다중 채널상으로 XML 데이터 방송을 위한 비트 좌표 색인 기법*

박상현, 류병걸, 이정현, 이상근
고려대학교 정보통신대학 컴퓨터통신공학부
e-mail:{condols, smart123, jhbslpd, yalphy}@korea.ac.kr

Bit Coordinate indexing for Multi-channel XML Data Broadcasting

Sang-Hyun Park, Byung-Gul Ryu, Jung-Hyun Lee, SangKeun Lee
Division of Computer and Communication Engineering, Korea University

요 약

본 논문에서는 무선 방송 환경에서 XML에 대한 다양한 사용자 질의에 대하여 다중 채널을 통해 효과적으로 질의의 결과를 전송하기 위한 색인 기법을 고려한다. 이를 위해 서버측에서는 질의 결과뿐만 아니라 원본 XML상에서 질의 결과가 위치하는 계층 정보까지 파악이 가능한 비트 좌표 기반 색인 기법을 제안한다. 제안 기법의 시뮬레이션을 통해 다중 채널의 효과뿐만 아니라 색인으로 인해 빠른 응답시간을 가짐을 보인다.

1. 서론

XML은 데이터 표현 및 정보교환의 표준으로서 유무선 환경의 구분 없이 다양한 정보제공 시스템의 핵심 기술로 사용되고 있다. 이처럼 XML의 사용이 일반화된 것은 XML 문서가 계층적으로 이루어져 있어 그 자체로서 정보의 의미를 전달할 수 있는 특징을 지니고 있기 때문이다. 최근 모바일 기기의 성능 향상과 더불어 스마트폰이 대중화됨으로 인해 유선환경 뿐만 아니라 무선 환경에서도 XML을 통한 정보교환은 증가하고 있다.

유선환경에서의 XML 데이터 색인 및 질의처리 기법은 전통적으로 오랫동안 많은 연구가 이루어져 왔다[1]. 최근에는 그 영역이 무선 환경으로 확대되어 무선 네트워크를 통한 XML 데이터 전송, XML 데이터 무선 방송, on-demand 무선 방송 환경에서의 XML 데이터 전송 등 그 연구영역이 기존의 모바일 컴퓨팅 분야의 다양한 연구분야와 접목되어 점차 확대되고 있다[2][3][4]. 그러나 기존의 모든 연구들은 XML 스트림을 생성하고 전송하는 서버가 단일 채널만을 사용하여 스트림을 전송하는 것을 가정하고 있으며 필요에 따라 동적으로 가용한 다중 채널로의 XML 전송에 적용하는데 있어 제약점을 가진다.

2. 관련연구

무선 환경에서 XML 컨텐츠를 전송하기 위한 미들웨어인 Xstream은 XML의 특성을 고려하여 전송 스트림을 생성하기 위한 단편화(fragmentation) 기법 및 패킷화(packetizaion) 기법을 제안하였다[2]. 또한, 무선 방송 환경에서 XML 데이터 전송시 XPath 처리기법을 제안한 초창기 연구에서부터 최근에는 on-demand 무선 방송 환경에까지 연구 영역이 확대되었다[3][4].

기존 모바일 컴퓨팅 분야에서는 다중 채널을 통해 동일한 크기를 가진 데이터 아이템들을 전송하기 위한 방송 프로그램 생성에 관한 많은 연구가 진행되었으며[5][6] 최근에는 전송하고자 하는 데이터 아이템의 크기가 서로 다른 이질적인 환경을 위한 다중 채널 할당 및 방송 기법이 제안된 바 있다[7].

하지만 XML의 특성상 기존 모바일 컴퓨팅에서의 다중 채널에서의 전송기법을 동일하게 적용하는 것이 용이하지 않다. 기존에 무선으로 방송되는 각 데이터가 독립적었던 것과는 달리 XML의 경우 내부의 노드들이 계층적인 구조로 긴밀하게 연관되어 있는 특성을 지니고 있기 때문에 사용자는 전송받은 해당 데이터가 서버상의 원본 XML 문서상에서 위치하는 계층정보를 동일하게 파악할 수 있어야 한다.

본 논문에서는 XML 질의 언어로 표현되는 사용자의 요청에 대하여 다중 채널을 통해 무선으로 XML 노드들로 이루어진 스트림을 전송하고 사용자측에서는 스트림으로부터 결과값이 포함된 원본 데이터의 일부를 재구성할 수 있는 비트 좌표 기반의 XML 스트리밍 색인 기법을 제안한다. 본 논문이 기여한 바는 아래와 같이 요약될 수 있다.

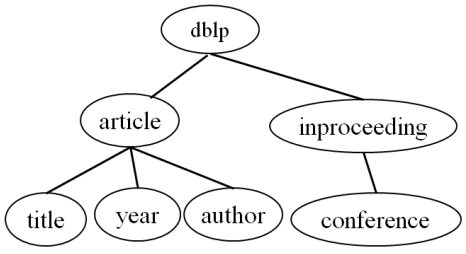
* 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 2009-0077925)

· 사용자가 요청한 질의에 대하여 원본 XML에 대한 질의의 결과값과 동시에 결과값이 원본 XML 상에서 위치하는 계층 정보까지 파악이 가능한 비트 좌표 기반 무선 XML 스트리밍 색인 기법을 제안한다.
· 질의 결과를 전송하고 이를 복원함에 있어 다중 채널로 인한 효과뿐만 아니라 색인의 효율성으로 인해 빠른 응답시간을 보임을 실험을 통해 검증한다.

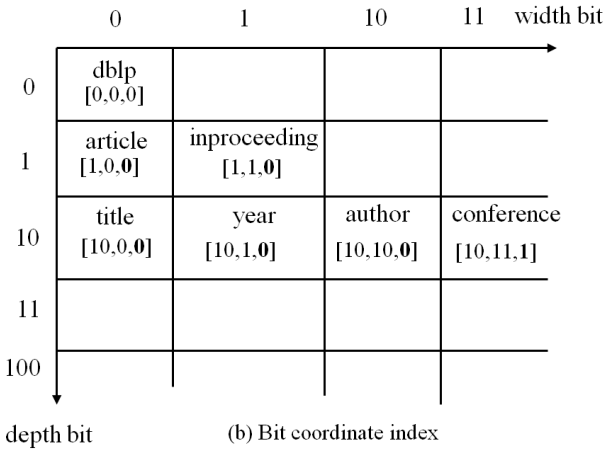
본 논문의 구성은 다음과 같다. 3장에서는 제안하는 비트 좌표 기반 XML 스트리밍 데이터 색인기법을 설명하고 4장에서는 다양한 사용자 질의 및 데이터에 대한 실험결과를 통해 제안기법을 검증한다. 끝으로 5장에서는 결론 및 향후연구에 대하여 기술한다.

3. 비트 좌표 기반 XML 스트리밍 데이터 색인

서버는 사용자들로부터 XPath 혹은 XQuery 로 기술된 질의를 통해 요청을 수신한다. 서버내에서의 질의 처리는 기존의 다양한 유선환경에서의 XML 색인 및 질의 처리 기법의 적용이 가능하다. 따라서 3장에서는 질의처리 결과를 사용자측으로 전송하기 위한 무선 방송 색인 및 전송 기법에 대해 기술한다.



(a) Sample XML data



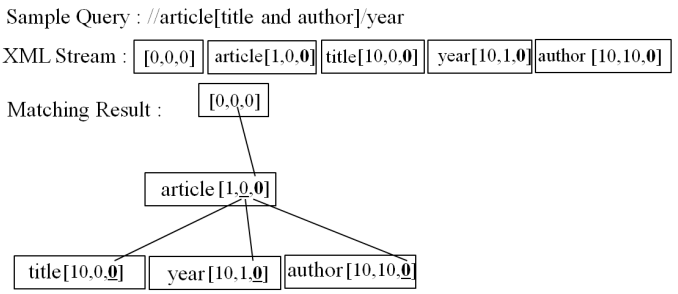
(b) Bit coordinate index

(그림 1) 예제 XML 데이터 및 비트 좌표 색인

서버에서는 원본 XML 문서의 각 노드들에 대한 비트 좌표 기반 색인을 작성한다. 먼저 그림 1(a)의 예제 XML 데이터에 대하여 SAX 이벤트 기반 깊이 우선 탐색을 통해 색인을 수행한다. 각 노드의 이벤트를 기반으로 비트 기반 좌표축에 노드의 좌표 할당을 수행하며 비트 좌표축에 위치하는 각 노드는 3개의 비트값인 [*depth bit, width bit, width bit of parent*] 을 가진다.

예를 들어 그림 1(b)에서 dblp 노드의 경우 루트노드로서 비트 좌표축의 [0,0]의 좌표에 기본적으로 매칭된다. 또한 article 노드의 경우 dblp 노드의 자식노드이므로 비트 좌표축에서 depth 비트를 하나 증가시켜 [1,0]의 좌표를 가진다. 이때 그림 1(b)에서 article 노드는 [1,0,**0**]으로서 세 번째 추가 비트인 '**0**'을 가지게 되는데 이는 dblp 노드의 width 비트에 해당하는 것으로 이를 통해 article이 dblp 노드의 자식 노드임을 식별할 수 있도록 한다. 앞의 예와 마찬가지로 conference 노드의 경우 자신의 비트좌표 [10, 11] 과 더불어 '**1**'의 추가 비트를 가지는데 이는 자신의 부모노드인 inproceeding 노드의 width 비트를 상속받아 최종적으로 [10,11,1]의 비트 좌표를 가지게 된다.*

예를 들어 그림 2의 예와 같이 사용자들의 질의중 하나가 //article[title and author]/year 라고 가정하자. 서버는 해당 질의에 대한 요청을 받은 후 원본 XML에서 해당 질의의 결과값을 도출할 것이다. 이때 서버에서는 article, title, author, year 노드들과 각 노드들의 원본 XML 상에서의 계층적 정보인 색인을 함께 전송한다. 이때 사용자는 article의 부모노드인 dblp 노드의 노드명은 필요로 하지 않으며 질의 결과값인 4개 노드와 동시에 결과값들이 원본 XML 상에서 위치하는 계층정보를 필요로 한다.



(그림 2) 예제 질의에 따른 해당 채널상의 XML 스트림 및 사용자 매칭결과

따라서 그림 2에서와 같이 XML 스트림을 구성하여 전송하게 되며 전송되는 스트림은 [0,0,0], article[1,0,0], title[10,0,0], year[10,1,0], author[10,10,0] 의 형태로 구성한다. dblp 노드는 노드명이 포함되지 않고 단순히 색인정보만을 전송하게 되는데 이는 사용자가 dblp 노드명은 필요로 하지 않으나 나머지 결과값들이 원본에서 위치하는 계층적인 위치를 파악할 수 있게 하기 위함이다. 이와 같이 한 사용자의 요청에 대하여 해당 응답 채널상에서 스트림을 구성함에 있어 사용자가 필요로 하는 최소한의 정보로 결과값을 전송한다.

* 실제 한 노드가 가지는 색인값은 10 11 1 과 같이 비트값들로만 구성되며 이때 서버와 사용자간에는 동일한 비트 좌표공간을 통해 매칭을 수행한다.

스트림을 수신한 사용자는 스트림상의 색인정보로부터 결과값과 동시에 계층정보를 복원하게 되는데 이때 각 노드의 색인 비트값들 중 자신의 부모노드 width 비트값 정보를 통해 노드간의 계층정보를 파악하여 질의 결과값을 포함한 원본 XML의 일부분을 복원한다.

4. 성능평가

제안한 기법의 성능평가를 위하여 제안된 스트림 색인 및 전송기법의 시뮬레이션을 수행하였다. 실험은 intel Core i5 CPU에 4GB 메모리를 가진 하드웨어 및 MS Windows 7 OS 상에서 Java 언어로 구현하였다. 실험에 사용한 XML 데이터는 각각 DBLP(130MB) 와 TreeBank(85MB)[8]를 사용하였으며 각 실험에 사용한 데이터의 통계 및 사용한 질의는 아래와 같다.

<표 1> 실험에 사용한 XML 데이터 및 질의

	Data size (MB)	Nodes (million)	Max./Avg. depth
DBLP	130	3.3	6/2.9
TreeBank	84	2.4	36/7.8

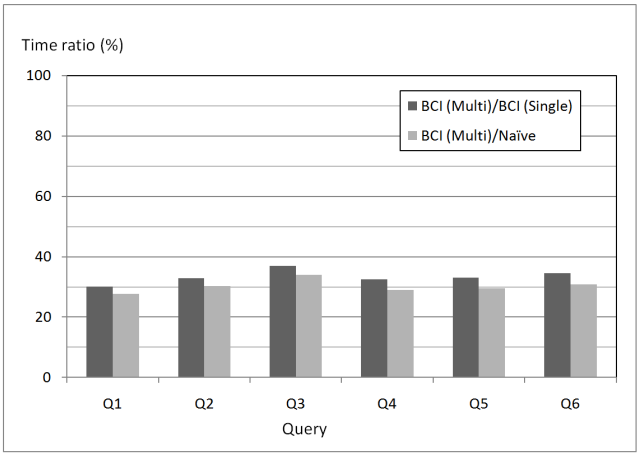
	DBLP (130MB)
Q1	/dblp/inproceedings[title]/author#
Q2	/dblp/inproceedings[title and booktitle]/author#
Q3	/dblp/inproceedings[title and booktitle and year]/author#

	TreeBank (84MB)
Q4	//S[.//VP[.//JJ][.//VBD]]//NP[.//WP]//DT#
Q5	//S//NP[.//IN][.//VBN]//JJ#
Q6	//S[.//NP[.//DT][.//NN]]//PP[.//TO]//NN#

각 XML 데이터에 대하여 3명의 사용자가 각자 질의를 통해 요청한다고 가정하였으며 각 질의에 대한 결과값을 비트 좌표 색인을 통해 스트림을 생성하여 각 응답 채널을 통해 전송한다. 모든 사용자가 질의에 대한 요청을 수행한 뒤 서버로부터 결과값이 도착하기까지의 대기 시간은 제외하였으며 응답시간은 최종적으로 스트림을 수신한 시점부터 사용자가 최종적으로 계층적 정보를 파악하여 스트림의 복원을 완료하는 시점까지로 측정하였다. 비교를 위하여 질의에 대한 결과값을 단일 채널상으로 전송하는 경우의 응답시간과 비교하였으며 이때 단일 채널로 전송되는 결과값은 세 질의의 결과값을 모두 포함하도록 한다.

위의 두 비교실험의 경우 XML 데이터의 특성을 고려하여 질의의 결과값을 전송한다. 그러나 이러한 XML의 계층적인 특성에 대한 고려가 없을 경우에 해당 질의가 요청하는 XML 문서 전체를 전송하는 극단적인 상황 또한 추가로 고려하였으며 이에 대한 비교실험을 수행하였다.

그림 3은 비교 대상이 되는 두 기법에 대하여 제안기법이 차지하는 응답시간의 비율을 측정한 결과이다. 그림에서 BCI(Multi)는 본 논문에서 제안하는 비트 좌표 기반 색인을 의미하고 BCI(Single)은 여러 질의의 결과를 단일 채널로 통합하여 함께 전송하는 경우이며 마지막으로 Naive는 전체 XML 데이터를 전송하는 경우를 나타낸다. BCI(Single) 및 Naive에 대하여 상대적으로 BCI(Multi)가 차지하는 응답시간 비율을 나타내었다.



(그림 3) 질의에 따른 제안기법의 응답시간 비율

그림 3에서 BCI(Multi)의 경우 상대적으로 BCI(Single) 및 Naive에 비하여 최대 40%미만의 응답시간을 나타냄을 알 수 있다. BCI(Single)의 경우 질의의 결과값들 사이에 중복이 존재하게 된다. 그러나 단일 채널로 전송된 스트림을 수신한 각 사용자가 스트림 상에서 자신의 질의 결과를 추출하는데 있어 다른 질의의 불필요한 결과값까지 읽게 되어 BCI(Multi)에 비하여 상대적으로 느린 응답시간을 보여주었다. 그러나 BCI(Multi)의 경우 질의의 결과가 아니지만 질의 결과와 연관된 계층정보를 가진 노드의 경우 해당 노드의 실제 노드명은 전송되지 않으며 단순히 비트값만을 전송한다. 따라서 결과값이 원본상에서 차지하는 계층정보를 파악함에 있어 불필요한 노드 정보를 읽지 않는다.

5. 결론 및 향후연구

본 논문에서는 무선 환경에서 여러 사용자 질의에 대하여 XML상의 질의 결과를 다중 채널상으로 전송하기 위한 비트 좌표 기반 색인기법을 제안하였다. 제안 기법은 결과값을 효율적으로 전송할 수 있을뿐만 아니라 원본 XML상에서의 결과값이 차지하는 계층적인 부분 정보를 복원할 수 있다는 특징을 가진다. 실험결과를 통해 제안기법이 단순히 다중 채널의 효과로 인한 잇점뿐만 아니라 색인의 특성에 의해 응답시간에서 효과적인 성능을 보임을 검증하였다.

향후 연구로서 사용자 질의의 패턴을 고려하여 다수의 사용자가 공통적으로 요구하는 데이터에 대한 응답시간을 보다 향상시키기 위한 연구를 수행할 것이다.

참고문헌

[1] G. Gou and R. Chirkova. Efficiently querying large xml data repositories: A survey, IEEE Transactions on Knowledge and Data Engineering, 19(10):1381-1403, 2007.

[2] E. Wong, A. Chan, and H. V. Leong. Xstream: a middleware for streaming xml contents over wireless environments, IEEE Transactions on Software Engineering, 30(12):918-935, 2004.

[3] Y. D. Chung and J. Y. Lee. An indexing method for wireless broadcast xml data, Information Sciences., 177(9):1931-1953, 2007.

[4] Y. Qin, W. Sun, Z. Zhang, P. Yu, and Z. He, Query-grouping based scheduling algorithm for on-demand xml data broadcast, Proc. 4th Int'l Conf. WiCOM, pages 1-4, 2008.

[5] C.H. Hsu, G. Lee, and A.L.P. Chen, An Efficient Algorithm for Near Optimal Data Allocation on Multiple Broadcast Channels, Distributed and Parallel Databases, 18(3):207-222, 2005.

[6] J.-L. Huang and M.-S. Chen, Broadcasting Dependent Data for Ordered Queries without Replication in a Multi-Channel Mobile Environment, Proc. 19th IEEE Int'l Conf. Data Eng., 2003.

[7] H.-P. Tsai, H.-P. Hung, and M.-S. Chen. On channel allocation for heterogeneous data broadcasting, IEEE Transactions on Mobile Computing, 8(5):684-708, 2009.

[8] XML Data Repository, http://www.cs.washington.edu/research/xmldatasets/