제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

XML 기반 고문서 편찬 관리시스템

진두석, 최윤수, 안성수
한국과학기술정보연구원 정보시스템연구실

e-mail:{dsjin, armian, sshan}@kisti.re.kr

XML based Classics Archive Management System

Du-Seok Jin, Yun-Su Choi, Sung-Soo Ahn
Dept. of giis, Korea Institute of Science and Technology Information

요 약

최근 고문서 전산화 작업에 대한 관심이 증가함에 따라 대규모의 고문서 전산화 작업이 진행 되어지고 있다. 그러나 현재의 표준화 되어있는 코드체계만을 가지고는 고문서를 표현 할 수 없으며, 문서의 구조에 포함된 의미적 특징을 손상시키지 않고 데이터베이스를 구축하기가 매우 어렵다. 또한 이러한 작업은 수개월에서 수년에 걸쳐 여러 차례의 교정 작업이 수행된다. 그러므로 효과적인 고문서 전산화를 위해서는 문서 편찬, 교정, 서비스가 동시에 수행되는 시스템이 필요하다. 따라서 본 논문에서는 기존 코드체계를 확장하여 고문서 전산화에 필요한 확장한자 처리가 가능한 유니코드 기반 입력기를 소개하고, 고문서의 의미적 특징을 손상시키지 않기 위해서 문서 구조정보의 표현이 가능한 XML을 이용한 실시간 문서 편찬 관리시스템을 소개한다.

1. 서론

최근 대규모의 고문서[1] 전산화 작업이 많이 진행 되고 있다. 이러한 수백만 혹은 수천만 페이지에 달하는 대규모 고문서 전산화 작업에서 가장 어렵고 비용이 많이 소요되는 분야는 고문서의 의미적 특징을 최대한 손상시키지 않고 데이터베이스를 구축하는 일이다. 따라서 본 논문에서는 고문서의 특성을 고려하여 데이터베이스를 구축하고 관리할 수 있는 고문서 편찬 관리시스템에 대하여 소개한다. 특히 고문서 전산화에 반드시 필요한 확장한자의 입력 및 검색기능과 문서의 전후관계를 고려한 문서 구조정보의 처리, 그리고 이러한 모든 기능을 효율적으로 수행하기위한 정보검색 시스템을 소개한다.

2. 관련 연구

2.1 고문서 전산화

고문서의 특징과 전산화 작업에서의 문제점에 대하여 살펴보면 첫째, 고문서 전산화를 위한 기존의 표준 코드체계(KSC5601)에서는 한사연 코드 체계에 따른 새로운 폰트를 제작하여, 대체로 1만 5천자의 한자 표현이 가능하도록 설계되었다. 그러나 실제 고문서의 전산화 과정에서 이보다 훨씬 많은 코드를 요구하므로, 기존의 표준 코드 체계로서는 적절하게 처리할 수 없다. 따라서 유니코드[2]를 이용한 추가적인 확장한자 폰트가 필요하다. 둘째, 대규모의 전산화작업이 진행되는 고문서의 경우 데이터의 양이 매우 방대하기 때문에 구축된 이후 계층적인 접근과 검색을 지원하기 위한 적절한 분류작업이 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, 고문서 전산화작업의 기간은 수년 또는 수십년이 소요되는 경우도 있다. 따라서 현재 까지 진행된 결과물을 편찬 시점에서 바로 사용가능한 서비스로 적용할 수 있어야 한다.

2.2 구조문서

고문서의 구조적 특징을 살펴보면 대체로 일정한 패턴을 가진 문장들이 반복적으로 나열된다. 또한 각 문장들 간의 상하 전후관계에 중요한 의미가 포함되어있다. 따라서 이러한 구조를 효과적으로 표현하기 위해서 XML[3]형태의 문서 제작이 필요하다. 예를 들면, 승정원일기의 경우 1월1일 기사는 정치에 관한 것이고, 1월2일 기사는 경제에 관한 것이라고 할때 현재의 기사가 효종 때의 일인지 현종 때의 일인지를 구분하기위해서는 현재 기사내용의 상위 구조를 알 수 있어야 한다. 이러한 문제를 해결하기 위해서 필요한 정보를 정확히 표현할 수 있는 DTD가 정의되어야 하며, 정의된 DTD에 맞는 XML문서의 제작이 필요하다.

2.3 정보검색시스템

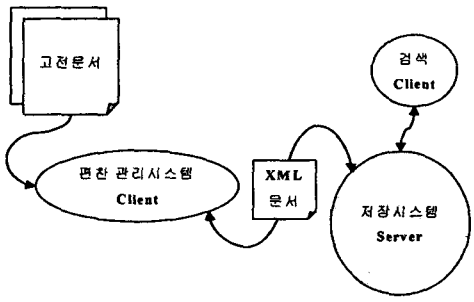
앞 절에서 설명한 기능들과 이를 위해 편찬된 XML문서를 처리하기위해서는 유니코드 기반의 저장 및 검색시스템이 필요하고 XML문서의 구조정보를 저장 및 검색 할 수 있을 뿐만아니라 실시간 문서 편찬 및 관리를 안정적으로 지원하는 정보검색 시스템이 필요하다.

3. 고전문서 편찬 및 관리시스템

3장에서는 본 논문에서 연구개발한 고문서 관리시스템의 각 모듈에 대하여 자세히 소개한다. 우선 전체적인 시스템의 구조를 설명하고, CJK-IME를 통한 확장한자를 처리하는 편찬시스템과 XML문서를 저장하고 검색 및 관리할 수 있는 정보검색 시스템에 대하여 설명한다.

3.1 전체적인 시스템 구조

[그림1]은 본 논문에서 제안하는 고문서 편찬 관리시스템의 전체적인 구조를 나타낸다. 본 시스템은 크게 3가지 모듈로 구성된다. 조선왕조실록이나 승정원일기와 같은 고문서를 저장시스템에 유효한 포맷 즉, 단순화된 XML문서로 변환하거나 이전에 저장된 문서를 검색하여 편집하는 작업을 수행하는 편찬관리시스템과 XML문서를 저장 및 색인하는 유니코드 기반 저장시스템, 그리고 검색질의를 분석하여 다양하고 최적화된 검색을 처리하는 검색시스템으로 구성된다.



[그림1] 전체 시스템 구조

3.2 편찬 관리시스템

고문서 편찬 관리시스템에서 사용하는 문서의 내용은 [그림2]와 같은 형식으로 구성된다. 즉, 하나의 XML 문서는 여러 개의 부분문서로 구분된다. 제안하는 시스템에서는 이 각각의 부분문서를 하나의 레코드로 저장하며, 삽입, 삭제, 수정, 이동을 레코드 단위로 수행한다. 또한 문서편찬에 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. 또한 편찬관리시스템의 클라이언트 모듈은 문서의 편찬뿐만 아니라 저장시스템의 데이터베이스 생성 및 관리 작업을 수행한다.

XML로 구성된 문서를 한국과학기술정보연구원에서 개발한 KRISTAL-2000[4] 정보검색시스템에 저장하기 위해서는 KRISTAL-2000이 지원하는 문서 형태로 변환하여 저장한다. KRISTAL-2000이 지원하는 문서 형태는 비정형 섹션들의 집합으로 구성된다. 각각의 섹션은 독립된 색인형식을 갖으며, 검색

대상이 된다. [표1]은 XML문서를 비정형문서로 변환하였을 때 사용하는 섹션들을 보여준다.

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet type="text/xsl" href="diary8.xsl"?>
<!DOCTYPE dataset SYSTEM "diary8.dtd">
<dataset>
    <record id="">
        <document>
          日有交彙兩珥
        </document>
    </record>
    <record id="">
        <document>
          當該堂上推考
        </document>
    </record>
</dataset>
```

[그림2] XML로 구성된 고전문서

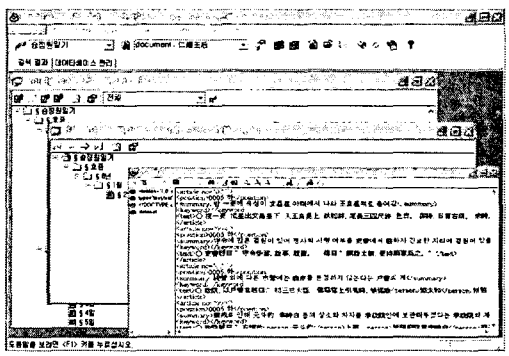
시스템 디비젼은 XML문서 또는 엘리먼트간의 유기적인 연결 관계에 대한 정보를 갖는다. dbase, partition, parent-db는 관리자에 의해 부여되며, 전 문서에 걸쳐 동일한 값을 갖는다. recid, title, type은 XML문서를 분석하여, 비정형 문서로 변환시에 독립적으로 할당된다. recid는 전체 문서내에서 유일한 값으로 부여된다. title은 내부 값에서 추출되어 할당되며, 검색시 간략보기에서 출력되는 섹션이다. level, parent, previous, next, firstchild, order는 자신과 관련된 부모와 형제노드들 간의 상관관계에 의해 할당된다. level은 루트노드로부터의 상대적인 위치 값이 부여된다. previous, next는 형제노드들에 대한 링크값(recid)이 부여된다. firstchild는 자신의 첫 번째 자식노드에 대한 링크 값이 부여된다. order는 형제노드들 사이에서의 자신의 순서 값이며, 간략보기시 출력순서를 결정하는데 사용된다. 문서 디비젼은 XML문서 전체 또는 구조문서 중 특정 엘리먼트의 내용으로 구성된다. 이 섹션은 색인되어 검색대상에 포함될 수도 있으며, 검색 후 뷰어를 통해 보여질 때, 스타일 쉬트와 함께 완전한 XML문서로 출력하기 위해 정의한다. 인덱스 디비젼은 문서 디비젼의 내용 중 일부 엘리먼트들에 대한 색인을 하기위해 정의한 영역으로, XML문서 관리자에 의해 파싱되어 각 엘리먼트별로 색인되며, 검색 시에 사용된다.

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

① 데이터베이스 생성 및 관리 기능
② 데이터베이스 벌크적재 및 백업 기능
③ CJK-IME를 이용한 유니코드 확장한자입력기능
④ XML문서 파싱 및 오류검사
⑤ XML문서 편집 및 브라우징 기능
⑥ 구조문서 검색기능
⑦ 고전문서 실시간 삽입, 삭제, 수정, 이동 기능

[표1] KRISTAL-2000에 저장되는 XML문서 구조

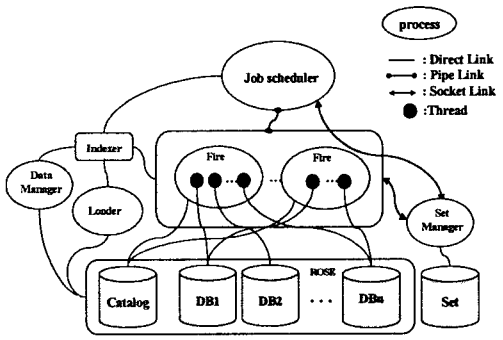
구 분	섹션	비 고
system division	dbase	데이터베이스 이름
	partition	데이터베이스 그룹 영역
	parent_db	부모 데이터베이스
	recid	문서식별자
	title	제목
	type	노드 타입
	level	루트노드로 부터의 위치
	parent	상위노드
	previous	이전노드
	next	다음노드
	firstchild	첫 번째 자식노드
	order	형제들간의 순서
document division	document	문서의 내용 전체
index division	user defined section lists	문서의 내용 중 색인할 엘리먼트



[그림3] CJK-IME를 이용한 문서편찬 관리기

3.3 저장시스템

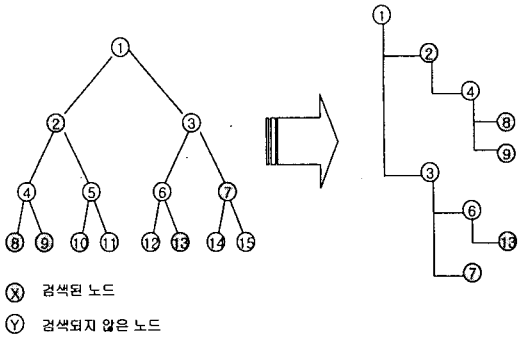
본 논문에서 사용한 KRISTAL-2000 저장시스템의 특징은 대용량의 문서에 빠른 적재능력이 있으며, 데이터베이스의 압축기능을 사용하여 저장 공간을 줄일 수 있다. 또한 고문서에 포함되어있는 멀티미디어 데이터의 저장이 가능하고 트랜젝션 처리를 통한 안정적인 문서의 삽입, 삭제, 수정을 보장한다. 그리고 빠른 검색성능을 위해서 멀티쓰래드를 이용한 분산검색을 수행하며, 셋 관리기를 통한 기존의 검색된 결과의 결과내 검색이 가능하다. 그리고 KRISTAL-2000의 색인기는 한글, 영문, 숫자, 한자 형태소 분석을 위한 다양한 형태의 색인타입을 지원하며, 보다 정확한 검색을 위해서 특정분야 전문사전을 이용한다. [그림4]는 KRISTAL-2000 정보검색 시스템의 구조를 보여준다.



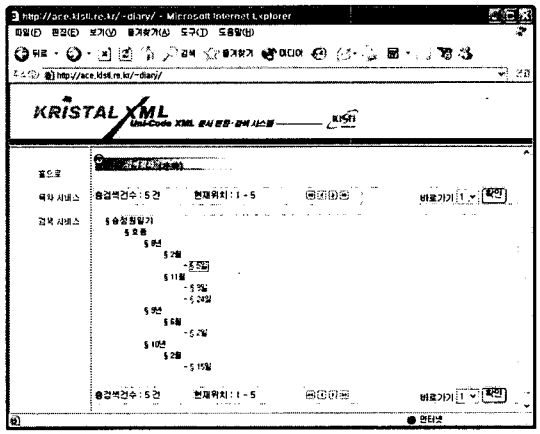
[그림4] KRISTAL-2000 정보검색시스템 구조

4 검색결과 및 브라우징

검색결과의 구조적 표현을 위해 저장되는 데이터는 시스템영역과 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. 검색된 결과의 브라우징을 위해서는 검색결과에 대한 문서구조 트리의 재구성이 필요하다. 따라서 검색된 노드의 시스템 영역의 정보를 이용하여 [그림5]와 같이 문서구조 트리를 재구성한다. 재구성된 트리를 이용하여 [그림6]과 같은 계층적 검색결과를 보여주고 검색된 결과 내에서 부모, 형제, 자식으로의 탐색 기능을 제공한다.



[그림5] 검색결과 브라우징을 위한 트리 재구성



[그림6] 검색결과 브라우징

5. 결론

본 논문에서 고문서 전산화 작업에 소요되는 비용을 절감하고, 고문서의 의미적 특징을 최대한 손상시키지 않고 데이터베이스를 구축 및 관리할 수 있는 고문서 편찬 관리시스템을 소개하였다. 특히, 고문서 처리에 꼭 필요한 CJK-IME를 이용한 확장한자의 입력기능, 문서의 전후관계를 고려한 문서의 구조정보 처리 그리고 실시간 문서 관리기능에 대하여 소개하였다. 향후에는 이러한 기능들을 보다 편리하게 사용할 수 있도록 본 논문에서 제시한 XML 기반 고문서 편찬시스템과 KRISTAL-2000 정보검색 시스템 계속 발전시켜 나갈 계획이다.

참고문헌

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