제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

경로 기반의 XML 문서 유사도 계산 기법

이동애, 장덕성
계명대학교 정보통신공학과
e-mail:dong_ae@naver.com

Technique for Path-based Similarity Evaluation of XML Documents

Dong-Ae Yi, Duck-Sung Jang
Dept. of Information and Communication Engineering, Keimyung University

요 약

XML은 의미적으로는 동일하거나 혹은 유사하지만 서로 다른 구조의 XML 문서들을 허용하므로 XML 문서들을 대상으로 하는 검색, 클러스터링 등의 응용에서는 XML 문서들간의 유사도 계산이 선행되어져야 한다. XML 문서간 유사도를 계산하기 위해서는 문서의 구조 정보인 엘리먼트들과 이들 엘리먼트들의 계층적 구조가 고려되어져야 한다. 본 연구에서는 두 XML 문서가 얼마나 유사한 경로들을 공통으로 가지냐를 두 문서간의 유사도로 보고, 경로 유사도 계산식과, 이를 기반으로 하는 문서 거리 및 문서 유사도 계산식을 정의하여, 유사도 계산 기법을 제안한다. 제안된 기법과 기존 유사도 계산 기법들을 예제 문서들을 통해 계산결과를 비교한다.

1. 서론

엘리먼트 이름과 엘리먼트들 사이의 계층적 구조의 임의적 선택을 허용하며 자기서술적 특징을 갖는 XML은 내용을 구조적으로 표현할 수 있는 장점을 가지는 반면, 의미적으로는 동일하거나 혹은 유사하지만 구조적으로는 서로 다른 XML 문서들의 양산을 초래하게 되어, 서로 다른 구조의 XML 문서를 대상으로 하는 검색을 포함한 여러 응용에서는 이들 간의 유사도 계산이 선행되저져야 한다. XML의 적용범위가 빠르게 넓어져가고 있고, 기존 데이터의 XML로의 변환도 가속화됨에 따라, 다양한 구조의 XML 문서들이 대량으로 생성되고 있지만 XML의 유사도와 관련한 연구는 국내외적으로 아직 초보단계라 할 수 있다.

본 연구에서는, 기존의 XML 유사도 계산 기법들에서 XML 문서의 구조를 충분히 반영하지 못하는 점을 개선하기 위해 XML 문서를 경로단위로 분할하고 경로간 유사도를 계산하고, 이를 기반으로 문서 전체의 유사도를 계산하고자 한다. 이를 위해 경로 유사도 계산식, 문서 거리 및 문서 유사도 계산식을 정의하고, 이에 따른 유사도 계산 기법을 제안한다. 또한 유사도 계산결과를 통해 제안된 알고리즘과 기존 유사도 계산 알고리즘들을 비교한다.

본 논문의 구성은, 2장에서 관련연구를 다루고, 3장에서 경로기반의 유사도 계산 기법을 소개하고, 기존 기법과의 계산결과를 비교하고, 4장에서 결론을 맺는다.

2. 관련연구

XML의 유사도와 관련한 연구로는, XML 문서를 대상으로 하는 연구들[1,2,4,6,7,8,10]과 XML 스키마를 대상으로 하는 연구들[3,9]이 있다. 연구[8]에서는 두 XML 문서에 해당하는 트리구조가 얼마나 공통부분을 가지느냐를 두 문서간의 유사도로 보고, 순차패턴마이닝(Sequential Pattern Minning) 알고리즘을 변형하여 유사도 계산 기법을 제안하였고, 연구[7]에서는 연구[8]에서 제안한 기법이 트리구조를 충분히 반영하지 못하는 점을 개선하여 새로운 기법을

제안하고 있으나, 이 두 기법 모두 두 문서사이에서 기준 문서가 누가 되느냐에 따라 유사도 계산결과가 달라지는 한계를 가지고 있다.

빠른 XML 문서 검색성능이 입증된 BitCube[2]에서는 XML 문서들을 경로단위로 분할하여 표1과 같이 비트맵 인덱스를 구성하는데, 비트맵 인덱스는 문서ID와 경로ID를 축으로 하는 비트연산이 가능한 필드로 구성된 2차원 배열로 빠른 연산이 장점이다. 또한 그림1에서와 같이 문서 거리와 문서 유사도를 정의하여 문서간의 유사도를 근거로 클러스터링을 수행한다. 그러나 표1에서 보는 바와 같이 해당 경로가 존재하면 1, 존재하지 않으면 0으로만 필드 값이 결정되므로, 두 경로간의 유사도가 정확히 반영되지 못하게 되어, 결국 문서간의 유사도 계산 나아가 문서 클러스터링에서 정확성을 기대할 수 없다[5].

표 1 BitCube의 비트맵 인덱스

Path / Document	p_1	p_2	p_3	...	p_n
d_1	1	1	0	...	0
d_2	0	1	1	...	1
d_3	0	1	1	...	0
popularity	0.33	1	0.66	...	0.33
center	0	1	1	...	0

정의 1. 문서 거리

$dist(d_i, d_j) = |xOR(d_i, d_j)|$

정의 2. 문서 유사도

$sim(d_i, d_j) = 1-|xOR(d_i,d_j)|/MAX([d_i],[d_j])$

그림 1 BitCube의 문서 유사도 계산식

본 연구에서는 BitCube에서 경로간 유사성이 고려되지 못하는 단점을 개선하기 위해, '경로간 유사도 계산식'을 정의하여 경로간 유사도를 반영하고, 이에 따라 '문서 유사도 계산식'을 정의하여 문서 전체의 유사도를 계산하여, 유사도 계산의 정확성을 높이고자 한다.

3. 경로 중심 유사도 계산 기법

XML 문서는 트리구조로 매핑되고, 트리는 루트 노드로부터 단말노드까지의 경로들로 구성된다. 표3은 그림2의 두 문서에서 추출한 경로들로, 계산의 편의를 위해 표2에서처럼 엘리먼트 이름 대신 매핑된 숫자로 처리한다. 그림3은 두 경로사이의 유사도 계산을 위해 정의된 '경로 유사도 계산식'이며, 그림4는 정의된 계산식에 따른 계산 예를 보여주고 있다. 경로 유사도를 근거로 문서 전체의 유사도를 고려하기 위해 표5와 같이 문서-경로 테이블을 생성하는데, 필드값은 임의의 경로가 해당 문서에서의 존재여부(1 혹은 0)로 표시되는 BitCude에서와는 달리 그림5에서 정의된 바대로 해당 문서에서 가장 유사한 경로의 유사도 값으로 취한다. 그림6은 정의된 '문서거리'와 '문서 유사도 계산식'을 보여준다.

```
<book>
  <bookinfo>
    <page></page>
    <paperback></paperback>
    <edition></edition>
    <date></date>
    <publisher></publisher>
    <ISBN></ISBN>
    <size></size>
  </bookinfo>
  <authorgrp>
    <name></name>
  </authorgrp>
  <buyinginfo>
    <price>
      <list></list>
      <our></our>
      <save></save>
    </price>
  </buyinginfo>
  <content>
    <section>
      <chap></chap>
    </section>
  </content>
  <reviews>
    <customs>
      <rating></rating>
    </customs>
  </reviews>
  <title></title>
</book>
```

(a) amazon.xml

```
<bookinfo>
  <reviews>
    <reviews_no></reviews_no>
    <avg_rating></avg_rating>
  </reviews>
  <price>
    <list></list>
    <our></our>
  </price>
  <author>
    <name></name>
  </author>
  <table_of_contents>
    <section>
      <chap></chap>
    </section>
  </table_of_contents>
  </title></title>
</bookinfo>
```

(b) b&n.xml

그림 2 amazon.xml, b&n.xml

표 2 amazon.xml, b&n.xml의 엘리먼트 이름

Node ID	엘리먼트 이름	amazon.xml	b&n.xml
1	book	o	
2	title	o	o
3	bookinfo	o	o
4	page	o	
5	paperback	o	
6	edition	o	
7	date	o	
8	publisher	o	
9	ISBN	o	
10	size	o	
11	buyinginfo	o	
12	price	o	o
13	list	o	o
14	our	o	o
15	save	o	
16	contents	o	o
17	section	o	o
18	chap	o	o
19	reviews	o	o
20	customer	o	
21	rating	o	o
22	authorgrp	o	o
23	name	o	o
24	review_no		o

표 3 amazon.xml, b&n.xm에서 추출된 경로

Document ID	Path ID	Path
d_1 (amazon.xml)	p_1	1 2
	p_2	1 3 4
	p_3	1 3 5
	p_4	1 3 6
	p_5	1 3 7
	p_6	1 3 8
	p_7	1 3 9
	p_8	1 3 10
	p_9	1 11 12 13
	p_{10}	1 11 12 14
	p_{11}	1 11 13 15
	p_{12}	1 16 17 18
	p_{13}	1 19 20 21
	p_{14}	1 22 23
d_2 (b&n.xml)	p_{15}	3 2
	p_{16}	3 12 13
	p_{17}	3 12 14
	p_{18}	3 22 23
	p_{19}	3 16 17 18
	p_{20}	3 19 24
	p_{21}	3 24 21

정의 1. 경로 유사도(path similarity)

$$p\_sim(\text{기준경로, 비교경로}) = \frac{MatchLength(\text{기준경로, 비교경로})}{Length(\text{기준경로})}$$

그림 3 경로 유사도 계산식

p_1 : 1 2 3 4 5 6 7
p_2 : 9 3 4 7 8 11 13 14

$$sim(p_1, p_2) = \frac{3}{8}$$

그림 4 경로 유사도 예제

value(i, j)
= Max(similarity(p_j,p_1), ...,similarity(p_j,p_n))
, p_1, p_n은 d_i 의 모든 경로
, $0 \le value(i, j) \le 1$

그림 5 문서-경로 테이블의 필드 값

정의 2. 문서 거리(document distance)

$$d\_dist = \sum_{k=1}^{n} |mik - mjk| \ / \ k$$

, k는 두 문서 경로수(document similarity)

정의 3. 문서 유사도(document similarity)

$$d\_sim(d_i, d_j) = 1 - d\_dist(d_i, d_j)$$

그림 6 본 연구의 유사도 계산식

표 4 amazon.xml, b&n.xml의 비트맵인덱스

Path / Document	p_1	p_2	p_3	p_4	p_5	p_6	p_7
d_1	1	1	1	1	1	1	1
d_2	0	0	0	0	0	0	0
Path / Document	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}
d_1	1	1	1	1	1	1	1
d_2	0	0	0	0	0	0	0
Path / Document	p_{15}	p_{16}	p_{17}	p_{18}	p_{19}	p_{20}	p_{21}
d_1	0	0	0	0	0	0	0
d_2	1	1	1	1	1	1	1
$d\_dist(d_1, d_2)$	21						
$d\_sim(d_1, d_2)$	1 - 21 /21 = 0						

표 5 amazon.xml, b&n.xml의 문서-경로 테이블

Path / Document	p_1	p_2	p_3	p_4	p_5	p_6	p_7
d_1	1	1	1	1	1	1	1
d_2	0.5	0.33	0.33	0.33	0.33	0.33	0.33
Path / Document	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}
d_1	1	1	1	1	1	1	1
d_2	0.33	0.5	0.5	0	0.75	0.25	0.66
Path / Document	p_{15}	p_{16}	p_{17}	p_{18}	p_{19}	p_{20}	p_{21}
d_1	0.5	0.66	0.66	0.66	0.75	0.33	0.33
d_2	1	1	1	1	1	1	1
$d\_dist(d_1, d_2)$	9.36/21 = 0.45						
$d\_sim(d_1, d_2)$	1 - 0.45 = 0.55						

표6는 본 연구에서 제안한 유사도 계산 기법과 기존에 제안된 기법들을 비교하기 위해 그림2의 amazon.xml, b&n.xml을 대상으로 한 계산결과이다. 우선 BitCube에서는 유사한 경로들이 있음에도 불구하고 동일한 경로가 없기 때문에 0으로 계산되며, 연구[7,8]은 두 문서에서 기준문서를 어느 것으로 하느냐에 따라 계산 결과가 달라지므로 amaxon,xml을 기준문서로 한 경우 각각 0.83, 0.27로 계산되며, 본 연구의 기법으로는 0.55로 계산되었다. 참고로 두 문서간의 공통 엘리먼트의 비율이 표2에서 보는 바와 같이 12/24이다.

표 6 amazon.xml, b&n.xml의 유사도 계산 결과

구분	연구[8]	연구[7]	BitCibe	본 연구
유사도	0.83	0.27	0	0.55

4. 결론

본 연구에서는 연구[7,8]의 기법들의 두 문서사이에서 어느 문서를 기준문서로 하느냐에 따라 유사도 계산 결과가 달라지는 문제를 해결하고, BitCube[2]에서 경로간 유사도가 고려되지 못하는 문제를 해결하기 위해 경로 유사도 계산식 과 문서 유사도 계산식을 정의하여 두 XML 문서간의 유사도를 계산하는 기법을 제안하였다. 향후 좀더 다양한 XML 문서들을 대상으로 유사도 계산 결과가 비교되어야 하겠고, 이 유사도 계산 기법에 근거하여 XML 문서 클러스터링 시스템을 구축할 계획이다.

참고문헌

[1] Andrew Nierman and H. V. Jagadish, "Evaluate Structural Simailarity in XML Documents", WebDB, pp. 61~66, 2002

[2] J. Yoon, V. Raghavan, V. Chakilam, and Kerschberg, "BitCube: A Three-Dimensional Bitmap Indexing for XML Documents", Journal of Intelligent Information System, Vol. 17, pp. 241~254, 2001

[3] M. Lee, L. Yang, W. Hsu and X. Yang, "XClust: Clustering XML Schemas for effective Integration", Proc. 11th Int. cong. on Information and Knowledge management, pp. 292~299, Nov., 2002

[4] Wang Lian, et al., "An Efficient and Scalable Algorithm for Clustering XML Documents by Structure", IEEE Transactions on knowledge and data engineering, Vol. 16, No. 1, January 2004

[5] 김연혜, 이재민, 황병연, "구조 유사도를 이용한 경로 기반의 색인 기법", 한국정보과학회 2003년 추계학술대회

[6] 이재민, 황병연, "xPlaneb: XML 문서 검색을 위한 3차원 비트맵 인덱스", 한국정보과학회논문지, 데이터베이스 제31권 제3호. 2004. 6

[7] 이동애, 장덕성, "XML 문서의 유사도", 제11회 한국정보과학회영남지부 학술발표논문집, 제11권 1호, 2003.12

[8] 이정원, 이기호, "유사성 기반 XML 문서 분석 기법", 한국정보과학회논문지 제29권, 제6호, pp. 367-376, 2002.

[9] 임태우, 이경호, "XML 스키마 클러스터링을 위한 효율적인 알고리즘", 한국정보과학회 2004년 춘계학술대회

[10] 황정희, 류근호, "순차패턴에 기반한 XML 문서 클러스터링", 한국정보처리학회논문지 D. 제10-D권 제7호, 2003.12