온톨로지 기반 DTD 필터링 및 정합에 의한 XML 질의 시스템

김명숙*, 노영주**, 공용해*
*순천향대학교 정보기술공학부, **청양대학 컴퓨터정보과
e-mail : krhkms@sch.ac.kr

Ontology based XML Query System by DTD Filtering and Matching

Myung Sook Kim*, Young Ju Noh**, Yong Hae Kong*
*Div. of Information Tech. Engineering, Soonchunhyang Univ.,
**Dept of Computer and Info., Cheongyang Provincial College.

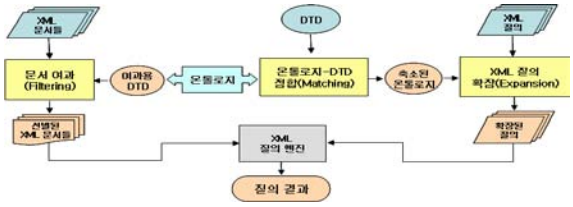
요 약

XML 문서의 논리적인 구조와 의미적 태그의 사용은 구조와 내용에 기반 한 검색을 가능하게 하는 반면, 동일한 정보라 하더라도 구조와 형식이 매우 다양하게 표현되므로 정보검색에 어려움을 초래한다. 효율적인 XML 정보검색을 위해, 본 논문은 온톨로지를 기반으로 검색에 적합한 문서만을 선별하는 문서여과 방법, 대상문서에 적합한 최소한의 질의생성을 위한 온톨로지 정합 방법 그리고 문서에 내재된 의미적 정보의 검색을 위한 정합된 온톨로지 기반의 질의확장 방법을 각각 제안하였다. 제안한 방법의 효과 및 효율은 예제 XML 및 DTD 문서를 대상으로 실험되었다.

1. 서론

XML(eXtensible Markup Language)이 문서교환의 표준으로 등장한 이후, XML로 저장되고 교환되는 데이터의 양이 늘어남에 따라 XML 문서에 대한 정보검색의 필요성이 점차 증대되고 있다. XML 문서의 논리적 구조와 의미적 태그의 사용은 구조와 내용에 기반 한 검색을 가능하게 하는 반면, 동일한 정보라 하더라도 구조와 형식이 매우 다양하게 표현되므로 정보검색에 어려움을 초래한다[1,2,3]. 이러한 XML 문서의 구조적 다양성에 기인한 정보검색의 한계를 극복하기 위해, 본 논문은 온톨로지 기반의 XML 문서선별 방법과 대상 XML 문서에 적중될 수 있는 최소의 질의생성 방법을 제안하였다. 온톨로지는 현재 W3C 등에서 영역에 대하여 표준화 작업이 진행되고 있으므로, 온톨로지를 기반으로 XML 문서를 접근하려는 시도는 보다 활발해질 것이며 XML 문서의 구조적 다양성을 해결하는 기반이 될 것이다[4].

효율적인 XML 정보검색을 위해, 본 연구는 그림 1과 같이 세 개의 알고리즘을 체계적으로 구성하였다. 우선, 검색에 불필요한 문서를 사전에 여과하고, 다음으로 온톨로지를 검색대상 문서의 구조와 일치하도록 DTD와 정합한다. 마지막으로, 정합된 온톨로지를 사용하여 의미적 정보검색이 가능한 최소한의 질의로 XML 질의를 확장한다.



(그림 1) 온톨로지 기반의 XML 질의검색 시스템

XML 정보검색은 관심영역에 포함되지 않는 부적합한 문서로의 검색을 제거하기 위해, 질의검색 전에 적합한 문서들만을 선별해야 한다. 이를 위하여 온톨로지를 기반으로 여과용 DTD를 생성하고, 여과용 DTD를 사용하여 검색 영역에 포함되는 XML 문서를 선별한다. 또한 XML 문서의 구조적 차이를 극복하고 보다 폭넓은 정보를 검색하기 위해 온톨로지로부터 질의를 확장하는 연구가 수행된 바 있다

본 논문은 정보통신부 정보통신연구진흥원에서 지원하고 있는 정보통신기초기술연구지원사업의 연구결과입니다.

[5]. 이는 온톨로지의 구조 및 상속 관계를 이용하여 질의를 하위로 확장함으로써 다수의 질의를 생성하였으며, 구조적 한계를 극복하고 보다 폭넓은 정보의 검색을 가능하게 하였다. 그러나 온톨로지 기반의 질의확장은 대상 문서의 구조를 고려하지 않으므로 XML 문서에 존재하지 않는 정보에 대한 부적합한 질의를 과도하게 생성함으로써 검색 적중률을 저하시켰다. 따라서 대상 XML 문서에 종속적인 질의의 생성을 위해서는 온톨로지뿐만 아니라 대상 문서의 구조를 정의한 DTD도 함께 고려되어야 한다. 따라서 본 논문은 온톨로지를 대상 DTD와 정합함으로써 온톨로지가 대상 문서의 구조와 부합되도록 축소한다. 축소된 온톨로지를 기반으로, XML 질의는 온톨로지의 개념상속 및 연관관계를 고려한 질의확장 알고리즘에 의해 최소로 확장된다. 확장된 질의들은 선별된 XML 문서에 내재된 의미정보를 효과적으로 검색하고 질의검색의 적중도를 증대시킴으로써 검색의 효율을 높일 수 있다. 제안한 알고리즘의 효율 및 효과는 다양한 구조의 예제 XML 문서들을 대상으로 실험된다.

2. XML 문서 필터링을 위한 여과용 DTD 생성

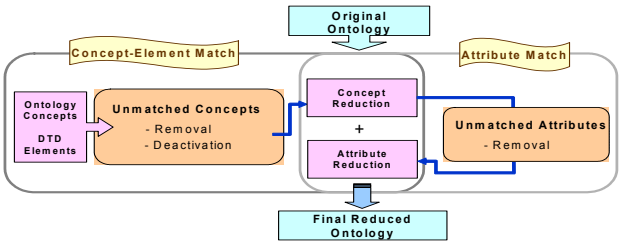
검색영역과 무관한 XML 문서를 여과할 수 있는 방법이 제공된다면, 불필요한 문서에 대한 XML 질의검색은 시도되지 않을 것이다. 질의검색 전에 대상문서를 선별하기 위해, 동일영역의 모든 문서를 포괄할 수 있는 여과용 DTD는 그림 2와 같은 과정에 의해 생성된다.

(그림 2) 여과용 DTD 생성 과정

3. 온톨로지-DTD 정합

온톨로지로부터 확장된 질의는 XML 문서의 폭넓은 정보를 검색할 수 있지만, 대상문서의 구조와 속성을 고려하지 않기 때문에 문서에 존재하지 않는 엘리먼트 및 속성에 대한 불필요한 검색을 시도할 수 있다. 온톨로지만을 기반으로 질의를 확장할 경우 다음과 같은 문제가 발생한다. 첫째는 온톨로지와 DTD의 구조적 차이에 의해 존재하지 않는 개념에 대하여 질의를 확장하는 것이고, 둘째는 온톨로지와 DTD 간의 속성의 불일치 및 온톨로지의 속성상속에 의해 부적합하게 질의를 확장하는 것이다.

이러한 문제들을 해결하기 위해, 대상 XML 문서의 구조를 고려한 온톨로지-DTD 정합 알고리즘을 제안하였다. 온톨로지-DTD 정합 알고리즘은 세부적으로 온톨로지와 DTD의 개념-엘리먼트 정합, 온토롤지와 DTD의 속성정합으로 구성된다. 개념-엘리먼트의 정합은 온톨로지를 XML 문서 구조인 DTD와 비교하여 일치하지 않는 개념을 찾아 모두 제거하는 것이다. 또한 속성정합은 온톨로지와 DTD의 일치하지 않는 속성을 찾아 제거하는 것이다. 그림 3은 온톨로지-DTD 정합 과정이며, 표 1과 2는 각각의 알고리즘을 나타낸 것이다.



(그림 5) 온톨로지-DTD 정합 알고리즘

(표 1) 온톨로지-DTD 개념-엘리먼트 정합 알고리즘

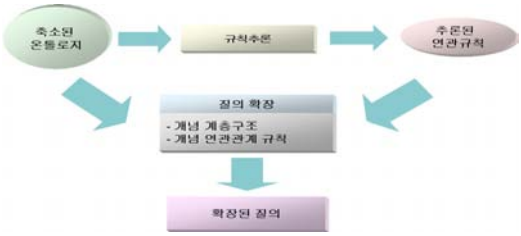
```
For all concepts, select a concept,
    Compare a concept with all elements of DTD.
    If the concept is not part of the DTD,
        If the concept is leaf concept,
            Remove the concept and concept's attributes from ontology.
        Otherwise, the concept is made inactive.
```

(표 2) 온톨로지-DTD 속성정합 알고리즘

```
For all the ontology concepts,
    For ontology concept A and its corresponding DTD element B,
        1. For each attribute P in common concept A and element B,
            Set association between attribute P and concept A.
        2. For each attribute Q of element B which is not contained in concept A,
            For each ancestor concept C of concept A,
                If attribute Q is included in concept C's attributes,
                    Set association between C's attribute Q and concept A.
```

4. 축소된 온톨로지 기반의 XML 질의확장

XML 문서의 정보를 검색하기 위해 축소된 온톨로지의 개념 계층구조와 연관관계를 이용하여 질의를 확장하였다. 온톨로지는 특정영역에 대한 개념과 속성을 정의한 것으로 개념 간에는 계층구조가 존재하며 이러한 구조를 이용하여 질의를 확장할 수 있다. 또한 개념 계층구조 사이에서 추론될 수 있는 연관관계는 의미정보를 내포하며, 이러한 연관관계 규칙을 이용하여 질의를 확장하면 XML 문서의 의미적 정보검색이 가능하다. 그림 4는 온톨로지를 이용한 XML 질의확장 과정이다.



(그림 4) 축소된 온톨로지를 이용한 XML 질의 확장

표 3은 계층구조에 의한 질의확장 알고리즘, 표 4는 연관관계 규칙추출 알고리즘, 그리고 표 5는 추론된 규칙에 의한 질의확장 알고리즘이다.

(표 3) 개념 계층구조에 의한 질의확장 알고리즘

1. If query contains no attribute at all, for all concepts of a query,
   1.1 Select a concept.
   1.2 Search for all its descendant concepts.
   1.3 If selected concept and its descendant concepts is not inactive concept, add the descendant concepts to the query.
2. If query contains any attribute,
   For all concepts associated with the selected attribute in reduced ontology,
   If each concept is a descendant or one of concepts within user query,
   add the concept to the query set.

(표 4) 연관관계 규칙추출 알고리즘

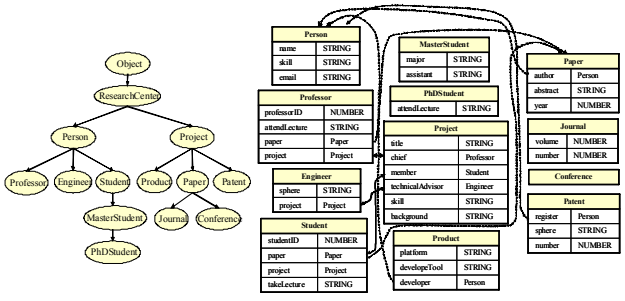
For all concepts of ontology,
1. Select a concept A.
2. For all attributes of concept A, select an attribute B,
3. Search concept C that matches attribute B.
4. Set association between concepts A and C.

(표 5) 연관관계 추론규칙에 의한 질의확장 알고리즘

For all concepts in query, select a concept
For the selected concept, if there is association rules,
add the associated concept to query recursively
For each all concepts in original query and added concepts,
Expand queries using query-expanding algorithm of table 3.

5. 실험

본 논문에서 제안한 XML 문서 필터링, DTD 정합 그리고 질의확장 알고리즘을 몇 개의 예제 XML 문서를 통하여 실험하였다. 본 실험을 위해 그림 5의 "대학연구센터" 온톨로지를 사용하였으며, 그림 5(a)는 개념 계층구조를 나타내며, 그림 5(b)는 개념 및 속성의 연관관계를 나타낸다.



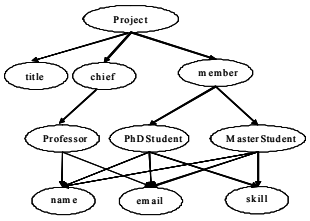
(a) 온톨로지의 개념 계층구조 (b) 개념에 대한 속성 및 연관관계

(그림 5) "대학연구센터" 온톨로지

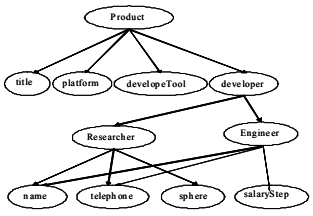
XML 문서 필터링을 위해 "대학연구센터" 온톨로지로부터 DTD 생성 알고리즘에 의해 여과용 DTD를 생성하였다. 표 6은 생성된 여과용 DTD의 일부이며, 그림 5(a), (b), (c), (d)는 문서필터링 실험에 사용될 예제 XML 문서들의 DTD 구조이다.

(표 6) XML 필터링을 위한 여과용 DTD의 일부분

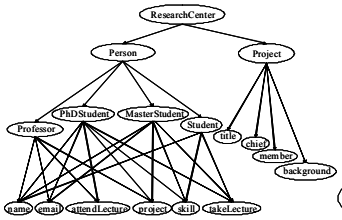
```
<!ENTITY % Person "Person|Student|PhDStudent|MasterStudent|Engineer|Professor" >
<!ENTITY % Student "Student|PhDStudent|MasterStudent" >
                                ...
<!ELEMENT Person (#PCDATA|name|skill|email)* >
<!ELEMENT Professor (#PCDATA|name|skill|email|professorID|attendLecture|paper|project)* >
                                ...
<!ATTLIST Project
   title                CDATA    #IMPLIED
   chief                CDATA    #IMPLIED
   member               CDATA    #IMPLIED
   technicalAdvisor     CDATA    #IMPLIED
   skill                CDATA    #IMPLIED
   background           CDATA    #IMPLIED >
                                ...
<!ELEMENT paper (#PCDATA | %Paper;)* >
<!ELEMENT project (#PCDATA | %Project;)* >
```



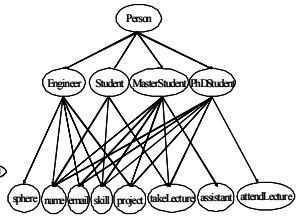
(a) 예제 DTD1



(b) 예제 DTD2



(c) 예제 DTD3



(d) 예제 DTD4

(그림 6) 예제 XML 문서들의 DTD 구조

그림 6의 예제 DTD의 구조를 살펴보면, 그림 6(a) DTD1과 그림 6(b) DTD2의 구조는 매우 유사하게 보인다. 그러나 두 DTD는 서로 다른 영역을 정의한 것으로서 온톨로지의 관점에서 서로 다른 구조를 가지고 있다. 반면에, 그림 6(c)와 6(d)의 DTD는 그림 6(a)의 DTD1 구조에 비해 매우 복잡하고 그 구조가 매우 다르게 보이지만. 모두 "대학연구센터" 영역을 정의한 동일한 DTD들이다. 따라서 그림 7과 같이 여과용 DTD에 의해 DTD2를 제외한 DTD1, DTD3, 그리고 DTD4는 모두 검색대상에 적합한 문서로 선별되었다.

(a) 선택된 DTD1

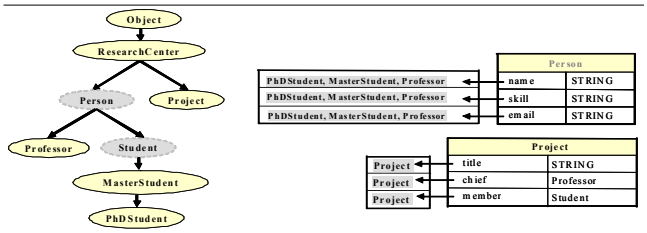
(b) 여과된 DTD2

(c) 선택된 DTD3

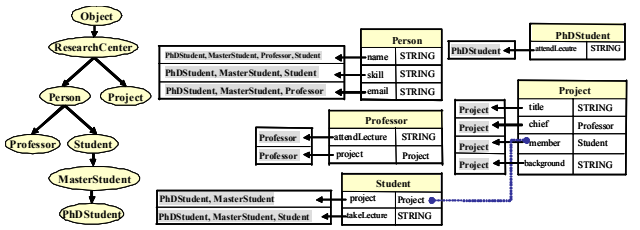
(d) 선택된 DTD4

(그림 7) 문서 필터링 결과

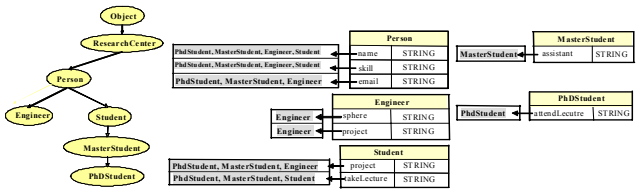
3장의 온톨로지-DTD 정합 알고리즘은 필터링 실험에서 선별된 DTD1, DTD3, DTD4와 그림 5의 온톨로지를 대상으로 적용되었으며, 온톨로지가 3개의 DTD들과 각각 정합되어 그림 8과 같이 축소되었다.



(a) DTD1-정합에 의해 축소된 온톨로지



(b) DTD3-정합에 의해 축소된 온톨로지



(b) DTD4-정합에 의해 축소된 온톨로지

(그림 8) 온톨로지-DTD 정합 결과

축소된 온톨로지에 의해 확장된 질의검색의 효율은 4장의 질의확장 알고리즘에 의해 실험되었다. 표 7은 단순질의를 원형 온톨로지와 축소된 온톨로지에 각각 인가하여 확장된 결과를 비교한 것이다. 인가된 단순질의에 포함된 총 3개의 질의에 대하여, 온톨로지를 사용한 경우 총 23개의 질의로 확장되었으며, 축소된 온톨로지의 경우 총 5개의 질의로 확장되었다. 이러한 결과는 온톨로지로부터 확장된 23개의 질의 중 단지 5개만이 XML 문서에 적합하며 나머지 18개의 질의는 대상문서의 구조에 적합하지 않은 불필요한 질의임을 나타낸다.

(표 7) 질의확장 결과 비교

질의항목 / 대상 DTD		(1) 인가된 단순 질의 (2) 원형 온톨로지에 의한 질의확장 결과 (3) 축소된 온톨로지에 의한 질의확장 결과
DTD1 (그림 5a)	(1)	//Person[name="Smith"]
	(2)	//(Person|Professor|Student|MasterStudent|PhDStudent)[name="Smith"]
	(3)	//(Professor|MasterStudent|PhDStudent)[name="Smith"]
DTD3 (그림 5c)	(1)	//Student[project]
	(2)	//(Student|MasterStudent|PhDStudent)[project], //(Project|Product|Paper|Patent|Journal|Conference)[member]
	(3)	//(PhDStudent|MasterStudent)[project]
DTD4 (그림 5d)	(1)	//Project[member]
	(2)	//(Project|Product|Paper|Patent|Journal|Conference)[member], //(Student|MasterStudent|PhDStudent)[project]
	(3)	No Query

결과적으로, 제안된 온톨로지-DTD 정합 방법은 확장된 질의가 검색 대상으로 선별된 XML 문서에 적중될 수 있도록 온톨로지를 효과적으로 축소하였으며, 축소된 온톨로지는 대상 문서에 불필요한 질의를 대부분 제거하고 검색에 성공할 수 있는 적합한 질의만을 생성함으로써 정보검색의 효율을 향상시켰다.

6. 결과

XML 문서는 그 구조와 속성이 매우 다양하므로 질의검색에 한계가 있다. 효율적인 XML 정보검색을 위해, 본 논문은 온톨로지를 사용하여 보다 체계적인 질의검색 방법을 제안하였다. 제안된 방법은 불필요한 문서의 필터링, 대상문서에 적합한 온톨로지 정합 그리고 효과적인 XML 질의확장 등으로 구성되었다. 필터링을 위한 여과용 DTD는 온톨로지로부터 생성되고, 생성된 여과용 DTD를 사용하여 검색영역에 포함되지 않는 XML 문서를 사전에 여과한다. 또한 대상문서에 종속적인 질의만을 생성하기 위해서, 온톨로지는 검색대상 문서의 구조인 DTD와 정합되어 축소된다. 그리고 축소된 온톨로지의 개념 상속 및 연관관계 특징을 고려하여 인가된 질의가 대상문서에 종속적인 최소한의 질의로 확장된다.

실험 결과, 제안된 방법은 온톨로지를 효과적으로 정합하였고 선별된 XML 문서에 적합하지 않은 불필요한 질의를 효과적으로 제거하였다. 결과적으로 검색 영역의 DTD와 정합된 온톨로지는 대상 XML 문서의 구조에 적합한 질의를 생성함으로써 검색 적중도를 높이고, 동일한 DTD를 갖는 사이트 내의 모든 XML 정보검색에 재사용됨으로써 효율적인 XML 정보검색을 가능하게 하였다.

참고문헌

[1] T. Bray, et al., Extensible Markup Language (XML) 1.0, W3C Recommendation 04 February 2004.

[2] M. Altinel, M.Franklin. "Efficient Filtering of XML Documents for Selective Dissemination of Informatuon", In VLDB, 2000.

[3] 김영란, "XML DTD의 효율적인 검색을 위한 구조 정보 및 인덱스 메커니즘", 한국컴퓨터정보학회 논문지, 8권, 3호, pp.80-86, 2003.

[4] D. L. McGuinness, et al., OWL Web Ontology Language Overview, W3C Recommendation, 10 February 2004.

[5] M. Erdmann and S. Decker, "Ontology-award XML Queries", WebDB, 2000.