시그니처를 이용한 XML 문서 질의 기법

김우생
광운대학교 컴퓨터과학과

XML Document Search Technique by Signature

Woosaeng Kim
Department of Computer Science, Kwangwoon University
kwsrain@kw.ac.kr

요 약

인터넷의 성장과 함께 문서 교환의 표준으로 자리 잡은 대량의 XML 문서들을 효율적으로 검색하기 위한 방법이 필요하다. 기존의 방법은 주어진 질의에 답하기 위하여 모든 XML 문서들을 검색하기에 비용이 많이 든다. 따라서 본 논문에서는 시그니처 기법을 사용하여 주어진 질의와 관련된 일부 XML 문서들만을 검색하는 방법을 제안한다.

Ⅰ. 서 론

최근 들어 인터넷에서 데이터 교환과 정보 관리에 사용되는 XML 문서들을 효율적으로 질의하는 방법들이 필요하나, 기존의 방법들은 주어진 질의에 답하기 위하여 모든 문서들을 검색하므로 많은 비용이 든다.

따라서 본 논문에서는 방대한 양의 XML 문서들 중에서 질의와 관련된 일부 문서들만을 효율적으로 검색하는 방법을 제안한다. 이를 위해 문서와 질의에 시그니처 기법을 적용해 후보 문서만을 선택하여 질의를 수행하는 방법을 사용한다.

논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 시그니처 기법을 살펴본다. 3장에서는 제안하는 기법을 설명하고 4장에서는 결론 및 향후 연구에 대해서 언급한다.

Ⅱ. 관련 연구

시그니처는 특정 키워드를 포함한 문서를 빠르게 검색하기 하기 위한 방법으로 제안되었다. 먼저 텍스트 문서에 있는 각 키워드에 해시함수를 적용해 시그니처를 생성하고, 키워드 시그니처들에 bit-wise OR 연산을 수행하여 하나의 문서 시그니처를 만든다[1]. 다음으로 질의문의 키워드에 대한 시그니처를 생성하고 문서 시그니처와 bit-wise AND 연산을 수행한다. 그 결과가 질의문의 시그니처와 일치한다면 텍스트 문서가 그 키워드를 포함할 가능성이 높다. 근래에 XML 문서에 시그니처를 적용하여 질의 최적화, XML 경로간 노드 비교 최소화, XML 인덱스 구조 등이 연구되고 있다[2,3,4,5].

Ⅲ. 시그니처를 통한 XML 문서 질의

본 연구에서는 XPath 질의와 관련된 XML 문서들을 효율적으로 찾기 위해 문서와 질의에 시그니처 기법을 적용한다. 이를 위해 문서나 질의를 구성하는 각 노드에 대한 시그니처를 먼저 구해야 한다. XML 문서에 대응하는 XML 트리가 그림 1과 같을 때 각 노드의 시그니처는 표 1과 같다고 가정한다.

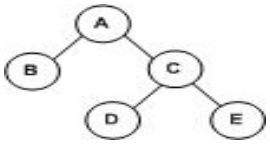

그림 1. XML 트리

표 1. 노드 시그니처

A	01010010
B	10001010
C	01001001
D	00101010
E	01000011

우선 문서의 경우, 문서 시그니처와 각 경로에 대한 시그니처를 구한다. 그림 1 문서 시그니처는 표 1의 각 노드에 bit-wise OR 연산을 수행한 11111011이 된다. 반면에 그림 1 문서의 경로 시그니처는 표 2와 같다.

표 2. 경로 시그니처

1	/A	01010010
2	/A/B	11011010
3	/A/C	01011011
4	/A/C/D	01111011
5	/A/C/E	01011011

다음으로 XPath 질의에 대한 시그니처를 구해 검색 대상 문서가 질의와 관련이 있는지를 조사한다. (1) 질의 시그니처와 문서 시그니처 간에 bit-wise AND 연산을 수행해 결과가 질의 시그니처와 같으면 문서가 질의에 포함된 노드들을 포함할 가능성이 있으므로 후보 문서로 간주하고, 만약 같지 않으면 질의 대상에서 제외시킨다. 예를 들어, XPath 질의가 /A/C라고 할 때 질의 시그니처는 01011011이 된다. 질의 시그니처 01011011과 그림 1 문서 시그니처 11111011의 bit-wise AND 연산의 결과는 01011011로 질의 시그니처와 같으므로 그림 1 문서는 후보 문서가 된다. (2) 각 후보 문서가 질의와 같은 경로를 포함하고 있는지를 검사한다. 이를 위해 본 연구에서는 질의에 대한 비교 대상이 많을 때 효율적인 비트 분할 시그니처 파일 기법을 이용한다[5,6]. 이 방법에서는 먼저 다음과 같은 방법으로 경로 시그니처로부터 비트 분할 시그니처를 구한다. 비트 분할 시그니처의 i 번째 행은, 각 경로 시그니처의 i 번째 비트로 이루어진 비트열이다. 즉, 경로 시그니처가 $N \times M$의 비트열 행렬이라면, 비트 분할 시그니처는 $M \times N$의 비트열 행렬이라 할 수 있다. 따라서 표 2의 경로 시그니처에 의해 표 3과 같은 비트분할 시그니처를 구한다.

표 3. 비트 분할 시그니처

1	01000
2	11111
3	00010
4	11111
5	01111
6	00000
7	11111
8	00111

질의 /A/C가 그림 1 문서의 경로에 포함되어 있는지는 다음과 같은 방법으로 구한다. 우선 질의 /A/C의 시그니처 01011011은 2,4,5,7,8 번째 비트가 1로 설정되어 있으므로, 표 3의 비트 분할 시그니처의 2,4,5,7,8 번째 행을 선택해 bit-wise AND 연산을 수행한다. 그 결과로 얻는 00111은 3,4,5 번째 비트가 1로 설정되어 있으므로, 표 2의 경로 시그니처의 3,4,5 번째 행이 후보 경로가 된다. 반면에 XPath 질의가 /E/A/B라면 질의 시그니처는 11011011이기 때문에 그림 1문서가 후보 문서로 간주된다. 그러나 질의 /E/A/B의 시그니처 11011011은 1,2,4,5,7,8 번째 비트가 1로 설정되어 있으므로, 표 3의 비트 분할 시그니처의 1,2,4,5,7,8 번째 행을 선택해 bit-wise AND 연산을 수행한 결과는 00000이다. 따라서 그림 1 문서는 /E/A/B 경로를 포함하고 있지 않음을 알 수 있다. 이와 같이 질의가 주어졌을 때, 문서 시그니처와 경로 시그니처를 통해 질의와 관련 있는 일부 문서들만을 선택해 효율적으로 질의를 수행할 수 있게 된다.

Ⅳ. 결 론

본 연구에서는 XML 문서와 질의에 시그니처 기법을 적용해, 질의와 관련이 있는 일부 XML 문서들만을 검사하는 방법을 제안하였다. 본 기법은 비트 분할 시그니처 파일 기법을 적용해 연산의 효율을 높였다. 향후 연구로는 시그니처를 통해 발생하는 허위 적중(false hit)의 문제를 해결하는 방법이 필요하다.

참고문헌

[1] C.Faloutsos, "Signature files: Design and Performance Comparison of Some Signature Extraction Methods," SIGMOD, 1985.

[2] 박상원, 박동주, 정태선. 김형주, "시그니처 기반 블럼 탐색을 통한 XML 질의 최적화 기법", 한국정보과학회 논문지: 데이터베이스 제 29권 1호, 2002.

[3] 임동혁, 정호영, 김형주, "OWL 질의 처리를 위한 시그니처 기반 최적화 기법", 한국정보과학회 논문지: 데이터베이스 제 32권제 6호, 2005.

[4] 장경훈, 황병언, "시그니처를 이용한 XML 경로간 노드 비교의 최소화", 한국정보과학회

가을 학술발표 논문집, 2011
[5] 강인선, 홍석진, 이태원, 이석호, "비트 분할 시그니처 화일을 이용한 XML 인덱스 구조", 한국정보과학회 가을학술발표 논문집, 2002.
[6] C.Falustsos and R.Chan, "Fast Text Access Methods for Optical and Large Magnetic Disks: Designs and Performance Comparison", VLDB, 1988.