유니코드와 SGML을 이용한 한국불교전서 데이터베이스 구축

Construction of Hanguk Pulgyo Chonso Database Using Unicode and SGML

설승진*, 이용규*, 이금석*, 홍영식*, 한보광**

Seung Jin Sul, Yong Kyu Lee, Keum Suk Lee, Young Sik Hong, Bo Kwang Han

*동국대학교 컴퓨터공학과

Dept. of Computer Engineering, Dongguk University

**동국대학교 선학과

Dept. of Seon Studies, Dongguk University

요약

한국 고문헌은 대부분 한자를 사용하여 기록되었으며, 이러한 한자들은 본자(本字)와 뜻은 같지만 모양이 틀린 이체자(異體字)와 오자(誤字)나 탈자(脫字)로 간주되는 파자(破字) 등을 포함하고 있으므로 고문헌의 입력이나 저장 관점에서 여러 문제점을 갖는다. 이러한 한국 고문헌의 효과적인 입력과 저장을 위해 본 연구에서는 유니코드(Unicode)를 사용하였으며, 고문헌의 문서 구조 표현과 효율적인 검색을 위해 SGML을 적용하였다.

유니코드를 사용하여 입력된 고문헌을 저장하는 방법은 일반적인 ASCII 코드나 완성형 코드를 사용하는 문헌과는 많은 차이가 있다. 본 논문은 유니코드 문서의 효율적인 저장 방법을 제시하고 SGML 문서 구조 분석, 저장 방법 및 색인 생성 방법에 대해 기술한다. 본 연구는 자체 개발된 유니코드를 사용하는 SGML 문서 편집기를 통해 입력된 한국 불교 전서를 대상으로 하였으며, 윈도 NT에서 운영되는 마이크로소프트사의 SQL Server를 데이터베이스 관리 시스템으로 사용하여 구현하였다.

1. 서론

이미 오래 전부터 미국과 중국어권 나라들을 중심으로 고문헌 전산화, 특히 불교경전의 전산화에 대한 연구가 활발히 진행되어 왔으며, 연구 목적뿐만 아니라 상업적 측면에서의 여러 산물들이 발표되었다. 또한 문화 유산의 가치를 드높인다는 측면에서 다양한 결과물들이 속속 발표되고 있다. 이에 비해 우리 나라는 찬란한 5천년의 역사 속에서 탄생한 귀중한 고문헌 자료의 전산화에 대해서는 등안시 해온 것이 사실이며, 일부 연구소에서만 몇 가지 작업을 진행중일 뿐이다[1,3,4,6].

한국 고문헌 전산화는 다른 중국어권 나라와 마찬가지로 여러 가지 극복해야할 문제점을 가지고 있다. 우선 입력 관점에서 바라보면, 대략 2~3만자에 이르는 한자를 사용하여 기록된 고문헌을 제대로 처리하기 위한 폰트와 코드 문제를 말할 수 있다. 우리 나라 실정은 아직까지 2~3만자나 되는 한자를 지원하는 코드 체계나 폰트를 개발하지 않은 상태이다. 일본의 JIS 코드 체계나 대만의 BIG5 코드 체계[7]는 이러한 문제를 해결하기 위한 노력의 결실이라고 할 수 있다.

고문헌에 사용된 한자 코드에 대한 문제는 이체자와 관련이 깊은데, 이체자란 뜻과 음은 동일하지만 다른 형태로 기록되는 한자를 의미한다[4]. 이체자는 오자나 탈자를 포함하는 파자와 함께 추가적인 코드와 폰트를 요구하므로 고문헌 전산화에 있어 필히 극복해야할 문제이다.

본 연구에서는 고문헌 전산화와 관련된 폰트 및 코드 문제를 유니코드(Unicode)[5, 8, 11]와 인터넷에 공개된 폰트[12]를 사용하여 해결하였으며, SGML(Standard Generalized Markup Language)과 마크업(markup)이 가능하며 유니코드를 기본 코드로 사용하는 한문 문서 편집기를 자체 개발하여 고문헌 전산화에 필요한 입력 작업을 수행하였다[14].

유니코드는 한자 입력 문제에 있어 해결책을 제시하는 반면, 검색 측면에서는 또 다른 문제를 야기하였다. 이것은 유니코드가 기존의 1바이트

형 문자가 아닌 2바이트형 문자라는 점에 기인한다. 따라서 본 논문에서는 적절한 전처리 과정을 통해 유니코드로 저장된 고문헌을 데이터베이스에 저장하고 색인 구축 과정을 거쳐 효율적인 검색이 가능하도록 하는 구현 과정을 기술한다.

본 논문은 2장에서 유니코드와 SGML에 대해 소개하고, 3장에서 데이터베이스 구축과 색인생성 과정에 대해 설명하며, 4장에서 구현 결과 및 분석, 그리고 5장에서 결론 및 향후 과제에 대해 기술한다.

2. 유니코드와 SGML

1) 유니코드

유니코드는 1980년대 말부터 미국의 몇몇 회사들에 의해 창설된 유니코드 콘소시움에서 만들어지기 시작하여 1990년대에 이르러서는 2바이트 문자 코드 체계를 정립하였다. 이러한 움직임에 ISO에서는 1980년대 중반부터 거의 10년에 걸쳐 UCS(Universal multi-octet coded Character Set) 표준화 작업을 수행하였으며, 1992년 6-7월에 열린 ISO/IEC JTC1/SC2/WG2 제22차 회의에서 ISO-10646-1을 국제 표준으로 발표하였다. UCS-2 코드 체계는 BMP(Basic Multi-lingual Plane)라 하며, 총 65,536개의 코드를 지원한다. 현재 그 가운데 약 2/3에 해당되는 코드가 이미 할당된 상태이며, 코드의 분할 영역은 (그림 1)과 같다. 고문헌 전산화를 위해 사용한 코드 영역은 '4E00'부터 '9FA5'까지이며 약 2만여개의 한자를 위한 코드가 정의되어 있다.

```
0000 - 1FFF : A-ZONE    ; Alphabets
2000 - 2FFF : A-ZONE    ; Symbols and Punctuation
3000 - 4DFF : A-ZONE    ; CJK Auxilary
4E00 - 9FFF : I-ZONE    ; CJK Unified Ideographs
A000 - DFFF : O-ZONE    ; Reserved
E000 - FFFD : R-ZONE    ; Restricted use
```

(그림 1) 유니코드의 코드 할당 영역

2) SGML

SGML은 ISO에 의해 발표된 국제 표준(ISO-8879[10])으로서 서로 다른 기종간에 효율적인 문서 전공을 가능하게 하며, 문서의 논리적 구조를 구조적으로 표현할 수 있도록 한다. 일반적인 SGML 문서는 다음과 같은 3부분으로 이루어진다.

- SGML 선언부(SGML Declaration)
- 문서 형 정의부(DTD; Document Type Definition)
- SGML 문서 실체(SGML Document Instance)

SGML은 독립성과 문서 구조에 대한 명확한 기술이 가능하기 때문에 미국 출판협회나 옥스퍼드 대학 출판부 등에서의 문헌 전산화에 널리 적용되고 있다. (그림 2)는 SGML을 이용하여 입력된 불교전서의 일부분을 예로 나타낸 것이다

```
<PAGE PAGENUM=4-528></PAGE>
<DAN DANNUM=1></DAN>
<JMOK>1大覺國師文集卷第一</JMOK>
序
新集圓宗文類序
新編諸宗敎藏摠錄序
刊定成唯識論單科序
八師經後序
消災經直釋詳定記

新集圓宗文類序
大 2□嚴之爲敎也 一眞妙蘊滿藏雄
詮 3□遍照之心源?普賢之行海誠
生靈之大本稱性之極談者歟 自景煥
龍宮風行 4□李聖賢繼踵述作連? 有
終南租師祉順 5□者歎曰 大哉 <KEYWORD>法界</KEYWORD>
之經也 自非登地 何6□披其文 見其
法哉 吾設其門以示之 於是 著法界
觀門 以授高弟<KEYWORD>知儼</KEYWORD>尊者 儼師得之
變之爲五敎 演之爲十玄 及乎賢首
<券一第一張> 組述於前 淸涼 憲章於
後 始可謂能事畢矣 故講大經者 咸
以儼藏淸涼三家義? 永爲標準 而旁
用諸家補焉 自我海東浮石尊者 求
...
```

(그림 2) 한국 고문헌의 SGML 문서 실체

3. 한국불교전서 데이터베이스 구축

1) 개요

SGML을 사용하여 마크업된 문서를 저장하기 위해서는 파일 시스템, 관계형 데이터베이스, 전문 데이터베이스(Full-text Database), 객체지향 데이터베이스 등이 사용될 수 있다. 본 연구에서는 가장 일반적으로 사용되는 관계형 데이터베이스인 마이크로소프트사의 SQL Server를 사용하는데, 관계형 데이터베이스는 비교적 높은 안정성과 신뢰성, 그리고 호환성을 제공하며, 무엇보다도 문서에 대한 중앙 집중적인 관리가 수월한 장점을 제공한다.
고문헌의 한자를 표현하기 위한 코드 체계는 2바이트를 사용하여 하나의 문자를 표현하는 유니코드 문자 체계이므로 일반적인 문서 파일과는 그 구조가 다르다. 특히, 문자형 데이터, 문자열 데이터, 그리고 유니코드 파일을 다루기 위해서는 일반 파일 처리와는 다른 방식을 사용해야 한다. 이러한 기술적인 내용에 대해서는 다음절에서 자세히 설명한다.

2) 유니코드 문서의 코드 구조

유니코드 문서를 효율적으로 저장하기 위해서는 우선 유니코드 문서의 구조를 파악해야 한다. (표 1)은 간단한 유니코드 문서를 파일로 저장했을 경우, 저장 코드 형태를 보여준다.

유니코드 문서의 내용	<JMOK>1大覺國師文集卷第一
파일 저장 형태	FF FE 3C 00 4A 00 4D 00 4F 00 4B 00 3E 00 31 00 27 59 BA 69 0B 57 2B 5E 67 65 C6 96 77 53 2C 7B 00 3E

(표 14) 유니코드 문서의 내용과 저장된 코드의 예

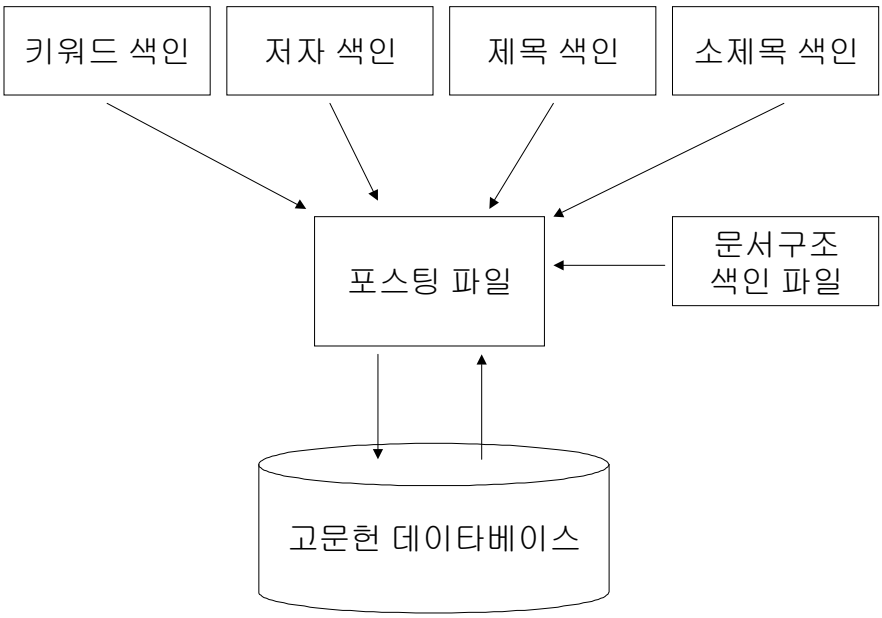
(표 1)에서 알 수 있듯이 유니코드 문서는 해당 문서가 유니코드임을 표시하기 위해 2바이트 특수 코드(FFFE)를 사용하며, 2바이트씩을 이용하여 하나의 문자를 나타낸다. 문서의 저장을 위해 중요한 특징은 기존의 ASCII 문자 코드 집합에 속하는 문자들은 기존의 ASCII 문자 코드에 '00'이 추가되어 2바이트 유니코드 체계로 변환된다는 점이다. 이것은 또한 ASCII 문자 코드에서의 문자열끝 문자인 '0D'와 줄바꿈 문자인 '0A' 경우에도 적용되어 각각 '0D00'과 '0A00'으로 바뀌게 된다.

3) 데이터베이스 저장을 위한 데이터형

일반적인 DBMS는 여러 종류의 데이터를 저장할 수 있도록 다양한 자료형을 제공한다. 이러한 자료형에는 대표적으로 문자와 문자열을 저장할 수 있도록 마련된 데이터형이 존재하지만, 유니코드 문서의 저장에는 적절하지 않다. 유니코드를 문자형으로 저장할 경우, 문서 중간에 존재하는 '00' 코드를 문자열의 끝으로 오인하게 되는 경우나 단일인용부호(')나 이중인용부호(")에 해당하는 코드가 존재하는 경우에는 정확한 문서 저장이 불가능하다. 이러한 문제들은 현재의 DBMS가 유니코드를 지원하지 않기 때문이다.

4) 데이터베이스 구성

한국불교전서의 저장 및 검색을 위한 데이터베이스 구성은 (그림 3)과 같다. 데이터베이스 구성은 차후의 검색 방법을 고려하여 결정하였다. (그림 4)에서 포스팅 파일(posting file)에 해당하는 것이 문서구조 저장 테이블이며, 각각의 태그별 색인이 생성된다. 색인 파일을 생성하는 방법은 다음절에 상세히 기술된다.



(그림 3) 한국불교전서 데이터베이스의 구성

일반적으로 SGML을 이용하여 마크업된 SGML 문서를 검색하는 방법은 문서 구조를 통한 검색과 문서 내용을 통한 검색 등 두 가지로 볼 수 있다[2]. 문서 구조를 통한 검색은 목차별 검색 혹은 디렉토리 검색이라고 말할 수 있으며, 문서의 상위 구조로부터 하위 구조로 탐색해나가는 방식으로 원하는 검색 결과를 찾는 방법이다.

이러한 검색 방법을 지원하기 위해 입력 SGML 문서를 파싱하여 태그 종류와 속성 데이터 그리고 태그의 문서내 위치 등 태그 관련 정보를 추출하고 이를 적절한 구조로 데이터베이스에 저장해야 한다.

본 연구에서는 유니코드를 사용한 입력 SGML 문서에서 태그 정보를 추출하기 위해 유니코드 문서를 ASCII 문서로 변화하는데, 이는 태그 정보를 추출하는 과정에서는 원문 데이터의 내용은 중요하지 않다는 점에 착안했다. 즉, 유니코드 문서 내의 태그들은 영문자를 사용하는 것이 일반적이므로 기존의 ASCII 코드에 '00'이 추가되는 형태로써 기록된다. 그러므로 입력 유니코드 문서에 대해 한 문자를 이루는 2바이트 코드 중의 후위 1바이트를 제거하면 원문 데이터의 의미는 사라지더라도 태그와 그 위치에 대한 정보를 완벽하게 추출할 수 있다. 이렇게 추출된 태그 정보는 (태그종류, 번호 데이터, 문서내 위치)의 형식으로 마스터 테이블에 저장되고 이 테이블에서 각각의 태그별, 번호 데이터별, 문서 내 위치별 뷰(view)를 구성하게 된다. 이러한 테이블과 뷰를 사용하여 문서구조를 통한 검색이 가능하게 된다.

우선 원문의 저장은 다음과 같은 테이블 구조를 이용하여 데이터베이스에 저장된다.

테이블 명: edocdata			
칼럼명	데이타형(길이)	제약사항	비고
nlinenum	integer(4 byte)	PK, NN	키워드 독음 번호
sdocdata	varbinary(255 byte)	NN	키워드 독음 (완성형)
npagenum	integer(4 byte)	NN	페이지 번호
npageline	integer(4 byte)	NN	페이지에서의 라인 번호
ndannum	integer(4 byte)	NN	단 번호

(그림 4) 원문 저장을 위한 테이블 구조

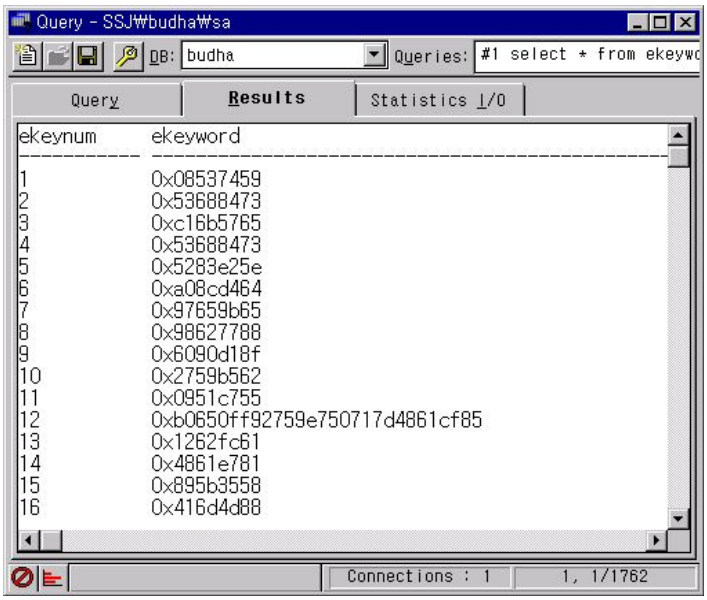
5) 색인 구성

문서 내용을 통한 검색의 경우, 키워드와 원문을 대상으로 색인 테이블을 구성하는 방법을 적용하였다. 이러한 방법에서 가장 중요한 작업은 원문의 구문을 분석하여 키워드를 자동 추출하고 이를 바탕으로 색인을 구축하는 것인데, 한문 문헌의 경우 한자 키워드를 의미론적으로 분석하여 자동 추출하는 것이 불가능하다. 따라서 한자를 사용하는 한국불교전서의 경우, 미리 키워드를 수작업으로 선정한 후, 특정 태그를 이용하여 본문내에 이를 표기하는 방법을 사용하였다. 이렇게 태그로 마크업된 키워드는 쉽게 추출될 수 있으며, 이를 바탕으로 색인을 생성할 수 있었다. (그림 5)는 추출한 유니코드 키워드를 저장하기 위한 데이터베이스 테이블 구조이다.

테이블 명: eKeyword			
칼럼명	데이터형(길이)	제약사항	비고
eKeynum	integer	PK, NN	키워드 번호
eKeyword	varbinary(40)	NN	유니코드 키워드

(그림 5) 유니코드 키워드 저장을 위한 테이블 구조

원문의 키워드 태그를 바탕으로 추출된 유니코드 키워드가 실제 저장된 화면은 (그림 6)에 나타내었다. (그림 7)은 추출된 유니코드를 메모장으로 출력한 결과이다.

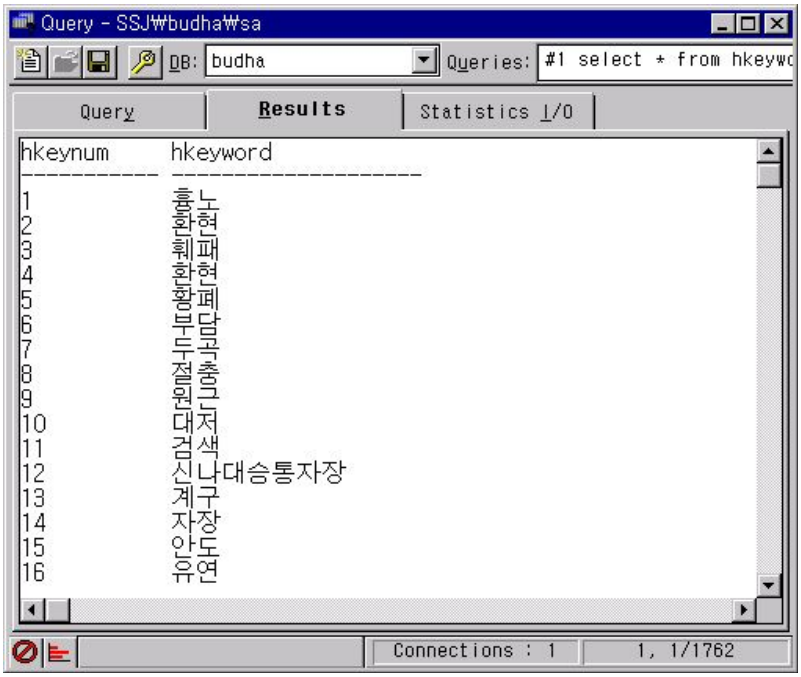


(그림 6) 저장된 유니코드 키워드

(그림 7) 추출된 키워드

이렇게 추출된 유니코드 키워드는 검색시 독음처리를 위해 독음으로 변

환해야 한다. 이러한 변환 과정은 훈글의 한자를 한글로 변환하는 기능을 이용하여 수작업으로 진행된다. 일단 유니코드 키워드가 한글 독음으로 변화되면 다시 데이터베이스 테이블에 저장하고 검색 과정에서 사용하게 된다. (그림 8)은 변환되어 데이터베이스에 저장된 한글 독음을 보여준다.



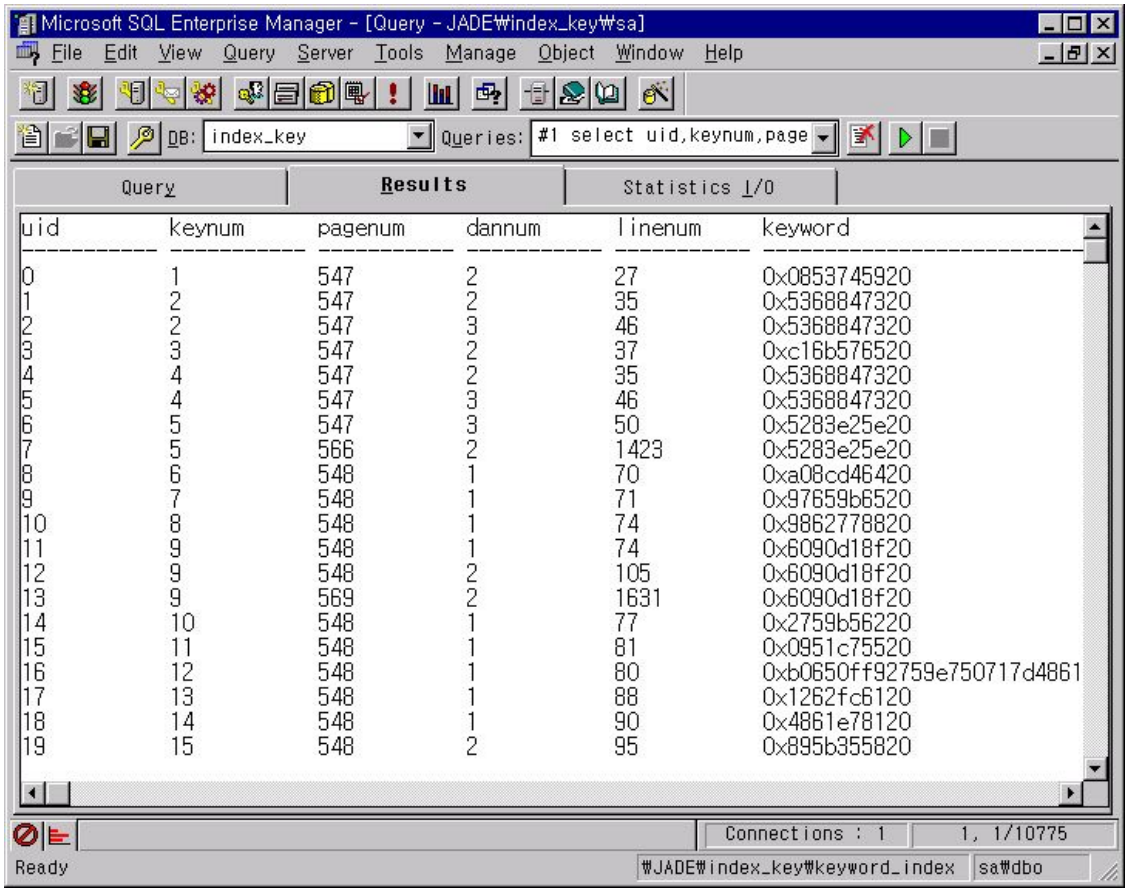
(그림 8) 테이블에 저장된 키워드 한글 독음

효율적인 검색을 위한 인덱스 생성은 원문과 추출된 키워드 파일을 이용하여 해당 키워드의 원문 내 위치를 조사하게 되며, 이때 (그림 9)와 같은 테이블 구조를 사용하여 인덱스를 표현한다.

테이블 명: keyword_index			
칼럼명	데이터형(길이)	제약사항	비고
uid	integer(4 byte)	PK, NN	일련번호
keynum	integer(4 byte)	NN	키워드 번호
keyword	varbinary(255 byte)	NN	키워드
pagenum	integer(4 byte)	NN	PAGE 번호
dannum	integer(4 byte)	NN	DAN 번호
linenum	integer(4 byte)	NN	LINE 번호

(그림 9) 인덱스 저장을 위한 테이블 구조

실제로 작성이 완료된 인덱스 테이블의 예는 (그림 10)에 나타내었다. 앞서 생성한 원문 및 키워드 테이블과 인덱스 테이블은 실제 웹에서의 검색시 사용된다.



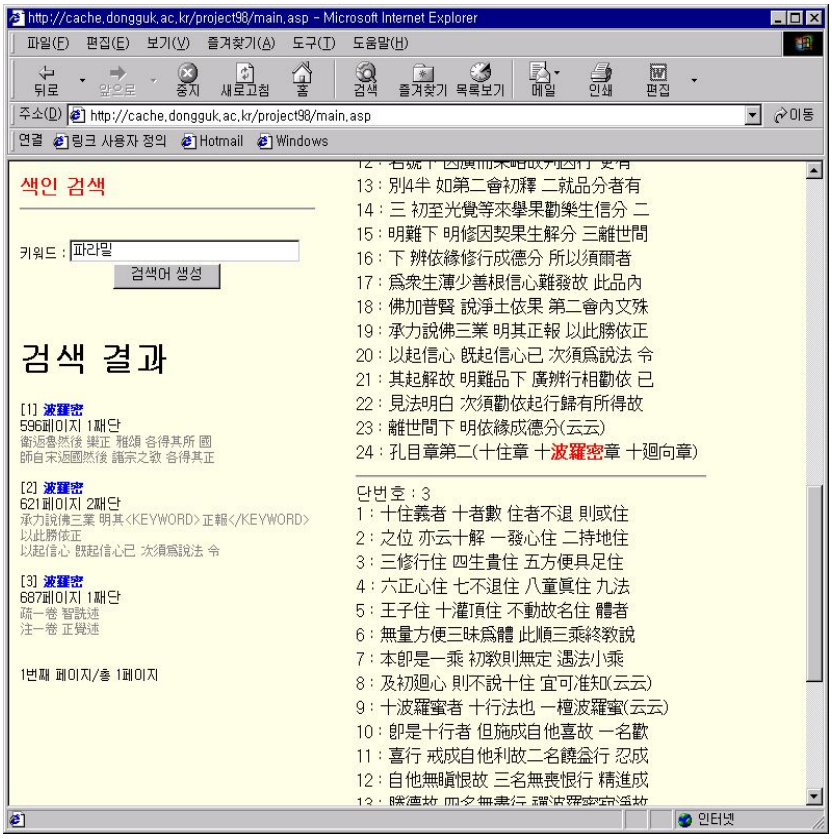
(그림 10) 생성된 인덱스 테이블

4. 구현 결과 및 분석

본 연구의 실험 환경으로는 주기억장치의 용량이 64MB이며 인텔 펜티엄 CPU를 탑재한 PC를 기반으로, 운영체제는 윈도우 NT, DBMS는 마이크로소프트사의 SQL Server 6.5 버전을 사용하였다. 데이터베이스 저장 및 색인 생성을 위한 클라이언트 프로그램의 작성은 마이크로소프트사의 비주얼 C++ 5.0 버전을 사용하였다.

입력 데이터는 동국대학교에서 편찬한 한국 불교 전서의 일부분을 자체 개발한 문서 편집기를 이용하여 입력하고 마크업하여 사용하였다. 약

400페이지에 해당되는 원문의 저장 및 인덱스 생성 과정에서 가장 큰 문제점으로 대두된 사항은 저장 및 인덱스 생성 시간이었다. 이는 유니코드의 특성상 저효율의 비교 연산을 사용할 수밖에 없기 때문에 야기되는 것으로 알고리즘의 개선을 통해 해결점을 모색 중이다.
구축된 한국불교전서 데이터베이스는 웹을 통해 접속하여 키워드를 이용한 자유로운 검색이 가능하였다[15]. 웹을 통한 검색 결과는 (그림 11)과 같다.



(그림 11) 웹을 통한 키워드 검색 화면

한국불교전서 데이터베이스 구축 과정에서 가장 큰 문제점은 현재의 DBMS가 유니코드를 지원하지 않는다는 점이었다. 이러한 문제는 여러 DBMS 회사들이 향후 유니코드를 지원할 것을 계획하고 있고 유니코드

를 지원하는 객체지향 DBMS를 사용함으로써 해결될 수 있을 것이다. 또한 유니코드로 표현할 수 없는 이체자 문제에 대한 해결책에 대한 연구도 진행되어야 할 것이다.

5. 결론 및 향후 과제

세계적으로 고문헌, 특히 불전 전산화에 대한 연구가 활발히 진행되고 있는 상황에서 한자의 입력 및 저장과 관련된 많은 문제점들이 알려지고 다양한 해결 방법이 연구되고 있다. 우리 나라의 경우, 해인사 대장경 연구소의 고려대장경 전산화가 대표적인 고문헌 전산화의 예일 뿐이며 아직도 한자 코드나 이체자 문제에 대한 명확한 해결책을 보유하고 있지 않다.

본 연구에서는 유니코드와 SGML을 이용하여 입력된 한국불교전서 데이터베이스 구축에 대한 문제점을 제시하고 해결 방안을 고안하였다. 현재 계속적인 과제 진행을 통해 더 많은 개선안이 고려되고 있으며 시범적으로 웹을 통해 서비스되고 있다.

더 효율적인 저장 방법 및 색인 생성 방법에 대한 연구가 진행 중이며 이체자의 처리 방안에 대해서도 집중적인 연구가 진행되고 있다.

[참고문헌]

[1] 김응철, “고려대장경 및 한자정보전산화에 관련한 문제제기,” [http://members.iWorld.net/hederein/menu22/Kim.html]

[2] 김규태, 현득창, 이수연, 정광철, “관계형 데이터베이스를 이용한 SGML 문서 처리”, 정보과학회논문지, 제3권, 제3호, 1997. 6.

[3] 노용균, “불전전산화와 SGML,” [http://members.iWorld.net/hederein/menu22/Dogam42.html]

[4] 이규갑, “고려대장경 전산화에 있어서의 이체자 처리문제,” [http://members.iWorld.net/hederein/menu22/Yi.html]

[5] 정주원, “ISO/IEC-10646 Universal Multiple-Octet Coded Character Set (UCS)에 대해서,” [http://simac.kaist.ac.kr/~jwjung/seminar/hangul-i18n/iso10646.html]

[6] 혜묵스님, “세계의 불교자료 전산화 계획과 고려대장경 전산화를 위한 몇 가지 문제들,” [http://[http://members.iWorld.net/hederein/menu22/Hte2.html]

[7] Christian Wittern, "Chinese Character Codes: an update," [http://www.iijnet.or.jp/iriz/irizhtml/multling/codes.html]

[8] "The Unicode Standard, Version 2.0," 1996, The Unicode Consortium

[9] Urs App, "A Look at the Korean Tripitaka Input Project," [http://www.iijnet.or.jp/iriz/irizhtml/ebti/samsung.html]

[10] ISO 8879:1986, "Standard Generalized Markup Language," 2nd Ed.

[11] ISO/IEC 10646-1:1993, "Information Technology - Universal Multiple-Octet Coded Character Set (UCS) -Part1: Architecture and Basic Multilingual Plane"

[12] Public Unicode Fonts [ftp://www.ifcss.org/ftp-pub/software/fonts/unicode/]

[13] “대정신수 대장경색인,” 제1권, 대장경학술용어연구회, 일본, 1975

[14] 한인, 이용규, 이금석, 홍영식 외, “유니코드 한자 지원 문법지시적 SGML 편집기의 설계 및 구현,” ‘98춘계학술발표논문집, 한국정보과학회

[15] 신훈철, 이용규, 이금석, 홍영식 외, “웹에서의 한국 고문헌 검색 시스템,” ‘98춘계학술발표논문집, 한국정보과학회