원본 XML 문서와 일관된 계층 구조를 가지는 축약본 생성 기법*

박상현, 하종우, 이정현, 이상근
고려대학교 정보통신대학 컴퓨터통신공학부
e-mail:{condols, okcomputer, jhbslpd, yalphy}@korea.ac.kr

XML Summarization Technique without Inconsistency in Hierarchical Structure

Sang-Hyun Park, JongWoo Ha, Jung-Hyun Lee, SangKeun Lee
Division of Computer and Communication Engineering, Korea University

요 약

XML이 다양한 환경에서 데이터 교환의 표준으로서 사용됨에 따라 보다 효과적인 데이터 교환을 위한 축약 기법이 요구된다. 본 논문에서는 기존의 경로 요약과 새로운 비트 시퀀스 기반의 색인을 통해 원본 XML 문서의 계층적인 구조 정보를 일관되게 유지하는 축약본 생성 기법을 제안한다. 실험을 통하여 계층적 구조 정보를 모두 유지하면서도 작은 크기의 축약본이 생성됨을 검증한다.

1. 서 론

XML은 데이터 저장 및 교환의 표준으로서 다양한 분야에서 사용되고 있다. 이렇게 XML의 활용도가 높은 것은 XML 문서가 계층적으로 이루어져 있으며 그 자체로서 데이터의 의미를 전달할 수 있는 특징을 지니고 있기 때문이다[1]. 그러나 XML 데이터의 특성상 부가적인 태그 정보를 내포함으로 인해 크기가 커지는 단점이 있다.

전통적으로 XML 문서의 축약 또는 압축을 통해 그 크기를 줄이면서 질의를 지원하기 위한 다양한 연구들이 이루어져 왔다[3, 4, 5]. 그러나 이러한 기존의 연구들에서는 지원하는 질의가 제한적이거나 축약을 위한 색인정보의 생성 및 유지에서 복잡한 알고리즘 혹은 자료구조를 요구하였다. 또한 무엇보다도 축약을 함으로써 원본 XML 문서의 계층적 구조를 손상시키는 문제가 발생하였다. 비록 XPath[2]와 같은 특정 질의 언어 대하여 원본 XML 문서와 그 축약본이 동일한 결과를 반환한다고 하더라도 원래의 계층적 구조를 완벽하게 복원하는 것이 불가능하다면 실제 응용에서의 적합성을 보장할 수 없는 것이다. 본 논문이 기여한 바는 아래와 같이 요약될 수 있다.

- 원본 XML 문서의 계층적인 구조정보를 일관되게 유지하면서도 축약이 가능한 새로운 색인 기법을 제안한다.
- 축약된 XML 데이터와 원본 XML 데이터의 크기 비교를 통해 제안 기법이 효과적임을 보인다.

2. 관련연구

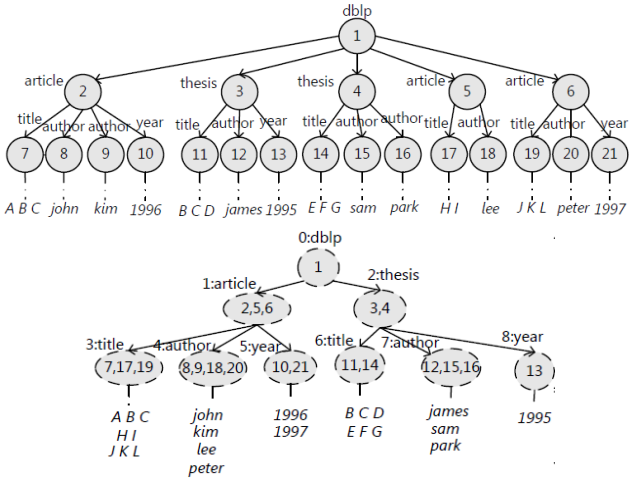
전통적으로는 DataGuide에서 XML과 같은 반구조적 데이터의 경로 요약 기법을 제안하였다[3]. 그러나 DataGuide의 경우 원본 문서의 계층적인 구조정보가 유지되지 않는 제약이 존재하였다.

XML 문서의 축약과 질의를 함께 고려한 가장 최근의 연구로서 BLAS가 있다[4]. BLAS의 핵심 아이디어는 기존 XPRESS[5]에서 제안한 reversed arithmetic 인코딩 기법을 기반으로 각각의 ReversedPath를 PLabel이라는 값으로 인코딩 함으로써 특정 Prefix를 가진 ReversedPath들을 B+ 트리상에 인접하게 모이도록 하여 질의 성능을 향상시키는 기법을 제안하였다. 이는 효과적인 질의 처리를 주요 목적으로 하여 원본 XML 문서를 재가공하여 색인을 구성하는 기법인 것이다.

기존의 이러한 연구와 본 연구의 차이점은 질의를 지원하기 위한 축약된 XML 문서의 색인이 아닌 유·무선을 통한 XML 문서의 전송 시 원본 문서의 계층적인 구조 정보를 모두 복원하는 것이 가능한 축약본 생성기법을 제안하는 것이다. 따라서 본 연구는 기존의 연구들이 기반으로 하는 시스템과의 XML 데이터 교환시 활용 가능한 기법이다.

3. XML 경로요약

DataGuide의 경우 경로 요약(path summary)을 활용하여 XML과 같은 반구조적 데이터의 축약기법을 제안하였다.



(그림1) 원본 XML 문서 트리 및 경로 요약본

* 이 연구에 참여한 연구자는 '2단계 BK21사업'의 지원비를 받았음.

그림 1은 원본 XML 문서와 그 요약본의 예를 보여주고 있다. 경로 요약본은 원본에 비해 그 크기가 줄어드는 장점은 있으나 계층 정보의 손실이 발생한다.

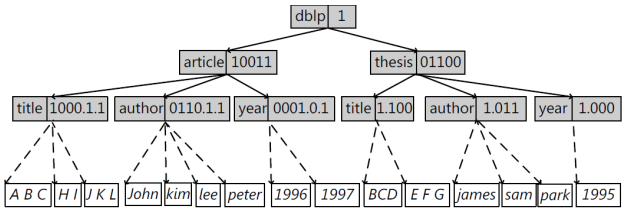
예를 들어 원본 문서의 첫 번째 깊이의 2번 *article* 노드의 경우 자식노드로서 두 개의 *author* 노드와 각각 하나의 *title*, *year* 노드를 가진다. 또한 5번 *article*의 경우 각각 하나의 *title*, *author* 노드만을 가진다. 그러나 경로 요약에 의해 축약하는 경우 *article* 노드가 각각 하나의 *title*, *author*, *year* 노드를 가지게 되어 원본 문서가 가지고 있는 계층적인 구조정보를 전혀 파악할 수 없게 된다.

3. 제안하는 XML 축약 기법

DataGuide의 장점을 활용하면서 계층적 구조정보를 유지하기 위한 새로운 비트 시퀀스 색인기법을 제안한다. 비트 1 (또는 0)은 노드의 존재 유무를 나타내며 비트의 위치를 통해 원본 문서에서의 노드간의 상대적인 위치를 파악하게 된다. 또한 원본으로의 복원을 위한 축약본의 트리 탐색은 넓이 우선 탐색을 취한다.

정의 1. *각 노드의 비트 시퀀스에 대하여 비트 시퀀스의 개수와 각 비트 시퀀스의 길이는 아래와 같이 결정된다.*

- *현재 노드 Nc와 부모노드 Np에 대하여 Nc의 비트 시퀀스의 개수는 원본 XML 문서에서 Np로의 동일한 레이블 경로의 수이며 각 부분은 점(.)으로 구분된다.*
- *비트 시퀀스의 각 부분에 대하여 만약 현재 노드와 동일한 형제노드가 존재한다면 비트 시퀀스의 길이는 전체 형제 노드의 수이다. 만약 동일한 형제 노드가 없다면 길이는 1이다.*



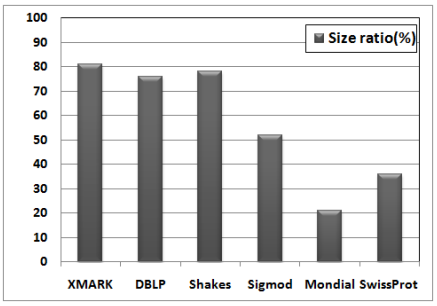
(그림 2) 비트 시퀀스 기반 원본 XML문서의 축약본

예를 들어 그림 2에서 *article*의 자식인 *author* 노드는 *0110.1.1*의 3부분으로 나누어진 비트 시퀀스를 가진다. 이는 그림 1의 원본 문서에서 *article*로의 동일한 레이블 경로가 3개가 존재하기 때문이다. 또한, 각 부분의 길이는 4, 1, 1이 되는데 비트 길이가 4인 경우에는 중복된 형제 노드가 존재하기 때문이다. 즉, 그림 1의 원본 문서에서 2번 *article* 노드의 경우 2개의 중복된 *author* 자식노드가 있으며 5번과 6번의 경우 자식노드 중 중복된 형제 노드가 없으므로 첫 번째 비트 시퀀스만 정의 1에 의하여 길이가 4가 된다.

따라서, 경로 요약을 기반으로 제안하는 비트 시퀀스 색인기법을 사용함으로써 축약본으로부터 원본의 계층적 구조정보를 정확하게 파악하는 것이 가능한 것이다.

4. 성능평가

제안 기법을 통한 축약본과 원본과의 크기를 비교하기 위해 원본에 대한 축약본의 상대적인 크기를 다양한 특성을 가진 실제 XML 데이터[6]에 대하여 측정하였다.



(그림 3) 원본 XML 문서에 대한 축약본의 크기 비율

그림 3에서 나타난 바와 같이 원본의 계층적 구조정보를 모두 유지하고 있음에도 불구하고 여전히 작은 크기의 축약본이 생성됨을 알 수 있다. 이는 경로요약 및 비트 시퀀스를 사용함으로써 원본에 비해 계층적 구조정보를 파악하기 위한 데이터의 크기가 감소되었음을 나타낸다. 이때, 중복된 레이블 경로가 많은 Mondial 과 SwissProt 과 같은 데이터의 경우 축약되는 비율이 매우 높음을 알 수 있다.

5. 결론

본 논문에서는 XML 기반의 정보교환 시 효율성을 높이기 위한 새로운 축약 기법을 제안하였다. 원본 XML문서의 계층정보를 기존의 경로 요약과 새로운 비트 시퀀스로서 표현함으로써 크기는 줄이면서도 원본의 구조정보를 모두 유지할 수 있으며 실험을 통해 원본보다 작은 축약본이 생성됨을 검증하였다. 향후 연구로서 방대한 XML 데이터의 스트리밍 환경에서 해당 기법을 활용하는 연구를 수행할 것이다.

참고문헌

[1] "Extensible Markup Language," http://www.w3.org/XML.

[2] "XML Path Language," http://www.w3.org/TR/xpath.

[3] R. Goldman and J. Widom, "DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases," *Proc. 23rd Int'l Conf., VLDB, 1997*, pp. 436 - 445.

[4] Y. Chen, S.B. Davidson, and Y. Zheng, "BLAS: An Efficient XPath Processing System," *Proc. 23rd ACM SIGMOD Int'l Conf., 2004*, pp.47-58.

[5] J.-K. Min, M.-J. Park, and C.-W. Chung, "XPRESS: A Queriable Compression for XML Data," *Proc. 22nd ACM SIGMOD Int'l Conf., 2003*, pp. 122-133.

[6] "XML Data Repository," http://www.cs.washington.edu/research/xmldataset/.