제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

XML 문서 클러스터링을 위한 경로 유사도의 계산

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

Path Similarity Calculation for Clustering of XML Documents

Bum-Suk Lee, Byung-Yeon Hwang
Dept. of Computer Engineering, The Catholic University of Korea

요 약

최근 DTD (Document Type Descriptor)를 포함하고 있지 않은 XML 문서의 사용이 증가하고 있다. 따라서 서로 다른 구조를 갖는 많은 양의 XML 문서를 관계형 DBMS에 저장하거나, 인덱스를 이용하여 매핑하는 등 보다 효율적으로 관리하기 위한 다양한 인덱싱 기법에 대한 연구가 진행되고 있다. 이러한 연구들 중 경로 비트맵 인덱싱 기법은 경로 구성 유사도를 기반으로 3차원 비트맵 클러스터를 생성하고, 클러스터 단위의 검색을 수행함으로서 빠른 검색 속도를 보여주었다. 그러나 이 기법은 비교하려는 두 경로 중 항상 짧은 경로가 기준 경로가 되는 한계점과, 같은 노드 구성을 가지는 두 경로에서도 노드의 위치에 따라 그 유사도가 크게 변하는 등의 여러 문제점을 가지고 있었다. 이러한 문제점을 해결하고, 정확한 클러스터링을 수행하기 위해서는 합리적인 경로 유사도 계산식이 필요하게 되었다. 본 논문에서는 기존 방법의 문제점을 해결하고, 보다 정확한 클러스터링을 수행할 수 있는 새로운 경로 유사도 계산식을 제안한다.

1. 서론

최근 XML [1]의 사용이 급격히 증가하면서, 많은 XML 문서들이 DTD를 포함하지 않은 상태로 사용되고 있다. 이처럼 서로 다른 구조를 갖는 다량의 XML 문서를 효율적으로 저장하고 검색하기 위한 다양한 기법에 대한 연구들이 진행되고 있다.

특히 문서, 경로, 단어를 축으로 가지는 3차원 비트맵 인덱스 기법을 사용한 BitCube[2]는 XYZFind, XQEngine 등 다른 XML 문서 검색 시스템과의 성능평가에서 빠른 속도를 보여주었다. 이 시스템은 동일한 경로를 많이 포함하는 문서들을 클러스터링하여 클러스터 단위로 문서 검색을 수행하였기 때문에 좋은 검색 성능을 보여줄 수 있었다. 그러나 XML 문서의 루트(root)로부터 말단 노드(leaf node)까지의 전체 경로를 하나의 필드(field)로 사용하기 때문에, 경로를 수정하면 그 유사성을 고려하지 않고 완전히 새로운 경로로 인식하는 단점이 있다.

이러한 단점을 개선하기 위해 유사한 경로를 이용한 클러스터링 기법에 대하여 수행한 연구들[3,4]이 있다. 이 연구들에서는 XML 경로들에 대해 유사도를 계산하고, 그 결과에 따라 경로 비트맵 인덱스 클러스터를 생성하였다. 그리고 유사도를 계산하는 식은 경로들 사이에 존재하는 서로 다른 노드의 수에 반비례하는 특성을 가진다. 이러한 특성은 의미적으로 차이가 나지 않을 수 있는 두 경로 구성에 대해 다른 유사도 값을 제시하거나, 서로 다른 노드 구성을 가지는 두 경로 사이의 유사도를 계산하는 것이 불가능한 문제점들을 가지고 있었다.

본 연구에서는 보다 정확한 유사도 계산을 위한 방법을 제안한다. 제안한 방법은 두 경로를 같은 형태로 만들기 위한 최소 비용을 계산하고, 이것을 이용하여 두 경로 사이의 거리(distance)와 유사도(similarity)를 계산한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 간략히 소개하고, 3장에서는 제안하는 경로 유사도를 계산하는 방법을 설명한다. 4장에서는 이 방법을 직접 구현하여 실험한 결과에 대해 논의하고, 마지막으로 5장에서는 결론과 향후 연구 계획에 대해 소개한다.

2. 관련 연구

2.1 경로 비트맵 인덱싱을 위한 경로 구성 유사도

XML 문서 검색을 위해 경로를 기준으로 하는 비트맵 인덱싱에서 경로는 인덱스 생성을 위한 기본 구성단위가 된다. 만약 문서의 구조가 변경되는 경우에는 변경되기 전의 경로와 변경된 후의 경로가 완전히 다른 경로로 인식되었다. 따라서 경로 기준의 비트맵 인덱싱은 유사한 경로의 검색이 불가능하였다. 이러한 문제점을 해결하기 위해 경로 유사도를 계산하는 다음과 같은 두 개의 정의를 제시하였다.

정의 2.1 경로 P를 구성하는 노드 N의 노드 값 $NodeValue(P, N)$은 다음과 같이 정의된다. d는 기준 경로의 노드가 비교하려는 경로의 노드를 만날 때 까지 일치하지 않는 노드의 개수이다.

$$NodeValue(P,N) = \frac{1}{2^d}$$

정의 2.2 서로 다른 경로 P_1과 P_2의 경로 구성 유사도 $P.C.Sim(P_1, P_2)$는 다음과 같이 정의된다. $NodeNum(P)$는 경로 P가 포함하고 있는 모든 노드의 개수이다. $NodeNum(P_1)$은 $NodeNum(P_2)$ 보다 작거나 같다.

$$P.C.Sim\ (P_1,P_2) = \frac{\sum_{i=1}^{NodeNum\ (P_2)} NodeValue\ (P_2,N_i)}{NodeNum\ (P_2)}$$

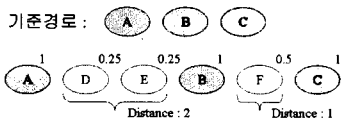


그림 1. 경로 구성 유사도의 계산

그림 1은 경로 구성 유사도를 계산하는 과정을 보여준다. 예를 들어, 두 경로 A.B.C와 A.D.E.B.F.C의 경로 구성 유사도를 계산하기 위해서 정의 2.1과 정의 2.2를 이용하여 두 경로 사이의 경로 구성 유사도를 구하면 (1+0.25+0.25+1+0.5+1)/6=0.67이 된다.

2.2 기존 경로 구성 유사도 계산식의 문제점

2.1에서 살펴본 경로 구성 유사도 계산식은 다음과 같은 여러 가지 문제점을 가지고 있다. 1) 다른 구성을 가진 경로 사이의 유사도 계산이 불가능하다. 즉, 그림 2의 (a)와 같은 두 경로에 대해서 위의 식으로는 계산이 불가능하다. 2) 그림 2의 (b)와 같이 노드의 순서가 다른 두 경로 사이의 유사도 계산도 불가능하다. 3) 거리 기반의 유사도를 계산하기 때문에 기준 경로 A.B.C와 그림 2의 (c)의 두 경로에 대한 각각의 유사도가 0.7과 0.8로 서로 다른 결과를 나타낸다.

그림 2. 경로 구성 유사도 계산식의 문제점

3. 새로운 경로 유사도의 계산

이 장에서는 2.2에서 설명했던 경로 구성 유사도 계산의 문제점을 보완한 새로운 경로 유사도 계산 방법을 제안한다. 이것을 설명하기 위해 우선 삽입, 삭제, 치환의 세 가지 연산에 대해 정의한다.

정의 3.1 $insert(x)$는 해당 경로에 새로운 노드 x를 삽입하는 연산이고, 이 연산을 수행하는데 드는 비용은 1이다.

$$Cost(insert(x)) = 1$$

정의 3.2 $delete(x)$는 해당 경로를 구성하는 노드 중 x를 삭제하는 연산이고, 이 연산의 수행비용은 1이다.

$$Cost(delete(x)) = 1$$

정의 3.3 $replace(x, y)$는 해당 경로의 노드 x를 y로 치환한다. 이 연산의 수행비용은 $delete(x)$와

*insert(y)*를 수행하는 비용을 합한 값으로 2이다.

$Cost(replace(x,y)) = Cost(delete(x)+insert(y)) = 2$

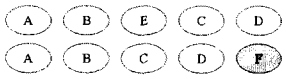
위의 정의를 이용하면 임의의 경로 (P_1)를 특정한 형태의 경로 (P_2)로 변경하기 위한 비용을 계산할 수 있다. 이 비용의 최소 비용을 $Min.Cost(P_1,P_2)$라고 하면, 두 경로 사이의 유사도 $S(P_1,P_2)$는 다음 정의와 같이 계산된다.

정의 3.4 경로 P_1과 P_2 사이의 유사도 $S(P_1,P_2)$는 다음 식과 같다. P_1을 P_2로 변경하는 최소 비용을 P_1과 P_2의 노드 개수를 합으로 나누고, 그 값을 1에서 뺀다.

$$S(P_1, P_2) = 1 - \frac{Min.Cost(P_1, P_2)}{NodeNum(P_1) + NodeNum(P_2)}$$

정의 3.4의 식에 따르면, 유사도 값은 0과 1 사이의 값을 가지게 되고, P_1과 P_2가 완전히 다른 경우 0, 완전히 일치하는 경우 1이다.

그림 3은 정의 3.4를 이용하여 실제 두 경로 사이의 유사도를 구하는 예를 보여준다. 두 경로 A.B.E.C.D (P_1)와 A.B.C.D.F (P_2)에서 P_1을 P_2로 변형하기 위한 최소 비용이 2이고, 두 경로에 포함된 노드의 개수 합이 10이다. 이를 정의 3.4의 유사도 계산식에 대입하면, 두 경로의 유사도는 0.8이 된다.



$Min.Cost(P_1,P_2)=2$
$NodeNum(P_1)+NodeNum(P_2)=10$
$S(P_1,P_2)=1-2/10=0.8$

그림 3. 두 경로 P_1을 P_2 사이의 유사도

제안하는 방법은 기존의 방법으로 유사도를 계산할 수 없었거나 문제점이 있었던 그림 2의 예에서도 사용될 수 있다. 우선 그림 2의 (a)는 경로가 서로 공통되지 않은 다른 노드를 포함한 경우이다. 제안하는 방법의 유사도 계산식을 이용하면, 그 유사도는 0.8이 된다. (b)는 두 경로를 구성하는 노드의 순서가 다른 경우이다. 이 경우의 유사도는 0.75이다. (c)는 거리 기반의 기존 방법으로 기준 경로 A.B.C에 대해 두 경로가 서로 다른 유사도를 가졌었으나, 제안하는 방법으로는 기준 경로에 대해 두 경로 모두 0.75의 일정한 유사도를 가진다.

경로 유사도 계산을 위한 시간 복잡도는 기준 경로의 노드 개수가 m, 비교하는 경로의 노드 개수가 n이라고 가정할 때, 기존의 방법은 $O(n)$이고 제안하는 방법은 $O(mn)$으로 제안하는 방법의 시간복잡도가 더 큰 것을 알 수 있다. 그러나 대부분 흔히 사용되는 XML 문서의 최대 깊이 (maximum depth)가 20개 이하인 것을 감안한다면, 실제 프로그램의 수행으로 걸리는 시간은 무시해도 좋을 정도의 수준이다.

4. 성능평가

본 논문에서 제안하는 계산식의 성능 평가를 위해 Java 1.5를 이용하여 프로그램을 구현하였고, 윈도우 2000 Server 플랫폼이 설치된 Pentium 4 1.7 GHz CPU와 RAM 512MB를 탑재한 PC환경에서 실험을 수행하였다.

표 1. 두 그룹의 경로 쌍과 그 특징

	예제 경로 쌍	특징
그룹 1	P_1 : A.B.C.D.E.F P_2 : B.C.D	두 경로 중 짧은 경로가 긴 경로에 포함
그룹 2	P_1 : A.B.C P_2 : A.D.E.B.C P_2' : A.D.B.E.C	서로 같은 경로 구성에 대해 다른 유사도를 가지는 경우

새로운 경로 유사도의 특성을 알아보기 위하여 임의의 특징을 부여한 두 그룹의 경로 쌍을 각각 10개씩 생성하였다. 표 1은 이 경로 쌍들의 예와 그 특징을 보여준다. 그룹 1의 특징은 두 경로 중 짧은 경로가 긴 경로에 포함되는 경우이고, 그룹 2의 특징은 서로 같은 경로 구성에 대해 다른 유사도를 가지는 경우이다. 그룹 2의 P_2는 P_1에 존재하지 않는 추가된 노드들이 서로 모여 있어 P_1에 대해 가장 낮은 유사도(기존 min.)를 가지도록 하였고, P_2'는 추가된 노드들이 분산되어 가장 높은 유사도(기존 max.)를 가지도록 생성하였다.

그림 4는 그룹 1의 경로 쌍에 대한 유사도 계산 결과 비교를 보여준다. 제안한 방법의 유사도 결과 값이 기존의 유사도 계산식과 비슷한 수준으로 0과 1 사이에서 고른 분포를 가지기 때문에, XML 문서의 경로 클러스터링을 위한 유사도 계산 방법으로 적합한 것을 알 수 있다.

그림 5는 그룹 2에 대한 유사도 계산 수행 결과를 보여준다. 그룹 2에서 기존의 유사도 계산식은 두 경로 쌍에 대해 서로 다른 유사도 계산 결과 값을

가지는 것을 알 수 있다. 또한 그 결과 값의 차이가 경로의 구성에 따라 매우 크게 변화한다. 이러한 이유는 기존의 방법이 정의 2.1에서 제시한 것과 같이 기준 경로의 노드가 비교하려는 경로의 노드를 만날 때 까지 일치하지 않는 노드의 개수가 늘어날수록 *NodeValue*가 줄어들기 때문에, 추가된 노드의 위치에 따라 유사도 값이 큰 차이를 보인다. 이에 비하여 제안하는 방법은 노드의 위치에 관계없이 기준 경로 P_1을 비교하는 경로 P_2 또는 P_2와 노드 구성이 같은 경로 P_2'와 비교할 때, 항상 일정한 유사도 값을 가지는 것을 알 수 있었다.

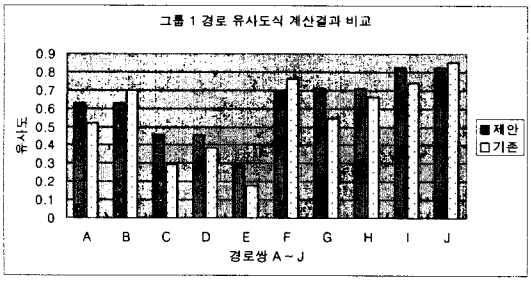


그림 4. 그룹 1의 유사도 계산식 결과

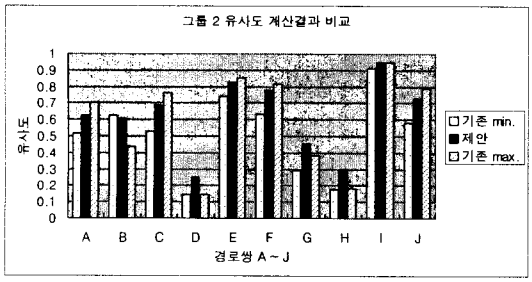


그림 5. 그룹 2의 유사도 계산식 결과

이 밖에도 두 경로에 공통되지 않는 다른 노드를 포함하거나, 두 경로가 서로 같은 노드를 가지지만 노드의 순서가 다른 경우에는 기존의 방법으로 유사도 계산이 불가하였다. 하지만 제안한 방법으로 이러한 경우에도 유사도 계산이 가능하였고, 이에 대한 실험도 수행하였으나 결과에 포함하지는 않았다.

5. 결론 및 향후 연구 계획

본 논문에서는 XML 문서의 경로 클러스터링을 위한 기존의 경로 구성 유사도 계산식에서 드러난 몇 가지 문제점을 개선하기 위해 새로운 경로 구성 유사도 계산식을 제안하였다. 기존의 계산식을 이용할 수 있는 범위가 기준 경로 P_1과 비교 대상 경로 P_2를 비교하기 위해, 1) 항상 짧은 경로가 기준 경로 P_1이 되어야 하는 점, 2) 짧은 경로 P_1을 구성하는 모든 노드가 비교하려는 P_2 경로에 포함되어야 하는 점과 같은 한계를 가지고 있었고, 3) 기준 경로 P_1과 비교하려는 경로 P_2의 노드 순서가 다른 경우 유사도 계산이 불가능하였고, 4) 기준 경로에 대해 경로 구성 순서가 다른 두 경로의 유사도가 다르게 계산되는 등의 문제점을 가지고 있었다.

이러한 문제점을 해결한 새로운 경로 구성 유사도 계산식은 두 경로를 동일한 형태로 바꾸기 위한 최소 비용을 계산하고, 이를 이용하여 0과 1 사이의 값을 가지는 경로 유사도를 계산하였다. 제안한 방법은 기존의 계산식이 가지고 있던 문제점을 해결하였고, 기존의 방법으로는 계산할 수 없었던 경로 쌍의 유사도까지도 계산할 수 있었다.

향후 연구로 여러 경로들의 대표 노드를 추출하고 그들의 순서를 결정하여 경로로 만드는 연구, 즉 경로 클러스터 생성을 위해 여러 경로에 대한 대표 경로를 추출하는 방법에 대한 연구를 진행할 것이다.

참고문헌

[1] http://www.w3.org/TR/2000/REC-xml-20001006

[2] J. Yoon, V. Raghavan, and V. Chakilam, "BitCube: Clustering and Statistical Analysis for XML Documents," In Proc of the 13th Int'l Conf. on Scientific and Statistical Database Management, Virginia, Jul. 2001.

[3] Jae-Min Lee and Byung-Yeon Hwang, "Path Bitmap Indexing for Retrieval of XML Documents," Lecture Notes in Computer Science, Vol. 3885, Springer-Verlag, Apr. 2006.

[4] Wang Lian, David Wai-lok Cheung, Mamoulis, Nikos, Siu-Ming Yiu, "An efficient and scalable algorithm for clustering XML documents by structure," Knowledge and Data Engineering, IEEE Transactions on Vol. 16, Issue 1, pp. 82-96, Jan. 2004.

[5] Antonio Robles-Kelly, Edwin R. Hancock, "Graph Edit Distance from Spectral Seriation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 3, pp. 365-378, Mar. 2005.