제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

경로 구성 유사도를 이용한 비트맵 인덱싱 기반 XML 문서 인식 기법

이재민, 황병연
가톨릭대학교 컴퓨터공학과
e-mail : {likedawn, byhwang}@catholic.ac.kr

An Identifying Method of XML Document based on Bitmap Indexing using Path Construction Similarity

Jae-Min Lee, Byung-Yeon Hwang
Dept. of Computer Engineering, Catholic University of Korea

요 약

XML의 대표적인 특징은 기존의 다른 컨텐츠와는 달리 문서의 구조를 기술할 수 있다는 것이다. 구조적 정보는 활용 방법에 따라 XML문서의 다양한 처리에 있어 성능을 향상시키는 핵심적인 요소가 될 수 있다. 그러나 XML 태그의 자기 서술적인 특성에서 비롯되는 구조적 표현의 차이는 오히려 문서의 식별을 어렵게 하는 원인이 된다. 본 논문에서는 기존의 비트맵 인덱스(Bitmap Index)를 이용한 XML 문서 검색 시스템이 다양한 구조적 유사성을 판별할 수 없는 단점을 보완 가능하도록 경로 중심의 유사 문서 인식 기법을 제안한다. 이 기법은 '경로 구성 유사도'와 '유사 경로 테이블'을 통해 기존의 비트맵 인덱스가 갖는 유사 경로를 인식하지 못하는 단점을 해결하고 검색의 유연성을 부여함으로써 보다 양질의 검색 결과를 도출할 수 있다. 또 이것은 기존 시스템의 Bit-wise 연산에 완전히 이식됨으로써 비트맵 인덱스의 장점인 빠른 성능을 그대로 유지할 수 있게 된다.

1. 서론

최근 웹 컨텐츠는 많은 수가 XML로 대체되어 가고 있다. 이것은 XML의 반구조적인 정보를 기술할 수 있는 특징에 기인한다고 할 수 있다[1]. 기존의 웹 문서나 컨텐츠들은 그 표현에 중점을 두고 있었으므로 시스템은 주로 전문에서 추출한 요약문을 사용해 정보를 저장하고 문서나 컨텐츠에 순위를 매기는 방법을 사용하여 질의에 대한 처리를 수행하였다. 한편, XML의 등장으로 시스템은 데이터베이스의 테이블에서 정보를 추출하듯이 문서의 구조가 되는 태그에서 정보를 추출하는 것이 가능하게 되었다. 그러나 XML태그의 자기 서술적 기능은 문서의 작성자가 임의로 태그를 정의하고 그것을 구성할 수 있게 함으로써 시스템이 유사한 구조를 인식하고 처리해야 하는 과제를 남겨주게 되었다[2].

기존의 비트맵 인덱스(Bitmap Index)[3]는 각각의 XML 문서에게 부여된 문서ID와 모든 문서들의 태그에 부여된 각각의 경로ID를 통해 검색을 수행한다. 이는 Bit-wise 연산을 통해 뛰어난 성능을 보장하지만 검색에 경로ID를 사용함으로써 유사한 경로에 대한 인식이 불가능할 경우에 유사한 구조를 갖는 문서를 인식할 수 없는 문제점을 갖게 된다. 이에 본 논문에서 제시하는 XML 문서 인식 기법은 유사한 경로를 식별할 수 있는 경로 구성 유사도를 설계하고 이것을 통해 구성된 유사 경로 테이블을 사용해 검색에 유연성을 부여하는 방법을 제시한다. 이 기법은 기존 시스템의 문제점을 해결하고 더 융통성이 있는 양질의 검색 결과를 제공할 것이다. 그리고 이를 Bit-wise 연산에 완전히 이식 가능하게 함으로써 기존 시스템의 장점인 빠른 성능을 그대로 유지하도록 한다.

2. 관련연구

비트맵 인덱스는 같은 경로를 갖는 비율이 높은 문서들을 묶어 놓은 문서ID와 경로ID를 축으로 하는 Bit-wise 연산 가능한 필드로 구성된 2차원 배열이다. 문서ID는 인덱스를 구성하는 모든 XML 문서들 각각에게 주어진 식별자이며 경로ID는 문서내에 존재하는 내용을 갖는 태그의 전체 경로들 각각에게 부여되는 고유한 식별자이다. 경로ID를 갖는 경로는 ePath(요소 경로)라 하며 태그의 내용을 제외한 모든 조상 태그들의 조합된 이름을 사용한다.

어떤 ePath 가 비트맵 상의 거의 모든 문서 내에 존재한다면 "ePath는 대중적이다"고 말하며 대중성(popularity)을 표현하는 식은 다음과 같다.

$$pop(pi) = |pi| / |[pi]$$

|pi|는 비트맵 pi의 크기, 즉 0 값이 아닌 값을 가지는 bit의 개수를 의미하고, [pi]는 비트맵 pi의 차수(cardinality), 0값을 가지는 bit의 개수와 0값이 아닌 값을 가지는 bit 개수의 합을 가진다. $pop(pi) \geq n$ ($0 \leq n \leq 1$,주어진 실수)이라면 pi 는 n-popular bit라고 하고, $pop(pi) \leq m$ 일 때는 m-unpopular bit

($0 \le m \le 1$,주어진 실수)라고 한다. 대중성을 위한 임계치(threshold)가 n($0.5 \le n \le 1$)으로 주어졌을 때, ePath들은 다음과 같이 세 가지 경우로 나뉠 수 있다. (1) $pop(pi) \ge n$ 인 경우 "popular bits"라고 한다. (2) $pop(pi) \le 1-n$ 인 경우 "weakening unpopular bits"라고 하며 (3) $1-n < pop(pi) < n$ 인 경우 "strengthening unpopular bits"라고 한다.

클러스터 혹은 문서의 중심(center)은 분할된 XML 문서들의 집합 안에서 popular bits와 strengthening un-popular bits가 1로 구성되고 weakening unpopular bits가 0으로 구성된 비트벡터(bit vector)이다.

xOR 가 exclusive or 연산이라고 할 때, 두 문서들 사이의 거리는 다음과 같다.

$$dist(di, dj) = \left| xOR(di, dj) \right|$$

두 개의 문서나 혹은 비트맵에서의 행들간의 유사도(Similarity)는 다음과 같다.

$$sim(di, dj) = 1 - |\, xOR(di, dj) \,| \,/\, MAX([di],[dj])$$

$sim(di, dj) > \xi$ ($0 \le \xi \le 1$, 주어진 실수)라면 문서 di, dj 는 ξ -similar라고 한다.

비트맵 인덱스를 기반으로 하는 XML 문서 검색 시스템인 BitCube는 XQEngine, XYZFind와 같은 동종의 검색 시스템과의 실험을 통해 그 성능이 입증되었으며 문서의 수가 증가하는 것에 따르는 성능 저하에도 매우 안정적이라는 것이 증명되었다.[3,4]

그러나 경로를 중심으로 문서의 클러스터링을 수행하는 비트맵 인덱스는 완전히 일치하는 경로에 대한 인식은 가능하지만 같은 의도로 작성된 경로가 갖는 구조적 유사성을 식별하지 못함으로써 정확한 클러스터링을 기대하기 어렵다. 그림 2.1은 비트맵 인덱스의 예이다.

PathID / DocID	1	3	4	7	9	…	N
1	1	1	1	1	1	…	1
2	1	0	1	1	1	…	0
4	0	1	1	0	1	…	1
5	0	0	1	1	0	…	1
7	1	0	1	1	0	…	0
Pop.	0.6	0.4	1	0.8	0.6	…	0.6
Cen.	1	0	1	1	1	…	1

Pop.≡ 대중성(popularity), *Cen.*≡ 중심(center)

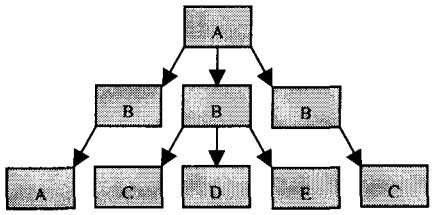
그림 2.1 비트맵 인덱스

3. 경로 구성 유사도

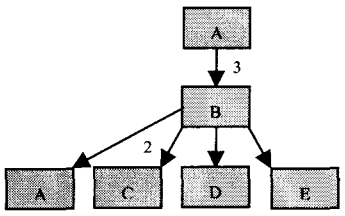
3.1 경로 축약 및 경로 추출

XML문서는 각각의 노드가 동일한 태그를 가질 수 있다. 그러나 비트맵 인덱스는 삽입될 XML문서 태그의 발생 횟수가 유사도 계산에 영향을 미치지 않는다. 그러므로 기존 경로 추출 과정에 경로 축약을 사용한 발생 빈도 수 측정을 통해 구한 확장된 경로 추출 테이블을 3.3절의 '비트맵 인덱스와의 이식'에서 사용한다.

경로 축약은 XML 문서 태그 트리의 가장 상위의 루트 노드부터 시작하여 BFS(너비 우선 탐색)와 같은 순서로 탐색을 진행한다. 해당 노드가 두 개 이상의 같은 이름을 갖는 자식 노드를 갖는다면 같은 이름을 갖는 모든 자식 노드의 수를 기록한다. 그리고 자식 노드들 중에서 임의의 하나를 대표 자식 노드로 선정하고 나머지 자식 노드들의 자손을 대표 자식 노드에 상속한 뒤에 삭제한다. 같은 연산을 상위 노드에서 하위 노드로 순차적으로 반복하여 수행한다. 그림 3.1의 (a)는 경로 축약을 거친 후에 그림 3. 1의 (b)와 같이 된다.



(a)XML문서의 태그 트리(축약 전)



(b)XML문서의 태그 트리(축약 후)

그림 3.1 XML 문서의 경로 축약

XML문서 안에 있는 태그 중에서 내용을 담고있는 태그를 대상으로 해당 태그의 전체 경로와 경로의 발생 빈도로 구성된 경로 추출 테이블을 구성한다. 예를 들어 그림 3.1과 같은 태그 트리를 갖는 문서가 태그 B, C 그리고 D에 내용을 가지고 있다면 표 3.1과 같은 경로 추출 테이블이 구성된다.

표 3.1 추출된 XML문서의 경로 테이블

경로명	발생 빈도
A.B	3
A.B.C	2
A.B.D	1

3.2 경로 구성 유사도 계산 및 유사 경로 테이블의 구성

XML문서는 동종의 문서라 해도 그 구조가 다를 수 있다. 예를 들어 기존의 문서와 완전히 같은 문서임에도 불구하고 삽입될 문서의 상위나 경로의 중간에 새로운 태그가 삽입되면 그 태그의 하위에 존재하는 모든 태그의 경로는 기존 경로와 같음에도 불구하고 다른 것으로 인식된다. 이를 위해 변형된 경로에 대하여 유사 경로 여부를 판별할 수 있는 기법이 필요하다.

본 논문에서는 경로 구성 유사도(P.C.Sim.)를 사용하여 비트맵 인덱스에 완전히 이식 가능한 유사 경로 인식 기법을 제안한다. 삽입 대상인 문서에 존재하는 경로를 기준 경로라 하고 요소 경로 테이블에 존재하는 경로들을 비교 대상 경로라 한다. 경로 구성 유사도는 기준 경로와 비교 대상 경로의 태그들이 얼마나 순차적으로 빠짐없이 연결되어있는지의 정도를 판별한다. 우선 내용이 있는 태그의 경로로 구성된 요소 경로 테이블에 존재하는 경로들의 경로ID를 이름으로 하는 각각의 유사 경로 테이블(S.P.Table)을 구성한다. 기준 경로는 비교 대상 경로와 경로 구성 유사도를 계산하여 특정 임계치 이상일 때 비교 대상 경로의 경로ID를 이름

으로 하는 유사 경로 테이블에 삽입된다. 경로 구성 유사도는 다음과 같다.

$$P.C.Sim. = Path.P.C.Cor. \bullet List.P.C.Cor.$$

- $Path.P.C.Cor.$ = 기준 경로의 경로 구성 정확도
- $List.P.C.Cor.$ = 비교 대상 경로의 경로 구성 정확도

경로 구성 정확도(P.C.Cor)는 경로에 속한 태그들의 가치의 평균 값으로 다음과 같다.

해당 경로를 구성하는 모든 태그의 수를 k라 하면,

$$P.C.Cor. = \frac{\sum_{i=1}^{k} Tag.Value_i}{k}$$

- $Tag.Value$ = 경로에 존재하는 태그의 가치

태그의 가치는 각각이 얼마나 양쪽 경로에 순차적으로 존재하는지를 측정한 값으로 다음과 같다.

i) 연결에 거리가 발생하지 않는다면,
$Tag.Value = 1$

ii) 연결에 n개의 거리가 발생한다면, 해당 n개의 태그의 가치는 다음과 같이 변경된다.

$$Tag.Value = \frac{1}{2^n}$$

경로 구성 유사도를 구하기 위한 태그 매칭은 경로의 가장 뒤에 있는 태그부터 순차적으로 수행한다. 탐색 여부 함수(Exp.)는 현재 탐색 대상인 태그가 탐색을 계속 수행할 것인지 아니면 비교 대상 경로에는 존재하지 않는다고 판별할 것인지를 결정하는 함수로 휴리스틱 탐색의 낙관적 추정자(optimistic estimator)[5]와 같다. 탐색 여부 함수는 어떤 임계치가 주어졌을 때, 임계치 이상이면 탐색을 계속 수행하고 임계치 미만이면 다음 태그의 탐색을 수행하며, 기준 경로의 태그가 모두 탐색이 완료되면 연산을 종료한다. 기준 경로에서 탐색이 완료된 모든 태그의 개수를 n_1이라 하고 완료되지 않은 태그의 개수를 n_2라고 한다. 비교 대상 경로에서 탐색이 완료된 모든 태그의 개수를 m_1이라 하고 완료되지 않은 태그의 개수를 m_2라고 하면 탐색 여부 함수는 다음과 같다.

$$Exp. = \frac{(\sum_{i=1}^{n1} Tag.Value_i) + n2}{n1 + n2} \times \frac{(\sum_{i=1}^{m1} Tag.Value_i) + n2}{m1 + n2}$$

표 3.2는 그림 3.2와 같은 기준 경로와 비교 대상 경로에 대하여 경로 구성 유사도를 계산하는 과정이다. 유사 경로 판별을 위한 임계치를 0.7이라 가정하면 경로(a)와 경로(b)는 경로 구성 유사도를 구하기 위한 순차적인 태그 매칭 과정에서 태그 'A'의 탐색 여부 함수의 값이 0.69가 되는 순간에 태그 'A'에 대한 탐색을 종료한다. 태그 'A'는 기준 경로의 마지막 태그이므로 시스템은 두 개의 경로가 서로 유사하지 않다는 판정을 내리고 경로 구성 유사도를 구하는 전체 연산을 종료한다.

표 3.2 경로 구성 유사도 계산

(a)	n1	1	2	2	3	3	4
	Tag.Name	D	C	C	B	B	A
	Tag.Value	1	1	1	1	1 0.5	1
	P.C.Cor.	1	1	1	1	1	0.87
(b)	m1	1	2	3	4	5	4
	Tag.Name	D	F	C	E	A	E
	Tag.Value	1	0.5	1	0.5 0.25	0.25	0.5
	P.C.Cor.	1	0.9	0.9	0.83	0.71	0.8
	Exp.	1	0.9	0.9	0.83	0.71	**0.69**

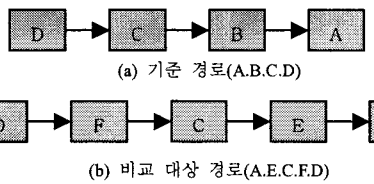


그림 3.2 경로 구성 유사도 계산

이와 같은 방법으로 기준 경로가 유사 경로로 판별이 된다면 기준 경로는 비교 대상 경로의 경로ID를 이름으로 하는 유사 경로 테이블에 삽입된다. 그림 3.3은 요소 경로 테이블과 유사 경로 테이블과의 관계를 나타낸다.

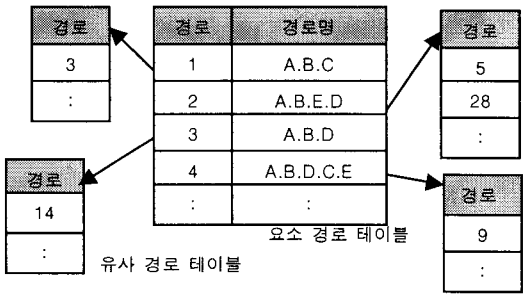


그림 3.3 요소 경로 테이블과 유사 경로 테이블

3.3 비트맵 인덱스와의 이식

본 기법을 비트맵 인덱스에 이식하기 위해서 모든 처리를 Bit-wise연산이 가능한 형태로 변환시킨다. 3.1절에서 추출한 경로 추출 테이블은 Bit-wise연산을 위해 비트맵 인덱스와 같은 형태로 변형되는데 각각의 추출된 경로는 발생 빈도수만큼 확장시킨다. 삽입될 문서의 비트맵 인덱스의 중심은 각각의 클러스터의 중심과 비교하여 유사도를 계산하는데 클러스터의 중심을 시스템에 로드할 때 유사 경로 테이블의 집합의 중심도 로드한다. 기존의 유사도 계산식인 $sim(di, dj) = 1 - |xOR(di, dj)| / MAX([di],[dj])$를 수행하는 과정 중에서 삽입 대상 문서의 중심과 클러스터의 중심을 xOR 연산하는 부분에서 결과가 $false(0)$으로 나왔을 때 유사 경로 집합의 중심과 같은 연산을 다시 수행한다. 그림 3.4는 비트

맵 인덱스와의 이식의 전체적인 도식을 나타내며 알고리즘은 다음과 같다.

유사도 $sim(di, dj) = 1 - | xOR(di, dj) | / MAX([di],[dj])$ 를 구하는 과정 중에서 $xOR(di, dj)$ 연산의 결과가 $false(0)$ 이면,

Method isSimilarPath(*di.ID*)
for i=1 **to** *Sim.P.S.Cen.Length*
  if *di.ID* = *Sim.P.S.Cen.tag[i]*
    return true
return false

- *di.ID* = 삽입 대상 문서의 현재 연산중인 경로ID
- *Sim.P.S.Cen.* = 유사 경로 집합의 중심
- *Sim.P.S.Cen.Length* = *Sim.P.S.Cen.* 의 크기
- *Sim.P.S.Cen.Path[n]* = *Sim.P.S.Cen.* 의 n번째 경로ID

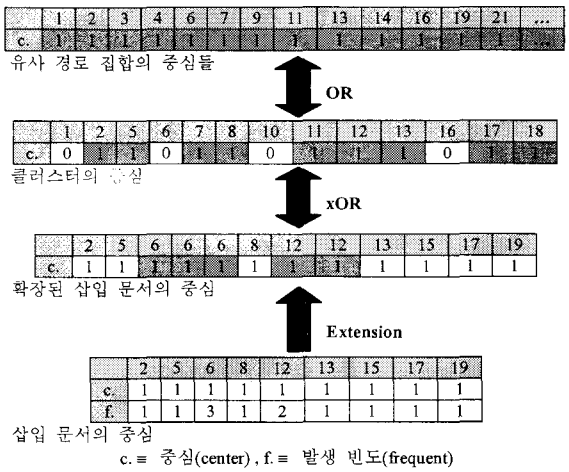


그림 3.4 비트맵 인덱스와의 이식

4. 실험 및 결과

경로 구성 유사도의 실용성을 평가하기 위해 '이력서'와 관련된 가장 대표적이라 여겨지는 8종의 문서를 대상으로 실험을 실시한 결과 해당 경로들은 유사 경로라 판별되는 경로에 대하여 평균적으로 73%의 인식률을 보였다. 이것은 서로 동일한 의도로 작성된 경로임에도 기존 시스템이 유사 경로를 전혀 판별할 수 없었던 오류를 제거함으로써 더 융통성있는 양질의 결과를 도출할 수 있음을 의미한다. 인식률이 예상보다 낮은 이유는 같은 의미로 사용된 경로가 그 구조적인 유사성의 판별은 가능했지만 태그의 직접적인 표현의 차이를 판별하는 것이 불가능했기 때문이다. 그림 4.1 은 실험의 결과 그래프이다.

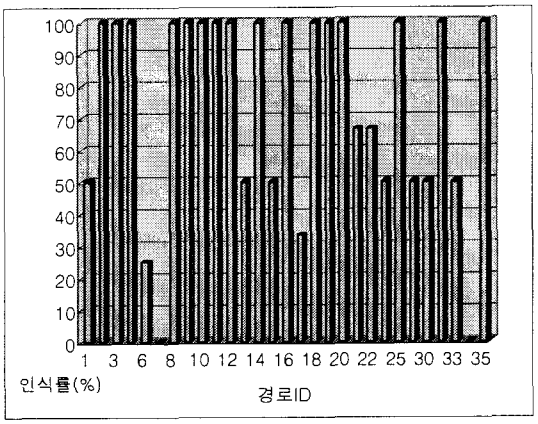


그림 4.1 경로들의 유사 경로 인식률

5. 결론

본 논문에서는 기존의 시스템이 유사 경로를 인식 하지 못함으로 발생하는 문제를 해결하기 위해 경로 구성 유사도를 이용한 비트맵 인덱싱 기반 XML 문서 인식 기법을 제시하였다. 이 기법은 경로 축약을 통해 얻은 정보로 경로의 발생 빈도를 추출하여 해당 경로를 연산시에 그 수만큼 확장하여 발생 빈도를 연산에 반영한다. 그리고 경로 구성 유사도를 사용하여 각각의 경로ID에 대한 유사 경로 테이블을 구성하고 이를 Bit-wise 연산이 가능한 기존 시스템에 완전히 이식함으로써 유사 문서 인식에 유연성을 부여하면서 동시에 기존의 빠른 성능을 저하시키지 않는 기법이다. 경로 구성 유사도는 실험 결과 유사 경로로 판단되는 경로들에 대하여 73%의 인식률을 보임으로써 기존 시스템보다 더 유연하게 양질의 결과를 도출할 것으로 기대된다.

앞으로 동일한 의미의 태그가 갖는 표현의 차이를 인식하지 못하는 문제와 현재 연구중인 3차원 비트맵 인덱싱을 사용하는 BitCube를 연결 리스트를 사용하여 축약하는 부분에 대해 연구를 계속하여 수행할 것이다.

참고문헌

[1] R. Bourret, "XML and Databases", http://www.rpburret.com/xml/XMLAndDatabases.htm, 2003

[2] L. Feng, E. Chang, and T. Dillon, "A Semantic Network-Based Design Methodology for XML Documents", ACM Transaction on Information System, Vol.20, No.4, 2002

[3] J. Yoon, V. Raghavan, and V. Chakilam, "BitCube: Clu-stering and Statistical Analysis for XML Documents", 13th International Conference on Scientific and Statistical Data-base Management, Virginia, July 18~20, 2001

[4] J. P. Yoon, V. Raghavan, V. Chakilam, and L. Ker-schberg, "BitCube: A Three-Dimentional Bitmap Indexing for XML Documents", Journal of Intelligent Information System, Vol.17, pp.241~254, 2001

[5] J. Nillson, Artificial Intelligence: A New Synthesis, MORGAN KAUFMANN, 1998