사용자 관심사를 고려한 어노테이션 기반의 XML 문서 트랜스코딩

이진상°, 송특섭, 임순범*, 최윤철
연세대학교 컴퓨터과학과
*숙명여자대학교 멀티미디어과학과
{gr20000, teukseob, ycchoy}@rainbow.yonsei.ac.kr
sblim@sookmyung.ac.kr

User Centered XML Document Transcoding Using Semantic Annotation

Jin-Sang Lee°, Teuk-Seob Song, Soon-Bum Lim*, Yoon-Chul Choy
Dept. of Computer Science, Yonsei Univ.
* Dept. of Multimedia Science, Sookmyung Women's Univ.

요 약

기존의 웹 컨텐츠를 휴대폰이나 PDA 등과 같은 개인용 단말기에 표현하기 위해서는 단말기 성능상의 제약(낮은 CPU 성능, 작은 출력 화면, 입출력 방법의 단순함 등)을 고려한 컨텐츠 변환의 과정이 필요하다. 트랜스코딩이란 기존의 웹 컨텐츠를 다양한 단말기의 환경에 따라 적합한 형태로 변환 하는 것을 의미한다. 지금까지의 트랜스코딩에 관한 연구는 단말기의 성능을 고려한 서비스 제공자 중심의 일방적인 컨텐츠 변환 기법으로 사용자 관심사의 반영이 어렵고 서비스 이용의 효율성이 떨어진다. 본 논문에서는 사용자의 관심사를 반영한 어노테이션 기반의 효과적인 XML 문서 트랜스코딩 기법을 제안한다. 제안하는 트랜스코딩 프레임워크는 3단계로 구성되며, 사용자 어노테이션의 생성, 어노테이션을 이용한 원본 문서의 재구성, 단말기 성능을 고려한 XSLT 변환으로 이루어진다.

1. 서론

최근 무선 인터넷 기술의 발전과 PDA나 스마트 폰 등의 다양한 모바일 단말기의 등장으로 웹의 접근 방식이 다양해지고 있다[1]. 그러나 소형 단말기를 통해 기존의 웹 컨텐츠를 이용할 경우, 단말기 성능상의 제약(대역폭, 화면크기, CPU 성능)으로 서비스 이용에 문제가 발생한다. 이러한 문제점을 해결하기 위해 단말기에 적합한 형태로 컨텐츠를 변환하는 트랜스코딩에 대한 연구가 다양하게 이루어지고 있다[4,5,6]. 기존의 트랜스코딩에 관한 연구는 주로 HTML 문서의 레이아웃 정보를 이용한 자동 변환 기법이다[4]. 또한 최근에는 W3C의 Device Independence 워킹그룹[1]을 중심으로 단일 저작된 웹 컨텐츠를 다양한 장비에서 이용하도록 하는 연구가 진행중이다.

지금까지의 트랜스코딩 기법은 주로 단말기의 제약을 고려한 서비스 제공자 중심의 일방적인 컨텐

"본 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음"(KFR-2003-041-D00533)

츠 변환으로, 사용자의 관심사를 반영하는 데에 어려움이 있다. 즉 사용자가 원하지 않는 불필요한 정보들도 전달하게 되므로 서비스 이용의 효율성이 떨어진다.

이에 본 논문에서는 사용자의 관심사를 반영한 어노테이션 기반의 효과적인 XML 문서 트랜스코딩 기법을 제안한다. 제안하는 트랜스코딩 프레임워크는 3단계로 구성되며, 사용자 어노테이션의 생성, 어노테이션을 이용한 원본 문서의 재구성, 단말기 성능을 고려한 XSLT 변환으로 이루어진다.

2. 관련연구

2.1 W3C Device Independence [1]

Device Independence 워킹 그룹의 목표는 각각의 디바이스에 맞게 제공되는 다양한 웹 접근 매커니즘을 단일화하여 웹 서비스의 노력과 비용을 최소화 하는 것이다. 즉, W3C의 주요 목표중의 하나인 월드와이드웹의 광역 접근성(Universal Access)을 이루는데 있다. Device Independence 워킹그룹에서 제안하는 컨텐츠 변환에 관한 주요 기술에는 Delivery Context, CC/PP(Composit Capability / Preference Profiles), Style Sheet/Transform이 있다.

- Delivery Context

  장비에 독립적인 웹을 구현하기 위해 필요한 의미정보로 사용자 디바이스의 정보, 웹 접근 매커니즘, 서비스 이용자의 선호도 등의 내용을 포함한다.

- CC/PP

  Delivery Context를 위한 수단으로 제정된 CC/PP는 광범위한 기기에 웹 컨텐츠 배포를 용이하게 하는 RDF(Resource Description Framework) 기반의 W3C 표준 프로파일 언어이다. CC/PP 프로파일은 컴포넌트들로 구성되며 컴포넌트는 하나 이상의 속성으로 이루어진다. 컴포넌트와 속성 값들을 이용하여 화면 사이즈, CPU, Operating System, 장비가 지원하는 HTML 버전 등 디바이스에 관한 포괄적인 정보 기술이 가능하다.

- Style sheet/Transform [2]

  Style Sheet/Transform 기술은 CC/PP와 함께 Device Independence에 있어서 가장 중요한 기술중의 하나이다. XSLT를 이용하여 XML 문서를 다른 단말기에서 이용 가능한 마크업(XHTML, cHTML, WML 등)으로 변환한다. XSLT 변환 기술은 서버 측 변환과 클라이언트 측 변환이 가능한데, 일반적으로 웹에서 제공하는 방식은 서버측 변환이며, 사용자 요청에 따라 문서의 일부 또는 전부를 변환하여 전송한다.

2.2 기존의 트랜스코딩 기법

지금까지의 트랜스코딩에 관한 연구는 HTML 문서의 테이블 구조 및 태그요소 분석을 통한 레이아웃 기반의 트랜스코딩[3]과 단말기의 프로파일 정보를 이용한 CC/PP 기반의 트랜스코딩[4]으로 분류할 수 있다. 전자는 단말기에 관한 고려사항이 크게 필요하지 않던 HTML 중심의 초기 트랜스코딩 기법이며, 후자는 모바일 기기의 다양화와 웹 접근 매커니즘이 다양해지고 있는 최근에 등장한 트랜스코딩 프레임워크이다. CC/PP 기반의 트랜스코딩은 W3C의 Device Independence 워킹그룹을 중심으로 다양한 연구가 진행되고 있으며, 단일 저작된 웹 컨텐츠를 다양한 단말기에 효과적으로 표현하는데 목적이 있다.

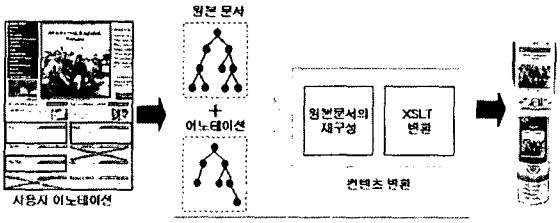
2.2 어노테이션 [8]

어노테이션은 일반적으로 '책이나 문서의 중요한 부분에 입력된 부가 정보'로 정의되며, 웹 페이지나 전자책 등에서 널리 사용되고 있는 기술이다. 일반적인 어노테이션 기법은 전자문서상에서 사용자가 중요하다고 생각하는 부분에 밑줄이나 특정 기호를 표시하는 방식으로 이루어진다. 본 연구에서는 이와 같은 어노테이션 기법을 적용하여 트랜스코딩에 필요한 컨텐츠 변환 규칙을 생성한다.

3. 어노테이션을 이용한 XML 문서 트랜스코딩

본 논문에서는 주제별 계층 구조가 명확한 웹 뉴스를 대상으로 하여, 사용자 중심의 트랜스코딩 프레임워크를 제안한다. 트랜스코딩 프레임워크는 3단

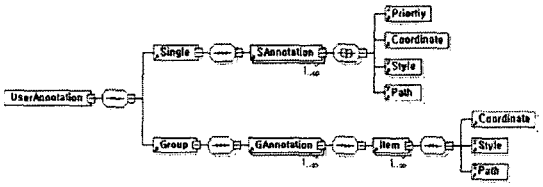
계로 구성되는데, 이는 사용자 어노테이션의 생성, 어노테이션을 이용한 원본 문서의 재구성, 단말기 성능을 고려한 XSLT 변환이다.



[그림 1] 트랜스코딩 프레임워크

3.1 사용자 어노테이션

사용자 관심사를 고려한 의미 있는 트랜스코딩을 위해서는 컨텐츠 변환 규칙을 기술한 메타 정보가 필요하다. 메타 정보 기술을 위해 정의한 사용자 어노테이션의 구조는 [그림 2]와 같다.

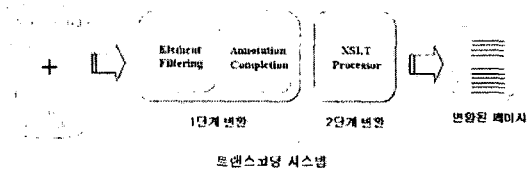


[그림 2] 사용자 어노테이션의 구조

사용자 어노테이션은 단일 어노테이션과 그룹 어노테이션의 집합으로 구성된다. 단일 어노테이션은 한 가지 주제에 대한 선택 및 삭제를 의미하며, 그룹 어노테이션은 두 가지 이상의 주제에 대한 연관성을 의미한다. 각각의 어노테이션은 아이디, 중요도, 화면상의 좌표, 어노테이션 유형, 원본 문서상의 경로 값을 갖는다. 특정 항목에 대한 중요도는 0~10 사이의 값으로, 0은 삭제를 의미하며 10에 가까울수록 중요도는 높아진다. 어노테이션의 대상이 되는 문서는 웹 뉴스의 인덱스 페이지이다. 일반적으로 뉴스 사이트의 인덱스 페이지는 헤드라인, 중요기사 목록, 주제별 최신기사 등 모든 정보를 요약해 놓은 페이지라고 할 수 있다. 따라서 사용자는 인덱스 페이지 상에서 자신이 원하는 정보를 선택할 수 있다. 사용자가 생성한 어노테이션 정보는 XML 문서로 저장된다.

3.2 어노테이션을 이용한 원본 문서의 재구성

사용자 어노테이션을 이용한 트랜스코딩의 과정은 엘리먼트 필터링과 미지정 부분의 처리를 위한 어노테이션의 완성으로 구성된다. 엘리먼트 필터링을 통해 선택, 삭제, 그룹, 중요도 등의 어노테이션 정보를 원본문서에 반영하고, 필터링 된 결과에 어노테이션 완성 규칙을 적용하여 사용자가 지정하지 않은 부분을 처리한다.



[그림 3] 어노테이션을 이용한 문서의 변환 과정

엘리먼트 필터링의 과정은 트랜스코딩을 위한 전처리 단계로 사용자 어노테이션 정보를 원본문서의 트리 구조에 반영한다. 파싱된 어노테이션 정보는 레코드 형태의 데이터 구조로 변환되어 순차적으로 처리 된다. 엘리먼트 필터링은 사용자 어노테이션의 개수만큼 반복되는데, 사용자가 선택한 항목에 해당하는 트리상의 노드를 찾아 어노테이션 유형과 중요도 값을 부여한다.

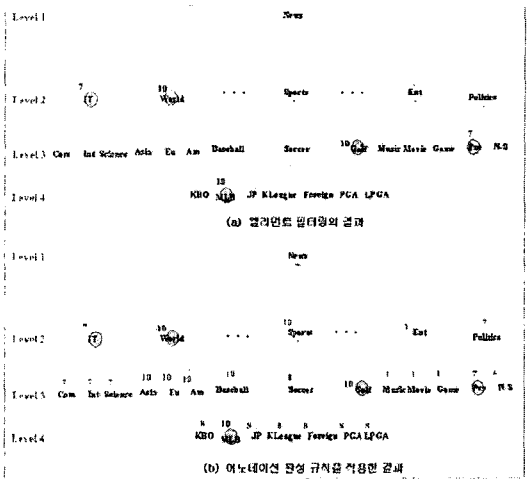
어노테이션 되지 않은 사용자 미지정 부분의 처리를 위해서는 엘리먼트 필터링 결과를 이용한 어노테이션 완성 규칙이 필요하다. 이는 사용자가 지정한 노드와 중요도 값에 근거하여 인접 노드의 중요도 값을 결정한다. 다음은 선택, 삭제, 그룹 및 지정된 노드의 깊이를 고려한 어노테이션 완성 규칙이다.

- 규칙1 :주제별 최상위 노드(레벨=2)에 대한 선택, 삭제, 그룹의 경우, 자식 노드의 중요도는 부모 노드의 중요도를 그대로 계승 받는다.
- 규칙2 :선택된 노드의 레벨이 3이고 자식 노드를 갖지 않은 경우, 부모 노드에는 중요도 값을 그대로 전달하고 선택되지 않은 형제 노드의 중요

도는 p(1≤p≤2)만큼 감소시킨다.

- 규칙3 : 선택된 노드의 레벨이 3이고 자식 노드를 갖는 경우, 부모 노드와 자식 노드에는 중요도 값을 그대로 전달하고 선택되지 않은 형제 노드의 중요도는 p만큼 감소시킨다.
- 규칙4 : 선택된 노드의 레벨이 4일 경우, 부모와 조상 노드에는 중요도 값을 그대로 전달하고 형제 노드의 중요도는 p만큼 감소시킨다. 또한 선택되지 않은 나머지 노드들은 모두 p 감소된 중요도 값을 갖는다.
- 규칙5 : 특정 주제에 대해 선택된 노드가 전혀 없는 경우, 모든 노드의 중요도 값은 1을 갖는다.
- 규칙6 : 레벨 2나 레벨 3의 노드들 간의 중요도가 같을 경우, (선택된 노드의 깊이 * 중요도)의 합이 큰 순서대로 노드의 우선순위가 결정된다.
-

[그림 4]는 어노테이션 완성 규칙을 적용한 예이다.



[그림 4] 어노테이션 완성의 예

3.3 단말기 성능을 고려한 XSLT 변환

원본 문서의 재구성 과정을 통해 노드들 간의 우선순위가 결정되면, 우선순위 정보를 반영하여 XSLT 변환 스크립트를 생성하고 단말기에 적합한 형태로 최종 변환한다. XSLT 스크립트는 일정한 출력 형식을 갖는 템플릿 형태로 제공되며, 우선순위 정보에 따라 출력 내용이 결정된다.

5. 결론

본 논문에서는 사용자의 관심사를 반영한 XML 문서 트랜스코딩 기법을 제안하였다. 이를 위해 웹 뉴스를 대상으로 하는 사용자 어노테이션 모델을 정의하였으며, 어노테이션 정보를 이용한 원본 문서 변환 기법을 제시하였다. 원본문서의 변환 과정은 엘리먼트 필터링과 미지정 부분 처리를 위한 어노테이션 완성으로 구성되며, XSLT 프로세서를 거쳐 단말기에 적합한 형태로 최종 변환된다.

향후 연구로는 다양한 모바일 기기의 웹 컨텐츠 배포를 위한 표준 프로파일 언어인 CC/PP를 이용하여 XML 문서를 해당 디바이스에 맞게 변환할 수 있는 XSLT 스크립트의 동적 생성 기법이 요구된다.

[참고문헌]

[1] W3C Device Independent Working group, http://www.w3.org/2001/di

[2] XSL Transformations (XSLT) Version 1.0 W3C Recommendation 16 November 1999, http://www.w3c.org/TR/xslt

[3] 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.

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

[5] T. Lemoluma and N. Layaida, "NAC: A Basic Core for the Adaptation and Negotiation of Multimedia Services", OPERA Project, INRIA, September 2001.

[6] Ilia A. Ovsiannikov,"Annotation Technology, International Journal of Human-Computer Studies," Vol.50, no.4, pp.329-362, 1999.

[7] M. Butler, F. Giannetti, R. Gimson and T. Wiley "Device Independence and the Web", IEEE Internet Computing, Vol:6, Issue:5, pp.81-86, September/October, 2002.

[8] Ilia A. Ovsiannikov, "Annotation Technology, International Journal of Human-Computer Studies," Vol.50, no.4, pp.329-362, 1999.