# DTD의 의미 구조 분석을 이용한 XML 문서의 변환

곽동규° 최종명* 조용윤 유재우
숭실대학교 컴퓨터 학과 *국립목포대학교 컴퓨터 공학과
coolman@ss.ssu.ac.kr°, *choijm@dreamwiz.com, yycho@ss.ssu.ac.kr, cwyoo@computing.ssu.ac.kr

## A Transformation of XML Documents With Semantic Constraints

Donggyu Kawk° Jongmyung Choi* Yongyoon Cho Chaewoo Yoo
Dept. of Computing, Soongsil University
*Computer Engineering, Mokpo National University

요 약

XML 문서를 변환하는 목적은 하나의 어플리케이션에서 사용되는 XML 문서를 다른 XML 어플리케이션에서 재사용하여 사용자에게 동일한 정보를 제공하는데 있다. XML 문서는 어플리케이션 특성에 따라 한 문서에서 전달할 수 있는 정보의 량이 다르다. 따라서 문서를 변환하기 전에 어플리케이션의 특성에 따라 문서를 분할하거나 병합하여야 한다[1]. 또한, XML 문서의 정보는 속성에 따라 문법적인 특성을 가진다. 본 논문은 문법적인 특성을 의미 속성이라 하고 의미 속성을 파악하기 위해 XML 문서의 의미 구조와 의미 구조 관계를 정의한다. 그리고 정의된 의미 구조와 의미 구조 관계를 이용하여 문서 정보의 속성을 분류하는 방법을 제안한다. 변환 규칙은 의미 구조 관계가 유사한 엘리먼트간의 대응으로 정의하고, 변환 규칙을 이용하여 문서 변환을 실행하여 변환 XML과 의미 관계 구조가 유사한 피 변환 XML 문서를 생성한다. 의미구조 분석을 이용한 변환은 기존의 사용 패턴을 분석한 변환에서 벗어나 DTD의 분석을 통한 자동화된 문서 변환 방법을 제공한다.

## 1. 서 론

XML 문서는 그것을 데이터로 사용하는 서로 다른 어플리케이션에 따라 많은 분야에서 다양한 목적과 형식으로 사용된다. 기존에 작성된 XML 문서 정보를 다른 XML 어플리케이션에서 사용하기 위해서는 XML 문서 변환이 필수적이다. 따라서 어떤 어플리케이션을 위한 XML 문서를 다른 XML 어플리케이션에서 재사용하기 위해서는 변환 XML 문서가 가지는 정보와 구조의 손실 없이 피 변환 XML 어플리케이션을 위한 구조와 스타일로 적절히 변환되어야 한다. 예를 들면, 웹 브라우저를 위해 작성된 XHTML의 정보를 음성 단말기에서 사용하기 위해서는 XHTML 문서를 VoiceXML[4]로 변환하여야 가능하다.

XML 문서의 구조는 엘리먼트(element)간의 종적, 횡적 결합으로 이루어져 있다. 따라서 XML 문서를 구성하는 엘리먼트의 결합 구조에 대한 분석정보는 동일한 의미를 갖는 다른 형태의 XML 문서 구조로 변환하기 위한 정보로 사용될 수 있다. 본 논문은 XML 문서를 구성하는 엘리먼트간의 결합을 분석하기 위해 대수적 표현 방법을 사용한다. 대수적 표현 방법을 이용한 XML 문서의 변환은 XML 문서를 어플리케이션의 특성에 따른 분할/병합 단계와 분할/병합된 문서를 변환하는 단계로 나누어진다. XML 문서의 분할과 병합 단계에서 XML 문서의 엘리먼트 구조 요소는 수학적 수식으로 표현될 수 있으며, 배분법칙과 같은 대수적 특성을 이용해 XML 문서를 어플리케이션의 특성에 따라 적절한 크기로 분할하고 병합 할 수 있다[1]. 대수적 표현 방법을 통해 분할된 XML 문서는 변환 규칙이 적용되어 원하는 형태의 XML 문서로 재작성 될 수 있다.

문서의 변환에서 적용하는 변환 규칙은 서로 다른 XML 문서 구조의 상응하는 엘리먼트간의 대응으로 이루어진다. 기존의 변환 규칙은 특정 XML에 대하여 변환 XML과 피 변환 XML의 사용패턴 분석을 통해 작성된 변환규칙을 작성한다. 그러나 이런 방법은 특정 XML 문서에만 적용할 수 있고, 새로운 XML에 따라 문서 변환 규칙을 생성할 수 없는 약점을 가지고 있다[5].

본 논문은 XML DTD 문법에 따라 XML 문서 데이터의 의미 속성을 파악하여 각 엘리먼트의 의미 구조를 분석하고, 이것을 통해 XML 문서를 변환하는 방법을 제안한다. 따라서 XML 문서의 의미 정보를 표현하기 위해 XML DTD에 대한 의미 제약(Sementic Constraints)[2] 방법을 이용하여 의미 구조 관계를 정의한다. 정의된 구조 관계를 이용해서 엘리먼트의 의미 구조 관계가 유사한 엘리먼트를 대응시켜 문서 변환 규칙을 정의할 수 있으며, 사용 패턴을 분석하지 않고 문서 변환 규칙을 생성할 수 있는 장점을 가질 수 있다. 또한, 변환 XML 문서와 피 변환 XML 문서의 의미 구조 관계를 비교하여 두 문서의 의미 구조 유사성을 비교할 수 있으며, 의미 구조 유사성을 통해 변환 규칙의 정확도를 확인할 수 있는 장점을 가질 수 있다.

본 논문은 2장에서 관련연구를 소개하고 3장 본론에서 의미 구조 관계와 의미 구조 유사도, 전체 시스템 구성에 대해 논한 후 4장에서 결론을 맺는다.

## 2. 관련 연구

### 2.1 마크 업 문서의 대수적 분석에 의한 분할과 변환[1]

이종 어플리케이션간의 XML 문서 변환에서 어플리케이

션의 특성에 따라 문서의 정보량의 차이가 있으므로, 문서를 분할하여 변환하는 방법을 제시하였다. XML 문서에서 태그의 결합은 연속된 결합과 포함된 결합으로 나누어진다. 연속된 결합은 DTD에서 SEQ와 OR로 나누어짐에 착안하여 포함된 결합은 연산 ()로 정의하고, 연속된 결합의 SEQ는 *로 OR는 +로 정의하였다. 또한, * 연산자는 엘리먼트들이 강하게 연결된 것을 의미함을 보이고 + 연산자는 엘리먼트들이 느슨하게 연결되어 있는 것을 보였다. 이 특성을 이용하여 배분법칙을 통해 문서를 분할하는 방법을 연구하였다.

2.2 Transform XML DTD to relational schema[2]

어떤 DTD에 유효한 XML 문서를 문서에 포함된 의미정보를 잃지 않고 데이터베이스에 저장하기 위해 Dongwon Lee[2]는 Relational Schema를 relational scheme과 semantic constraint의 순서쌍으로 정의하여 DTD의 문법적 구조를 의미적으로 Domain Constraints와 Cardinality Constraints, IDs, EGDs, TGDs로 나누어 데이터베이스 테이블 생성 시 각각 엘리먼트의 구조에 맞게 열을 생성하는 방법에 대해 제안하였다. 이 방법은 XML 문서의 의미적 구조를 잃지 않고 다른 형태의 데이터 포맷으로 변환할 수 있음을 보였다.

## 3. 본 론

일반적인 XML 문서는 DTD를 기술하는 개발자나 XML 문서를 기술하는 사용자의 의도가 각기 달라 일반적인 특성을 찾아내기가 어렵다. 그러므로 일반적으로 사용하는 XML의 사용 패턴을 다음과 같은 공리로 정의한다.

**공리**

1. 모든 XML 문서는 어떤 DTD에 유효한 문서이다.
2. 한 개체는 한 엘리먼트로 표현되고 개체의 정보는 개체를 표현한 엘리먼트의 자식 엘리먼트에 표현된다. (자식 엘리먼트를 포함하고 있는 엘리먼트는 부모 엘리먼트이다.)
3. 개체의 속성은 반드시 한번만 기술된다.(문서 전체의 속성도 문서 전체에서 반드시 한번만 기술된다.)

### 3.1 의미 구조 관계

XML 문서는 정보의 내용에 따라 보편적으로 사용하는 구조가 있다. 일반적으로 DTD 문서에는 문서의 정보나 속성을 표현하는 구조가 있고, 한 개체의 추가적인 정보는 엘리먼트의 포함으로 표현한다. 또한, 포함된 구조에서는 부모 엘리먼트와 자식 엘리먼트의 관계가 있고, 부모와 자식 엘리먼트 간에는 옵션(?), 0번 이상 반복(*), 1번 이상 반복(+)과 같은 정보의 속성에 따라 알맞은 문법 구조를 갖는다.

본 논문은 위와 같은 XML 문서의 특성을 구분하여 연구하기 위해 Relational Schema[2]를 변형하여 정의 1과 같은 의미 구조 관계를 정의한다.

**정의 1.** 의미 구조 관계

의미 구조 관계 R = (S, Δ)

의미 관계 Δ = (e, t)

$$e=\begin{cases} \text{if} & P_i\text{로부터 } C\text{가 한번을 초과하여 등장할 수 있으면 } 0 \\ other & 0 \end{cases}$$

$$t=\begin{cases} \text{if} & P_i\text{로부터 } C\text{가 등장하지 않을 수 있으면 } 0 \\ other & 1 \end{cases}$$

구조 관계 S = (Pi, C), P, C ∈ TAG, Pi는 i번째 부모 엘리먼트, C는 자식 엘리먼트.

정의 1의 의미 구조 관계는 하나의 개체를 표현하는 속성 엘리먼트의 구조를 파악하기 위한 정보를 준다. 예를 들어 XHTML의 "<head>"와 "<title>" 엘리먼트는 문서의 속성을 기술하는 엘리먼트로 문서에 반드시 한번만 등장한다. 그러므로 "<title>"엘리먼트는 루트 엘리먼트인 "<html>"로 부터 Δ = (1, 1)이다. 이 엘리먼트는 문서 전체에 반드시 한번만 등장하는 데이터이므로 문서 전체의 속성을 나타내는 데이터일 가능성이 높다. 같은 방법으로 어떤 태그에 대하여 Δ = (1, 1)인 엘리먼트는 등장하는 개체의 속성을 나타내는 데이터일 가능성이 높다. 또한 Δ = (1, 0)인 엘리먼트는 한번 등장하거나 등장하지 않는 엘리먼트의 의미 관계 값이다. 이런 속성을 갖는 데이터의 예는 XHTML의 "<style>"과 같은 엘리먼트가 있다. 이 엘리먼트는 문서의 속성을 표현하지만 필요에 의해서만 존재하는 엘리먼트이다. Δ = (0, 0)인 엘리먼트는 일반적으로 정보를 나열하기 위한 태그이고, Δ = (0, 1)인 엘리먼트는 하나 이상 반드시 가지고 있는 정보를 표현하기 위해 사용된다. 즉, Δ값은 문서의 의미적 구조를 나타내는 값이다.

문서의 변환 규칙은 Δ값이 근사한 엘리먼트간의 대응으로 이루어져야한다. 문서의 변환 규칙은 Δ값이 일치하는 엘리먼트간의 대응으로 이루어져야 한다. 하지만, 모든 XML 문서의 구조가 동일하지 않기 때문에 모든 엘리먼트의 Δ값이 완전히 일치하는 두 XML 문서는 찾기 어렵다. 그러므로 가장 근사한 엘리먼트간의 대응으로 대응시켜 변환 규칙을 생성한다.

### 3.2 의미 구조 유사도

전장에서 제안한 변환 규칙은 근사한 정도가 높은 엘리먼트로만 이루어진 문서는 높은 유사성을 가지는 문서를 생성하고 근사한 정도가 낮은 엘리먼트로만으로 이루어진 문서는 낮은 유사성을 가지는 문서를 생성한다. 왜냐하면 3.1에서 제안한 변환 규칙 생성 방법은 Δ값이 근사한 엘리먼트간의 대응으로 구성하기 때문이다. 그러므로 본 논문은 문서의 의미 구조의 유사성을 측정을 제안한다.

**정의 2.** 데이터 의미 구조 관계

데이터 의미 구조 관계 R = (S, Δ), Δ는 의미 관계

데이터 구조 관계 S = (Pi, C, D), P, C ∈ TAG, Pi는 i번째 부모 엘리먼트, C는 자식 엘리먼트, D는 데이터.

정의 2의 관계 구조 Δ의 e와 t는 DTD의 반복과 옵션을 나타내는 심볼에 따라 "1"이나 "0"의 값을 갖게 되는데 i=1일 때, 그 값은 표 1과 같다.

표 1. i = 1 일 때 관계 구조 Δ의 e와 t의 값

| 심볼 | e | t |
|---|---|---|
| ? | 1 | 0 |
| ε | 1 | 1 |
| * | 0 | 0 |
| + | 0 | 1 |

의미 구조의 유사성을 측정하기 위한 의미 구조의 유사도는 각 데이터의 관계 구조 Δ 값이 일치하면 1, 하나만 일치하면 0.5, 모두 일치하지 않으면 0의 값을 주어 그 평균으로 계산한다.

**정의 3.** 의미 구조의 유사도

$$\text{유사도 } S=\frac{\sum_{j=0}^{t}\delta_j}{t}$$

$$\delta_j=\begin{cases}\text{if } \Delta_{1i}=\Delta_{2i} \text{ that } 1\\ \text{if } (\Delta_{1i}.e=\Delta_{2i}.e \text{ and } \Delta_{1i}.t\neq\Delta_{2i}.t) \text{ or } (\Delta_{1i}.t=\Delta_{2i}.t \text{ and } \Delta_{1i}.e\neq\Delta_{2i}.e) \text{ that } 0.5\\ \text{other } 0\end{cases}$$

$t$ 는 한 문서의 모든 의미 관계 구조의 개수.

$\Delta_{ji}$에서 j는 문서 번호, i는 관계 번호.

정의 3에서 정의한 XML 문서간의 의미 구조 유사도를 측정할 수 있는 XML 문서 변환 시스템의 개략적인 구조도는 그림 1과 같다.

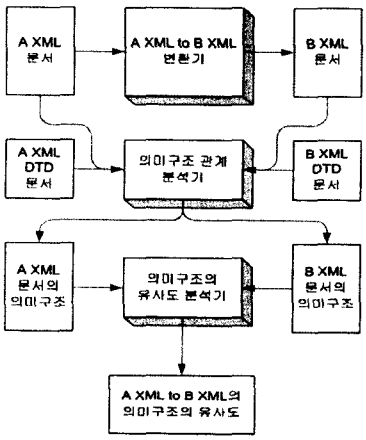


그림 1. XMLtoXML 개략적인 시스템 구조도

의미 구조를 분석하기 위해서는 XML 문서와 DTD를 입력으로 받는다. 변환 전 의미 구조와 변환 후 의미 구조의 유사도를 측정하여 XML 문서의 의미 구조가 얼마나 유지되었는지를 확인한다. 의미 구조의 유사도가 클수록 의미 구조는 유사하고 변환 전 XML 문서에 대하여 변환 규칙이 잘 정의된 변환기이다.

### 3.3 전체 시스템 구조

본 논문은 위와 같은 공리에 만족하는 XML 문서에 대하여 XML 문서 변환기를 그림 2와 같이 제안한다.

그림 2에서 의미 구조 분석기는 DTD에 포함되어 있는 의미 구조를 생성한다. 또한, 문서 변환 규칙 생성기는 생성된 두 의미 구조를 이용하여 XSLT[3]과 유사한 문서의 변환 규칙을 생성한다. XML 문서 분할기와 문서 변화기는 "마크 업 문서의 대수적 분석에 의한 분할과 변환"[1]에 따른다.

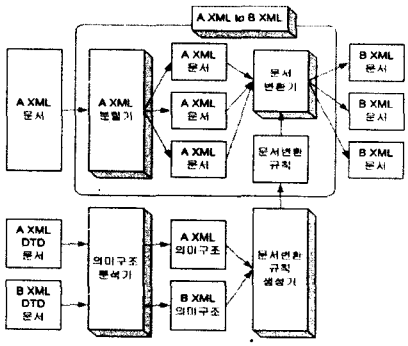


그림 2. DTD 의미 구조 분석기를 이용한 XMLtoXML 시스템 구조도

## 4. 결 론

XML의 사용이 증가함에 따라 변환이 필수적이다. 이에 따라 본 논문은 XML 문서의 정보가 문법적 특성에 따라 기술됨을 보이고 이런 문법적 특성을 연구하기 위해 의미 구조와 의미 구조 관계를 정의하였다. 의미 구조와 의미 관계는 DTD의 옵션(?) 그리고 0번 이상의 반복(*), 1번 이상의 반복을 대수적인 관계로 정의하였다. 문서 변환 규칙은 관계가 유사한 엘리먼트간의 대응으로 정의하여 문서간의 변환을 반자동으로 생성할 수 있는 방법을 제시하였다. 즉, 대수적 분석에 의한 변환 방법은 일반적인 XML 문서간의 DTD만을 이용한 변환 규칙 생성 방법이다. 그러므로 대수적 분석에 의한 변환 방법은 기존의 사용패턴을 이용한 방법[5]이 가지고 있던 일반적인 XML에 적용할 수 없는 약점을 극복하여 새로운 XML의 등장과 함께 사용할 수 있는 장점을 가지고 있다.

### 참고 문헌

[1] 곽동규, 최종명, 유재우, "마크 업 문서의 대수적 분석에 의한 분할과 변환", *한국정보과학회 프로그래밍언어 논문지* 제17권 제3호, pp.57~66, 2003.

[2] Lee, D. & Chu, W. W "Constraints-preserving transformation from XML document type definition to relational schema", *Proc. of Int'l Conf. on Conceptual Modeling* (ER), 2000.

[3] XSL Transformations (XSLT) Version 1.0, http://www.w3.org/TR/1999/REC-xslt-19991116.

[4] Voice Extensible Markup Language (Voice XML) Version 2.0, http://www.w3.org/TR/2004/PR-voiexml20-20040203.

[5] 최훈일, 장영건, "HTMLtoVoiceXML 변환기의 설계 및 구현", *한국정보과학회 논문지C*, 7권 6호, pp. 559-569, 2001.