일반 텍스트 문서의 XML 문서 변환시스템

A Translation System from General Text to XML Document

이현실, 원광대학교 중앙도서관
최유순, 원광대학교 컴퓨터공학과
한성국, 원광대학교 컴퓨터공학과

Hyun-Sil Lee, Yue-Soon Choi and Sung-Kook Han
Central Library of Wonkwang Univ., Dept. of Computer Engineering, Wonkwang Univ.

21세기 지식기반사회를 맞이하여 도서관은 정보를 지식화하고, 지식화된 정보를 자동으로 추출하여 제공할 수 있는 사용자 편의를 지향한 정보서비스를 필요로 하고 있다. 정보의 지식 처리를 위해서는 문서가 다양한 의미를 표현할 할 수 있는 XML 문서의 형태로 되어야 한다. 본 연구는 문서의 효율적인 교환과 제공을 위하여 XML 문서의 데이터 모델링 개념을 활용하여, 일반 텍스트 문서를 XML 문서로 변환하는 시스템을 구현하였다.

1. 서론

인터넷 기술의 발전과 더불어 웹 기술에 의한 정보의 혁명 시대를 맞이하였다. 이로 인해 많은 양의 정보에 쉽게 접근할 수 있게 되었고 인터넷은 모든 정보처리의 기반이 되었다. 또한 웹의 우수한 정보가공과 표현의 능력은 다양한 정보와 멀티미디어정보 처리의 표시가 되었다. 그러나 지금까지의 웹은 정보표현 능력이 뛰어나지만, 지식표현능력은 부족하다. 이에 따라, 정보의 구조화와 의미 표현을 중시하는 의미웹(Semantic web)으로의 이전이 필요하게 되었고 의미웹은 XML을 기반으로 한 RDF, DAML 등으로 실현되고 있다. 이와 같이 정보지식을 정제하고 조직화하는 의미기반의 웹으로 지식처리의 새로운 시대가 열리게 되었다.

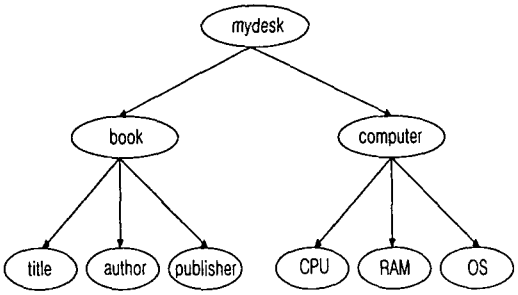
웹이 지식 처리 능력을 갖기 위해서는 문서가 XML 문서의 형태로 되어야 한다. 그런데 HTML로 작성된 웹 문서나 일반 텍스트 문서를 XML로 자동 변환하는 것은 어려움이 있다. 왜냐하면, XML 문서는 의미 기반이므로, HTML이나 텍스트 문서에 내포된 지식을 자동으로 추출하는 것은 불가능하기 때문이다. 따라서, HTML이나 텍스트 문서를 XML문서로 효과적인 변환의 방법이 필요하다. 이와 같은 XML 문서변환은 지식기반 사회를 맞이하여 도서관에서 문서처리를 위해 반드시 필요하다. 따라서 본 연구에서는 XML 문서의 데이터 모델링 개념을 활용하여, 일반 문서를 XML 문서로 변환하는 시스템을 구현하고자 한다.

2. XML 데이터의 모델링

XML 문서를 만들 때 사용자의 필요에 따라 임의의 확장과 구조적인 정의가 가능하다. 이러한 융통성으로 누구나 훌륭한 XML 문서를 만들 수도 있지만, 이해하기 어려운 문서도 만들 것이다. 여기에 모델링이라는 부분이 문제의 핵심이다. XML은 모델링의 효율성에 따라 새롭게 만드는 것에서부터 저장과 변환 및 교환되는 지식의 처리가 원활하기 때문이다.

2.1 XML 모델링의 개념

XML 데이터 모델링은 XML 문서가 데이터를 취급하는 기본 방법이다. XML에서 데이터는 트리(tree) 형태의 계층적 구조로 다루어진다. 즉 XML에서 문서나 정보의 기본 모델링은 트리를 기반으로 하고 있다는 뜻이다. XML은 문서나 정보를 구조화하는 기본적인 방법을 제공하는데, XML에서는 문서나 정보를 계층적인 관점에서 조명한다. 문서의 각 인스턴스(instance)는 맨위에 하나의 루트 요소가 있고 다른 요소들은 거기에 계층적인 구조로 종속이 되는데 아래에 오는 요소를 자식 요소라 하고 위쪽에 오는 요소를 부모 요소라고 한다. 이것은 마치 하나의 뿌리가 많은 가지와 나뭇잎들을 가지고 있는 트리의 구조와 같은데 나뭇잎은 실제 데이터의 값들을 표현한다고 할 수 있다.



[그림 1] XML문서의 트리구조

요소의 특성에 관한 정보를 제공하기 위해, 시작태그 내에 속성(attribute)을 가질 수 있다. 속성은 이름과 값의 쌍으로 구성되며 필요에 따라 여러 개의 속성을 가질 수 있다. 속성의 목적은 요소에 추가적인 정보, 즉 정보에 대한 정보를 정의하고자 할 때 사용하는데 이것을 메타데이터라고 한다. [그림 2]는 book이라는 루트요소에 title, author, publisher라는 하위요소를 두어 구조화하였다.

```
<book size="25cm" price="8000won">
  <title subtitle=" 과학적 경영론">도서관 경영론</title>
  <author email=" yhpark@moak.ac.kr">박영희</author>
  <publisher phone= "02-850-1111"> 한솔</publisher>
</book>
```

[그림 2] XML 문서의 예

요소 "book"은 두 개의 속성(size, price)과 각기 속성을 가지고 있는 세 개의 하위요소 "title" "author" "publisher"를 가지고 있다. 그리고 각 하위요소는 속성을 가지고 있다. XML에서 요소와 속성의 이름은 임의로 만들어질 수 있지만 반드시 시작과 끝 태그의 한 쌍으로 사용되어야 한다.

2.2 모델링과 DTD

XML은 데이터 구조에 대한 정보를 DTD (Document Type Definition)로 기술하고 있다. DTD는 XML 문서 모델링을 위한 가장 기본적인 지식 중의 하나이다. DTD는 문서를 구성하는 요소, 요소의 속성값, 요소간의 관계 등 문서의 형태를 지정하는데 필요한 규칙을 지정한다. DTD에서 정의한 문서 조직 규칙을 만족하는 XML 문서를 유효한 XML문서라 한다.

XML을 기반으로 원문정보를 제공하고자 하는 도서관에서는 원문을 제공하려는 모든 유형에 대한 DTD를 개발하여야 한다. 문서의 구조 규칙인 DTD는 문서의 종류에 따라, 적합한 문서 구조 규칙을 작성할 수 있다. DTD는 기본

적으로 관련된 XML문서의 논리적 구조를 위한 템플릿을 형성하는데 데이터의 계층구조를 표현한다. [그림 3]은 앞에서 모델링한 구조에 대한 DTD표현이다.

```
<!ELEMENT book (title, author, publisher)>
<!ATTLIST book size CDATA REQUIRED>
<!ATTLIST book price CDATA REQUIRED>
<!ELEMENT title (#PCDATA)>
<!ATTLIST title subtitle CDATA REQUIRED>
<!ELEMENT author (#PCDATA)>
<!ATTLIST author email CDATA IMPLIED>
<!ELEMENT publisher (#PCDATA)>
<!ATTLIST publisher phone CDATA REQUIRED>
```

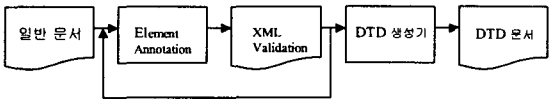
[그림 3] DTD 문서의 예

DTD는 크게 요소 선언, 속성 선언, 엔티티 선언 노테이션 선언의 네가지 형식으로 나눌 수 있으나, 여기에서는 간단한 개념설명을 위하여 요소 선언과 속성 선언만을 사용하여 "book"과 그의 하위 요소인"title", "author", "publisher"에 대한 정의를 하였다.

이기종 시스템들간의 구조적인 문서 교환을 위한 XML 문서의 사용이 증가함에 따라 이에 대한 모델링 방법들이 연구되고 있다. 그러나 기존의 모델링 방법들은 인위적으로 DTD를 생성하기 위하여 일일이 요소들을 설정해주어야만 했다. 따라서 일반 문서를 모델링하기 위한 자동화된 방법이 필요하다.

3. XML 문서 구조로의 변환

사용자에게 정보 검색의 용이함이나 지식표현의 능력을 볼 때, 기존의 일반 문서를 의미기반의 XML 문서로 변환할 필요가 있다. 이때 XML로 표현하기 위해 문서를 다시 작성한다면 시간과 노력, 비용이 많이 소요될 것이다. 본 연구에서는 기존의 일반 문서를 XML 문서로 변환할 수 있는 방안을 제시하고, T2XG (Text to XML Generator)시스템으로 구현하였다. [그림 4]에 시스템의 흐름도를 나타내었다.



[그림 4] T2XG 시스템 흐름도

3.1 XML 문서표현

본 연구에서는 한국정보관리학회에서 회원들에게 발송한 학술대회 개최공문을 일반텍스트 문서 샘플로 XML 문서구조에 적용하여 변환하였다.

한국정보관리학회

본 학회는 제 8 회 한국정보관리학회 학술대회, 2001을 아래와 같이 개최하고자 하오니 모두 참석하시어 발표의 장고 유익한 토론의 시간을 가지시길 바랍니다. 상세한 세부 일정은 추후 알려드리겠습니다.

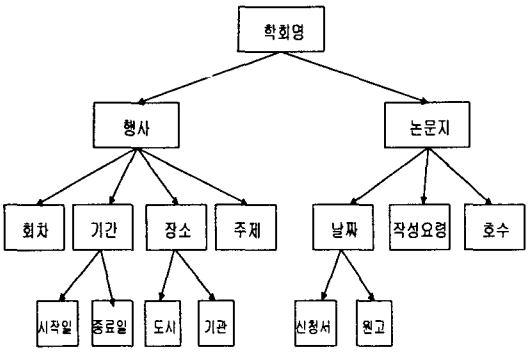
- 아 래 -

일시 : 2001년 8월 23일(목) - 24일(금) 9:30 - 17:30
장소 : 한국과학기술정보연구원(KISTI) 대전본원
주제 : 문헌정보학, 정보관리 전분야

학술대회에 발표를 원하시는 회원 여러분들은 발표할 논문에 관한 내용을 기일 내에 신청하여 주시고 첨부된 원고 작성요령에 맞추어 준비하여 주시기 바랍니다. 본 대회에서 발표된 논문은 제8회 한국정보관리학회 학술대회 논문집(Proceedings of the 8th Conference of Korean Society for Information Management, 2001"에 게재될 것입니다.
신청서 제출 기한 : 2001년 7월 13일(금)
원 고 제출 기한 : 2001년 8월 3일(금)

[그림 5] 일반 문서

일반 문서를 XML 문서로 변환하기 위해서는 먼저 문서에 대한 정보의미 모델을 설정할 필요가 있다. 앞서 서술한 바와 같이 XML은 트리 형식으로 모델화하고 있으며, [그림 5]에 대해서는 [그림 6]과 같이 모델화 할 수 있다. <학회명>태그를 루트 요소로 하고, 학회명 아래 <행사>와 <논문지>를 하위 요소를 첨가하였고, <행사>와 <논문지>는 속성을 부여하여 XML 문서를 만들었다.



[그림 6] XML로 표현하기 위한트리 구조

[그림 6]과 같이 문서 모델화 하였을 때, [그림 5]의 문서는 [그림 7]과 같이 XML 문서로 표현된다. 여기서는 편의상 DTD를 생략하고 문서 본체의 구조화 예만을 보였으며, 본 연구에서는 문서 구조화에 수반되는 DTD의 자동 생성 방법을 제시한다 .

```
<학회명> 한국정보관리학회
  <행사> 학술대회
   <회차> 8 </회차>
    <기간>
     <시작일> 2001년 8월 23일</시작일>
     <종료일> 2001년 8월 24일</종료일>
    </기간>
    <장소>
     <도시> 대전직할시 </도시>
      <기관> 한국과학기술정보연구원 (KISTI) </기관>
    </장소>
   <주제> 문헌정보학 </주제>
   <주제> 정보관리 </주제>
  </행사>
  <논문지> 학술발표논집
    <날짜>
     <신청서> 2001년 7월 13일 </신청서>
     <원고> 2001년 8월 3일 </원고>
   </날짜>
   <작성요령> 제목, 저자, 소속, 초록, 본문 </작성요령>
   <호수> 9집 </호수>
  </논문지>
</학회명>
```

[그림 7] 일반 문서의 XML 변환 문서

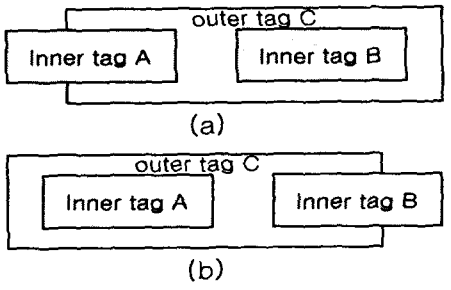
3.2 변환 알고리즘

일반텍스트 문서를 XML 문서로 변환하기 위해서는 사용자가 문서의 의미와 구조에 해당하는 요소 태그를 지정해 주어야 한다. 사용자가 문서구조의 모델에 관계없이 임의로 요소 태그를 부여할 수 있도록 본 연구에서는 다음과 같은 알고리즘을 구상하여 사용하였다.

Algorithm:Document Model Tree Generation

Input : marked text region, element and attributes
Output : DTD tree Structure
Variables : *root, mtr, tag,* inner tag list *TagList*

① Create document node called *root.*
② Set the marked text region into *mtr.*
③ If *mtr* contains any tags of elements
then make an inner text region list called *TagList*
else return *mtr* attached to parent node
④ Create a new element node *mtr* with attributes.
⑤ for *tag* = *First(TagList)* to *End(TagList)*
if *mtr* and *tag* are crossing in their regions
then return Error("Crossing Regions")
else modify the attachment of *tag* to *mtr*
⑥ Attrach *mtr* to parent node.
⑦ Repeat step2 until no more marked text region.



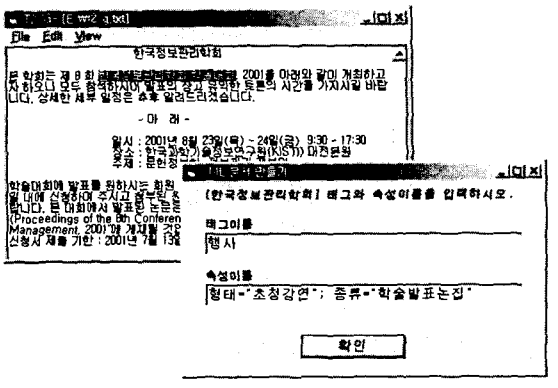
[그림 8] 중첩된 태그에서의 충돌

XML 태그를 구성할 때 주의할 점은 중첩된 태그의 경우 서로 충돌이 일어나지 않아야 한다. 내부태그는 온전히 외부태그의 일부분이어야 하고, 외부태그는 내부태그를 완전히 포함하여야 한다. [그림 8]은 중첩된 태그를 구성할 때 서로 충돌이 일어나는 경우이다. (a)는 내부태그 A가 외부태그 C에 일부만 포함되어 정상적으로 처리되지 못하는 예이고, (b)는 내부태그 B가 외부태그 C에 일부만 포함되어 있는 예를 보였다. 이런 경우, 외부태그를 정의할 때 내부태그가 존재하는지를 확인하고, 존재한다면 내부태그를 모아 TagList를 만든다. TagList는 새로 정의할 외부태그와 교차하지 않는지를 확인할 수 있도록 해준다.

4. XML 문서 변환 시스템

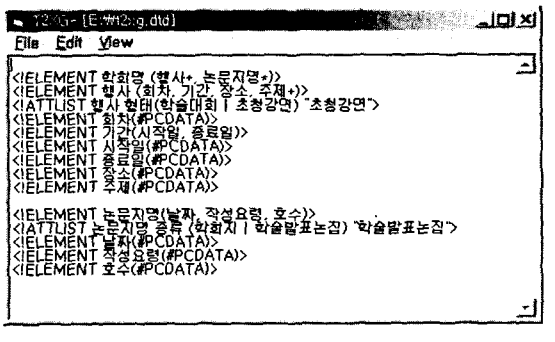
T2XG는 [그림 9]에서 보는 바와 같이 텍스트 형태의 문서를 입력하거나 읽어올 수 있고, 그러한 문서는 단어를 블록으로 설정하므로서 XML 태그 이름을 입력할 수 있다. XML로 표현된 태그는 열린 태그와 닫힌 태그가 자동으로 삽입되어 XML 폼에 맞는 문서를 구성하게 된다.

[그림 9]에서는 t2xg.txt 문서를 읽어서 XML 문서로 만드는 과정 및 결과를 보이고 있다. 그 과정을 보면 처음에, 학회 이름인 '한국정보관리학회 학술대회'에 요소 이름을 부여하기 위하여 블록을 설정하였다. 다음에, 설정된 블록에 대하여 태그이름과 속성 이름을 삽입할 수 있는 메뉴가 보인다. 이 때, 태그이름 입력란에 요소 이름을 입력하도록 한다. 속성이 필요한 경우, 속성이름 입력란에 속성도 같이 입력한다.



[그림 9] T2XG를 이용한 일반 문서변환

요소 이름을 입력하여 View 메뉴에서 XML을 선택하면 [그림 10]과 같은 XML 문서를 볼 수 있다.



[그림 10] 일반문서의 DTD 표현

이러한 과정으로 일반 텍스트 문서를 XML 문서로 쉽게 변환할 수 있다. XML로 변환된 문서는 태그 이름을 요소로 하고, 속성 이름을 지니는 DTD 문서를 자동으로 생성하게 된다. DTD로 형성된 결과가 [그림 10]에 나타나 있

다. T2XG는 일반 문서를 XML로 변환하면서 '태그보이기' 기능을 이용하여 편집 상태에서 조판부호로 삽입된 태그를 볼 수 있도록 구현하였다.

5. 결론

본 연구에서는 정보의 지식처리를 위해서 일반 텍스트 문서를 의미처리가 가능한 XML문서로 변환하는 시스템을 구현하였다. 방법은 XML 문서의 모델링 개념을 활용하였고 변환 알고리즘을 개발하여 XML 문서가 추출되도록 설계하였다.

구현된 시스템은 일반문서에 활용은 물론 도서관에서 학위논문이나 연구보고서의 원문구축 또는 가상문서 브라우징 등 다양하게 응용할 수 있는 기초적인 모델을 제시하였다.

이 시스템에 효율을 높이기 위하여 향후에 추가 설계하여야 할 사항은 다음과 같다.

첫째, 사용자 인터페이스를 강화하여 개발한다.

둘째, 자동 요소영역을 추출하는 자동생성 능력을 향상시킨다.

셋째 HWP, PDF, DOC 등의 다양한 파일 형식을 지원할 수 있는 호환성을 높인다.

<참 고 문 헌>

1] 김채미, 김심석, 최학열. 2000.『XML DTD와 스키마로 객체지향 모델링하기』 Microsoftware 207. pp. 344-353.

2] 김현희. (1999). 『XML을 이용한 가상대학 교육 문헌 구조 설계 및 데이터베이스 구축에 관한 연구』. 명지대학교 인문과학연구논총 19. pp. 307-336.

3] 양중식. (2000). 『디지털 도서관 시스템에서 가상문서 브라우징을 위한 XML 기반 문서처리』. 석사학위논문. 충남대학교 대학원.

4] 우항준. (1998). 『전자도서관을 위한 연구보서용 XML DTD개발』. 석사학위논문, 동국대학교 산업기술환경대학원.

5] 이준섭, 유정연, 권석훈, 나재열, 이규철, 구경철, 박기식, 박치항. (2001). 『XML을 적용한 표준 문서 관리시스템의 설계 및 구현』. 한국문헌정보학회지 35(1). pp. 77-99.

6] Devedzic, V. (1999). 『A survey of modern knowledge modeling techniques』. Expert Systems with Applications 17. pp. 275-294

7] Miller, Dick R. (Summer 2000). 『XML: libraries' strategic opportunity』. Library Journal Net Connection. <http://www.librar yjournal.com/xml.asp>

8] Mylopoulos, John. (1998). 『Information Modeling in the time of the revolution』.Information Systems 23(3/4). pp.127-155.