# 관계형 데이터베이스 기반 색인을 이용한 XML 데이터의 저장 기법

손대준° 정병수
경희대학교 컴퓨터공학과
e-mail: thsdowns@dblab.khu.ac.kr

# Ordered Indexing Technique for Storing XML Data Using Relational Databases

Dae-Jun Son° Byeng-Soo Jung
Dept of Computer Engineering, Kyung-Hee University

**요 약**

인터넷의 급속한 발전으로 인해 다양한 종류의 데이터들이 증가 하게 되었으며 이러한 데이터의 표현과 데이터 교환을 위해 XML이 사실상의 표준으로 빠르게 자리 잡아 가고 있다.XML문서를 데이터로 저장 시 오랜 기간에 걸쳐서 성숙된 RDBMS를 사용하여 XML데이터를 RDBMS로 저장 시 발생할 수 있는 단편화 방지와 XML질의의 효과적인 질의 처리에 대한 많은 연구들이 제안되었다. 본 논문에서는 XML 문서를 관계형 데이터베이스 스키마로 저장 시 발생할 수 XML문서내의 엘리먼트의 관계에 대한 정보를 저장하기 위해서 추가적으로 발생하는 정보의 수를 줄이고 엘리먼트간의 관계를 효과적으로 저장할 수 있는 방법에 대해서 연구한다.

## 1. 서론

인터넷의 급속한 발전으로 인해 다양한 종류의 데이터들이 증가 하게 되었으며 이러한 데이터의 표현과 데이터 교환을 위해 XML이 사실상의 표준으로 빠르게 자리 잡아 가고 있다.

XML 문서는 새롭고 자유로운 태그를 사용하고 고정된 스키마가 없는 점에서 반구조적 데이터(Semistructured data)의 성격을 가진다.

현재 많은 데이터들이 XML로 작성되고 있으며 이러한 대량의 데이터들에 대한 효율적인 관리와 검색, 저장에 관한 많은 연구들이 진행되고 있다.

XML 문서의 질의어로는 XPath, XQuery등이 제안 되었으며 이들 중 W3C(World Wide Web Consortium)에서 개발 중에 있는 XML질의 언어로서 다양한 형태의 XML데이터 소스에 폭넓게 적용할 수 있도록 설계되어 있는 XQuery가 표준으로 간주 되고 있다.

XML데이터를 저장하기 위해서는 XML전용 데이터베이스가 가장 적당하지만 고가의 도입비용과 검증되지 않은 안정성으로 인해 널리 사용되지 못하고 있는 상황이다.

본 논문에서는 대신 오랜 기간에 걸쳐 안정성을 인정받고 여러 분야에서 이미 사용하고 있으며 대량의 정보처리가 가능하고 회복, 동시성 제어 등의 성숙된 기술들을 가지고 있는 RDBMS를 사용하여 XML 데이터를 RDBMS로 효과적으로 저장하는 방법과 저장된 XML문서들의 정보를 효과적으로 검색할 수 있는 방법을 제안한다.

## 2. 관련 연구

XML문서들은 구조적 특성을 가지며 문서의 스키마 정보 없이 사용될 수도 있고 사용의 편리성으로 인터넷에서의 표준이 되어가고 있다.

```
<? xml version="1."?>
<PLAY play_num="S-001">
  <TITLE>
    The Tragedy of Hamlet,Prince of Denmark
  </TITLE>
  <ACT>
    <TITLE>ACT 1</TITLE>
    <SCENE>
      <TITLE>
        SCENE I. Elsinore. A platform before the castle.
      </TITLE>
      <STAGEDIR>
        FRANCISCO at his post. Enter to him BERNARDO
      </STAGEDIR>
    </SCENE>
  </ACT>
  ....
```

그림 1 XML 문서의 구조

XML 문서들에 대한 질의는 내용에 대한 질의뿐만 아니라 구조에 대한 질의도 가능해야 한다. XQuery와 같은 XML 질의어들은 정규경로식 형태의 질의를 취하기

때문에 이러한 구조를 처리하기 위해서는 XML 문서들이 평면적인 구조의 관계형 테이블에 저장될 때 문서의 내용정보뿐만 아니라 문서의 구조정보도 추출하여 관계형 테이블에 함께 해야 한다.

구조 질의 처리에는 저장된 구조 정보를 효율적으로 활용하는 SQL문으로 변환하는 것이 중요하다. 동일한 XML 문서와 질의라 하더라도 관계형 데이터베이스의 테이블들에 저장된 구조정보의 형태와 활용 방법에 따라 질의처리의 효율성이 다르기 때문이다.

### 2.1 저장기법

RDBMS를 이용해서 XML 문서나 데이터들을 저장하는 방법에는 크게 DTD를 이용하여 XML 문서를 저장하는 방법과 DTD를 이용하지 않는 방법이 있다[1-6].

DTD를 이용하는 경우에는 XML 문서에 대해서 기계적으로 DTD를 생성할경에 문서의 작성자의 의도와는 다른 DTD를 생성할 수 있기 때문에 이런 경우 DTD를 이용하여 데이터베이스 스키마를 생성할 때 의도하지 않은 결과가 초래된다. 따라서 본 논문에서는 XML 문서가 DTD를 제공하지 않는다고 가정한다.

위와 같은 방법을 이용한 XML 저장 시스템에 대한 기존의 연구들에서는 구조정보로서 각 노드간의 연결 정보, 노드의 경로 정보 그리고 노드간의 포함관계에 대한 영역정보를 이용하는 방법들이 있다[2],[3].

[4]에서는 노드간의 연결 정보로서 각 노드에 ID를 부여하고 부모 노드의 ID와 내용정보를 단일 테이블에 저장, 이용하였다.
이럴 경우에 정규경로식 질의에 대해서 경로식 만큼의 테이블에 대한 조인이 필요하게 되므로 경로식이 길어지게 될 경우에 조인의 수가 커지므로 결과적으로 성능의 저하를 초래하게 된다.

### 2.2 경로 문자열 저장

이러한 정규경로 질의의 처리 효율을 높이기 위해서 [2],[3]에서는 다음과 같이 문서상의 경로를 테이블로 구성하여 처리의 효율을 높였다.

표 1 경로 문자열 저장

| 경로 ID | 경로표현 |
|---|---|
| 1 | "/PLAY" |
| 2 | "PLAY/@play_num" |
| 3 | "/PLAY#/TITLE" |
| 4 | "PLAY#/ACT" |
| 5 | "PLAY#/ACT#/TITLE" |

### 2.3 엘리먼트 식별자 부여

기존의 연구들에서는 단순히 각각의 엘리먼트의 구분을 위해서 엘리먼트들에 고유한 식별자만을 부여하였지만 이와 같은 경우 XML문서에 대한 순서정보 같은 정보의 손실을 초래하게 된다. 이럴 경우에 검색을 위한 질의나 질의에 의해 선택된 값들에 대한 재조립 시에 기존의 XML 문서의 구조에 대한 정보에 대해서 추가적인 비용이 발생하지만 [1]에서는 넘버링 방법을 이용해서 테이블에 엘리먼트의 순서에 대한 정보를 값으로 저장하여 효율성을 높였다.

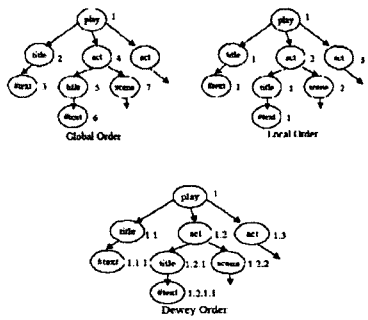


그림 2 XML 트리에 대한 번호 부여의 예

[1]에서는 XML 트리에 대해 번호를 부여하는 방법으로 세 가지 방법을 제안하였는데 hybrid 형태인 dewey order 방법이 일반적인 방법이며 효율적이라 할 수 있다.

### 3. 제안한 문서 저장 모델

XML 문서를 관계형 데이터베이스에 저장하기 위해서는 먼저 XML문서를 파싱하여 XML 트리를 생성하고 트리 정보를 바탕으로 데이터베이스 스키마를 생성한 후 테이블을 작성한다. 트리 작성 시 정보검색을 위해 XQuery를 사용하므로 W3C에서 정의한 7가지 노드 형식에 따라 XML문서를 모델링한다. .

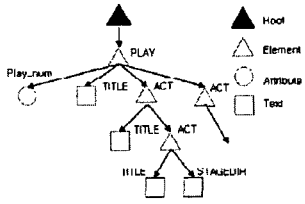


그림 3 그림 1의 XML Tree

[4]에서 제안한 edge테이블 방식은 정규경로식이 길어질 경우에나 특히 와일드카드(*)나 재귀적 내림연산자(//) 처리 시에 조인에 드는 비용이 커진다.
따라서 본 논문에서는 조인의 비용을 줄이기 위해서 [2],[3]에서 제안한 정규경로식에 대한 패스테이블을 이용한다.

"/PLAY/ACT[1]/SCENE/STAGEDIR"과같이 XQuery질의에 조건이 포함되어 있을 경우 [2],[3]에서는 결과 값으로 선택된 노드가 경로조건에 만족하는 노드인지를 계산하기 위해서 영역넘버링 방법(그림 4)을 이용한 θ-조인을 사용하여 선택된 노드의 적합여부를 판별한다. 하지만 경로상의 부모-자식 관계에서도 동등-조인 보다 효율이 떨어지는 θ-조인을 이용함으로서 질의처리의 효율이 떨어지게 된다.

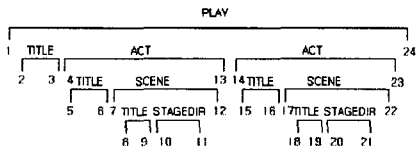


그림 4 XML 문서의 영역넘버링

이러한 문제점을 극복하고자 [7]에서는 XQuery를 분석하여 부모-자식 관계에는 동등-조인을 사용하고 조상-자손 관계에는 θ-조인을 이용하여 문제를 해결하고자 하였다. 하지만 이 경우 경로식에 분석에 추가적인 비용이 들고 조상-자손관계에는 여전히 θ-조인을 사용하는 문제점이 있다.

본 논문에서는 그림 3에 순서화된 색인번호를 [2],[3]에서 사용한 영역넘버링 값 대신 사용한다.

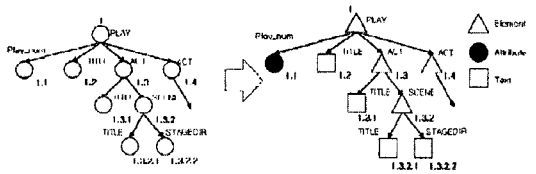


그림 5 순서번호를 XML 트리에 적용

영역 넘버링값 대신 색인번호를 사용함으로써 "/PLAY/ACT[1]/SCENE/STAGEDIR"와 같은 조상-자손관계에서도 조건질의에서도 포함관계를 찾기 위해서 θ-조인을 사용하는 것이 아니라 단지 순서화된 색인 값만을 비교하면 되고 [7]에서 사용한 노드의 관계분석에 대한 비용을 줄 일수 있다. 또한 스키마 구성 시에 문서의 구조를 위해 사용되는 속성 값들을 하나의 속성으로 묶음으로서 문서의 사이즈를 줄 일수 있는 장점도 있다. 그림 5에서 나타낸 트리를 기본으로 노드의 형태에 따라 다음과 같이 3개의 스키마로 구성하고 정규경로식 표현에 대한 질의의 효율을 높이기 XML 문서에 대한 경로 스키마를 추가 적으로 사용한다.

(표 2) 본 논문에서 제안한 테이블 스키마

| |
|---|
| Element(dewey_ID, docID, pathID,) |
| Attribute(dewey_ID, Document, pathID, value) |
| Text(dewey_ID, pathID, value) |
| Path(pathID, pathexp) |

본 논문에서는 [2][3]에서 XML문서를 관계형 데이터베이스에 저장시 XML 문서의 구조정보와 엘리먼트간의 관계를 유지하기 위해 추가 적으로 발생하는 데이터의 수를 줄이고 XQuery에 조건이 포함되어 있을경우 순서화된 색인방법을 사용하여 질의의 효율성을 높였다.

## 4. 결론 및 향후 연구 방향

본 논문은 XML문서의 데이터를 RDBMS로 효과적으로 저장하고자 하는 생각에서 출발 하였다.

따라서 XML을 RDBMS로 저장할 경우 서로간의 구조의 상이함으로 인해 발생될 수 있는 구조정보의 손실이나 테이블의 단편화 방지와 XQuery와 같은 정규경로식을 사용하는 구조기반의 검색어에 대해서 효율적인 검색을 지원할 수 있는 모델 방법을 제시하였다.

향후 연구 과제로는 XML 문서를 본 논문에서 제시한 관계형 데이터 베이스모델에 저장하였을 경우에 효과적으로 XQuery문을 SQL문으로 변환할 수 있는 알고리즘의 구현이 필요하고 이에 대한 쿼리 최적화에 대한 연구가 필요하다.

## 참 고 문 헌

[1] Igor Tatarinov, et al, "Storing and Querying Ordered XML Using a Relational Databases System" Proc, of SIGMOD , 2002

[2] H.Jiang, et al, "Path Materialization Revisited: An Efficient Storage Model for XML Data." Proc. of ADC , 2002.

[3] M. Yoshikawa, et al, "XRel : A Path-based Approach to Storage and Retrieval of XML Documents Using Relational Databases." ACM TOIT 1(1), 2001

[4] D. Florescu, et al, "Storing and Querying XML Data using an RDBMS." IEEE Data Engineering Bulletin 22(3),1999

[5] Jayavel Shanmugasundaram, et al, "Relational Databases for Querying XML Documents: Limitations and Opportunities."Proc. In Proc, of VLDB ,1999.

[6] D. Florescu, et al, "A performance evaluation of alternative mapping schemes for storing XML data in a relational database", In Proc, of VLDB ,1999

[7] 김대일 외, "XML 정규 경로식을 위한 유연한 질의 처리 시스템" , 정보과학회지 논문지 D , 2003