통합 DTD를 이용한 폼(Form) 기반 XML 편집 및 변환 시스템 개발*

손원성, 이현찬°, 김재경, 유범종†, 최윤철
연세대학교 컴퓨터과학과,
†연구개발정보센터

Development of a Form-based XML Editing and Converting System Using Integrated DTD

Won-Sung Sohn, Hyun-Chan Lee, Jae-Kyung Kim, Beom- Jong You†, Yoon-Chul Choy
Dept. of Computer Science, Yonsei University
†Korea Research & Development Information Center

요 약

전자문서를 정의하기 위해서는 구조문서 정의가 가능한 SGML(Standard Generalized Markup Language) 및 XML(eXtensible Markup Language), 그리고 웹문서를 위한 HTML을 이용할 수 있다. 그러나 SGML은 매우 복잡한 구문을 포함하기 때문에 일반 사용자들이 쉽게 사용하기가 어렵고 HTML은 비 구조적인 문서정보 및 스타일 정보를 포함하기 때문에, 근래에 개발된 대부분의 전자문서 시스템에서는 XML을 응용하고 있다. 전자문서 시스템은 복잡한 구조의 다양한 문서들을 포함하고 있으며 이러한 문서들은 DTD(Document Type Definition)로 정의된다. 그러나 기존의 전자문서 시스템에서는 각 문서의 구조마다 별도의 DTD를 정의하고 있기 때문에 DTD의 구조 정의 및 관리가 매우 비 효율적이다. 따라서 본 연구에서는 다양한 구조를 모두 표현할 수 있는 통합 DTD를 정의하고 이를 효율적으로 관리 및 처리하기 위한 폼 기반 구조문서 처리 방법을 제안한다. 통합 DTD와 폼을 통하여 사용자는 폼 단위의 문서 인스턴스만을 입력 및 편집함으로써 복잡한 DTD 구조에 독립적으로 문서를 처리할 수 있으며, 처리된 문서는 DTD에 위배되지 않는 정확한 XML 인스턴스가 된다. 또한 기존의 HTML문서를 XML로 변환하기 위하여 본 논문에서는 누구나 손쉽게 사용할 수 있는 중간단계(semi-auto)의 XML 변환시스템을 제공한다. 그 결과 본 연구에서 개발한 시스템에서는 다양하고 복잡한 문서에 대하여 효율적인 문서구조가 가능하고, XML문서를 폼을 이용하여 누구나 쉽고, 정확하게 작성할 수 있다. 그리고 웹에서 사용된 HTML 문서들, 본 연구에서 정의한 통합 DTD 구조에 인지하는 XML 문서로 간단히 변환할 수 있다.

1. 서론

다양한 전자문서를 처리하기 위해서는 구조문서 정의가 가능한 SGML[1] 및 XML[2], 그리고 웹문서를 위한 HTML을 이용할 수 있다. SGML은 매우 복잡한 구문을 포함하기 때문에 일반인들이 쉽게 사용할 수 없으며, SGML을 완벽하게 처리하는 시스템을 개발하기 어렵다. 또한 HTML은 비 구조적인 문서정보 및 스타일 정보를 동시에 포함하기 때문에 SGML과 같은 체계적인 문서 구조 정의가 불가능하다. 이러한 문제점들을 절충한 것이 XML이며, 현재 매우 다양한 연구 및 적용 사례를 살펴볼 수 있다[3,4,5].

특히 전자정부 및 사내 전자문서처리 시스템등에서는 XML을 표준으로 지정한 경우가 많으며, 이러한 시스템에서는 방대한 규모의 정보를 처리할 수 있어야 한다. 이러한 방대한 정보를 XML형태로 처리하기 위해서는 효율적인 구조 정의가 요구된다. 예를 들어 전자정부 및 사내 전자문서처리 시스템을 위한 DTD는 각 부처별로 정의하거나 통합한 형태로 정의할 수 있다. 각 부처별로 구조를 정의한 경우에는 구조 정보를 표현하기가 용

* 본 연구는 과학기술부 정책연구사업의 연구비 지원에 의한 것임

이하나, 복잡하고 다양한 정보들을 각각의 DTD에 따라 처리해야 하기 때문에 복수의 처리 루틴 및 별도의 관리 방법이 요구되는 비효율적 방법이 된다.

따라서 본 논문에서는 다양한 구조를 모두 표현할 수 있는 통합 DTD를 정의하고 이를 효율적으로 관리 및 처리하기 위한 폼 기반 구조문서 처리 방법을 제안한다. 폼 기반 처리 방법이란, 방대한 정보를 통합적으로 정의한 통합 DTD를 기반으로, 부처 또는 세부 단위의 정보를 XML 형태의 중간 파일을 이용하여 처리하는 기법이다. XML 형태의 폼 입력 양식은 각 부처에 대해 고정된 입력 양식을 제공하기 때문에 사용자들은 XML 문서를 간단하고도 정확히 작성할 수 있다.

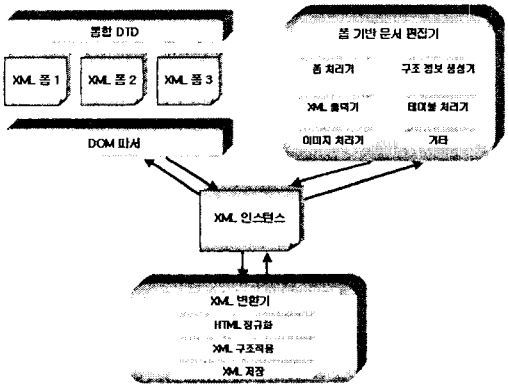
또한 기존의 HTML문서를 XML로 변환하기 위하여 본 연구에서는 누구나 손쉽게 사용할 수 있는 중간단계(semi-auto)의 XML 변환시스템을 제공한다. 그 결과 본 연구에서 개발한 XML 시스템에서는 다양하고 복잡한 문서에 대하여 효율적인 문서구조가 가능하고, XML문서를 폼을 이용하여 누구나 쉽고, 정확하게 작성할 수 있다. 그리고 웹에서 사용된 HTML 문서를, 본 연구에서 정의한 통합 DTD 구조에 일치하는 XML 문서로 쉽게 변환할 수 있다.

2. 통합 DTD 기반 XML 편집 및 변환 시스템 설계

본 연구에서 제안한 통합 DTD 및 폼 기반 XML 입력 시스템의 설계 내용은 다음과 같다.

2.1 시스템구조

XML 문서를 위한 편집 시스템의 구조는 [그림 1]과 같으며 다음과 같은 기능들로 구성된다.



[그림 1] 폼 기반 XML 문서 편집 시스템의 구조

- 통합 DTD와 Form 생성
- 문서 구조 정보를 위한 구조 생성기
- XML 문서 출력 기능
- 테이블, 이미지, 특수문자 입력 기능
- HTML 문서의 XML 변환 기능

2.2 통합 DTD 정의와 폼 기반 XML 처리 기법

통합 DTD 처리기법은 앞서 살펴본 바와 같이 대용량의 복잡한 문서 구조를 효율적으로 저장, 관리, 처리하기 위하여 제시된 방법이다. 특히 본 연구에서는 과학기술부, 정보통신부, 산업자원부의 특정연구개발과제, 선도기술개발사업, 산업기반기술개발사업 보고서의 구조를 모두 포함하는 통합 DTD를 개발하였고, 이러한 통합 DTD의 설계 기준은 다음과 같으며 상세 구조는 [표1]과 같다.

- 구조 파악의 용이성
- 문서 구조의 확장성
- 사용의 용의성
- 폼 정의 기능

<표 3.1> 3개 부처 사업보고서에 대한 통합 DTD의 구조

	구조정보	세부구조정보	구조명
문서 메타정보	수행부처		performAgency
	과제명		projectName
	이하 생략		
서지정보	제어번호		controlnumber
	과제관리번호		tasknumber
	이하 생략		
요약정보	초록	초록-국문	abstract(k)
		초록-영문	abstract(e)
	소개, 본문	장	chapter
	장	절	section
	이하 생략		
본문	본문	구분	part
		장	chapter
	장, 절	장	chapter
기타	부록, 맺음말		epilog

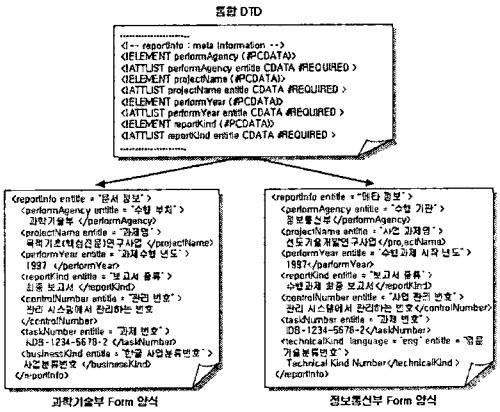
위와 같은 통합 DTD는 다양한 구조의 여러 문서들을 효과적으로 구조화할 수 있으며, 일반 사용자들도 쉽게 XML 문서를 편집할 수 있도록 폼 정의 기능을 제공한다

XML 및 SGML과 같은 구조정보를 입력하기 위한 기존의 편집 시스템에서는 사용자들이 작성하려는 문서 구조에 대하여 상세히 알고 있어야 한다. 그 결과 기존 시스템에서의 SGML/XML 작성은 구조언어의 구문과 DTD에 대한 지식이 필수적으로 요구된다. 따라서 이러한 지식이 부족한 일반 사용자들은 쉽게 구조 문서를 작성할 수 없을 뿐만 아니라, 작성된 문서에 대한 검증과정이 필요하다. 이러한 문제를 해결하고자 본 논문에서는 일반 사용자들도 쉽게 올바른 구조문서를 작성할 수 있는, 태그 없는 폼 (Form) 기반 입력 방식 시스템을 제안한다.

이러한 폼 기반 입력 방식을 위해서 본 연구에서는 통합 DTD에 기반한 부처별 폼을 정의하였다. 폼이란 통합 DTD에 정의된 다양한 부처의 구조 정보를 XML 문서

로 정의한 최소단위의 입력 문서 단위를 의미하는 것이다. 이렇게 정의한 폼을 부처별로 제공한다면 사용자는 DTD에 독립적으로 XML 문서를 쉽게 작성할 수 있게 된다.

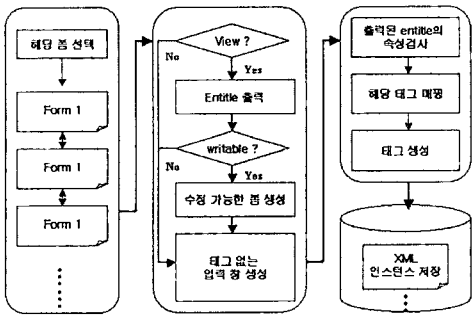
본 논문에서는 3개 부처 3개 과제의 최종 보고서를 지원하는 폼을 정의하였고, 정의된 폼은 입력기에서 부처별 해당 보고서 작성에 사용된다. 이러한 통합 DTD에서 정의된 폼을 [그림 2]에서 나타내고 있다.



[그림 2] 3개 부처 통합 DTD와 각 부처별 폼 양식의 예

2.3 태그 없는 폼 기반 XML 편집 시스템

기존 시스템에서의 구조문서 작성은 구조언어의 구문이나 DTD에 대한 지식이 필수적으로 요구된다.

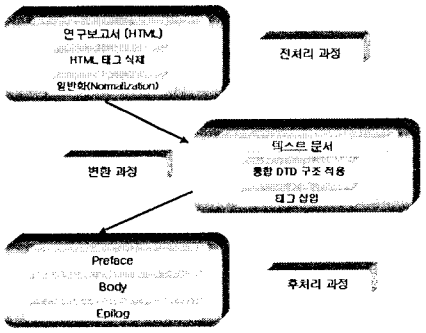


[그림 3] 폼 기반 입력 시스템의 세부 처리과정

따라서 본 논문에서는 위와 같은 문제를 해결하기 위하여 폼 기반 입력 방식을 제안하며 이에 기반한 태그 없는 입력 양식을 지원하는 XML 입력기를 개발한다. 폼 기반 입력 방식은 2.2절에서 살펴본 부처별 폼을 기반으로 XML 구조 문서를 작성하는 방법이며, 이러한 폼은 편집 시스템에서 태그 없는 입력방식으로 처리된다. 이러한 기법을 통하여 일반 사용자들도 DTD 및 문서 구조에 독립적인 XML 문서를 쉽고 정확하게 작성할 수 있다. 이러한 폼 기반 입력 방식의 XML 입력기 처리 구조는 [그림 3]과 같다.

2.4 XML 변환 시스템

과학기술평가원에서 구축한 기존 연구보고서는 HTML 문서 형태이다. HTML 문서는 웹에서 출력될 스타일 정보를 중심으로 구성되어 있는 문서이기 때문에 XML로 변환하기 위한 구조 정보가 부족하다. 따라서 본 연구에서는 HTML과 같은 비구조적인 문서에서 구조문서인 XML 형태로 변환하기 위한 새로운 변환 규칙을 제시하고, 이에 기반한 XML 변환기를 개발한다. 즉, 개발된 XML 변환기는 기존의 과학기술평가원 연구보고서를 본 연구에서 제안한 XML 문서 형태로 변환시켜주며, 변환 후 사용자가 추가적인 작업을 손쉽게 수행할 수 있는 환경을 제공한다. 전체적인 문서 변환 과정은 [그림 4]와 같으며 전처리 과정, 변환 과정, 후처리 과정으로 분류할 수 있다.



[그림 4] XML 변환기의 내부 처리 과정

전처리 과정은 기존의 연구보고서를 변환하기 쉽게 HTML 태그들을 제외하는 과정이다. 변환과정은 연구보고서의 가장 큰 부분인 본문 부분을 사용자가 쉽게 변환하는 과정이다. 이 과정이 끝난 후 변환기는 변환된 부분 이외에 서지정보(preface)와 에필로그(epilog)부분을 덧붙여서 유효한(valid) 연구보고서용 문서를 생성한다. 변환 과정에서 사용하는 구조 정보는 연구보고서 마다 다르기 때문에 사용자는 먼저 어떤 연구보고서로 변환할 지를 선택하여야 하고 이에 해당하는 서지정보와 에필로그 부분이 추가된다. 후처리 과정은 본문을 제외한 나머지 부분에 대한 정보를 입력하는 과정으로 편집기를 통하여 수행된다.

3. 시스템 개발

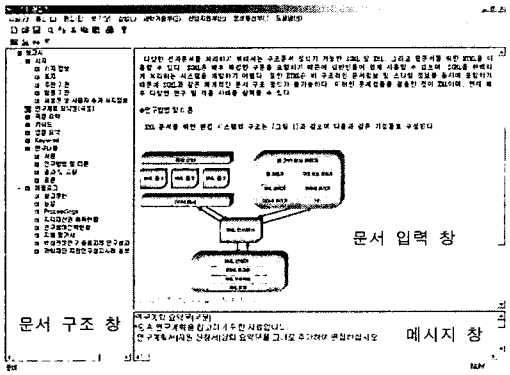
본 시스템은 Windows 98환경에서 Visual C++ 및 MSXML 파서를 이용하여 개발되었다. 시스템의 세부

기능은 다음과 같다.

3.1 XML 입력기 개발

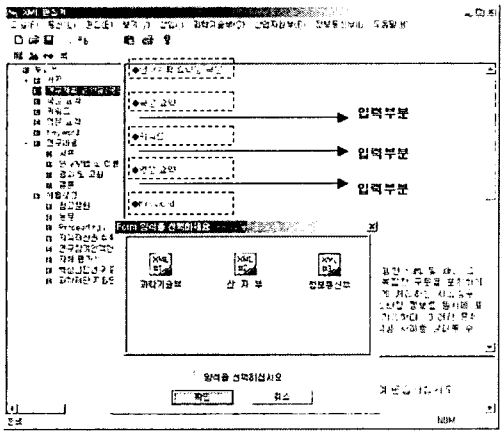
3.1.1 XML 입력기의 구성

본 논문에서 개발한 폼 기반 XML 편집기는 XML 문서 구조를 트리로 출력하는 문서 구조 창, 태그 없는 XML 문서 입력 창, 각 항목의 내용을 세부 설명하는 메시지 창으로 구성된다. 폼 기반 XML 편집기의 전체 실행화면은 [그림 5]와 같다.



[그림 5] 폼 기반 XML 입력기의 실행 화면

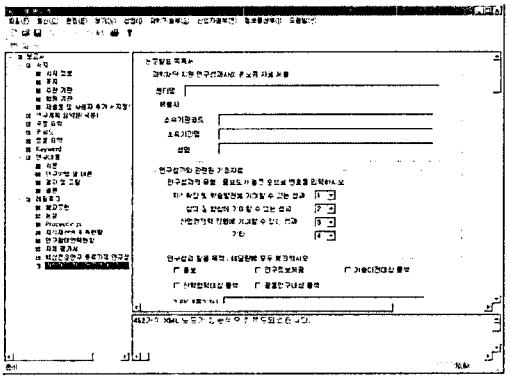
3.1.2 태그 없는 폼 기반 XML 입력 기능



[그림 6] XML 입력기의 입력 환경

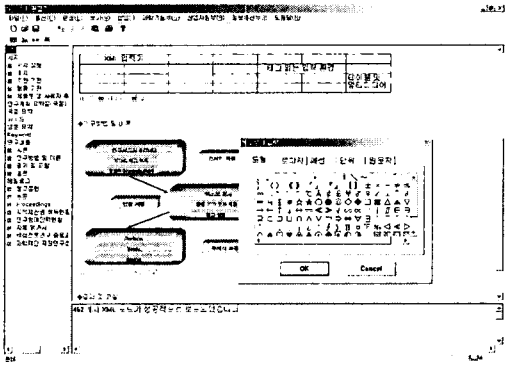
본 연구에서 개발한 XML 입력기는 본 논문에서 제안한 통합 DTD의 폼 양식을 이용하여 사용자에게 부처 별 양식을 제공한다. 사용자는 문서 입력창에서 변경이 불가능한 제목을 제외한 나머지 부분에 일반 워드프로세서처럼 내용을 입력하기만 하면 쉽고 정확하게 XML 문서를 작성할 수 있다. 이러한 과정을 [그림 6]에서 나타내고 있다.

또한 연구보고서의 서지나 부록부분의 복잡한 내용을 쉽게 처리하기 위하여 본 연구에서는 별도의 입력 양식을 제공한다. 이러한 입력양식은 텍스트 박스, 콤보박스 및 체크박스 등으로 구성되어 있기 때문에 복잡한 XML 구문을 이해하지 않아도 누구나 쉽게 XML 문서를 입력할 수 있다. 본 입력기의 서지 및 에필로그 처리 과정은 [그림 7]과 같다.



[그림 7] 서지 및 에필로그 입력 화면

3.1.3 테이블 및 멀티미디어 기능



[그림 8] 입력기의 다양한 기능 예

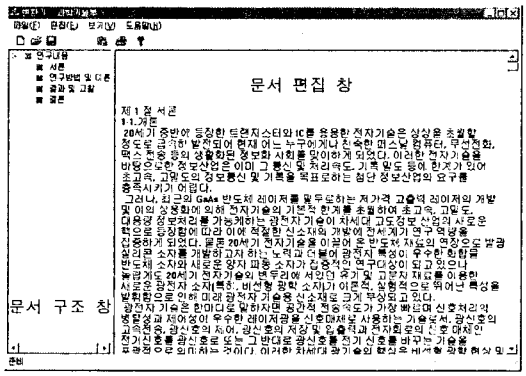
본 입력기는 태그 없는 XML 입력기능 외에도 일반 워드프로세서의 테이블, 이미지, 수식, 특수 문자 등의 삽입 및 편집 기능을 제공한다. 테이블 편집, 이미지, 특수 문자 등의 삽입 예는 [그림 8]과 같다.

3.2 XML 변환기 개발

3.2.1 XML 변환기의 구성

XML 변환기는 변환 문서의 구조를 나타내는 문서 구조창, XML 구조적용 및 편집 기능의 문서 편집 창으로

구성되며 내용은 [그림 9]와 같다.



[그림 9] XML 변환기의 기본 구조

본 연구의 변환기는 전처리 과정으로 HTML을 입력 받아 이를 텍스트 형태로 변환하게 된다. 이 과정에서 HTML 태그 삭제 및 일반화(Normalization)가 수행되며, 특히 HTML의 테이블 및 이미지 태그는 삭제하지 않는다. [그림 9]의 문서 편집 창에서 HTML 문서를 텍스트로 변환한 결과를 나타내고 있으며, 이러한 텍스트에 XML 구조를 적용하게 된다.

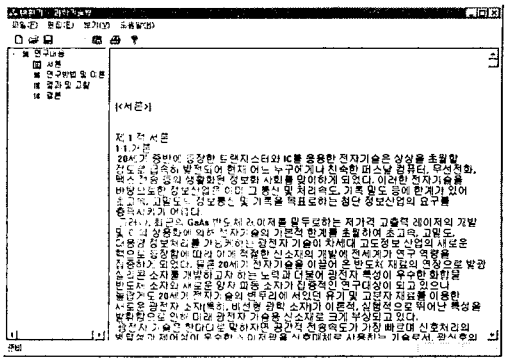
3.2.2 XML 구조 적용

HTML 문서를 본 논문의 통합 DTD에 기반한 XML로 변환하기 위해서는 먼저 변환하려는 부처의 양식을 선택하여야 한다. 사용자의 부처 선택에 따라 변환기에서 제공하는 XML 구조는 상이하며, 문서 구조 창에 엘리먼트 형태로 구조를 제공한다. 사용자가 편집창의 위치를 지정한 후에 구조 창의 엘리먼트를 선택하면 해당 구조가 적용되며, [그림 10은] 과학기술부, [그림 11]은 산업자원부의 HTML 연구보고서에 XML 구조를 적용한 예이다.

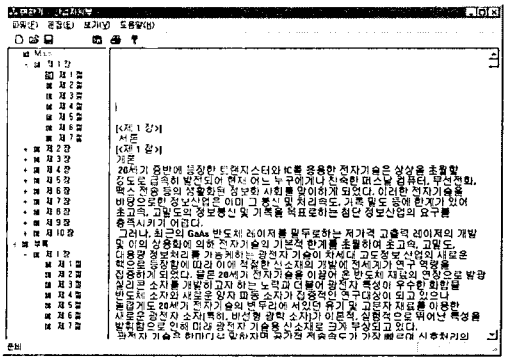
4. 결론

본 논문에서는 기존 구조문서 시스템에서의 문제점을 해결기 위하여 통합 DTD 정의 기법과 Form 기반의 구조문서 처리 방법을 제안하였고, 이에 근거한 Form 기반 XML 문서 편집 시스템을 구현하였다.

그 결과 본 연구에서 개발한 XML 시스템에서는 다양하고 복잡한 문서에 대하여 효율적인 문서구조가 가능하고, XML문서를 폼을 이용하여 누구나 쉽고, 정확하게 작성할 수 있다. 그리고 웹에서 사용된 HTML 문서를, 본 연구에서 정의한 통합 DTD 구조에 일치하는 XML 문서로 쉽게 변환할 수 있다



[그림 10] 과학기술부 문서에 대한 XML 구조 적용의 예



[그림 11] 산업자원부 문서에 대한 XML 구조 적용의 예

[참고문헌]

[1] ISO 8879:1986. Information Processing - Text and Office System - Standard Generalized Markup Language (SGML), Oct. 15 1986.

[2] Tim Bray and Jean Paoli, "Extensible Markup Language(XML) 1.0," http://www.w3c.org/TR/1998/REC-1998210.html, 1999.

[3] 윤승구, 이경호, 최윤철, "DSSSL 기반 SGML 문서 편집기의 설계 및 구현," 정보과학회논문지(C), Vol. 4, No. 6, 1999.

[4] 조영환, 강지훈, "SAX를 이용한 XML 문서 편집기," 정보과학회 춘계학술발표논문집, Vol.26, No.1, 1999.

[5] Techno2000 Inc. "CLIP ! V 1.5," http://www.t2000-usa.com, 2000