웹 문서 재구조화를 위한 XML 기반의 대화식 코드 변환 시스템 설계

송동리° 황인준

아주대학교 정보통신전문대학원 정보통신공학과

{harley, ehwang}@madang.ajou.ac.kr

XML-based Interactive Transcoding System for Reconstructing Web Pages

Dong-rhee Song° Een-jun Hwang

Graduate School of Information and Communication, Ajou University

요 약

최근에 웹을 통한 기업 이미지의 부각과 기업 활동의 홍보는 기업 경영에 중요한 수단이 되고 있다. 이를 위해, 웹사이트는 사용자에게 다양한 정보와 사용의 편리함을 제공하여야 하지만 아직 많은 웹사이트들이 비규칙적으로 구성되어 있기 때문에 사용자에게 편리한 정보 전달을 제공하지 못하고 있다. 이러한 문제를 해결하기 위한 방법 중 하나는 웹 사이트 내의 문서들간의 공통적인 구성 요소를 알아내고 문서내의 정보를 중심으로 재구성하는 것이다. 본 논문에서는 XML을 이용한 문서의 재구성과 사이트를 구성하는 정보의 분류를 통하여 원하는 유형의 정보를 추출해 내는 대화식 코드 변환 시스템을 제안한다.

1. 서론

정보의 대량 생산은 급격한 정보 변화를 초래하였고, 그 결과로 많은 웹 페이지들이 구조를 빈번하게 바꾸고 있다. 게다가 대량의 정보를 표현하기 위해 웹 문서의 구조는 점점 더 복잡하여 지고 있다. 이것은 웹의 구조를 모르는 사용자에게는 원하는 정보를 접근하는 데 상당한 장애 요소가 될 수 있다. 사용자는 단순히 문서의 시각적 구성(layout)이나 배열을 따라 원하는 정보에 접근한다. 예를 들어 신문사 홈페이지의 경우, 특정한 정보를 찾기 위해 다양한 종류의 기사와 많은 광고 사이를 오가며 시간을 소비해야 하는 경우가 발생할 수 있다. 이러한 문제를 해결할 수 있는 방법 중 하나가 문서의 재구조화이다. 일반적으로 웹 문서의 제작에 많이 쓰이는 HTML(Hypertext Markup Language)은 평면적이고 비구조적이기 때문에 정보 추출과 재배열에 많은 어려움이 있다[2], [3], [6]. 본 논문에서는 웹 문서가 갖는 문제점을 해결하기 위한 방안으로 문서 특성에 따른 재구조화를 위해 구조적 언어인 XML(eXtensible Markup Language)[7] 기반의 대화식 코드 변환 시스템을 제안한다. 코드 변환은 크게 두 과정으로 나뉘는데 첫 번째 과정에서는 웹 페이지의 정보들에 대해 각각을 같은 유형의 최대 크기 구역들로 나누고, 그 의미에 따라 내용과 역할을 XML로 명시한다. 두 번째 과정에서는 이러한 명세를 기반으로 사용자의 의도에 따라 정보를 재구성해서 사용자에게 전달해 주게 된다. 본 논문의 구성은 다음과 같다. 2절에서는 본 시스템의 코드 변환 배경 기술과 관련 연구에 대해 알아보고 3절에서는 구역화 된 정보와 이에 따른 XML의 필요성을 통해 문서의 재구성을 설명한다. 4절에서는 재구조화를 위한 XML 기반의 대화식 코드 변환 방법과 이를 위한 전체적인 시스템 구조에 대해 논하며 5절에서는 향후 활용 방안 및 계획 등을 제시하며 결론을 맺는다.

2. 관련 연구

주석 기반의 웹 내용 코드 변환 시스템[1]은 작은 화면을 지원하는 기기들 상에 웹 정보를 제공하기 위한 방법을 다루고 있다. 그들은 웹 페이지 정보를 조각내어 그곳에 외부 주석을 붙이는 방법으로 HTML 문서를 코드 변환한다. 이 시스템이 시각적으로 구역화된 정보를 인식하는 점은 비슷하지만 PDA와 같은 작은 화면 기기를 위한 것으로서 화면 크기, 메모리 크기 등과 같은 기기의 물리적인 특성과 성능 부분을 고려해야 하는 반면 본 논문은 이와는 상관없이 각 구역의 명세를 제공하여 사용자의 의도에 따라 각 구역들로 구성된 페이지의 구조를 변환한다.

전통적인 웹 구조는 일 차원적인 평면적 구조만 고려되어 왔다. 이에 의미론적 코드 변환[2]은 외부 주석을 이용해서 다 계층으로 구성된 삼 차원적인 구조로의 확장을 위한 쉽고 간단한 방법을 제공한다. 따라서 정보의 구조화에 따른 추출과 재배열이 가능 해지지만 이 방법은 [1]의 기본원리에서 주어진 것에 대한 일반적인 접근이다.

또 다른 방법으로는 웹 사이트의 시각적인 정보는 무시하고 텍스트만으로 정보를 재구성하며 폰트의 크기나 색깔 등의 다양한 효과를 사용하여 해당 정보를 사용자에게 전달한다[4]. 하지만 이 방법은 전적으로 서버쪽에서 일을 처리하게 되므로, 클라이언트쪽의 사용자 환경이 고려된 본 논문과 구별되며, 목적 또한 웹 페이지의 재구성을 위한 것이 아니다.

3. XML을 이용한 문서의 재구성

3.1 구역화 된 정보

웹 제작자들은 가능한 한 비슷한 유형의 정보들에 대해서 동일한 시간적 공간적 특성을 부여하려는 경향이 있으며 이를 위해 테이블이나 이미지, 배경색, 폼과 같은

다양한 효과를 사용한다. 동일하게 구성된 같은 유형의 정보가 포함된 페이지의 각 부분을 구역(region)이라 부른다. 시각적 공간적으로 구역화 된 정보들은 시작과 끝을 가지며, 대개 주 내용, 광고, 차례 등으로 구분이 된다. 하지만 이렇게 제작된 웹 페이지는 정보의 순서와는 상관없는 다양한 시각적 구성을 가지므로 웹 페이지를 접하는 사용자들에게 의도와는 달리 혼란을 초래할 수 있다. 따라서 각 구역들의 경계를 파악할 수 있게 구역에 포함된 정보의 의미를 미리 정의하여 사용자에게 알려준다면 사용자는 원하는 정보만으로 웹 페이지의 정보를 재구성 할 수가 있다.

3.2 XML의 적용

재구조화란 문서의 평면적이고 비구조적인 정보에서 문서를 이루는 구성 요소들의 내용과 역할을 분석하여 이를 토대로 문서의 내용을 구조적으로 재배열하는 과정을 말한다. 그리고 XML은 이러한 구조적인 정보를 표현하는데 적합한 언어이다. 웹 문서제작에 있어 제작자들은 시각적인 효과나 공간 활용의 극대화를 위해 버튼과 같은 작은 이미지를 자주 사용하는데 이는 웹 페이지의 구성을 재배열하는 데에 많은 어려움을 초래한다. 한 예로 "사용자 등록"이란 글이 담긴 이미지가 있다고 가정할 때 이 이미지가 속하는 구역의 역할을 시스템이 스스로 예측하기는 어렵다. 따라서 구조의 재배열 시 XML을 이용하여 이미지에 "등록"이란 의미 부여를 한다면 자료의 구분에 따른 관리가 명확해지며, 사용자의 의도에 따른 페이지의 재구성 또한 가능해진다. 즉 구역의 명세가 선택 되었을 때 그와 일치하는 페이지의 XML 요소 값을 찾는다면 원하는 정보의 구역을 찾을 수 있고, 문서의 재구성이 가능하다.

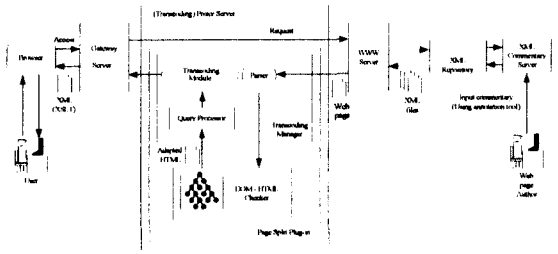

그림 1. 코드 변환 시스템 구조도

4. XML 기반의 대화식 코드 변환 시스템

4.1 시스템의 구조

그림 1은 코드 변환 시스템의 전체적인 구조를 보여준다. 전체 시스템은 크게 코드 변환 관리자(transcoding manager), XML 저장소(XML repository), DOM-HTML 검사기(DOM-HTML checker)로 구성되어 있다. 코드 변환 관리자는 HTML 문서의 코드 변환을 담당하는 구성 요소이며 사용자는 page split plug-in 기능이 지원되는 프록시 서버에 게이트웨이를 통해서 접근한다. 코드 변환 관리자가 웹 서버로부터 접근하고자 하는 HTML 문서를 받으면 이 문서에서 URL(Uniform Resource Locator)을 얻는 후 일치하는 URL을 XML 저장소로부터 찾는다. 코드 변환 관리자는 저장소로부터 받은 이 XML 문서를 기반으로 요청한 HTML 문서를 파싱한다. 일반적으로 하나의 XML 파일은 하나의 URL을 가지는 게 원칙이나 페이지가 서로 유사한 구조를 가지는 형태이면 페이지 사이에 서로 공유할 수 있다. 이는 DOM-HTML 검사기에서 그래프 형태에 의한 DOM tree 유사도 계산[6]에 의해 다른 페이지에서도 사용 가능한지를 판단하게 된다. 그리고 질의 처리기 (query processor)를 통해 사용자의 의도대로 코드 변환 모듈(transcoding module)에서 코드 변환을 하면 웹 페이지의 재구성이 이루어지게 된다. XML 주석 서버(XML commentary server)에는 주석 처리 도구를 이용해서 웹 페이지의 제작자들이 각 구역에 포함된 정보에 대한 의미를 부여하는 주석처리를 하며, 이는 XML 저장소로 저장된다.

4.2 주석 처리 도구

웹 페이지 문서를 XML로 주석 처리하기 위해 그림 2와 같이 브라우저의 DOM(Document Object Model)[10] 인터페이스와 XML 파서[9]를 이용한 주석 처리 도구(annotation tool)를 사용한다. 여기에 사용되는 XPath(XML Path Language)[8]는 브라우저의 DOM 구조에 의해 결정된다.

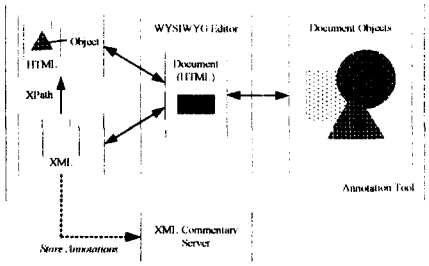

그림 2. 주석 처리 도구

웹 페이지의 저작자들은 웹 페이지의 구역들을 보면서 선택할 수 있고, 이에 나타나는 주석 입력 창을 통해 주석 서버에 주석 처리를 하게 된다. 저작도구는 선택된 정보와 일치하는 XPath의 표기를 내부에서 찾아오기 위해 각 태그의 속성을 포함한 XML 파일을 읽어오는데, 이는 특정한 지역에 있는 노드들의 경로를 지정하여 특정한 값을 갖는 노드를 찾아내는 것이다. 예를 들어, "stock"의 <title>이 선택된다면 이는 "region[title = 'stock']"로 "stock"이라는 값을 갖는 <title>을 자식으로 가진 모든 <region>의 요소를 찾게 된다. 이는 "[ ]" 괄호가 단지 필터처럼 작동하기 때문에 오직 특정한 <region>의 요소들만이 선택되어지기 때문이다[8]. 그리고 이에 따른 XML 주석 처리를 위해 DOM은 내부적으로 수정 또는 생성될 것이다. 또한 사용자가 웹 페이지의 정보 구역들을 선택할 때 각 구역 제목의 XPath는 모든 객체를 포함하는 가장 작은 상위 노드로부터 내부적으로 찾기 시작한다. 따라서 단락의 내용이 바뀌더라도 이에 해당하는 주석은 계속 유용할 것이고, 그것은 그림 3과 같이 XML 파일에서 <contents> 태그 안의 region의 속성으로 정의가 된다. 이런 방식으로 주석 처리 도구는 코드 변환 프록시 서버를 통합하고 사용자는 주석 처리의 결과를 볼 수가 있다.

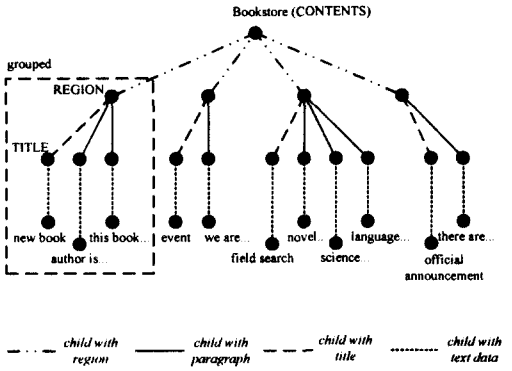

그림 3. XML로 처리된 파일의 구조

4.3 사용자를 위한 페이지 재구성

웹 페이지의 재구성 과정을 설명하기 위해, 여러 개의 구역으로 구성되며 각 구역은 의미에 맞게 XML로 주석 처리가 된 웹 페이지가 있다고 가정하자.

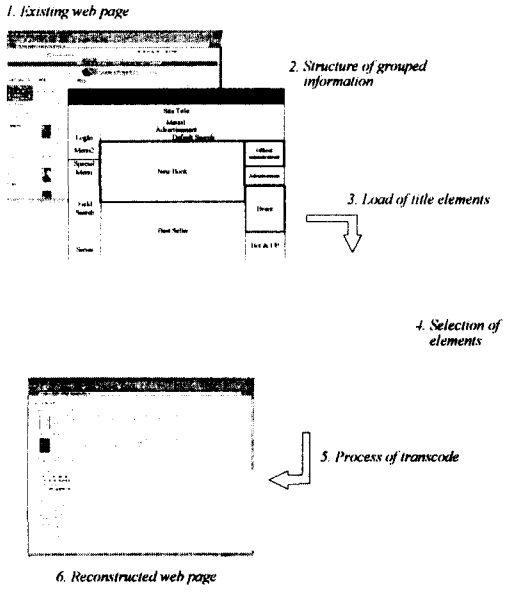

그림 4. 웹 페이지의 재구성 과정

그림 4의 1단계는 현재 존재하는 웹 페이지를 보여주며, 2단계는 이를 동일하게 구성된 같은 유형의 정보가 포함된 각 부분으로 나눈 것이다. 3단계는 XML 주석으로부터 추출된 각 구역의 요소들의 값을 읽어오는 과정이고, 이것들로 구성된 리스트의 모습이 4단계이다. 사용자는 이 구역 명세 리스트를 보면서 의도대로 재구성을 하게 되는데 5, 6단계는 사용자의 선택에 의해 정해진 구역들로 재구성된 웹 페이지를 보여준다. 최종적으로 사용자는 XSLT(XSL Transformation)[11]를 통해 선택된 구역만으로 웹 페이지가 재구성되어 보여지게 된다.

5. 결론

본 논문에서는 XML을 사용하여 웹 문서의 특성에 따른 재구조화를 위한 대화식 코드 변환 시스템을 제안하였다. 그리고 XML로 주석 처리된 웹 페이지 구역들의 재구성 과정을 위한 방법을 살펴보았다.

한편, XML 주석 처리 서버에 웹 페이지의 각 구역에 대한 정확한 의미를 부여하기 위해서는 웹 문서를 제작했던 제작자 스스로가 문서를 재구성하는 것이다. 그러나 이것은 많은 시간을 필요로 하는 과정이므로 여건상 지원자들로부터 작성된 XML 파일을 받아서 XML 저장소에 저장시킨다.

제안된 재구조화 과정을 통한 코드 변환을 사용함으로써 사용자는 편리하게 원하는 정보들을 검색할 수 있다. 하지만 각 웹 정보 구역들이 XML로의 의미가 부여되지 않은 상태에서 웹에 있는 수많은 HTML 문서들의 예기치 못한 구조 때문에 아직도 웹상의 많은 페이지들이 원하는 대로 보여지지 않는다. 그러므로 제안한 코드 변환의 방법을 시도하기 전에 많은 페이지의 테스트가 필요하며, 제안된 시스템을 효과적으로 운영하기 위해서는 자발적인 참여가 필요하다.

또한 사용자가 언제든지 자유로운 구조 형태로 볼 수 있게 이웃 하는 페이지와의 유사점을 검사하여 가능한 한 많은 페이지들이 하나의 XML 주석 파일을 공유할 수 있게 해야 한다.

6. 참고 문헌

[1] Hori M, Kondoh G, Ono S, Hirose S, and Singhal S., "Annotation-Based Web Content Transcoding", Proceedings of The 9th International World Wide Web Conference (WWW9/Computer Networks 33 (1-6)), pp.197-211, 2000.

[2] Nagao K. Semantic Transcoding: Making the World Wide Web More Understandable and Usable with External Annotations, in Proceedings of International Conference on Advanced in Infrastructure for Electronic Business, Science, and Education on the Internet, 2000

[3] P. Brusilovsky, "Methods and techniques of adaptive hypermedia", Journal of User Modeling and User-Adaptive Interaction, Vol.6, No. 2-3, pp. 97-129, 1996

[4] Web Access Gateway, "the Association of C and C++ Users", http://www.accu.org/cgi-bin/access/access

[5] F. Rousseau, J. A. G.-Macias, J.V.de Lima, and A. Duda, "User Adaptable Multimedia Presentations for the WWW", Proceedings of The 8th International World Wide Web Conference, Toronto, Canada, 1999

[6] K. Oh, D. Park, and E. Hwang , "Automatic XML Conversion of Web Pages with Common Pattern", Proc. of Int'l Conference on Computer and Information Science, Orlando, Florida, Oct 2001

[7] Extensible Markup Language (XML) 1.0 (Second Edition) W3C Recommendation, http://www.w3.org/TR/REC-xml

[8] XML Path Language (XPath) Specification Version 1.0. W3C Recommendation, http://www.w3.org/TR/xpath

[9] Xerces, "Apache XML Project", The Apache Software Foundation, http://apache.org

[10] Document Object Model (DOM) Level 2 Specification Version 1.0. W3C Recommendation, http://www.w3.org/TR/DOM-Level-2-Core

[11] XSL Transformation (XSLT) Version 1.1 W3C Working Draft, http://www.w3.org/TR/xslt11