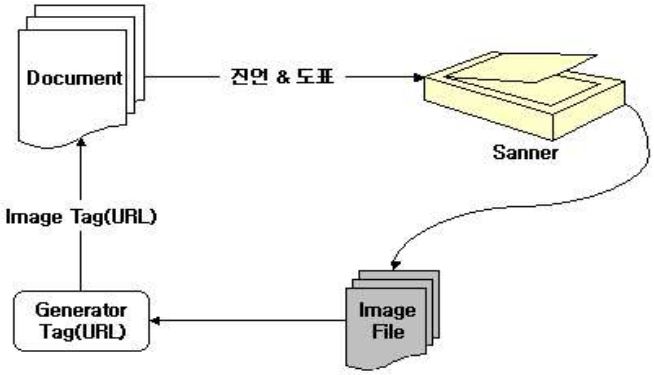
[그림 34]는 실제 웹 페이지 상에서 누락 문자 이미지 사용의 결과를 나타낸다. 굵은 붉은 실선이 그어진 부분에 대한 HTML 코드를 살펴보면 다음과 같다. 유니코드로 작성된 문자열 사이에 이미지 태그가 들어 있는 것을 확인할 수 있다. ‘&#’ 이후에 쓰인 정수는 웹 페이지에서 각각의 문자에 해당하는 유니코드를 나타낸다.

```
<tr>
<td class=content_type08>
  &#46041&#51652&#40&#26481&#26187&#41&#32&#52380&#52629&#49340&#51109&#40&#22825&#31482&#19977&#34255&#41&#32&#48520&#53440&#48156&#53440&#46972&#40&#20315&#39345&#36299
<IMG SRC=http://ebti.dongguk.ac.kr/images/&#107&#48&#55&#49&#54.gif width=12 height=13 align=texttop>
&#32645&#41&#54620&#50669
</td>
</tr>
```

4.3 진언 및 도표 처리

한글 대장경에서 존재하는 진언 및 도표도 누락문자와 같이 이미지 파일을 생성하여 웹에서 볼 수 있도록 처리하였다. 진언의 사전 풀이는 '진실하여 거짓됨이 없는 불교의 비밀스런 주문. 부처와 보살의 서원이나 덕, 그 별명이나 가르침을 간직한 비밀의 어구.'를 뜻한다. 진언은 부처님 재세시 인도에서 쓰이던 고대 언어인 범어(산스크리트어)를 한자로 음차한 것으로 우리나라를 포함해 중국과 일본 등에서는 진언을 번역하지 않고 그대로 읽고 있다.

범어를 한자로 음차하였기 때문에 많은 수의 누락문자가 존재하고 또한 진언의 독특한 형태를 그대로 웹상에 보여주기에는 많은 문제점을 가지고 있기 때문에 한글 대장경 원문에 들어 있는 진언을 하나의 이미지로 생성하여 웹상에 보여주는 것이 사용자에게는 더 자연스럽게 보여진다. 그리고 한글 대장경 원문에 존재하는 도표도 화면 구성시 불균형을 해소하기 위해 이미지 처리를 한다.

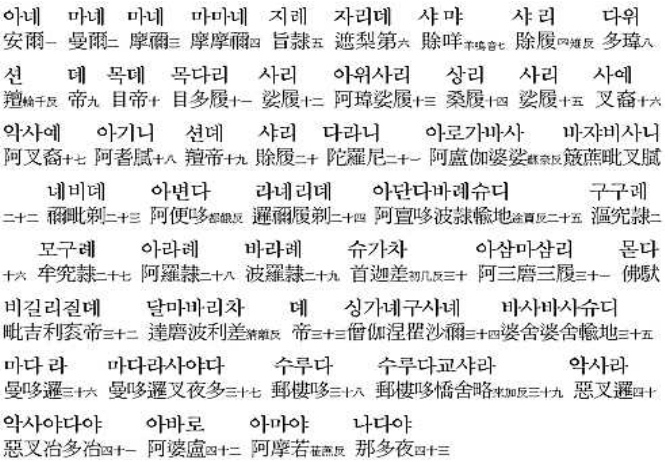


[그림 35] 진언과 도표 이미지 처리 과정

진언 및 도표의 이미지 처리과정은 [그림 35]와 같다. 먼저 진언과 도표가 있는 원문 페이지를 찾아 스캐너로 해당 부분을 스캔하여 이미지 파일을 생성한다. 그 후 해당 이미지 파일에 이름을 다음과 같은 형식으로 변환한다.

[경이름의 영문약자][페이지번호]-[등록 순서].png

마지막으로 이미지 태그를 작성하여 원문 파일에 해당 진언 및 도표의 이미지 태그를 작성한다.

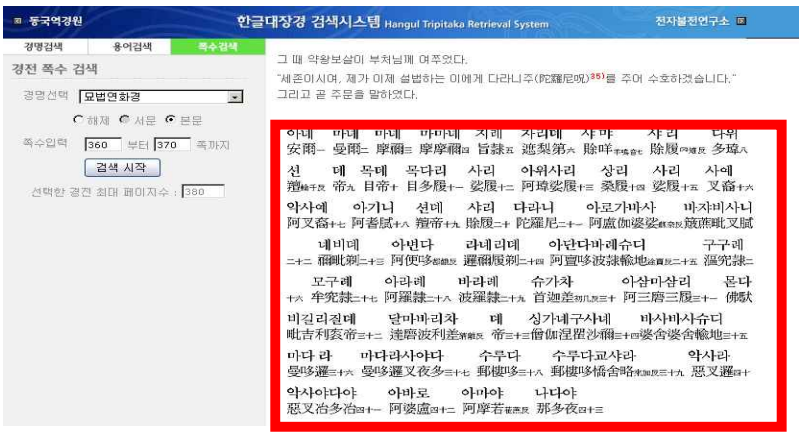

아네 마네 마네 마마네 지레 자리데 샤먀 샤리 다위
安爾一 曼爾二 摩禰三 摩摩禰四 旨隷五 遮梨第六 賒咩平填音七 除履四姪反 多璋八
션 뎨 목뎨 목다리 사리 아위사리 상리 사리 사예
遍編千反 帝九 目帝十 目多履十一 娑履十二 阿璋娑履十三 桑履十四 娑履十五 叉裔十六
악사예 아기니 션뎨 샤리 다라니 아로가바사 바자비사니
阿叉裔十七 阿耆膩十八 遍帝十九 賒履二十 陀羅尼二十一 阿盧伽婆娑麒奈反簸蔗毗叉膩
네비뎨 아면다 라네리뎨 아단다바례슈디 구구례
二十二 禰毗剃二十三 阿便哆都緤反 邏禰履剃二十四 阿亶哆波隷輸地金寶反二十五 溫究隷二
모구례 아라례 바라례 슈가차 아삼마삼리 몬다
十六 牟究隷二十七 阿羅隷二十八 波羅隷二十九 首迦差初几反三十 阿三磨三履三十一 佛默
비길리젼뎨 달마바리차 뎨 싱가녜구사녜 바사바사슈디
毗吉利蓁帝三十二 達磨波利差第賴反 帝三十三 僧伽涅瑟沙禰三十四婆舍婆舍輸地三十五
마다라 마다라사야다 수루다 수루다교샤라 악사라
曼哆邏三十六 曼哆邏叉夜多三十七 郵樓哆三十八 郵樓哆憍舍略求加反三十九 惡叉邏四十
악사야다야 아바로 아마야 나다야
惡叉冶多冶四十一 阿婆盧四十二 阿摩若翟蕉反 那多夜四十三

[그림 36] 진언 스캔 이미지 예

[그림 36]은 “묘법연화경”에 존재하는 진언의 한 부분을 스캔한 이미지이다. 이 이미지 파일을 이름은 법원의 영문 표기의 약자인 mb와 진언이 존재하는 페이지인 365 페이지를 붙이고 해당 페이지에서 스캔된 이미지가 하나이므로 등록 순서는 붙이지 않으면 mb365.png가 된다.

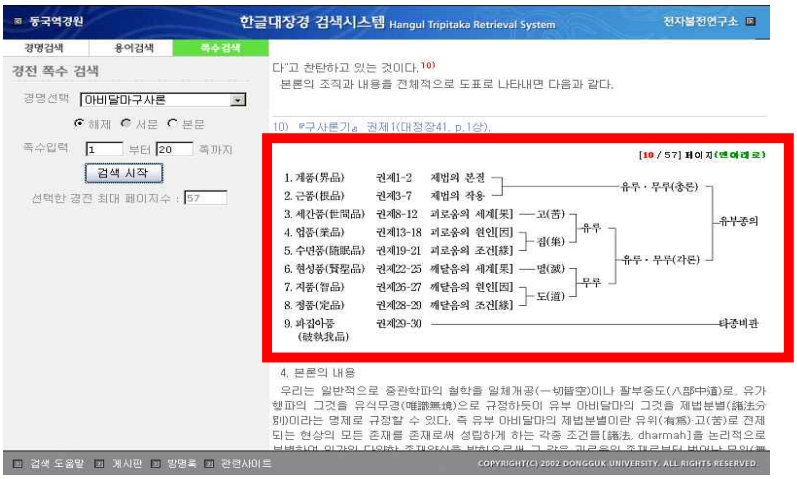
[그림 37] 도표 스캔 이미지의 예

[그림 37]은 아미달마구사론에 있는 도표를 스캔한 이미지이다. 이 이미지 파일의 이름도 [그림 38]의 이미지 파일과 같이 아미달마구사론의 영문표기의 약자와 페이지수를 결합하고 해당 페이지 내에 스캔된 이미지 파일이 두개 이상이므로 등록순서 1을 입력하여 ami10-1.png로 표시한다.



[그림 38] 실제 웹상에서 보이는 진언 이미지

[그림 38]은 실제 웹상에서 스캔 과정을 거치 진언 이미지가 보이는 것을 나타낸다.



[그림 39] 실제 웹상에서 보이는 도표 이미지

[그림 39]는 실제 웹상에서 보이는 도표 이미지를 나타낸다. HTML 코드를 살펴보면 다음과 같다. 유니코드로 작성된 문자열 사이에 이미지 태그가 들어 있는 것을 확인할 수 있다.

```
<tr><td height=8><img src=image/common/dot_gray1.gif
width=100% height=1></td></tr>
<tr><td height=18 valign=middle align=right class=content_type03>
[<span class=content_type04><a name=10>
<a href=./page/index_page.asp?bookNum=214&pageNum=10010
&endNum=10019&maxPageHidden=1397 target=_top>10</a>
</span> / 57] 페이지<a href=#quickSearch><span class=quick>(맨아래로)
</span></a></td></tr><tr><td class=content_type08>  
<IMG SRC=http://ebti.dongguk.ac.kr/images/
&#97&#98&#105&#49&#48&#45&#49.png align=texttop>
<tr><td class=content_type08>  <tr><td>
```

III. 결론

본교는 불교학을 중심으로 한 한국학과 컴퓨터 정보통신 두 분야를 특성화의 큰 축으로 하고 있으며, 불교 자료의 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 한국 불교 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 한글대장경을 전산화하여 이를 연구하는 연구자들이나 열람을 원하는 사람들에게 도움이 될 뿐만 아니라 우리의 귀중한 문화유산을 전 세계에 널리 알릴 수 있다.

한글대장경의 전산화를 위하여 가장 필요한 것은 워드프로세서 입력형태로 되어있는 한글대장경 원문을 데이터베이스에 저장하는 기술, 저장된 데이터베이스에서 원하는 부분을 검색하는 기술 및 이를 인터넷에서 사용할 수 있도록 하는 인터페이스 처리 기술이다.

본 연구에서는 한글 워드 프로세서로 작업한 형태의 파일을 일반 유니코드 텍스트로 변환하여 이것을 유니코드 형태 그대로 데이터베이스에 저장하는 기술을 개발 및 구현하였다. 또한 검색 구조를 위하여 문서의 논리적 구조를 표현할 수 있는 XML을 도입하여 재구성하였으며, 이러한 XML 형태의 문서에서 실제 검색에 필요한 조건들을 추출하여 데이터베이스를 구축하였다.

또한 이렇게 구축된 데이터베이스를 인터넷상에서 열람 및 검색이 가능하도록 웹-기반 프로그램을 작성하였으며, 이를 통하여 인터넷

환경에서 직접 한글대장경을 열람할 수 있도록 하였다. 그리고 여러 가지 검색 기능을 추가하여 사용자가 손쉽게 한글대장경을 열람하고 검색할 수 있도록 하였다.

그리고 유니코드로 표현되지 않는 한자를 인터넷에서 사용할 수 있도록 누락문자를 이미지화하고 원문에 해당 이미지의 URL를 입력할 수 있는 누락문자 관리기를 개발하였다.

본 연구에서 개발된 한글대장경 90권본에 대해 인터넷을 통해 검색하고자 한다면 다음의 URL을 이용하면 된다. URL은 'http://ebti.dongguk.ac.kr'이다. 향후 연구 과제는 확장한자에 대한 처리를 위해 기존에 작업했던 누락문자를 찾아 확장한자를 입력하는 작업이 필요하다. 누락문자가 이미지 파일이기 때문에 화면상 불균형이 생기는 문제를 해결할 수 있다. 그리고 한글대장경의 더욱 많은 부분을 빠른 시일 내에 전산화하는 일이 필요하다. 그리고 데이터베이스에 저장된 내용을 검색을 위해 더 다양한 검색 기법의 도입이 필요하다.

참고문헌

[1] The Unicode Consortium, The Unicode Standard, Version 2.0, Addison Wesly, 1996.

[2] 황기태역, 어드밴스 윈도우 NT, 도서출판 대림, pp. 757-784, 1995.

[3] 김응철, "고려장경 및 한자정보전산화에 관련한 문제제기", http://members.iWorld.net/hederein/menu22/Kim.html

[4] 심재룡, "정보화 사회와 불교 전산화", http://members.iWorld.net/hederein/menu22/Dogam32.html

[5] 강석진, "팔만사천대장경 전산화를 위한 제언, 한자위주 문헌의 워드프로세서 데이터베이스, 탁상출판 시스템 개발을 위해", http://members.iWorld.net/hederein/menu22/Kang.html

[6] 혜묵스님, “세계의 불교자료 전산화 계획과 고려대장경 전산화를 위한 몇가지 문제들”,
http://members.iWorld.net/hederein/menu22/Hye. html
[7] 종림스님, “팔만대장경 전산화 추진경과와 이후 계획”,
http://members.iWorld.net/hederein/menu22/
[8] 노용균, “불전 전산화와 SGML”,
http://members.iWorld.net/hederein/menu22/Dogam42.html.
[9] 이규갑, “고려대장경 전산화에 있어서 이체자의 처리 문제”,
http://members.iWorld.net/hederein/menu22/Yi.html
[10] 정주원, “한글 코드에 대하여”, 1995,
http://simac.kaist.ac.kr/~jwjung/seminar/hangul-i18n/ko-code.html.
[11] Urs App, “A Look at the Korean Tripitaka Input Project”,
http://www.iijnet.or.jp/iriz/irizhtml/ebit/samsung.htm
[12] Urs App, “The Importance of Markup”,
http://www.iijnet.or.jp/iriz/irizhtml/maketext/foguang.html
[13] Urs App, “A Look at the Korean Tripitaka Input Project”,
http://www.iijnet.or.jp/iriz/irizhtml/ebit/samsung.htm.
[14] ISO 8879:1986, Standard Generalized Markup Language, 2nd Edition.
[15] ISO/IEC 10646-1:1993, “Information Technology - Universal Multiple-Octet Coded Character Set(UCS) - Part I : Architecture and Basic Multilingual Plane”.
[16] Public Unicode Font,
ftp://www.ifcss.org/ftp-pub/software/fonts /unicode.
[17] 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.
[18] Unicode Support in Win32, Microsoft Developer's Network, 1997.

[19] Unicode enabling, Microsoft Developer's Network, 1997.
[20] True Type and Unicode,
http://truetype.demon.co.uk:80/unicode.htm
[21] CJK Codes-Unicode/ISO-10646 Unicied "Ideographs",
http://www.mit.edu:8001/afs/athena.mit...r/a/k/akbar/www/Unicode-ideographs.html.
[22] "Installing Bitstrean Cyberbit Version 1.1",
http://www.bitstrean.com/cyberbit.html.
[23] Panorama, http://www.softquad.com, Softquad 사.
[24] 인터넷으로 만나는 불교,
http://members.iWorld.net/hederein/menu23/Pogyu121.html
[25] 김숙자, SGML의 모든 것, 성안당, 1997.
[26] 장회창, 현득창, 이수연, SGML 가이드, 사이버출판사, 1997.
[27] 현득창, 임광택, 이수연, "SGML 기본 파서를 이용한 SGML 문서 편집기의 구현", 한국정보과학회, 정보과학회 눈문지, Vol 25, No. 1, 1998.
[28] 대장경학술용어연구회, 대정신수 대장경소인, 제1권, 대장경학술용어연구회, 1975.
[29] 한국불교신문, 시방시계 1월 27일자, 현대불교신문사, 1999.
[30] 김정숙, 유응구, 이용규, 이금석, 홍영식, "유니코드를 기반으로 한 한자 입력 시스템 개발", 한국정보과학회, '98 춘계학술발표논문집, Vol 25, No. 1, 1998.
[31] 김태규, 유병인, 한인, 이용구, 이금성, 홍영식, "유니코드 한자 지원 문법 지시적 SGML편집기의 설계 및 구현", 한국정보과학회 '98 춘계학술발표논문집, Vol. 25, No. 1, 1998.
[32] 주신탁, 설승진, 이용규, 이금석, 홍영식, "유니코드와 SGML을 이용한 한국 고문헌 데이터베이스 구축", 한국정보과학회, '98 춘계학술발표논문집, Vol. 25, No. 1, 1998.
[33] 조은정, 신훈철, 이용규, 이금석, 홍영식, "웹에서의 한국 고문헌

검색시스템", 한국정보처리학회, '98 춘계학술발표논문 CD, 1998.
[34] 홍영식 외 13인, "웹에서의 한국 고문헌 관리 및 검색 기술 개발", 정보통신부, '97 초고속 정보통신 응용기술개발사업 최종 연구 보고서, 7월 1998.
[35] 이용규, 홍영식, 이금석, 김정숙, 한인, 설승진, 신훈철, "한국 고문헌 데이터베이스", 동국대학교, 동국논총 제 三十七 , 12월 1998.
[36] 한보광 외 6인, "한글대장경 전산화", 동국대학교 전자불전연구소, 전자불전, 한글대장경의 성립과 전개, 제4집, 2002.

키워드(Keyword)

한글대장경, 한글대장경 검색 시스템, 한글 대장경 전산화, 유니코드, XML
Hangul Tripitaka, Hangul Tripitaka Retrieval System, Hangul Tripitaka Digitalization, Unicode, XML

Abstract.

The Present State of the 3rd Hangul Tripitaka Digitalization Project

Jin Hong No*, Hyun Woo Koo*, Eung Gu Ryu*, Sung Eun Park*, Young Hee Park**, Yong Kyu Lee*, Keum Suk Lee*, Young Sik Hong*, Bo Kwang Han**

*Dept. of Computer Engineering, Dongguk University
**Dept. of Seon Studies, Dongguk University

This research aims for constructing the retrieval system by digitizing a quantity of the 30 Hangul Tripitaka books in the 3rd Hangul Tripitaka Digitalization Project.

By revising and digitalizing the Hangul Tripitaka which is a Korean version of the Tripitaka Korean we can input, store in database, and search the archaic documents through the Internet. Since the archaic documents of the Hangul Tripitaka includes extension characters of Chinese origin, missing characters and special characters, etc, we use Unicode and make the image fonts that cannot be represented by Unicode. And we apply XML for the efficient representation of document structure and the retrieval. So people can search the same contents as the archaic documents. Moreover we

developed the search engine which provides the efficient and easy search method, the archaic documents saved as Unicode can access from the whole world using the Internet.

The retrieval system developed in this research uses Microsoft SQL Server and IIS(Internet Information Server) on Windows 2000 Server.