UTF-8을 이용한 인명한자의 표현과 한글 독음 처리

조영철, 유정원, 변정용
동국대학교 전자계산학과

A Representation of Korean Human Names and Their Written in Chinese Characters Pronunciation Processing Using UTF-8

Young-Choel Cho, Jeong-Won Yoo, Jeong-Yong Byun
Dept. of Computer Science, Dongguk University

요 약

인물 정보 검색 시스템은 사용자에게 종합적이고 정확한 정보와 편리한 사용자 인터페이스를 제공해야 한다. 하지만 현재 웹 상에서 이 시스템의 방대한 확장 한자 자료를 표현하는데 EUC-KR은 많은 어려움을 가지고 있다. 그리고 어려운 인명용 한자와 전문 한자 용어로 인해 일반인들의 사용이 어려웠다. 이를 해결하기 위해 본 논문에서는 확장된 한자를 표기하기 위하여 UTF-8 인코딩 방식을 사용한다. 그리고 사용자가 알기 원하는 한자의 한글 독음 변환 처리를 통해 데이터베이스의 효율성과 사용자가 쓰기 편한 인터페이스를 제공한다.

1. 서론

인터넷과 웹의 활용이 확산됨에 따라 웹에서 제공되는 정보의 범위는 다양해지고 있다. 현재 많은 웹사이트에서 법조, 정계, 재계 등 사회 각 분야의 인물에 대한 정보 서비스 시스템이 개발되고 있다. 이러한 정보 서비스 시스템들은 법조계, 정계, 재계 및 사회 문화의 주요 인사 및 연예인들을 주축으로 하거나 문중 및 역사 인물 분야에서 그 연구가 활발해지고 있다. 이러한 연구들은 각 분야를 연구하는 학자 및 일반인들에게 유용한 정보로 활용된다[1].

각 분야별 인물 정보 검색 시스템은 사용자에게 종합적이고 다양한 정보를 제공할 수 있어야 하고, 시스템 환경의 변화에 따른 업그레이드와 편리한 사용자 인터페이스가 제공되어야 한다[2].

역사 인물 검색 시스템 중의 하나인 조선시대 인물 정보 검색 시스템(HiPIS: Historical People Information System for Chosun Dynasty)[3]은 웹에서 조선시대 인물과 문과 급제자들을 대상으로 각 인물의 관직기록, 개인신상, 거주사항에 대한 정보를 검색하거나 각 인물의 가족 관계를 통해 형성되는 가계도[4]를 살펴볼 수 있는 검색 서비스를 제공한다.

조선시대 인물에 대한 정보를 데이터로 가지고 있는 이 시스템에서는 정보의 대부분이 한자로 저장되어 있어서 검색 결과의 대다수가 한자로 보여지고 있다. 이 한자를 웹 상에 표현하고자 할 때 깨지는 문제가 발생한다. 이를 위해 KS X 1001 영역의 한자를 모두 포함하고 있는 유니코드 영역의 한중일 통합 한자(CJK Unified Ideographs)와 확장 A를 포함한 한자를 표기할 수 있도록 하는 가변적 인코딩 방식인 UTF-8 인코딩을 제안한다.

또한, 어려운 인명용 한자와 전문 한자 용어가 대다수를 차지하고 있어서 사용자가 찾고자 하는 검색 결과를 얻어도 쉽게 읽을 수 없는 문제를 보이고 있다. 이를 위해, 현재 한자와 독음을 같이 병거하는 방법을 쓰고 있다. 하지만 이것은 계속 증가되고 수정되는 데이터로 인해 데이터베이스의 저장용량을 더욱 증가시키고 수정 / 유지보수를 어렵게 한다. 본 논문에서는 한글 독음에 대한 매핑 테이블을 통해 위 문제점을 해결하는 방법을 제안하고자 한다.

본 논문의 구성은 2장에서는 확장 한자 표기에 관한 기존 연구현황을 분석하고, 3장에서는 확장 한자의

표기와 한자의 한글 독음 변환을 위한 시스템 설계에 관한 내용을 기술한다. 4장에서는 확장 한자 표기와 한자의 한글 독음 변환을 위한 시스템 설계에 대한 구체적인 구현과 결과를 기술하고, 마지막 5장에서는 결론 및 향후 나아갈 방향과 과제에 대해 기술한다.

2. 기존 연구현황

본 절에서는 현재 웹 상에서 제공되고 있는 인물 정보 검색 시스템의 확장 한자 처리와 독음에 대한 방식을 알아본다.

현재 웹 상에서 제공되고 있는 서비스 중 디지털 한국학[5]의 "조선조방목"은 한국의 역대 인물, 조선조방목, 삼국사기, 한국의 왕, 성씨, 가문에 관한 검색 시스템을 제공한다. 조선조방목에서는 한자의 한글 독음 표기를 "한글(한자)"의 형태로 나타내어 데이터 자체에 한글과 한자를 병거하고 있어서 따로 한자를 한글로 변환해 주는 번거로움을 덜어주고 있다.

하지만, 확장 한자 표기에 있어서는 문제점을 보이고 있다. 조선조방목의 인코딩 기준이 EUC-KR (Exetend UNIX Code)[7]로 되어 있어서, KS X 1001에 해당하는 한자만 표기되고 유니코드의 확장 한자 영역의 한자는 16진수 코드값 그대로 화면에 표기되고 있다.

또 다른 서비스인 한국 정신문화연구원의 "사마방목"[6]의 총 입력한자는 1,455,609자로 그 구성을 보면 상용 한자가 1,443,158자, 비상용 한자가 12,451자 이다. 사마방목에서는 데이터베이스 이용의 편의성을 높이기 위해 모든 한자는 한글 독음으로 병거하여, 사용자는 원하는 정보를 한글과 한자의 두 가지 형태로 볼 수 있다. 확장 한자 표기에 있어서는 KS X 1001의 한자는 제대로 표기되고 있지만, 유니코드의 확장 한자 영역의 한자와 유니코드에 없는 한자는 이미지로 처리하여 웹 상에 출력하고 있다.

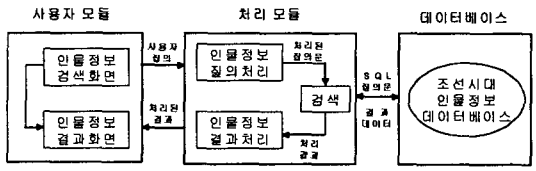
확장 한자의 표기는 문자 세트와 인코딩 방법에 따라서 표기 방법이 달라지는데 기존 시스템에서는 KS X 1001의 한자 영역 4,888자를 사용하고 있어서 확장 한자 표기가 제대로 이루어지지 않았다.

특히, 인명용 한자의 경우 한자의 종류가 많고, 유니코드(Unicode)[8] 영역에 없는 한자도 대다수이기 때문에 검색 서비스를 제공함에 있어서 많은 어려움을 갖는다.

3. 설계

본 시스템의 구성은 크게는 그림 1과 같이 사용자 모듈, 처리모듈 그리고 검색모듈로 나누어져 있다. 본 논문에서 독음처리를 위해서 사용자 모듈을 확장하였고, 확장한자 표기를 위해서 처리모듈과 검색모듈 또한 확장하였다.

아래의 그림 1은 전체 시스템의 처리 과정을 나타내고 있다.



[그림 1] 조선시대 인물 정보 검색 시스템 구성도

3.1 독음처리를 위한 사용자 모듈 확장

조선시대 인물정보검색 시스템의 특성상 대부분의 데이터가 한자로 표기되어 있다. 편리한 사용자 인터페이스 구현을 위해 검색 결과의 한자를 한글로 변환하는 과정이 필요하다. 이를 위해서 확장 한자의 맵핑 테이블을 설계하였고 결과비교를 통해 독음이 이루어지게 설계하였다. 아래의 그림 2는 전체적인 독음처리 과정을 나타내고 있다.

```
① sql 질의문에 대한 검색 결과의 한자 입력
② for( 질의문에 검색 결과의 한자 길이만큼 검색루틴 수행){
      if(결과의 한자 == 매칭 테이블의 한자의 비교){
         결과를 버퍼에 저장
      }
   }
③ 결과를 화면에 출력
```

[그림 2] 독음처리를 위한 알고리즘

3.2 UTF-8 인코딩을 통한 확장 한자 처리

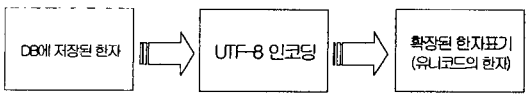
기존의 웹 상에서 제공되고 있는 인코딩 방식이 EUC-KR로 되어 있어서 유니코드 영역에 있는 한자와 유니코드에 없는 한자는 표기 할 수 없다. 유니코드는 프로그램 내부에서 사용하거나 특정 프로그램만이 이해하는 파일 혹은 교환되는 자료의 인코딩으로 적당할 뿐, 일반 텍스트 문서의 인코딩과 다양한 프로그램 사이에 전달되는 텍스트 자료의 인코딩으로서는 부적당하다. 현재 우리나라에서 일반화된 인코딩 표준인 EUC-KR을 대체하기는 곤란하여, 유니코드와 1:1 대응하며 여러 가지 우수한 성질을 갖는 UTF-8 인코딩이 EUC-KR의 대체 인코딩으로 변경되는 추세이다.

UTF-8은 유니코드 문자집합을 8비트 스트림으로 1 바이트에서 6바이트까지 가변적으로 인코딩하는 규약이다. 유니코드를 인코딩하는 방법으로는 UTF-7, UTF-8, UTF-16 등 여러 가지 방법이 있지만, CJK 문자를 사용하는 경우 UTF-8은 상당히 효율적인 인코딩 방법이 될 수 있다. UTF-8 변환 포맷은 세계의 모든 언어를 지원할 수 있고, 지금으로서는 아스키 파일과 역행 호환성을 가지고 있기 때문에 국제화 텍스트 정보를 교환하는데 있어서 가장 유력한 방법이 되고 있다. 가변적 인코딩을 위해서 하나의 유니코드 문자가 몇 바이트로 인코딩 되느냐 하는 것은 유니코드에 할당된 코드의 정수값에 의존한다. 유니코드의 각 문자를 1 ~ 4개의 바이트로 다음의 [표 1]과 같이 인코딩 한다.

[표 1] UTF-8 인코딩

유니코드 (16진수)	인코딩된 바이트	UTF-8로 인코딩 된 바이트 수(2진수)
0000 - 007F	1	0XXXXXXX
0080 - 07FF	2	110XXXXX 10XXXXXX
0800 - FFFF	3	1110XXXX 10XXXXXX 10XXXXXX

본 시스템에서는 UTF-8 인코딩을 적용함으로써 확장한자 처리를 가능하게 하였다. 아래의 그림 3은 데이터베이스의 한자를 UTF-8로 인코딩하여 맵핑 테이블과 비교하여 매칭되는 결과를 웹상에 출력하도록 하는 과정을 보여준 것이다.



[그림 3] 확장 한자 표기를 위한 UTF-8 인코딩 처리 모듈

4. 구현

4.1 한글 독음 처리 루틴 및 인코딩 처리

본 논문에서는 기존의 한글 독음 정보를 데이터베이스에 직접 넣은 방법과는 달리 한자 대 한글 매핑 테이블을 구현하여 이를 통해 검색된 독음을 웹 상에 표현한다.

그림 4의 루틴을 웹 화면에 있는 한자한글변환 버튼에 적용함으로써 사용자가 원하는 독음을 알 수 있게 구현하였다.

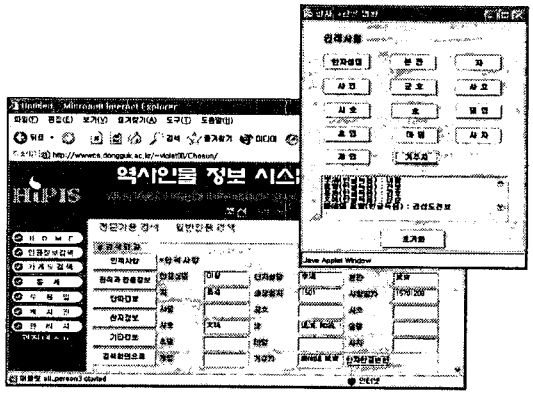
```
public void Hanja_To_Hangul_Search(){
  TextArea.append(all_person.chinese_name.trim()
    +"("+"한글독음"+")"+" "+":"+ " ");

  for(int k=0; k<=all_person.chinese_name.length(); k++){
   for(i=0; i<maching_table범위; i++){
     if(all_person.chinese_name.charAt(k)==matching_table[i][0]){
       str1 = matching_table[i][1];
       TextArea1.append(""+str1);
       str1 = ' ';
     }
   }
  }
   TextArea1.append(""+"\n");
}
```

[그림 4] 한자의 한글 독음 변환 루틴

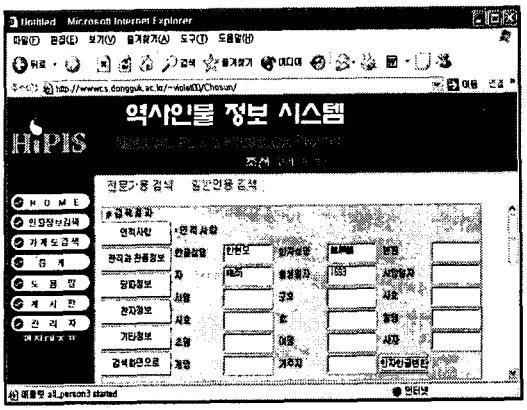
또한, KS X 1001의 범위를 넘어서는 확장 한자의 표기를 위해서 자바언어에서 지원하는 getBytes() 메소드를 이용하여 UTF-8 인코딩을 처리한다. 이를 통해 기존의 이미지를 통해 웹 상에서 출력하는 것과는 달리 텍스트로 출력이 가능하게 되었다.

4.2 실험 결과



[그림 5] 한자의 한글 독음 변환 결과 화면

위의 그림 5는 독음처리에 대한 결과이다. 각 항목에 대한 한글한자변환 버튼을 클릭하면 각 한자에 대한 독음을 확인할 수가 있다. 이를 통해 데이터베이스 저장공간 사용의 효율성을 높이고 수정 및 유지보수를 효율적 관리가 가능하게 되었다.



[그림 6] UTF-8 인코딩을 통한 확장 한자 표기 출력 화면

위의 그림 6은 KS X 1001의 한자 영역에는 있지만 표기되지 않은 한자에 대한 실험결과이다. 데이터베이스에 저장된 "한현모"라는 인물의 한자이름 "韓顯모" 중 한글로 표기 된 부분 "모"에 해당하는 한자를 찾아서 수정하여 UTF-8로 인코딩 과정을 거치면 그림 6에서 처럼 "韓顯謨"라고 표기된다.

5. 결론

본 논문에서는 정보 검색 시스템 중의 하나인 조선시대 인물 정보 검색 시스템의 인명 한자에 초점을 맞추어서 연구를 진행하였다.

확장 한자 표기 실험에서는 확장 한자를 표기할 때 효율적인 인코딩 방식인 UTF-8을 이용하여 KS X 1001의 한자를 포함한 유니코드 영역의 한중일 통합 한자와 확장 한자 표기를 가능하게 하였다. 확장 한자 표기로 인해 한자 데이터의 올바른 입력이 가능해졌고, 한자 출력 부분에 있어서도 확장 한자의 출력이 가능해졌다.

또한, 기존의 인물 정보 검색 시스템의 한자와 독음을 같이 병거하는 방법은 그 특성상 계속 증가되고 수정되는 데이터로 인해 데이터베이스의 저장용량을 더욱 증가시키고 수정 / 유지보수를 어렵게 한다. 본 논문에서는 한글 독음에 대한 매핑 테이블을 통해 데이터베이스의 효율성을 높이고 사용자가 쓰기 편한 인터페이스를 제공하였다.

본 논문에서는 데이터의 양이 방대하여 인명용 한자의 일부를 확장 한자 처리에 적용하였다. 앞으로의 과제는 인명용 한자뿐만 아니라, 전문 용어 한자 및 지명 한자 등 모든 데이터에 대해서 확장 한자 처리가 구현되어야 한다.

[참고문헌]

[1] 엄채임, 웹 기반 조선시대 인물 정보 서비스 시스템의 설계 및 구현, 동국대학교 대학원 컴퓨터학과 석사논문, 1999

[2] 강태욱, 인물 정보 검색 시스템을 위한 사용자 인터페이스 개선, 동국대학교 공학대학 컴퓨터학과 학사논문, 2002

[3] HiPIS 홈페이지, http://wwwcs.dongguk.ac.kr/~violet00

[4] 도효진, 조선시대 인물 정보 검색 시스템에서의 전체 가계도 검색 시스템의 구현

[5] 디지털 한국학, http://www.koreandb.net/

[6] 사마방목, http://www.koreaa2z.com/sama1/

[7] 문자코드 · 인코딩, http://camars.kaist.ac.kr/dtkim/java/encoding.html

[8] 유니코드 컨소시움, http://www.unicode.org

[9] 한자 한글 변환의 문제점 분석, http://kldp.org/KoreanDoc/html/Hanja2Hangul/Hanja2Hangul.html