XML 기반의 고문헌 검색 시스템 설계

0이근우*, 이근무**
부경대학교 역사학과* 경주대학교 컴퓨터공학과**

The Design for Ancient Literature Retrival Sytem Using XML

Kun-Woo Rhee*, Kun-Moo Rhee **
Department of History , Pukyung University.**
Department of computer Science , Kyungju University.**

요 약

논문에서는 최근 인터넷 상에서 표준 공통 포맷으로 대두되는 XML을 이용하여 웹 기반의 역사자료 의 데이터베이스 검색 시스템을 설계 및 구현하였다. 원격 교육 시스템의 참조문헌 지원 시스템을 설계 하였다. 컴퓨터 상에서 고문헌 자료 주로 한자를 입력하고 검색할 수 있는 환경이 점차 나아지고 있다고는 하지만, 여전히 원전사료의 한자를 일반적인 환경에서 자유롭게 구현하는 일은 결코 용이하지 않다.
종래의 연구자들은 텍스트 별로 수작업으로 작성한 색인류를 이용하여 연구를 해왔다. 그러나 이러한 색인 이용방법에는 문제가 있다. 색인어로 삼고자 하는 단어를 텍스트에서 추출하는 과정에서 누락되는 경우가 있기 때문이다. 전산화된 데이터는 이른바 '발견적 이용'이라는 관점에서 활용될 수 있다

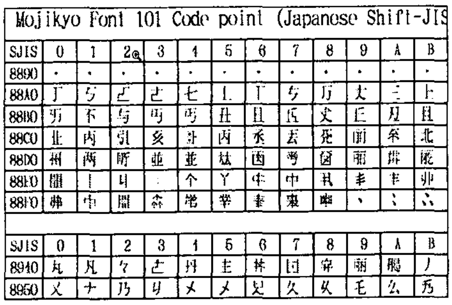
1. 서론

논문에서는 최근 인터넷 상에서 표준 공통 포맷으로 대두되는 XML을 이용하여 웹 기반의 역사자료의 데이터베이스 검색 시스템을 설계 및 구현하였다. 원격 교육 시스템의 참조문헌 지원 시스템을 설계 하였다. 컴퓨터 상에서 고문헌 자료 주로 한자를 입력하고 검색할 수 있는 환경이 점차 나아지고 있다고는 하지만, 여전히 원전사료의 한자를 일반적인 환경에서 자유롭게 구현하는 일은 결코 용이하지 않다.
종래의 연구자들은 텍스트 별로 수작업으로 작성한 색인류를 이용하여 연구를 해왔다. 그러나 이러한 색인 이용방법에는 문제가 있다. 색인어로 삼고자 하는 단어를 텍스트에서 추출하는 과정에서 누락되는 경우가 있기 때문이다. 전산화된 데이터는 이른바 '발견적 이용'이라는 관점에서 활용될 수 있다. 『속일본기』의 경우에, 전체 텍스트를 1일부터 월말까지 날짜별로 집계해 본 결과, 6일마다 기사가 적어진다는 사실을 확인할 수 있었다고 한다. 이는 당시 관인들의 휴일이 6일 간격이었기 때문이라고 한다. 아울러 종래에는 생각할 수 없을 정도로 많은 자료를 대상으로 한 검토와 연구도 가능해진다. 많은 자료들을 이용할 수 있게 되면, 자신의 학문적인 가설이나 논의를 광범위하게 검증할 수 있게 된다. 본 연구는 2 장에서는 관련연구를 3 장에서는 시스템 설계 및 구현상의 문제들을 4 장에서는 실제 설계과정을 5 장에서는 결론을 맺었다.

2. 관련연구

유니코드의 제정으로 컴퓨터에서 나타낼 수 있는 한자의 수가 26,000자 이상으로 확장되었다고 하지만, 이들 한자들을 일반적인 pc환경에서 항상 자유롭게 쓸 수 있는 것은 아니다. 한걸음 더 나아가서 인터넷 상에서 중국 일본에서도 역사 데이터베스를 이용할 수 있는 환경을 만들고자 하면 더욱 문제는 복잡해진다. 일본의 경우 현재 인터넷 환경에서는 JIS에 제정된 한자 6,349자에 대한 읽기 획수 등의 속성정보에 대한 자료를 제공하고 있다. 그러나 일본 내에서도 JIS, S-JIS, EUC 등의 한자코드가 혼재하고 있는 상황이다. 중국과 대만의 경우에도 각각 間體와 繁體로 다른 문자코드를 사용하고 있다. 한자코드의 문제를 좀더 생각해 보도록 하자. 현재

우리나라를 비롯해서 중국 대만 일본에서 일반적으로 사용되는 한자 코드는 서로 다르다. 예를 들어 一이라는 글자에 대해서 우리나라의 KS코드로는 6c69(16진수)인데 대하여, 중국의 GB에서는 523b, 대만의 CNS에서 4421, 일본의 JIS에서는 306c에 해당한다(유니코드에서는 4E00이다). 그러므로 예를 들어 『일본서기』에 대한 한문자료를 입수해서 읽으려고 해도 서로 다른 문자코드 때문에 자료를 직접 해독할 수는 없다. 특히 가능한 한 많은 문자를 수용하기 위하여 독자적으로 개발된 한자코드로 입력된 데이터인 경우에는 더욱 어려움이 따르고 인터넷을 통한 정보교환에는 문제가 많다. 예를 들어 일본에서 역시 독자적인 한자입력 체계로 개발된 Mojikyo(今昔文字鏡)의 경우가 그렇다. 이들 문자는 gif파일로 출력해서 글자를 볼 수는 있지만 문자데이터 형태로 사용하기에는 어려움이 있다. (그림1참조) 그래서 세계적으로 공통되는 코드를 제정할 필요성이 생겼고, 그 결과가 제정된 것이 ISO10646이라는 규격이다. 이 규격은 1문자를 2바이트 혹은 4바이트로



< 그림 1> Mojikyo(今昔文字鏡) 입력 시스템

나타내는 것이며, 현재 구체적으로 코드가 지정된 것은 2바이트 부분이다. 이를 ISO에서는 BMP(Basic Multilingual Plane)이라고 하고, 유니코드 콘소시움(Unicode Consortium)에서는 유니코드(Unicode)라고 한다. 물론 유니코드는 프로그램을 작성하는 데 드는 비용과 노력을 줄이기 위해서 상업적인 목적에 만들어진 것이기 때문에 편의성을 위주로 하여 한글코드 등에도 적지 않은 문제점도 안고 있고, 또 유니코드에 지정된 한자는 국제한자위원회에서 선정한 한자를 모두 충족시키지는 못하지만, 현재로서는 가장 많은 한자를 지원하는 코드인 셈이다. 유니코드는 세계의 대부분의 문자를 2바이트로 처리할 수 있는 공통적인 코드를 목적으로 만들어졌으며, 전체 문자영역에는 65,536문자를 수용할 수 있다. 그 속에 우리나라 중국 일본에서 사용되는 한자 27,136자가 지정되어 있다. 그러나 현실적으로 인터넷 특히 www 상에서 유니코드로 기록된 텍스트를 보거나 조작하거나 인쇄하는 일은 여전히 용이하지 않다. 단순히 유니코드 텍스트를 보는 것은 웹 브라우저 들에서도 다중언어지원프로그램(Multilanguage Support program) 설치하면 가능하다. 그러나 현재는 UTF-7나 UTF-8 [1]체계로 된 유니코드 문서만을 지원하며, 16비트 유니코드 문서는 아직 다른 프로그램의 도움을 받아야 읽을 수

〈 그림 2 UNI CODE 한자 코드체계〉

있다. 또 이러한 브라우저에서는 한 문서 안에 여러 개의 다른 문자 세트에 의거해서 기록된 문서, 예를 들어 한글과 일본어가 함께 들어있는 문서를 볼 수는 없다.

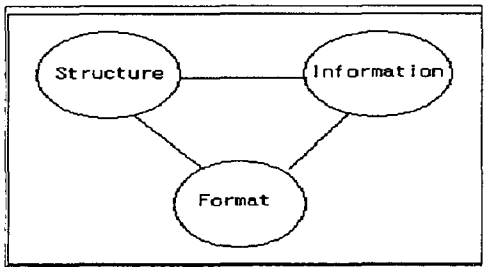
현재까지 정해진 기준 중에서 한자를 가장 많이 표현할 수 있는 문자체계가 유니코드임에도 불구하고 인터넷 환경에서 원활하게 사용하는데는 아직도 문제점이 남아있는 상황이다. 그러나 이미 유니코드에서 정해진 한자를 입력하는 툴도 개발되어 있는 상황이므로[2][3], 역사 학술연구 데이터베이스도 데이터들을 유니코드로 입력하여 정리해야 할 것이다. <그림 2 참조>

3. 시스템구성이 기반 기술과 문제사항

현재 WINDOW NT환경에서의 데이타베이스 서버로서의 SQL서버 그리고 WINDOW에서 흔히 사용되는 데는 데스크 탑용 데이터베이스 도구인 ACCESS에서도 유니코드가 지원되므로, 유니코드를 이용한 데이터베이스 구축환경은 갖추어진 셈이다. 그러나 ACCESS에서도 유니코드를 제대로 다

루기 위해서는 입력및 검색과 관련되어 추가적인 작업이 필수적이다.

XML(eXtensible Markup Language)은 웹 기반 애플리케이션을 통해 데이터를 표현하고 교환하기 위한 표준 공통 포맷인 마크업 언어이다. 전통적인 문서는 정보(Contents 혹은 Information), 구조(Organization 혹은 Structure), 형식(Format 혹은 Display)이 하나의 형태로 서로 묶여져 있어 효과적인 처리가 어려웠으나 XML은 <그림3>와 같이 문서의 구성요소를 분리하여 다룸으로써 인터넷 상에서의 성능 향상을 가져온다.[4][5]

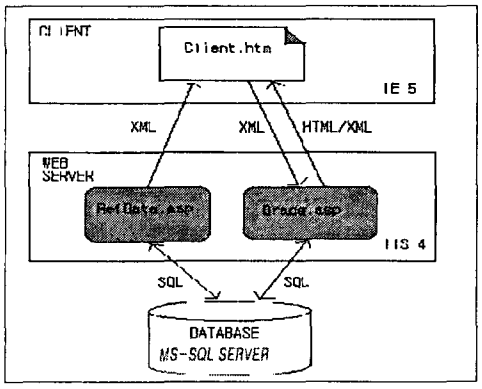


〈그림3〉 XML 문서구조 모델

XML은 자료의 내용을 바탕으로 구조를 생성할 수 있으며, 사용자 인터페이스를 구조화된 데이터로부터 분리하여 다룰 수 있으므로 다양한 데이터 소스로부터 혹은 데이터 소스로의 다방향 정보 교환 및 처리가 가능하며 인텔리전트한 데이터 통합이 쉬운 기술이다. 현재 3-tier 시스템 구조에서 대부분의 정보는 특정 데이터베이스 포맷으로 저장되어 있으며 RDBMS는 데이터 저장과 복구에는 매우 훌륭하지만 정보 공유의 문제를 안고 있다.[6] 웹을 통해 내부 데이터베이스에 저장된 정보를 공유하고자 할 때 필요한 데이터를 쿼리함으로써 결과를 얻으며 또한 그 결과 데이터는 어떠한 의미구조를 가질 수 있는 데이터여야 한다. 이러한 웹 상에서의 의미있는 정보를 공유하고 처리할 수 있는 해결책은 미들티어에서 XML 데이터 포맷을 사용하는 것이다. 즉 클라이언트와 서버는 플랫폼에 관계없이 XML로 정보를 교환하고, 서버와 데이터베이스 엔진 사이에는 기존 데이터베이스 구조 변경해야 하는 부담없이 번역자를 두어 SQL로 쿼리한 결과를 XML 정보로 변환한다. XML 스트림은 전세계적으로 파싱 가능한 유니코드 체계를 사용하므로 브라우저들 사이에서 쉽게 전송 가능하고, 일단 클라이언트에 보내진 데이터는 사용자가 자신의 단일 애플리케이션에서 원하는 대로 데이터를 편집, 조작, 렌더링할 수 있다. 이것은 가용성과 사용자당 성능이 증가할 수 있다.

4. 시스템 구성도

설계한 역사자료 데이터베이스 시스템의 구성은 다음 <그림4>와 같다. MS사의 IE5는 XML과 XSL을 위한 파서를 가지고 있어 XML문서 지원이 가능하지만 웹 기반 교육 시스템에서 모든 학습자들이 IE5를 사용한다고 볼 수 없으므로 미들티어인 웹 서버에서 사용자의 브라우저를 검사하여 IE5이면 XSL로 처리하여 XML 문서를 그 외의 브라우저이면 CSS로 처리하여 HTML문서를 전송하도록 한다. 또한 MSXML 파서는 서버에서 W3C 표준 DOM 메소드나 속성 또는 MS사의 확장된 MSXML메소드와 속성이 자바스크립트와 함께 사용되어 클라이언트에 정보를 표시할 수 도 있다.[7][1]본 논문에서는 DOM을 사용하여 IIS 웹 서버에서 데이터베이스에 저장된 참고 문헌 정보에 접근하여 XML포맷으로 변환해 DOM으로 처리한 후 클라이언트에 표시되게 할 것이다. 빠른 개발 프로토타입을 위해 데이터베이스는 MSSQL-SERVER를 사용했다.

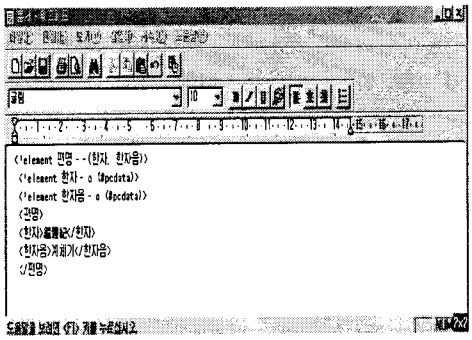


<그림4> 시스템구성도

[<그림5>는 역사자료 DB시스템에서 XML 문

1) W3C DOM WG, http://www.w3.org/DOM/,2000

서를 정의한 DTD를 보인 것이다



<그림5> 데이터베이스를 검색을 위한 DTD

2) 동작 원리 및 구현

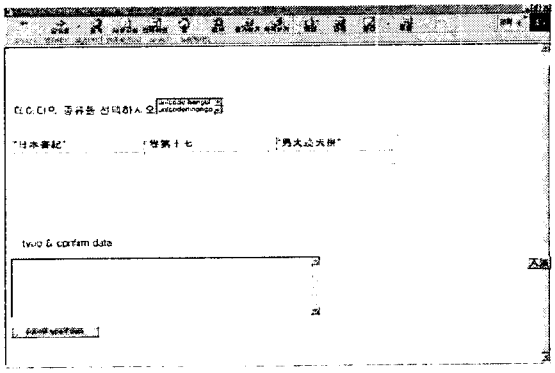
미들티어로 동작하는 웹 서버는 IIS이며 데이터베이스 내의 참조정보를 XML데이터로의 변환 처리는 ASP 스크립트와 DOM을 사용하여 구현하였다.

사용자는 필요한 자료를 검색 및 자료의 형태를 선택할 수 있는 client.htm으로 이동한다. 원하는 검색단어의 타입을 선택하면 SQL 쿼리문으로 선택된 데이터베이스 테이블에 접근할 수 있다. 이 때 DOM API를 이용하여 서버의 데이터베이스에 저장된 참고문헌 정보 MS SQL-SERVER에 접근하여 데이터베이스 테이블의 필드를 DOM의 Element 노드로 생성하여 XML포맷으로 변환시킨다. 각 Element 노드에 실제 저장된 각 필드 값을 Text노드로 생성시켜 Element노드의 자식 노드로 붙인다. 이것은 레코드의 끝에 도달할 때 까지 루프로 돌려 웹 상에 테이블 형식으로 표시되게 구현한다.

<그림6>은 역사자료 DB를 DOM으로 처리하여 검색 결과를 확인하고 변경하는 절차에 대한 예이다. 위와 같은 형태로 논의 한 시스템에 대한 개략적 설계를 보였다.

5. 결론

논문에서는 최근 인터넷 상에서 표준 공통 포맷으로 대두되는 XML을 이용하여 웹 기반의 역사자료



<그림6> 역사자료 DB를 DOM으로 처리한 테이블의 예

의 데이터베이스 검색 시스템을 설계하였다. 연구는 나아가, 첫째 역사자료로서의 도면, 그림, 소리 등 고고학자료등에 관한 멀티미디어 베이스에 대한 개발연구가 진행되어야 할 것이다. 이는 의미있는 정보(특징)을 최대한 포함한 데이터의 축적을 목표로 하여, 해상도 색상정보 박리정보 마연정보 투명부의 화상화, 삼차원정보, 입체시 등을 고려. 이러한 정보를 일종의 개념 네트워크(내용표현 네트워크)을 사용해서 표현, 이를 사용해서 검색을 행하는 다양하게 제안된 기법들에 대한 평가와 이용이 진행되어야 할 것이다. 예를들면 .일반 형태를 인식하는 윤곽선 정보만이 아니라, 문양정보 함몰정보 문양과 함몰의 배치관계의 정보라고 하는 고고학적으로 의미있는 정보를 확인 하고 축적하여야 할 것이다. 나아가. 이 정보를 바탕으로 한 파편의 정보를 표현하는 내용표현 네트워크의 구조화, 파편의 촬영정보에서 내용표현 네트워크를 생성하는 프로그램의 개발및 촬영화상과 내용표현 네트워크의 관련정보에 대한 데이터베이스 저장 및 이용되어야 할 것이다. 또한 고전텍스트는 일반적으로 여러 차례에 걸쳐 필사되어, 잔존하는 諸本에는 異同이 있다. 그래서 연구자들은 어떤 사본(底本이라고 한다)을 기준으로 하여, 다른 사본과의 이동을 조사하여, 연구자가 타당하다고 생각하는 텍스트를 정한다. 이를 校訂이라고 한다. 당연히 완성된 텍스트는 연구자에 따라서 다를 수 있다. 연구의 대상이 되는 異本의 수가 100여 가지에 달하는 경우도 있다. 각 사본의 문자를 충실하게 표현하려고 하면 필요한 문자의 수가 증가하므로, 같은 종류의 문자 즉 이체자들은 하나의 자형으로 통일하여, 문자의 수를 줄이는 방법도 생각할 수 있다. 이에 대해서 가능한 한 텍스트에 충실

해야 한다는 주장도 있다. 이 또한 연구의 목적에 따라서 필요한 경우가 있으므로, 반드시 어느 쪽이 낫다고 단언할 수는 없다. 그러나 의미상 차이가 없이 자형만 조금 다른 경우에는 구별하지 않아도, 통상적인 역사연구에서는 큰 지장이 없다. 개개의 자형이 문제가 되는 경우에는 원래의 글자를 참조할 수 있도록 주를 달거나 화상으로 처리하여 링크하는 방법을 생각할 수 있다. 『속일본기』의 경우는 전체 문자수가 30만자에 달하고 있어 이러한 필요성은 절실하다. 자료의 검색과 관련해서는 다음의 사항들이 향후 시스템개발에 고려되어야 할 것이다. 텍스트 검색에서는 자료가 편년체인 경우는 일별도 나누어 각각을 문헌정보처럼 다룰 수 있다. 검색이 대상이 되는 단어가 예상되는 경우에는 미리 텍스트를 분할하여, 인덱스 파일을 만들어 검색에 이용한다. 텍스트를 순서대로 검색하는 경우에는 pattern matching수법을 이용할 수 있다. 『일본서기』나 『속일본기』 등의 텍스트 파일은 수백 kb정도의 크기이므로 이러한 검색으로도 충분하다.

[참고문헌]

[1]http://www.unicode.org/unicode/uni2errata/UTF-8_Corrigendum.html

[2] 유웅구 외, 「유니코드를 기반으로 한 한자 입력 시스템」, 『한국정보과학회지』 26호

[3] 윤지헌 . 변정용, 유니코드3.0 한자 입력시스템. 한국정보과학회 27회 학술발표논문,2000년 봄.

[4] Alex Coponkus, Faraz Hoodbhoy, The applied XML : A Toolkit for Programmers, JOHN WILEY & SONS, INC., 1999.

[5] 신행자, 박경환, 웹 기반 교육 시스템에서 강의 컨텐츠를 위한 XML 문서 설계 및 구현 동아대학교 부설 정보기술연구소 논문지 제7권 1호, 1999.

[6] Michael Morrison, et al., XML Unleashed, SAMS, 1999.

[7] W3C DOM WG, http://www.w3.org/DOM/,2000

[8]. 김태규 외, "유니코드 한자 지원 문법지시적 SGML 편집기의 설계 및 구현", 한국정보과학회 98년 봄 학술발표논문. 1998.