XML의 View DTD를 이용한 웹 상의 정보통합 및 검색

홍석일° 양정욱 홍동완 윤지희
한림대학교 컴퓨터공학부
{sihong, jerry, dwhong, jhyoon}@iris.ce.hallym.ac.kr

Information Integration and Search on the Web using XML View DTD

Seok-Ill Hong° Jeong-Uk Yang Dong-Wan Hong Jee-Hee Yoon
Dept. of Computer Engineer, Hallym University

요 약

인터넷에 산재해 있는 분산 이질의 대규모 정보들을 통합, 관리하기 위한 기본 모델로서 최근 정보교환 표준으로 인정 받고 있는 XML을 사용할 수 있다. HMS(Hallym Mediator System)는 XML을 기본 데이터 모델로 하는 미디에이터 시스템으로서 뷰(View) DTD의 정의, 추론 기능을 이용하여 웹 상의 분산, 이질 정보의 통합 기능을 제공한다. 뷰 DTD는 관리자에 의하여 정의되어, 추론 및 보완 과정을 거쳐 생성되며, 웹 상의 통합된 가상 정보 구조를 표현한다. 웹 상의 사용자는 이와 같이 생성된 뷰 DTD를 근거로 분산이질 정보에 대한 구조정보 및 내용정보에 대한 질의를 수행할 수 있다. HMS에서는 DTD 구동형 비주얼 사용자 인터페이스를 제공하여, 관리자와 일반 사용자에게 직관적이고 간편한 웹 정보 브라우징 및 질의검색 환경을 제공한다.

1. 서론

국내 인터넷의 사용이 최근 몇 년 동안 급격하게 증가하여 그 보급율이 세계 7 위까지 오르게 되었으며, 인터넷 사용의 증가에 기인하여 웹 기반 데이터의 증가 속도도 기하급수적으로 증가하고 있는 추세이다. 이 들 방대한 양의 데이터들은 일반 텍스트 데이터, 이미지, 음성, 동영상, RTF(Rich Text Format)등의 멀티미디어 데이터들로서 구조화 되어지기 힘든 자료형을 가지고 있다. 이런 데이터들은 대부분 유용한 자료들로서 서로 공유되어져야 하며, 이러한 공유에 의하여 새로운 데이터의 창출을 가져올 수 있다. 따라서 정보의 통합은 필수적인 요구조건으로 대두되고 있으며, 최근 구조적 문서를 생성하는 XML[1] 이 제안되어 인터넷상의 전자문서 표준으로 자리잡아가고 있다. 그러나 기존의 시스템은 XML 문서와는 다른 데이터 타입을 기본 데이터 모델로 가지고 있으므로 각 기존 시스템의 데이터를 XML 문서로 전환하는 데에 어려움이 있다.

최근 분산 환경에서 이질적인 데이터를 통합하기 위한 방법론으로서 미디에이터[2] 개념이 널리 활용되고 있으며, 이들 대규모 정보들을 통합, 관리하기 위한 기본 모델로서 XML을 사용할 수 있다[3, 6].

본 논문에서는 XML을 기본 데이터 모델로 하는 HMS 미디에이터 시스템의 개발에 대하여 논한다. HMS 에서는 분산 이질 정보의 가상적인 통합 뷰의 생성과 질의 처리에 XML의 DTD를 이용한다. 관리자는 각종 응용 목적에 적합하도록 여러 개의 소스 DTD로부터 임의로 뷰 DTD를 선언적으로 정의할 수 있으며, 시스템은 DTD 추론 과정에 의하여 뷰 DTD를 생성한다. 생성된 뷰 DTD는 관리자와의 보완 과정을 거쳐, 최종적으로 XML 저장소에 저장되며, 사용자는 이 들 DTD를 웹 상의 정보 구조로 인식하여 이에 대한 질의를 수행함으로써 원하는 결과를 얻게된다. 여기에서는 시스템 구성 및 각 모듈의 특성에 대하여 설명하고, 관리자에 의한 뷰 DTD 생성 인터페이스와 일반 사용자에 의한 질의 인터페이스를 이용한 웹 상에서의 정보 통합 및 검색 예를 보인다.

2. HMS (Hallym Mediator System)

2.1 시스템 구성

본 시스템의 기본 구조는 그림 1 과 같다. 시스템 기능은 크게 두 가지로 나누어 미디에이터 관리자에 의한 통합 뷰 생성 및 관리기능과 일반 사용자에 의한 정보검색 기능으로 나눌 수 있다. HMS 에서는 뷰의 정의와 정보 검색을 위한 미디에이터 언어 HMLL(HML Language) 을 제공한다. HMLL은 XML-QL[4], Yat, UnQL 등의 XML 질의언어와 동등한 기능을 가지는 고 수준 선언적 질의언어로서, 정보 구조를 정의하는 생성절(construct clause) 과 검색조건을 정의하는 조건절로(where clause) 로 이루어져 있으며 뷰의 정의와 정보검색을 위하여 동일 언어를 사용한다.

통합 뷰 생성 기능 : 미디에이터 관리자는 웹상의 각종 소스에 산재되어 있는 정보를 통합, 가공하여 응용 목적에 적합한 가상의 통합 뷰를 정의, 구축하여야 한다. 일반적으로 이 일은 시간을 요하는 전문적인 일로서 전체 시스템 운영에 큰 부담으로 작용할 수 있다. HMS 에서는 웹 상의 정보 소스에 대한 각종 정보(소스 내용, 속성정보, 제약조건, 소스 신뢰도, 질의처리 능력 등)를 메타 데이터형태로 수집, 저장, 관리한다[3]. 관리자 인터페이스는 이 들 메타데이터를 기반으로 소스 DTD 브라우징 기능, 뷰 DTD 편집/생성 기능, 뷰 DTD 추론기능, 설명 기능 등을 제공하여 관리자의 통합 뷰 생성을 지원한다. 관리자는 전용 인터페이스를 통하여 뷰 DTD 를 정의하게 되며, DTD 추론 모듈은 정의에 의하여 뷰 DTD 를 자동 생성한다. 생성된 뷰 DTD 는 관리자에 의하여 다시 보완될 수 있으며, XML 저장소에 저장된다.

정보 검색 기능 : 사용자는 일반 사용자 인터페이스를 이용하여 응용 목적에 적합한 뷰 DTD 를 구동 시킨 후, 분산이질 정보에 대한 구조정보 및 내용정보에 대한 질의를 수행할 수 있다. HMS 에서는 정보검색을 위한 방법으로 가상 접근기법(virtual approach) 을 이용한다. 이 방식은 웨어하우징(warehousing) 기법에 비하여 최신 데이터를 얻을 수 있다는 점, 데이터 중복이 불필요하다는 점 등의 장점이 있지만, 효율면에 문제가 있을 수 있으므로 질의 평가를 위한 최적화 작업이 병행되어야 한다[3]. HMS 에서는 뷰 DTD 를 근거로 작성된 사용자의 질의를 입력 받아 각종 정보 소스에 대한 메타 데이터를 이용하여 질의의 효율적 평가를 위한 실행 전략을 수립한 후, 실행전략에 의하여 실제로 정보가 산재하여 있는 정보소스로부터 래퍼를 통하여 정보를 추출하고, 이를 다시 통합 가공하여 사용자에게 결과로 전송한다.

시스템 구성요소인 각 모듈의 기능 및 특징을 간단히 설명하면 다음과 같다. 사용자 인터페이스 기능은 2.2 절에서 예제와 함께 제시한다.

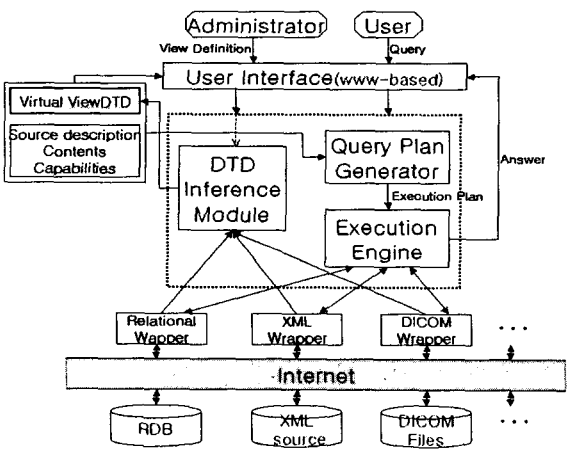

그림 1. HMS 의 시스템 구성도

(1) Query Plan Generator :
사용자의 질의(뷰 정의, 정보검색 질의)를 입력 받아 시스템 내부자료 구조로 변형 한 후, 정보 소스의 메타 데이터를 이용하여 세부 실행전략을 작성한다. 사용자 인터페이스를 통해 입력된 질의는 우선 파서를 통해 구문분석 작업을 거쳐 시스템 전반에서 사용할 수 있는 내부구조로 저장되어지며 이 구조정보는 서브 질의 생성시 질의 분할의 기초로 사용된다. 파싱된 질의에 의한 구조체는 그림 2 와 같이 이진 트리를 기본 구조로 사용하며, 트리 노드는 엘리먼트 이름, 자식 노드, 형제노드의 엘리먼트 관련 정보를 가진다. 트리가 여러 개일 경우에는 연결-리스트로 연결된다.

(2) Execution Engine :
실행 전략에 따라 분할된 서브 질의를 각각의 정보 소스의 래퍼로 전송한다. 래퍼는 서브 질의에 대한 검색 결과를 HMS 시스템 내부 자료구조로 변형하여 전송하게 되며, 이 결과는 트리구조에 저장된다. 이후 저장된 정보에 대하여 트리를 순회하면서 적합한 XML 문서를 생성하게 된다. 생성된 XML 문서는 사용자 인터페이스로 전달되어 사용자에 의하여 브라우징 혹은 저장될 수 있다.

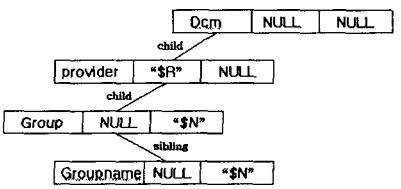

그림 2. 질의에 의해 생성되는 트리 노드의 예

(3) DTD Inference Module :
관리자의 뷰 DTD 정의에 의하여 뷰 DTD 를 자동 생성한다. 생성된 뷰 DTD 의 정확성(tightness)를 제공하기 위하여 참고문헌 [5] 에서 제시한 Tightening 알고리즘을 구현하였다. 단 이 알고리즘은 특수 형태의 부분 질의(pick-element query)에 대하여만 tight 한 DTD 를 보장하므로, 본 시스템에서는 정확한 DTD 작성을 위하여 사용자에 의한 보완작업을 필요로 할 수 있다.

2.2 시스템 구현

본 시스템은 윈도우 NT 기반에서 JDK 1.2 를 이용하여 구현하였으며, XML 파서로는 IBM Parser for java[6] 를 이용하였다. 하부 데이터베이스로는 RDBMS 인 Oracle 8i 를 사용하였으며 데이터베이스와의 연결은 JDBC 를 사용하였다.

본 시스템에서는 관리자를 위한 뷰 정의 인터페이스와 일반 사용자를 위한 정보 검색 인터페이스를 기본으로 제공한다. 그림3의 예제를 이용하여 사용자 인터페이스의 기능을 설명한다. 그림 3 의 예제에서 사용되는 DTD 는 환자 기초정보에 대한 DTD 와 영상정보에 대한 DTD 의 일부를 나타낸다.

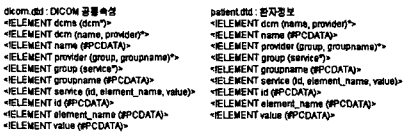

그림3. 예제 DTD

그림 4 는 관리자 인터페이스에 의한 통합 뷰 생성 과정을 나타낸다. 왼쪽 상단의 화면은 소스 DTD 를 읽어 들이는 화면으로 임의의 개수의 DTD 를 읽어 들일 수 있으며, 구조를 나타내는 디렉토리형의 트리 형태로 출력된다. 왼쪽 하단은 뷰 DTD 생성 질의를 입력하는 부분이다. 드래그앤드롭 방법으로 완성된 질의는 실행 버튼에 의해서 뷰 DTD 로 변환되어 오른쪽 화면에 결과가 출력된다. 이 예에서는 그림 3 에 보인 2개의 DTD 를 이용하여 영상정보를 포함한 환자정보에 대한

뷰 DTD를 생성하는 과정을 나타내고 있다.

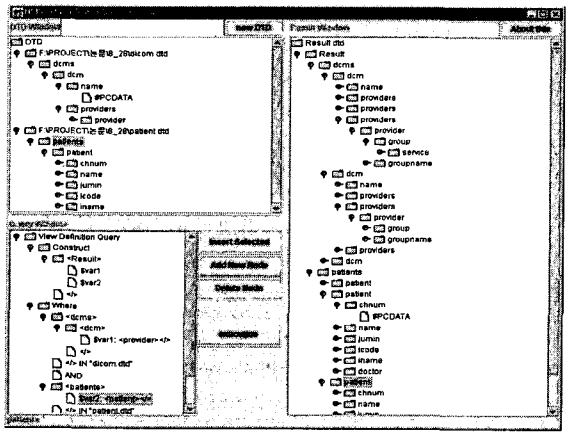

그림 4. 뷰 DTD 생성 예

그림 5는 일반 사용자 인터페이스에 의한 질의검색 과정을 나타낸다. 왼쪽 화면은 역시 소스 DTD를 읽어 들이는 화면으로 앞에서 생성된 뷰 DTD를 읽어 들일 수 있다. 이 예에서는 그림 4에서 생성된 뷰 DTD를 이용하여 '영상사진분류가 ct.001 인 환자의 환자정보 중 이름, 주민등록번호, 의사명을 출력하라' 는 사용자 질의를 생성하고 처리되는 과정을 보이고 있다. 오른쪽 화면은 질의 작성 화면이며, 보기 버튼을 클릭하여 HMLL로 생성된 질의를 확인할 수 있다. 실행버튼을 클릭하면 검색된 결과가 XML 문서형태로 보여진다. 그림 6에서는 그림 4와 5에서 사용된 뷰 정의문, 생성된 뷰 DTD, 사용자 질의문을 보인다.

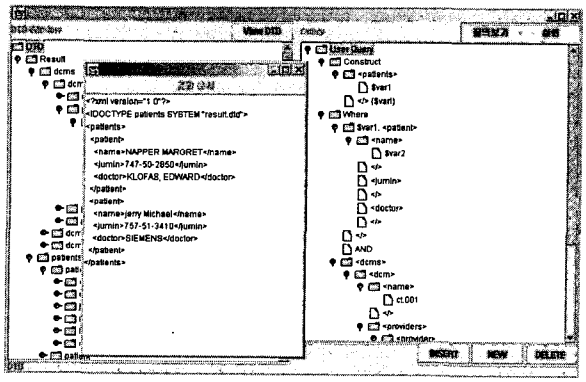

그림 5. 사용자의 질의 검색 예

3. 결론

본 시스템에서는 XML DTD를 기반으로 웹 상의 이질분산 정보에 대한 통합 뷰를 제공하고, 이들 정보 자원들에 대한 구조/내용 검색을 수행하기 위한 간편하고 직관적인 비주얼 사용자 환경을 제공한다. 향후 연구과제로는 DTD 추론 알고리즘의 개선을 통하여 뷰 DTD에 대한 최적화를 수행해야 하며 질의 언어의 확장을 통해 하위 정보에 대한 검색뿐만 아니라 변경, 삭제등과 같은 관리 작업을 수행하기 위한 연구가 필요하다.

그림 6. 뷰 정의, 뷰 DTD, 사용자질의

참고 문헌

[1] "Extensible Markup Language(XML) 1.0 ", http://www.w3.org/TR/1998/REC-xml-19980210.

[2] G. Wiederhold. "Mediators in the Architecture of Future Information Systems ", IEEE Computer, 25(3) :pp. 38-49, 1992.

[3] D . Florescu, A. Levy, A. Mendelzon, "Database Techniques for the World-Wide Web: A Survey ", SIGMOD Record 27(3):pp.59-74, 1998.

[4] A. Deutsch, M. Fernandez, D. Florescu, A. Levy, and D. Suciu, "XML-QL – A query language for XML". In Proc. Of the Int. World Wide Web Conf., Canada, 1999.

[5] B. Ludascher, Y. Papakonstantinou, P. Velikhov, V. Vianu, "View Definition and DTD Inference for XML ", Post-ICDT Workshop on QPSD.

[6] "XML Parser for Java Version 3.0.1", IBM alphawork, http ://www.alphaworks.ibm.com/tech/xml4j/.

[7] C. Baru. A Gupta. B. Ludascher, R. Marciano, Y. Papakonstantinou, P. Velikhov, V. Chu, "XML Based Information Mediation with MIX", exhibition program, ACM Conf. On SIGMOD'99, Philadelphia, USA, 1999.

[8] 이경하, 조정수, 이강찬, 이규철, "XMF:XML 기반 분산 이질 정보 자원의 통합 프레임워크", 2000 봄 데이터베이스 학회, pp. 262-270, 2000.

[9] B. Ludäscher, Y. Papakonstantinou, P. Velikhov, "A Brief Introduction to XMAS ", XMAS sub-group of MIX, February 9, 1999.

[10] Y. Papakonstantinou, P. Velikhov. "Enhancing semistructured data mediators with document type definitions", In Proc. ICDE Conf., 1999.

[11] S. Nestorov, S. Abiteboul, R. Motwani, "Inferring