한국불교전서 데이터베이스의 분석

이 용 규*

목 차

요 약

이 논문에서는 한국불교전서 전산화 사업을 통해 구축된 데이터베이스 시스템을 간략히 소개하고, 데이터베이스에 저장된 한국불교전서 원문과 인덱스 관련 데이터를 분석하여 데이터베이스 테이블의 규모와 각 요소별 구성 비율, 그리고 키워드 테이블의 크기와 사용 현황 등 여러 가지 관심 있는 통계 분석 결과를 제시하였다.

* 동국대학교 컴퓨터공학과 교수

1. 서 론

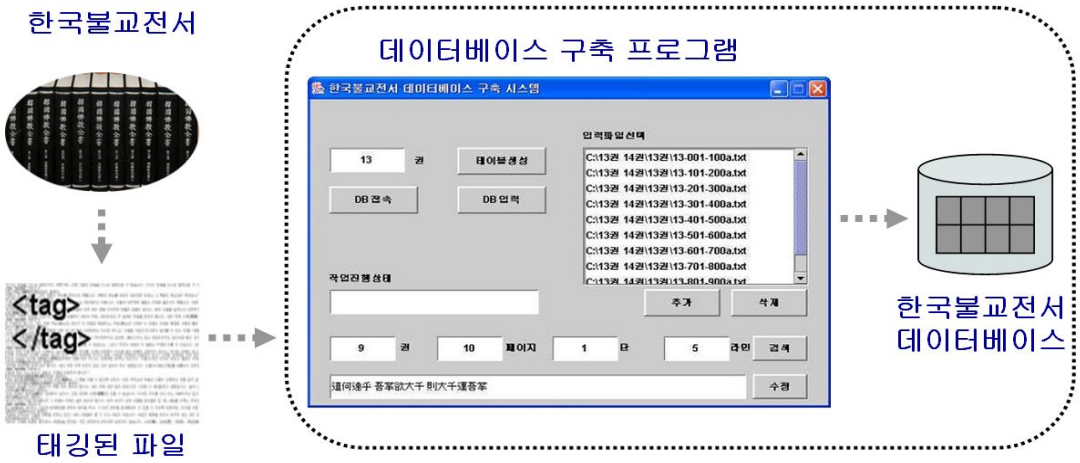
한국불교전서 전산화 사업은 1999년 7월 1일부터 2000년 6월 30일까지 1년간 수행된 한국불교전서 전산화 시범사업을 바탕으로 2000년 7월 1일부터 2007년 6월 31일까지 총 7년간 수행되었으며 현재까지 출간된 한국불교전서 제1책부터 제14책까지 14책 모두를 전산화 하였다[1][2][3].

전산화를 통해 한국불교전서의 텍스트와 인덱스 데이터베이스가 구축되었으며, 이 논문에서는 이 데이터베이스를 분석하여 여러 가지 통계 자료를 제시하고자 한다. 이처럼 데이터베이스의 분석을 통해 관심 있는 통계 현황을 추출할 수 있는 것 또한 전산화의 부수적인 효과라고 할 수 있다.

2. 한국불교전서 데이터베이스 시스템

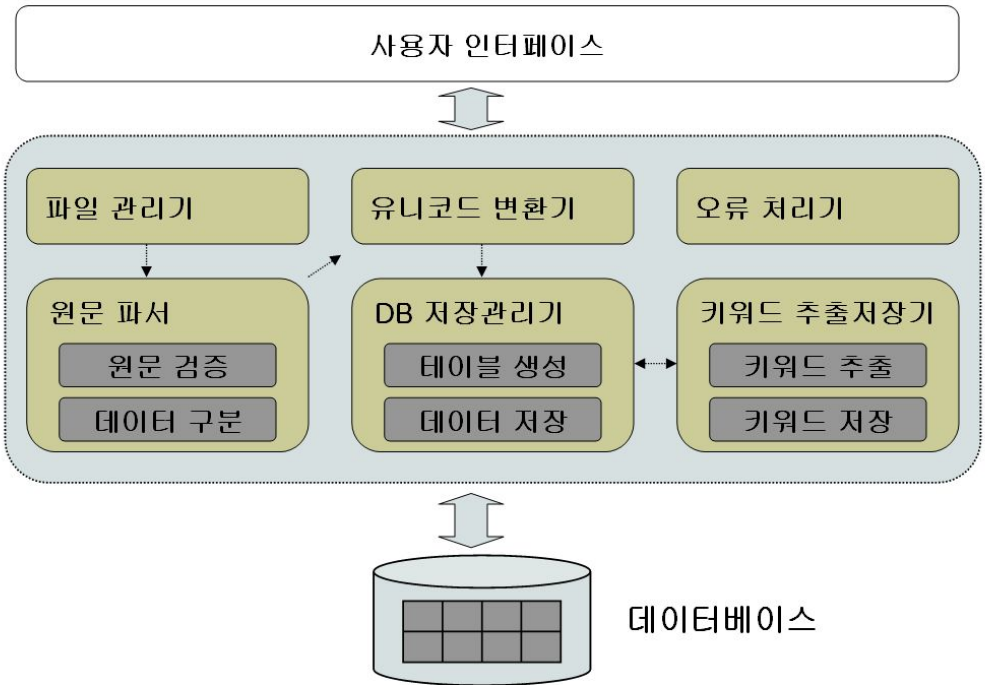
한국불교전서 데이터베이스는 마이크로소프트의 SQL Server 2000 DBMS를 사용하여 구축된 관계형 데이터베이스이다.

[그림 1]에 도시된 바와 같이 데이터베이스에 저장하기 위해서 먼저 한국불교전서 원문을 XML을 사용하여 마크업 하고, 이를 데이터베이스 구축 프로그램을 이용하여 저장한다.



[그림 1] 한국불교전서 데이터베이스 구축 시스템

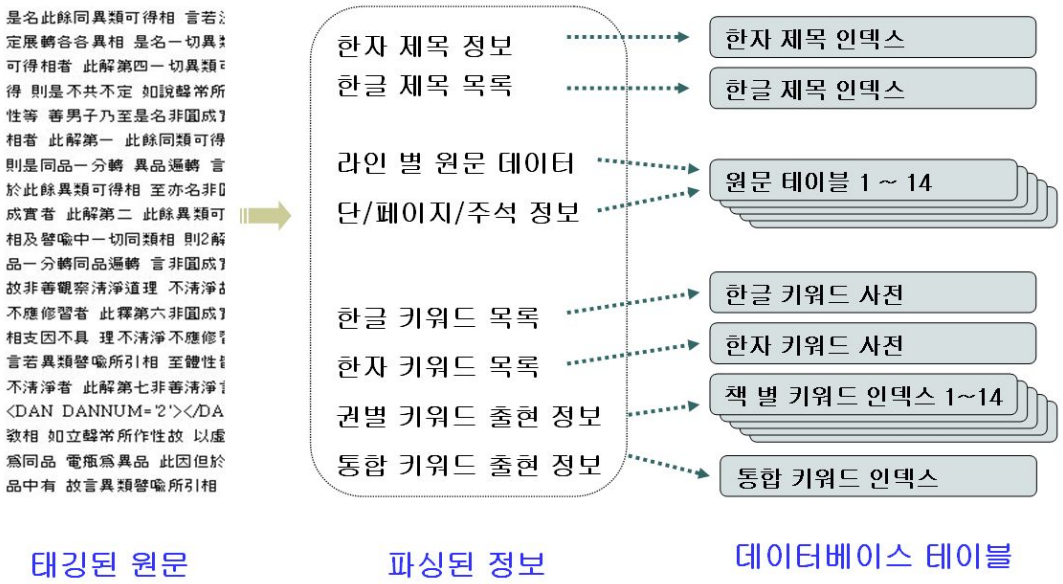
데이터베이스 구축 프로그램은 [그림 2]에서 보는 바처럼 몇 가지 모듈들로 구성되고, 파일관리기, 파서, 유니코드 변환기, DB 저장 관리기, 오류처리기, 키워드 추출 저장기 등이 유기적으로 동작하며 텍스트를 비롯하여



[그림 2] 한국불교전서 데이터베이스 구축 프로그램

검색에 필요한 정보를 추출하여 데이터베이스의 테이블에 저장한다.

마크업이 된 원문 텍스트를 파싱하여 검색에 필요한 정보를 추출하여 테이블에 저장하는 과정이 〔그림 3〕에 도시되어 있다. 제목 정보, 위치 정보, 키워드 정보 등이 추출되어 해당 테이블에 저장된다.



[그림 3] 한국불교전서 데이터베이스 테이블 구성

한국불교전서 데이터베이스의 구성이 〔표 1〕에 요약되어 있다. 원문을 1책부터 14책까지 각각 나누어 저장하는 원문 테이블이 있으며, 원문의 키워드 검색이 가능하도록 한자 키워드 사전, 한글 키워드 사전, 그리고 키워드 출현 정보를 저장하는 키워드 인덱스 테이블이 있다. 한글 및 한자로 제목 검색이 가능하도록 한글과 한자 각각의 제목 인덱스 테이블을 가지고 있다.

[표 1] 한국불교전서 데이터베이스 주요 테이블

분류	데이터베이스 테이블	테이블 설명
원문	원문테이블 1~14	책 별 원문 데이터 저장
키워드	한자 키워드 사전	키워드에 대한 유니코드와 첫 글자의 획수 저장

인덱스	한글 키워드 사전	키워드에 대한 한글독음과 첫 글자의 획수 저장
	통합 키워드 인덱스	전책에 대한 키워드 출현 정보 저장
	책 별 키워드 인덱스 1~14	책 별 키워드 출현 정보 저장
제목 인덱스	한글 제목 인덱스	한글 제목 검색을 위한 제목 정보 저장
	한자 제목 인덱스	한문 제목 정보 저장

※ 임시 테이블을 포함하여 총 50개의 테이블로 구성됨

3. 한국불교전서 데이터베이스 분석

한국불교전서 데이터베이스는 크기가 617MB로 구성 비율은 원문이 26.0%, 제목 인덱스가 0.1%, 키워드 인덱스가 73.9%를 차지하고 있다. 상세한 내역은 〔표 2〕에 나타나 있다.

[표 2] 한국불교전서 원문 데이터베이스 규모

분류	테이블 이름	테이블 행의 수	테이블 크기(KB)
원문	원문테이블 1~14	885,011	160,344
	소계	885,011	160,344
키워드 인덱스	한자 키워드 사전	58,245	2,952
	한글 키워드 사전	58,245	1,736
	통합 키워드 인덱스	6,595,732	225,520
	책 별 키워드 인덱스 1~14	6,595,732	226,128
	소계	13,307,954	456,336
제목 인덱스	한글 제목 인덱스	326	56
	한자 제목 인덱스	1,576	392

	한자 제목 인덱스 요약	1,576	48
	소계	3,478	496
총계		14,196,443	617,176

※ 임시 테이블과 여유 공간을 포함한 데이터베이스 크기 : 878,216KB

한국불교전서 원문 데이터베이스의 시대별 구성 비율을 살펴보면 신라시대편이 20.3%, 고려시대편이 22.6%, 조선시대편이 28.4%, 보유편이 28.6%를 차지하고 있다. 이의 자세한 내역은 〔표 3〕에서 살펴볼 수 있다.

[표 3] 한국불교전서 원문 데이터베이스 시대별 구성

구분	책 번호	페이지수 (개수)	단수 (개수)	총 라인수 (개수)	주석수 (개수)	글자수 (개수)	텍스트크기 (KB)	DB 테이블 크기 (KB)
신라시대	1책	843	2,527	61,242	2,367	811,354	2,588	10,568
	2책	846	2,538	61,373	2,929	837,657	2,711	11,400
	3책	782	2,523	57,447	3,479	793,786	2,446	10,504
	소계	2,471	7,588	180,062	8,775	2,442,797	7,745	32,472
고려시대	4책	869	2,338	63,532	3,354	828,832	2,818	11,976
	5책	924	2,607	67,816	4,715	827,168	2,637	11,200
	6책	902	2,674	66,244	5,073	853,592	2,991	13,000
	소계	2,695	7,619	197,592	13,142	2,509,592	8,446	36,176
조선	7책	830	2,454	59,548	4,565	745,974	2,498	11,584
	8책	796	2,431	48,659	2,315	805,545	2,506	10,568
	9책	801	2,394	55,670	1,366	678,130	2,210	9,608
	10책	1,145	3,406	77,608	2,876	1,028,746	3,260	13,832
	소계	3,572	10,685	241,485	11,122	3,258,395	10,473	45,592
보유편	11책	871	2,457	55,992	1,085	723,185	2,228	9,536
	12책	875	2,621	62,868	2,144	780,607	2,544	10,824
	13책	1,002	3,006	74,995	5,177	897,288	2,993	12,808
	14책	962	3,062	72,017	5,353	889,078	2,900	12,936
	소계	3,710	11,146	265,872	13,759	3,290,158	10,665	46,104
총계		12,448	37,038	885,011	46,798	11,500,942	37,329	160,344

한국불교전서 키워드 데이터베이스의 시대별 구성 비율은 신라시대편 25.6%, 고려시대편 20.6%, 조선시대편 23.4%, 보유편 30.4%로 되어 있다. 이의 자세한 내용은 〔표 4〕에서 보여주고 있다.

[표 4] 한국불교전서 키워드 데이터베이스 시대별 구성

구분	책 번호	키워드 수(개)	키워드 평균 출현횟수(회)	전체 출현 횟수(회)	DB 테이블 크기(KB)
신라 시대	1책	11,050	52	571,392	19,592
	2책	12,818	44	567,242	19,464
	3책	11,505	48	547,051	18,752
	소계	35,373	48	1,685,685	57,808
고려 시대	4책	12,938	40	519,293	17,800
	5책	8,682	46	397,904	13,640
	6책	13,996	31	440,729	15,104
	소계	35,616	38	1,357,926	46,544
조선 시대	7책	11,364	33	377,616	12,936
	8책	12,727	29	371,165	12,744
	9책	10,296	30	307,094	10,568
	10책	12,954	38	487,987	16,712
	소계	47,341	33	1,543,862	52,960
보유편	11책	12,459	32	404,492	13,896
	12책	13,335	31	404,167	13,832
	13책	9,111	65	594,535	20,352
	14책	8,275	73	605,065	20,736
	소계	43,180	47	2,008,259	68,816
전체		58,245	113	6,595,732	226,128

한국불교전서 전체에서 가장 빈번히 출현하는 키워드들을 1위부터 10위까지 추출하였다. 이들은 전체 키워드 출현의 13.94%를 차지하며, 상세한 내용은 〔표 5〕에 나타나 있다. 출현빈도 1위부터 10위까지 모두 모두 한 글자로 구성된 키워드이다.

[표 5] 한국불교전서 10대 키워드

순위	키워드	출현횟수	출현 비율(%)
1	有(유)	125,552	1.90
2	者(자)	121,579	1.84
3	無(무)	120,911	1.83
4	故(고)	100,972	1.53
5	如(여)	82,155	1.25
6	三(삼)	79,343	1.20
7	中(중)	75,650	1.15
8	法(법)	74,990	1.14
9	所(소)	72,147	1.09
10	生(생)	66,186	1.00
총계		919,485	13.94

한국불교전서 전체에서 가장 빈번히 출현하는 두 글자로 구성된 키워드들을 1위부터 10위까지 추출하였다. 이들은 전체 키워드 출현의 1.54%를 차지하며, 자세한 내역은 〔표 6〕에서 보여주고 있다.

[표 6] 한국불교전서 10대 키워드(2자)

순위	키워드	출현 횟수	출현 비율(%)
1	菩薩(보살)	16,421	0.25
2	如是(여시)	16,117	0.24
3	一切(일체)	15,769	0.24

4	第二(제이)	10,188	0.15
5	衆生(중생)	9,214	0.14
6	第三(제삼)	7,190	0.11
7	分別(분별)	6,932	0.11
8	煩惱(번뇌)	6,771	0.10
9	差別(차별)	6,606	0.10
10	如來(여래)	6,349	0.10
총계		101,557	1.54

한국불교전서 전체에서 가장 빈번히 출현하는 세 글자로 구성된 키워드들을 1위부터 10위까지 추출하였다. 이들은 전체 키워드 출현의 0.148%를 차지하며, 상세한 내역은 〔표 7〕에 나타나 있다.

[표 7] 한국불교전서 10대 키워드(3자)

순위	키워드	출현 횟수	출현 비율(%)
1	如何是(여하시)	1,239	0.019
2	一切法(일체법)	1,197	0.018
3	作麽生(자마생)	1,085	0.016
4	阿彌陀(아미타)	1,009	0.015
5	彌陀佛(미타불)	929	0.014
6	無分別(무분별)	906	0.014
7	善男子(선남자)	902	0.014
8	補特伽(보특가)	887	0.013
9	三摩地(삼마지)	826	0.013
10	曹溪宗(조계종)	800	0.012
총계		9,780	0.148

한국불교전서 전체에서 가장 빈번히 출현하는 네 글자로 구성된 키워드들을 1위부터 10위까지 추출하였다. 이들은 전체 키워드 출현의 0.068%를 차지하며, 자세한 내역은 〔표 8〕에서 보여주고 있다.

[표 8] 한국불교전서 10대 키워드(4자)

순위	키워드	출현 횟수	출현 비율(%)
1	阿彌陀佛(아미타불)	863	0.013
2	補特伽羅(보특가라)	762	0.012
3	遍計所執(변계소집)	664	0.010
4	發菩提心(발보리심)	374	0.006
5	無分別智(무분별지)	322	0.005
6	三世諸佛(삼세제불)	320	0.005
7	瑜伽師地(유가사지)	316	0.005
8	毗鉢舍那(비발사나)	292	0.004
9	八萬四千(팔만사천)	280	0.004
10	波羅蜜多(바라밀다)	279	0.004
총계		4,472	0.068

한국불교전서에서 가장 빈번히 출현하는 키워드들의 순위를 시대별로 추출하였다. 신라시대편, 고려시대편, 조선시대편, 보유편의 키워드들의 순위가 〔표 9〕, 〔표 10〕, 〔표 11〕, 〔표 12〕에 각각 나타나 있다.

[표 9] 신라시대편 10대 키워드 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	故(고)	41,315	0.63
2	有(유)	36,815	0.56
3	者(자)	36,238	0.55
4	無(무)	32,718	0.50

5	如(여)	24,085	0.37
6	三(삼)	23,766	0.36
7	所(소)	22,794	0.35
8	中(중)	22,522	0.34
9	法(법)	21,156	0.32
10	名(명)	20,983	0.32
총계		282,303	4.28

[표 10] 고려시대편 10대 키워드 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	者(자)	30,184	0.46
2	無(무)	24,524	0.37
3	有(유)	22,881	0.35
4	故(고)	19,904	0.30
5	中(중)	17,446	0.26
6	三(삼)	16,835	0.26
7	法(법)	16,340	0.25
8	如(여)	16,165	0.25
9	人(인)	14,194	0.22
10	大(대)	13,223	0.20
총계		191,696	2.91

[표 11] 조선시대편 10대 키워드 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	無(무)	26,578	0.40
2	者(자)	23,174	0.35
3	有(유)	23,044	0.35
4	人(인)	21,283	0.32

5	山(산)	16,523	0.25
6	三(삼)	16,315	0.25
7	如(여)	16,121	0.24
8	大(대)	15,445	0.23
9	心(심)	15,286	0.23
10	法(법)	15,021	0.23
총계		188,790	2.86

[표 12] 보유편 10대 키워드 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	有(유)	42,812	0.65
2	無(무)	37,091	0.56
3	者(자)	31,983	0.48
4	故(고)	27,759	0.42
5	所(소)	27,240	0.41
6	如(여)	25,784	0.39
7	法(법)	22,473	0.34
8	三(삼)	22,427	0.34
9	中(중)	21,615	0.33
10	生(생)	21,125	0.32
총계		280,309	4.25

한국불교전서에서 가장 빈번히 출현하는 두 글자 키워드들의 순위를 시대별로 추출하였다. 신라시대편, 고려시대편, 조선시대편, 보유편의 두 글자 키워드들의 순위를 〔표 13〕, 〔표 14〕, 〔표 15〕, 〔표 16〕에서 각각 보여주고 있다.

[표 13] 신라시대편 10대 키워드(2자) 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	菩薩(보살)	6,164	0.09
2	第二(제이)	4,788	0.07
3	一切(일체)	4,256	0.06
4	如是(여시)	4,130	0.06
5	分別(분별)	3,418	0.05
6	第三(제삼)	3,331	0.05
7	中有(중유)	2,804	0.04
8	煩惱(번뇌)	2,763	0.04
9	一切(일체)	2,683	0.04
10	衆生(중생)	2,651	0.04
총계		36,988	0.56

[표 14] 고려시대편 10대 키워드(2자) 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	如是(여시)	2,572	0.04
2	菩薩(보살)	2,465	0.04
3	第二(제이)	2,268	0.03
4	一切(일체)	2,207	0.03
5	第一(제일)	2,124	0.03
6	衆生(중생)	2,122	0.03
7	什麽(습마)	1,930	0.03
8	雲門(운문)	1,657	0.02
9	和尙(화상)	1,609	0.02
10	三乘(삼승)	1,582	0.02
총계		20,536	0.31

[표 15] 조선시대편 10대 키워드(2자) 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	衆生(중생)	2,339	0.04
2	一切(일체)	2,185	0.03
3	如來(여래)	2,017	0.03
4	菩薩(보살)	1,933	0.03
5	菩薩(보살)	1,874	0.03
6	大師(대사)	1,805	0.03
7	禪師(선사)	1,278	0.02
8	金剛(금강)	1,235	0.02
9	二十(이십)	1,164	0.02
10	上人(상인)	1,110	0.02
총계		16,940	0.26

표 16. 보유편 10대 키워드(2자) 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	如是(여시)	7,541	0.11
2	一切(일체)	7,121	0.11
3	菩薩(보살)	5,859	0.09
4	差別(차별)	2,934	0.04
5	淸淨(청정)	2,862	0.04
6	煩惱(번뇌)	2,858	0.04
7	第二(제이)	2,525	0.04
8	因緣(인연)	2,462	0.04
9	有情(유정)	2,392	0.04
10	解脫(해탈)	2,385	0.04
총계		38,939	0.59

한국불교전서에서 가장 빈번히 출현하는 세 글자 키워드들의 순위를 시대별로 추출하였다. 신라시대편, 고려시대편, 조선시대편, 보유편의 세 글자 키워드들의 순위가 〔표 17〕, 〔표 18〕, 〔표 19〕, 〔표 20〕에 각각 나타나 있

다.

[표 17] 신라시대편 10대 키워드(3자) 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	一切法(일체법)	831	0.013
2	無分別(무분별)	449	0.007
3	無自性(무자성)	364	0.006
4	善男子(선남자)	363	0.006
5	薩婆多(살바다)	331	0.005
6	三摩地(삼마지)	290	0.004
7	菩提心(보리심)	277	0.004
8	所知障(소지장)	276	0.004
9	圓成實(원성실)	271	0.004
10	依他起(의타기)	256	0.004
총계		3,708	0.056

[표 18] 고려시대편 10대 키워드(3자) 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	如何是(여하시)	814	0.012
2	作麽生(자마생)	786	0.012
3	華嚴經(화엄경)	274	0.004
4	法華經(법화경)	174	0.003
5	善知識(선지식)	171	0.003
6	一切法(일체법)	156	0.002
7	不可得(불가득)	148	0.002
8	初發心(초발심)	139	0.002
9	波羅蜜(바라밀)	135	0.002
10	菩提心(보리심)	133	0.002
총계		2,930	0.044

[표 19] 조선시대편 10대 키워드(3자) 출현 빈도

순위	키워드	출현 횟수	출현 비율(%)
1	阿彌陀(아미타)	733	0.011
2	彌陀佛(미타불)	712	0.011
3	金剛山(금강산)	257	0.004
4	一切法(일체법)	241	0.004
5	彌陀佛(미타불)	240	0.004
6	如來禪(여래선)	238	0.004
7	如何是(여하시)	225	0.003
8	須菩提(수보리)	223	0.003
9	祖師禪(조사선)	214	0.003
10	善男子(선남자)	203	0.003
총계		3,286	0.050

[표 20] 보유편 10대 키워드(3자) 출현 빈도

순위	키워드	출현 횟수	출현 빈도(%)
1	補特伽(보특가)	2,718	0.041
2	梵網經(범망경)	1,470	0.022
3	阿羅漢(아라한)	1,009	0.015
4	三摩地(삼마지)	960	0.015
5	世俗智(세속지)	953	0.014
6	般涅槃(반열반)	849	0.013
7	去來今(거래금)	643	0.010
8	無顚倒(무전도)	620	0.009
9	嗢柁南(올타남)	602	0.009
10	賴耶識(뢰야식)	556	0.008
총계		10,380	0.157

4. 결 론

한국불교전서 데이터베이스의 분석 결과 전체 크기는 원문 텍스트와 키워드 파일들을 모두 합하여 617MB이고, 기타 파일들을 모두 포함하면 878MB를 차지한다. 원문 텍스트는 14책, 12,448쪽, 37,038단, 885,011행, 11,500,942자로 구성되어 160MB를 차지한다. 키워드는 58,245개로 구성되어 전체 출현 빈도가 6,595,732회이며, 모든 인덱스 테이블의 크기는 456MB를 차지하고 있다. 키워드들의 출현 빈도도 분석하였으며, 전체 14책에서 상위 10대 키워드들의 출현 빈도와 시대별 상위 10대 키워드들의 출현 빈도를 조사하였다. 분석 결과 키워드들은 시대별로 상이한 출현 양상을 보였다. 앞으로 이 결과를 바탕으로 보다 심도 있는 분석이 필요하다.

〔참고문헌〕

〔1〕 한보광, 한국불교전서 성과 및 향후 과제, 제9회 동국대학교 전자불전문화재콘텐츠연구소 학술세미나 자료집, 2007. 11. 2.
〔2〕 동국대학교 전자불전문화재콘텐츠연구소, 한국불교전서 전산화 사업 결과보고서, 2007. 6.
〔3〕 구현우 외 7인, 한국불교전서 검색 시스템 개발, 전자불전, 제8집, 2006. 12.

키워드(Keyword)

한국불교전서, 한국불교전서 전산화, 한국불교전서 데이터베이스
Korea Bulgyo Chonso, Korea Bulgyo Chonso Digitalization, Korea Bulgyo Chonso Database