분산 데이터 통합을 위한 색인기반의 매핑 시스템

설진안° 김운용 정계동 최영근
광운대학교 컴퓨터과학과
{nicolas°, wykim, gdchung, ygchoi}@kw.ac.kr

Mapping System based on Indexing for Integrating Distributed Data

Jin-An Seol° Woon-Yong Kim, Kye-Dong Jung, Young-Keun Choi
Dept. of Computer Science, Kwang-Woon University

요 약

분산된 데이터는 이질적인 시스템 환경으로 인하여 공유하기 어렵고, 데이터의 형식 및 데이터 모델이 서로 다르게 정의되어 사용함으로서 통합하기 또한 어렵다. 본 논문에서는 이러한 문제를 해결하기 위해 분산된 데이터를 XML문서로 변환한다. 또한 색인기법으로 문서의 구조 및 컨텐츠 정보를 추출하여 서로 다르게 정의된 의미정보를 데이터 사전과 비교하여 표준문서로 통합할 수 있는 색인 기반의 매핑 시스템에 대해 기술한다. 제안된 매핑 시스템은 DOM이나 SAX와 같은 표준 인터페이스를 사용하여 XML문서를 통합하는 것보다 효율적으로 통합할 수 있다.

1. 서 론

정보추출은 한 문서에서 그 문서의 중심적 의미를 나타내는 특정 구성요소를 인식하여 추출하는 작업을 가리킨다.[1] 정보추출의 대상은 구조화(structured) 문서, 그리고 준구조화(semi-structured) 문서이다. 준구조화 문서나 구조화 문서는 실제 데이터와 그 데이터가 나타내는 의미를 메타 데이터를 통해 명시적으로 표현하는 문서 형태이다. 이러한 데이터들은 분산 환경에서 독립적으로 데이터를 관리함으로서 접근의 제한성, 이질적인 시스템 환경, 구현언어의 다양성 등으로 인하여 데이터 공유 및 통합에 있어서 어려움은 물론 많은 시간과 비용을 투자해야한다.

따라서 사용자의 요구에 맞게 분산된 데이터를 이용하여 통합적인 결과를 제공하고, 데이터를 공유함으로서 모두에게 편리함과 동시에 효율성을 증가시킬 수 있는 연구가 필요하다.[6]

본 논문에서는 이질적인 시스템으로부터 서로 다른 의미로 정의된 데이터를 수집하여 통합할 수 있는 매핑 시스템을 위해 분산된 데이터를 XML문서로 변환하고, 색인기법으로 문서의 구조 및 컨텐츠 정보를 추출하여 서로 다르게 정의된 의미정보를 데이터 사전과 비교하여 표준문서로 통합할 수 있는 색인 기반의 매핑 시스템에 대해 기술한다. 2장에는 이와 관련된 연구에 대해 기술하고, 3장에는 색인 기반의 매핑 시스템 구성과 색인정보 추출 및 매핑기법에 대해 기술하며 마지막 4장에서는 결론 및 향후 연구방향에 대해 기술한다.

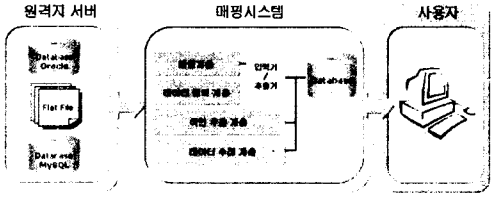
2. 관련연구

XML은 일반적으로 기업 환경뿐 아니라 웹상에서 데이터 교환을 위한 선도적인 언어로 부각되어 왔다. 또한 이기종 간에 포괄적인 접근을 가능하게 해주는 기술로서, 시스템 환경에 제한받지 않고 데이터 교환이 가능하다[2]. 이러한 기술을 이용하여 XML 기반으로 통합하려는 연구가 활발하게 진행되고 있다[3][4]. 따라서 구조화된 XML문서를 통합하고 효율적으로 검색 및 관리하기 위해서는 문서의 내용 및 구조정보가 필요하다. 색인기법은 이러한 문서의 내용정보 및 구조정보를 색인화하는 방법으로 K-ary 완전트리 기법, ID Assignment기법 등이 있다[5]. K-ary 완전 트리(K-ary Complete Tree) 색인 기법은 SGML 문서를 K-ary 완전 트리 매핑 과정을 통해 구조검색을 지원하기 위한 색인기법으로 부모노드와 자식노드의 관계를 간단한 수식을 이용하여 빠르게 검색할 수 있 장점을 갖고 있지만 가상 노드까지 노드 번호를 부여하는 단점을 가지고 있으며 조상/부모/형제/자손에 대한 계층정보와 순서 정보를 알기 어렵다. ID Assignment 색인 기법은 구조화된 문서의 문법적인 구조를 부모와 자식의 관계를 색인 내에서의 경로 표현을 구하기 위해 문서 구조의 특정 형태를 취하는 추상화에 기반한 색인 기법으로 자신의 노드 ID 앞에 부모의 노드를 추가하는 방법으로 자동적으로 자신의 조상노드의 정보를 알 수 있음은 물론 형제 노드에 대해서는 순서를 부여 하므로서 조상/부모/형제/자손에 대한 계층정보와 순서 정보를 간단하게 알 수 있다[5].

본 논문에서는 ID Assignment기법을 기반으로 문서를 통합할 수 있는 매핑기법 및 과정에 대해 기술한다.

3. 색인 기반의 매핑 시스템 구성

매핑 시스템은 XML문서를 통합하기 위해 미들웨어를 갖춘 3-Tier시스템으로 구성되어지며 크게 세 부분으로 나누어진다. 첫째 데이터 소스를 제공하는 원격지 서버 구역, 둘째 추출된 데이터를 표준문서와 매핑하여 통합된 정보를 제공하기 위한 매핑 시스템 구역, 셋째 사용자가 매핑 서비스를 이용하기 위한 사용자 구역으로 나누어진다. [그림1]은 시스템의 전체 구성도를 나타내며 구성요소로는 다음과 같다.



[그림1] 전체 시스템 구성

(1) 데이터 수집 계층: 각각의 분산된 서버로부터 데이터를 수집하여 XML문서로 변환하는 등 수집과 관련된 모든 작업은

"원격접속 관리자"(RCM: Remote Connection Manager)에 의해 이루어진다.

(2) 색인 추출 계층: XML의 구조적인 정보와 데이터 정보를 추출하기 위한 계층으로 "색인정보 추출기"에 의해 이루어진다.

(3) 데이터 정의 계층: 같은 의미지만 서로 다르게 정의된 이름을 검색 및 비교할 수 있도록 "데이터 사전"을 제공하는 계층이다.

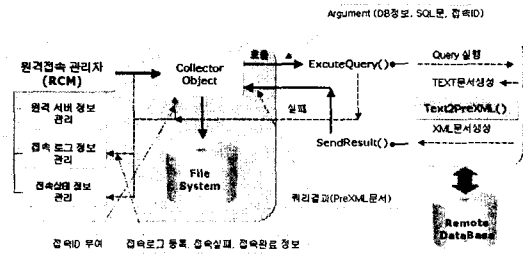
(4) 매핑 계층: "매핑 테이블 작성기"를 이용하여 표준문서의 엘리먼트를 기준으로 각각의 XML문서를 통합하기 위해 "매핑 테이블"을 작성하는 계층으로 데이터 사전을 이용한다.

(5) 입력기/추출기: 작성된 매핑 테이블을 기초로 데이터베이스에 저장 및 추출 기능을 제공하는 계층이다.

3.1. 원격접속 관리자 (RCM: Remote Connection Manager)

RCM은 원격지 서버 정보관리에서부터 실제 데이터를 추출하는 Collector Object 생성까지 데이터 추출과 관련된 모든 정보를 관리한다.

RCM은 다양한 원격지 서버로부터 데이터를 추출하기 위해 객체를 생성하여 주기적으로 데이터를 추출한다. 추출된 각각의 데이터는 콤마로 분리되어 어느 정도 구조화된 텍스트 문서를 생성하게 되며, 이 텍스트 문서는 통합을 위해 하나의 레벨만 갖는 스타(Star)형의 XML문서로 변환되어 Collector Object에게 리턴한다. [그림2]는 이러한 추출을 위한 모든 과정을 관리하는 원격접속 관리자의 구성이다.



[그림2] 원격접속관리자(RCM)

원격접속 관리자에서 가장 중심적인 역할을 수행하는 객체는 RCM에 의해 생성되는 Collector Object로서 생성과 동시에 접속로그에 기록되어 "접속 ID"를 부여받는다. 또한 원격지 서버에 접속 및 데이터를 추출하기 위해 인자값 - DB정보, SQL문, 접속ID - 을 이용하여 ExecuteQuery()를 호출하게 되며 접속실패 또는 접속완료 등의 이벤트를 핸들링 한다. 이를 위해 호출하는 주요 함수와 기능은 다음과 같다.

(1) ExecuteQuery(): 이 함수는 쿼리를 수행하여 텍스트 파일로 결과를 저장하고, 저장된 텍스트 파일을 다시 XML문서로 변환하기 위해 Text2XML()함수를 호출한다.

ExecuteQuery()는 연결된 원격지 서버를 확인하고 SQL문을 이용하여 다음과 같은 두가지 쿼리를 실행한다. 첫째, 해당 테이블의 필드명을 추출하고, 둘째 사용자가 입력한 SQL문으로 데이터를 추출한다. 실행 결과는 콤마(,)라는 구분자(Delimit)를 이용하여 해당 필드의 집합인 레코드 단위로 추출한다. 또한 파일명은 접속ID를 파일명으로 갖는 텍스트 파일을 생성하며, 형식은 다음과 같다.

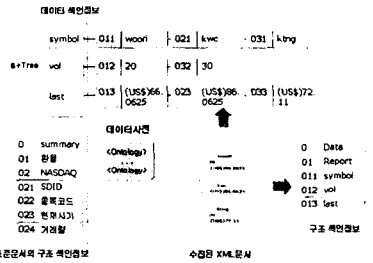
File Name: 접근ID.txt(예, 9200.txt)
Field $Name_1$, . . . Field $Name_n$ //테이블 필드명 표시.
Field $Value_1$, . . . Field $Value_n$ //하나의 레코드 표시

(2) Text2XML(): 생성된 텍스트 파일은 XML문서로 변환해야 한다. 변환하는 이유는 대부분의 분산된 시스템은 다양한 시스템 환경에서 운영되기 때문에 다른 시스템과 데이터 교환이 쉽지 않다. 따라서 XML은 이러한 문제를 해결할 수 있는 방안으로서 Text2XML()는 텍스트 문서를 XML문서로 변환한다.

(3) SendResult(): 변환된 XML문서는 이 함수에 의해 Collector Object에게 파일을 리턴한다.

3.2. 색인정보 추출기

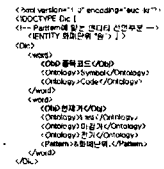
"색인정보 추출기"의 목적은 계층적인 구조를 갖는 XML문서의 효율적인 검색과 표준문서를 기준으로 변환된 XML문서를 매핑하기 위해 문서의 구조인 엘리먼트의 색인정보와 데이터 정보인 컨텐츠 정보를 분리해 추출하는 것이다. 본 논문에서는 문서의 구조 및 데이터 정보를 추출하기 위해 색인기법을 기반으로 한다. 이 기법은 조상/부모/형제/자손에 대한 정보와 순서정보를 알 수 있다. 즉 자신의 EID(Element ID) 앞에 부모의 EID를 추가하며 자신의 부모 및 조상 엘리먼트의 계층정보를 알 수 있음은 물론 같은 레벨에 있는 형제(Sibling) 엘리먼트의 경우 ID에 순서를 지정함으로서 형제 엘리먼트 간의 순서정보를 알 수 있다[7]. [그림3]은 수집된 XML문서를 데이터와 구조정보를 각각 분리해 색인정보를 추출해 저장한 형태이다.



[그림3] 데이터와 구조의 색인정보 추출

3.3. 데이터 사전

데이터 사전은 데이터의 의미는 같지만 이름이 다르게 정의되어 사용되는 엘리먼트의 이름을, 표준문서의 각 엘리먼트에 대한 유사어를 조사하여 정의한다. 다음은 XML문서로 작성된 "데이터 사전"이다



데이터 사전은 다음과 같이 두 가지 기능을 제공한다. 첫째 표준문서에서 정의한 엘리먼트 이름과 유사한 용어 사이의 관계를 정의한다. 데이터 사전에서 <Obj>엘리먼트들은 표준문서

에 정의된 이름이며 <Ontology>엘리먼트들은 해당 <Obj>와 유사한 의미를 정의한 것이다. 따라서 데이터 사전에 <Obj>엘리먼트의 컨텐츠가 "지역"이면, 해당<Ontology>엘리먼트의 내용과 추출된 XML문서의 엘리먼트를 비교하여 같으면 매핑하게 된다. 또한 <Obj>에 대한 <Ontology>엘리먼트가 없을 경우 사용자에 의해 매핑되고 이 것을 추가한다.

둘째 데이터 사전은 데이터 패턴 매칭 기능을 제공한다. 화폐단위의 경우 국가별로 다르다. 따라서 컨텐츠 정보의 패턴을 통일하기 위해서 <Obj>엘리먼트가 <Pattern>엘리먼트를 포함하는 경우 DTD에 정의된 해당 ENTITY를 참조하여 컨텐츠 정보의 패턴을 통일한다.

3.4. 매핑 테이블 작성기

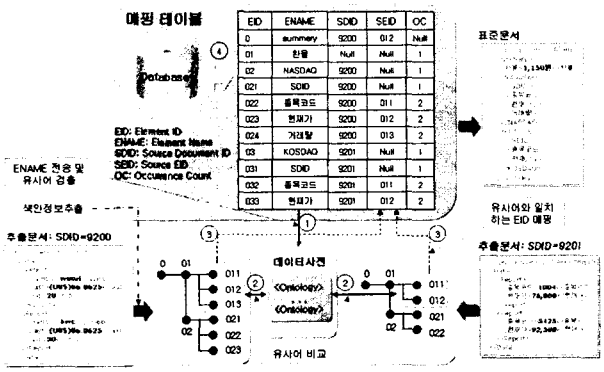
추출된 XML문서를 표준문서에 매핑하여 저장할 경우 각각의 데이터 색인정보와 구조 색인정보, 데이터 사전이 필요하다. 매핑 테이블 작성기의 역할은 표준문서에서 엘리먼트 이름의 유사어를 데이터 사전에서 추출하여 추출된 XML문서의 엘리먼트와 비교하여 매핑한다. 추출된 문서는 고정된 엘리먼트의 크기로 반복되기 때문에 부모 EID가 "011"인 엘리먼트에 한정하여 검색함으로서 빠르게 해당 엘리먼트로 접근할 수 있다. 또한 반복되는 회수(OC) 정보는 "01"이라는 EID의 형제 엘리먼트 수를 지정하면 된다. [그림4]는 두개의 문서를 통합하기 위해 매핑 테이블을 작성하기 위한 과정이다.

①매핑 테이블에서 해당 ENAME의 유의어를 데이터 사전에서 추출한다.

②추출된 유의어와 동일한 이름이 있는지 수집된 문서와 비교한다.

③일치하는 엘리먼트의 구조정보 및 데이터 색인정보를 매핑 테이블에 전송한다.

④ 매핑된 결과를 테이블을 저장한다.



[그림4] 매핑 테이블 작성기에 의한 매핑테이블 작성

3.5. 비교분석

본 논문에서는 구조화된 XML문서를 통합하고 효율적으로 관리하기 위해 DOM이나 SAX와 같은 표준 인터페이스를 이용하지 않고 색인기법과 데이터 사전을 이용하여 매핑하였다. [표1]은 여러 개의 XML문서를 비교 및 통합할 경우 제안된 기법과의 차이점을 나타낸다.

	DOM	SAX	제안된 매핑기법
접근기반	객체기반	이벤트 기반	색인정보 기반
임의접근	가능	불가능	가능
메모리 사용량	높음	낮음	낮음
검색효율	느림	느림	빠름

[표1]매핑기법의 차이점

따라서 여러 개의 XML문서를 통합하는 경우 제안된 기법이 더 효율적임을 알 수 있다.

4. 결론 및 향후 연구방향

본 논문에서는 사용자의 요구에 맞게 분산된 데이터를 이용하여 통합적인 결과를 제공하고, 데이터를 공유함으로서 개발자 및 사용자에게 편리함과 동시에 효율성을 증가시킬 수 있는 색인기반의 매핑 시스템을 설계하였다. 매핑 시스템은 각각의 문서에 대한 색인정보를 추출함으로서 문서전체를 메모리에 로드하지 않기 때문에 메모리 낭비를 줄일 수 있으며, 색인정보를 이용하여 임의접근이 가능함으로서 비교 및 검색효율을 증가 시킬 수 있다. 그러나 제안된 매핑기법에 있어서 비교기준은 엘리먼트 단위로 비교했기 때문에 엘리먼트가 속성을 포함하는 경우 비교할 수 없는 문제점과 데이터 사전은 학습에 의한 유사어 정의가 아니라 사용자가 사전에 조사하여 정의함으로서 많은 수작업이 요구되는 단점을 갖고 있다. 따라서 비교기준 및 데이터 사전에서 의미 정의 방법에 관한 연구가 필요하다.

참고문헌

[1] N. Kushmerick, Gleaning the Web, IEEE Intelligent Systems, vol.14, no.2, pp. 20-22, 1999.

[2] Patrick Cauldwell, Vivek Chopra et al , "Professional XML Web Services", Wrox Press, pp. 77-87, 2001 9.

[3] G. Gardarin, A. Mensch, T. Tuyet Dang-Ngoc, L. Smit, "Integrating Heterogeneous Data Sources with XML and XQuery.", Proceedings of 2002 13th IEE(DEXA'02) Workshop, Page 839-846, 2002

[4] Wustner, E., Hotzel, T., Buxmann, P., "Converting business documents:a classification of problems and solutions using XML/XSLT", Advanced Issues of E-Commerce and Web-Based Information Systems(WECWIS 2002), Page 54 -61, 2002.

[5] Y.K. Lee., "Index structures for structured document.", Proceedings of the first ACM international conference on Digital libraries, Page 91-99, 1996. 4.

[6] C. Petrou, S. Hadjiefthymiades, D. Martakos, "An XML-based, 3-tier scheme for integrating heterogeneous information sources to the WWW", pp , 1999

[7] 설진안, 정계동, 최영근, "유사구조를 갖는 XML 문서의 재구성을 위한 점진적인 시스템 설계.", 한국정보처리학회지 하권, Page 3-9, 2003. 5.