다중 문서에서 구조 정보를 이용한 XML 조인 질의 처리

정성호°*, 김병곤**, 정헌석*, 이재호***, 임해철*
*홍익대학교 컴퓨터공학과, **부천대학 사무자동학과
***인천교육대학교 컴퓨터교육과
{shjung, hschung, lim}@cs.hongik.ac.kr, bgkim@bc.ac.kr, jhlee@mail.inue.ac.kr

XML Join Query Processing using Structured Information from Multiple Documents

Sungho Jung°*, ByungGon Kim**, HunSuk Chung*, Jaeho Lee***, HaeChull Lim*
*Dept. of Computer Eng., Hong Ik Univ., **Dept. of OA, Bucheon College
***Dept. of Computer Education, Inchon National University of Education

요 약

XML 문서에 대한 다양한 질의를 위해서 W3C에서는 XQL, XML-QL, XML-GL, XQUERY와 같은 질의어를 제안하였다. 이들 질의어는 다양한 질의 유형의 분류와 표현은 가능하나, 조인 질의의 경우 단순 조인 질의만을 지원할 뿐, XML 문서의 구조나 텍스트 정보의 유사성을 이용한 보다 다양한 조인 질의에 대한 연구가 미비하였다. 본 논문에서는 다중 문서에 대한 조인 질의를 체계적이고 효과적으로 표현하기 위해, 문서에 대한 조인 질의를 여러 타입으로 분류하였다. 또한 효율적인 질의처리를 위하여 다양한 일반 조인 질의 및 정보검색 기능을 지원하는 유사성 조인 연산자(similarity join operator), 순수 구조 기반 조인을 지원하는 구조 조인 연산자(structured join operator)를 지원하도록 XML 질의어인 QUILT를 확장하였다. 특히, 구조 정보만을 이용한 질의시 구조의 깊이(depth)정보를 이용하여 사용자의 요구에 맞게 질의 검색 범위를 설정하고, XML 문서에 대한 질의문을 좀더 간결하게 표현할 수 있도록 설계하였다.

1. 서 론

최근 정보의 양이 급증하면서 수많은 문서 정보를 인터넷 전자문서로 만들고 관리하는 연구가 많은 분야에서 활발히 진행되고 있다. 특히, 구조적 데이터 표현이 가능한 XML은 대표적인 전자문서의 표준으로 각광받고 있다.

XML 문서에 대한 다양한 질의를 위해서 W3C에서는 LOREL[1], XQL[2], XML-QL[3], XML-GL[1], QUILT[4], XQUERY[5]와 같은 XML 질의어들을 발표하였다. 이러한 질의어들은 엘리먼트 내용 정보 검색, 애트리뷰트 정보 검색과 같은 XML 문서의 특성에 맞는 검색을 지원한다. 이 외에도 XML 질의의 다양한 유형을 분류하고 표현하기 위한 많은 연구가 있었다. 하지만 단순히 단일 문서 내에서의 조인 질의나 여러 문서간의 일반적인 조인에 대한 연구가 대부분 이었다.

따라서 본 논문에서는 다중 문서에서의 조인 질의 유형을 체계적으로 분류하고, 질의 내용에 따른 다양한 문서 조인을 통하여 복잡한 질의를 처리 할 수 있는 질의문을 설계하였다. 또한 유사성 조인 연산자(similarity join operator)를 이용하여 정보 검색을 지원하고, 일반적인 엘리먼트 기반의 조인이 아닌, XML 문서의 구조적 정보인 조상/자손 관계, 부모/자식관계, 형제관계등과 같은 순수 구조 정보를 이용한 조인 질의를 지원하도록 확장하였다. 즉, 두 문서의 구조 관계만을 가지고 조인할 수 있기 때문에 구조는 같지만 상이한 내용을 가지고 있는 문서에서도 질의 검색을 할 수 있도록 구조 조인 연산자(structured join operator)를 이용하여 XML 질의어인 Quilt 문법을 새롭게 정의하였다. 또한, 구조 질의시 깊이(depth) 정보를 이용하여 질의 대상 범위의 설정이 가능하다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 여러 XML 질의어의 특징과, XML 질의어에 추가된 정보검색기능을 소개하고 그 제한점을 지적한다. 3장에서는 다양한 유형의 조인 예제를 제시하고, 본 논문에서 제시한 유사성 조인 질의와 구조 조인 질의에 대하여 알아보고, 결론을 맺는다.

본 연구는 정보통신연구진흥원 대학기초연구지원사업 (과제번호 : C1-2002-122-3) 의 지원을 받았음

2. 관련 연구

XML 문서의 질의어들은 LOREL[1], XQL[2], XML-QL[3], QUILT[4], XQUERY[5]등이 있다. 이 질의어들은 많은 양의 XML 문서로부터 데이터를 추출하거나, XML 데이터를 서로 다른 DTD들 사이에서 해석하기 위해서, XML 문서의 특성을 반영한 내용/구조 기반 검색들을 지원한다. 조인 질의의 경우에 있어서는 질의 내용에 따른 단일 문서 또는 다중 문서에서 값을 통하여 엘리먼트 매칭으로 조인을 표현하였다. <표 1>은 제안된 XML 질의어를 지원하는 조인 항목별로 분류·비교한 것이다. XQL[2]는 1998년 W3C에 의해 제안된 범용적 질의 언어로 경로 표현(path expression)을 이용하여 질의 대상을 표시하며, XQL은 세미-조인만을 지원하였으나, 최근에는 다중 문서의 경로(path)내에서 상관관계 변수를 이용하여 원 문서와 다른 문서들과의 조인을 지원한다. XML-QL[3]은 임의의 조인 조건을 정의하여 단일 문서나 여러 다른 문서에서 데이터를 통한 엘리먼트 매칭을 통하여 같은 변수 이름을 이용하여 조인을 표현한다. Quilt[4]는 기존 여러 질의어들의 특징을 통합하기 위해 시도한 새로 제안된 XML 질의어로, 다중문서에서 내부조인과 외부조인을 완전하게 지원한다. 그러나 현재까지 제안된 질의어들[1]은 XML 질의시 조인 조건이 같은 문서에 속해 있거나 2개의 다른 문서에 속해 있는 2개 또는 그 이상의 애트리뷰트나 데이

● 가능 ▲ 확장 가능
▼ 인급하지 않음 ✖ 불가능

질의 구조 형태		경로표현 + 필터표기	Where-Construct절	경로표현, FLWR (FOR, LET, WHERE, RETURN)	XML-QL + IR	XQL + IR
조인	내부	▲	●	●	●	▲
	외부	▲	●	●	●	▲
	세미	●	▼	▼	▼	▼
	유사성	✖	✖	✖	●	▼
	구조	✖	✖	✖	✖	✖

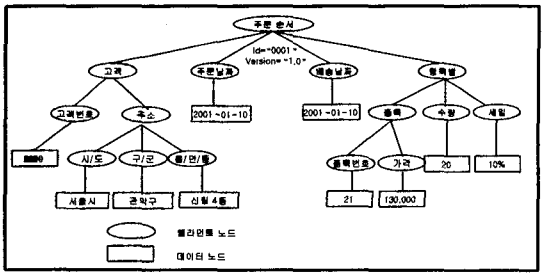
< 표1 > XML 질의어 조인 비교

터를 이용한다. 같은 값을 포함하는 2개 또는 그 이상의 엘리먼트들을 매칭함으로써 조인을 표현하였는데, 이러한 XML 조인 질의 표현은 XML의 구조적 특성을 모두 반영하지 못하는 단점이 있다.

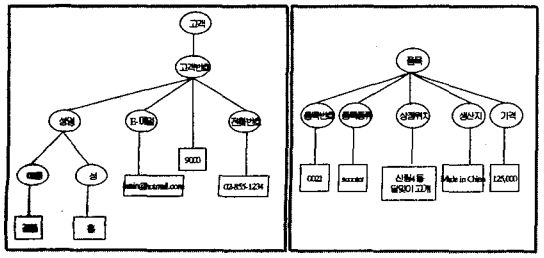
XML 질의어에 정보검색 기능을 지원하는 ELXIR[6]은 텍스트 유사성 조인 연산자(similarity join operator)를 가지고 XML-QL을 확장하였다. 하지만 XML-QL에서의 확장은 XML 질의 처리시 한정된 기능만을 제공하므로, 중첩 질의 또는 그룹핑 조인 질의시 문제점을 가지고 있다.

3. XML 조인 질의

본 논문에서는 XML 문서의 조인 질의 유형을 질의 내용에 따라 단순히 엘리먼트나 애트리뷰트를 이용한 조인뿐만 아니라, 텍스트의 유사성을 통한 조인과 XML 문서 내의 수직적, 수평적 구조만을 이용한 조인으로 보다 체계적으로 분류하였다. 특히 부모/자식, 조상/자손의 수직적 구조와 형제의 수평적 구조만을 이용한 구조 질의의 개념을 새롭게 제안하고, 효과적이고 간결한 질의 표현이 가능하도록 Quilt[4]를 확장하였다. 다음 <그림 1>과 <그림 2>는 다양한 조인 질의의 예를 위해 사용할 XML 문서의 일부분만을 표현한 주문서(주문서.xml), 고객 리스트(고객.xml), 물품 리스트(품목.xml)의 각 문서 트리 구조이다.



<그림 1> 주문서.xml 문서 트리 구조

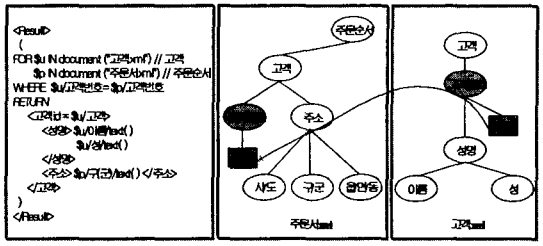


< 그림 2 > 고객.xml 과 품목.xml

3.1 내부-조인 (Inner-Join)

정의 *엘리먼트나 애트리뷰트를 통해, 2개 또는 그 이상의 문서로부터 조인 조건 즉, 값을 가지는 엘리먼트 매칭을 통해 조인을 표현하여 정보를 검색한다.*

<그림 3>은 물품을 구매한 모든 고객의 주소와 이름을 검색하기 위하여 주문서와 고객 리스트 문서를 고객번호 엘리먼트 값을 이용해 내부-조인 질의한 예를 나타낸다.



<그림 3> 내부 조인

단일 문서에 포함되어 있지 않은 정보를 두 문서의 내부 조인을 통하여 질의 검색을 수행할 수 있다.

3.2 외부-조인 (Outer-Join)

정의 *엘리먼트나 애트리뷰트를 통해 2개 또는 그 이상의 문서로부터 정보를 검색하되, 조인의 과정에서 조인할 상대 문서에 매칭하는 정보가 없을 경우에도 널(null)값을 만들어 결과 문서에 모두 포함한다. 즉, 내부-조인의 확장된 개념이다.*

모든 고객의 이름과 구입 항목 및 주문 번호를 검색하되, 주문하지 않은 고객에 대한 정보도 검색하기 위하여 고객 리스트와 주문서, 물품 리스트를 고객번호 엘리먼트, 품목번호 엘리먼트를 주문서.xml, 고객.xml, 품목.xml 세 문서에 매칭 하여 외부-조인을 수행한다. 외부-조인을 통하여 세 문서의 모든 정보를 결과 문서에 반환한다.

3.3 세미-조인 (Semi-Join)

정의 *엘리먼트나 애트리뷰트를 통해 2개 또는 그 이상의 문서로부터 정보를 검색하는데, 결과 문서에는 한 문서 내의 정보만이 포함된다.*

가격이 10,000미만인 물품을 구입한 모든 고객의 이름, 전화번호를 검색하기 위해 고객 리스트, 주문서, 물품 리스트를 고객번호, 품목번호 엘리먼트 매칭을 통해 세미-조인한다. 이 세미-조인의 결과 문서에는 고객 리스트에 존재하는 정보의 종류만이 나타난다.

3.4 유사성-조인 (Similarity-Join)

정의 *텍스트의 유사성을 이용하여 한 문서 또는 여러 문서로부터 정보를 검색한다.*

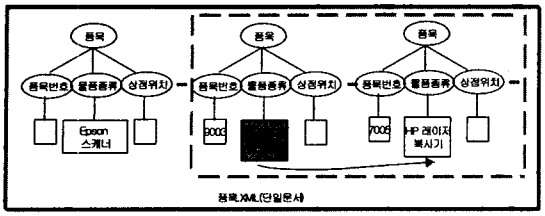
아래는 유사성 조인 연산자(similarity join operator)의 정의를 나타낸다.

Operator	::=	'≒'

같은 문서 내에서의 유사성 조인

같은 문서 내에서의 유사성 조인 질의는 상이한 엘리먼트가 가지는 유사한 텍스트 정보를 이용해 단일 문서를 스스로 조인한 후 정보를 검색한다.

<그림 4>는 물품 리스트 내에서 물품항목이 9003인 물품종류의 이름과 유사한 이름을 가지는 물품들을 검색하는 예를 보여준다. <그림 4>의 질의는 물품 리스트에서 고객이 요구한 물품과 유사한 이름을 가지고 있는 다른 물품들을 검색할 수 있도록, 이름 텍스트 값을 통해 조인한다.

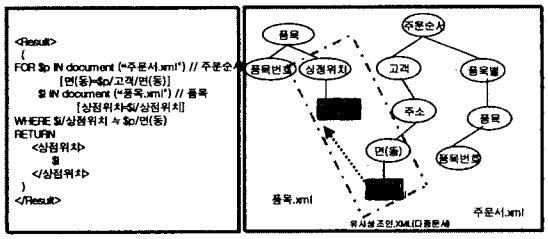


<그림 4> 유사성 조인 (단일문서)

여러 문서에서의 유사성 조인

여러 문서에서의 유사성 조인은 같은 문서 내에서의 유사성 조인과 마찬가지로 서로 다른 구조, 서로 다른 이름의 엘리먼트나 애트리뷰트를 가지고 있는 여러 문서들을 단지 텍스트 정보의 유사성만을 이용해 조인하여 보다 다양한 정보를 검색할 수 있도록 한다.

<그림 5>는 특정 물품을 구입한 고객이 거주하는 주소와 비슷한 주소 내에 있는 상점이 보유하고 있는 물품들을 검색하기 위하여 물품 리스트와 주문서 내의 주소 텍스트 정보의 유사성을 기반으로 조인한 예를 보여준다. <그림 5>의 왼쪽 부분은 조인 조건에 유사성 조인 연산자 " ≒ "를 가지고 Quilt XML 질의어에 구문을 작성한 예이다.



<그림 5> 유사성 조인 (다중 문서)

이러한 다중 문서에서의 유사성 조인은 각 문서에 대한 정확한 정보 없이 애매모호하게 질의를 하는 경우에도 관련성이 있는 정보 검색을 할 수 있는 장점이 있다.

3.5 구조-조인 (Structured-Join)

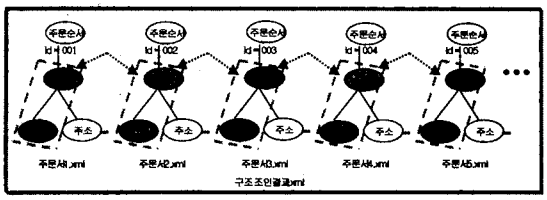
정의 *엘리먼트나 애트리뷰트 또는 텍스트 정보의 유사성이 아닌 XML 문서의 구조 특히, 조상/자손·부모/자식관계로 구성된 수직적 구조, 형제간의 관계로 구성된 수평적 구조를 통해 2개 또는 그 이상의 문서로부터 정보를 검색한다. 구조의 동일성의 정도를 깊이(depth) 정보를 이용하여 유연하게 조정 가능하다.*

다음은 구조 조인 연산자(structured join operator)에 대한 정의이다.

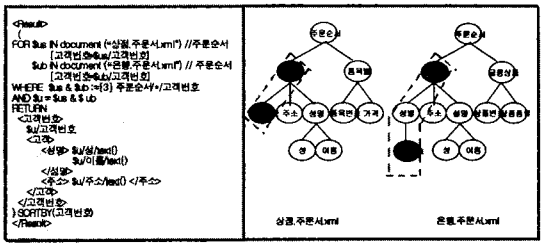
Operator	::=	':= [depth]'

<그림 6>은 여러 개의 주문서를 고객과 고객번호 엘리먼트의 부모/자식 구조의 동일성을 이용하여 조인하여 모든 고객의 주소를 검색하는 질의 예를 보여준다. 이 질의에서는 조인 시 깊이 조건을 1로 하여 중첩 레벨이 1 이내, 즉 고객과 고객번호 간의 부모/자식 관계의 구조만을 조인에 참여하도록 하였다. 이러한 구조-조인을 이용하면 많은 XML 문서에 대해 각각 질의를 할 필요 없이 질의문을 좀 더 간결하게 표현할 수 있고, 보다 유용하게 필요한 정보를 관리할 수 있다는 장점을 가진다.

<그림7>은 같은 지역에 존재하는 은행과 상점의 고객 정보를 통합 검색하기 위하여 구조 조인한 예를 보여준다. <그림 6>에서와 마찬가지로 고객과 고객번호 엘리먼트의 구조를 이용한 조인 질의이지만 <그림 7>에서는 깊이를 3으로 제한하여, 고객과 고객번호 엘리먼트가 부모/자식인 경우를 포함하여 중첩 레벨이 3 이내인 조상/자손 관계에 있는 구조를 대상으로 하는 조인 질의라는 차이가 있다. 본 논문에서 제안한 구조 질의의 개념은, 서로 다른 내용을 가지고 있으나 일정 중첩 레벨 이내로 그 구조에 유사성이 있는 경우 구조 조인을 통하여 흩어져있거나 중복되는 정보들을 하나의 결과 문서에서 보다 간결한 질의문을 이용하여 검색할 수 있다.



<그림 6> 구조 조인 (1)



< 그림 7 > 구조 조인 (2)

위의 구조 질의 예제에서는 간단한 질의문을 예로 들었으나, 구조 조인 질의는 복잡한 문서일 경우에 더욱 효율적이다. "조상/*/자손" 관계와 같은 복잡한 문서 구조가 깊은 중첩 정도를 가지는 경우에, 사용자가 요구대로 깊이를 지정하여 질의 내용과 관련성이 많은 부분을 우선적으로 처리하고, 그 결과에 다른 질의 내용을 재 질의함으로써 검색 결과에 대한 정확도를 높이고 질의 처리 시간을 효율적으로 관리할 수 있다.

5. 결 론

본 논문에서는 다중 문서에서의 구조 정보를 이용한 효율적인 조인 질의를 처리하기 위하여 구조 문서에 적용 가능한 조인 질의를 여러 타입으로 분류하였다. 또한 텍스트 정보의 유사성을 이용하는 유사성 조인 연산자(Similarity Join Operator)와 순수 구조 정보를 이용하는 구조 조인 연산자(Structured Join Operator)를 XML 질의어인 Quilt에 확장하여, 정보 검색 기능과 순수구조 관계인 조상/자손, 부모/자식, 형제 관계를 이용하여 다중 문서로부터 정보를 검색할 수 있는 기능을 지원하는 질의문을 설계하였다. 순수 구조 정보만을 이용한 질의시 구조의 일치 정도를 깊이(depth) 정보를 이용하여 질의자가 제한할 수 있게 하여, 유연성 있는 질의가 가능하도록 하였고, 질의문을 간결하게 표현 할 수 있게 하였다.

참고 문헌

[1] Angela Bonifati, Dongwon Lee., "Technical Survey of XML Schema and Query Language", 2001.

[2] Jonathan Robie, Joe Lapp, David Schach, "XML Query Language (XQL)", "http://www.w3.org/Tands/QL/QL98/pp/xql.html".

[3] Alin Deutsch, Mary Fernandez, Alon Levy, Dan Suciu, "XML-QL: A Query Language for XML", World Wide Web Cons, 1998.

[4] Don Chamberlin, Jonathan Robie, Daniela Florescu, " Quilt: An XML Query Language for Heterogeneous Data Sources", In Int'l Workshop on the Web and Database(WebDB), 2000.

[5] Scott Boag, Don Chamberlin, Mary F. Fernandez, Daniela Florescu, Jonathan Robie, "XQuery 1.0: An XML Query Language" , W3C Working Draft 16 August 2002

[6] T.T. Chinenyanga and N. Kushmerick. An Expressive and Efficient Language for XML Inf. Retrieval. In Proc. of the 2001 SIGIR Conf., pp. 163-171, 2001.

[7] N. Fuhr and K. Grossjohann. XIRQL: A Query Language for Information Retrieval in XML Documents. In Proc. of the SIGIR 2001 Conf., pp. 172-180, 2001.