Searching Missing Characters from the Hanguk Pulgyo Chonso Database

Young Sik Hong, Keum Suk Lee, Yong Kyu Lee
Computer Engineering Dept. and Electronic Buddhist Text Institute(EBTI)

Tae Sik Han(Ven.Bo Kwang Sunim)
Seon Studies Dept. and Electronic Buddhist Text Institute(EBTI)

Dongguk University, Korea

Abstract: Several approaches handling missing characters in digitalization of ancient documents written in Chinese characters, such as Hanguk Pulgyo Chonso, have been attempted for the past several years. Since 1998, The Electronic Buddhist Text Institute(EBTI) at Dongguk University has conducted a research project to digitalize the Hanguk Pulgyo Chonso based on the Unicode standard, which can be accessed on the World Wide Web.

In building the database for storing web pages for the Hanguk Pulgyo Chonso, missing characters represented in the corresponding image files were also included in HTML documents as a form of image tags, which will be displayed with the Unicode characters on a monitor by the web browser. Even though it was possible to retrieve a part of Chinese texts including missing characters, until now, we could not retrieve documents by keywords with missing characters.

We redesigned the the retrieval system in which keywords with missing characters were also entered into the keyword dictionary. That is, by storing keywords including the image tags for missing characters in the index table, we can retrieve documents by keywords with missing characters.

Most technical problems concerning digitalization of the Hanguk Pulgyo Chonso were resolved by redesigning our retrieval system. Therefore, we can access the Hanguk Pulgyo Chonso on the WWW, regardless of missing characters.

1. Introduction

Most Korean ancient documents including the Hanguk Pulgyo Chonso and the Koryo Buddhist Canon have been written in Chinese characters. Several institutions, such as the Research Institute for Tripitaka Koreana and the Electronic Buddhist Text Institute of Dongguk University(the Dongguk EBTI) build their own full-text databases of the Korean ancient documents.

The Dongguk EBTI has developed a new text editor based on the Unicode to digitalize the

Hanguk Pulgyo Chonso[2][3] Even though the Unicode supports around 27,000 CJK characters, many CJK characters were not assigned their codes. In order to handle the missing character problems, the Dongguk EBTI developed the missing character manager where missing characters were registered and their font image files stored in a local database[5]. But keywords with missing characters could not be entered into the keyword dictionary for the database, that is, texts including missing characters could not be retrieved directly by keywords with missing characters.

We redesigned our retrieval system in which keywords including missing characters were also entered into the keyword dictionary. That is, by storing keywords including the image tags for missing characters in the index table, we can retrieve documents by keywords with missing characters.

Section 2 describes the missing character manager and section 3 presents the index table of keywords with missing characters. In section 4, we explain the retrieval of Buddhist documents by keywords with missing characters. Finally, conclusions and future work appear in section 5.

2. Management of missing characters

The Unicode standard uses two bytes to represent a Chinese character and has codes for around 27,000 CJK characters. We chose the Unicode system to input all CJK characters in the Hanguk Pulgyo Chonso. Until now, more than 100 CJK characters have been found to be missing characters.

A font image file has been prepared for each missing character and these image files have been stored in the local database. If a new missing character is found, then it is searched from the local database for missing characters. If the corresponding font image is not found in the local database, then a new font image file is created and is added to the local database. Figure 1 shows the overall control flow for th missing character manager.

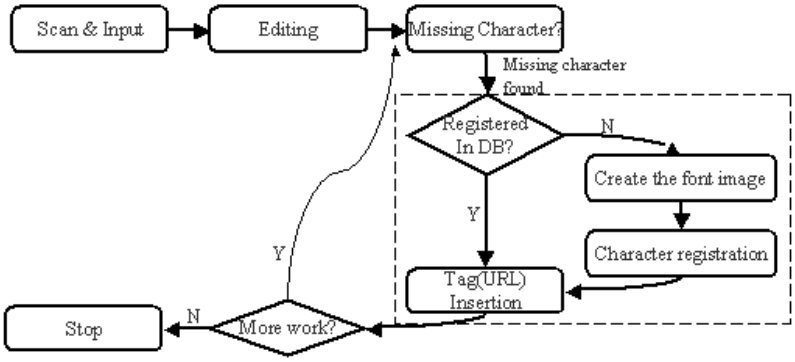

Figure 1. Control flow of the missing character manager

The input texts for the Hanguk Pulgyo Chonso have been stored as the form of HTML texts with image tags for each missing characters and thus could be displayed in the web browsers.

3. Building the index table of keywords with missing characters

The database for the retrieval system consists of several tables such as the table of source texts and the index table of keywords. Keywords extracted from the input source texts are inserted into the keyword table. Figure 2 represents briefly the process of creating the database for the retrieval system.

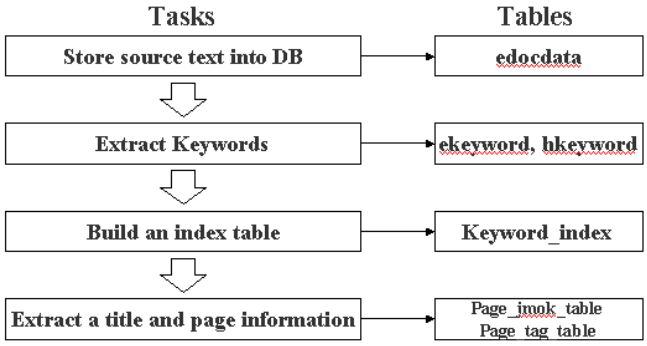

Figure 2. Creating tables for DB

An entry in the keyword table has other positional information for retrieval such as a volume number, a page number, a paragraph number and a line number in addition to the keyword. Figure 3 shows a part of the actual keyword table.

Keyword	Volume	Page	Paragraph	Line
伽經	4	187	2	12
大覺國師	4	549	3	15
法海	1	69	1	14
如來	4	808	1	7
淨土	1	151	1	8
華嚴論	4	768	3	1

Figure 3. Keyword table

With the tables of keywords including missing characters, it is possible to retrieve texts by keywords with missing characters. In order to create the index table with keywords including missing characters, the image tag corresponding to the missing character should be allowed to enter into the keyword table. For example, a keyword containing a missing character, 疸斯 can be represented as follows:

<img src=http://ebti.dongguk.ac.kr/images/K1005.GIF>斯

A part of the source text including the above missing character can be represented as follows:

於一時在婆羅***<img src=http://ebti.dongguk.ac.kr/images/K1005.GIF>***斯仙人墮處施鹿林

Figure 4 shows a part of the table of keywords with missing characters, where the circled missing characters are represented as the corresponding image tags.

Keyword	Volume	Page	Paragraph	Line	Ref
<img src=http://ebti.dongguk.ac.kr/images/K1005.GIF>**斯**	1	281	2	19	疸斯
羅<img src=http://ebti.dongguk.ac.kr/images/K1412.GIF>**佉**	1	223	2	1	羅佉法
<img src=http://ebti.dongguk.ac.kr/images/K904.GIF>**比丘**	1	202	1	9	鵨比丘
於波羅<img src=http://ebti.dongguk.ac.kr/images/K1108.GIF>	1	218	3	19	於波羅奈

Figure 4. Table of keywords with missing characters

In order to reduce the required amount of storage for the keyword table, each long image tag can be represented as a short form. Figure 5 shows a part of simplified keyword table.

Keyword	Volume	Page	Paragraph	Line	Ref
<K1005.GIF>斯	1	281	2	19	寔斯
羅<K1412.GIF>佉	1	223	2	1	羅[illegible]佉
<K904.GIF>比丘	1	202	1	9	鴝比丘
於波羅<K1108.GIF>	1	218	3	19	於波羅[illegible]

Figure 5. Simplified keyword table

By using the simplified keyword table, the keyword searching speed can be improved in addition to saving the amount of the needed storage for the keyword table.

4. Searching Database

Since the extended table of keywords with missing characters has been created, the documents stored in the database can be retrieved by the string searching algorithm. A new interface has been redesigned to support the extended search operation with keywords having missing characters. Figure 6 shows the redesigned interface of the retrieval system, where a list of keywords with missing characters are displayed.

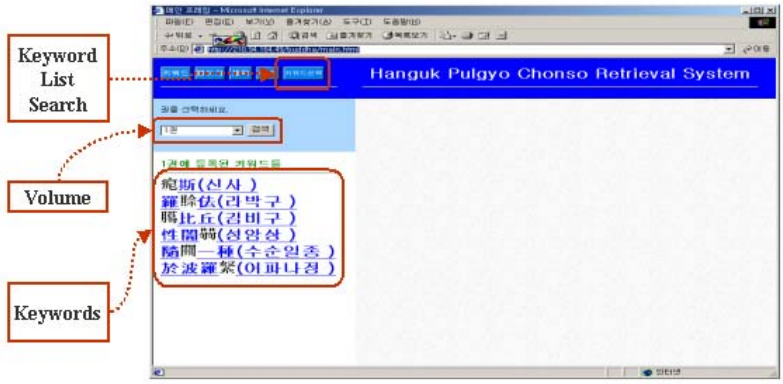

Figure 6. Subwindow for selecting a searching key

Figure 7 shows the retrieved documents using a keyword with a missing character.

Figure 7. Retrieved texts using a keyword with missing characters

Circled words in the search results correspond to the given keyword and the lower left part of the subwindow shows the positional information for them.

5. Conclusions

A research project to digitalize the Hanguk Pulgyo Chonso has been started in 1998. Until now, we have developed a text editor based on the Unicode to input the documents, a retrieval engine working on the Internet and the missing character manager.

In order to add a searching operation using keywords with missing characters, the retrieval system was redesigned. A missing character was represented by the corresponding image tag and keywords including the image tags were stored in the keyword table. A new searching algorithm was developed with this keyword table. Most technical problems concerning digitalization of the Hanguk Pulgyo Chonso seem to be resolved by adding the extended searching operation to the retrieval system.

Various kinds of font images with the different size and color for missing characters need to be added to the retrieval system in the future.

References

[1] Dongguk University Press, The Korean Ancient Buddhist Corpus, 1979.

[2] Y. S. Hong, et al., ”Development of the technologies for Korean Ancient Document Management and Retrieval on the Web”, Project final report, Ministry of Information and Communications, 1998.

[3] Y. S. Hong, et al., “Development of a Syntax-directed SGML editor for processing Korean Ancient Documents”,

[4] Chu-Ren Huang, Keh-Jiann Chen and Shin Lin, “Corpus on Web: Introducing The First Tagged and Balanced Chinese Corpus”, PNC Special Meeting in Tapei, Feb. 17-19, 1997.

[5] Y. K. Lee, et al., “The Hanguk Pulgyo Chonso and the Hangul Tripitaka(the Korean Ancient Buddhist Corpus and the Korean Translation of the Koryo Buddhist Canon) on the WWW”,

[6] Michael Murry, “Unicode Issues and the Input of ‘Han’ Character Texts”, The 4th EBTI Meeting Kyoto, Oct.23-26, 1997.

[7] C. Witten, “SMART Project and the Database of Chinese Buddhist Texts”, The 4th EBTI Meeting Kyoto, Oct.23-26, 1997.