무선 인터넷 서비스를 위한 XML기반 컨텐츠 변환기 설계

김미영*, 정헌, 강병욱
영남대학교 컴퓨터공학과
e-mail : mykim13@yumail.ac.kr

A Design of the XML-based Contents Converter for Wireless Internet Services

Mi-Young Kim*, Heon Jeong, Byung-Wook Kang
Dept. of Computer Engineering, Yeungnam University

요 약

현재 무선 인터넷 컨텐츠는 유선 컨텐츠와 호환이 되지 않으며 다양한 종류의 무선 마크업 언어들로 구성되어 있어 무선 컨텐츠를 재구축해야 하는 문제점이 있다. 본 논문에서는 이를 해결하기 위해 유선 컨텐츠를 재사용하는 방안으로 XML 기반의 유무선 컨텐츠 변환기를 설계한다. 컨텐츠 변환의 중간 표준포맷으로 XML 문서를 생성하므로 새로운 언어로의 변환을 고려할 때 해당 프리젠테이션 변환 모듈만 추가시키면 확장 가능하므로 유지보수가 쉽고 효율적인 변환이 가능하다. 따라서 새로운 표준과 기술 변화에 대응이 용이하며 또한 유선 컨텐츠의 재사용으로 인해 개발비용이 절감되고 개발기간이 단축되므로 빠른 무선 인터넷 시장 진입이 가능해진다.

1. 서론

최근 이동통신 산업의 급속한 발전에 따라 무선 인터넷 시장의 성장이 가속화되고 있으며 이로 인해 무선 컨텐츠 개발에 대한 수요가 증가하고 있다. 그러나 작은 메모리, 낮은 네트워크 대역폭, 소형 브라우저 등과 같이 제한된 자원을 가진 무선 단말기에 기존의 유선 컨텐츠 기술 언어인 HTML로 표현하기엔 많은 제약이 따른다. 무선 컨텐츠 개발의 또 다른 장애요소로 무선 마크업 언어의 다양성을 들 수 있다. 현재 무선 컨텐츠 포맷은 WAP 기반의 WML, HDML과 ME 기반의 mHTML, cHTML등이 혼재되어 있어서 다양한 클라이언트 브라우저 환경에 맞는 컨텐츠를 각각 구축해야 하는 문제점을 가지고 있으며, 재구축에 따른 개발 기간과 비용 또한 낭비되고 있다[2][6]. 따라서, 유무선 컨텐츠간의 호환이나 통합 필요성이 대두되고 있다.

본 논문에서는 기존의 HTML기반 유선 컨텐츠를 재사용하는 방안으로 XML 기반의 유무선 컨텐츠 변환기를 설계한다. 컨텐츠 변환의 중간 표준 포맷으로 이용할 XML(eXtensible Markup Language)은 W3C에서 제정한 "국제 표준 전자문서의 메타언어"로서 데이터와 프리젠테이션이 분리되어 컨텐츠의 동적표현이 가능하며 정보 교환과 관리에 효율적이다[1].

본 논문의 구성은 다음과 같다. 2장에서는 컨텐츠 변환을 중심으로 한 기존의 관련연구를 살펴보고 3장에서는 변환기의 전체적인 구성과 각각의 모듈에 대한 특징 및 기능에 대해 설명하고 마지막으로 4장에서는 결론 및 향후 연구 방향을 제시한다.

2. 관련연구

현재 무선 마크업 언어의 다양성으로 인해 각각의 형태에 맞는 컨텐츠를 모두 구축해 주어야 하는데 이것은 비효율적이다. 또한 아무리 우수한 언어가 개발되었다 하더라도 기존의 언어체계를 무시하는 것 역시 비현실적이다. 따라서 기존의 문서를 재

사용하는 방안이 절실히 요구되어진다.

2.1 컨텐츠 변환

컨텐츠 변환방법으로 Automated Converting 방법과 Configurable Converting 방법이 있다. 미리 설정한 표준 변환 규칙에 따라 자동 변환되는 Automated Converting 방법에는 레이아웃과 태그를 동시에 변환시키는 HTML Reformatting 방법과 태그 변환만을 지원하는 Tag Converting 방법이 있다. 그리고 개발자가 변환 규칙을 설정할 수 있는 Configurable Converting 방법의 대표적인 예는 변환 영역을 추출하여 변환하는 Web Clipping 방법이 있다[7].

현재 연구되어 온 컨텐츠 변환기의 형태로는 HTML 필터를 이용한 무선 언어로의 매핑 방식, HTML 파일을 WML이나 HDML 등의 파일로 각각 변환하는 방식, 무선 언어간의 변환 방식이 대부분이다[6]. 이러한 방식은 변환될 언어에 대한 모든 모듈을 포함해야 하므로 복잡하고 유지보수가 어렵다. 그리고 일대일 태그 매핑과 필터링 방식은 변환의 정확성이 떨어지고 컨텐츠의 손실을 가져오므로 실제 변환된 결과의 QoS를 보장하기 어렵다.

2.2 XML(eXtensible Markup Language)

W3C에서 HTML의 한계를 극복하기 위해 인터넷 상의 문서를 다양하게 표현할 수 있는 XML을 새로이 정의하였다.

XML은 단순히 문서의 내용을 기술하는 표준뿐만 아니라 컨텐츠를 포함할 수도 있으며 특정 컨텐츠를 표현하는 태그와 속성을 설명하는 DTD (Document Type Definition)를 정의할 수도 있다. 이러한 XML의 특성은 데이터를 다양한 형태로 표현하고 변환 가능하게 할 뿐 아니라, 이기종 간의 시스템 환경에서도 문서를 효과적으로 상호교환할 수 있다는 장점을 갖고 있다[1][5].

그리고, 서버와 클라이언트 사이에서 XML 문서의 형태를 변환하는 메커니즘을 Transcoding이라고 부른다.

3. Contents Converter의 설계

Converter의 적용기준과 전체적인 구성, 그리고 각각의 모듈에 대한 특징 및 기능에 대해 설명한다.

3.1 Converter 적용기준

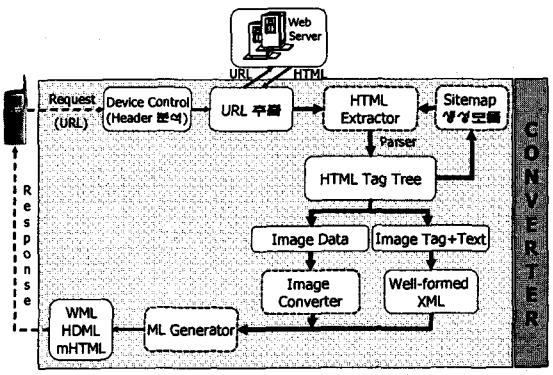
다양한 유선 컨텐츠 중 어떤 유형의 컨텐츠를 Converter를 통해 무선화 할 것인지 결정하기 위해 적용될 컨텐츠에 대한 기준이 있어야 한다. 사용자의 최소 입력을 통해 최대의 정보를 추출할 수 있는 컨텐츠가 변환의 대상이 된다. 다음의 기준에서 벗어나는 컨텐츠는 새로 무선 컨텐츠를 재구축하는 것이 더 바람직할 것이다.

- 작은 양의 텍스트 정보를 유선 컨텐츠에서 압축해서 추출할 수 있는가?
- 적은 단계의 링크로 완결된 정보 제공이 가능한가?
- 컨텐츠의 갱신 빈도가 높은가?

3.2 시스템의 전체 구성도

본 시스템은 Client, Converter, Web Server로 구성되어 있다. Client의 요구(URL)가 발생하면 해당 Web Server로부터 가져온 컨텐츠를 Converter에서 자동 인식된 단말기 타입에 맞는 마크업 언어로 변환해주는 시스템이다.

Server로부터 추출된 HTML 문서는 HTML Extractor를 거치면서 파싱(Tidy)되어 HTML Tag Tree를 생성한다. 이 Tree에서 분리된 이미지 데이터는 Image Converter를 통해 변환되어 무선 마크업 언어 변환 모듈인 ML Generator에서 재조합된다. 또한 이미지 태그를 포함한 나머지 태그들로 Well-formed XML 문서를 생성한다. 생성된 XML 문서에 무선 마크업 언어 스타일시트(XSL)를 적용한 후 무선 단말기에 적합한 문서로 변환하여 Client에게 응답(Response)한다. <그림 3.1>은 Contents Converter를 포함한 전체 시스템 구성도를 나타낸다.



<그림 3.1> 전체 시스템 구성도

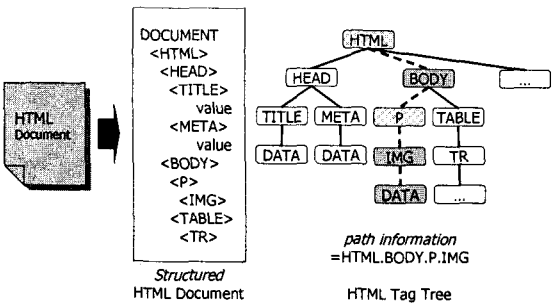
3.2.1 HTML Extractor

HTML 문서를 무선 단말기에 맞게 지원하려면 우선 무선 단말기의 제한된 환경을 고려한 HTML 문서로 재구성해야 한다. 컨텐츠의 질(quality)를 고려하지 않고 변환하여 전송하게 되면 과도한 양의 컨텐츠가 전달되어 사용자의 이용에 오히려 방해요소가 된다.

Client가 요구한 HTML 문서를 로드한 후 문서가 프레임으로 되어있는지, 사이트의 전체 구성 정보를 알 수 있는 Sitemap 문서가 있는지 판별한다. 만약 Sitemap 문서가 존재하지 않는다면 Sitemap 자동생성 모듈에서 생성하여 모든 변환의 초기 문서로 사용한다. 그리고 무선 마크업에서 지원하지 않는 스타일 관련 내용이나 태그, 애플릿, 자바 스크립트, 주석, 광고, 외부링크 등을 필터링 시킨다.

3.2.2 HTML Tag Tree

HTML 문서를 파싱하여 HTML Tag Tree를 생성하고 HTML 문서내에 존재하는 태그 오류를 찾아 수정하는 모듈이 필요하다.



<그림 3.2> HTML Tag Tree

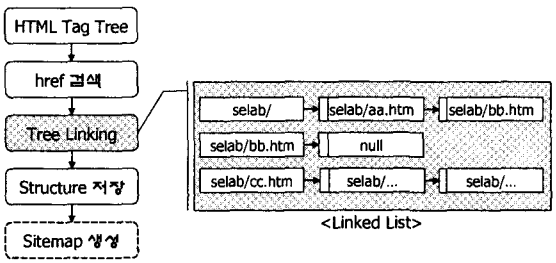
Top-Down 방식으로 문서 트리의 Root에서 시작하여 깊이 우선 탐색(DFS:Depth First Search)을 시행한다. 각 노드의 자식노드를 차례로 파싱하여 태그와 데이터를 분리시켜 나간다. 파싱 과정을 거친 태그들은 계층적으로 표현이 되며 트리구조의 말단에는 추출될 실제 데이터들이 위치하고 있다. 예를 들어, <그림 3.2>의 Tag Tree에서 <IMG> 태그 밑에 위치한 데이터를 추출해 Image Converter로 넘겨주게 된다.

이런 과정을 거쳐 Well-formed HTML 문서가 생성되면 구문 분석을 통해 DOM(Document Object Model) 형태의 XML 문서로 변환해준다.

3.2.3 Sitemap 생성모듈

HTML Tag Tree로 형성된 구조화된 문서에서 링크 정보(href)를 검색하여 링크된 문서들을 Linked List로 저장해 가면서 Sitemap을 구성해 나간다. 링크된 문서 간에 사이클이 형성될 수 있기 때문에 이전에 방문한 문서의 경우 하위 노드로 연결하고 다시 방문하지 않도록 한다.

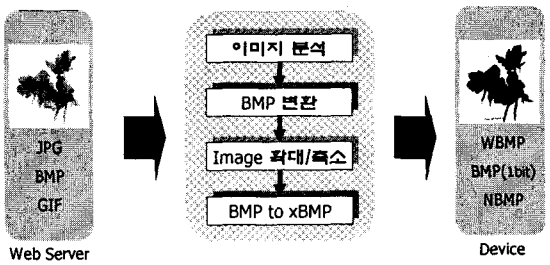
HTML Extractor에서 추출되거나 자동생성 모듈에서 생성된 Sitemap 문서는 Converter를 거치면서 Client 단말기의 시작 페이지로 제공된다.



<그림 3.3> Sitemap 자동생성

3.2.4 Image Converter

JPG, BMP, GIF와 같은 이미지 형태를 무선 단말기에서 지원 가능한 포맷으로 변환하며, 색상 수, 브라우저 크기, 메모리 용량 등을 고려하여 이미지의 크기를 조절하거나 링크로 변환하는 모듈이다.



<그림 3.4> Image Converter

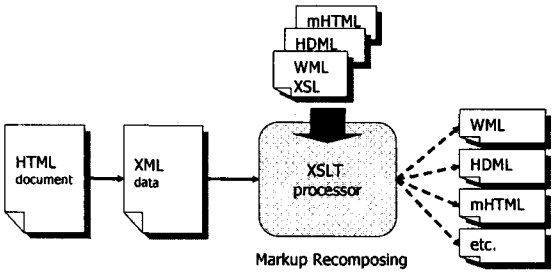
입력되는 이미지 포맷을 분석하여 압축이 되지 않은 형태인 BMP 파일로 변환한다. 각 이미지 포맷을 BMP로 통일함으로써 xBMP를 생성할 때 중복되는 과정을 생략할 수 있고, GIF나 JPG가 갖는 이미지 압축에 대한 문제를 해결할 수 있다. 변환된 이미지의 크기를 조정한 후 xBMP로 변환한다.

반면 이미지 태그 처리는 <IMG> 태그에 ALT 속성값이 없다면 이미지의 확장자를 xBMP 형태로

변환한 후 부여된 이미지 이름으로 저장하고 ALT 속성값이 있거나 변환이 불가능한 이미지는 텍스트 링크로 대체하거나 삭제한다.

3.2.5 ML Generator

XML 데이터와 Style Manager를 통해 생성된 무선 마크업 XSL이 XSLT(XSL Transformation) processor를 통해 결합하여 각각의 단말기에서 요구하는 마크업 언어로 변환되는 모듈이다.



<그림 3.5> ML Generator

HTML은 굵은 글씨를 위한 <B>나 가운데 정렬을 위한 <CENTER>와 같은 표현적인 요소들을 추가하면서 발전해 왔으나, 무선 단말기에서 그러한 요소들을 처리하기에는 용량이나 성능 면에서 한계가 있다. 이를 위해 W3C에서 문서의 내용과 표현을 분리시키기 위한 XSL이라는 것을 정의하였다. 즉, 스타일을 기술하고 있는 부분만 바꾸어주면 원본 XML 문서를 수정하지 않고도 다양한 스타일로 문서를 표현할 수 있다.

기반 데이터는 데이터베이스에 저장하고 사이트에 접근하고자 하는 Client 단말기 유형에 맞는 XSL(eXtensible Stylesheet Languge)을 결합함으로써 단말기에 적합한 컨텐츠를 제공해준다.

4. 결론 및 향후과제

무선 인터넷 기술의 발전과 그에 따른 사용증대는 무선 컨텐츠 이용자들에게 편리성 및 유용성을 제공하기 위한 유무선 통합의 새로운 서비스 개발을 촉진시켜 왔다. 이것은 기존 PC환경에서 제공되었던 유선 컨텐츠를 재사용하여 무선 컨텐츠로 활용하고자 하는 것이다.

본 논문에서는 이를 근거로 한 XML 기반의 컨텐츠 변환 시스템을 설계한다. 이 시스템은 기존의 유선 컨텐츠를 무선 컨텐츠로 자동 변환시킴과 동시에 다양한 단말기에 구애받지 않고 한번 생성된 XML 문서를 이용하여 동적으로 해당 마크업 언어를 생성한다. 컨텐츠 변환의 중간 표준포맷으로 XML 문서를 이용하므로 새로운 언어로의 변환을 고려할 때 해당 프리젠테이션 변환 모듈만 추가시키면 확장 가능하므로 유지보수가 쉽고 효율적인 변환이 가능하다. 따라서 새로운 표준과 기술 변화에 대응이 용이하며 또한 유선 컨텐츠의 재사용으로 인해 개발비용이 절감되고 개발기간이 단축되므로 빠른 무선 인터넷 시장 진입이 가능해진다.

HTML Extractor를 통해 불필요한 정보가 제거되어 컨텐츠의 질이 향상되고, Sitemap 생성으로 네비게이션이 최적화된다. 또한 이미지 변환으로 인해 시각적 효과가 커져 정보 전달력이 증대된다.

향후 과제로서 Converting 과정의 오버헤드로 인해 응답시간이 증가하는 것을 고려하여 기존의 컨텐츠를 보존하면서 전송 용량을 최소화시키는 방안에 대한 연구를 필요로 한다.

참고문헌

[1] W3C, "Extensible Markup Language(XML) Version1.0 (Second Edition)", http://www.w3.org/TR/REC-xml, Oct. 6, 2000.

[2] WAP Forum, "Wireless Markup Language", http://www.wapforum.org, 2002.

[3] H. Ouahid, A. Karmouch, "Converting Web Pages into Well-formed XML Documents", Proceedings of the 1999 IEEE International Conference on Communicatios, Vol.1, pp.676-680, June. 6, 1999.

[4] Marcin Metter, Dr Robert Colomb, "WAP enabling existing HTML applications", User Interface Conference, pp.49-57, Feb. 3, 2000.

[5] 이귀남, "무선 인터넷을 위한 마크업 언어 변환 엔진의 구현", 원광대학교 대학원 컴퓨터공학과 학위논문집, 2002.

[6] 이미경, "XML기반의 유무선 인터넷 문서 변환 시스템의 설계", 정보과학회 논문지, Vol. 28, No. 2, 2001.

[7] 이승진, "확장성 있는 웹 서비스를 위한 무선 응용 프로토콜 기반의 HTML Filter 구현", 정보과학회 봄학술발표논문집(A), pp.391-393, 2001.

[8] 조수선, "모바일 웹 서비스를 위한 컨텐츠 재작성 기술", 인터넷 정보학회 논문지, Vol. 3, No. 5, 2002.