한국불교전서 전산화에 있어서 누락문자 관리 시스템

홍 영 식*

목 차

요 약

유니코드를 기반으로 표현된 한자를 입력할 때 유니코드로 표현할 수 없는 한자를 관리하는 문제는 한국불교전서 디지털화에 있어서 매우 중요한 문제에 속한다.

본 논문에서는 누락문자관리 시스템의 기능적 구성과 지난 8년간 한국불교전서 디지털화 과정에서 개발한 누락문자 처리기법의 변화에 대해서 기술한다.

또한, 그동안 누적된 누락문자에 대한 통계와 누락문자 관리기법과 검색을 위한 사용자 인터페이스에 대해서 검토한다.

* 동국대학교 컴퓨터공학과 교수

1. 서 론

1999년에 동국대학교 100주년 기념사업의 일환으로 한국불교전서 디지털화 과제수행에서 한자입력은 중요한 사항이었다. 한자입력과 관련된 사항은 한자표기를 위한 코드의 결정과 폰트의 확보, 문서편집기의 선택 및 기존 문서편집기에서 입력되지 않는 특수 한자, 즉 누락문자의 처리에 관한 것이다.

이 시기는 한자문화권에 속하는 한국, 대만, 중국 및 일본에서 한자로 기술된 문서의 디지털화작업이 시작되어 활발하게 진행되고 있었다. 특히, 한자표기를 위한 코드관련 문제를 해결하기 위해서 한국, 대만, 중국 및 일본에서 국가별로 독자적인 노력을 해오다가 한국, 대만 및 일본이 중심이 되어 CJK 코드가 정해졌고, 이것이 현재 표준코드로 정해진 2바이트 유니코드의 기반이 되었다.

국내에서는 해인사 대장경연구소에서 추진한 대장경 디지털화를 위한 4바이트 코드체계가 시도되었고, 동국대학교 전자불전연구소에서는 한국불교전서 전산화를 위해서 2바이트 유니코드를 채택했다. 2000년 2월에 발표된 유니코드 3.0의 경우 27,786개의 한자용 코드가 할당되었지만, 그 당시 국내 폰트제작 업체에서 공급하는 폰트 수는 유니코드 2.1에 기반한 21,204개의 코드에 대한 한자 폰트가 제공되고 있었다[1,6,7,8,9,10].

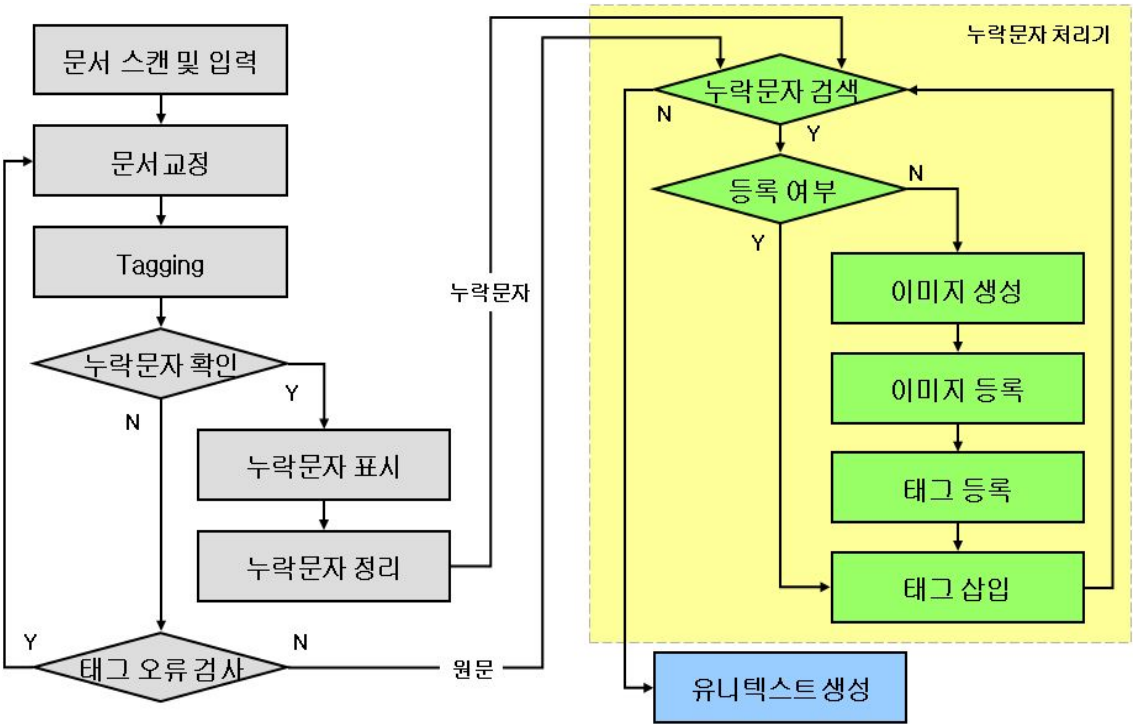
한역 불교경전을 위시한 한자로 기술된 고문헌의 디지털화에 필요한 한자 폰트의 수가 절대적으로 부족하여 코드가 할당되지 않은 누락문자에 대한 별도의 처리방안이 필요하다.

인터넷 환경에서 웹서비스를 위해서 유니코드가 할당되지 않은 누락문자를 표현하기 위해서 누락문자에 대한 이미지파일을 생성하여 XML로 표현된 웹문서에 포함시킨다[2,3,5].

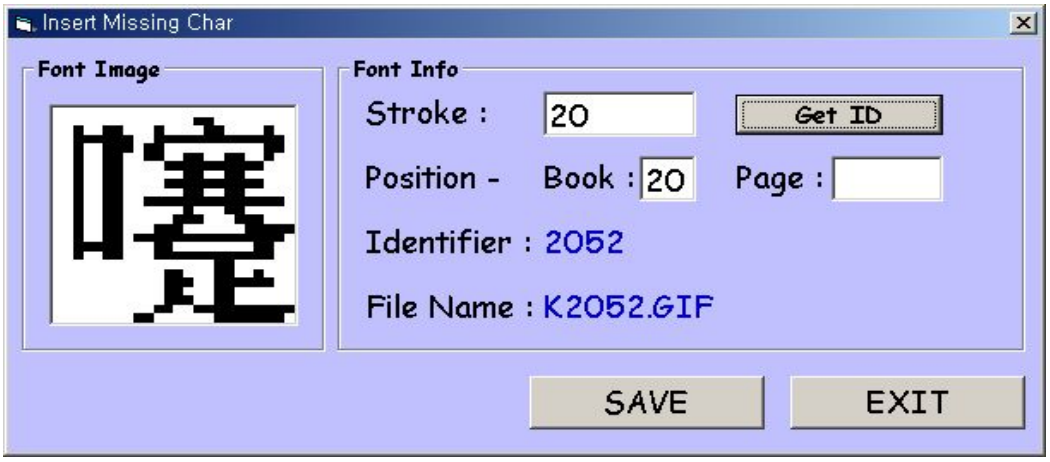
본고에서는 한국불교전서 디지털화사업에서 한자문서입력과 관련한 누락문자처리 방식과 한국불교전서 디지털화사업 추진과정에서 누락문자관리와 관련한 기술적 변화과정을 고찰하고자 한다.

2. 누락문자처리절차 및 관리방식

인터넷 기반 웹서비스를 위한 한국불교전서 문서의 데이터베이스구축과정에서 일차적으로 초기입력이 필요하며, 초기입력은 워드프로세서로 직접 입력하거나, 혹은 스캐너를 사용해서 입력파일을 생성한다. 그 다음에 이것을 원문과 대조하여 누락문자로 판단된 글자에 대해서 누락문자 데이터베이스를 검색하고, 누락문자 데이터베이스에 아직 존재하지 않는 글자의 경우 이미지 폰트파일을 생성하여 누락문자 데이터베이스에 등록한다. 〔그림 1〕은 누락문자를 데이터베이스에 등록하는 절차를 도시한 것이고, 〔그림 2〕는 이미지 폰트를 생성하여 누락문자 데이터베이스에 입력하는 예를 보인 것이다.

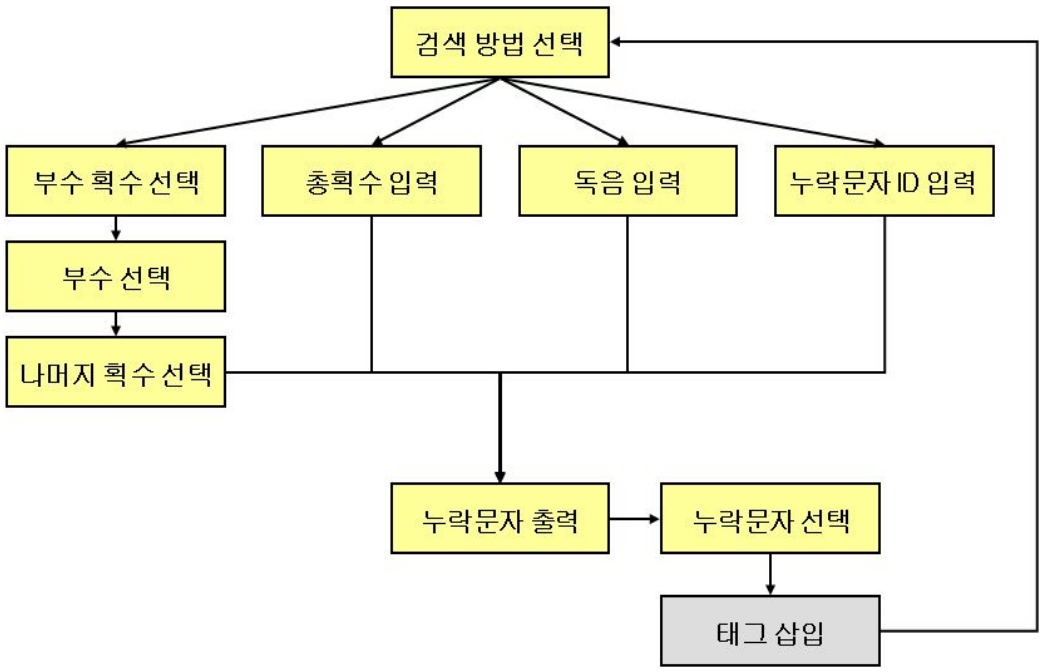


[그림 1] 누락문자 데이터베이스 등록절차

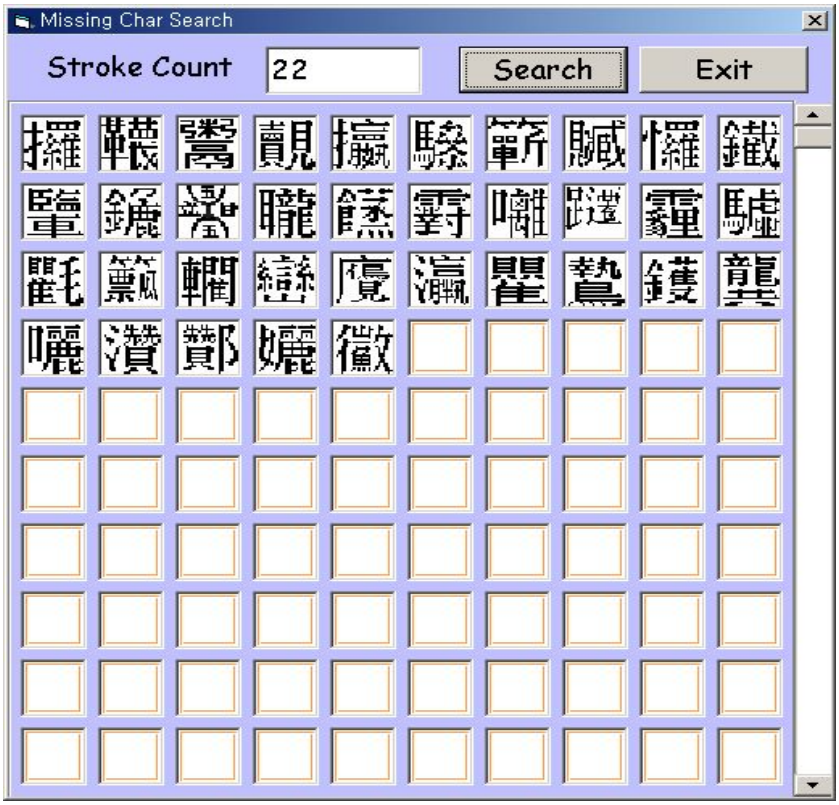


[그림 2] 누락문자 입력 예

그리고, 누락문자 데이터베이스에 이미 등록되어 있는 누락문자인 경우에는 데이터베이스를 검색한다. [그림 3]은 누락문자 검색절차를 도시한 것이고, [그림 4]는 누락문자의 총획수 22로 검색된 누락문자들을 나타낸 것이다.



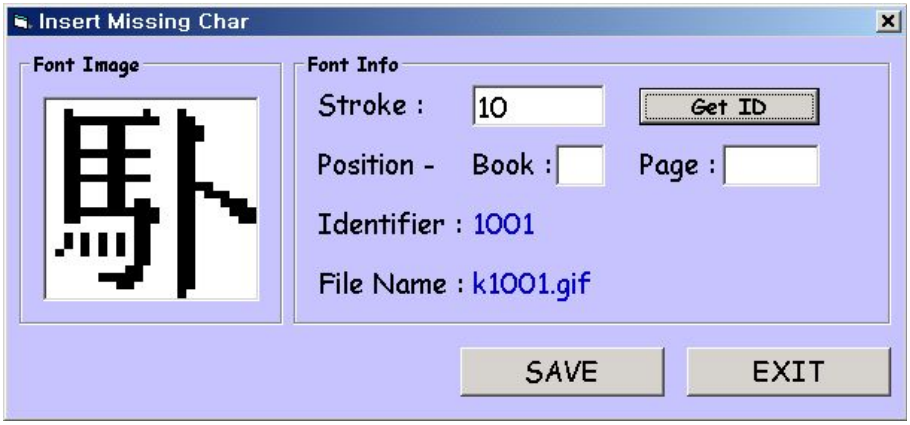
[그림 3] 누락문자 검색 절차



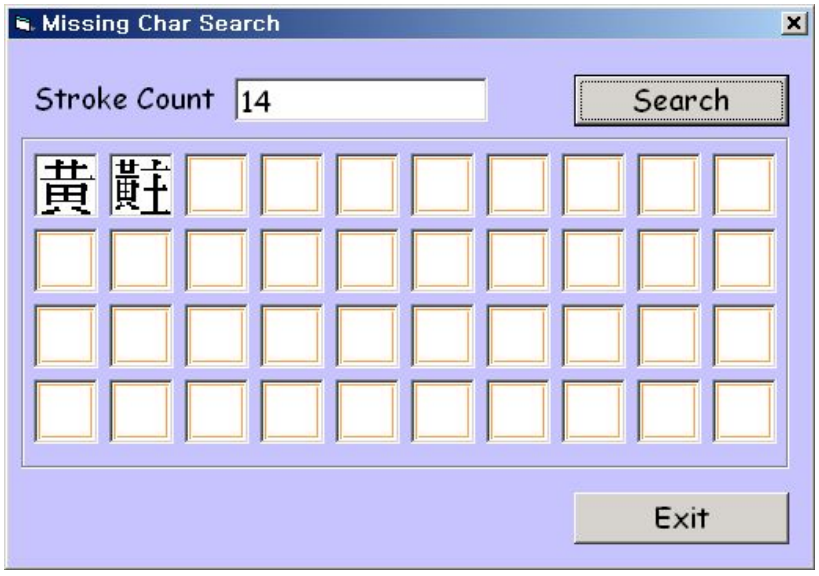
[그림 4] 누락문자 검색 예

한국불교전서 디지털화사업 초기에는 누락문자 데이터베이스에 존재하지 않는 새로 발견된 누락문자에 대한 이미지를 생성하기 위해서 문자경연구회에서 제공하고 있는 모직교 폰트의 문자표에서 해당 문자를 검색하고 검색된 이미지의 URL을 누락문자 위치에 "붙여넣기"를 수행했다.

그러나, 모직교 폰트사이트에 매번 접속하는 것이 비효율적이서 2001년부터 누락문자 관리기를 자체 제작하게 되었다. 누락문자 관리기를 사용하여 누락문자에 대한 이미지 파일을 생성하고 누락문자의 폰트에 해당하는 이미지파일을 관리하는 누락문자 데이터베이스를 구축하게 되었다. 〔그림 5〕와 〔그림 6〕은 누락문자 파일을 만든 후 누락문자 데이터베이스에 등록하고 검색하는 예에 관한 것이다.



[그림 5] 문자의 등록

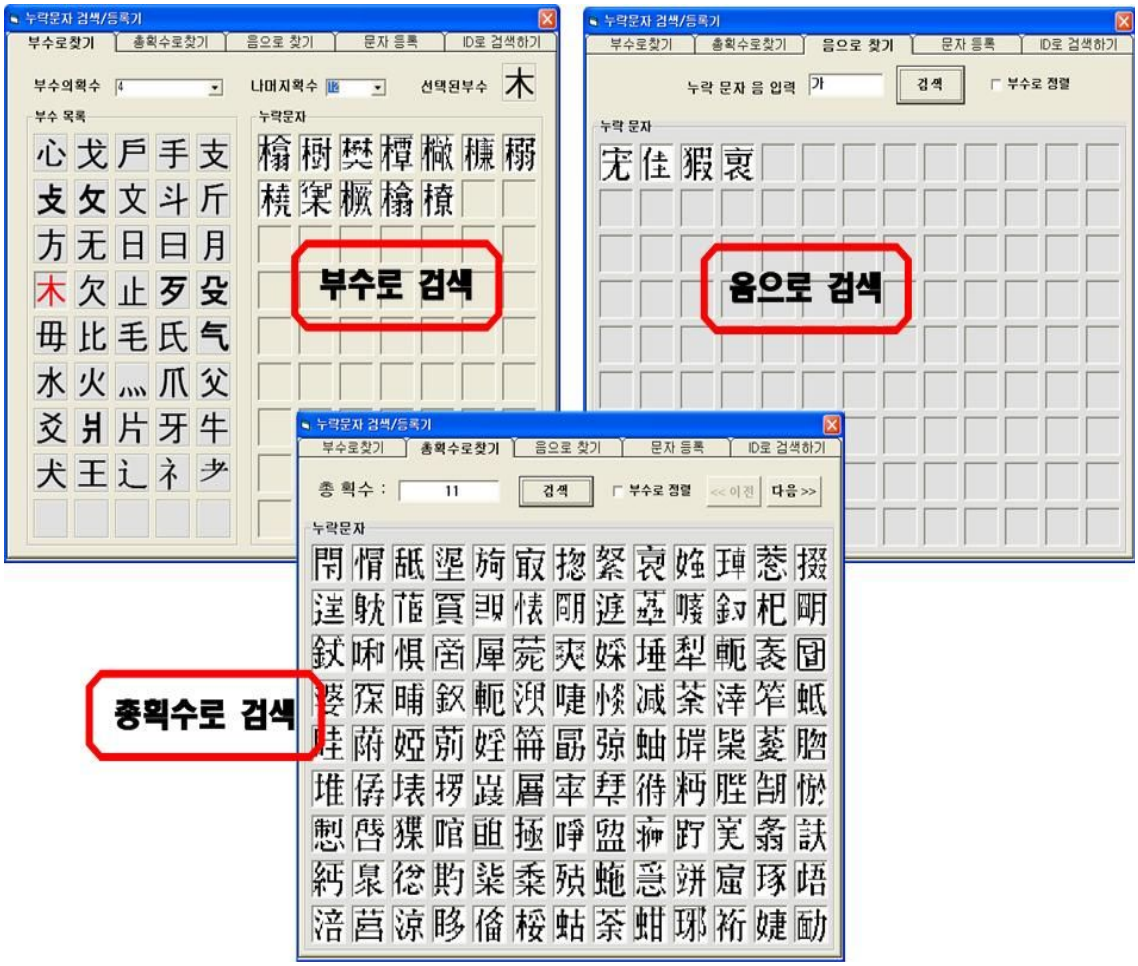


[그림 6] 문자의 검색

누락문자 디스플레이 효과를 높이기 위해서는 누락문자의 검정색 문자 이미지에 다양한 색상의 이미지를 생성할 수 있는 누락문자 색상 변환 기능이 필요하며, 〔그림 7〕은 누락문자 관리기의 색상변환 기능에 관한 것이다.

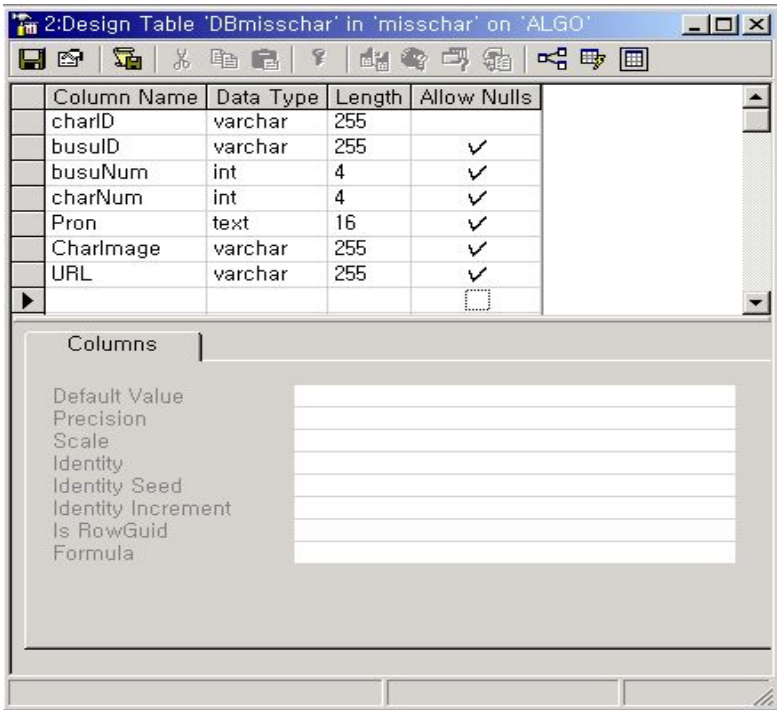
[그림 7] 누락 문자 색상 변환기

또한, 누락문자 검색에서 기존의 획수에 의한 검색기능에 추가해서 부수나 발음을 이용한 검색기능을 추가함으로써 누락문자 검색기능을 다양화시키는 것이 효과적이며, 〔그림 8〕은 누락문자에 대한 총획수, 부수 및 독음을 이용한 검색 예를 보여준다.



[그림 8] 누락 문자 검색 기능들

그리고, 다양한 검색 방법을 지원하려면 누락문자 데이터베이스 구조를 확장하고 누락문자 등록 인터페이스를 수정할 필요가 있다. 〔그림 9〕는 누락문자 검색기능의 다양화를 위한 누락문자 데이터베이스의 확장된 내용에 관한 것이고, 〔그림 10〕은 이와 관련한 누락문자 등록 인터페이스를 보여준다.

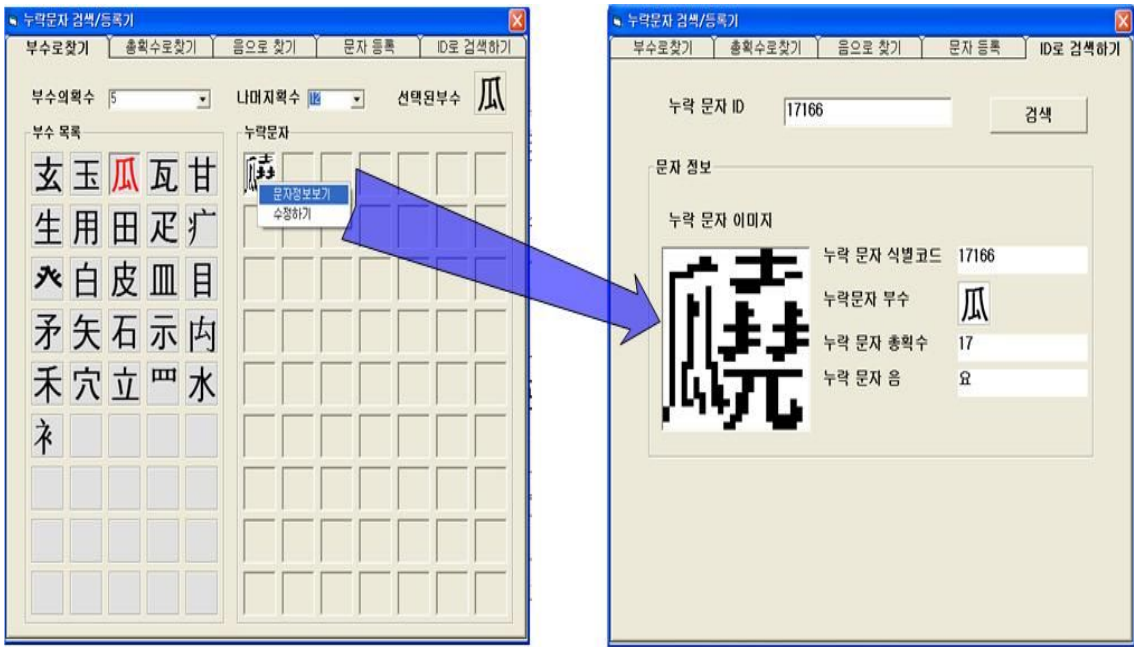


[그림 9] 개선된 누락 문자 데이터베이스 구조

[그림 10] 누락문자 등록 인터페이스

누락문자관리기는 누락문자 데이터베이스를 효과적으로 관리하기 위한 것으로서 기존의 누락문자 등록인터페이스를 확장할 필요가 있다. 기존 누락문자 관리기의 검색 기능에 누락문자 ID를 이용한 검색기능을 추가하여 동일한 누락문자의 연속 검색에 따른 검색단계를 간소화하여 입력효율을 향상시킨다. 또한, 수년간의 누락문자 데이터베이스 관리 과정에서 발생한 등록된 누락문자에 이미지 등 등록된 정보의 수정기능을 추가할 필요가 있다.

〔그림 11〕은 부수를 사용한 누락문자 검색 예이고, 〔그림 12〕는 누락문자 데이터베이스에 이미 등록되어 있는 글자 중에 이미지 오류가 있는 경우에 이것을 수정하는 예를 보여준다.



[그림 11] “부수를 이용한 검색”에서 누락문자 추가 정보 보기

[그림 12] 누락문자 수정기

3. 누락문자 통계

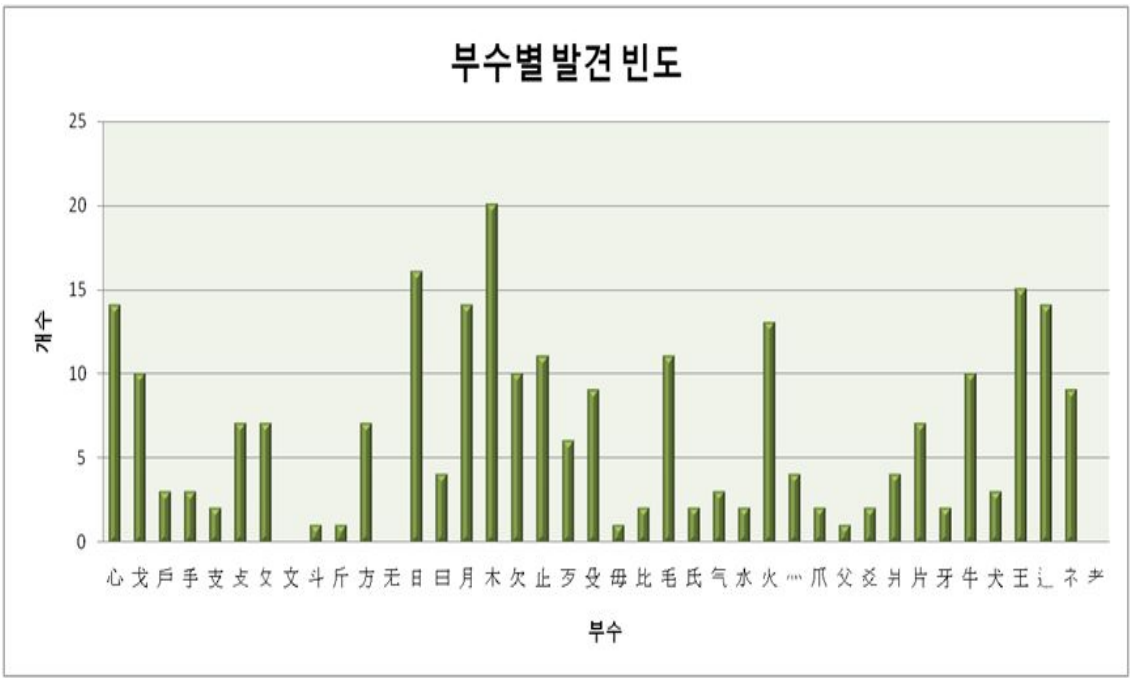
누락문자관리기는 한글대장경 전산화에서도 사용되었으며, 한국불교전서 디지털화와 한글대장경 전산화 과정에서 현재까지 누락문자 데이터베이스에 등록된 문자의 총수는 3,318 자로서 이것을 총획수별로 분류한 통계는 〔표 1〕과 같다.

[표 1] 누락문자들에 대한 총 획수별 통계

획수	빈도수	획수	빈도수	획수	빈도수	획수	빈도수
1	118	11	310	21	89	31	2
2	30	12	313	22	63	32	0
3	25	13	278	23	55	33	1
4	34	14	300	24	36	34	1
5	61	15	262	25	13	35	0
6	76	16	297	26	19	36	1

7	148	17	211	27	9	37	1
8	178	18	173	28	9	38	0
9	201	19	149	29	0	39	0
10	251	20	94	30	5	40	0

또한, 누락문자 데이터베이스에 등록된 전체 문자 중에서 4획 부수의 문자들에 대한 개략적 분포는 〔그림 13〕과 같다.



[그림 13] 4획 부수인 누락문자들의 분포

4. 누락문자 관리 시스템 활용과 향후 과제

앞에서 기술한 바와 같이 누락문자 관리시스템은 8년간 한국불교전서 디

지털화 과제 수행과정에서 기술적 처리방식이 지속적으로 개선되었으며, 초창기의 누락문자관리 방식에서 그동안 개선된 사항은 다음과 같다.

1) 누락문자 관리 인터페이스의 개선
2) 다양한 누락문자 검색방법 제공
3) 검색속도 향상

누락문자 관리 시스템은 한국불교전서 디지털화뿐만 아니라 한글대장경 전산화에서도 사용되고 있으며, 이것은 인터넷 검색을 기반으로 한 여타 고문헌 전산화에서도 활용될 수 있슴을 입증하고 있다. 또한, 누락문자 처리에 소요되는 시간이 고문헌 입력 및 편집시간에 미치는 영향이 크기 때문에 보다 효율적인 누락문자 처리시스템 선택이 필요하다.

5. 결 론

한국불교전서의 디지털화 과정에서 개발된 누락문자관리 시스템에 대한 기술적 발전과정을 개략적으로 살펴보았다. 또한, 누락문자 관리 시스템이 인터넷을 기반으로 한 웹서비스를 전제로 한 여타 고문헌 전산화에 활용될 수 있음을 알 수 있었다. 한국불교전서 디지털화와 한글대장경 전산화 과정에서 구축된 누락문자 데이터베이스에 앞으로 발견될 누락문자를 계속 추가함으로써 향후의 고문헌 디지털화에 효과적으로 활용할 수 있을 것이다.

인터넷 사용이 보편화된 현 시점에서 한국불교전서의 디지털화 과제를 수행하면서 축적한 누락문자 관리 기술과 누락문자 데이터베이스의 활용방안을 지속적으로 모색할 필요가 있다. 특히, 누락문자 처리에 대한 중복 투자를 피하기 위해서 누락문자 이미지 파일과 누락문자 ID 및 누락문자 검색 인터페이스의 표준화는 향후에 필히 해결되어야 할 과제 중 하나라 하겠다.

〔참고문헌〕

〔1〕 Richard Cook , "UniHan Variation: Issues and Solutions", 23rd I nternationailzation & Unicode Conference, 2002.
〔2〕 Y.S.Hong, et al. "Searching Missing Characters from the Hanguk Bulgyo Chonso Database", 2001 EBTI International Conferenc e, 2001.
〔3〕 Y.S.Hong, et al. "Development of a Syntax-directed SGML Edito r for Processing Korean Ancient Documents", Prodeedings of 1999 EBTI, ECAI, SEER & PNC Joing Meeting in Taipei, 19 99.
〔4〕 구현우, 김영희, 박미화, 이재수, 신병삼, 이금석, 이용규, 홍영식, 한보광, "한국불교전서 검색 시스템 개발", 동국대학교 전자불전.문화재콘텐츠연구소, 제8집, 2006.
〔5〕 구현우, 박성은, 노진홍, 김영희, 박영희, 이금석, 이용규, 홍영식, 한보광, "한국불교전서 전산화 5차 사업", 동국대학교 전자불전연구소, 제6집, 2004.
〔6〕 노진홍, 유응구, 박성은, 이용규, 이금석, 홍영식, 한보광, "한글대장경 전산화", 전자불전, 제4집, 동국대학교 전자불전연구소, 2002.
〔7〕 윤용석, "고려대장경의 인터넷 검색 및 열람", 전자불전, 창간호, 동국대학교 전자불전연구소, 1999.
〔8〕 이금석, 이용규, 홍영식, 한태식, "한국불교전서 전산화를 위한 누락문자 처리방안", 동국대학교 전자불전연구소 제 2회 세미나, 2000.
〔9〕 한인, 이용규, 이금석, 홍영식, 한보광, "유니코드 한자처리를 위한 입력기와 편집기의 설계 및 구현", 전자불전, 창간호, 동국대학교 전자불전연구소, 1999.
〔10〕 허인섭, "전산화본 고려대장경 2000 완성의 학술적 의미와 미래전망", 동국대학교 전자불전연구소 제 2회 세미나, 2000.

키워드(Keyword)

유니코드, 한국불교전산화, 누락문자, 사용자 인터페이스, 누락문자 검색, 이미지 파일, 폰트

Unicode, digitalization of Hanguk Bulgyo Chonso, Missing character, User interface, Retrieval of missing character, Image file, Font