# 웹 스키마를 이용한 HTML 문서의 XML 변환

오금용° 박동문 황인준
아주대학교 정보통신전문대학원 정보통신공학과
{pickboy, minerva, ehwang}@madang.ajou.ac.kr

# XML Conversion of HTML Documents Using Web Schema

Keumyong Oh° Dongmoon Park Eenjun Hwang
The Graduate School of Information and Communication, Ajou University

**요 약**

최근에 웹(Web) 사용의 지속적인 증가로 인하여 정보가 급증하고, 이로 인하여 웹은 정보교환의 의미뿐 아니라 정보 저장이라는 중요한 의미를 지니게 되었다. 하지만 현재 많은 웹 페이지들이 HTML(Hyper Text Markup Language)문서로 제작되어 있어 정보관리의 의미에서 많은 부족함이 있고 이를 보완하기 위한 방법 중에 하나가 구조적이고 기능적 언어로 부상하고 있는 XML(eXtensible Markup Language)을 기반으로 하여 문서를 제작하거나 변환하는 것이다. 본 논문은 HTML문서를 XML문서로 변환하는데 있어 HTML문서 구조를 분석하고 분석결과를 토대로 형성되는 웹 스키마(Schema)를 이용하여 구조 중심의 변환이 이루어지도록 하는 방법에 대해서 제안한다.

## 1. 서론

웹은 이미 정보 교환의 기능을 뛰어넘어 정보 저장소의 역할까지 하고 있다. 현재 웹 정보 저장에 가장 많이 사용되는 웹 문서인 HTML은 간편하고 사용하기 쉽다는 장점을 가지고 있는 반면 제한된 태그들로 인하여 구조화되지 못한 문서의 형태를 보이고 있어 정보축적의 한계가 있다. 이러한 HTML의 단점을 보안하기 위해 W3C(World Wide Web Consortium)는 1996년 웹 문서의 표준으로 다양한 기능들과 구조적인 표현 능력을 가진 SGML(Standard Generalized Markup Language)에 기반을 둔 XML을 제안하였다. XML은 SGML의 기능성과 구조적 표현성을 지원할 뿐 아니라 사용의 편리성을 강조하였다. 사용자가 문서상에 사용될 태그를 자유롭게 정의할 수 있으며 또한 다른 사람들도 그 태그를 사용할 수 있다. 즉 XML은 본질적으로 다른 언어를 기술하기 위한 메타언어이다.

XML이 가진 확장성과 편리함 때문에 많은 웹 문서들이 XML로 작성되고 있지만, 기존에 있는 다수의 웹 문서들은 HTML로 만들어져 있기 때문에 웹 데이터의 통합적이고 효율적인 관리를 위하여 XML 문서로의 변환이 필요하다. 이런 방향으로 현재 진행중인 연구로는 W4F, XWRAP, XML Wrapper등이 있다. 본 논문에서는 이제까지의 연구에서 나타난 HTML 문서의 일부를 자바 객체로의 매핑(Mapping)이나 단순한 정보 축출에 그치지 않고 문서의 구조를 분석하고 데이터들간의 관계를 정의하여 이를 토대로 HTML 문서들을 탐색하면서 XML 문서로 변환하는 방법을 제안한다.

## 2. 관련연구

### 2.1. W4F

W4F(WysiWyg World Wide Web Factory)는 특정 웹에 있는 자원을 쉽고 간편하게 래퍼(Wrapper)로 만들 수 있도록 해주는 툴킷(toolkit)이다. W4F는 반 자동식(semi-automatic)으로 웹 문서의 데이터를 자바 클래스로 생성한다. 사용자는 이 클래스 파일을 변형하여 다른 애플리케이션에 재사용 할 수 있다. 그렇기 때문에 웹 문서에 대한 접근이 필요한 데이터베이스 시스템이나 소프트웨어 에이전트(software agent) 부분에 래퍼로 사용되기도 한다. 하지만 이 툴킷은 HTML 문서를 파싱(parsing)해서 그 태그를 자바 객체로 매핑시키는 정도에만 그치고 있어서 사용자가 필요한 용도에 따라 가공해야 한다는 단점이 있다.

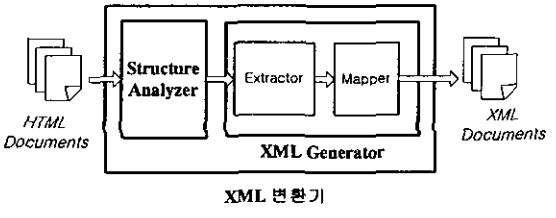
### 2.2. XWRAP

XWRAP(eXtensible Wrapper Generation System)는 웹 문서를 위한 래퍼 프로그램의 반자동적인 생성을 위한 소프트웨어이다. XML로 제작되지 않은 웹 문서를 XML 문서로 표현하는 기능과 불필요한 부분에 대한 제거를 통한 내용선택 기능을 제공한다. 자바로 구현된 이 프로그램의 장점은 사용자 인터페이스가 제공되어지고 래퍼의 생성이 용이하다는 것이다. 그러나 XML이 아닌 문서에서 구조를 분석하는데 있어 미흡한 면이 있고, 사용자가 개입하여 변환을 할 경우 다수의 문서를 위한 공통적인 DTD의 생성이 힘들어지는 단점이 있다.

### 3. XML 변환기

기본적인 변환과정은 사이트의 구조를 중심으로 이루어진다. 변환하기 위한 HTML 문서가 저장되어있는 사이트의 전체 구조를 트리로 나타낼 경우 트리의 첫 출발점은 사이트의 초기페이지이거나 혹은 문서를 링크하고 있는 페이지이다. 변환과정은 트리를 탐색하면서 마지막 노드에 해당되는 HTML 문서에 도착하게 되면 이루어진다. 한 노드의 XML 변환이 끝나면 링크를 통해 전후로 연결되어 있는 페이지를 추적하여 위의 변환과정을 반복하게 된다.

[그림 1]은 XML 변환기의 전체구조를 나타내는 그림으로 기능에 따라 HTML 문서의 구조를 분석하는 구조 분석기(Structure Analyzer)와 XML 문서를 생성하는 XML 생성기(XML Generator)로 구성되어 있다. 구조 분석기는 HTML 문서를 파싱하여 구조화된 HTML 문서를 생성하고 이를 기반으로 문서들간의 공통적인 구조를 파악하여 변환을 위한 웹 스키마를 정의한다. XML 생성기는 구조화된 HTML 문서로부터 XML 문서에 들어갈 데이터를 추출하는 추출기(Extractor)모듈과 웹 스키마로부터 만들어진 DTD(Document Type Definition)와 추출된 데이터를 가공해서 XML 문서를 생성하는 매퍼(Mapper)모듈로 나누어진다.




[그림 1] XML 변환기의 전체 구조

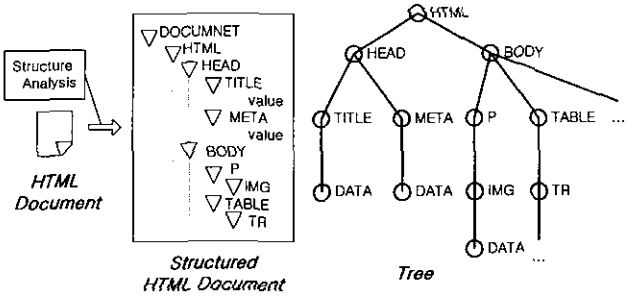
#### 3.1. HTML 문서의 구조 분석

구조 분석기는 HTML 문서의 유효성에 대한 검증을 우선적으로 한다. 변환될 HTML 문서는 다음의 항목을 갖추고 있어야 한다.

- 모든 엘리먼트는 시작태그와 끝태그를 가지고 있어야 한다.
- 모든 엘리먼트는 올바른 포함 구조를 가져야 한다.
- 모든 속성 값들은 따옴표 속에 나타나야 한다.

HTML 문서에 대한 유효성이 검증되면 파싱과정을 통해 구조분석이 이루어지고, 각 태그의 트리(Tree)구조가 생성된다. 파싱된 HTML 문서 안에는 단순히 태그와 텍스트뿐 아니라 다른 웹 문서로의 링크(Link)나 테이블(Table) 혹은 이미지, 오디오, 비디오 파일과 같은 멀티미디어 컨텐트가 있을 수 있다. 따라서 구조 분석에서는 태그뿐만이 아니라 HTML 문서에 사용되는 링크나 멀티미디어 컨텐트까지도 객체로 취급해서 모델링(modeling)한다.

[그림 2]는 구조화된 HTML 문서 생성의 예를 보여준다. 트리에 들어가는 엘리먼트로는 루트노드와 내부노드, 그리고 말단노드가 있고 각 노드는 해석된 HTML 태그로 표현된다. 루트노드는 HTML 문서를 대표하며, 내부노드는 HEAD, BODY를 나타내는 노드에서부터 HTML의 가장 작은 단위의 태그를 나타내는 노드들을 구성되며 각 노드들은 태그가 가지는 이름, 속성, 값의 정보를 포함하게 된다. 말단노드에는 추출될 데이터가 저장된다. 태그에서 사용되는 URI(Uniform Resource Indicator)나 URL(Uniform Resource Locator)은 태그를 나타내는 노드의 자식노드로 표현된다.




[그림 2] 구조화된 HTML 문서

#### 3.2. 스키마 생성

스키마 생성은 구조화된 HTML 문서를 통해서 이루어진다. 웹 상에서 다루어지는 HTML 문서는 일반적으로 주제(subject)에 따라 분류할 수 있으며 분류된 문서들은 공통적인 형식을 지니고 있다. 이러한 문서들은 구조분석 과정을 거치면 유사한 트리 구조를 보이고 이를 바탕으로 스키마 분석이 이루어진다. 정확한 스키마 분석을 위하여 말단노드가 가지고 있는 실질적인 데이터를 분석하여 스키마에서 사용될 의미를 부여할 수 있다. 사용자에게 스키마 분석을 위한 인터페이스가 제공되어지며 사용자는 구조화된 문서를 통해 각각의 데이터에 의미를 부여할 수 있다.

예를 들어, 공공기관에서 사용하는 기안문을 나타내는 HTML 문서를 구조화하였을 때 각각의 기안문에서 발신자의 데이터가 *html.body.table[5].tr.td.p.font.font*에 위치하고 있다면 변환을 위한 스키마는 <문서><기안문서><문서번호><발신><발신인>으로 표현되며 이러한 구조는 변환되어질 기안문서에 공통적으로 사용될 수 있다. <문서>나 <기안문서>등의 요소는 문서뿐 아니라 전체적인 구조를 고려하여 생성되어지는데, 이러한 스키마 분석과정은 변환과정에서 한번만 수행되며 이를 토대로 변환을 위한 DTD가 생성되고, 구조가 같은 문서들은 이 DTD를 따라 자동으로 변환되어 진다.

#### 3.3. DTD 생성

구조화된 HTML 문서를 기반으로 생성된 스키마 구조는 변환과정에서 사용될 DTD와 흡사하다. 하지만 웹 스키마는 XML 문서가 참조하게 될 DTD가 아니기 때문에 변환후 XML 문서로 저장된 다른 문서들이 참조해야할 DTD를 생성해야 한다. 이 과정은 다른 XML 문서와의 통합등을 고려하여 사용자나 개발자의 개입을 통해 원하는 DTD를 생성할 수도 있고, 위의 과정에서 얻어진 스키마의 구조를 그대로 반영하는 자동적인 방법을 사용할 수도 있다. 자동적인 방법은 일정하게 정해진 문서의 위치 판단 방법이나 스트링 매치 방법 등을 통해 데이터 추출이 가능하다는 점을 고려하여 웹 스키마의 요소들을 그대로 XML 문서에 사용될 DTD의 요소로 사용하는 것이다.

이렇게 DTD는 구조화된 HTML 문서를 바탕으로 형성된 웹 스키마를 토대로 하여 작성되어질 수 있다. 생성된 DTD의 요소정보는 추출기와 매퍼에 전달되어지고, 일련의 과정을 통해 XML 문서로 변환하게 된다. 또한 생성된 DTD를 참조하여 링크되어 있는 다른 문서의 변환을 시도할 때 문서의 구조 분석 판단을 통해 유사한 문서로 판단되어지는 경우 반복적으로 변환을 하게 된다.

### 3.4. XML 문서의 생성

변환과정 중 마지막 단계인 문서 생성단계에서 데이터 추출기는 HTML 문서의 트리구조 말단에 위치한 데이터를 추출하고 이를 객체로 만들어 매퍼에게 전달한다. 매퍼는 DTD에서 정의한 요소들과 객체들을 매핑하고 파일기록기를 호출하여 매핑된 데이터 기록을 통해 XML 문서를 작성하게 된다. 이때 사용되어진 DTD의 정보를 문서에 포함시킨다.

## 4. 예제 - 뉴스사이트

뉴스 정보 제공은 시간에 따라 분류가 되고, 고정된 날짜의 뉴스페이지를 보면 여러 개의 하위 목록으로 연결되어 있다. 그 목록들은 연계된 또 다른 하위페이지로 반복적으로 연결되어 있다. 예를 들어 뉴스 사이트를 보면 정치, 경제, 스포츠 등의 여러 가지 목록을 볼 수 있고, 스포츠는 야구, 축구, 농구 등의 여러 목록으로 나누어져 연결되어 있다. 이러한 중간 페이지들은 또 다른 기준에 맞게 반복적으로 연결되어 있으며 그 끝은 특정한 사건중심인 HTML 문서임을 알 수가 있다. 이 구조를 바탕으로 하여 변환을 위한 탐색이 가능하고, 탐색과정에서 말단에 위치한 문서를 분석하여 XML 문서로의 변환을 위한 웹 스키마를 정의 할 수가 있다. 또한 링크의 특성을 이용하여 한 문서의 변환이 끝나면 링크주소를 따라가 또 다른 변환이 반복적으로 이루어지게 된다. 각 노드에 대한 접근방법은 점 접근방법(Dot Notation), 화살표 접근방법(Arrow Notation)과 같은 방법이 사용된다.

- Dot_notation : (newspaper[1].sports[2].baseball[30].xxx....)
- Arrow_notation : (xx->x[10])

위와 같은 방법으로 트리의 각 노드에 접근해서 대응하는 태그의 이름, 속성과 값을 얻어낸다.

변환의 대상은 기사문서를 중심으로 이루어지고, 이는 구조분석기에서 구조화된 HTML 문서를 형성하게 된다. 구조화된 HTML 문서의 형식을 트리로 표현하면 기사문서마다 얻어지는 구조가 유사하게 된다. 이를 분석하여 변환된 XML 문서가 사용할 DTD를 위한 웹 스키마를 얻게 된다. 다음은 몇 개의 기사를 구조화하여 얻어진 결과 중 공통적인 구조의 일부분을 나타내고 있다.

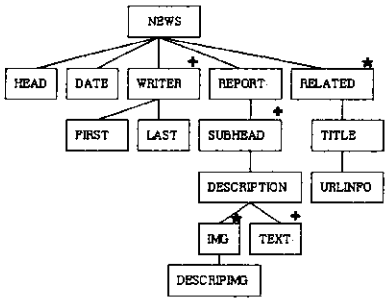
- 제목 : html.body.p.font.b
- 글 : html.body.p[i]
- 사진 : html.body.table[j].tr[k].td.img
- 사진설명 : html.body.table[j].tr[k+1].td.div

이러한 구조는 신문 전체의 구조를 고려하여 다음과 같이 스키마를 형성할 수 있다.

<뉴스><스포츠><국내><프로축구><기사><제목>
<뉴스><스포츠><국내><프로축구><기사><글>
<뉴스><스포츠><국내><프로축구><기사><사진>
<뉴스><스포츠><국내><프로축구><기사><사진설명>

스키마 정보중 <뉴스>, <스포츠>, <국내>, <프로축구> 등은 사이트의 전체 구성에 대한 정보를 바탕으로 얻어지게 되고, 트리의 첫 노드인 html요소는 <기사>가 된다.

이를 토대로 변환된 XML 문서의 바탕이 되는 DTD를 만들 수 있다. [그림 3]은 CNN의 뉴스사이트에 대해 HTML 문서의 구조분석결과로 나온 웹 스키마를 기반으로 DTD의 일부 중 기사(article)를 위한 부분을 트리구조로 표현하였다.




[그림 3] 기사를 위한 DTD 트리구조의 예

변환되어진 XML 문서가 사용할 DTD를 기반으로 구조분석기를 통해 나온 트리구조의 말단노드들에 위치한 데이터들을 추출하게 된다. 이렇게 추출된 데이터들은 매퍼를 통해 새로이 생성되어지는 XML 문서상에 들어가게 되고 DTD에서 정의한 요소를 태그로 부여받아 XML 문서가 생성되고 링크를 따라서 반복적으로 변환이 진행되어진다.

## 5. 결론

본 연구에서는 HTML 문서를 구조분석을 통한 XML 문서로 변환하는 과정을 제안하였다. 변환과정은 HTML 문서의 구조를 분석하여 스키마를 정의하고, 그에 따른 데이터 추출과 매핑방법을 사용하였다. 이러한 방법을 응용하면 정보를 제공하는 사이트의 특성과 관련된 HTML 문서의 구조를 분석을 통하여 스키마와 DTD를 정의 할 수 있고 이를 기반으로 XML 문서로의 자동적인 변환을 지원할 수 있다. 향후 연구과제로는 저장되어져 있는 XML 문서를 대상으로 하여 질의 시스템과 연계하고, XSLT(eXtensible Stylesheet Language Transformations) 등을 적용하여 사용자의 취향에 맞는 맞춤형 정보제공을 자동적으로 이루어지게 하는 시스템을 구현하는 것이다.

### 참고문헌

[1] N.Ashish, C.Knoblock, "Wrapper Generation for Semi-structured Internet Sources", Proceeding of the Workshop on Management of Semi-structured Data, Tucson, Arizona, 1997
[2] Ling Liu, Wei Han, David Buttler, Calton Pu, Wei Tang, "An XML-based Wrpper Generator for Web Information Extraction"
[3] Arnnaud Sahuguet, Fabien Azavant, "WysiWyg Web Wrapper Factory (W4F)"
[4] D.Taniar, Y.Jiang, J.W. Rahayu, L. Bishay, "Structured Web Pages Management for Efficient Data Retrieval"
[5] World Wide Web Consortium: eXtensible Markup Language (XML) 1.0, Feb.10.1998