XML 문서 검색을 위한 한국어 질의 처리 시스템

박춘용*, 이현영*, 윤보헌**, 강현규**, 이용석*
전북대학교 컴퓨터과학과*, 한국전자통신연구원 지식정보연구부 문서정보연구팀**
cypark@cs.chonbuk.ac.kr

Korean Query Processing System for XML Document Retrieval

Chun-Yong Park*, Hyeon-Yeong Lee*, Bo-Hyun Yun, Hyun-Kyu Kang**, Yong-Seok Lee*
Dept. of Computer Science, Chonbuk National University*
Document Information Research Team, Dept. of Knowledge Information, ETRI**

요 약

인터넷 문서의 표준 사양인 XML 문서가 늘어나면서 XML 문서를 효과적으로 관리하고 검색하기 위한 시스템이 개발되고 있다. 그러나 정형화된 질의언어를 사용한 XML 문서의 검색 방법은 질의언어의 구조를 이해하고 사용법을 숙지해야 하는 어려움이 있어 일반 사용자에게는 적합하지 않다. 따라서 사용자가 쉽게 사용할 수 있으면서도 정확한 결과를 가지는 시스템이 요구된다. 본 논문에서는 XML 문서를 검색하기 위해 자연어로 질의를 입력하면 이를 XML 구조 검색을 위한 XQL 언어로 자동 변환해 주는 시스템을 제안한다. 제안한 시스템은 자연어를 이용하기 때문에 사용하기가 쉽고 XML 문서 구조가 변경되어도 쉽게 확장할 수 있는 장점을 가진다.

1. 서론

인터넷 기반 문서의 표준으로 자리잡은 XML은 인터넷을 이용한 문서의 결재나 공문 발송을 하기 위한 수단으로 사용되는 추세이다[1]. 따라서 인터넷에 분산된 XML 문서 중에서 자신에게 필요한 문서만을 검색하는 기술이 요구되고 있다. 그러나 지금까지 개발된 XML 문서의 검색 시스템은 XQL이나 XML-QL과 같은 질의언어를 이용하기 때문에 질의어의 구조를 이해하지 못한 일반 사용자가 사용하기에는 매우 어렵다. 또한 브라우저를 이용하는 경우에도 사용방법을 알아야 하며 새로운 구조의 DTD(Document Type Definition)가 만들어지면 브라우저를 다시 만들어야 한다.

```
논문 -> 전방 -> 논문지정보 -> 잡지명
                          -> 출판일
                            :
             -> 제목그룹 -> 제목
                          -> 부제목
                            :
             -> 저자그룹 -> 저자
                          -> 소속
             -> 날짜
             -> 출판전방 -> 후원
                          -> 출판내력 -> 접수일
                                      -> 승인일
                            :
             -> 요약
      -> 본문 -> 부 -> 번호
        :
```

[표 1]정보과학회 논문지 DTD 논리 구조의 일부

XML은 추상화된 정보 표현의 기본 단위가 문서가 아니라 엘리먼트(element)이다. 엘리먼트는 문서의 구조를 논리적으로 정의하는 DTD에 의해 알려질 수 있으므로 문서의 내용에 기반한 검색 방법 이외에 문서 구조에 의한 검색을 지원한다[2,3]. 예를 들어 [표 1]은 정보과학회 논문지에 수록된 논문을 구조 검색하기 위해서 논문지 DTD에 포함된 엘리먼트들 중 일부를 논리적 구조로 표현한 것이다[4].

[표 1]과 같이 구조화되어 있는 DTD 정보를 활용하면 XML 문서를 검색하기 위해 자연어 질의를 활용할 수가 있다. 자연어 질의를 사용하게 되면 사용자가 이해하기 쉬울 뿐만 아니라 DTD의 구조가 바뀌어도 쉽게 확장할 수가 있다. [표 1]의 DTD 지식을 이용하여 구조화된 자연어 질의의 예는 다음과 같다.

1) 정보과학회논문에서 **저자**가 홍길동인 논문을 찾아라
2) 정보과학회논문에서 한국과학재단의 **후원**을 받는 논문을 모두 찾아라
3) 저자의 **소속**이 한국대학이고 **제목**이 문서검색인 논문을 찾아라

이와 같이 구조 검색의 정확성과 확장성 및 사용자의 편의를 고려하여 XML 문서를 검색하기 위해서는 사용자가 자연어로 질의를 하면 질의 처리 시스템이 이를 구조 질의로 변환하는 방법이 현재로서는 가장 적절한 대안으로 떠오른다

[4]. 본 연구에서는 인터넷에 분산된 XML 문서를 위한 효율적인 검색 방법의 대안으로 특정 구조 질의 방법이 아닌 자연어에 기반한 XML 구조 질의 처리 시스템을 제안한다.

2. XML 구조 질의 유형

XML은 문서의 계층적 구조를 표현하기 위해 문서의 구조를 논리적으로 표현한 DTD(Document Type Definition)를 가지고 있다[1]. DTD는 문서 내에 있는 요소들간의 구조 정보로 XML 문서를 검색할 때에 문서 전체가 아닌 부분 항목들로 처리함으로써 사용자 질의에서 원하는 특정 영역에 바로 접근할 수 있는 구조 기반 정보 검색을 가능하게 한다[3,5,6].

그러나 구조화된 DTD 정보를 활용하기 위해서는 구조화된 자연어 질의를 이용해야 한다. 즉, 구조화된 DTD 정보를 활용할 수 없는 유형의 질의는 의미가 없기 때문에 DTD 정보를 활용해서 XML 문서를 검색할 수 있는 구조화된 자연어 질의를 사용한다. 본 논문에서는 구조화된 자연어 질의 형태로 사용자가 해답을 요구하는 형태와 구체적인 객체에 대한 질의 유형을 사용한다. 한국어 질의 처리 시스템은 구조화된 자연어 질의를 분석해서 XML 문서를 검색하기 위한 질의언어인 XQL로 변환한다.

XML 문서의 구조 검색 질의를 위해 사용되는 한국어 질의는 다음과 같은 유형을 가질 수 있다.

가) ~에서 ~이 있는(들어있는) ~을 찾아라/ 보여라/ 보여줘
- *논문에서 정보 그리고 검색이 들어있는 장을 보여줘*

나) ~에서 (숫자)이내에 ~:(가중치)가 들어있는 ~을 찾아라.
- *논문에서 5이내에 정보:0.5가 들어있는 문단을 찾아라*

다) ~에서 애트리뷰트 ~가 값으로 ~를 갖는 ~엘리먼트를 찾아라.
- *논문에서 애트리뷰트id가 값으로 ch01을 갖는 모든 엘리먼트를 찾아라.*

라) ~를 제외한/포함하지않는 ~을 찾아라
- *논문에서 시스템을 제외한 장을 찾아라*

마) 가) - 라)에서 용언이 생략된 유형
- *논문에서 정보 그리고 검색이 들어있는 장은*

3 XML 구조 질의 시스템

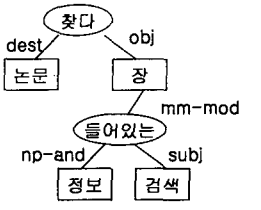
XML 구조 질의를 XQL로 변환하기 위해서는 DTD 정보를 이용해서 구조질의를 XQL로 사상하는 과정이 필요하다. 이를 위해서는 XML 구조 질의를 분석해서 키워드를 추출하고 연산자를 분리한 후에 이들 정보를 이용해서 XQL로 변환해야 한다.

3.1 키워드 추출

형태소 해석 결과를 이용해서 한국어 질의를 불리언 질의로 변환하는 방법은 주어진 질의어에 대해 신속하게 적절한 검색어로 변환할 수 있지만 결과의 정확성은 저하된다. 본 연구에서는 구문 형태소를 이용한 구문 분석[6]을 통해 문법적 관계까지를 고려하여 정확한 검색 질의어를 생성한다.

키워드를 추출하는 방법은 다음과 같다. 질의어의 구문분석 결과인 파스트리를 순회하며 단말노드인 경우 범주(category) 정보를 검사하여 보통 명사나 고유 명사 또는 이들을 포함하는 체언구로 분석된 경우 키워드로 추출한다. 관형격으로 사용된 비단말 노드인 경우는 후행하는 체언구를 키워드로 추출한다. 예로 예문 4)의 경우 파스트리는 [그림 1]과 같고 단말노드에 의해서는 "논문, 정보, 검색"이 추출되고 비단말 노드에서는 "장"이 키워드로 추출된다.

4) 논문에서 정보 그리고 검색이 들어있는 장을 찾아라



[그림 1]예문 4)의 파스트리

3.2 연산자 결정

키워드들 사이의 연산자는 명사에 붙는 조사나 관형형 어미, 부사격 어미 또는 접속사에 의해 결정된다. 예를 들면 "정보 그리고 검색"에서는 접속사 '그리고'에 의해 AND 연산자로 결정되고 "정보나 검색"에서는 조사 '나'에 의해 AND 연산자로 결정된다. 또한 문장에서 사용되는 체언구의 기능을 보고 결정하는 경우도 있다 예를 들면, [그림 1]에서 '논문에서'는 'dest'격을 가지므로 DTD를 지정하도록 한다. 이와 같이 XML 구조 질의를 분석한 결과 필요한 연산자는 [표 2]와 같다.

종류	설 명	
지정 연산자	엘리먼트 지정	IN
	DTD 지정	OF
	모두(all) 지정	*
불리언 연산자	AND 연산	and, AND, &
	OR 연산	or, OR, |
	NOT 연산	andnot, ANDNOT, !
애트리뷰트명 표시	@	
선택	[ ]	

[표 2]자연어 구조 질의를 XQL로 변환하기 위한 연산자

불리언 연산자를 결정하기 위해서는 [표 3]의 정보를 활용해서 체언구의 기능을 설정한다. 예를 들어 [그림 1]에서 "정보 그리고 검색"과 같이 'AND'의 구조이면 '정보'에 'np-and'라는 기능을 할당해서 파스트리를 탐색할 때 이용하

도록 한다. 또한 지정 연산자는 [표 4]의 구문 정보를 이용한다.

OR	AND	ANDNOT
A 나 B	A와 B // A의 B	A를 제외한
A 거나 B	A B // A, B	A가 제외된
A 혹은 B	A 그리고 B	A가 아닌
A 또는 B	A 방식의 B	A를 포함하지 않는
	A를 이용한 B	A가 포함되지 않는
	A를 위한 B	A를 뺀
	A에 사용되는 B	A가 빠진
	A (분야) 중 B	A 이외의
	A에 대한 B	A를 갖지 않는
	A 및 B // A 또 B	A가 들어있지 않는

[표 3]불리언 연산자의 구성

연산자	구문 기능	예
IN	mm-mod	들어있는, 포함하는, 있는, ...
OF	dest	에서, 서, 에, ...
*	all	모든, 모두, 전부, ...
AND	np-and	그리고, 와, 의, 또, 및, ...
OR	np-or	나, 또는, 혹은, ...
ANDNOT	np-not	-를 제외한, -가 아닌, ...
NEAR	np-near	-이내의, -내외의, -내의, ...
=	np-value	값으로, 값인, 값을 갖는, ...

[표 4]구문적 기능에 의한 연산자 결정

3.3 XQL로의 변환

자연어 질의를 분석해서 키워드가 추출되고 연산자가 결정되면 XQL로 변환할 수가 있다. XQL로의 변환은 구문분석 결과인 파스트리를 좌에서 우로 순회하면서 키워드와 연산자를 추출한다. [그림 1]의 결과를 이용해서 XQL로 변환하는 과정을 살펴보면 다음과 같다. 참고로 본 논문에서 사용하는 XQL의 BNF 표기는 [표 5]와 같고 [그림 1]의 XQL은 (a)의 구조로 변환된다.

가) 키워드로 '논문'을 추출한 후 키워드의 기능정보인 'dest'에 의해 지정 연산자 'OF'를 결정하고 DTD_name으로 키워드인 '논문'을 할당한다. XQL의 변환 결과는 **'OF 논문'**이 생성된다.

나) 키워드로 '정보'를 추출하고 기능정보 'np-and'에 의해 불리언 연산자 'AND'를 결정한다. 'AND'는 이항연산자이므로 다음의 체언구 '검색'을 읽어 **'정보 AND 검색'**을 생성한다.

다) '들어있는'이 'mm-mod'이므로 연산자 'IN'을 선택한다. 'IN'은 element 값을 필요로 하므로 관형구가 수식하는 체언구 '장'을 읽어 element 값으로 할당한다. 그러면 **'IN 장'**의 결과가 구해진다.

라) 구해진 결과를 [표 5]의 (a)에 맞도록 조정한다. 결과는 다음과 같다.

결과) **정보 AND 검색 IN 장 OF 논문**

```
Start := Vec_Query
Vec_Query := Query Vec_Query
Query := query AND query [IN element_name] [OF DTD_name] -- (a)
   | query OR query [IN element_name] [OF DTD_name]
   | query ANDNOT query [IN element_name] [OF DTD_name]
   | query_weight [IN element_name] [OF DTD_name]
   | literal [IN element_name] [OF DTD_name]
   | element_name{@attribute_name = attribute_value} [OF
     DTD_name]
query_weight := query :weight | query
literal := term | near term term number
        | withins term term number
term := keyword | keyword :weight
element_name := keyword
attribute_name := keyword
DTD_name     := keyword
Attribute_value := keyword
```

[표 5]XQL BNF Form

다른 유형의 자연어 질의를 XQL로 변환한 예를 들면 다음과 같다.

- 정보과학논문에서 애트리뷰트 id가 값으로 ch01를 갖는 장 엘리먼트를 찾아라.

  XQL : 장{@id = ch01} OF 정보과학논문

- 정보과학논문에서 정보:0.7와 검색:1.0이 들어있는 장을 찾아라.

  XQL : 정보:0.7 and 검색:1.0 IN 장 OF 정보과학논문

4. 결론 및 향후 연구

XML 문서를 검색하기 위해서 구조화된 한국어 질의를 구문 분석하여 구조 검색 질의 처리가 가능한 XQL로 변환해 주는 시스템을 개발했다. 이는 자연어로 질의를 표현함으로써 얻는 장점과 XQL을 이용한 구조질의가 가지는 XML 문서 검색의 장점을 결합할 수 있다. 이러한 방법은 XML 문서의 검색뿐만 아니라 정보검색을 이용한 다양한 응용분야에서 한국어 인터페이스의 한 모듈로써 이용될 수 있을 것이다.

또한 본 시스템은 구문 분석 결과를 이용하기 때문에 보다 정확한 키워드의 추출과 다양한 연산자의 생성이 가능했다. 그러나 XML 문서검색뿐만 아니라 다양한 문서를 검색하기 위해서는 구조화된 질의어뿐만 아니라 다양한 질의를 처리할 수 있도록 시스템의 확장이 필요하다. 아울러 구문 정보를 이용한 다양한 연산자 정보 추출에 대한 연구가 필요하다.

참고문헌

[1]http://www.w3.org/XML/

[2]http://www.w3.org/TandS/QL/QL98/pp/xql.html

[3]이희주, 장재우, 심부성, 주종철, "구조화 문서를 위한 정보검색 인덱스의 설계", 한국정보과학회 추계학술발표논문집, Vol. 24, No. 2, pp. 337-340, 1997.

[4]한국전자통신연구원, "자연어 질의의 구문분석을 통한 XML 구조 검색 질의 처리기의 개발", 한국전자통신연구원 최종보고서, 2000.

[5]이계준, 신동욱, 권택근, "XML 문서의 검색을 위한 효율적인 색인 기법과 질의언어(TQL)의 설계", 한국정보과학회 추계학술발표논문집, Vol. 26, No. 2, pp. 57-59, 1999.

[6]Hyeon-Yeong Lee, Yi-Gyu Hwang, Woo-Jeong Bae and Yong-Seok Lee, "Unification Based Korean Parsing Using Sentence Patterns Information", NLPRS'99, pp.150-155, 1999.