국내 도서관 한자검색의 문제점에 관한 연구

[1]최세종, [1]박성혁, [1]서주리, [1]황성진, [1]이창훈*)
[1]서울과학기술대학교 컴퓨터공학과
e-mail: kosejong@naver.com

A study on the problem of library Chinese character searching system in korean libraries

[1]Sejong Choi, [1]Sunghyeok Park, [1]Juri Seo, [1]Seongjin Hwang, [1]Changhoon Lee*
[1]Department of Computer science & Engineering, Seoul National University Of Science & Technology

요 약

본 논문에서는 국내 도서관의 한자검색의 문제점에 대해 집중적으로 분석해 보았으며, 특히 간체자와 이체자의 문제점에 대하여 유니코드와 관련시켜 해결방안을 모색해 보았다. 간체자와 이체자를 정자로 인식하지 못 하는 문제점이 나타나는데 이는 매핑테이블을 이용하여 해결할 수 있을 것이다.

1. 서론

컴퓨터가 문자를 인식하는 과정의 핵심은 문자 인코딩에 있다. 과거 ASCII 코드로부터 시작되어 현재는 유니코드를 전 세계적으로 쓰고 있다. 유니코드에는 영어부터 시작해서 한글과 한자까지도 등록이 되어있다. 하지만 이 유니코드에는 문제점이 존재한다. 이는 사람이 생각하기에는 동일한 문자이지만 컴퓨터의 입장에서는 다른 문자로 인식하는 문제점 이다. 이 문제점은 특히나 한자검색에서 두드러진다.

유니코드에 등록된 한자들은 같은 모양일지라도 서로 다른 유니코드를 가진 한자들이 존재한다. 또한 하나의 대표자에서 떨어져 나온 간체자 그리고 이체자들이 있다. 이러한 문자들은 동일한 뜻을 가지지만 이 또한 서로 다른 유니코드로 등록되어 있다. 이 같은 상황은 같은 뜻으로 검색하여도 다른 검색결과가 나오는 문제점을 초래하게 된다. 이렇게 초래된 문제로 인하여 결과적으로 문서나 자료들을 검색하는 과정에서 많은 어려움을 가지게 만든다. 특히 고서자료가 많은 도서관에서 두드러지게 나타나며 해결이 시급한 상황이다.

본 논문에서는 이 문제점을 해결하고자 유니코드를 이용하면서도 부가적인 기능을 사용하는 방법을 제시하고자 한다. 이 부가적인 기능은 완벽한 해결책이 될 수는 없지만 현재 시급한 문제점을 해결하기에는 적합한 해결책이 될 것이다.

2. 간체자, 이체자, 유니코드

본 절에서는 간체자, 이체자 그리고 유니코드에 대하여 내용을 구성하였다.

2-1. 간체자

간체자는 간화자라고도 불리며 1960년대 중화인민공화국에서 중국공산당의 주도로 만들어진 간략화한 한자이다. 즉, 복잡한 한자를 간단하게 변형시켜 쓰고 있는 문자이다. 여기서 문제되는 점은 간체자로 검색을 했을 경우와 대표자로 검색을 할 경우 결과가 다르게 나온다는 것 이다. 이러한 결과가 나오는 이유는 유니코드에서 간체자와 대표자의 유니코드가 다르게 배정되었기 때문이다.

2-2. 이체자

이체자는 소리와 뜻이 모두 같지만 글자 모양만 다른 한자이다. 일반적으로 우리는 한자가 하나의 모양만 가지고 있다고 생각하지만 실제로는 그렇지 않다. 즉, 하나의 글자가 지역이나 시대에 따라 다른 모양으로 사용된다는 것이다. 이러한 이체자는 글자에 따라 몇 개에서 수십 개가 있는 경우도 있다. 여기서 문제되는 점은 간체자와 마찬가지로 이체자 또한 대표자의 유니코드가 다르기 때문에 한자검색을 하는데 문제가 생기게 된다.

2-3. 유니코드와 관련된 문제점

위에서 언급된 간체자와 이체자의 문제점들을 보면 모두 배정된 유니코드에 의해서 문제가 발생하게 된다. 이 문제가 어떻게 발생되는지 국립중앙도서관과 서울대학교중앙도서관에서 검색해 보았다. 이 때 한자검색에서 문제가 발생하는 간체자와 이체자를 주로 검색하였다. 또한 간체자와 이체자 그리고 대표자에 대한 유니코드도 첨부하였다.

아래의 표는 간체자(医, 与, 爲)와 이 대표자(醫, 與, 爲)의 검색결과를 비교한 결과 그리고 이체자(吉, 僊, 來)와 대표자(世, 仙, 來)의 검색결과를 비교하여 표로 나타내보았다.

*) 교신저자

<표 1> 간체자와 이체자의 검색 결과 비교

도서관 사이트			국립중앙도서관	서울대학교 중앙도서관
간체자	医	U+533B	X	O
	醫	U+91AP		
	与	U+4E0E	X	O
	與	U+8207		
	爲	U+4E3A	X	O
	爲	U+7232		
이체자	帋	U+534B	X	O
	世	U+4E16		
	僊	U+50CA	X	O
	仙	U+4ED9		
	來	U+6765	X	O
	來	U+4F86		

*검색결과가 같다면 O, 같지 않다면 X로 표기

위의 결과를 보면 국립중앙도서관은 간체자와 이체자를 지원하지 않는 반면 서울대학교중앙도서관은 간체자와 이체자를 지원하고 있다. 이러한 결과가 나오는 까닭은 위의 표에서 나오는 유니코드의 값 때문이다. 유니코드 값을 살펴보면 간체자와 대표자 그리고 이체자와 대표자의 유니코드 값이 다르다는 것을 알 수 있다.

아래의 그림들은 차례대로 국립중앙도서관과 서울대학교 중앙도서관에서 가예학, 가례학 그리고 홍무예제, 홍무례제에 대한 검색 결과이다.

그림을 보면 두 사이트 모두 서로 다른 검색결과를 보여준다. 하지만 가예학, 가례학 그리고 홍무예제, 홍무례제에 대해서 한자로 바꾸어 검색한다면 정상적인 결과가 나온다

(그림 1) 국립 중앙 도서관 검색결과(가예학)

(그림 2) 국립 중앙 도서관 검색결과(가례학)

(그림 3) 국립 중앙 도서관 검색결과(홍무예제)

(그림 4) 국립 중앙 도서관 검색결과(홍무례제)

이러한 결과가 나오는 것은 가예학, 가례학 그리고 홍무예제와 홍무례제에서 예와 례가 유니코드가 다르기 때문이다. 하지만 위에서도 말했듯이 한자로 바꾸어 검색하면 같은 검색결과가 나오는데 이는 다중코드자에 대한 매핑테이블이 되어있기 때문이다.

그렇다면 <표 1>에서 국립중앙도서관과 서울대학교 중앙도서관의 결과가 다른 이유는 무엇일까? 이것 또한 매핑테이블로 설명이 가능하다. 이에 대해서는 개선방안에서 자세하게 다룰 것이다.

3. 개선방안

앞서 보았던 도서관 데이터베이스 검색의 가장 큰 문제점은 간체자 그리고 이체자 검색이 아직 원활하지 않다는 것이다. 서로 표기법이 다른 간체자, 이체자와 대표자를 같은 유니코드로 표현할 수는 없으므로 가장 근본적인 해결방안은 신뢰성이 있는 매핑 테이블을 제작하여 표준화될 수 있도록 공개하는 것이다. 여기서 말하는 매핑 테이블은 간체자와 대표자, 이체자와 대표자의 유니코드를 연결시키는 역할을 한다. 매핑테이블로 연결된 유니코드들은 서로 달라도 컴퓨터에서는 같은 글자로 받아들여 같은 검색결과가 나오게 유도한다.

매핑테이블은 다음과 같이 만들 수 있다. 먼저, 간체자와 이체자의 대표자에 해당하는 한자들의 유니코드를 대표값으로 지정한다. 그리고 이 한자와 같은 뜻을 가지고 있거나 파생된 한자들의 유니코드는 참조값으로 지정한 후에. 지정된 한자들을 하나의 테이블을 만들어 연결시킨다. 이때 테이블은 간체자와 이체자를 구분할 수 있어야 한다.

<표 2> 매핑테이블 예시

	No.	유니코드(대표자)	유니코드(참조되는 한자)			
간체자	1	U+91AP	U+533B	U+512A	U+525E	…
	2	U+8207	U+4E0E	U+475E		
	3	U+7232	U+4E3A			
	…	…	…	…	…	…
이체자	No.	유니코드(대표자)	유니코드(참조되는 한자)			
	1	U+4E16	U+534B	U+533A		
	2	U+4ED9	U+50CA	U+51C1	U+5032	…
	3	U+4F86	U+6765			
	…	…	…	…	…	…

*유니코드는 가상의 값임.

<표 2>와 같이 매핑테이블을 만든 후에는 검색엔진에 테이블을 적용하여 같은 번호에 해당하는 행은 같은 검색결과가 나오도록 만든다. 그리고 매핑테이블을 사용 할 경우에 검색결과가 너무 많아질 수도 있기 때문에, 검색 시 특정기호를 입력하여 매핑테이블 기능을 생략할 수도 있어야 한다.

서울대학교중앙도서관의 검색결과를 보면 이미 매핑테이블은 적용이 되어있다. 이처럼 완벽하지는 않지만 기능적인 개선방안은 이미 나와 있다. 하지만 이러한 매핑 테이블을 적극적으로 사용하고 보완하는 데에는 아직 노력이 필요해 보인다. 이러한 노력이 적극적으로 기울여진다면 한자검색이 편리해질 뿐만 아니라 사용자가 원하는 결과를 정확하게 얻을 수 있게 될 것이다.

4. 결론

본 논문에서는 국립중앙도서관, 서울대학교 중앙도서관에서 한자검색 현황을 분석하여 문제점을 파악하고 개선 방안을 도출해 보고자 하였다.

먼저 간체자, 이체자는 매핑정보가 빈약하여 검색에 장애가 되는 경우가 있다. 앞의 자료에서 보이듯이 실제로 각 데이터베이스에서 간체자, 이체자에 대한 검색을 부분적으로 지원하고 있었다.

현재 간체자와 이체자에 대한 검색 기능을 개선하기 위해서는 유니코드 컨소시엄에서 제공하고 있는 호환한자-통합한자의 매핑정보와 같이 신뢰할 수 있는 매핑 테이블을 표준화하는 것이다. 또한 표준화된 매핑 테이블은 국가나 검색엔진에 상관없이 누구나 이용할 수 있어야 할 것이다.

본 논문에서 제안한 해결방안 중 가장 중요한 부분은 바로 매핑 테이블을 상용(常用)화 하는 것이다. 기능을 개선하였음에도 상용화 되지 못한다면 이전과 달라지는 것이기 때문이다. 만약 이러한 매핑 테이블이 상용화 된다면 도서관 데이터베이스 발전에 좋은 결과가 있을 것으로 기대된다.

참고문헌

[1] 이 정 현 “유니코드 한자 검색의 문제점 및 개선방안”(2012)

[2] 국립 중앙 도서관 “활용사이트”
http://www.nl.go.kr/nl/index.jsp (검색일:2013.10.09)

[3] 서울대학교 중앙 도서관 “활용사이트”
http://library.snu.ac.kr/index.ax (검색일:2013.10.09)

[4] 유니코드 문자 변환기 “활용사이트”
http://www.nl.go.kr/kolisnet/convert/convert.php
(검색일:2013.09.12)