Unicode 기반 고전문서 편찬 관리시스템

최윤수[0] 진두석 안성수
한국과학기술정보 연구원
{armian[0], dsjin, ssahn}@kisti.re.kr

Unicode based Classics Archive Management System

Yun-Soo Choi[0] Du-Seok Jin, Sung-Soo An
Dept. of giis, Korea Institute of Science and Technology Information

요 약

고전문서는 우리가 상상할 수 없을 만큼의 문화와 지식의 깊이를 지니고 있다. 이러한 문화와 지식을 바탕으로 새로운 지식을 창출해내기 위한 고전문서의 전산화 작업은 필수적인 과제이다. 따라서, 최근 대규모의 고전문서 전산화 작업이 많이 진행 되고 있다. 이러한 수백만 혹은 수천만 페이지에 달하는 대규모 고전문서 전산화 작업에서 가장 어렵고 비용이 많이 소요되는 분야는 고전문서의 의미적 특징을 최대한 손상시키지 않고 데이터베이스를 구축하는 일이다. 그러므로 본 논문에서는 고전문서의 특성을 고려하여 데이터베이스를 구축하고 관리할 수 있는 고전문서 편찬 관리시스템에 대하여 소개한다. 특히 고전문서 전산화에 반드시 필요한 확장한자의 입력 및 검색기능과 문서의 전후관계를 고려한 문서 구조정보의 처리, 그리고 이러한 모든 기능을 효율적으로 수행하기위한 정보검색 시스템에 대하여 소개한다.

1. 서 론

최근 고전문서 전산화 작업에 대한 관심이 증가함에 따라 대규모의 고전문서 전산화 작업이 많이 진행 되어지고 있다. 이러한 분야는 고전문서의 의미적 특징을 최대한 손상시키지 않으면서, 서비스할 수 있어야 하며, 이에 고전문서 전산화에 반드시 필요한 기능들을 효과적으로 처리하기위한 방법을 제시한다. 즉, 현재 표준화 되어있는 코드체계를 확장하여 확장한자의 입력이 가능하도록 설계하였고 문서의 전후관계를 고려한 문서 구조정보를 저장함으로서 고전문서의 의미적 특징을 손상시키지 않고 데이터베이스를 구축하는 방법을 제시한다. 마지막으로 이러한 모든 기능을 효율적으로 수행하기위한 정보검색 시스템에 대하여 소개한다.

2. 관련 연구

2.1 고전문서

고전문서의 특징과 전산화 작업에서의 문제점에 대하여 살펴보면 첫째, 고전문서 전산화를 위한 기존의 표준 코드체계(KSC5601)에서는 한사연코드체계에 따른 새로운 폰트를 제작하여, 대체로 1만 5천자의 한자 표현이 가능하도록 설계되었다. 그러나 실제 고전문서의 전산화 과정에서 이보다 훨씬 더 많은 코드를 요구하므로, 기존의 표준 코드 체계로서는 적절하게 처리할 수 없다. 따라서, 유니코드를 이용한 추가적인 확장한자 폰트가 필요하다. 둘째, 대규모의 전산화작업이 진행되는 고전문서의 경우 데이터의 양이 매우 방대하기 때문에 구축된 이후 계층적인 접근과 검색을 지원하기 위한 적절한 분류 작업이 필요하다. 따라서 적용된 분류법에 따라 자료를 조각냈을 경우 각각의 조각된 문서의 전후관계에 포함된 의미를 상실하는 경우가 발생하기 때문에 이러한 의미적 손상을 최소화 할 수 있는 방법이 필요하다. 셋째, 고전문서 전산화작업의 기간은 수년 또는 수십년이 소요되는 경우도 있다. 따라서 현재 까지 진행된 결과물을 편찬 시점에서 바로 사용가능한 서비스로 적용할 수 있어야 한다.

2.2 유니코드

현재 유니코드는 유니코드3.2.0가 배포되었고, 여러 운영체제에서 유니코드 지원에 대한 노력을 하고 있다. Windows같은 운영체제에서는 현재 유니코드3.0을 지원하고 있으며, 약 2만 7천여자의 한자를 표현할 수 있다

한자와 한글고어 등을 저비용으로 처리하기 위하여는 유니코드를 채택해야 하며, 이러한 유니코드를 자유자재로 입력할 수 있는 입력기에 대한 개발이 필요하다. 물론 유니코드에 존재하지 않는 한자나 고어의 편집을 위해서는 유니코드의 개인사용영역에 폰트 등록 및 관리를 위한 기능이 필요하다.

2.3 구조문서

고전문서의 구조적 특징을 살펴보면 대체로 일정한 패턴을 가진 구조들이 반복적으로 나열된다. 또한 각 문장들 간의 상하 전후관계에 중요한 의미가 포함되어있다. 따라서 이러한 구조를 효과적으로 표현한기 위해서 XML형태의 문서 제작이 필요하다. 예를들면, 승정원일기의 경우 임금, 년, 월, 일 형태의 구조적 정보에 기사가 나열되어 있다.

이러한 구조적인 형식은 DTD의 구성이 수월하고,

XML문서로 쉽게 제작될 수 있으며, 이렇게 제작된 XML문서를 분석 저장할 수 있는 기능이 필요하다.

2.4 정보검색시스템

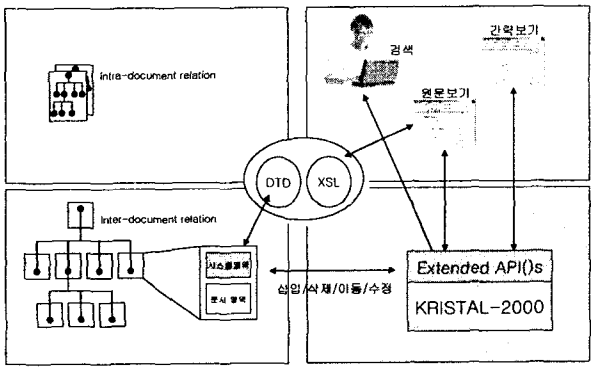
앞 절에서 설명한 기능들과 이를 위해 편찬된 XML문서를 실시간으로 검색 및 관리하기 위해서는 기존의 일반적인 정보검색 시스템을 이용할 경우 많은 제약점을 가지고 있다. 위에서 언급한 기능을 처리하기 위해서는 첫째, 유니코드 기반의 저장 및 검색시스템이 필요하다. 둘째, XML문서의 구조정보를 저장하고 검색할 수 있어야 된다. 셋째, 실시간 문서 편찬 및 관리를 안정적으로 지원해야 한다.

3. 고전문서 편찬 및 관리시스템

3장에서는 본 논문에서 연구개발한 고전문서 관리시스템의 각 모듈에 대하여 자세히 소개한다. 우선 전체적인 시스템의 구조를 설명하고, CJK-IME를 통한 확장한자를 처리하는 편찬시스템과 XML문서를 저장하고 검색 및 관리할 수 있는 저장시스템에 대하여 설명한다. 그리고 본 시스템에서 사용하는 검색시스템에 대하여 소개한다.

3.1 전체적인 시스템 구조

[그림1]은 고전문서 편찬 관리시스템의 전체적인 구조를 나타낸다. 본 시스템은 크게 3가지 모듈로 구성된다. 조선왕조실록이나 승정원일기와 같은 고전문서를 저장시스템에 유효한 포맷 즉, 단순한 XML문서로 변환하거나 이전에 저장된 문서를 검색하여 편집하는 관리자를 위한 편찬관리시스템, XML문서를 저장 및 색인하는 유니코드 기반 저장시스템, 그리고 사용자를 위한 검색질의를 분석하여 다양하고 최적화된 검색을 처리하는 검색시스템으로 구성된다.



[그림1] 전체 시스템 구조

3.2 편찬 관리시스템

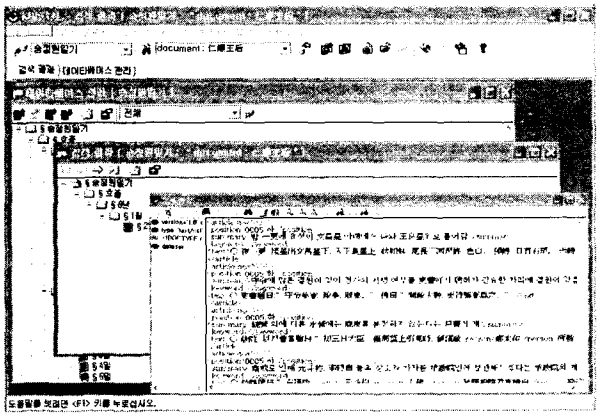
고전문서 편찬 관리시스템에서 사용하는 문서의 내용은 다음과 같은 형식으로 구성된다. 즉, 하나의 XML 문서는 여러 개의 부분 문서로 구분되며, 제안하는 시스템에서는 이 각각의 부분 문서를 하나의 레코드로 저장하며, 삽입, 삭제, 수정, 이동을 레코드 단위로 수행한다. 본 시스템에서 사용되는 XML문서의 한 예를 [그림2]에서 볼 수 있으며, 이 문서는 편찬관리시스템이나 추가 확장한자 입력이 가능한 편집기를 통해 제작될 수 있다.

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/xsl" href="diary8.xsl"?>
<!DOCTYPE dataset SYSTEM "diary8.dtd">
<dataset>
    <record id="">
        <document>
          日有交暈兩珥
        </document>
    </record>
    <record id="">
        <document>
          當該堂上推考
        </document>
    </record>
</dataset>
```

[그림2] XML로 구성된 고전문서

또한 문서 편찬에 사용되는 입력문자는 Unicode CJK IME(Unicode Chinese, Japanese, Korean Input Method Editor)를 사용한다. 이는 한국과학기술정보연구원(Korea Institute of Science & Technology Information, KISTI)과 평양정보센터(Pyongyang Informatics Center, PIC)가 남북 정보 기술 협력 사업의 일환으로 공동 개발한 Windows 용 유니코드 다국어 문자 입력 프로그램이다.

Unicode CJK IME는 Windows 2000 및 XP 체계에서 동작하는 각종 응용 프로그램에서 한국어, 일본어, 중국어, 영어, 러시아어를 다양한 방법으로 입력할 수 있게 하고, 특히 한 중 일 통합 한자(CJK Unified Ideographs)와 확장 한자(CJK Ideographs Extension A)를 비롯한 유니코드 전영역의 문자와 기호들을 쉽게 입력할 수 있는 방법을 제공한다.



[그림3] CJK-IME를 이용한 문서편찬 관리기

편찬관리시스템의 클라이언트 모듈은 문서의 편찬 뿐만 아니라 저장시스템의 데이터베이스 생성 및 관리작업

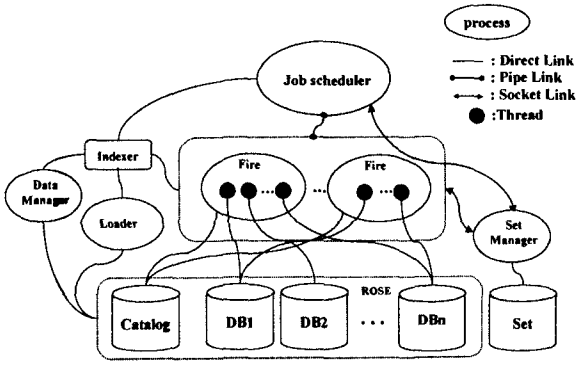
을 수행한다. [그림3]은 문서편찬을 위한 클라이언트 측의 관리자용 도구를 보여준다. 이 관리기는 [그림2]와 같은 형태의 XML문서를 생성,편집하여 저장시스템에 저장하고, 검색결과를 브라우징하는 역할을 수행하며, 아래와 같은 기능을 포함한다.

- 데이터베이스 생성 및 관리 기능
- 데이터베이스 벌크적재 및 백업 기능
- CJK-IME를 이용한 유니코드 확장한자 입력기능
- XML문서 파싱 및 오류검사
- XML문서 편집 및 브라우징 기능
- 구조문서 검색기능
- 고전문서 실시간 삽입, 삭제, 수정, 이동 기능

고전문서에 대한 편집을 담당하는 관리자는 관리도구를 통하여, 새로운 문서를 생성하고 저장된 문서에 대한 실시간 문서교정이 가능하다.

3.3 저장시스템

본 논문에서 사용한 저장시스템은 한국과학기술정보연구원에서 개발한 KRISTAL-2000 정보검색 시스템을 이용한다. 저장시스템의 구조는 [그림4]와 같다. 저장시스템의 특징은 대용량의 문서에 빠른 적재능력이 있으며, 데이터베이스의 압축기능을 사용하여 저장 공간을 줄일 수 있다.



[그림4] KRISTAL-2000 정보검색 시스템 구조

고전문서에 포함되어있는 멀티미디어 데이터의 저장이 가능하고 트랜잭션 처리를 통한 안정적인 문서의 삽입, 삭제, 수정을 보장한다. 또한 편찬 관리시스템과 소켓으로 연결되어 있어 편찬 관리시스템의 요구사항이 실시간으로 처리하여 검색결과에 즉시 반영한다. 검색측면에서는 전문화된 검색문법을 제공하며 다양한 검색모델을 지원하고 검색결과의 랭킹이 가능하다.

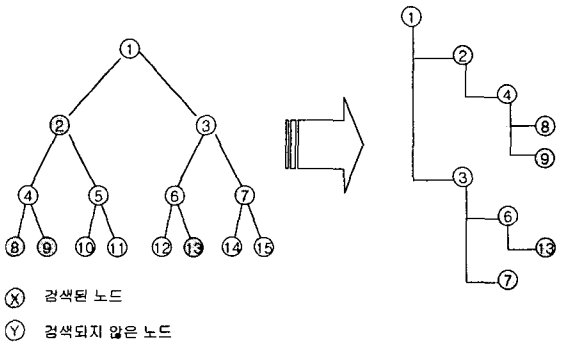
그리고 빠른 검색성능을 위해서 멀티 쓰레드를 이용한 분산검색을 수행하며, 셋 관리기를 사용하여 기존의 검색된 결과에 대한 결과 내 재검색이 가능하다.

KRISTAL-2000 정보검색 시스템의 색인기는 한글, 영문, 숫자, 한자 형태소 분석을 위한 다양한 형태의 색인 타입을 지원하며 어절분석속도 향상을 위한 최적의 알고리즘을 사용하여 구현하였고 보다 정확한 검색을 위한 특정분야 전문사전을 이용한다.

3.4 검색결과 및 브라우징

검색결과의 구조적 표현을 위해 저장되는 데이터는 시스템 영역과 문서영역으로 구분되어 저장된다. 시스템 영역에는 문서의 엘리먼트 또는 문서간의 구조정보가 저장된다. 즉, 자신과 관련된 부모와 형제 노드들 간의 상관관계 등이 저장된다. 문서영역은 KRISTAL-2000의 검색결과의 단위가 저장되는 영역이다. 또한 문서영역에서 특정 엘리먼트를 선택하여 검색을 위한 색인을 생성할 수 있다.

검색은 노드단위로 수행되며, 검색된 결과는 검색된 노드와 노드의 상위노드들을 모두 포함하여 보여준다. 이를 해서는 검색결과에 대한 구조문서 트리의 재구성이 필요하다. 따라서 검색된 노드의 시스템 영역의 정보를 이용하여 [그림5]와 같이 구조 문서 트리를 재구성한다. [그림5]의 오른쪽 그림에서 색깔이 있는 노드가 검색된 노드이고, 그렇지 않은 노드는 검색된 노드의 부모노드이다.



[그림5] 검색결과 브라우징을 위한 구조 트리 재구성

5. 결론

본 논문에서 제시한 고전문서 편찬 관리시스템은 최근 고전문서 전산화 작업에 소요되는 비용을 절감하고, 고전문서의 의미적 특징을 최대한 손상시키지 않고 데이터베이스를 구축 및 관리하는 시스템임을 소개하였다. 특히 CJK-IME를 이용한 확장한자의 입력기능, 문서의 전후관계를 고려한 문서 구조정보처리 그리고 저장 및 검색 기능을 효율적으로 수행하기위한 정보검색 시스템에 대하여 KRISTAL-2000을 이용한 고전문서 편찬 관리시스템을 제시한다.

6. 참고문헌

[1] KRISTAL2000, http://ace.kisti.re.kr/~dairy
[2] XML 문서, http://www.w3.org/XML/
[3] 유니코드, http://www.unicode.org/
[4] 고전문서, http://www.2bytefont.com/