

XML 기반 다목적 콘텐츠 출판 시스템의 설계

김연혜⁰ 황영연

가톨릭대학교 컴퓨터공학과

yhkim@ntechs.com⁰, byhwang@catholic.ac.kr

Design of XML Based Publish System for Multipurpose Contents

Yun-Hye Kim⁰ Byung-Yeon Hwang

Dept. of Computer Science & Engineering, The Catholic University of Korea

요 약

개인용 컴퓨터의 대중화가 이루어지면서 수많은 전자 문서가 작성되어졌다. 그러나 전자 문서의 검색 문제와 문서의 3가지 구성요소인 내용, 구조, 표현의 혼재로 인한 문서의 재작성성이 많이 일어났다. 이러한 소모적인 문서의 재작성은 특히 동일한 콘텐츠를 가진 여러 목적의 출판에서 두드러진다. 본 논문에서는 이러한 전자 문서의 재작성성을 줄이고자 XML 기반 다목적 콘텐츠 출판 시스템을 설계한다. 본 논문에서 설계된 시스템은 전자 문서 변환을 2단계로 분리함으로써 변환기 중복 구현을 최소화하고 전자 문서 편집기간의 호환성 문제를 해결한다.

1. 서 론

개인용 컴퓨터의 대중화가 이루어지면서 수많은 전자 문서가 만들어졌고 그 중 일부는 종이로 출판되어지거나 웹을 통해 전파되어졌다. 이 중 많은 양의 문서들이 같은 내용을 담고 있음에도 원하는 목적에 맞추어 매번 새롭게 만들어야만 하는 문제를 가지고 있다.

그 중요 원인 중 하나는 문서의 3가지 구성요소인 내용, 구조, 표현이 혼재되었기 때문이다. 예를 들면, 신문출판을 위해 만들어진 문서는 단지 신문출판을 위해서만 사용하고 이를 웹 문서로 만들기 위해서는 단순히 웹 문서로 변환하는 것이 아니라 다시 재작성을 해야만 하는 것이다.

이것은 워드프로세서와 웹의 초기 제작 목적 때문이다. 워드 프로세서와 웹은 문서의 내용보다는 문서를 읽게 되는 사람에게 효과적으로 내용을 전달하기 위해서 표현이라는 측면을 강조하게 되었고 그에 따라 내용과 구조, 표현이 뒤섞인 전자 문서가 나타난 것이다. 또한 워드프로세서는 각기 자신만의 고유 포맷을 가지고 있기 때문에 다른 포맷의 워드프로세서와 호환성 문제가 발생하였다.

이러한 문서 및 데이터의 저장 포맷으로 인한 문제가 계속 발생하자 1986년 SGML(Standard Generalized Markup Language)이 발표되었으나 매우 복잡하여 실제 응용에는 많이 사용되지 않았다. 1991년에 SGML로 만들어진 HTML은 워드프로세서의 호환성 문제를 없애고 웹에서 공통적으로 사용하는 포맷으로 자리 매김 했으며 이로 인해 인터넷 사용이 폭발적으로 증가하게 되었다. 그러나 HTML 역시 표현을 강조하였기 때문에 중복된 내용을 담고 있는 문서를 매번 다시 제작해야하는 문제가 발생하게 된다. 1996년 W3C에서 제안한 XML(Extensible Markup Language)은 문서의 내용과

표현을 분리시킴으로써 자원의 재활용이 용이하게 되었으며 표현의 변화에 내용이 독립적이 되어 전자 문서의 전체 재작성이 필요 없게 되었다[1, 2].

문서 재작성의 또 다른 원인으로 여러 전자 문서들의 문서간의 검색을 들 수 있다. 전자 문서는 하나의 파일로 이루어져 있으므로 문서간의 검색이 용이하지 않다. 그에 따라 같은 내용이 있음에도 불구하고 찾지 못하여 새로 전자 문서를 만들게 되는 경우가 자주 발생하게 되었다. 또한 원하는 내용이 하나의 문서에 있지 않음으로 여러 문서에서 필요한 부분만을 발췌해 새로운 문서를 작성하는 경우도 빈번하게 발생하고 있다.

이러한 문제는 모든 문서를 XML 문서로 작성한 후 각기 다른 XML 문서에 접근하여 필요한 요소만을 원하는 위치에 보여주도록 하는 표현을 사용하면 문서의 내용을 변화시키지 않고도 새로운 문서를 볼 수 있게 된다. 본 논문에서는 이러한 다목적 콘텐츠 출판 시스템을 설계하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 XML 출판 모델과 CMS(Contents Management System)에 대해 소개하고, 3장에서 다목적 콘텐츠 출판 시스템을 설계한다. 끝으로 4장에서 결론과 추후 연구과제를 기술한다.

2. 관련 연구

2.1 XML 출판 모델

가장 간단한 출판 모델은 인간의 생각을 종이에 적어 출판하는 양식이였다. 이러한 방식은 컴퓨터의 등장으로 전자 문서를 통한 출판으로 발전되었으며 SGML/XML의 등장으로 패러다임의 변환을 맞게 되었다[3].

XML 출판 모델은 인간의 생각을 문서 데이터의 추상화, 지식화, 장식화 과정으로 분리한다. 따라서 똑같은 내용을 담은 CD-ROM, 신문, 책, 웹, PDF 등 많은 출판물들을 장식화과정만을 변화시킴으로써 콘텐츠의 효율적인

재사용이 가능하게 되었다.

2.2 CMS

XML 출판 모델에서 문서 데이터의 추상화, 지식화를 거친 자료는 단순히 텍스트가 아닌 멀티미디어 자료이다. 또한 상품 목록, 재고, 위치와 같은 특정 의미를 가진 데이터가 된다. 이러한 모든 데이터들을 콘텐츠로 생각하고 통합하여 관리한다면 다음과 같은 이점을 가지게 된다.

- 정보 저장/검색/변경 비용 감소
- 콘텐츠 상태에 따른 효율적인 관리
- 항상 사용 가능한 콘텐츠
- 출판 비용 감소

따라서 CMS는 다음과 같은 구성 요소를 지닌다.

- 데이터 저장소(Repository)[4, 5]
관리할 정보를 저장하고 정보를 요청하면 이를 제공하는 구성 요소이며, 정보의 저장과 제공을 신속하고 효율적으로 수행할 수 있어야 한다.

- 사용자 인터페이스

관리하고 있는 정보를 사용자가 손쉽게 브라우징하고 찾아볼 수 있도록 하는 구성요소이다.

- 편집 도구

관리하게 될 정보의 생성 또는 편집을 위한 도구이다.

- 워크플로우

관리할 정보는 다수의 사용자에게 의해 생성, 변경, 검색되게 된다. 이러한 다수 사용자 환경에서 각 콘텐츠를 구성하고 있는 요소들이 현재 어떠한 상태에 있는가를 적절히 추적하고 통제, 관리하는 것은 정보의 적합성 유지를 위해 필수적인 일이며 이러한 기능을 제공하는 구성요소가 워크플로우이다.

- 검색 엔진

관리하고 있는 정보를 손쉽게 활용하기 위해서는 다수의 정보 중 자신이 원하는 정보를 정확하고 신속하게 검색할 필요성이 절대적이며, 이 경우 지원되어야 할 검색 단위는 전체로서의 정보가 아니라 정보를 이루고 있는 각 콘텐츠 단위로 검색할 필요성이 있다. 따라서 각 콘텐츠간의 관계성 및 이들에 대한 관리 정보 등을 포함하여 복합적인 형태의 검색 기능이 되어야 한다[6, 7].

- 출력 도구

출력 도구는 데이터 저장소에 저장되어 있는 정보를 특정한 사용 유형에 맞도록 포맷해 주는 도구이다. 즉 동일한 정보가 인쇄, CD-ROM, 인터넷 등을 통해 다양한 형태로 출판될 수 있게 하고, 필요시 정보의 일부분만을 선택적으로 출판할 수 있게 한다. 또한 사용자 요구에 맞는 출판이 가능하도록 다양한 형태의 출력 도구를 요구하게 된다.

3. 다목적 콘텐츠 출판 시스템의 설계

다목적 콘텐츠 출판 시스템은 XML 편집기로 작성된 XML 문서뿐만 아니라 기존의 작성되어진 각종 전자 문서, 데이터베이스 안의 데이터 등을 문서의 제작성 또는 데이터의 중복 없이 통합된 콘텐츠 관리를 통해 인쇄,

CD-ROM, 인터넷과 같은 여러 출력 형식을 갖도록 하는 시스템이다. 따라서 기존의 전자 문서를 원하는 XML 문서로 변환하는 변환기가 필수적이다.

특정 벤더에 종속된 기존 전자 문서의 XML 문서 변환기는 현재 특정 벤더 또는 여러 프로젝트를 통해 개발되어졌다. 변환기를 통한 결과물, 즉 XML 문서의 유형은 다음과 같다.

- 특정 벤더 종속 XML 문서
- XML 응용(application) 문서

특정 벤더에 종속된 XML 문서는 역시 벤더에 종속적인 시스템에서만 사용 가능하다. 이 문제를 해결하기 위해 XML 문서를 범용적으로 사용하는 XML 응용 등으로 변환하기 위해서는 역시 벤더에 종속적인 변환기와 매핑이 필요해지므로 원하는 XML 문서를 만들기 위해서 여러 번의 변환과 매핑이 발생된다. 이것은 비효율적이며 호환성이 보장되지 못한다.

XML 응용 문서는 기존의 전자 문서를 XML 응용 문서로 바로 변환시킴으로써 앞에 방식보다는 좋은 호환성을 가진다. 그러나 그림 1과 같이 많은 XML 응용과 많은 종류의 전자 문서 포맷을 가진다면 변환기를 매번 제작해야 한다는 단점을 가진다.

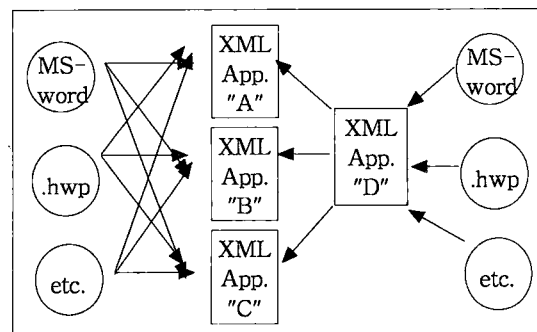


그림 1. XML 응용으로 직접 변환하는 변환기와 공통의 XML 응용을 통한 변환기

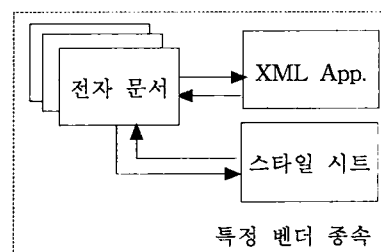


그림 2. 특정 벤더에 종속적인 전자 문서와 공통의 XML 응용으로의 변환

따라서 전자 문서 편집기에서 사용될 XML 응용을 설

계하고 전자 문서 편집기 벤더 측에서는 그림 2와 같이 설계된 공통의 XML 응용으로의 변환기만을 제공하면 다른 XML 응용에서 제공하는 XML To XML 변환기를 통해 쉽게 어떤 XML 응용으로도 변환기 가능해진다.

예를 들어 여러 전자 문서 편집기로 작성된 기사를 신문, TV 뉴스, 라디오 뉴스 등 여러 매체에서 출판되는 뉴스를 위한 XML 응용인 NewsML과 텍스트 뉴스를 위한 NITF(News Industry Text Format)로 변환한다고 하자[8, 9].

사용자들은 이전부터 쉽게 사용되어진 전자 문서 편집기로 기사를 작성한 후 벤더 측에서 제공된 변환기로 편집기 공통의 XML 응용 문서로 변환시킨다.

그 후 NewsML이나 NITF 측에서 제공한 편집기 공통의 XML 응용과의 변환기와 매핑기를 사용하면 기존의 어떠한 전자 문서라도 쉽게 원하는 XML 응용으로 변환시킬 수 있다.

이렇게 생성된 XML 응용 문서들은 하나의 컨텐츠로 통합, 관리되고 XML 문서들은 역 변환을 통해 전자 문서 출판이 가능해진다.

이러한 다목적 출판 시스템의 핵심인 XML 응용 변환기는 그림 3과 같다.

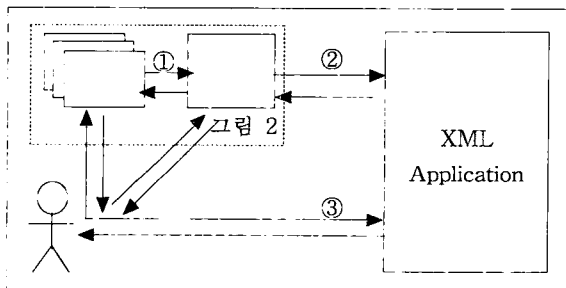


그림 3. XML 응용 변환기

단계 ①은 기존의 전자 문서를 전자 문서 편집기 공통의 XML 응용 문서로 변환시키는 변환기이다. 이 변환기는 이전의 문제점이던 각 전자 문서 편집기간의 호환성 문제를 해결한다. 그러나 이러한 공통의 XML 응용은 아직 연구되어지고 있지 않다.

단계 ②는 전자 문서 편집기 포맷에서 추출된 XML 문서를 각 목적, 각 분야에 맞춘 XML 응용으로 변환시키는 것이다. 이러한 변환기는 XML 응용이 새로 만들어질 때 함께 연구되어져 제공된다면 변환기의 중복 제작 문제를 해결할 수 있다.

따라서 기존 전자 문서 편집기(①, ②) 및 XML 편집기(③)를 통해 생성된 다양한 문서들은 XML 응용으로 변환되어 하나의 컨텐츠로 통합되고 관리되어 문서의 재작성 없이 인터넷과 여러 방송 매체를 통해 출판될 수 있게 된다.

또한 XPoint, XLink 등을 사용하여 문서 중 필요한 일부분만을 가져와 새로 장식함으로써 전자 문서의 컨텐츠 중복을 감소시킨다. 따라서 수정/저장 비용이 감소되며 문서간의 관련 정보를 가지게 됨으로써 더욱 지능화된

컨텐츠 관리가 가능하게 된다.

4. 결론 및 추후 연구과제

현대인들은 수많은 전자 문서들과 여러 매체를 통한 출판물들의 홍수 속에서 살고 있다. 또한 똑같은 내용을 가진 여러 방식의 출판물들을 위해 문서의 재작성이 빈번하게 발생하고 있다. 이러한 문제점을 해결하기 위하여 전자 문서 변환기를 포함하는 다목적 컨텐츠 출판 시스템을 설계하였다.

다목적 컨텐츠 출판 시스템의 가장 중요한 단계인 전자 문서 변환을 2단계로 분리함으로써 변환기 중복 구현을 최소화하였고 전자 문서 편집기간의 호환성 문제도 해결하였다.

그러나 이 변환기가 실제로 사용되기 위해서는 우선 전자 문서 편집기를 만드는 벤더들 간의 공통적인 XML 응용이 필요하며 두 번째로 NewsML, MathML 등과 같이 특정 분야 및 의미를 가진 XML 응용과의 변환이 허용되어야만 한다.

앞으로 연구될 것은 본 논문에서 제안된 설계안을 확장해서 다양한 기능이 추가된 XML 기반 다목적 출판 시스템을 구현하는 것이다.

참고문헌

- [1] <http://www.xml.org/>
- [2] <http://www.w3.org/Style/XSL/>
- [3] K4M, "TheoCMS SGML/XML enabled CMS(Contents Management System)," Technical Publication Team, 1999.
- [4] J. Mchugh, S. Abiteboul, R. Goldman, D. Quass, and J. Widom, "Lore: A Database Management System for Semistructured Data," ACM SIGMOD Record, Vol. 26, No. 3, pp. 54-66, Sep. 1997.
- [5] M. Yoshikawa, T. Amagasa, and S. Uemura, "XRel: A Path-Based Approach to Storage and Retrieval of XML Documents Using Relational Databases," ACM Transactions on Internet Technology, Vol. 1, No. 1, pp. 110-141, Aug. 2001.
- [6] T. Milo and D. Suci, "Index Structures for Path Expressions," ICDT 99, Jan. 1999.
- [7] J. P. Yoon, Y. Raghavan, and V. Chakilam, "BitCube: A Three-Dimensional Bitmap Indexing for XML Documents," Journal of Intelligent Information Systems, Vol. 17, Nov. 2001.
- [8] <http://www.NewsML.org/>
- [9] <http://www.nitf.org/>