XML 문서 관리 시스템의 순환적 DTD 구조 저장 기법 및 질의 변환 전략

김정은*, 신판섭*, 정헌석*, 이재호**, 임해철*
*홍익대학교 컴퓨터공학과
**인천교육대학교 컴퓨터교육과
{jekim, psshin, hschung, lim}@cs.hongik.ac.kr, jhlee@mail.inue.ac.kr

Method of storing nested DTD structure and Query translation strategy in XML Repository system

JeongEun Kim*, PanSeop Shin*, Hunsuk Chung*, Jaeho Lee**, HaeChull Lim*
*Dept. of Computer Engineering, Hong Ik University
**Dept. of Computer Education, Inchon National University of Education

요 약

XML은 문서의 구조를 독립적으로 작성할 수 있어 문서의 체계적인 구조화가 가능하다. 이러한 이유로 최근, XML 문서를 구조화하여 데이터베이스에 저장, 관리하는 XML 문서 관리시스템 연구가 활발하다. XML문서 관리 시스템은 XML의 구조 정보를 효과적으로 표현하기위해 여러 가지 기법을 사용하고 있다. 그러나, 기존의 방법들은 XML문서 구성에 따라 저장 스키마가 유동적이거나 문서 정보 검색의 제약을 가지고 있을 뿐만 아니라, DTD의 문서 구조가 순환 관계와 같이 복잡한 형태를 지닐 때, 그 구조를 적절히 반영하지 못하거나, 구조를 반영하더라도 검색 시, 모든 요소를 순차적으로 탐색해야 하는 등의 문제점을 지니고 있다. 따라서 본 연구에서는 XML의 내용이나 구성에 영향받지 않는 저장 스키마를 설계하고 정보검색의 제약을 해결가능한 경로 정보를 제안한다. 또한 순환 관계를 갖는 DTD의 구조 정보를 비 순환 구조 부분과 순환 구조 부분으로 분리, 정의하고 질의처리 시, 입력되는 XML-QL을 SQL로 변환하기 위하여 XML-QL의 패턴을 분류하고 이에 따른 중간 단계의 SQL을 정의하여 질의어 변환기법을 제안한다.

1. 서 론

XML은 HTML에 비해 문서의 내용과 구조, 표현 형식, 링크정보를 독립적으로 작성할 수 있어 문서의 좀더 완벽하고 체계적인 구조화가 가능하다. 따라서, 이러한 XML 문서를 데이터베이스와 연동하여 체계적으로 저장, 검색 및 관리하는 XML 문서 저장 시스템의 연구가 활발히 진행되고 있다. 최근까지의 XML문서 저장 시스템 연구를 정리하면 다음과 같다. 첫째, XML의 모델링 및 질의어를 설계하는 데이터 표현 표준으로써의 XML 연구, 둘째, 데이터 교환 포맷으로의 XML 연구, 셋째, XML의 분산 저장 기법에 대한 연구[1]로 나누어진다.

지금까지 일반적인 XML 모델링 연구는 관계형 데이터베이스를 후위에 적용한 저장 시스템 연구가 주류를 이루고 있다. 저장 시스템 설계 시, 가장 중요한 고려 사항은 XML 문서 내용과 구조 정보의 손실 없는 저장과 검색을 들 수 있는데, 기존 XML 문서 저장 시스템에서는 내용 정보를 저장하기 위해 DTD 의존적 스키마를 사용하거나, DTD 독립적 스키마 생성 방법에 구조 정보를 저장하기 위한 경로 정보나 id, 위치정보 등을 적용하였다. 먼저, DTD 의존적 스키마 생성 방법은 DTD에 정의된 요소의 구조에 따라 스키마를 동적으로 생성하는 방법으로 구축 비용이 많이 들며, 생성되는 테이블의 수가 가변적이므로 검색 시, XML 질의어의 SQL 변환을 정형화시키기 어렵고 XML 문서의 내용 정보가 중복되는 단점이 있다. 반면 DTD 독립적 스키마는 스키마 형태를 미리 정의하고 DTD 정보를 테이블의 인스턴스로 삽입하는 방법으로, 테이블 수가 고정되어 있어 DTD 의존적 스키마 생성 방법에 비해 검색 시 테이블 조인 수가 줄어드나 구조적 검색이 쉽지 않다는 단점이 있다.

본 연구는 한국과학재단 특정기초연구과제 (과제번호 : 98-0102-09-01-3)의 지원을 받았음.

그리고, 기존 XML 문서 구조의 표현 방법으로 경로 정보를 사용하는 방법은 DTD를 참조하여 엘리먼트나 애트리뷰트를 나열하여 구조 정보를 표현하는 형태로 특정 엘리먼트에 대한 검색시 직접 접근이 가능하여 항해 비용을 줄이고, DTD 구성 요소의 계층정보(부모, 자식)의 표현이 쉬운 장점을 가지고 있으나 DTD의 구조가 순환 구조일 경우에는 경로 정보가 무한정 늘어날 가능성이 있다. id를 이용하는 방법은 XML문서의 각 요소에 id를 부여하여 구조 정보를 나타내는 방법으로 복잡한 구조 정보를 표현하는 것이 용이하나 XML 문서 갱신 시 요소의 모든 id 갱신이 필요하고, 모든 검색에 대해 순차적인 접근만이 가능한 단점이 있다. 마지막으로 위치 정보를 사용하는 방법은 문서의 각 요소의 시작 위치와 오프셋 또는 종료위치를 지정하는 것으로, 문서의 갱신에 효율적이지 못한 단점이 있다.

따라서, 본 논문에서는 관계형 데이터베이스를 기반으로 내용 정보를 저장하기 위한 DTD 독립적 스키마를 제안하여 저장 구조를 정형화시키며, 구조 정보의 저장과 검색이 가능하도록 경로 정보를 사용한다. 그리고, 기존의 경로 정보로는 표현하기 힘들었던 순환 속성 구조를 효율적으로 저장하기 위한 확장 경로 정보를 제안하고 DTD 독립적 데이터 모델을 기반으로 한 XML 질의어의 정형화된 SQL 변환기법을 제안한다.

2. 관련연구

현재까지의 XML 문서 저장 시스템 연구를 살펴보면 저장 스키마 구성 방법과 XML 문서의 논리 구조 표현 방법을 적절히 융합하여 성능 개선을 수행하고 있는데, 대표적인 연구를 서술하면 다음과 같다.

[2,3]에서는 각 요소별로 테이블을 생성하거나 모든 요소들을 통합하

여 하나의 테이블을 구성하는 등 DTD 의존적인 스키마를 생성하여 내용 정보를 저장하고, 논리 구조를 표현하기 위해서 각 요소와 요소사이의 간선에 id를 부여하는 방법을 사용한다. 즉, <그림1>과 같이 paper와 reference사이의 순환 구조를 가지는 문서를 저장하기 위해서 paper 요소에서 reference요소로, reference요소에서 paper요소사이의 간선에 id를 부여하고 각 요소별로 시작간선 id와 종료간선 id값을 정의한다.

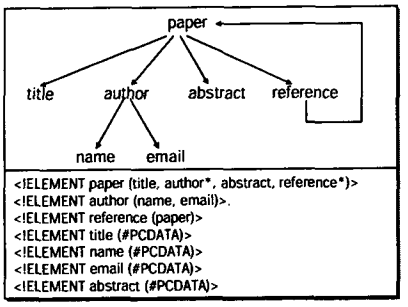


그림 1. Paper DTD와 DTD 그래프

그러고, [4]는 [2,3]처럼 구조 정보를 함해하여 *DTD* 의존적인 스키마를 생성하지만, 스키마의 에트리뷰트로 논리 구조 정보를 나타내기 위해 경로 정보의 변형된 형태를 사용한다. [4]에서 제시한 Shared 스키마를 구성할 경우 <그림1>의 구조를 저장하기 위해서는 **paper**(paperID, paper.title,isroot, paper.title, paper.abstract, paper.reference.isroot),author(authorID, author.parentID, author.name.isroot, author.name, author.email.isroot, author.email)의 스키마 구조가 생성된다. 종합해 보면 [2,3,4]는 XML 내용을 저장할 스키마가 *DTD* 의존적이므로 스키마 생성 비용이 큰 단점이 있다.

그리고, [5]에서는 DTD 독립적인 스키마를 사용하여 XML 문서 내용을 저장하며 경로 정보와 위치 정보를 이용하여 구조 정보를 표현하는데, <그림1>의 순환 구조를 나타내기 위해서는 저장 시 순환에 관계된 모든 요소를 나열해야만 한다. 즉, 순환이 3번 반복된다면 경로 정보가 paper.reference.paper.reference.paper.reference식으로 표현되어야 하므로 [5]에서 정의된 경로 정보의 표현력만으로는 복잡한 순환 구조를 나타내는데 한계가 있다.

저장기법 연구와 더불어 중요한 것은 XML 문서에 대한 질의어 연구는 질의어 설계연구와 XML질의어의 변환연구로 분류된다. **첫째**, 지금까지 설계된 질의어로는 XML-QL, LOREL, XML-GL, XSL, XQL등이 있는데, 이들 질의어를 [6,8]에서 제시 된 XML 질의어의 요구항목(선언적 질의형태, 다양한 표현력, 사용의 용이성 등)에 따라 비교, 분석하면 관계형 데이터베이스에서는 XML-QL의 사용이 가장 적절하다.

둘째, 기존의 XML 질의어 변환 연구는 [5]처럼 DTD 독립적 스키마 생성 방법을 사용하여 XQL을 SQL 변환시키는 방법과 [7]처럼 DTD 의존적 스키마 생성 방법을 사용하여 XML-QL을 SQL로 변환시키는 방법 등이 있는데, 지금까지 서술한 저장 방법과 질의어 요구사항을 종합해 보면 검색 시, XML-QL을 사용한 SQL의 변환 형태를 정형화시키기 위해서는 고정된 수의 테이블을 구성하는 DTD 독립적 스키마를 사용하는 것이 가장 적절하다.

따라서, 본 논문에서는 구조적 검색이 용이하도록 경로 정보 방법을 사용하여 내용 정보를 저장하기 위해서 테이블 생성 수와 조인 수를 줄이기 위해 DTD 독립적 스키마 생성방법을 제안한다. 또한, 순환 구조를 가지는 복잡한 구조 정보의 표현을 가능하게 하여 기존의 경로 정보 방법의 한계를 극복한 확장 경로 정보 방법을 제시하고, 정형화된 XML-QL의 SQL 변환기법을 제시한다.

3. 저장 구조

본 장에서는 [9]에서 제시한 DTD 독립적인 관계 데이터 모델을 확장하고, 경로 정보 표현 방법을 기반으로 하여 순환 구조를 갖는 DTD의 구조 정보를 저장하기 위해 경로 정보를 이용한 확장 기법을 제안한다.

3.1 관계형 스키마 설계

XML 문서의 내용 정보와 DTD 정보를 분리하여 독립적 스키마를 생성하는 것은 테이블 속성정보와 테이블 수가 고정되어 있어, 검색 패턴을 분석하여 질의어 변환을 정형화시킬 수 있는 장점이 있다. 따라서, 본 연구에서는 내용 정보와 구조 정보가 독립적으로 저장되는 스키마를 제안하였다. 그 구성은 다음과 같다.

XML 스키마
- DTD_Table(DTD_id, DTD_name, DTD_content)
- DTD_Structure_Table(DTD_struct_id, DTD_id, path)
- DTD_Level_Table(DTD_level_id, DTD_id, element, level)
- Attribute_Table(Attr_id, DTD_id, Att_name, default)

DTD 스키마
- XML_Table(XML_id, DTD_XML_id, XML_name, add_info)
- Element_Instance_Table(Element_id, DTD_struct_id, XML_id, Instance_id, contents)
- Attribute_Instance_Table(Attr_Inst_id, Attr_id, XML_id, Instance_id, value)

3.2 확장 경로 정보

본 연구에서 제안하는 확장 경로 기법은 부분 순환 구조를 갖는 DTD를 경로 정보를 이용하여 표현하기 위해서 DTD 그래프를 비 순환 부분과 순환 부분으로 나누어 구조 정보를 저장하는 것으로, 제안 기법으로 <그림1>의 DTD를 표현한 것이 <그림2>이다.

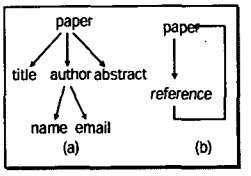


그림 2. 분할 DTD 구조

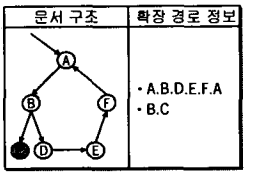


그림 3. 확장경로정보표현

(a)는 DTD의 요소들 중에서 트리 구조를 이루고 있는 부분만을 분리해서 표현한 것으로, (a)와 같은 구조정보는 루트요소에서부터 내용정보가 저장 될 수 있는 각 요소까지의 경로정보를 각 요소별로 구분하여 저장한다. 그리고, (b)는 DTD의 요소들 중에서 순환 구조를 이루고 있는 부분만을 분리해서 나타낸 것이다.

DTD가 순환 구조를 가지기 위해서는 내용을 갖지 않는 요소들간의 순환 관계가 성립되어야 한다. 따라서 본 연구의 순환 구조 내에는 순환 관계를 직접 형성하는 요소들과 그 요소들과 관계된 순환 구조내의 내용 정보를 저장하는 요소들이 포함되며, 경로 정보의 구성은 순환관계를 형성하는 간선의 구성 요소들의 나열로 표현되는데, 이를 도식화하면 <그림 3>과 같다.

다시 말하면, 확장 경로 정보 표현은 기존의 문서 구조 표현 방법 중 id를 이용하는 형태의 확장으로, 순환 구조를 가진 그래프의 특정 노드에서 시작하여 순환 요소들을 차례대로 나열하고 다시 시작노드를 명시하여 표현한다.

이와 같이 복잡한 구조의 문서 구조 정보를 순환 부분과 비 순환 부분으로 나누어 지정하여, 순환 부분에서는 기존의 경로 정보를 그대로 사용하여 구조적 탐색과 특정 위치의 요소에 대한 직접 접근이 가능하도록 하였고, 비 순환 부분에서는 간선을 이루는 요소들의 나열로 순환 관계를 표현함으로써 기존 경로 정보의 표현 방식의 제약을 해결하였다.

<그림1>의 DTD 구조를 본 연구에서 제안한 확장 경로 정보 표현을 사용하여 나타내면 <그림4>와 같다.

DTD_struct_id	path
1	paper.title
2	paper.author.name
3	paper.author.email
4	paper.abstract
5	paper.reference.paper

그림 4. 확장경로정보표현 예

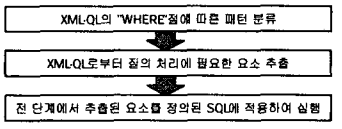
4. XML-QL to SQL 변환 전략

본 장에서는 앞서 제시한 DTD 독립적인 관계데이터 모델을 사용함으로써 관계형 데이터베이스에 저장된 내용에 대한 검색을 정형화한다.

XML-QL은 검색 조건을 명시하는 "WHERE" 절과 검색 결과를 XML형태로 재 구성하기 위한 "CONSTRUCT" 절로 구성되어 있고, 기본연산으로 선택, 축소, 추출, 재 구조화, 조합이 있다. 5 가지 기본 연산 중에서 본 연구에서는 XML-QL의 "WHERE"절의 선택, 추출, 조합 연산만을 고려하여 XML-QL to SQL의 변환 기법을 제안한다.

4.1 XML-QL 패턴 분류

본 연구에서의 XML-QL을 SQL로 변환하는 질의 처리 전략은 다음과 같다.



우선, XML-QL를 "WHERE"절의 형태에 따라 검색 타겟이 엘리먼트인 경우와 애트리뷰트인 경우 그리고, 태그 변수를 사용하여 질의를 하는 경우, 변수값을 사용하여 엘리먼트의 조인을 하는 경우와 regular-path 형식을 사용하는 경우에 따라 5 가지로 분류한다. 그리고, XML-QL로부터 SQL에 변환에 필요한 요소(문서명, 검색요소의 경로정보, 검색요소의 내용값 등)들을 추출하고, 미리 정의된 중간 단계의 실행 SQL 시퀀스에 이를 추가하여 완전한 SQL문을 생성한다.

본 연구에서 제안한 "중간단계의 SQL 시퀀스"는 <그림5>와 같다.

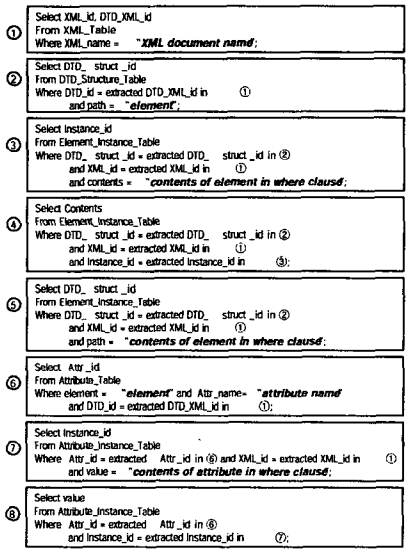


그림 5 XML-QL to SQL 변환 단계

<그림5>의 중간단계 SQL 시퀀스를 사용한 질의 처리 전략의 예를 들면, 타겟이 엘리먼트인 경우에는 먼저, XML-QL의 "WHERE" 절에서 문서명과, 엘리먼트의 경로정보, 조건값을 추출한다. 그리고, ①단계를 적용하여 전 단계에서 추출한 문서의 id와 dtd_id를 구하고, 추출된 엘리먼트의 경로정보를 이용하여 ②단계에서 경로정보 id를 구하고, 단계 ③에서 질의 시 제시되었던 조건값으로 검색하고자 하는 엘리먼트 인스턴스 id를 구하여 ④ 단계에서 검색하고자 했던 엘리먼트의 값을 구한다.

그 외에 검색 타겟이 애트리뷰트인 경우에는 ①~③ 단계에 ⑥~⑧ 단계가 추가되고 태그 변수를 사용하는 경우에는 ①~③ 단계에 ⑤단계를 추가한다. 그리고, 변수 값을 사용하여 엘리먼트를 조인할 때에는 ①~③ 단계를 적용한 후 검색 뷰을 생성하는 단계가 추가되고 마지막으로 ④단계를 사용하여 검색한다. 마지막으로, regular-path 표현을 사용할 경우에는 ①~④단계를 차례로 적용시킨 후 경로 정보를 계산하기 위한 모듈이 추가된다.

5. 결 론

기존의 관계형 데이터베이스 기반의 XML 저장 시스템 연구는 저장 스키마 구성 방법으로 DTD 의존적 스키마를 기반으로 id를 사용하거나 위치정보를 이용하여 구조 정보를 표현하거나 DTD 독립적 스키마를 사용하여 문서를 저장하고 경로 정보를 이용하여 구조 정보를 나타냈다. 전자의 경우는 순환 정보를 지니는 DTD를 저장할 수 있으나, 저장 테이블의 수가 가변적이고 특정 요소에 대한 탐색이 순차적으로 이루어져 검색 속도가 떨어지고 DTD 의존적인 스키마를 구성하므로, XML질의어의 SQL변환을 정형화시키기 어려운 단점이 있고. 후자의 방법은 경로 정보를 이용하여 구조 정보를 저장하는 방법으로, 문서의 갱신이 자유롭고 구조적인 검색이 쉽고, 질의어의 정형화 된 SQL변환이 가능한 장점이 있으나, 순환 구조를 갖는 복잡한 DTD에 대한 구조 정보의 저장이 어려운 단점이 있었다.

따라서, 본 논문에서는 DTD 독립적인 스키마를 사용하여 문서를 저장함으로써 저장 테이블의 수를 고정화시키고, XML 질의어인 XML-QL을 "WHERE"절의 패턴에 따라 5가지 형태로 분류하여 SQL 변환을 정형화 시켰다. 그리고, 확장 경로 정보 표현 방법을 제안하여 순환 관계를 갖는 DTD를 비 순환 부분과 순환 부분으로 나누어 비 순환 부분은 기존의 경로 정보를 이용하여 구조적 검색이 쉽게 이루어 질 수 있도록 하였고, 순환 부분은 순환 관계를 형성하는 간선의 구성 요소들의 나열로 표현하여 기존의 경로 정보 표현의 한계를 개선하였다.

6. 참고 문헌

[1] Stefano Ceri, Piero Fraternali, Stefano Paraboschi, "XML: Current Developments and Future Challenges for the Database Community" EDBT pp.3-17, 2000

[2] Daniela Florescu, Donald Kossmann, "Storing and Querying XML Data using an RDMBS" *IEEE Data Engineering Bulletin 22(3)*: pp.27-34, 1999

[3] Daniela Florescu, Donald Kossmann, and Ioana Manolescu, "Integrating Keyword Search into XML Query Processing" Proc. of the 9th WWW Conf. 2000

[4] Jayavel Shanmugasundaram, Kristin Tufte, Chun Zhang, Gang He, David J. DeWitt, Jeffrey F. Naughton, "Relational Databases for Querying XML Documents: Limitations and Opportunities" *VLDB 99*, pp.302-314, 1999

[5] Takeyuki Shimura, Masatoshi Yoshikawa, Shunsuke Uemura "Storage and Retrieval of XML Documents Using Object-Relational Databases" *DEXA99*, pp. 206-217, 1999

[6] Angela Bonifati, Stefano Ceri, "Comparative Analysis of Five XML Query Languages" SIGMOD Record 29(1) pp.68-79

[7] Daniela Florescu and Donald Kossmann, "A Performance Evaluation of Alternative Mapping Schemes for Storing XML Data in a Relational Database" INRIA TR, 1999

[8] Mary Fernandez, Jerome Simeon, Philip Wadler, "XML Query Languages: Experiences and Examplers" draft manuscript, communication to the XML Query W3C Working Group 1999.

[9] 김정은, 신관섭, 이재호, 임해철, "XML 문서를 위한 DTD 독립적인 데이터 모델 설계" pp.69-71 춘계정보과학회, 2000