전자불전 누락문자의 관리

홍 영 식*

목 차

1. 서론
2. 유니코드와 누락문자 및 특수기호
3. 누락문자 통계
4. 누락문자 이미지 편집
5. 누락문자 관리 시스템
6. 결론

요 약

한국불교전서와 한글대장경의 디지털화에서 유니코드로 표현될 수 없는 문자에 대한 누락문자관리는 해결해야 할 중요한 과제 중의 하나이다.

본 논문에서는 누락문자관리 시스템의 구조와 누락문자 관리를 위해 웹서비스와 관련한 기술적 기법들에 관해서 기술한다. 또한, 지난 10년간 한국불교전서와 한글대장경의 디지털화 과제를 진행하면서 개발된 누락문자 이미지 생성, 등록 및 이미지 편집 등의 연산에 대해서 검토한다.

1. 서 론

한국불교전서의 디지털화 과제는 동국대학교 전자불전문화콘텐츠연구소에서 1999년부터 10년간 진행된 후 완료되었고, 한글대장경의 디지털화 작업은 2005년부터 지금까지 진행되고 있다. 이러한 고문헌 디지털화 대형과제를 진행하는데 있어서 일반 사용자들을 위한 문헌검색 서비스는 웹서비스를 기본으로 해야 하고, 한자를 웹에 표현하기 위해서는 유니코드를 사용하여 한자를 표현할 수 있다.[1, 2, 3, 4, 10]

한편, 누락문자 해결을 위해서 글립표현모델(glyph expression model)을 사용해서 한자를 기본요소로 조립하는 방법도 발표되었지만[5, 6], 고문헌에 사용된 특수기호는 조립식 표현이 쉽지 않다. 제한 된 개수의 유니코드로서 한자 위주로 쓰여진 고문헌의 디지털화에서 유니코드 폰트가 존재하지 않는 한자 및 특수문자는 유니코드가 할당된 한자 폰트 이미지와 동일한 크기의 이미지를 만들어서 표현한다.

누락문자 관리시스템은 누락문자의 이미지를 생성하고, 이것을 누락문자 데이터베이스에 등록함으로써 이후에 발견되는 누락문자를 누락문자 데이터베이스에서 검색하여 원문에 대한 HTML 문서에 태깅(tagging)하여 추가한다. 새로 발견된 누락문자 이미지를 데이터베이스에 계속해서 추가하게 되면 시간이 흐름에 따라 데이터베이스의 크기가 늘어난다. 그렇게 되면 누락문자 이미지 처리 과정에서 잘못 만들어진 이미지나 선명도가 낮은 누락문자 이미지들이 발생하게 된다. 이러한 문제를 해결하기 위한 누락문자 이미지 편집기능이 추가로 필요하게 되었고, 이와 같이 누락문자관리 시스템의 기능은 개선되어 왔다.

본 논문에서는 누락문자와 유니코드에 관하여 기술하고, 특수기호의 누락문자 처리와 누락문자에 대한 이미지 파일의 편집과 누락문자 관리시스템의 구성에 대해서 서술한다.

2. 유니코드와 누락문자 및 특수기호

2바이트 유니코드로 표현되는 한자는 한국, 대만 및 일본이 중심이 되어 공용 한자표현을 위한 CJK 코드가 정해져서 유니코드의 기반이 되었다. 2000년에 발표된 유니코드 3.0에서는 27,786 개의 한자용 코드가 할당되었고, 2009년에 발표된 유니코드 5.2에는 기본 한자를 위한 20,944개의 코드와 확장A 영역에 6,582개, 확장B에 42,711 및 확장C에 4,149개의 코드가 할당되어 총 74,386 개의 통합 한자용 코드가 할당되어 있다. 그러나, 현재 웹 환경에서 HTML 문서를 작성하는데

실제로 사용 가능한 유니코드 한자 폰트는 27,496개이다.[7, 8]

한국불교전서를 비롯한 한적위주의 고문헌에는 한자 외에 유사 한자 기호나 원문자 등 다양한 형태의 기호들도 발견되었다. 그림 1은 현재 누락문자 데이터베이스에 등록되어 있는 한자 이미지와 특수기호에 대한 이미지의 예를 볼 수 있다. 이러한 특수기호는 누락문자로 간주하고 이미지로 처리한다.

3. 누락문자 통계

한국불교전서와 한글대장경의 디지털화 과정에서 누락문자 데이터베이스에 등록된 누락문자 이미지는 2003년에 총 2,313 개였으나 2010년에는 총 5,252 개로 늘어났으며, 글자의 총획수는 1획에서 36획까지 다양하며 획수별 빈도수는 <표 1>과 같다.

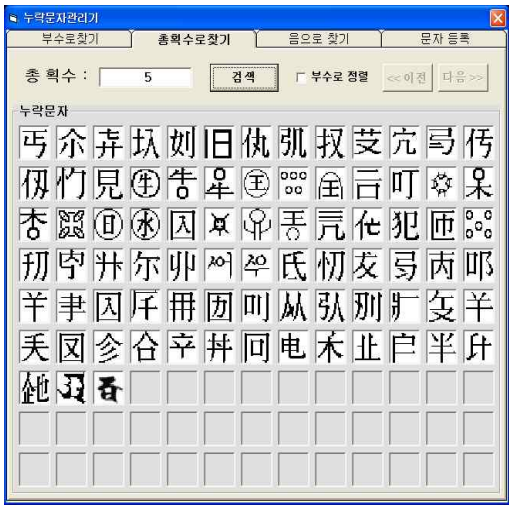
<표 1> 누락문자 이미지 통계(2010.12 현재)

획수	개수	획수	개수
1	137	18	214
2	66	19	167
3	50	20	116
4	54	21	95
5	81	22	70
6	105	23	63
7	178	24	38
8	219	25	15
9	246	26	19
10	936	27	10
11	372	28	10
12	371	29	0
13	337	30	5
14	358	31	3
15	323	34	2
16	340	36	1
17	250	합계	5,252

이들 누락문자 이미지 가운데 800개 이상은 확장된 유니코드에 포함된 글자로 판명되었으나 입력시점에 폰트가 존재하지 않았던 글자에 해당한다.

즉, 원문에 대한 HTML 문서에는 어떤 글자는 누락문자 이미지로 표현된 것도 있고, 유니코드로 표현된 것도 있다. 즉, 동일한 글자가 약간 다른 이미지로 표현된 경우가 있으며, 웹에 표현되는 것에는 특별한 문제가 없는 것으로 판명되고, 원문 데이터베이스 수정 시에 발생하는 비용 등의 이유로 원문에 대한 HTML 문서에서 누락문자 이미지 태그를 삭제하지는 않았다.

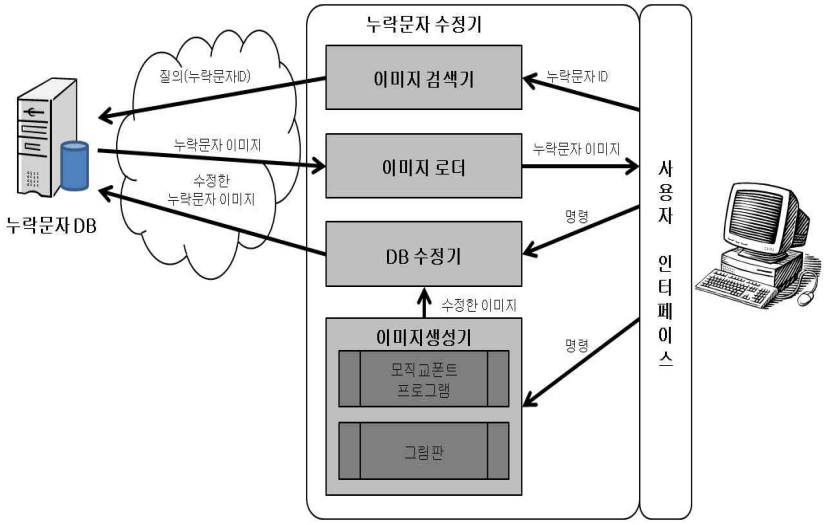
4. 누락문자 이미지 편집



<그림1. 특수기호 및 누락한자 이미지>

누락문자 데이터베이스에 등록된 이미지의 수가 5,000 개 이상으로 늘어나면서 등록된 누락문자의 이미지 가운데는 저장된 이미지 파일을 수정할 필요가 생긴다. 이미지를 수정해야하는 중요한 이유로는 누락문자에 대한 이미지 생성과정에서 이미지의 손상이나 잘못된 이미지의 저장 등의 오류나 이미지를 생성하는 도구의 취약성으로 이미지의 선명도가 낮아진 경우를 들 수 있다.

누락문자 데이터베이스에 저장된 이미지를 수정하려면, 저장된 이미지를 검색하고, 검색된 이미지를 수정하고, 수정된 이미지로 데이터베이스를 수정한다. 누락문자 이미지의 수정기의 구조는 그림 2와 같다.

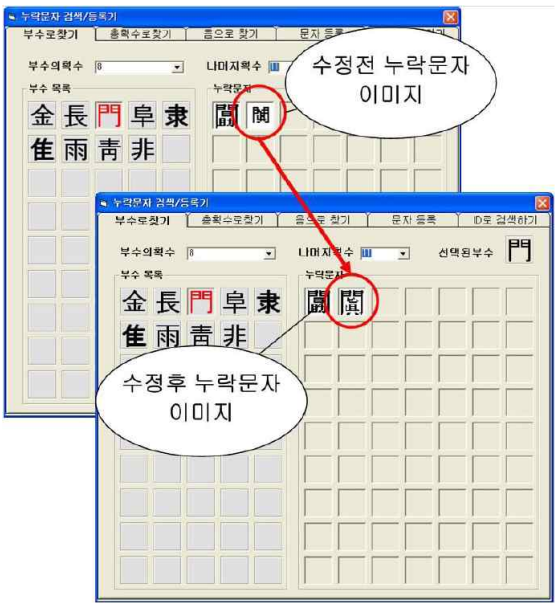


<그림 2 누락문자 이미지 수정기>

그림 3은 오류 이미지를 수정하기 위해서 누락문자 이미지를 검색하고, 수정한 예를 나타낸 것이다.

<그림3 누락문자 이미지 검색 및 수정 예>

그림 4는 수정된 누락문자 이미지를 데이터베이스에 저장하고, 다시 검색한 예를 나타낸 것이다.



<그림4 누락문자 데이터베이스 변경 예>

또한, 누락문자 데이터베이스에 저장된 이미지의 선명도가 낮을 때는 저장된 이미지의 선명도를 개선할 수도 있다. 특히 누락문자의 획수가 많고, 이미지 생성에 사용된 도구 프로그램의 기능상의 한계로 해당 문자에 대한 이미지의 선명도가 낮은 경우에 해당한다. 이러한 경우에 해당 누락문자 이미지의 표현크기를 (24 x 24)비트에서 (100 x 100)비트의 크기로 생성한 다음에 다시 (24 x 24)비트의 크기로 축소하는 변환 프로그램으로 문자 이미지의 선명도를 개선할 수 있다.

그림5는 누락문자 이미지의 선명도를 개선한 예를 보여준다.

<그림 5 누락문자 이미지의 선명도 개선 예>

5. 누락문자 관리시스템

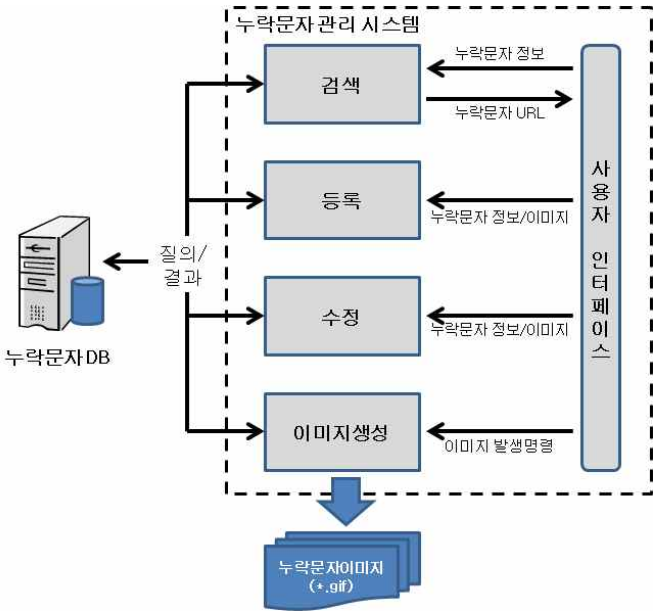
한국불교전서와 같은 한자로 쓰여진 고문헌의 디지털화 과정에서 원문을 HTML로 태깅한 문서의 데이터베이스를 구축한다.

이 과정에서 일차적으로 초기입력이 필요하며, 초기입력은 워드프로세서로 직접 입력하거나, 혹은 스캐너를 사용해서 입력파일을 생성한다.

그 다음에 이것을 원문과 대조하여 누락문자로 판단되면 누락문자 데이터베이스를 검색한다. 만약 데이터베이스에 존재하지 않으면 누락문자로 인정하고, 누락문자 이미지를 생성한 다음에 누락문자 데이터베이스에 등록하고 나서 이미지파일로 원문에 태깅한다.

만약 데이터베이스에 존재하면, 해당 누락문자의 이미지파일로 원문에 태깅한다.

그림 6은 누락문자 관리시스템의 구조를 보여주는 것으로 검색, 등록, 수정 및 이미지 생성 기능으로 구성되어 있다.



<그림6 누락문자 관리시스템>

그리고, 누락문자 데이터베이스에 저장되는 이미지파일의 파일이름은 누락문자 검색을 위한 누락문자 ID 와 연계해서 부여한다.

파일이름 형식은 “K[m][n].gif” 이고, 여기서 m은 해당 누락문자의 총획수에 해당하고, n은 동일한 획수의 누락문자에 부여된 일련번호에 해당한다.

그림 7은 누락문자 이미지파일 이름과 누락문자 ID의 관계를 보여준다.

ID	이미지파일 이름	이미지
0501	k0501.gif	
0502	k0502.gif	
0503	k0503.gif	
0504	k0504.gif	
0505	k0505.gif	
0506	k0506.gif	
0507	k0507.gif	
0508	k0508.gif	
0509	k0509.gif	
0511	k0511.gif	

<그림 7 이미지파일 이름과 누락문자 검색 ID 예>

5. 결 론

1999년 이래 10여년 간 진행된 한국불교전서 와 한글대장경의 디털화 과정에서 유니코드로 표현될 수 없는 누락문자들은 별도의 이미지파일을 생성함으로써 웹서비스를 성공적으로 할 수 있었다. 그 당시의 인터넷 기술이 미비한 시점에서 고문헌 디지털화를 위한 대형 과제수행에서 누락문자 및 특수기호를 이미지로 처리한 것은 기술적 판단이 바르게 되었음을 입증한다.

누락문자 처리를 위한 누락문자 관리시스템을 설계하고, 누락문자 데이터베이스를 구축하는 과정에서 누락문자 이미지파일의 효과적인 관리를 위한 문제들을 검토했다. 특히, 고문헌에 나타난 문자 중 한자로 볼 수 없는 특수문자나 한자에 부가하여 표시된 원문자 등의 문자들도 누락문자

로 간주하고 이미지파일을 생성하여 처리함으로써 고문헌의 웹서비스를 위한 원만한 해결방안이 되었다.

참고문헌

[1] Y.S. Hong, "Searching Missing Characters from the Hanguk Pulgyo Chonso Database", 2001 EBTI International Conference, 2001.

[2] Y.S. Hong, et al. "Development of a Syntax-directed SGML Editor for Processing Korean Ancient Documents"", Prodeedings of 1999 EBTI, ECAI, SEER & PNC Joing Meeting in Taipei, 1999.

[3] Y.S. Hong, "Missing character management for electronic Buddhist texts", 2010 EBTC in Dongguk International Academic Conference, 2010.

[4] Keum Suk Lee, "International cooperation in electronic Buddhist Texts Retrieval Sysrem", 2010 EBTC in Dongguk International Academic Conference, 2010.

[5] D. Juang, J.-H. Wang, C.-Y. Lai, C.-C. Hsieh, L.-F. Chien, and J.-M. Ho, "Resolving the unencoded character problem for chinese digital libraries", JCDL'05, June 7-11, Denver, Colorado, p.311-319, 2005.

[6] S. Wu and S. Zheng, " An XML-based approach for Chinese character glyph description", First international Workshop on Education Technology and Computer

[7] Richard Cook , "UniHan Variation: Issues and Solutions", 23rd Internationalization & Unicode Conference, 2002.

[8] The Unicode Consortium, http://unicode.org/, Science, 2009.

[9] In Sub Hur, "Report on the Digital Tripitaka Koreana 2001", 2001 EBTI International Conference, 2001.

[10] 홍영식, “한국불교 전산화에 있어서 누락문자 관리 시스템”, 전자불전, 제 9 집, p.29-44, 동국대학교 전자불전문화재콘텐츠연구소, 2007.

[11] 이금석, 이용규, 홍영식, 한태식, “한국불교전서 전산화를 위한 누락문자 처리방안“, 동국대학교 전자불전연구소 제 2회 세미나, 2000.

[12] 이금석, 이용규, 홍영식, 한태식, “한글대장경 검색시스템”, 동국대학교 전자불전연구소 제 4회 세미나,

2002.

[13] 이용규, 이금석, 홍영식, 한보광, “한국불교전서 전산화”, 동국대학교 전자불전연구소 설립기념 세미나, 1999.

키워드: 유니코드, 누락문자 관리 시스템, 한국불교전서, 한글대장경, 웹기반 검색 시스템

Abstract

Missing Character management for the Electronic Buddhist Texts

Young Sik Hong
Dept. of Computer Science & Engineering
Dongguk Univ-Seoul

Managing missing characters is one of the big issues in digitalization of the Hanguk Bulgyo Chonso and the Hangul Tripitaka due to the limited number of Unicode for Chinese characters.

In this paper, we describe the functional structure of the missing character management system and the related technical schemes for handling missing characters for Web services. Also, we present sub-operations for managing missing characters such as character image creation, registration, retrieval and image editing, which have been developed during the last 10 years in digitalization of the Hanguk Bulgyo Chonso and the Hangul Tripitaka.

Keywords: Unicode, missing character management system, Hanguk Pulgyo Chonso, Hangul Tripitaka, web-based retrieval system

논문투고일: 2010.11.30. 심사완료일: 2010.12.16. 게재확정일: 2010.12.18.