엘리먼트 빈도수 정보를 이용한 XML 문서 매칭

고승규° 강명수 임순범* 최윤철
연세대학교 컴퓨터과학과
*숙명여자대학교 멀티미디어학과
{skko°, mskang, ycchoy}@rainbow.yonsei.ac.kr, sblim@sookmyoung.ac.kr

An XML Document Mathcing using Element Frequency Information

Seung-Kyu Ko° Myoung-Soo Kang Yoon-Chul Choy
Dept. of Computer Science, Yonsei University
Soon-Bum Lim*
Dept. of Multimedia, Sookmyoung Women's University

요 약

XML이 널리 사용됨에 따라 많은 정보가 XML 형태로 표현되고 있다. 또한 인터넷의 대중화로 다양한 정보를 통합하여 처리하거나 교환, 변환하는 경우가 빈번하게 발생한다. 따라서 XML로 표현된 정보도 교환되거나 통합되는 경우가 많이 발생하게 된다. 이와 같은 XML 문서 간의 통합이나 변환에서는 XML의 특징인 문서의 논리적인 구조가 적절하게 반영되어야 한다. 그리고 이를 위해서는 XML 문서의 기본적인 구성 요소인 엘리먼트 간의 매칭이 필수적이다. 기존의 XML 문서 매칭 기법에서는 엘리먼트 이름과 계층 정보 등 명시적으로 표현된 최소한의 정보만을 이용하여 매칭을 수행한다. 이러한 최소한의 제한된 정보를 최대로 이용하여 많은 매칭을 수행하기 위하여 기존의 방법에서는 동의어 사전이나 구조 정보를 과도하게 이용하는 경향이 많다. 따라서 많은 대응을 생성할 수 있지만 동시에 잘못된 대응의 수도 증가한다. 이에 본 논문에서는 명확한 대응을 생성시키기 위하여 XML의 명시적인 정보 이외에 엘리먼트의 빈도수 정보로부터 엘리먼트 간의 연결성 정보를 정의하고, 이를 이용한 매칭 방법을 제안한다. 제안 방법은 엘리먼트 이름이나 계층 구조 등의 명시적인 정보뿐 아니라 엘리먼트의 연결성을 이용하기 때문에 매칭의 정확도가 향상될 수 있다. 최근에 발표되는 XML 기반의 표준들은 크기가 방대하고 점점 더 복잡해지고 있다. 이같은 환경에서는 잘못된 대응으로 인해 발생하는 비용이 무척 크다. 제안 기법은 매칭의 정확도가 높으므로 이러한 환경에서 좋은 성능을 발휘할 것으로 기대된다.

1. 서 론

웹 문서 표준인 XML은 재사용성, 확장성, 상호 운영성의 여러 장점을 지니고 있어서 점차 널리 사용되고 있다. 또한 관계형 데이터베이스나 객체지향 데이터베이스 등의 기존의 데이터 처리 시스템에서도 XML을 지원해 나가고 있다. 이와 같이 XML이 널리 사용됨에 따라 정보 교환 시 XML로 표현된 정보를 교환하거나 통합하는 경우가 많이 발생한다. XML 문서를 교환하거나 통합할 경우에는 한 쪽 XML 문서의 구성 요소인 엘리먼트를 다른 쪽 XML 문서의 구성 요소인 엘리먼트에 대응시키는 매칭(matching)이란 과정이 필수적이다. 이러한 매칭은 현재 대부분 수작업으로 이루어지고 있으며, 근래에 자동으로 매칭을 생성하는 방법에 관해 연구 중에 있다. 기존의 XML 문서 매칭 기법에서는 엘리먼트 이름과 계층 정보 등 명시적으로 표현된 최소한의 정보만을 이용하여 매칭을 수행한다. 이러한 최소한의 제한된 정보를 최대로 이용하여 많은 매칭을 수행하기 위하여 기존의 방법에서는 동의어 사전이나 구조 정보를 과도하게 이용하는 경향이 많다. 즉, 일반적인 동의어 사전을 이용하거나 구조 정보를 과도하게 이용하면 많은 대응을 생성할 수 있지만 동시에 잘못된 대응의 수도 증가한다. 이러한 잘못된 대응은 탐색과 수정 등의 고비용이 발생하기 때문에 대응을 생성시키지 않는 것보다 더 안좋은 결과를 유발한다.

이에 본 논문에서는 명확한 대응을 생성시키기 위하여 XML의 명시적인 정보 이외에 엘리먼트의 빈도수 정보로부터 엘리먼트 간의 연결성 정보를 정의하고, 이를 이용한 매칭 방법을 제안한다. 제안 방법은 엘리먼트 이름이나 계층 구조 등의 명시적인 정보뿐 아니라 엘리먼트의 연결성을 이용하기 때문에 매칭의 정확도가 향상될 수 있다.

2. 관련 연구

기존의 XML 문서 매칭 시스템은 사용하는 정보에 따라 다음과 같이 분류할 수 있다.

- 엘리먼트 이름 매칭
  이 방법은 엘리먼트 이름을 대응시키는데 중점을 둔 방법이다. 이 방법에서는 스테밍(stemming), 동의어 사전, 약어 사전, WORDNET 등을 이용하여 엘리먼트 이름 간에 대응을 생성한다. 그러나 일반적인 사전을 이용함으로써 잘못된 동의어를 이용할 수 있고, 모든 엘리먼트에 대하여 대응을 생성시키지 못한다. 이 방법은 대부분의 매칭 시스템에서 기본적으로 사용하는 방법이다.
- 부가 정보를 이용한 매칭
  이 방법은 스키마에서 정의된 타입, 키 등을 이용하여 매칭하는 방법이다. 따라서 단순히 엘리먼트

이름만을 이용한 매칭보다 정확도가 높다. 그러나 XML 문서에서는 이러한 타입이나 키 등을 이용할 수 없기 때문에 데이터베이스[1] 등에서 이용되는 방법이다.

- 인스턴스 정보를 이용한 매칭
  엘리먼트가 실제 내용을 전부 또는 적절하게 표현하지 못하기 때문에 실제 문서들로부터 해당 엘리먼트를 적절하게 표현할 수 있는 키워드를 추출하여 이를 엘리먼트 이름과 같이 이용하여 매칭하는 방법이다. 이 방법은 내용이 유사한 문서가 많을 경우에 효과적일 수 있지만, 실제 문서 내용이 다양할 경우에는 성능이 떨어질 것으로 예상된다. 이 방법은 [2]에서 이용된다.
- 구조 정보 매칭
  이름을 이용하여 매칭을 할 경우에 생성되는 대응에 한계가 있으므로 이름 외에 사용할 수 있는 구조 정보를 이용한다. 예를 들어 부모와 형제, 자식 엘리먼트가 대응이 되면, 이름이 다르더라도 대응을 생성시킨다. 이 방법은 [3]에서 이용된다.
- GUI 기반의 매칭
  자동이 아닌 매칭 방법으로 수작업 매칭을 손쉽게 하기 위하여 GUI를 이용하는 방법이다. 이 방법은 문서의 스키마를 효과적으로 출력하여 전문가가 대응을 명확하게 정의할 수 있도록 한다. 이 방법은 [4]에서 이용된다.
- 기타
  위의 정보 이외에 위치적으로 인접해 있는 엘리먼트의 대응 정보를 이용한 매칭 방법이 있다. 이 방법은 한 엘리먼트가 대응이 될 경우에는 그 상위나 하위의 엘리먼트가 대응될 가능성을 높여주는 방법이다. 이는 매칭된 정보를 재사용하는 것으로도 볼 수 있다. 이 방법은[5]에서 사용되었다.

대부분의 시스템은 하나의 방법만을 이용하지는 않고, 엘리먼트 이름 매칭을 기반으로 두서너가지 방법을 혼합하여 이용한다.

이와 같이 대부분의 시스템에서는 XML 문서에서 명시적으로 표현되는 엘리먼트 이름이나 계층정보만을 이용하고 있다. [2]에서는 실제 문서의 내용으로부터 키워드를 추출하여 이용하고 있으나, 내용이 문서마다 많이 상이할 경우에 효과적이지 못한 단점이 있다.

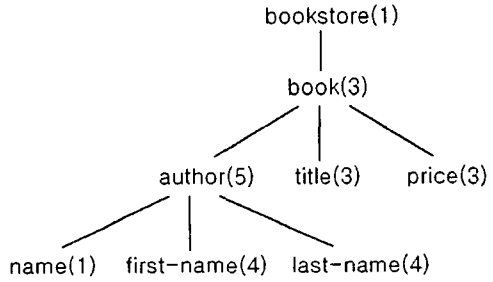
3. 엘리먼트 간의 연결성

XML 엘리먼트 간에 명시적으로 표현되는 관계는 계층 관계와 링크 관계가 있다. 이 중 계층 관계는 부모-자식 관계나 형제 관계만을 표현할 수 있으나 실제 인접한 엘리먼트 간의 관계는 표현하지 못한다. 이에 비하여 XML 문서에서 엘리먼트 간의 빈도수는 인접한 엘리먼트 간의 연결성 정도를 표현할 수 있다. 예를들어 두 엘리먼트의 빈도수가 같으면, 두 엘리먼트가 동시에 발생함을 의미하고, 이는 두 엘리먼트가 강하게 연결되어 있음을 나타낸다. 이와 같은 엘리먼트 간의 연결성 정보는 한 문서가 아닌 여러 문서 구조를 요약하면서 추출될 수 있으며, 일정 수준 이상의 문서에서 추출되는 연결성 정보는 엘리먼트 간의 연결성을 적절하게 표현할 수 있다.

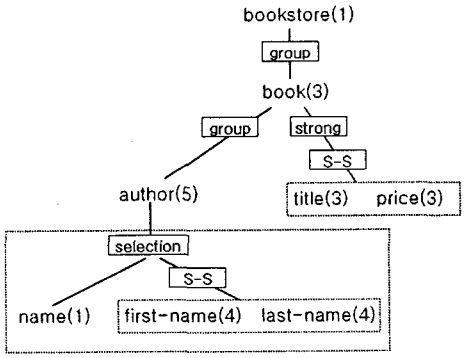
XML 문서에서 현재 엘리먼트 E_n이라고 하고, 부모 엘리먼트를 P_n, 자식 엘리먼트는 C_{n1}, C_{n2}, C_{n3}, …, C_{nm}, 형제 엘리먼트를 S_{n1}, S_{n2}, S_{n3},…, S_n이라 하고, 각각의 빈도수는 $|P_n|$, $|C_n|$, $|S_n|$이라 하면, 빈도수로부터 추출할 수 있는 관계는 [표 1]과 같다.

[표 1] 빈도수에 기반한 엘리먼트 간의 관계

인접 엘리먼트	관 계	조 건
부모	강한 연결(strong)	$\|E_n\| = \|P_n\|$
	선택 연결 (selection)	$\|E_n\| + \|S_{nk}\| = \|P_n\|$, *for some k*
	모임 연결 (group)	$\|E_n\| > \|P_n\|$
	기타 (etc)	*Othercases*
자식	강한 연결	$\|E_n\| = \|C_{nk}\|$, *for some k*
	선택 연결	$\|E_n\| = \|C_{nk}\| + \|C_{nl}\|$, *for some k, l*
	모임 연결	$\|E_n\| < \|C_n\|$
	기타(etc)	*Othercases*
형제	강한 연결(S-S)	$\|E_n\| = \|S_{nk}\|$, *for some k*
	기타(child etc)	*Othercases*



[그림 1] 문서 구조의 예 (()는 빈도수)



[그림 2] 관계가 표현된 문서 구조의 예

현재 엘리먼트를 기준으로 인접한 엘리먼트는 [표 1]에 나타난 바와 같이 상위의 부모 엘리먼트, 하위의 자식 엘리먼트, 그리고 동등한 형제 엘리먼트로 나눌 수 있다. 부모 엘리먼트와의 관계는 빈도수가 같은 강한 연결, 다른 형제 엘리먼트와 빈도수를 더하여 부모 엘리먼트의 빈도수가 같게 되는 선택 연결, 부모 엘리먼트 빈도수보다

많은 모임 연결 그리고 기타로 분류할 수 있다. 자식 엘리먼트 간의 관계도 부모 엘리먼트와의 관계와 유사하고, 형제 엘리먼트 간의 관계는 형제 엘리먼트 중에서 빈도수가 현재 엘리먼트와 같은 강한 연결과 기타로 나눌 수 있다. 부모와의 관계와 자식과의 관계는 따로 발생하기보다는 형제 관계와 동시에 발생한다. 즉, 형제-강한 연결과 부모-선택 연결, 형제-기타와 부모-강한 연결 등이 동시에 발생하게 된다. 따라서 매칭 시, 이러한 관계를 어떻게 이용할 것인지 명확하게 지정해야 한다.

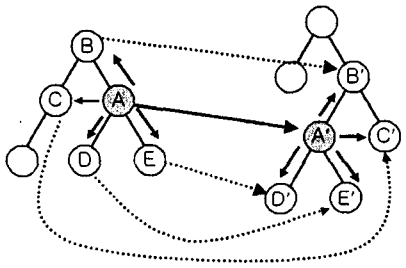
[표 1]의 관계를 예를 들어 설명하면 [그림 1], [그림 2]와 같다. [그림 1]은 개별 문서로부터 추출한 문서 구조이고, [그림 2]는 빈도수를 이용하여 추출한 엘리먼트 간의 관계를 표현한 예이다. [그림 2]에서와 같이 엘리먼트 간의 관계는 형제 관계와 부모/자식 간의 관계와 동시에 발생하는 경우가 많다.

4. XML 문서 매칭

제안된 엘리먼트 간의 연결성을 이용하여 XML 문서 간의 매칭을 수행하는 과정은 두 단계로 이루어진다. 첫번째 단계에서는 WordNet을 이용하여 동의어, 약어 정보 등을 이용한 엘리먼트 이름 간의 매칭을 수행한다. 그리고 두번째 단계에서는 엘리먼트 간의 연결성을 이용하여 인접한 매칭 정보를 이용하여 매칭이 안된 엘리먼트의 유사도를 다시 계산한다. 처음 단계에서 엘리먼트 이름을 비교하여 나오는 유사도 값은 다음과 같다.

$$Similarity_{name} = \{$$
$$1; perfect\ matching,$$
$$0.8; similarity\ matching,$$
$$0.5; partial\ matching,$$
$$0.3; etc\ matching,$$
$$0\}$$

이러한 초기 유사도 값을 이용하여 유사어 사전을 이용하여 매칭된 가장 높은 엘리먼트들을 대응시킨다. 그리고 대응되지 않는 엘리먼트들이 존재하면 인접한 매칭된 엘리먼트의 연결성 정보를 이용하여 유사도를 증가시킨다. 이 방법은 [5]와 유사하지만 [5]은 fixed point iteration을 이용하여 임의로 유사도를 증가시키는 반면에 제안 기법은 엘리먼트 간의 연결성을 추출하여 선별적으로 유사도를 증가시킨다. 즉, [그림 3]에서와 같이 첫번째 단계에서 A만 대응이 되었을 경우에 인접한 엘리먼트들의 유사도 값을 증가시킨다.



[그림 3] 매칭 선택 개념도

두번째 단계에서 사용된 매칭의 유사도는 다음과 같이 계산된다.

$$ModifiedSimilarityOfNonmatched = sim \times (w_c + w_p + w_s)/3$$

$$w_p = \{1; parent\ strong,$$
$$0.8 \times \frac{\sum(matched\ sibling\ group\ nodes)}{\sum(sibling\ group\ nodes)}; parent\ group,$$
$$0.8 \times \frac{\sum(matched\ sibling\ selection\ nodes)}{\sum(sibling\ selection\ nodes)}; parent\ selction,$$
$$0\}$$

$$w_c = \{\frac{\sum(matched\ strong\ children\ nodes)}{\sum(strong\ children\ nodes)}; strong\ children,$$
$$0.8 \times \frac{\sum(matched\ group\ children\ nodes)}{\sum(group\ children\ nodes)}; group\ children,$$
$$0.8 \times \frac{\sum(matched\ children\ selection\ nodes)}{\sum(children\ selection\ nodes)}; selection\ children,$$
$$0\}$$

$$w_s = \{\frac{\sum(matched\ strong\ sibling\ nodes)}{\sum(strong\ sibling\ nodes)}; strong\ sibling,$$
$$0\}$$

5. 결론

XML 문서로 표현된 정보들 간의 교환이나 통합 또는 필요한 정보를 추출하기 위해서는 XML 문서 간의 매칭이 필수적이다. 기존의 XML 문서 매칭은 XML 문서에서 명시적으로 표현되는 정보만을 이용하여 매칭을 수행하고, 문서로부터 파생되는 부가적인 정보들을 이용하지 않고 있다. 이에 본 논문에서는 XML 문서에서 엘리먼트 빈도수 정보를 이용하여 엘리먼트 간의 연결성을 추출한 후 이를 이용한 XML 문서 매칭 기법을 제안하였다. 엘리먼트 간의 연결성은 형제 연결과 다른 연결들이 복합적으로 발생하는 경우가 많으며, 각각의 연결이 매칭에 어떠한 영향을 끼치는 지는 향후 실험을 통하여 확인되어야 한다.

제안 매칭 기법은 XML 문서를 교환, 변환, 통합하거나 이질적인 XML 문서에서 필요한 정보를 추출할 때에도 사용될 수 있다. 또한 엘리먼트의 연결성은 기존의 매칭 기법에서 엘리먼트 간의 유사도를 계산하는데 보완적으로 사용될 수 있다.

참고문헌

1. Castano, S., V.D. Antonellis, and S.D.C.d. Vimercati, *Global Viewing of Heterogeneous Data Sources.* IEEE TKDE, 2001. **13**(2): p. 277~297.
2. Doan, A., P. Domingos, and A.Y. Halevy. *Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach.* In Proc. of *SIGMOD.* 2001..
3. Madhavan, J., P.A. Bernstein, and E. Rahm. *Generic Schema Matching with Cupid.* In Proc. of *VLDB.* 2001..
4. Hernandez, M.A., R.J. Miller, and L.M. Hass. *Clio: A Semi-Automatic Tool For Schema Mapping.* In *Proc of SIGMOD.* 2001.
5. Melnik, S., H. Garcia-Molina, and E. Rahm. *Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching.* In Proc. of *ICDE.* 2002.