링크 질의를 통한 XML 문서의 검색 기법

문 찬 호[†]·강 현 철[††]

요 약

Web 문서를 기술하기 위해 차세대 표준으로 제안된 XML은 Web 기반의 여러 응용 분야에서 널리 사용되고 있으며, Web 상의 XML 문서들은 서로 하이퍼링크를 통해 연결되어 있다. 현재까지 대부분의 XML 관련 연구들은 XML 문서의 효율적인 저장, 관리 및 검색을 위한 XML 저장 시스템을 대상으로 하고 있으며, XML 링크를 지원하는 질의어의 개발이나 링크를 활용한 XML 검색 시스템의 개발에 대한 연구는 미흡하다. 본 논문에서는, XML 링크 질의 표현을 위한 XML 질의어의 확장과 링크 질의 처리 기법을 제시한다. 링크 질의는 하나의 XML 문서(질의 문서)와 질의 문서 내의 링크로 참조되는 XML 문서(참조 문서)들의 내용을 검색하는 것이다. 참조 문서의 검색을 위해서 현재는, 참조 문서에 대한 질의를 수작업으로 생성, 처리, 그리고 그 결과의 리턴을 반복적으로 수행하는 방법이 사용되고 있다. 본 논문의 링크 질의 처리 목적은 한번의 질의 입력을 통해 추가적인 수작업 없이 참조 문서(들)에 대한 검색 결과까지 얻을 수 있는 기능을 제공하는 것이다. 기존 수작업 기반과 본 논문의 링크 질의 처리의 성능을 비교, 분석한 결과, 참조 문서로의 링크가 많을수록 수작업 기반에 비하여 질의 처리 시간이 줄어들고, 질의 문서가 저장된 사이트에 참조 문서가 많이 있을수록, 질의 처리 시간이 줄어들었다.

Retrieval Scheme of XML Documents Using Link Queries

ChanHo Moon[†]·Hyunchul Kang[††]

ABSTRACT

The XML that was proposed as a next-generation standard for describing Web documents is widely used in various Web-based applications. In addition, XML documents on the Web link each other by hyperlinks. The current works on XML focus on the XML storage system that can efficiently store, manage, and retrieve XML documents. However, the research on the query language that supports the XML links and on the XML retrieval systems to process the XML links, is little conducted until now. In this paper, we propose an extension of an XML query language for expressing the XML link query and its processing scheme. A link query is to retrieve contents from an XML document (a query document) and from the XML documents (referenced documents) that are referred to by the links in the query document. As far as retrieving from the referenced documents is concerned, the current practice is to manually generate queries to get the partial results, and to repeat such a procedure. The purpose of link query processing in this paper is to eliminate the manual work altogether in getting the complete query result. The performance analysis shows that our link query processing strategy outperforms the conventional approach including the manual tasks. The more links to the referenced documents and the more referenced documents there are in the site storing the query document, the more query processing time decreases.

키워드 : Web, XML 문서 검색(retrieval of XML document), 링크 질의(link query), 부질의(subquery)

1. 서 론

현재의 컴퓨팅 환경에서는 Web의 발전과 더불어 Web 문서의 활용 분야가 점점 확대되고 있다. 특히 전자 도서관, 전자 상거래와 같은 응용 분야에서는 Web 문서를 통해 다양한 형태의 정보를 제공하고 있다. 이에 따라 Web 문서들에 대한 효과적인 관리 기능의 필요성이 대두되고 있다. Web 문서들을 효율적으로 저장하고 검색, 이용하기 위해 W3C에서는 차세대 Web 문서를 기술하기 위한 표준 마크업 언어인 XML (eXtensible Markup Language)을 제안하였다[1]. HTML과 비교하여 XML의 기능상 특징을 몇 가지 정리하면 다음과 같다.

- XML은 간단하면서 다루기 쉽기 때문에 Web 상에서 제공하는 방대한 양의 Web 문서들을 다양하게 표현할 수 있다. 특히 HTML과는 달리 새로운 태그와 속성을 정의할 수 있다.
- 문서 구조의 검증이 필요한 응용개발 과정에서 XML은 문법적 오류에 관한 판단을 문서 내에서 제공한다.

다양한 정보 형태를 가진 Web 문서의 효과적인 관리를 위해 XML 관련 연구들이 현재 국·내외에서 활발히 진행되고 있다. XML 문서들을 저장, 관리 및 검색할 수 있는

※ 본 연구는 한국과학재단 목적기초연구(2001-1-303-001-3)지원으로 수행되었음.
† 준 회 원 : 중앙대학교 대학원 컴퓨터공학과
†† 정 회 원 : 중앙대학교 컴퓨터공학과 교수
논문접수 : 2000년 11월 17일, 심사완료 : 2001년 7월 3일

XML 저장 관리(repository) 시스템 개발에 대한 연구[2-5], XML 관련 질의어에 관한 연구[6-9], 기존 데이타베이스 시스템에 저장된 데이타를 XML 문서로 변환하는 도구 개발에 관한 연구(예를 들어, 관계 데이타베이스에서 레코드들의 집합을 XML 문서로, 객체 지향형 데이타베이스에서 객체를 XML 문서로)[10, 11], Web 환경에서 이질적인 정보에 대해 데이타의 추출 및 통합 기능을 위한 wrapper-mediator 기반의 저장 시스템 개발[12-14] 등이 그 예이다.

Web 상에는 HTML 문서와 같이 XML 문서들도 서로 하이퍼링크로 연결되어 있는 상태로 분산되어 있다. 즉, 서로 다른 Web 사이트들 간에도 XML 문서에 대한 링크가 존재한다. 특히 HTML과는 달리 XML에서는 다양한 링크 형식을 제공하고 있다. XML이 제공하는 다양한 링크의 기능을 활용하면 XML 문서를 좀더 효율적으로 검색할 수 있다. 본 논문에서는 XML 링크를 지원하는 질의를 **XML 링크 질의**(이하, **링크 질의**)라고 부른다. 현재까지 개발된 XML 질의어들은 링크 질의 기능을 제공하지 못하고 있다. 본 논문에서는, Web 상에 분산되어 저장된 XML 문서를 대상으로 효율적인 검색을 지원하기 위해, 링크 질의를 정의하고, 링크 질의 표현을 위한 XML 질의어의 확장과 링크 질의 처리 기법을 제시하고, 이의 성능을 분석한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 XML 질의어와 XML에서의 링크 표현 방법 및 링크를 지원하기 위해 현재 진행 중인 XML 브라우저 개발 현황에 대해 기술한다. 3장에서는 링크 질의를 정의하고 링크 질의 처리 환경을 기술한다. 4장에서는 링크 질의 표현을 위한 XML 질의어의 확장을 제안하고, 5장에서는 링크 질의 처리 기법을 제안한다. 6장에서는 링크 질의를 통한 XML 문서 검색의 성능을 분석한다. 마지막으로 7장에서 결론 및 향후 연구에 대해 기술한다.

2. 관련 연구

현재 XML 관련 질의어는 XQL(XML Query Language)[6], XML-QL[7], XPath[8], XQuery[9] 등이 있으며, 이들은 XML 문서의 구조적 특성을 반영한 구조/내용 기반 검색들을 지원하고 있다. <표 1>은 XML 관련 질의어들을 서로 비교한 것이다. XQL은 1998년 W3C에 의해 제안된 범용적 질의 언어로 경로 표현(path expression)을 이용하여 질의 대상이 되는 입력 노드를 표시하고 그 외 조건을 표시하는 필터 연산 필터 연산자([ ]), 비교 연산자(=, !=, lt, gt, …) 등을 이용하여 질의를 표현한다. XML-QL은 다수의 XML 문서로부터 데이타를 추출하거나 관계형 데이타베이스, 객체 지향 데이타베이스 또는 XML 데이타 간의 데이타를 변환하거나 여러 소스로부터 데이타를 통합하기 위해 W3C에 의해 제안된 XML 질의어이다. XML-QL은 4.1절에서 더 자세히 설명한다. XPath은 XML 문서의 엘리먼트(element), 애트리뷰트(attribute), 주석(comment), 처리 지시문(processing instruction) 등(이들을 XPath에서는 노드(node)로 정의하고 있음)에 대해 주소를 지정하여 필요로 하는 정보에 해당하는 정확한 노드(들)를 얻을 수 있도록 해준다. XQuery는 Quilt [15]에 기반한 질의어로 기존의 XML 질의어인 XQL, XML- QL, XPath 뿐만이 아니라 SQL/OQL과 같은 질의어의 특징을 빌려 구조화 문서 및 반구조화 문서, 그리고 관계 데이타베이스 및 객체 데이타베이스를 포함하는 다양한 데이타 소스에 대한 질의를 표현할 수 있다.

이들 XML 질의어를 활용하면 내용 기반 검색, 구조 기반 검색 등과 같은 다양한 형태의 XML 문서 검색 질의를 표현할 수 있다. 그러나 현재까지 제안된 XML 질의어들은 XML 링크를 지원하는 질의를 표현하지 못하고 있다.

<표 1> XML 관련 질의어 비교

질의어 / 비교항목	XQL	XML-QL	XPath	XQuery
제안	W3C	W3C	W3C	W3C
질의 구조 형태	경로 표현+필터 연산자+비교 연산자 등	WHERE-CONSTRUCT 구조	위치 경로	경로 표현, 엘리먼트 생성자, FLWR (FOR, LET, WHERE, RETURN) 표현식 등
결과 형태	노드, 노드리스트, XML 문서, 배열, 기타 구조	XML 데이타	노드의 집합	XML 데이타
내용/구조 기반 검색 지원 여부	지 원	지 원	지 원	지 원
질의 가능 영역	• 한 문서 또는 XML 저장소 내 모든 문서에 대한 질의 지원 • 문서 간의 연산은 지원하지 않음	• 한 문서 또는 XML 저장소 내 모든 문서에 대한 질의 지원 • 문서 간의 연산 지원	• 한 문서 또는 XML 저장소 내 모든 문서에 대한 질의 지원 • 문서 간의 연산은 지원하지 않음	• 한 문서 또는 XML 저장소 내 모든 문서에 대한 질의 지원 • 문서 간의 연산 지원
검색 결과의 후처리 가능 여부	불 가	가 능	불 가	가 능
링크 지원 여부	지원하지 않음	지원하지 않음	지원하지 않음	지원하지 않음

XML에서의 링크는 XLink[16]와 XPointer[17]를 사용하여 각각 문서와 문서 사이의 링크, 문서 내부에서의 연결 정보를 표현하고 있다. HTML에서는 <A> 태그를 이용하여 문서로의 링크를 만들 수 있다. 즉, 한 자원(<A>와 </A>로 둘러싸이는 텍스트 및 이미지)과 URL이 1 : 1로 부합(match)된다. XML에서도 HTML에서와 같은 링크 형식을 제공하는 데 이를 단순 링크(simple link)라 한다((그

림 1) 참조). 그리고 XML에서는 한 자원에 대해 여러 개의 URL 정보와 링크할 수 있는 형식을 제공하는 데 이를 확장 링크(extended link)라 한다((그림 2) 참조). Web 상의 두 XML 문서가 단순 링크 또는 확장 링크로 연결되어 있을 경우, 이들 문서는 각각 링크로 상대 문서를 참조하는 문서와 링크로 참조되는 문서로 나눌 수 있다.

```
< xdoc >
< para >
This is the report on the origins of the cold war.
        < locator xml : link = "simple" href = "intro.xml" >
                Introduction< /locator >
        < locator xml : link = "simple" href = "propos.xml" >
                Proposal< /locator >
        < locator xml : link = "simple" href = "summary.xml" >
                Summary< /locator >
< /para >
< /xdoc >
```

(그림 1) XML에서의 단순 링크 예

```
< xdoc >
< para >
This is the report of
    < mylink xml : link = "extended" inline = "true" title =
                "Debate" >
    < locator xml : link = "locator" href = "intro.xml" title =
                "Introduction"/ >
    < locator xml : link = "locator" href = "propos.xml" title =
                "/>
    < locator xml : link = "locator" href = "rebut.xml" title =
                "Rebuttal" / >
        a debate
    < /mylink >
on the origins of the cold war.
< /para >
< /xdoc >
```

(그림 2) XML에서의 확장 링크 예

현재까지 진행중인 XML 관련 연구를 정리해 볼 때, XML과 관련하여 링크를 지원하는 질의어 개발이나 링크를 활용한 검색 시스템 개발에 대한 연구는 미비하다. 특히, 링크를 통해 자유롭게 항해(navigation)할 수 있는 HTML 브라우저의 발전과 비교했을 때, XML 브라우저의 개발은 아직 미흡하다. 현재 XML 브라우저로는 MS사의 인터넷 익스플로러 5.5, 영국 U. Nottinghan의 Jumbo3(http://www.xml-cml.org), 넷스케이프 사의 Mozilla(http://www.mozilla.org/source.html)와 넷스케이프 6.0, IBM 사와 Alphaworks 사의 XML Viewer(http://www.alphaworks.ibm.com/tech/xmlviewer) 그리고 일본 Fujitsu사의 HyBrick(http://www.fsc.fujitsu.com/hybrick/) 등이 선보이고 있다. 지금까지 개발된 XML 브라우저는 스타일 문서(stylesheet)를 적용하여 XML 문서를 보여 줄 수 있는 기능 중심으로 개발되었고, 제한적으로 링크에 의한 항해 기능을 제공하고 있다. HTML의 경우에는 업계 표준(de facto standard)으로 인터넷 익스플로러나 넷스케이프 네비게이터와 같은 브라우저가 있으나, XML의 경우에는 아직 표준적인 브라우저가 없는 상황이어서 XML 링크의 지원이 향후 XML 브라우저 상에서 어떤 형태로 이루어질 지가 아직 확립되어 있지 않은 상태이다.

국내에서도 정보 검색 시 XML 링크 정보를 이용한 색인 기법을 제시한 연구들이 있다[18, 19]. [18, 19]는 XML 문서의 각 링크마다 고유한 식별자를 부여하고 각 식별자별로 링크의 가중치를 부여하여 가중치 값을 기준으로 색인하는 방법을 제시하고 있다. 그러나 이들은 기존 정보 검색 시스템이 갖던, 문서 내 색인어 발생 빈도에 따른 문서 검색 방법의 대안으로 XML 링크로 많이 연결된 문서 순으로 문서를 검색하는 방법에 관한 연구이다. 그러므로, 이들 연구들은 Web 상에서 XML 문서의 링크를 따라 항해하면서 문서를 검색하는 방안을 제시하고 있지 않다.

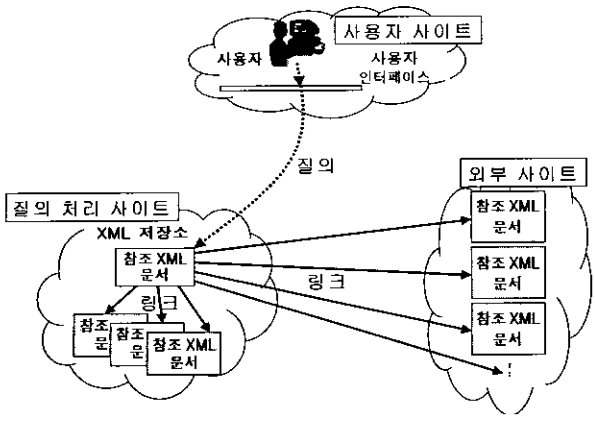
3. XML 링크 질의

본 장에서는 XML 링크 질의를 정의하고 링크 질의 처리 개요에 대해 기술한다. 그리고, 링크 질의 처리 환경에서 발생 가능한 링크 질의들을 유형별로 분류한다.

3.1 링크 질의 처리 개요

링크 질의는 특정 URL을 참조하는 XML 문서(**질의 XML 문서**(이하, **질의 문서**))를 대상으로 질의 문서 전체 또는 일부 내용을 검색할 뿐 아니라 질의 문서에 의해 링크로 참조되는 XML 문서(**참조 XML 문서**(이하, **참조 문서**))의 전체 또는 일부 내용을 검색하는 질의로 정의된다. (그림 3)은 링크 질의를 처리하기 위한 환경을 나타낸 것이다. Web 상의 사이트는 사용자 사이트, 질의 처리 사이트와 외부 사이트로 분류된다. **사용자 사이트(user site)**는 사용자에게 질의를 입력받기 위한 사용자 인터페이스를 제공하고, 사용자로부터 링크 질의를 입력받는다. 질의 문서가 저장되어 있고 질의 처리가 수행되는 주(main) 사이트를 **질의 처리 사이트(query processing site)**라 한다. 하나의 질의 문서와 링크로 연결된 참조 문서는 다수가 될 수 있으며, 이 참조 문서는 질의 처리 사이트 내의 XML 저장소(repository)에 있거나 질의 처리 사이트가 아닌 제 3의 사이트(**외부 사이트(external site)**)에 있을 수 있다. XML의 확장 링크를 사용한 경우에는 참조 문서가 질의 처리 사이트와 외부 사이트 모두에 있을 수도 있다.

기존 XML 질의어를 이용하여 링크 질의를 수행하는 경우, 질의 문서를 대상으로 질의(**초기 질의**)를 처리하여 하나 또는 그 이상의 참조 문서 URL(들)을 사용자에게 리턴



(그림 3) 링크 질의 처리 환경

한다. 사용자는 리턴된 URL(들)로 새로운 질의(URL 질의)를 입력하고 원하는 결과(URL 또는 원하는 최종 데이타)를 얻을 때까지 이 과정을 반복하게 된다.

<표 1>에서 설명한 기존의 XML 질의어를 이용하여 링크 질의를 수행하기 위해서는 수작업 기반의 링크 질의 처리가 필요하다. 수작업 기반의 링크 질의 처리는 사용자가 원하는 결과를 얻기 위해 사용자가 직접 수작업에 의해 질의를 생성하여 해당 질의를 처리함을 의미한다. 수작업 기반의 링크 질의 처리 과정은 다음과 같다.

① 사용자는 사용자 인터페이스를 통해 질의 문서를 대상으로 한 질의(초기 질의)를 질의 처리 사이트로 전달된다.
② 질의 처리 사이트는 초기 질의를 처리하여 그 결과인 참조 문서의 URL(들)을 사용자에게 리턴한다.
③ 사용자는 리턴된 URL(들)을 사용하여 새로운 질의(URL 질의)를 생성한다. 즉, 수작업으로 사용자 인터페이스를 통해 URL 질의를 입력한다.
④ URL 질의는 그 대상이 되는 사이트로 전달되고 처리되어 그 결과가 사용자에게 리턴된다.
⑤ ②의 과정에 의해 리턴된 참조 문서의 URL이 다수인 경우, 원하는 결과(URL 또는 원하는 최종 데이타)를 사용자가 얻을 때까지 ③, ④ 과정을 반복한다.

그러나, 본 논문에서 제시하는 링크 질의 처리는 한번의 질의 입력을 통해 원하는 최종 결과를 얻는 것을 의미한다. 이를 위해 본 논문에서는 기존의 XML 질의어를 구문 확장하여 링크 질의를 표현할 수 있는 기법을 제시하고, 링크 질의를 처리하는 방안을 제시한다.

3.2 링크 질의의 유형

3.1절에서 기술한 링크 질의 처리 환경에서 발생 가능한 링크 질의는 다음과 같이 크게 세 가지 유형으로 분류할 수 있다.

(유형 1) 참조 문서가 질의 처리 사이트 내에만 있는 경우

A 쇼핑몰 사이트는 자신의 XML 저장소에 판매 물품 전체에 대한 카탈로그 XML 문서와 개개 판매 물품의 상세 정보 XML 문서들을 저장/관리할 경우, (유형 1)의 예는 다음과 같다.

> **A 쇼핑몰 사이트 내 카탈로그 XML 문서를 대상으로, x 물품에 대한 상세한 정보를 찾으시오.**

(유형 2) 참조 문서가 외부 사이트에만 있는 경우

소비자 평가 자료 사이트는 회원들에 대한 정보를 갖는 XML 문서를 관리하고, 각 회원들의 홈페이지는 모두 소비자 평가 자료 사이트가 아닌 다른 사이트에 저장되어 있는 경우, (유형 2)의 예는 다음과 같다.

> **소비자 평가 자료 사이트 내 회원 정보 XML 문서를 대상으로, 회원 중 이름이 홍길동이라는 사람의 홈페이지를 찾아 publication 정보를 찾으시오.**

(유형 3) 참조 문서가 질의 처리 사이트와 외부 사이트에 모두 존재할 경우

각 쇼핑몰 사이트는 자신의 XML 저장소에 판매 물품에 대한 카탈로그 XML 문서들을 저장하고 있다고 가정하자. 물품에 대한 가격 비교를 위해, A 쇼핑몰에서는 타 쇼핑몰에서 판매하는 동일 물품에 대한 정보도 제공할 경우, (유형 3)의 예는 다음과 같다.

> **A 쇼핑몰 사이트 내 카탈로그 XML 문서를 대상으로, x 물품에 대한 상세한 정보와 타 쇼핑몰에서 판매되는 x 물품에 대한 상세한 정보를 찾으시오.**

4. 링크 질의 표현을 위한 XML 질의어의 확장

본 장에서는 링크 질의 표현을 위한 XML 질의어의 확장 구문을 제시하고, 확장된 질의어로 표현된 링크 질의의 예를 기술한다. 기존 XML-QL[7]은 서로 다른 문서를 대상으로 질의를 처리하여 각 문서에 대한 결과를 합성하는 조인 연산 기능을 제공한다. 본 논문에서는 XML-QL의 이러한 기능을 확장하여 서로 다른 문서들 간의 링크 질의를 표현할 수 있도록 한다.

4.1 XML-QL

(그림 4)는 XML 문서(a)를 대상으로 XML-QL로 표현된 질의어(b)를 수행하여 얻어지는 결과(c)를 나타낸 것이다. XML-QL의 구문은 크게 WHERE 절과 CONSTRUCT 절로 구성되며, CONSTRUCT 절 내에 다시 WHERE 절과

CONSTRUCT 절이 중첩될 수 있다. WHERE 절은 SQL의 SELECT-WHERE-FROM 구조에서의 WHERE절에 해당하고, XML 문서를 대상으로 한 검색 조건을 나타낸다. CONSTRUCT 절은 질의 처리 결과를 XML 문서로 나타내기 위한 결과 형식을 나타낸다. WHERE 절과 CONSTRUCT 절은 모두 XML 구조를 따른다.

```
<bib>
  <book year = "1995">
    <title> An Introduction to Database Systems </title>
    <author>
        <lastname> Date </lastname>
    </author>
    <publisher>
        <name> Addison-Wesley </name>
    </publisher>
  </book>
  <book year = "1998">
    <title> Foundation for Object/Relational Databases:
          The Third Manifesto </title>
    <author>
        <lastname> Date </lastname>
    </author>
    <author>
        <lastname> Darwen </lastname>
    </author>
    <publisher>
        <name> Addison-Wesley </name>
    </publisher>
  </book>
</bib>
```

(a) XML 데이타

```
WHERE <book>
              <publisher>
                <name>Addison-Wesley</name>
              </publisher>
              <title> $t</title>
              <author> $a</author>
            </book> IN "www.a.b.c/bib.xml"
CONSTRUCT <result>
                <author> $a</author>
                <title> $t</title>
              </result>
```

(b) XML-QL로 표현된 질의

```
<result>
  <author> <lastname> Date </lastname> </author>
  <title> An Introduction to Database Systems </title>
</result>

<result>
  <author> <lastname> Date </lastname> </author>
  <title> Foundation for Object/Relational Databases:
        The Third Manifesto </title>
</result>

<result>
  <author> <lastname> Darwen </lastname> </author>
  <title> Foundation for Object/Relational Databases:
        The Third Manifesto </title>
</result>
```

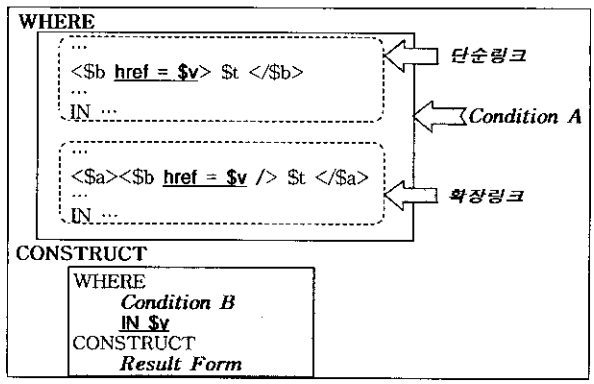
(c) (a)를 대상으로 (b)를 수행한 결과

(그림 4) XML-QL로 표현된 질의의 예 및 그 수행 결과

4.2 XML 질의어 확장 구문

(그림 5) (a)는 링크 질의를 위한 XML 질의어의 구문(syntax)을, (b)는 링크 질의의 예를 나타낸 것이다. 링크 질의를 위한 XML 질의어의 구문은 크게 WHERE 절과 CONSTRUCT 절로 구성되며, CONSTRUCT 절 내에 다시 WHERE 절과 CONSTRUCT 절이 중첩되는 구조를 갖는다. 바깥쪽의 CONSTRUCT 절은 질의 문서에 대한 검색 조건을 기술하며, 안쪽의 CONSTRUCT 절은 참조 문서에 대한 검색 조건을 기술한다. 이때, 사용자는 질의 문서 내 참조 문서로의 링크가 단순 링크로 구성되었는 지, 확장 링크로 구성되었는 지를 알 수 없다. 따라서 바깥쪽의 CONSTRUCT 절에는 단순 링크와 확장 링크에 대한 검색 모두를 표현한다.

질의 구문 중 "href = $v"와 "IN $v"은 질의가 링크를 지원하는 질의임을 나타낸다. "$v"는 질의 처리 시 검색 대상이 되는 XML 문서(참조 문서)의 URL을 저장하는 변수이다. "href = $v"는 애트리뷰트로 링크를 포함하고 있는 XML 문서의 엘리먼트를 찾아 변수 $v에 해당 URL을 저장하라는 의미를 갖는다. IN $v는 변수 $v가 갖는 URL의 XML 문서를 대상으로 검색을 처리하라는 의미를 갖는다.



(a) 링크 질의를 위한 XML 질의어의 구문

```
WHERE
  <Register>
      <Author href = $v> A. Einstein
      </Author>
  </Register> IN "www.a.v/register.xml"
  <Register>
      <Author>
          <$a href = $v/>
          A. Einstein
      </Author>
  </Register> IN "www.a.b.c/register.xml"
CONSTRUCT
    WHERE
      <Homepage>
          <Pub> $b </Pub>
      </Homepage> IN $v
    CONSTRUCT
      <Pub> $b </>
```

(b) XML 링크 질의의 예

(그림 5) 링크 질의를 위한 XML 질의어의 구문 및 그 예

(그림 5) (a)에 기술한 질의어 구문은 <Condition A>에 부합되는 엘리먼트의 링크와 연결된 참조 문서를 찾아 참조 문서 내에서 <Condition B>에 부합되는 내용을 찾아 그 결과를 <Result Form>으로 리턴하라는 검색 요청을 표현하고 있다. (그림 5) (b)는 (그림 5) (a)에서 기술한 구문에 맞게 작성된 질의의 예로써, URL이 "www.a.b.c/homepage.xml"인 질의 문서를 대상으로 Author 엘리먼트의 내용이 A. Einstein인 링크와 연결된 참조 문서를 찾아 참조 문서를 대상으로 Hompage/Pub 엘리먼트 내용을 찾아 Pub를 루트 엘리먼트로 하는 XML 문서를 생성하라는 링크 질의다.

본 논문에서 제안하는 링크 질의 처리의 최종 목표는, 기존 검색 엔진의 키워드 검색 결과에서와 같이 참조 문서의 URL을 찾는 것이 아니라, 참조 문서 내에서 검색 조건에 맞는 내용을 새롭게 구성하여 XML 문서의 형식으로 최종 결과를 리턴하는 것이다. 이를 위해, 질의 처리 사이트는 사용자 인터페이스를 통해 받아들인 질의를, 참조 문서만을 대상으로 검색하는 질의(들)로 재구성하여 사용한다. 이를 **부질의**라 한다. 부질의는 질의 문서가 참조 문서를 가리키는 링크 수만큼 생성된다. 각 부질의는 참조 문서를 관리하는 사이트로 전송되어 해당 사이트에서 처리되고, 그 결과는 질의 처리 사이트로 리턴된다. (그림 6) (a)는 링크 질의 처리 시 생성되는 부질의의 템플릿 형태를, (b)는 그 예를 나타낸 것이다. 부질의 템플릿이란 사용자 질의로부터 부질의를 쉽게 만들 수 있도록 부질의의 형식을 갖는 템플릿이다.

```
WHERE
        Condition B
        IN $URL(Link Info.)
CONSTRUCT
        Result Form
```

(a) 부질의 템플릿

```
WHERE
    <Homepage>
        <Pub> $b </Pub>
    </Homepage> IN "www.a.b.c/homepage.xml"
CONSTRUCT
    <Pub>$b</Pub>
```

(b) 부질의의 예

(그림 6) 부질의 템플릿의 형태 및 그 예

부질의를 생성하는 과정은 다음과 같다. 먼저, 사용자 인터페이스를 통해 입력받은 질의로부터 부질의 템플릿을 만든다. (그림 6) (a)는 (그림 5) (a)의 링크 질의에서 CONSTRUCT 절(음영 표시 부분) 내용을 추출하여 참조 문서만을 검색하는 부질의를 생성할 수 있도록 만든 부질의 템플릿이다. 부질의는 생성된 부질의 템플릿의 "$URL(Link Info.)"에 참조 문서의 URL로 지정(assign)함으로써 생성된다. (그림 6) (b)는 (그림 5) (b)의 링크 질의에서 부질의 템플릿을 생성하여, 참조 문서의 URL인 "www.a.b.c/homepage.xml"을 지정한 부질의의 예이다.

링크 질의 표현을 위해 확장된 XML-QL은 기존의 XML-QL이 갖던 기능에 링크를 활용한 검색 기능을 갖는다. 기존의 XML-QL은 질의 문서를 대상으로 구조/내용 기반 검색과 속성 기반 검색, 그리고 질의 결과에서 같은 엘리먼트 이름이 존재하거나 엘리먼트의 검색 결과가 서로 같은 내용을 포함할 경우, 중복을 제거하여 하나의 결과만 리턴할 수 있는 기능을 제공한다. 또한, 검색된 결과를 정렬하는 기능도 제공하고 있다. 링크 질의 표현을 위해 확장된 XML-QL도 질의 문서에 의해 링크로 참조되는 참조 문서를 대상으로 구조/내용 기반 검색, 속성 기반 검색, 그리고 중복 회피, 정렬을 지원하는 검색을 지원한다. 즉, 링크 질의는 서로 링크로 연결된 XML 문서들에 대해서도 기존 질의어가 갖는 기능을 지원함으로써 XML 문서 검색을 더 효율적으로 할 수 있다. 그리고, 사용자가 검색된 URL(들)에 대해 수작업으로 질의를 여러 번 작성하여 입력하지 않고도 한 번의 질의를 통해 원하는 최종 결과를 얻을 수 있다.

4.3 확장된 질의어로 표현된 링크 질의의 예

다음은 다양한 기능을 지원하는 링크 질의를 확장된 질의어로 표현한 예를 나타낸 것이다.

(예 1) 구조/내용 기반 검색

www.a.b.c/register.xml에 기록된 사람 중에서 저자 이름이 "A. Einstein"인 사람의 homepage를 찾아 publication 정보를 찾으시오.

```
WHERE
    <Register>
        <Author href = $r > A. Einstein </>
    </Register> IN "www.a.b.c/register.xml"
    <Register>
        <Author>
            <$a href = $r /> A. Einstein
        </Author>
    </Register> IN "www.a.b.c/register.xml"
CONSTRUCT
    WHERE
        <Homepage>
            <Pub> $b </Pub>
        </Homepage> IN $r
    CONSTRUCT
        <Pub>$b</Pub>
```

(예 2) 속성 기반 검색

www.x.y.z/catalog.xml에 포함된 물품 중에서 흰색 BMW에 대한 상세한 정보 찾으시오.

```
WHERE
    <Catalog>
        <Car href = $r color = white>
            BMW </Car>
    </Catalog> IN "www.x.y.z/catalog.xml"
```

```
    <Catalog>
        <Car color = white>
            <$a href = $r />
            BMW
        </Car>
    </Catalog> IN "www.x.y.z/catalog.xml"
CONSTRUCT
    WHERE
        <$t> $b </$t> IN $r
    CONSTRUCT
        <$t> $b </$t>
```

■

(예 3) 중복 회피, 정렬을 지원하는 검색

www.a.b.c/register.xml에 기록된 사람 중에서 저자 이름이 "K. Smith" 인 사람의 homepage를 찾아 최근에 출판된 publication 정보를 엘리먼트 중복없이 년월 오름차순으로 찾으시오.

```
WHERE
    <Author>
        <Author href = $r > K. Smith </Author>
    </Author> IN "www.a.b.c/register.xml"
    <Author>
            <$a href = $r >
            K. Smith
    </Author> IN "www.a.b.c/register.xml"
CONSTRUCT
    WHERE
        <Pub>
            <year> $y </year>
            <month> $z </month>
        </Pub> ELEMENT_AS $m IN $r
        ORDER-BY $y, $z
    CONSTRUCT
        $m
```

■

5. 링크 질의 처리

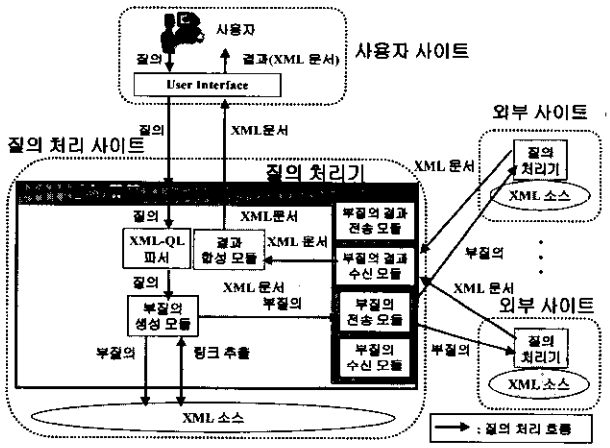
링크 질의 기법과 관련된 기술적 고려 사항은 다음과 같다. 본 장에서는 이들에 대해 설명한 후, 링크 질의 처리의 예를 기술한다.

(1) 링크 질의를 위한 질의 처리기의 모델링
(2) 링크 질의 결과에 대한 합성 방법

5.1 링크 질의를 위한 질의 처리기의 모델링

링크 질의를 위한 **질의 처리기**는 (그림 7)과 같은 구성을 갖는다. (그림 3)에서와 같이 XML 문서의 검색을 위해 Web 상의 사이트들은 사용자 사이트, 질의 처리 사이트, 외부 사이트들로 구성되며, 각 사이트는 XML 문서를 저장하고 있는 **XML 소스**(source)와 사용자 질의를 처리할 수 있는 질의 처리기를 갖는다. 질의 처리기는 XML-QL 파서, 부질의 생성 모듈, 그리고 통신 모듈 등으로 구성된다. 통신 모듈은 결과 전송/수신 모듈, 부질의 전송/수신 모듈로 구성되며, 다른 사이트들과의 메시지(질의 및 결과) 전송 기능을 담당한다. 질의 처리기를 구성하는 각 모듈의 기능은 다음과 같다.

- XML-QL 파서 : 사용자 인터페이스를 통해 입력받은 사용자 질의를 파싱하고 그 결과를 부질의 생성 모듈로 전달한다.
- 부질의 생성 모듈 : 파싱된 링크 질의로부터 부질의 템플릿을 생성하고 링크를 지정하여 부질의를 생성한다.
- 부질의 전송 모듈 : 부질의 생성 모듈에 의해 생성된 부질의들 중, 참조 문서가 외부 사이트에 존재하는 부질의를 해당 외부 사이트로 전송한다.
- 부질의 수신 모듈 : 질의 처리 사이트의 부질의 전송 모듈로부터 전송된 부질의를 수신하여 자신의 부질의 생성 모듈로 전달한다.
- 부질의 결과 전송 모듈 : 외부 사이트에서 처리된 부질의 결과를 질의 처리 사이트로 리턴한다.
- 부질의 결과 수신 모듈 : 외부 사이트로부터 전송된 요청한 부질의 결과를 수신하여 결과 합성 모듈로 전달한다.
- 결과 합성 모듈 : 부질의 결과를 합성하여 사용자에게 리턴한다.



(그림 7) 링크 질의를 위한 질의 처리기 모델링

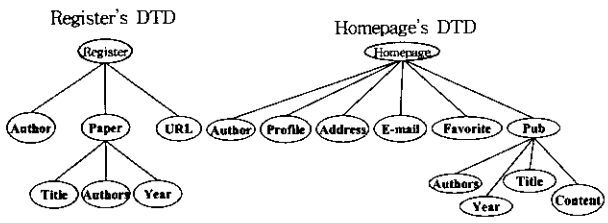
질의 처리기의 링크 질의 처리 과정은 다음과 같다. 통신 모듈을 통해 전송받은 링크 질의는 XML-QL 파서를 통해 문법적 오류를 검증받고, 파싱된 결과는 부질의 생성 모듈로 전달된다. 부질의 생성 모듈은 파싱된 링크 질의를 대상으로, XML 소스 내 질의 문서의 링크 정보를 추출한다. 부질의 생성 모듈은 사용자 질의로부터 부질의 템플릿을 생성하고, 추출된 질의 문서의 링크 정보를 이용하여 부질의를 생성한다. 부질의의 참조 문서(들)가 질

의 처리 사이트에 있는 경우, 부질의 생성 모듈은 부질의(들)를 처리하여 그 결과(들)를 질의 합성 모듈로 전달한다. 부질의의 참조 문서(들)가 외부 사이트(들)에 있는 경우, 부질의 생성 모듈은 부질의(들)를 부질의 전송 모듈로 전달하고, 부질의 전송 모듈은 해당 외부 사이트의 질의 처리기로 부질의(들)를 전송한다. 각 외부 사이트의 질의 처리기는 자신의 XML 소스에 저장된 참조 문서를 대상으로 부질의를 처리하고 자신의 부질의 결과 전송 모듈을 통해 부질의 결과를 질의 처리 사이트로 리턴한다. 리턴되는 부질의 결과(들)는 질의 처리 사이트의 부질의 결과 수신 모듈을 통해 결과 합성 모듈로 전달된다. 질의 처리 사이트나 외부 사이트(들)로부터 리턴된 부질의 결과(들)는 결과 합성 모듈에 의해 합성되어 XML 문서 형식의 결과로 사용자에게 리턴한다. 결과 합성 모듈은 단지 여러 XML 문서를 단순히 첨가하여 하나의 XML 문서로 합성할 뿐만이 아니라, 사용자의 요청에 따라 링크 질의 결과 내용 중 같은 엘리먼트 이름이 존재하거나 엘리먼트의 검색 결과가 서로 같은 내용을 포함할 경우, 중복된 내용을 제거하고 하나의 합성된 결과만을 생성할 수 있다. 또한, 결과 합성 과정 중 검색된 결과를 정렬하는 기능을 제공한다.

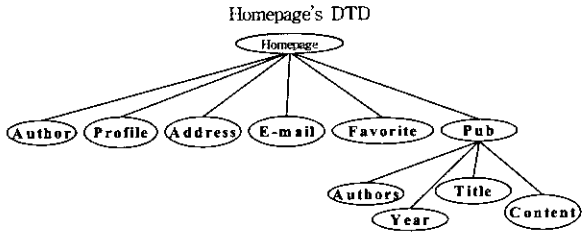
5.2 링크 질의 처리 예

본 절에서는 링크 정보 관리 테이블을 활용한 부질의 생성 및 처리와 그 결과의 합성 과정으로 구성되는 링크 질의 처리의 예를 기술한다. 다음은 링크 질의 처리 과정을 기술하기 위해 사용된 링크 질의의 예이다.

www.a.b.c/register.xml 내 기록된 사람 중 저자 이름이 "A. Einstein"인 사람의 homepage를 찾아 publication 정보를 찾으시오.



(a) 질의 처리 사이트 내

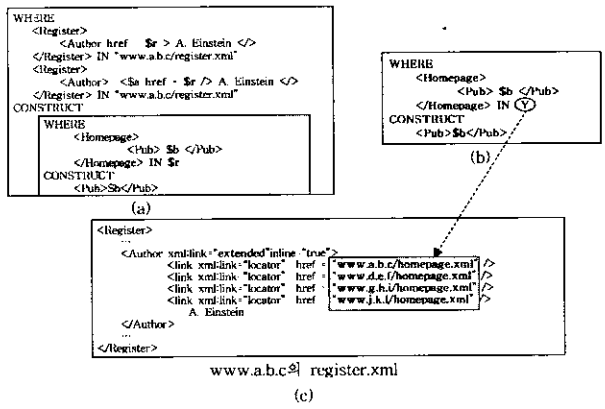


(b) 외부 사이트 내

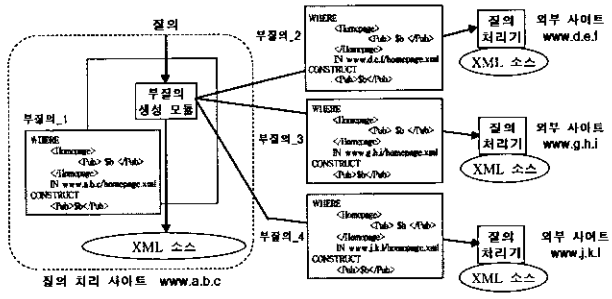
(그림 8) DTD의 구조

"www.x.y.z" 사이트가 관리하는 컴퓨터 단말기를 통해 사용자가 위의 질의를 입력한다고 가정하면, 사용자 사이트는 "www.x.y.z"이고, 질의 처리 사이트는 "www.a.b.c"이며, 질의 문서의 URL은 "www.a.b.c/register.xml"이 된다. 위의 링크 질의에 대해 외부 사이트들은 모두 3개로 "www.d.e.f", "www.g.h.i" 그리고 "www.j.k.l"이라고 가정하자. (그림 8)는 질의 처리 사이트와 외부 사이트가 관리하는 XML 문서의 DTD 구조를 나타낸 것이다.

사용자 사이트의 사용자 인터페이스를 통해 입력받은 링크 질의는 질의 처리 사이트로 전송되고, 부질의 생성 모듈은 링크 질의로부터 부질의를 생성한다. (그림 9)에서 (a)는 링크 질의, (b)는 링크 질의로부터 추출된 부질의 템플릿, (c)는 "www.a.b.c"에 저장된 register.xml 문서의 링크 내용을 나타낸 것이다.



(그림 9) 링크 질의, 부질의 템플릿, 그리고 링크 정보 관리 테이블의 예

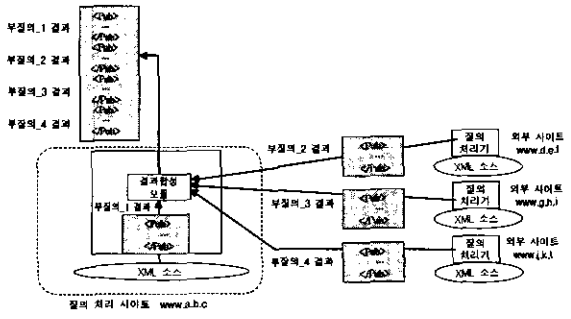


(그림 10) 부질의의 내용 및 부질의가 전송되는 사이트들

(그림 9)에 나타난 것과 같이 생성되는 부질의의 총 개수는 4개이다. 질의 처리 사이트 내 부질의 생성 모듈에 의해 생성된 부질의들은 질의 처리 사이트 내에서 처리되거나 외부 사이트로 전송하게 된다. (그림 10)은 부질의의 내용과 부질의가 전송되는 사이트를 표시한 것이다. (그림 10)에서 부질의_1은 질의 처리 사이트 "www.a.b.c"로, 부질의_2, 부질의_3, 부질의_4는 각각 외부 사이트 "www.d.e.f", "www.g.h.i", "www.j.k.l"로 전송되어 처리된다.

부질의를 전송받은 외부 사이트는 부질의를 처리하

여 그 결과를 XML 문서로 작성한다. 이 작성된 XML 문서는 결과 전송/수신 모듈을 통해 질의 처리 사이트의 결과 합성 모듈로 전송되고, 결과 합성 모듈은 부질의 결과를 합성하여 XML 문서 형식의 최종 결과를 작성하고 이를 사용자에게 리턴한다. (그림 11)은 부질의 처리 후 최종 결과 XML 문서로 합성되는 과정을 나타낸 것이다.



(그림 11) 부질의 처리 후 최종 결과 XML 문서로의 합성

6. 성능 분석

본 장에서는 3.1절에서 기술한 링크 질의 처리 환경 하에서 링크 질의 처리의 성능을 평가하기 위해, 링크 질의 처리 시간에 영향을 미치는 파라미터를 분류한 후, 질의 처리 시간을 수식화하여 분석하였다.

먼저, 링크 질의 처리 시간은 기본적으로 아래 식 (1)에서 열거한 4가지 항들의 합으로 구성된다.

링크 질의 처리 시간 = 사용자 수작업에 의한 질의 생성 시간 (A)
+ 네트워크 상에서의 사이트들 간 메시지 전송 시간 (B)
+ 관련 사이트 내 질의 처리기에서의 질의 처리 시간 (C)
+ 질의 결과 합성 소요 시간 (D) (1)

첫 번째 항(이하, A항)은 사용자가 요구하는 결과를 도출할 수 있도록 사용자 인터페이스를 통해 수작업으로 질의를 생성하는 시간이다. Web을 대상으로 하는 질의는 기존 DBMS를 대상으로 하는 질의(예를 들어 SQL 또는 OQL과 같은 질의어로 생성된 질의)와는 달리 명료하지 않고 그 표현과 의미가 모호하다. 그렇게 때문에 사용자가 원하는 결과를 얻을 때까지 수 차례의 질의 처리와 결과 리턴, 그리고 결과 리턴에 따른 재질의(결과를 사용자가 분석하여 새롭게 생성한 질의)의 생성 및 처리를 반복해야 한다. 그러므로, Web 환경에서 질의 처리를 신속하게 하기 위해서는 수작업에 의한 질의 생성을 최소한으로 수행하는 방법이 필요하다. 따라서 Web 상에서의 링크 질의 처리 시간은 사용자 수작업에 의한 질의 생성 시간을 포함한다.

두 번째 항(이하, B항)은 서로 다른 사이트 간의 질의 전송과 결과 리턴, 그리고 본 논문에서 링크 질의 처리를 위해 제시한 부질의 전송과 부질의 결과 리턴 과정에서 발생하는 메시지 전송 시간이다. 세 번째 항(이하, C항)은 XML 소스를 관리하는 사이트 내 질의 처리기가 전송된 질의 또는 부질의를 처리하는 시간이다. 마지막 네 번째 항(이하, D항)은 링크 질의 처리 결과를 합성하는 데 드는 시간이다. 현재 Web을 대상으로 하는 질의는 여러 사이트에서 처리된 결과(들)를 사용자에게 리턴한다. 그러면 사용자는 수작업에 의해 그 결과(들)을 합성하여 최종 결과를 얻게 됨으로써 질의 처리 시간이 길어지게 된다. 그러므로, Web 환경에서 질의를 신속하게 처리하기 위해서는 수작업에 의한 질의 결과 합성을 최소한으로 수행하는 방법이 필요하다. 따라서 A 항과 마찬가지로 링크 질의 처리 시간은 질의 결과 합성 소요 시간을 포함한다.

식 (1)을 기본으로 하여 기존의 XML 질의어를 이용하여 수작업 기반으로 링크 질의를 처리할 경우(3.1절 참조)와 4.2절에서 제시한 확장된 질의어를 이용하여 부질의 기반으로 링크 질의를 처리할 경우와의 질의 처리 시간을 비교한다. 다음은 성능 비교를 위해 가정한 내용들이다.

① 사용자 사이트와 질의 처리 사이트는 서로 다른 사이트이다.
② 수작업 기반으로 링크 질의를 처리할 때, URL 질의가 여러 개인 경우, 수작업에 의해 한 URL 질의를 생성하는 데 드는 시간은 서로 동일하다.
③ 각 사이트 내 질의 처리기에서의 질의(또는 부질의) 처리 시간과 부질의를 생성하기 위해 XML 소스에 접근하여 링크 정보를 추출하는 데 소요하는 시간은 서로 동일하다.
④ 사이트 간 질의(또는 부질의) 전송 시간은 동일하다.
⑤ 질의(또는 부질의)의 결과가 XML 문서 형식을 갖는 경우, 사이트 간 질의(또는 부질의) 결과 전송 시간은 서로 동일하다. 또한 수작업 기반 링크 질의 처리의 결과가 URL인 경우, URL 전송 시간도 질의(또는 부질의)의 결과 전송 시간과 동일하다. 그러나 사이트 간 질의 전송(또는 부질의) 전송 시간보다는 오래 걸린다.

링크 질의 처리에 관련된 성능 파라미터는 <표 2>와 같다.

6.1 수작업 기반 링크 질의 처리

기존의 XML 질의어를 이용하여 수작업 기반으로 링크 질의를 처리할 경우의 질의 처리 시간 T_N은 식 (2)와 같다.

$$T_N = T_H \times (1 + L) + T_{QT} \times (1 + L) + T_{RT} \times (1 + L)$$

$$+ T_P \times (1+L)$$
$$+ T_{IN} \qquad (L \geq 1) \qquad (2)$$

질의 처리 시간은 링크 질의 처리 과정 시 참조되는 링크의 수가 많을 수록 길어진다. 이와 관련하여 사용자 수작업에 의해 생성되는 총 질의의 수는 초기 질의(사용자가 최초로 질의 처리 사이트로 전송하는 질의)와 초기 질의 처리 결과에 따른 링크(질의 처리 사이트와 외부 사이트를 대상으로 하는 질의) 수의 합(1+L)이 된다. 그러므로, T_N의 A항은 수작업에 의해 한 질의를 생성하는 시간 T_H와 총 질의 수와의 곱으로 구할 수 있다.

네트워크 상에서의 사이트들 간 전송되는 총 메시지 수는 사용자 사이트와 질의 처리 사이트, 사용자 사이트와 외부 사이트 간의 질의 전송과 결과 리턴으로 인해, 사용자 수작업에 의해 생성된 총 질의 수의 2배(2×(1+L))가 된다. T_N의 B항은 사이트 간 총 질의 전송시간($T_{QT}\times(1+L)$)과 사이트 간 총 질의 결과/URL 전송시간($T_{RT}\times(1+L)$)의 합으로 구할 수 있다. T_N의 C항은 질의 처리 사이트와 외부 사이트에서의 질의 처리 시간의 합($T_P\times(1+L)$)으로 구할 수 있다. T_N의 D항은 수작업에 의한 전체 결과 합성 소요 시간 T_{IN}이다.

〈표 2〉 링크 질의 성능 파라미터

파라미터	내 용
T_N	수작업 기반 링크 질의 처리 경우의 질의 처리 시간
T_S	부질의 기반 링크 질의 처리 경우의 질의 처리 시간
P_i	사용자로부터 입력받은 링크 질의가 3.2절의 (유형 i)에 해당하는 질의일 확률 (i = 1, 2, 3) $\sum_{i=1}^{3} P_i = 1$
T_H	수작업 기반의 경우 수작업에 의해 한 질의(초기 질의 또는 URL 질의)를 생성하는 데 드는 시간, 또는 부질의 기반의 경우 수작업에 의해 한 링크 질의를 생성하는 데 드는 시간
T_{QT}	사용자 사이트, 질의 처리 사이트와 외부 사이트 간의 한 질의(부질의) 전송 시간
T_{RT}	사용자 사이트, 질의 처리 사이트와 외부 사이트 간의 한 질의(또는 부질의) 결과 XML 문서/URL의 전송 시간 ($T_{QT} < T_{RT}$)
T_Q	부질의 기반 링크 질의 처리의 경우, 사용자 질의를 전송받은 질의 처리 사이트가 전체 링크 질의에 대한 일련의 처리(본문 6장 (2) 참조)를 수행하는 데 드는 시간
T_P	질의 처리 사이트 또는 외부 사이트에서 한 질의(또는 한 부질의)의 처리 시간, 또는 부질의 생성하기 위해 XML 소스에 접근하여 링크 정보를 추출하는 데 소요하는 시간
T_G	한 개의 부질의 생성 시간($T_G << T_H$)
T_{IN}	수작업 기반 링크 질의 처리 경우의 전체 결과 합성 소요 시간
T_{IS}	부질의 기반 링크 질의 처리 경우의 전체 부질의 결과 합성 소요 시간($T_{IS} << T_{IN}$)
L	링크 질의 내 링크 수 또는 링크 질의 당 생성된 총 부질의 수 ($L \geq 1$)
L_l	질의 처리 사이트에서 처리하는 부질의의 개수
L_r	외부 사이트에서 처리하는 부질의의 개수

6.2 부질의 기반 링크 질의 처리

4.2절에서 제시한 확장된 질의어를 이용하여 부질의 기반으로 링크 질의를 처리할 경우의 질의 처리 시간 T_S는 식 (3)과 같다. 식 (3)은 식 (1) 형태의 질의 처리 시간을 유도하기 위해 간략화한 식이다.

$$T_S = T_H + T_{QT} + T_{RT} + T_Q \qquad (3)$$

수작업 기반의 경우와는 달리, 부질의 기반의 경우에는 사용자 수작업에 의해 생성되는 총 질의의 수는 1개이다. 이는 링크를 참조하여 부질의를 재구성하고 관련 사이트로 부질의를 전송함으로써 사용자가 원하는 결과를 얻기까지 한번의 질의를 생성하면 되기 때문이다. 식 (3)는 링크 질의 처리 시간이 사용자 수작업에 의한 질의 생성 시간(T_H), 사용자 사이트와 질의 처리 사이트 간의 질의 전송 시간(T_{QT}), 결과 전송 시간(T_{RT}), 그리고 사용자 질의를 전송받아 질의 처리 사이트에서 일련의 질의 처리(부질의 재구성, 부질의를 해당 사이트로의 전송 및 부질의 처리, 부질의 결과 리턴, 그리고 부질의 결과의 합성 등)를 하는 시간(T_Q)의 합으로 구성됨을 나타내고 있다.

식 (3)에서 사용된 T_Q을 기술하면 식 (4)와 같다.

$$T_Q = T_P + T_G \times L + T_{QT} \times L_r + T_{RT} \times L_r + T_P \times L + T_{IS} \qquad (L_l + L_r \geq 1) \qquad (4)$$

질의 처리 사이트에서 일련의 질의 처리를 수행하는 시간은 부질의 생성하기 위해 XML 소스에 접근하여 링크 정보를 추출하는 데 소요하는 시간(T_P), 링크 정보를 활용하여 부질의를 생성하는 시간($T_G\times L$), 질의 처리 사이트와 외부 사이트 간의 부질의 전송 시간($T_{QT}\times L_r$), 결과 전송 시간($T_{RT}\times L_r$), 각 부질의를 해당 사이트에서 처리하는 시간($T_P \times L$), 그리고 전체 부질의 결과 합성 소요 시간 T_{IS}의 합으로 계산된다. 이때, 참조 문서가 질의 처리 사이트에 저장된 경우에는 네트워크 상에서의 메시지 전송 시간이 소요되지 않는다. 이는 질의 처리 사이트 내 부질의 생성 모듈이 직접 부질의를 처리하기 때문이다.

식 (3)에 식 (4)를 대입하여 식 (1) 형태로 표현하면 식 (5)와 같다.

$$T_S = T_H + T_G \times L + T_{QT} \times (1+L_r) + T_{RT} \times (1+L_r) + T_P \times (1+L) + T_{IS} \qquad (5)$$

T_S의 A항은 사용자 수작업에 의한 질의 생성 시간 T_H와 부질의를 생성하는 시간($T_G \times L$)의 합으로 구할 수 있다. T_S의 B항은 질의 처리 사이트와 외부 사이트 간의 부질의 전송 시간($T_{QT} \times (1 + L_r)$)과 결과 전송 시간($T_{RT} \times (1 + L_r)$)의 합으로 구할 수 있다. T_S의 C항은 부질의 생성하기 위해 XML 소스에 접근하여 링크 정보를 추출하는 데 소요하는 시간(T_P)과 각 부질의를 해당 사이트에서 처리하는 시간($T_P \times L$)의 합이 되며, T_S의 D항은 전체 부질의 결과 합성 소요 시간 T_{IS}가 된다.

6.3 수작업 기반 링크 질의 처리의 경우와 부질의 기반 링크 질의 처리의 경우와의 링크 질의 처리 시간 비교

식 (6)은 수작업 기반 링크 질의 처리의 경우와 부질의 기반 링크 질의 처리의 경우의 링크 질의 처리 시간을 비교하기 위해 식 (2)에 식 (5)를 뺀 식 $T_N - T_S$을 나타낸 것이다.

$$\begin{aligned} T_N - T_S &= (T_H - T_G) \times L \\ &\quad + T_{QT} \times L_l + T_{RT} \times L_l \\ &\quad + T_{IN} - T_{IS} \end{aligned} \tag{6}$$

식 (6)을 3.2절에서 분류한 링크 질의 유형별로 나누어서 나타내면 식 (7)과 같다. 링크 질의가 (유형 3)일 때는 참조 문서가 질의 처리 사이트와 외부 사이트 모두에 있는 경우로서 식 (6)과 동일하다.

$$\begin{aligned} \textbf{(유형 1)} \quad T_N - T_S &= (T_H - T_G) \times L_l \\ &\quad + T_{QT} \times L_l + T_{RT} \times L_l \\ &\quad + T_{IN} - T_{IS} \\ \textbf{(유형 2)} \quad T_N - T_S &= (T_H - T_G) \times L_r \\ &\quad + T_{IN} - T_{IS} \\ \textbf{(유형 3)} \quad T_N - T_S &= (T_H - T_G) \times (L_l + L_r) \\ &\quad + T_{QT} \times L_l + T_{RT} \times L_l \\ &\quad + T_{IN} - T_{IS} \end{aligned} \tag{7}$$

링크 질의가 3가지 유형에 해당할 확률을 곱해서, 수작업 기반 링크 질의 처리의 경우와 부질의 기반 링크 질의 처리의 경우와의 링크 질의 처리 시간의 차이 Δ를 나타내면 식 (8)과 같다.

$$\begin{aligned} \Delta = T_N - T_S = \\ P1[(T_H - T_G) \times L(=L_l) + (T_{QT} + T_{RT}) \times L_l] \\ + P2[(T_H - T_G) \times L(=L_r)] \\ + P3[(T_H - T_G) \times L + (T_{QT} + T_{RT}) \times L_l] + T_{IN} - T_{IS} \end{aligned} \tag{8}$$

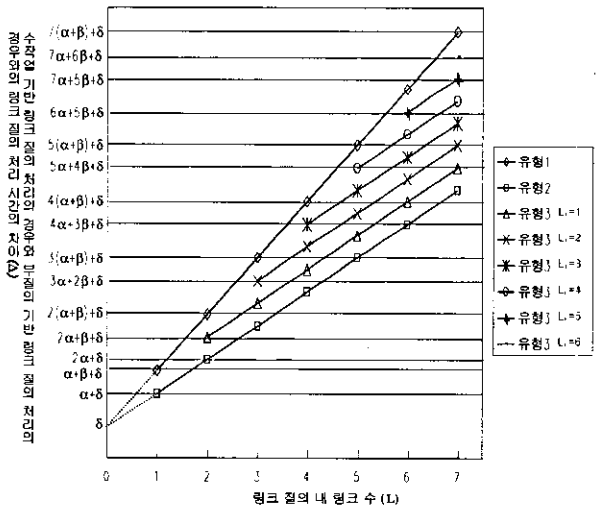
식 (6)과 식 (8)을 분석하면 다음과 같이 수작업 기반 링크 질의 처리 대비 부질의 기반 링크 질의 처리 시간의 성능을 정리할 수 있다.

① (수작업 기반 링크 질의 처리 경우의 전체 결과 합성 소요 시간 - 부질의 기반 링크 질의 처리 경우의 전체 부질의 결과 합성 소요 시간)의 합만큼 질의 처리 시간이 줄어든다.

② (수작업에 의해 한 질의를 생성하는 데 드는 시간 - 한 개의 부질의 생성 시간)의 부질의 개수 배만큼 질의 처리 시간이 줄어든다.

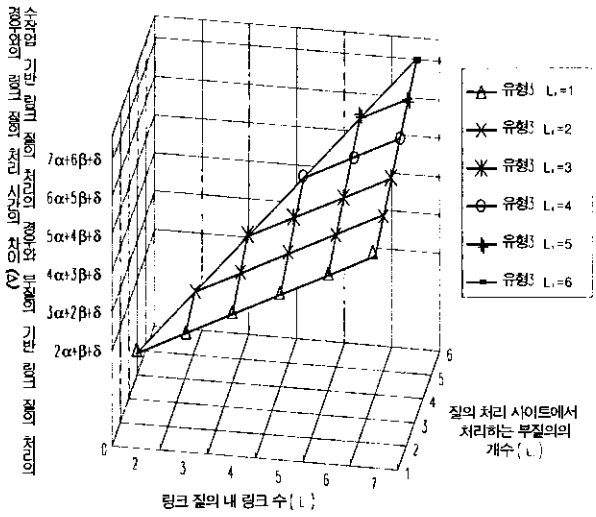
③ 참조 문서가 질의 처리 사이트에 많이 있을수록, 사이트 간 부질의 전송 시간과 부질의 결과 전송 시간이 소요되지 않기 때문에 이에 비례하여 질의 처리 시간이 줄어든다.

다음은 식 (6)과 식 (8)을 이용한 성능 비교의 몇 가지 예를 기술한다. 먼저, (그림 12)는 링크 질의 내 링크 수의 증가함에 따라 수작업 기반 링크 질의 처리의 경우와 부질의 기반 링크 질의 처리의 경우와의 링크 질의 처리 시간의 차이 Δ를 3.2절에서 분류한 링크 질의 유형별로 비교한 예이다. 링크 질의 유형 3의 경우에는 L_l의 개수가 1, 2, 3, 4, 5, 6인 경우를 나누어 비교하였다. Δ를 간략화시키기 위해 $(T_H - T_G)$는 α로, $(T_{QT} + T_{RT})$는 β로, $(T_{IN} - T_{IS})$는 δ로 대치하였다. (그림 12)는 α, β, δ 간의 크기 비교를 고려하지 않았다. (그림 12)에서 보는 것과 같이 질의 처리 유형 2, 유형 3, 유형 1 순서로 질의 처리 시간 차이가 커졌다. 참조 문서가 질의 처리 사이트에 많이 있을수록 부질의 기반 링크 질의 처리 시간이 줄어들기 때문이다. 또한 (그림 12)에서는 유형 2, 유형 3보다 유형 1의 경우에 링크 질의 내 링크 수 대비 질의 처리 시간 차이의 증가 비율이 더 크게 나타났다. 유형 1은 참조 문서가 질의 처리 사이트에만 있으므로 유형 2, 유형 3보다 부질의 기반 링크 질의 처리 시간이 더 줄어들기 때문이다.



(그림 12) L의 증가에 따른 Δ 비교

(그림 13)은 (그림 12)에서 질의 처리 유형 3에 해당하는 내용을 질의 처리 사이트에서 처리하는 부질의 개수 L_l의 증가에 따라 3차원으로 나타낸 것이다. L 내에 L_l이 많을수록 질의 처리 시간의 차이가 커짐을 알 수 있었다.

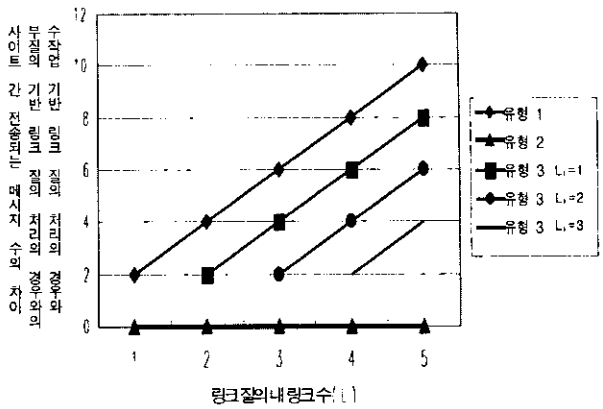


(그림 13) L, L_l의 증가에 따른 Δ 비교

다음은 네트워크 상에서의 사이트들 간 전송되는 메시지 수에 따른 수작업 기반의 경우와 부질의 기반의 경우의 성능을 비교 실험한 예로서, 실험을 위해서 설정된 파라미터는 다음과 같다.

질의 유형 \ 링크 질의 내 링크 수		L=1	L=2	L=3	L=4	L=5
유 형 1	L_l	1	2	3	4	5
	L_r	0	0	0	0	0
유 형 2	L_l	0	0	0	0	0
	L_r	1	2	3	4	5
유 형 3 (L_l = 1)	L_l	-	1	1	1	1
	L_r	-	1	2	3	4
유 형 3 (L_l = 2)	L_l	-	-	2	2	2
	L_r	-	-	1	2	3
유 형 3 (L_l = 3)	L_l	-	-	-	3	3
	L_r	-	-	-	1	2

상기 파라미터는 3.2절에서 기술한 링크 질의 유형 별로 링크 질의 내 링크 수 L에 따라 질의 처리 사이트에서 처리하는 부질의의 개수 L_l와 외부 사이트에서 처리하는 부질의의 개수 L_r의 분포를 나타낸다. (그림 14)에서 보는 것과 같이 질의 처리 유형 2인 경우에는 두 질의 처리 간의 메시지 수 차이가 없으나 유형 3, 유형 1인 경우에는 (L_l×2)에 비례하여 메시지 수 차이를 나타냈다. 참조 문서가 질의 처리 사이트에 있는 경우에는 부질의/부질의 결과 전송에 따른 네트워크 비용이 들지 않기 때문이다.



(그림 14) L의 증가에 따른 두 질의 처리 간의 메시지 수 차이 비교

7. 결론 및 향후 연구

본 논문에서는, Web 상에 분산되어 저장된 Web 문서들 중 하이퍼링크로 연결된 XML 문서들을 대상으로 효율적인 검색을 지원하기 위해, 링크 질의에 대해 정의하고, 링크 질의 표현을 위한 XML 질의어의 확장과 링크 질의 처리 기법을 제시하고, 이의 성능을 분석하였다.

링크 질의는 특정 URL을 갖는 XML 문서(질의 문서)를 대상으로 전체 또는 일부 내용을 검색할 뿐 아니라 질의 문서에 의해 링크로 참조되는 XML 문서(참조 문서)의 전체 또는 일부의 내용을 검색하는 질의로 정의하였다. 그리고 참조 문서가 저장된 위치에 따라 링크 질의를 3가지 유형으로 분류하였다.

링크 질의를 표현하기 위해 XML-QL을 확장하였는데 확장된 XML 질의어의 특징을 정리하면 다음과 같다.

- XML 문서 내의 링크를 지원하는 질의어이다.
- XML 질의 구조는 XML 문서 형태를 갖는다.
- XML 질의 실행 결과는 XML 문서 형태를 갖는다.
- XML 질의어는 내용/구조/속성 기반 검색 기능 및 중복 회피, 정렬 기능을 지원한다.
- XML 질의어의 질의 가능 영역은 한 사이트 내에 저장된 XML 문서 간의 연산을 지원할 뿐만 아니라 타 사이트에 분산되어 있는 XML 문서 간의 연산도 지원한다.
- XML 질의 실행 결과에 대한 재구성이 가능하다.

링크 질의를 효율적으로 처리하기 위해 사용자 인터페이스를 통해 받아들인 질의를, 참조 문서만을 대상으로 검색하는 부질의로 재구성하여 사용하였다. 부질의는 참조 문서를 관리하는 사이트로 전송되어 해당 사이트에서 부질의를 처리하고, 그 결과는 질의 처리

사이트로 리턴된다. 부질의를 생성하는 과정은 사용자 인터페이스를 통해 입력받은 질의로부터 부질의 템플릿을 만들고, 참조 문서의 URL을 지정함으로써 자동으로 생성한다.

링크 질의는 단순히 부질의의 결과를 첨가하여 하나의 XML 문서로 합성할 뿐만 아니라, 사용자의 요청에 따라 링크 질의 결과 내용 중 같은 엘리먼트 이름이 존재하거나 엘리먼트의 검색 결과가 서로 같은 내용을 포함할 경우, 중복된 내용을 제거하고 하나의 합성된 결과만을 생성할 수 있다. 또한, 결과 합성 과정 중 검색된 결과를 정렬하는 기능을 제공한다.

다음의 표는 수작업 기반의 질의 처리 방법과 부질의 기반의 링크 질의 처리 방법을 항목별로 비교한 내용이다.

처리 방법 / 비교 항목	수작업 기반 링크 질의 처리	부질의 기반 링크 질의 처리
사용되는 질의어	기존 XML 질의어 사용	링크 질의를 표현하기 위해 확장된 XML-QL 사용
수작업으로 사용자 인터페이스에 입력되는 질의	초기 질의, URL 질의(들)	한번의 링크 질의
질의 결과 형태	URL(들)	질의 결과에 부합되는 XML 문서
질의 처리 방법	기존의 XML 검색 시스템 활용	링크 질의를 처리하기 위한 질의 처리기 활용

수작업 기반으로 링크 질의를 처리할 경우와 대비하여 부질의 기반으로 링크 질의를 처리할 경우의 질의 처리 시간의 성능을 비교하면, 참조 문서로의 링크가 많을수록, 링크 수에 비례하여 수작업에 의한 질의 생성 시간과 결과 합성에 소요 시간만큼 부질의 기반의 질의 처리 시간이 줄어들고, 참조 문서가 질의 처리 사이트에 많이 있을수록, 이에 비례하여 부질의 기반의 질의 처리 시간이 줄어들었다.

향후 연구로는 여러 질의 문서를 대상으로 참조 문서가 질의 문서와 직접 링크된 것이 아니라 여러 XML 문서를 거쳐 링크되어 있는 경우의 복잡한 링크 질의에 대해 유형을 분류하고 그에 따른 질의 처리 모델 및 질의 처리 방법을 제시하고, 또한 질의 처리 결과 합성에서 중복을 회피하고 정렬을 수행하는 알고리즘을 개발하는 것이다. 그리고, 외부 사이트에 질의 처리기가 존재하지 않는 경우에서의 링크 질의 처리 방법에 대해 연구하는 것이다.

참 고 문 헌

[1] T. Bray et al., "Extensible Markup Language (XML) 1.0," http://www.w3.org/TR/1998/REC-xml-19980210, 1998.

[2] C. Baru et al. "XML-Based Information Mediation with MIX," Proc. of the 1999 ACM SIGMOD Int'l Conf. on Management of Data, pp.597-599, Jun. 1999.

[3] 이경하 외, "XMF : XML기반 분산 이질 정보자원의 통합 프레임워크," KDBC 2000 학술발표논문집, pp.262-270, 2000.

[4] 오준환 외, "3계층 XML문서저장 시스템의 설계", 2000 춘계 학술대회논문집, 한국정보처리학회, 2000.

[5] 이용석 외, "XML 문서저장 시스템의 설계 및 구현", '98 가을 학술발표논문집, 한국정보과학회, Vol.25, No.2, pp.347-349, 1998.

[6] J. Robie et al., "XML Query Language (XQL)," http://www.w3.org/TandS/QL/QL98/pp/xql.html, 1998.

[7] A. Deutsch et al., "XML-QL : A Query Language for XML," http://www.w3.org/TR/NOTE-xml-ql/, 1998.

[8] J. Clark and S. DeRose., "XML Path Language (XPath) Version 1.0," http://www.w3.org/TR/xpath, 1999.

[9] D. Chamberlin et al., "XQuery : A Query Language for XML," http://www.w3.org/TR/xquery, 2001

[10] D. Florescu and D. Kossmann, "Storing and Querying XML Data Using an RDBMS," Bulletin of the Technical Committee on Data Engineering, Vol.22, No.3, pp.27-34, 1999.

[11] 윤정희 외, "객체지향 데이타베이스의 XML로의 표현", 2000 봄 학술발표논문집, 한국정보과학회, 제27권, 제1호, pp.143-145, 2000.

[12] M. Fernandez et al., "Catching the Boat with Strudel : Experience with a Web-Site Management System," SIGMOD Record, Vol.27, No.3, pp.414-425, 1998.

[13] A. Sahuguet and F. Azavant, "W4F : A WysiWyg Web Wrapper Factory," http://cheops.cis.upenn. edu/ sahuguet/WAPI/wapi.ps.gz, 1999.

[14] D. Florescu et al., "Database Techniques for the World-Wide Web : A Survey," SIGMOD Record, Vol.27, No.3, pp.59-74, 1998.

[15] D. Chamberlin et al., "Quilt : An XML Query Language for Heterogeneous Data Sources," http://www.almaden.ibm.com/cs/people/chamberlin/quilt_lncs.pdf, 2000.

[16] S. DeRose et al., "XML Linking Language(XLink)," http://www.w3.org/TR/xlink, 1999.

[17] S. DeRose et al., "XML Pointer Language (XPointer)," http://www.w3.org/TR/WD-xptr, 1999.

[18] 김은정, 배종민, "XML 링크정보를 이용한 정보검색 색인기법의 설계," 정보처리논문지, 제7권, 제7호, pp.2020-2027, 2000.

[19] 김상준 외, "XML 링크의 메타 데이타를 이용한 검색 시스템의 설계," 2000 봄 학술발표논문집, 한국정보과학회, 제27권, 제1호, pp.157-159, 2000.

문 찬 호

e-mail : moonch@rose.cse.cau.ac.kr

1997년 중앙대학교 컴퓨터공학과 졸업 (공학사).

1999년 중앙대학교 컴퓨터공학과 대학원 졸업(공학석사).

현재 중앙대학교 컴퓨터공학과 박사과정 재학중.

관심분야 : Web 데이타베이스, XML

강 현 철

e-mail : hckang@cau.ac.kr

1983년 서울대학교 컴퓨터공학과 졸업 (공학사)

1985년 U. of Maryland at College Park, Computer Science(M.S.)

1987년 U. of Maryland at College Park, Computer Science(Ph.D.)

1988년~현재 중앙대학교 컴퓨터공학과 교수

관심분야 : 이동 데이타베이스, 웹 데이타베이스, DBMS 저장 시스템