시맨틱 어노테이션을 이용한 XML 문서 트랜스코딩

이진상° 송특섭, 손원성, 고승규, 임순범*, 최윤철
°연세대학교 컴퓨터과학과,
*숙명여자대학교 멀티미디어학과

{gr20000°, teukseob, sohnws, skko, ycchoy}@rainbow.yonsei.ac.kr

*sblim@sookmyung.ac.kr

XML Documents Transcoding using Semantic Annotation

Jin-Sang Lee°, Teuk-Seob Song, Won-Sung Sohn, Seung-Kyu Ko,
Soon-Bum Lim*, Yoon-Chul Choy
Dept. of Computer Science, Yonsei University
*Dept. of Multimedia, Sookmyung Women's University

요 약

기존의 웹 컨텐츠를 휴대폰이나 PDA등과 같은 개인용 단말기에 표현하기에는 단말기 성능상의 제약(낮은 CPU 성능, 작은 출력 화면, 입출력 방법의 단순함 등)이 따르게 되므로 컨텐츠 변환의 과정이 필요하게 된다. 트랜스코딩이란 기존의 웹 컨텐츠를 단말기의 환경에 따라 적합한 형태로 변환 하는 것을 의미하며, HTML 문서의 레이아웃 정보를 이용하여 변환하는 연구가 다양하게 이루어져 왔다. 본 논문에서는 사용자 의견을 반영한 XML문서의 정확한 트랜스 코딩을 위하여 시맨틱 어노테이션 기법을 제안한다. XML 문서의 트랜스코딩에는 IPTC(International Press Telecommunications Council)에서 정한 NewsML을 기반으로 하였으며, 본 논문에서 제안하는 트랜스코딩 프레임워크는 크게 3단계로 나뉘어 진다. 어노테이션 생성 및 인식, 어노테이션의 구조 정보를 활용한 페이지 생성 및 페이지 맵 구성, 디바이스에 따른 페이지의 변환으로 구성된다. 향후 연구로는 어노테이션과 페이지 생성 기법을 통해 생성된 XML 문서를 CC/PP를 이용하여 PDA나 휴대폰 등의 시스템에 적합하게 변환하는 기법 등이 요구된다.

1. 서 론

PDA나 핸드폰 등의 무선 단말기를 통해 기존의 웹 컨텐츠를 이용할 경우, 단말기 성능상의 제약성((낮은 CPU 성능, 작은 출력 화면, 입출력 방법의 단순함 등)[1][2] 이 존재하게 된다. 그러므로 이러한 소형 단말기에서 웹 컨텐츠를 이용하기 위해서는 단말기의 환경에 따라 적합한 형태로 컨텐츠를 변환하는 과정이 필요하다.[3]

트랜스코딩에 관한 기존 연구의 문제점과 한계를 몇 가지로 요약해 보면 다음과 같다. 첫째, HTML의 레이아웃 중심의 연구가 이루어져 왔다[3]. HTML 문서의 레이아웃 정보를 고려하여 트랜스코딩 할 경우, 불필요한 정보나 상호 연관성이 없는 정보가 같은 카테고리 내에 포함되는 문제가 발생하게 된다. 둘째, 컨텐츠에 종속적인 개별문서 위주의 트랜스 코딩 기법을 사용하였다. 개별문서 위주로 컨텐츠를 변환하게 되면 변환하고자 하는 모든 문서를 처리해야 하는 불편함이 따르게 된다. 셋째, 사용자의 의도를 반영하지 못하는 서비스 제공자 위주의 기법을 사용하였다.[4] 사용자의 의견을 반영하지 못하는 일방적인 트랜스코딩 기법은 불필요한 정보를 전송하게 하는 문제점이 발생하게 된다.

본 논문에서는 앞서 제시한 문제점들을 보완하며, XML 구조문서를 효과적으로 트랜스코딩 할 수 있는 어노테이션 트랜스코딩 기법을 제안하고자 한다. XML 문서의 트랜스코딩을 위하여 IPTC(International Press Telecommunications Council)에서 정한 표준 뉴스 포멧인 NewsML[5]을 기반으로 하였다. NewsML은 AFP, Reuter등의 언론사들이 실제 적용을 실험하고 있는 차세대 뉴스 표준으로 사용자의 의견을 반영하는 본 논문의 트랜스코딩 기법에 적합하다고 할 수 있다.

본 논문에서는 사용자의 의견을 반영하기 위한 어노테이션 인터페이스를 제안하고, 생성된 어노테이션 정보를 기반으로 XML 문서를 트랜스코딩 하는 기법을 제안하고자 한다.

2. 관련연구

2.1 트랜스코딩

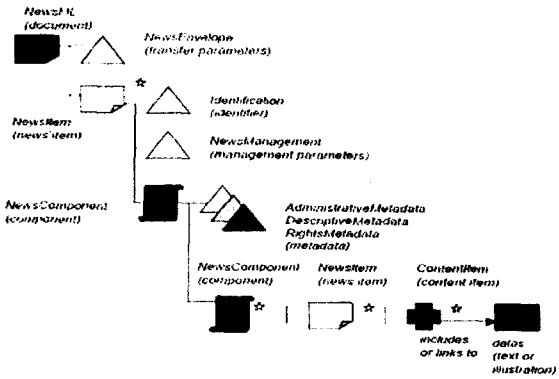
기존의 웹 문서를 소형화면 단말기로 변환하는 트랜스코딩 기법에 대한 연구는 주로 HTML 기반의 연구로서 텍스트를 요약, 추출하거나 수작업을 동반한 변환에 대한 연구가 진행되었으며[4][7], 최근에는 디바이스의 성능이 향상됨에 따라 디바이스

에서 변환을 실시하는 연구결과도 발표된 바 있다.[8] 이러한 내용을 정리하면 HTML 기반의 트랜스코딩[4], 문서 요약 및 필터링 기반 트랜스코딩[6], 어노테이션 기반 트랜스코딩으로 요약할 수 있다.[4]

HTML 기반의 트랜스코딩 기법은 HTML 스타일 정보를 그룹핑하여 이를 축약된 형태의 정보로 변환하는 방법이다[3]. 많은 양의 정보 전달은 가능하나 스타일에 기반함으로써 관련이 없는 정보들이 그룹핑 되는 문제가 발생할 수 있으며, XML 문서 환경에 적합하지 않다. 문서 요약 및 필터링 기반 트랜스코딩 기법[6]은 시스템 관리자에 의해 HTML 문서의 직접적인 변환을 수행하고 별도의 필터링에 의한 반자동 기법이다. 이 기법은 원본 문서의 내용이 손실됨으로써 완벽한 정보 전달이 어렵고, 원본 문서가 변경됨으로써 저자의 의도가 변경되는 문제점이 있다. 어노테이션 기반 트랜스코딩 기법은 HTML 및 텍스트 등의 문서를 어노테이션을 이용하여 Palm OS 등의 디바이스에 맞게 트랜스코딩하기 위한 프레임워크 (IBM WebSphere)[4]이다. XML과 같은 구조정보를 트랜스코딩하기 어려우며, 트랜스코딩을 위해서는 복잡한 단계의 수작업 어노테이션이 필요한 단점이 있다.

2.2 NewsML

NewsML은 IPTC(International Press Telecommunications Council)[5] 가 정의한 또 다른 XML 언어 체계로 멀티미디어 뉴스의 저작, 저장, 전송을 위한 표준이다. NewsML은 문자, 영상, 사진, 음성 할 것 없이 모든 뉴스 포멧과 미디어가 동등하게 처리된다. 또한 한 건의 뉴스를 다른 뉴스와 자유롭게 결합시키고 뉴스의 일부분을 다른 뉴스에 삽입시키는 등 뉴스를 쪼개서 다루는 기능이 뛰어나다. 그래서 NewsML은 멀티미디어 시대에 맞는 차세대 뉴스 표준이라 불리며 세계의 여러 언론사들이 실제 적용을 실험하고 있다. IPTC에서 발표한 NewsML의 스펙은 DTD와 XML 스키마 형태로 발표되었으며 전체적인 구조는 다음과 같다.



[그림 1] NewsML의 구조 [5]

NewsML의 가장 핵심적인 구성요소는 NewsItem으로 뉴스 컨텐츠와 관리 정보 등 모든 정보를 담는다.

NewsML은 뉴스의 논리적인 구조나 메타데이터를 담고 있을 뿐 뉴스 서비스 최종 단계에서 어떤 형태로 표출될 것인가에 관한 레이아웃이나 디자인에 관한 정보는 전혀 없다. 레이아웃이나 디자인은 XSL과 같은 스타일시트를 적용하여 전혀 별개의 작업으로 이뤄져야 한다.

NewsML은 현재 v1.2로 업데이트 되었으며, 머지 않아 전세계 뉴스 표준으로 자리잡을 것으로 보인다. 특히 XML 웹상에서 널리 보급되고 XML을 지원하는 브라우저가 상용화되면 NewsML을 적용한 뉴스 매체들은 디지털 뉴스 본연의 다양하고 입체적인 서비스를 선보이게 될 것이다.

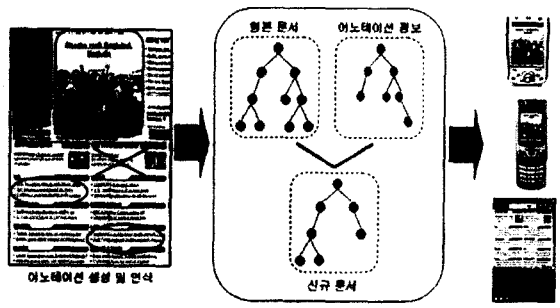
2.3 어노테이션 (Annotation)

어노테이션은 일반적으로 책이나 문서의 중요한 부분에 입력된 부가 정보로 정의되며, 웹 페이지 및 Adobe Acrobat 등의 전자문서 소프트웨어에서 사용하고 있는 기술이다. 어노테이션의 저작 및 인식, 표현 등의 기법은 다양하게 연구되어 왔으며, 여러 분야에서 이용되고 있다[4].

일반적인 어노테이션 기법은 전자문서상에서 사용자가 중요하다고 생각하는 부분에 밑줄이나 특정 기호를 마킹하는 방식으로 이루어 진다. 본 연구에서는 이와 같은 어노테이션 기법들을 사용하여 XML 구조문서 상에서 사용자의 원하는 정보를 추출할 수 있는 수단으로서의 어노테이션 기술을 활용하고자 한다.

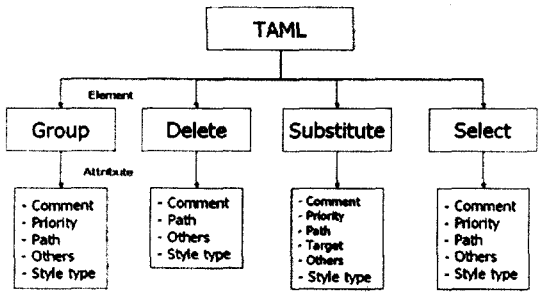
3. 어노테이션 트랜스코딩 프레임워크

본 논문에서 제안하는 트랜스코딩 프레임워크는 크게 3단계로 나뉘어 지며 어노테이션 생성 및 인식, 어노테이션의 구조 정보를 활용한 페이지 생성 및 페이지 맵 구성, 디바이스에 따른 페이지의 변환으로 구성된다. 앞의 두 단계는 원본 문서를 사용자의 목적에 맞게 추출하여 재구성하는 기본 단계이고 디바이스에 따른 실제 변환은 마지막 단계에서 이루어 진다.



[그림2] 어노테이션 트랜스코딩의 개요

어노테이션 생성 및 인식 단계에서는 문서상에 마킹한 어노테이션을 통해 해당 정보를 인식하고 어노테이션 DTD에 유효한 XML 문서를 생성한다. 트랜스코딩 의도를 반영하기 위한 어노테이션 모델은 다음과 같다.

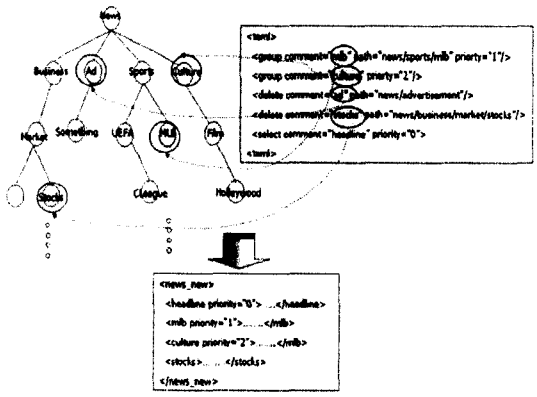


[그림3] 어노테이션 모델

어노테이션은 크게 네 가지 앨리먼트로 구성되어 있으며 각각의 앨리먼트 내에는 앨리먼트를 특징 짓는 속성들로 이루어져 있다. Group 노드는 각 항목들 간의 연관관계를 설정, Delete 노드는 트랜스코딩에 제외시킬 항목을 설정, Select 노드는 개별적인 구조요소 선택, Substitute 노드는 대체요소 지정, Priority 노드는 우선순위를 수동으로 지정하는데 이용한다.

4. 페이지 생성 및 어노테이션 확산

어노테이션을 이용한 페이지 생성 기법은 구조정보를 이용한 앨리먼트 필터링과 사용자 지정 부분에 근거한 어노 테이션 확산으로 구성되어 있다. 구조정보를 이용한 앨리먼트 필터링은 원본문서에 어노테이션 정보를 반영하여 트랜스코딩을 위한 XML 페이지를 생성하는 과정으로 다음과 같다.



[그림4] 앨리먼트 필터링의 예

앨리먼트 필터링의 기본 과정은 [그림4]에서 보는 바와 같이 어노테이션 문서를 파싱하여 원본 문서의 해당 앨리먼를 비교하고, 사용자가 지정한 우선순위 정보를 고려하여 새로운 XML 문서를 생성한다.

어노테이션 확산은 사용자가 어노테이션을 입력할 때 발생하는 모호성을 해결하는 방안으로 XML 문서의 구조정보를 활용하여 어노테이션 되지 않은 노드를 정해진 규칙에 따라 처리한다. 어노테이션 확산 규칙은 크게 세 가지로 선택, 대체, 삭제로 나눌 수 있다.

- 선택

  자식 노드가 없는 경우(Leaf 노드인 경우) 해당 노드의 정보를 그대로 보여준다. 자식 노드가 있고 제목 노드인 경우 자식 노드를 고려하지않는다. 부모 노드가 있고 자식 노드가 있는 경우 해당 노드와 자식 노드의 정보를 보여준다.

- 대체

  디바이스의 성능에 따라 내용이나 그림을 대체하도록 하는 규칙으로, 해당 기사의 사진이나 다양한 그래픽 정보를 저용량의 이미지나 텍스트로 대체한다.

- 삭제

  선택된 앨리먼트의 하위 데이터가 없을 경우 해당 노드를 삭제 한다. 자식노드가 있으나 자식노드 중 선택된 노드가 없으면 삭제한다. 자식노드가 있고 자식 노드 중 선택된 것이 없으면 선택된 노드의 상위 자식노드까지 모두 삭제한다.

5. 결론 및 향후 연구 방향

웹 컨텐츠를 소형 디바이스에 표현하기 위해서는 트랜스코딩의 과정이 필수적이다. 본 논문에서는 시맨틱 어노테이션을 이용하여 사용자의 원하는 정보를 추출할 수 있는 트랜스코딩 프레임 워크를 제안하였다. 트랜스코딩 프레임워크는 크게 세 가지로 어노테이션 생성 및 인식, 페이지의 생성 및 페이지 맵의 구성, 디바이스에 따른 페이지의 변환으로 구성된다. 본 논문에서는 사용자의 어노테이션을 기반으로 XML 문서를 생성하는 두 번째 단계까지 다루었으며, 어노테이션과 페이지 생성을 위하여 어노테이션 모델링, 앨리먼트 필터링 및 어노테이션 확산 기법을 적용하였다.

향후 연구로는 다양한 모바일 기기의 웹 컨텐츠 배포를 위한 표준 프로파일 언어인 CC/PP를 이용하여 기존에 생성된 XML 문서를 해당 디바이스에 맞게 변환하는 기법 등이 요구된다.

참고문헌

[1] E. A. Brewer, R. H. Katz, Y.Chawathe, et al. "A Network Architecture for Heterogeneous Mobile Computing",IEEE Personal Communications, Vol.5, No.5, pp.8-24, October 1998.

[2] 배찬권, "정보통신산업동향 정보통신기기편 제7절 PDA", 정보통신정책연구원, 2001.

[3] T. Bickmore, W. Schilit, "Digestor : Device-Indepen dent Access to the World Wide Web", Computer Networks and ISDN Systems, Vol.29, No.8, pp.1075-1082, 1997.

[4] M. Hori, G.Kondoh, K. Ono, S Hirose and S. Singhal, "Annotation-Based Web Content Transcoding", 9th World Wide Web Conference, 2000.

[5] IPTC, http://www.iptc.org

[6] IBM, Websphere Transcoding Publisher, www-3.ibm.com/software/webservers/transcoding/index.html

[7] T. Bickmore, A.Girgensohn and J.W. Sullivan, "Web Page Filtering and Re-Authoring for Mobile Users", The Computer Journal, Vol.42, No.6, pp.534-546, 1999.

[8] N. Milic-Frayling and R. /sommerer, "SmartView : Flexible Viewing of Web Page Contents", World Wide Web Conference 2002, 2002