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

Yong Kyu Lee, Keum Suk Lee, Young Sik Hong
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, Republic of Korea

Abstract

The Electronic Buddhist Text Institute (EBTI: **電子佛典研究所**) at Dongguk University (**東國大學校**) was established in 1997 with the mission of the digitalization of Buddhist documents of Korea. As the first step, EBTI has planned to digitalize the Hanguk Pulgyo Chonso (the Korean Ancient Buddhist Corpus, **韓國佛教全書**) and the Hangul Tripitaka (the Korean Translation of the Koryo Buddhist Canon, **韓譯 高麗大藏經**) on the World-Wide Web by the year of 2006.

The Hanguk Pulgyo Chonso, published by Dongguk University, is a collection of 288 documents written in CJK characters by 171 monks and scholars throughout Silla (**新羅**), Koryo (**高麗**), and Chosun (**朝鮮**) dynasty. The Chonso consists of twelve volumes totalling around 10,000 pages. Also published by Dongguk University, the Hangul Tripitaka is a Korean translation of the Hae-in Monastery version of the Koryo Buddhist Canon, which is composed of 280 volumes with about 140,000 pages.

In order to digitalize the Chonso and the Tripitaka, we are currently developing a text editor based on the Unicode and a retrieval engine on the WWW. Using the editor and the retrieval system, we are planning to input some parts of the Chonso and the Tripitaka and to open to the public through the Internet.

1. Introduction

Dongguk University (**東國大學校**), founded by the Chogye Order of Korean Buddhism (**大韓 佛教 曹溪宗**) in 1906, has a large number of Buddhist documents in the library, and has published many Buddhist texts for the past dozens of years including two invaluable books, the Hanguk Pulgyo Chonso (the Korean Ancient Buddhist Corpus, 韓國佛教全書) [1] and the Hangul Tripitaka (the Korean Translation of the Koryo Buddhist Canon, 韓譯 高麗大藏經) [2].

The Hanguk Pulgyo Chonso is a collection of 288 documents written in CJK characters by 171 monks and scholars throughout Silla (新羅), Koryo (高麗), and Chosun (朝鮮) dynasty. It consists of twelve volumes totalling around 10,000 pages. The Hangul Tripitaka is a Korean translation of the Hae-in Monastery (海印寺) version of the Koryo Buddhist Canon, which is composed of 280 volumes with about 140,000 pages.

With the mission of the digitalization of Buddhist documents of Korea and the construction of Buddhist digital libraries, the Electronic Buddhist Text Institute (EBTI: 電子佛典研究所) at Dongguk University was established in 1997. EBTI is interdisciplinary containing 17 professors at various departments, such as Buddhist Studies, Seon Studies, Korean Language and Literature, Chinese Language and Literature, Computer Engineering, etc. As the first step, EBTI has planned to completely digitalize the Hanguk Pulgyo Chonso and the Hangul Tripitaka on the World-Wide Web by the year of 2006 in commemoration of the centennial anniversary of the opening of Dongguk University.

For the digitalization of the Chonso and the Tripitaka, we are currently developing a text editor [4, 9]. The text editor is based on the Unicode [10] and includes a CJK character input system in which a user first inputs a Korean character from a keyboard and then converts it into a CJK character. That is because every CJK character has a corresponding Korean character and all of the Korean word processors use this approach. We have adopted the Unicode since it supports more than 20,000 CJK codes while the Korean standard KSC-5601 [6] supports only 4,888 CJK codes. Our editor has limited basic functions compared to commercial word processors. However, it contains some additional functions to mark up XML tags. Inputted documents are processed to find out keywords and partitioned to be stored into the SQL Server DBMS. Indexes are also stored in the database.

Also we are developing a retrieval system [9] which can retrieve documents from the database using Netscape and Internet Explorer on the World-Wide Web. The search can be performed by using keywords or the table of contents. Using the editor and the retrieval engine, we are planning to input some parts of the Chonso and the Tripitaka and to open to the public through the Internet.

In Korea, the digitalization of Buddhist documents was begun for several years ago when the

Research Institute of Tripitaka Koreana (高麗大藏經研究所) was established to computerize the Hae-in Monastery version of the Koryo Buddhist Canon (海印寺 高麗大藏經, Tripitaka Koreana) [7, 8]. The input of the canon has been almost finished using a commercial word processor called Hun-min-jung-eum. Now, the project is in the proofreading stage with a plan to open through the Internet in the future.

The contents of this paper are organized as follows according to the sections: In the next section, we discuss about the Hanguk Pulgyo Chonso and the Hangul Tripitaka. Section 3 presents the character code systems of Korea. A Unicode character input system developed in this research is represented in Section 4. Section 5 describes the storage of the Buddhist documents using a DBMS. In Section 6, we explain the retrieval of Buddhist documents on the WWW. Finally, our conclusions and future work appear in Section 7.

2. The Hanguk Pulgyo Chonso and the Hangul Tripitaka

The Hanguk Pulgyo Chonso (韓國佛教全書), published by Dongguk University Press (東國大學校 出版部) from 1979 to 1998, is a collection of 288 invaluable documents written in CJK characters by 171 monks and scholars, such as Wonhyo (元曉), Uichon (義天), and Jinul (知訥), throughout Silla (新羅), Koryo (高麗), and Chosun (朝鮮) dynasty of Korea. The Chonso consists of twelve volumes totalling around 10,000 pages with each page containing about 1,000 characters. Thus, the number of characters of it reaches to about 10 million.

Published by Dongguk University Translation Center of Buddhist Scriptures (東國 譯經院) from 1966 to 1998, the Hangul Tripitaka (韓譯 高麗大藏經) is a Korean translation of the Hae-in Monastery version of the Koryo Buddhist Canon (海印寺 高麗大藏經), which is composed of 280 volumes totalling around 140,000 pages. With about 500 Korean characters in a page, the total number becomes around 70 million characters.

The digitalization of the Hanguk Pulgyo Chonso is more tricky than that of the Hangul Tripitaka as the former is written in CJK characters only while the latter is written in Korean characters and handling these on the word processors and the web browsers causes no problem. In the remainder of this paper, the main issues will be related to the digitalization of the Hanguk Pulgyo Chonso. Computerization of the Chonso contains many problems to be solved. The major problem is that the number of CJK codes supported by the Korean standard KSC-5601 [6] is not enough for the Chonso. This makes us consider another character code system and it will be discussed in the next section. The other stereotypical problems occurred during this kind of projects can be found in the literature [8, 11, 12, 13].

3. Character Code Systems of Korea

The KSC-5601 code system [6], standardized in 1987, uses two bytes to represent a character. It has codes for 2,350 Korean characters, 4,888 CJK characters, 1,128 special characters, and other 470 characters. In 1991, the standard was extended as the name of the KSC-5657 code system to add new characters to the KSC-5601 standard including 1,930 Korean characters, 2,856 CJK characters, and 1,677 ancient Korean characters. However, the standard has two drawbacks. The first one is that it cannot represent all Korean characters. The second one is that it can represent only 7,744 CJK characters and this is not enough to input the Hanguk Pulgyo Chonso.

To overcome the first problem, another standard, called the KSC-5601 composition code system, has been standardized in 1992. Even though the standard also uses two bytes to code a character, the coding method is different from the previous ones. Here, one character is represented by combining three 5-bit codes each of which representing a Korean alphabet. For example, the Korean character (HAN) is the composition of (H), (A), and (N). However, this standard is not used by commercial systems because the code is not compatible with the KSC-5601. Currently, most commercial word processors adopt the KSC-5601 standard.

To fix the second problem, the commercial word processors found their own solutions. That is, they have extended the KSC-5601 standard to represent more CJK characters. Now, widely used word processors have around 15,000 CJK characters. However, their extended codes are not compatible with each other and the extended CJK characters can not be displayed on the web browsers. Thus, we have adopted another character code system, the Unicode standard [10].

The Unicode standard also uses two bytes to represent a character and has codes for 20,902 CJK characters, which is much more than those of the KSC-5601 standard. Even though this number is not enough yet for inputting all CJK characters in the Hanguk Pulgyo Chonso, we have chosen it since it could represent most of the frequently appearing characters. Thus, infrequent missing characters found in the Chonso could be inputted using unique identifiers in a Chinese dictionary. However, we have to develop a new text editor for ourselves as we have failed to find a Unicode editor which can be used for the project.

4. Unicode CJK Character Input System

We have defined an XML DTD describing the logical structure of the Hanguk Pulgyo Chonso as shown in Figure 1. The DTD is rather simple because it is defined to present only major constructs of the Chonso, which has to be extended in the future. The DTD includes elements for book, author, article, title, body, page, column, comment, keyword, etc. A part of a sample document instance according to the DTD is presented in Figure 2. The document instance is a part of the collection of works by Uichon (**義天**, **大覺國師文集**).

```
<!ELEMENT Book (BookTitle,Author,(Article)+)>
<!ELEMENT BookTitle (Line)+>
<!ELEMENT Author (Line)+>
<!ELEMENT Article (Title,Body)>
<!ELEMENT Title (Line)+>
<!ELEMENT Body (Page)+>
<!ELEMENT Page (Column)+>
<!ATTLIST Page PageNum CDATA #REQUIRED>
<!ELEMENT Column (Content,Comment?)>
<!ATTLIST Column ColumnNum (1|2|3) #REQUIRED>
<!ELEMENT Content (Line)+>
<!ELEMENT Comment (Line)+>
<!ELEMENT Line (Keyword?,Text)+>
<!ATTLIST Line LineNum CDATA #IMPLIED>
<!ELEMENT Keyword (#PCDATA)*>
<!ATTLIST Keyword Continue (True|False) "False">
<!ELEMENT Text (#PCDATA)*>
```

Figure 1. An XML DTD for the Hanguk Pulgyo Chonso

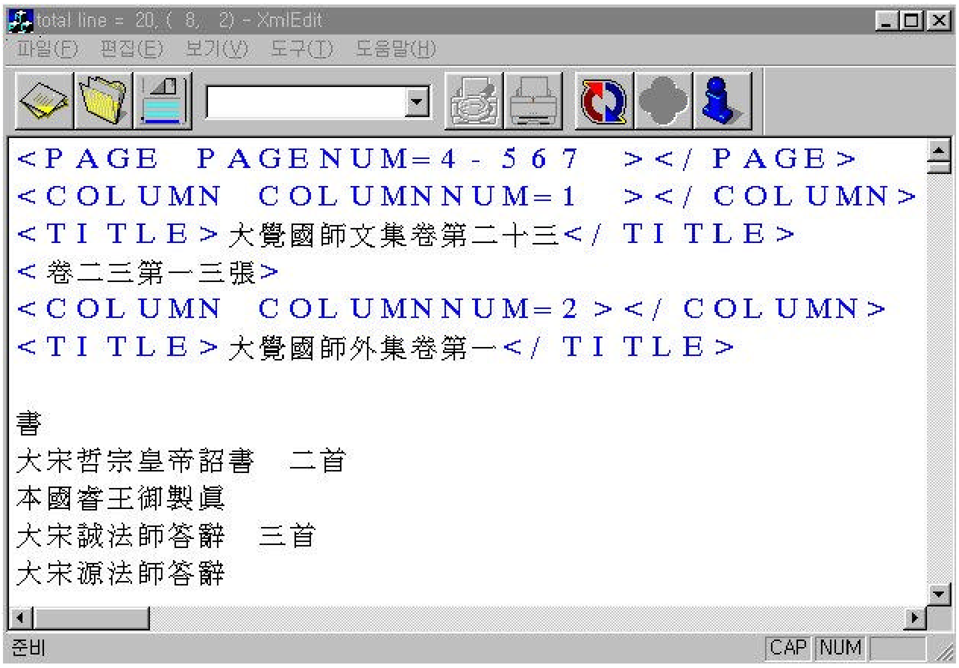
```
<?xml  version="1.0"?>
<!DOCTYPE Book SYSTEM "Chonso.dtd">
<Book>
    <BookTitle> <Line> <Text>大覺國師文集</Text> </Line></BookTitle>
```

```
      <Author> <Line> <Text>大覺國師  義天</Text> </Line> </Author>
      <Article>
           <Title>
            <Line LineNum="1"> <Text>大覺國師外集卷第一九</Text></Line>
         </Title>
           <Body>
          <Page  PageNum="561">
          <Column  ColumnNum="2">
           <Content>
          <Line  LineNum="2"><Text>和金官侯讀金剛經</Text></Line>
          <Line  LineNum="3"><Text>以起信論本贈瀧西相國</Text></Line>
           </Content>
           </Column>
           </Page>
         </Body>
      </Article>
   </Book>
```

Figure 2. A Sample Document According to the DTD

In order to input and process the Hanguk Pulgyo Chonso, we have developed a text editor as shown in Figure 3. The editor includes menus for some basic functions as follows:

- File menu - New, Open, Save, Save-As, Print, Print-Preview, Exit
- Edit menu - Cut, Copy, Paste, Find-Convert
- Tool menu - Korean-CJK Convert
- Help menu
- Tag List Box - XML DTD Tags

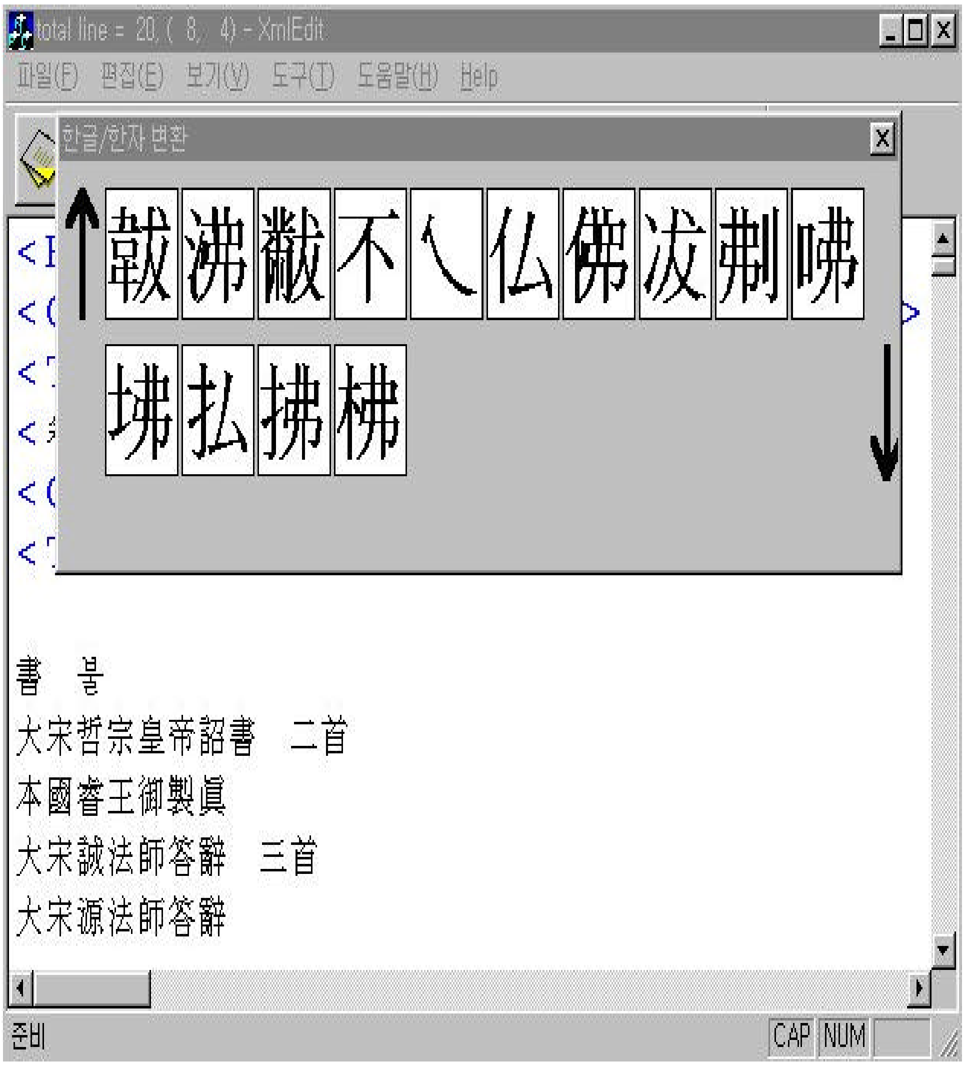


```
＜ＰＡＧＥ　ＰＡＧＥＮＵＭ＝４－５６７　＞＜／ＰＡＧＥ＞
＜ＣＯＬＵＭＮ　ＣＯＬＵＭＮＮＵＭ＝１　＞＜／ＣＯＬＵＭＮ＞
＜ＴＩＴＬＥ＞大覺國師文集卷第二十三＜／ＴＩＴＬＥ＞
＜卷二三第一三張＞
＜ＣＯＬＵＭＮ　ＣＯＬＵＭＮＮＵＭ＝２＞＜／ＣＯＬＵＭＮ＞
＜ＴＩＴＬＥ＞大覺國師外集卷第一＜／ＴＩＴＬＥ＞

書
大宋哲宗皇帝詔書　二首
本國睿王御製眞
大宋誠法師答辭　三首
大宋源法師答辭
```

Figure 3. Text Editor Window

As can be seen above, the editor contains a list box for predefined XML tags as well as menus for edit functions. From the Tag List Box, a user can select an XML tag in order to mark up the Chonso according to the DTD. The text editor is based on the Unicode as mentioned before. It includes a CJK character input system in which the user first inputs a Korean character from a keyboard and then converts it into a CJK character as shown in Figure 4. That is because we cannot input a CJK character directly from the keyboard. More detailed explanation about the input system and the editor can be found in [4].



書　불
大宋哲宗皇帝詔書　二首
本國睿王御製眞
大宋誠法師答辭　三首
大宋源法師答辭

Figure 4. Text Input System Window

Using the text editor, we have started inputing the collection of works by Uichon (**義天, 大覺國師文集**) and Jinul (**知訥, 普照國師集**), which is in Volume 4 of the Hanguk Pulgyo Chonso.

5. Storage of Buddhist Documents Using a DBMS

The inputted document using the text editor has to be processed to be stored into the database and to build an index structure for content based retrieval using keywords. We have used the SQL Server 6.5 relational database management system on the Windows NT 4.0 for the storage of documents and indexes. The procedure of this stage is described in Figure 5.

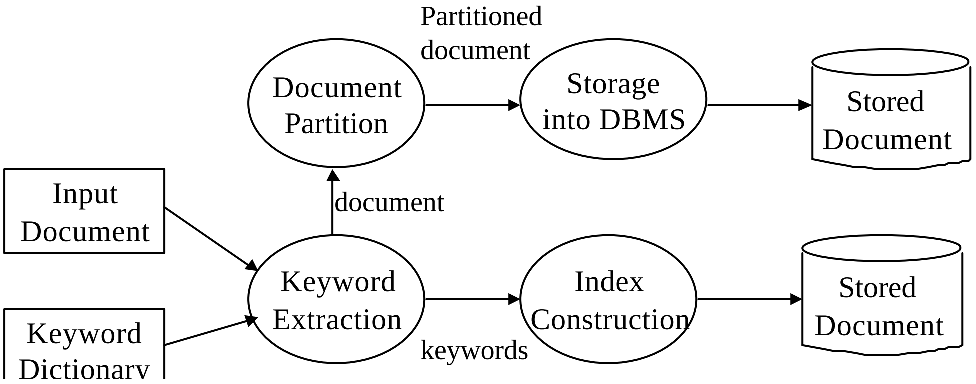
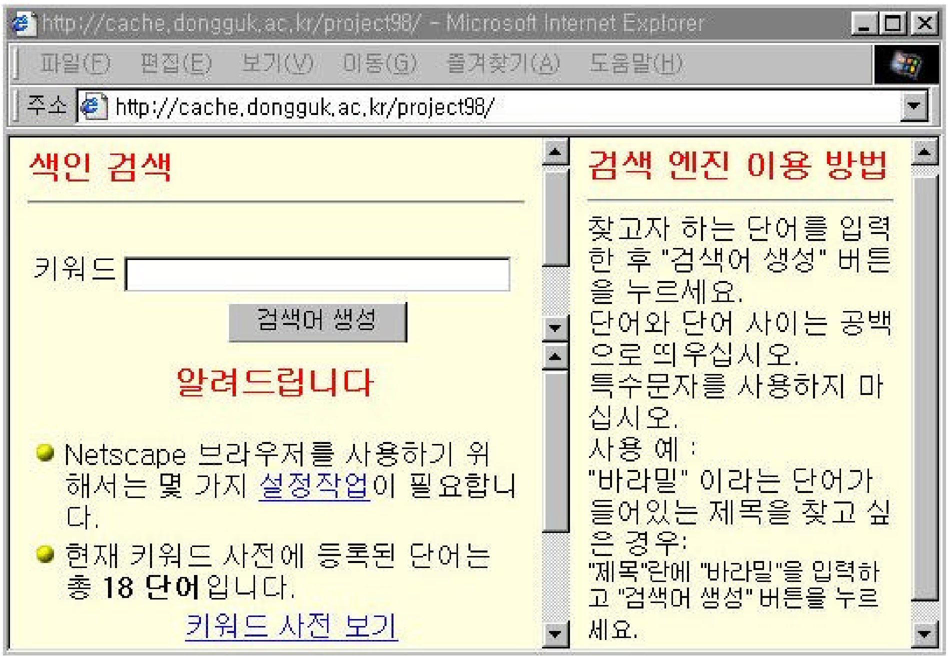


Figure 5. Storage of a Document into the Database

First, the input document is scanned to find out keywords using the keyword dictionary. The extracted keywords are inserted into the index structure. For easy implementation, we have used the DBMS for the index storage. Thus, the index is also a table in the database. During this stage, the keyword dictionary, a text file in the system, may be augmented if a new keyword marked up by the user is found. Then, the keyword-extracted document is partitioned to be stored into the database. The partition is performed based on the logical structure and the storage unit is a column in a page. That is, a column is stored separately in a field of the database table and it becomes the retrieval unit from the database. In the Hanguk Pulgyo Chonso, a page consists of three columns.

6. Retrieval of Buddhist Documents on the WWW

The documents stored in the database can be retrieved by using keywords or by using the table of contents. Until now, we have implemented the retrieval by only keywords and the other part will be added later. The interface uses Korean language even though it will be replaced by English in the future. Figure 6 shows the interface of the retrieval system. Both of Netscape and Internet Explorer can be used to access the database server. Presently Boolean *and* queries are supported and Boolean *or* queries will also be supported. A search example by using a keyword is shown in Figure 7. In the edit line, the user can enter a keyword or a group of keywords in Korean, which have to be converted into CJK characters.


http://cache.dongguk.ac.kr/project98/ - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 이동(G) 즐겨찾기(A) 도움말(H)
주소 http://cache.dongguk.ac.kr/project98/

색인 검색

키워드

알려드립니다

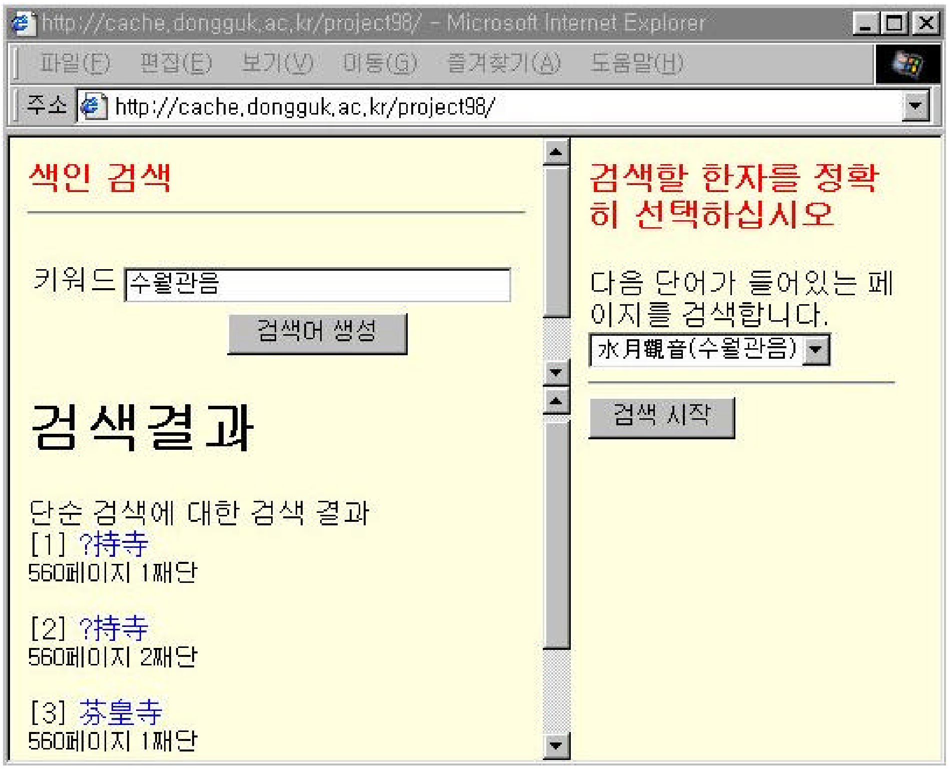
- Netscape 브라우저를 사용하기 위해서는 몇 가지 설정작업이 필요합니다.
- 현재 키워드 사전에 등록된 단어는 총 **18 단어**입니다.

키워드 사전 보기

검색 엔진 이용 방법

찾고자 하는 단어를 입력한 후 "검색어 생성" 버튼을 누르세요.
단어와 단어 사이는 공백으로 띄우십시오.
특수문자를 사용하지 마십시오.
사용 예 :
"바라밀" 이라는 단어가 들어있는 제목을 찾고 싶은 경우:
"제목"란에 "바라밀"을 입력하고 "검색어 생성" 버튼을 누르세요.

Figure 6. User Interface of the Retrieval System


http://cache.dongguk.ac.kr/project98/ - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 이동(G) 즐겨찾기(A) 도움말(H)
주소 http://cache.dongguk.ac.kr/project98/

색인 검색

키워드 수월관음

검색결과

단순 검색에 대한 검색 결과
[1] ?持寺
560페이지 1째단

[2] ?持寺
560페이지 2째단

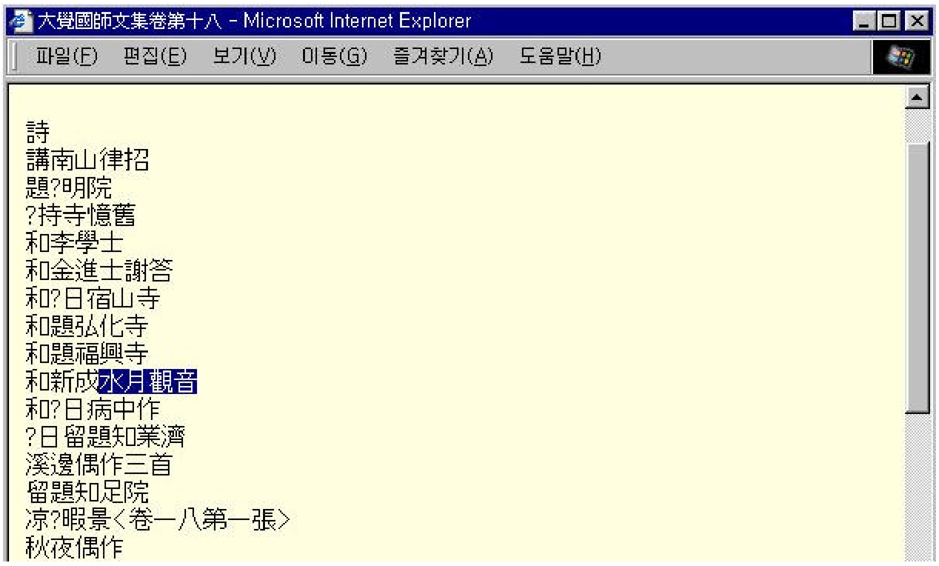
[3] 芬皇寺
560페이지 1째단

검색할 한자를 정확히 선택하십시오

다음 단어가 들어있는 페이지를 검색합니다.
水月觀音(수월관음)

Figure 7. Search by Keywords

The search results appear in the bottom of the left window of Figure 7. Among the results, the user may select one to browse as shown in Figure 8.


大覺國師文集卷第十八 - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 이동(G) 즐겨찾기(A) 도움말(H)

詩
講南山律招
題?明院
?持寺憶舊
和李學士
和金進士謝答
和?日宿山寺
和題弘化寺
和題福興寺
和新成水月觀音
和?日病中作
?日留題知業濟
溪邊偶作三首
留題知足院
凉?暇景<卷一八第一張>
秋夜偶作

Figure 8. Search Result

7. Conclusions

In July 1998, we started a project to digitalize the Hanguk Pulgyo Chonso (**韓國佛教全書**) and the Hangul Tripitaka (**韓譯 高麗大藏經**), both of which were published by Dongguk University (**東國大學校**). Until now, we have implemented a text editor based on the Unicode to input the documents and a retrieval engine for search on the Internet. We have adopted the Unicode since it has 20,902 CJK characters, which is much larger than 4,888 CJK characters supported by the KSC-5601 standard. The system has been developed using Microsoft Visual C++ 5.0 on the Windows NT 4.0 with SQL Server 6.5 relational DBMS. The retrieval system can be accessed by using Netscape or Internet Explorer on the World-Wide Web.

Using the editor, we have inputted most parts of the collection of works by Uichon (**義天, 大覺國師文集**) and Jinul (**知訥, 普照國師集**), which is in Volume 4 of the Hanguk Pulgyo Chonso. After completing the input of the collection, we are going to input the collection of works by Wonhyo (**元曉**).

The system has many things to be improved, as it is in the beginning stage. We have a plan to complete the whole project by 2006, and we hope to open the digitalized parts of the documents through the Internet next year.

References

[1] Dongguk University Press, "The Hanguk Pulgyo Chonso (韓國佛教全書)," Vol. 1-12, 1979-1998. (In CJK)

[2] Dongguk University Press, "The Hangul Tripitaka (韓譯 高麗大藏經), Vol. 1-280, 1966-1998. (In Korean)

[3] C. F. Goldfarb and P. Prescod, *The XML Handbook,* Prentice Hall PTR, 1998.

[4] Y. S. Hong, K. S. Lee, and Y. K. Lee, "Development of a Syntax-directed SGML Editor for Processing Korean Ancient Documents," 1999 EBTI, ECAI, SEER & PNC Joint Meeting, Taipei, Taiwan, January 18-21, 1999.

[5] 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 Communication, 1998. (In Korean)

[6] S. H. Jeon, "Character Code Systems of Korea," MicroSoftware, Nov. 1998. (In Korean)

[7] Jong Lim Sunim, "Past and Future of the Korean Tripitaka Input Project," [http://members.iWorld.net/hederein/menu21/Report.html], 1997. (In Korean)

[8] Hye Muk Sunim, "Some Problems in Digitalizing the Koryo Buddhist Canon," [http://members.iWorld.net/hederein/menu22/Hye2.html], 1997. (In Korean)

[9] Y. K. Lee, et al., "Construction of a Korean Ancient Text Database," Dongguk Journal, Vol. 37, Dongguk University, Dec. 1998. (In Korean)

[10] The Unicode Consortium, *The Unicode Standard, Version 2.0,* Addison-Wesley Developers Press, 1996.

[11] Urp App, "Guidelines for the Creation of Large Chinese Text Databases," [http://www.iijnet.or.jp/iriz/irizhtml/maketext/guidelin.htm], 1997.

[12] Urp App, "A Look at the Korean Tripitaka Input Project," [http://www.iijnet.or.jp/iriz/irizhtml/ebti/samsung.htm], 1997.

[13] C. Wittern, "Chinese Character Codes: an Update," [http://www.iijnet.or.jp/iriz/irizhtml/multling/codes.htm], 1997.