경로 서명 : XML문서 검색을 위한 경로-지향 질의처리 시스템

박희숙* · 박주현* · 조우현*

Path Signatures : Path-oriented Query Processing System for XML document Retrieval

Hee-Sook Park* · Ju-Hyun Park* · Woo-Hyun Cho*

요 약

최근 인터넷의 폭발적인 성장과 인기로 인하여, 인터넷상에서 정보의 교환이 매우 빠르게 증가하고 있다. 또한 XML은 인터넷상에서 정보교환의 수단인 동시에 표준으로서 자리를 잡아 가고 있다. 따라서 본 논문에서는 경로-지향 질의어를 평가하기 위한 새로운 인덱싱 기법과 사용자들에게 편리한 XML문서 검색 결과를 제공하기 위한 경로-지향 질의 처리 시스템의 설계 및 구현을 제공한다. 제안된 인덱싱 기법은 XML문서의 검색 성능을 향상시키기 위하여 이진 트라이 구조와 경로서명 파일을 결합하였다.

ABSTRACT

Recently, due to the popularity and explosive growth of the Internet, the information exchange is increasing so rapidly over the Internet. Also the XML is becoming a standard as well as a major tool of data exchange on the Internet and thus we propose the new indexing technique for evaluating a path-oriented query and design and implementation of Path-oriented Query Processing System to give useful for users. In proposed indexing technique, which combined a binary trie structure with a path signature file to improve performance of XML document retrieval.

키워드

Path Signature, Binary Trie, XML, Path-oriented Query Language

Ⅰ. 서 론

인터넷의 발달과 함께 그 중요성이 부각된 XML은 네트워크상에서 문서정보를 교환하기 위한 표준으로서 그 위치를 더욱 확고히 하고 있다. 또한 관계형 데이터베이스에 저장된 문서를 효과적으로 검색하기 위한 방법들에 대한 연구의 필요성이 대두됨에 따라 현재 이들에 대한 연구가 활발히 진행되고 있다. 현재 XML은 인터넷을 포함한 다양한 장소에서 데이터 교환의 실질적인 표준으로 자리 잡고 있다.

XML은 구조(structure), 데이터(data) 그리고 표현(description)이 분리된 데이터 정의를 위한 언어로서 HTML과 SGML이 갖는 단점을 보완한 것이다[4]. XML은 원래 대규모 전자출판 문제를 해결하기 위한 목적으로 설계되었으나 현재 e-비지니스/전자상거래와 관련한 마크업 언어, 수학 공식, 화학 분자 구조, 그래픽, 회계 데이터등과 같은 데이터를 마크업 하기 위해 사용되고 있다[1][9]. 따라서 XML문서의 활용 분야가 더욱 광범위하게 됨에 따라서 오늘날 가장 일반적으로 사용되고 있는 관계형 데이터베이스에 XML문서를 저장하는 문제

* 부경대학교 전자컴퓨터정보통신공학부

접수일자 : 2007. 3. 9

와 저장된 XML문서를 경로-지향 질의어를 통해 효율적으로 검색하기 위한 질의 평가 속도 개선 문제는 중요한 연구과제가 되고 있다.

경로-지향적 질의 언어(Path-oriented Query Language)의 종류로는 XPath(XML Path Language), XML-QL, XQL 등이 있다[3][5][6].

본 논문에서는, XPath와 같은 경로-지향 질의가 입력될 때 구조적으로 저장된 XML문서에 대한 질의 평가 속도를 효과적으로 개선하기 위한 새로운 인덱싱 기법을 제안한다. 또한 제안한 인덱싱 기법을 적용하는 경로-지향 질의 처리 시스템의 설계 및 구현을 한다.

본 논문에서 제안한 인덱싱 기법은 기존의 경로서명 이론과 이진 트라이 구조를 접목하여 사용한다.

본 논문은 다음과 같이 구성된다. 2장에서는 XML기술과 경로-지향언어들에 대하여 간략히 설명하고, 3장에서는 본 논문에서 제안한 인덱싱 기법을 위한 경로서명 파일과 이진 트라이 구조에 대하여 기술하고, 4장은 경로-지향 질의처리 시스템의 설계 및 구현에 관하여 설명한다. 마지막 5장에서 결론 및 향후과제에 대하여 논의한다.

Ⅱ. XML기술과 경로-지향 언어

XML은 SGML(Standard Generalized Markup Language)의 부분집합으로 설계되었으며 W3C에 의해 권장되고 있다. XML은 단순하고 매우 유연한 텍스트 형식을 가지고 있으며, 이것의 기본 구문은 HTML과 유사하지만 그 목적은 다르다. XML의 가장 큰 특징은 XML이 메타언어(Meta Language)라는 것이다. 이것은 XML이 HTML과 같이 어떤 문서를 기술하는 문서유형을 제공하는 것이 아니라 문서유형을 만드는 역할을 하기 때문이다[1][9].

XML문서는 DTD(Document Type Descriptor) 또는 스키마(Schema)에 의해 미리 기술된 구조와 태그 규칙을 따라야만 한다. 작성된 XML문서의 유효성을 검사하기 위해 파서를 사용하며 DOM기반의 파서는 W3C의 표준 권고안으로서 XML문서를 트리(tree)와 같이 표현할 수 있다. 표현된 트리에서 노드의 형태는 각각 엘리먼트(Element)와 애트리뷰트(Attribute) 그리고 텍스트(Text) 중에 한 가지이다[8][9].

그림 1과 그림 2는 예제 XML문서와 트리구조를 표현한 것이다.

```
<?xml version="1.0" ?>
<bookinfo xmlns="x-schema:book-schema.xml">
 <title type="computer">Data Structure</title>
 <price>18000</price>
 <isbn>89-88412</isbn>
 <date> July 1st 2006 </date>
 <writer>
  <name>Chang </name>
  <email>Ch@mail.net</email>
 </writer>
</bookinfo>
```

그림 1. 예제 XML 문서
Fig. 1. A sample XML document

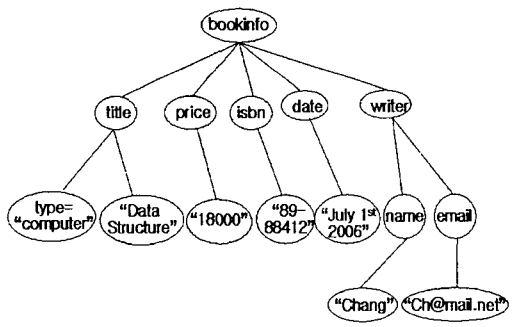

그림 2. 예제 XML 문서의 트리구조
Fig. 2. Tree structure of a sample XML document

트리와 같은 구조로 표현된 XML문서의 애트리뷰트와 엘리먼트등을 다루기 위해 XQL, XML-QL과 같은 몇 가지 경로-지향 질의 언어들이 제안 되었다. 다음의 XQL 질의는 그림 1에서 XML문서에 대한 간단한 경로-지향 질의에 대한 예이다.

```
/bookInfo/writer[name$contains$'Chang']
```

여기서 '/bookInfo/writer'는 경로이고, [name $contains$ 'Chang']는 술어이며, 엘리먼트 name이 단어 'Chang'를

포함하는지 아닌지 질의한다. 다음은 속성값을 포함하는 질의 예를 나타낸 것이다.

/bookinfo/title[@type='computer']

여기서 <bookinfo> 루트 엘리먼트에 포함되어 있는 모든 <title> 엘리먼트 중에서 type속성의 값이 'computer'인 <title>엘리먼트를 포함하는 아닌지를 질의한다.

Ⅲ. 경로서명 파일과 이진 트라이 구조

경로 서명 파일내의 각 경로 서명을 생성하기 위해 사용되는 서명 값은 1로 설정된 m개의 비트를 가진 길이 F인 해시코드화 된 비트패턴으로 이루어져 있으며 경로 서명 파일을 생성하기 위해 중첩기호법을 사용한다. 경로서명 파일은 부정확한 여과장치(inexact filter)의 개념을 기반으로 하고 있다. 따라서 그들은 많은 부적합한 값들을 버리는 빠른 테스트를 제공하지만 적합한 값들은 확실히 테스트를 통과한다. 그러나 몇몇의 값들은 실제로는 검색 요구 조건을 만족하지 않는다 할지라도 여과장치를 통과 한 경우도 있다. 이들을 허위드롭(False Drop)이라 한다. 서명을 생성하기 위한 방법으로는 전형적으로 중첩기호법(Superimposed ORing)을 사용한다. 어떤 경로지향 질의어가 도착했을 때 먼저 경로-지향 질의어내의 포함된 경로(Path)에 대한 경로서명을 변환한다. 그런 다음 이 경로서명을 포함하는 모든 경로서명 파일에 저장된 엘리먼트의 경로들을 검색하고 이 과정에서 많은 부적합한 엘리먼트들이 버려진다[2][7].

이진 트라이는 외부노드인 정보노드(information node)와 중간노드인 분기노드(branch node)로 구성된다. 모든 정보는 외부노드에만 저장되며 분기노드는 정보는 없고 링크만을 가진다. 다른 검색 트리들이 입력되는 노드의 순서에 따라 모양이 변하는 것과 달리 트라이는 입력순서에 관계없이 항상 같은 모양을 나타낸다[10][11].

본 논문에서는 정보검색을 위한 인덱싱 기법으로 이진 트라이와 경로서명 파일을 결합하는 방식을 사용한다.

본 논문의 예제 XML문서에 대한 경로서명 파일의 예는 표 1과 같다.

표 1. 엘리먼트 경로서명 파일의 예
Table 1. An example of path signature file for elements

위치	Element-Path	Path signature(F=7)
1	bookinfo	0101000
2	bookinfo-title	0101100
3	bookInfo-price	0111000
4	bookInfo-isbn	1101000
5	bookInfo-date	0111010
6	bookInfo-writer	0101010
7	bookInfo-writer-name	0111011
8	bookInfo-writer-email	0101111

본 논문에서는 경로-지향 질의를 처리하기 위한 인덱싱 구조로 이진 트라이 구조를 이용한다. 그림 3은 표 1의 경로서명 파일에 대한 이진 트라이 구조를 표현한 것이다.

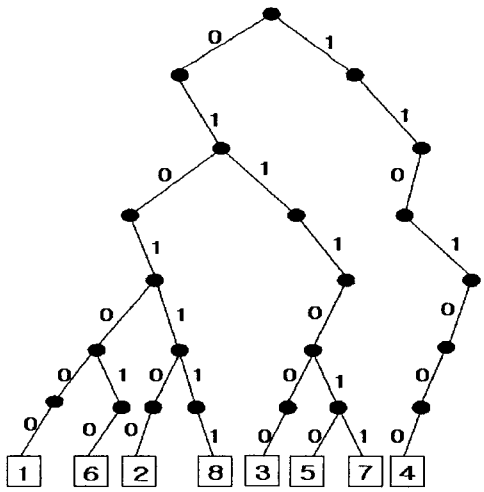

그림 3. 경로서명 파일에 대한 이진 트라이 인덱싱 구조
Fig. 3 Binary trie indexing structure for path signature file

그림 3에서 인덱싱 구조의 동작은 어떤 경로-지향 질의어의 경로서명값에서 각 비트값이 0이라면 왼쪽과 오른쪽 자식 노드 모두를 따라 이동하고 각 비트값이 1이라면 오른쪽 자식 노드를 따라 이동한다.

만약 어떤 경로-지향 질의어의 경로서명값이 "0111010" 이라면 질의결과는 경로서명 파일에서 주소 5와 주소 7에 해당하는 경로서명값이 검색결과로 반환된다.

질의처리기는 다음과 같은 질의를 생성한다.

SELECT * FROM ElementTable e

WHERE e.단축경로 IN 결과로 반환된 경로-서명;

만약 질의가 /bookInfo/writer[name$contains$'Chang']의 형태로 입력된다면 질의처리기는 다음과 같은 질의를 생성한다.

SELECT * FROM ElementTable e TextTable t
WHERE e.엘리먼트명='name'
AND e.단축경로 IN 결과로 반환된 경로서명
AND e.문서ID=t.문서ID
AND e.엘리먼트ID=t.부모ID
AND t.텍스트값그'Chang';

Ⅳ. 경로-지향 질의처리 시스템의 설계 및 구현

그림 4는 본 논문에서 제안한 경로-지향 질의처리 시스템의 논리적인 구성도이다. 전체 시스템 구성은 사용자, 질의처리시스템, 관계형 데이터베이스로 이루어진 3-tier구조이다.

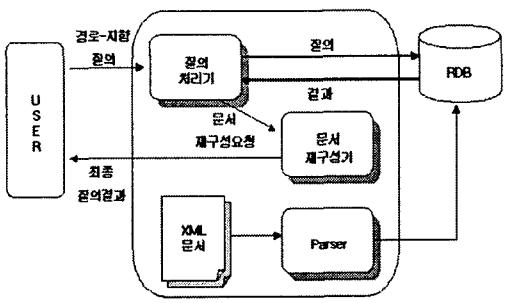

그림 4. 질의처리 시스템의 논리적 구성도
Fig. 4. Logical architecture of query processing system

본 논문의 경로-지향 질의처리 시스템은 파서와 질의처리기 그리고 문서재구성기로 구성된다.

모든 XML문서는 경로-지향 질의를 수행하기 전에 파서에 의해 파싱되어 그 결과는 데이터베이스 내에 미리 저장되어 있어야 한다. 질의처리기(Query Processor)는 사용자로부터 입력된 경로-지향 질의의 경로서명 값을 생성하고 사용 가능한 표준 관계형 질의 형태로 변환하여 데이터베이스에 저장되어 있는 XML 문서에 대한 질의를 수행한다.

문서 재구성기(Document Reorganizer)는 질의처리기에 의해 수행된 질의 결과 값이 존재하는 경우 해당 XML문서에 대한 모든 정보를 DB로부터 추출한 다음 이들을 원래의 XML 문서 형태로 재구성하여 사용자에게 경로-지향 질의처리 결과로 되돌려준다. 또한 사용자가 원하는 경우 사용자측의 로컬 컴퓨터에 XML파일로 저장하는 기능을 제공하기도 한다.

그림 5는 XML문서가 파서에 의해 파싱되어 관계형 데이터베이스에 저장된 형태를 보여준 것이다. 엘리먼트 테이블, 애트뷰트 테이블, 텍스트 테이블로 구성된다. 엘리먼트 테이블의 구조는 다음과 같다.

{DocID:<integer>,EID:<integer>,Ename:<string>,ParentID:<integer>,Path_Sig:<string>,Depth:<integer>}

표 2. 엘리먼트 테이블
Table 2. Element table

DocID	EID	Ename	ParentID	Path_Sig	Depth
1	1	bookinfo	*	0101000	1
1	2	title	1	0101100	2
1	3	price	1	0111000	2
1	4	isbn	1	1101000	2
1	5	date	1	0111010	2
1	6	writer	1	0101010	2
1	7	name	6	0111011	3
1	8	email	6	0101111	3

텍스트(Text) 테이블은 좀 더 단순한 구조를 가지며 다음과 같다.

{DocID:<integer>, ParentID:<integer>,TVal:<string>}

표 3. 텍스트 테이블
Table 3. Text table

DocID	ParentID	TVal
1	2	Data Structure
1	3	18000
1	4	89-88412
1	5	July 1st 2006
1	7	Chang
1	8	Ch@mail.net

애트리뷰트(Attribute) 테이블은 다음과 같은 구조를 가진다.

{DocID:<integer>,ParentID:<integer>,Att_name:<string>,Att_value:<string>}

표 4. 애트리뷰트 테이블
Table 4. Attribute table

DocID	ParentID	Att_Name	Att_Value
1	2	type	"computer"

본 논문에서 제안한 경로-지향 질의처리 시스템은 Windows 2000® 운영체제를 탑재한 Intel® Pentium® 4 시스템 상에서 구현 및 실험이 이루어졌다. 시스템의 사양은 메모리 1GB, CPU 속도 1.7GHz, HDD 80GB로 구성되어 있으며, 데이터베이스 시스템으로 Oracle® 9i를 사용하여 수행하였다. 시스템의 구현을 위해 JAVA언어를 사용 하였으며, 실험에 사용된 데이터는 표 5와 같다.

표 5. 실험에 사용된 데이터들
Table 5. Simulation Data

경로서명의개수	해시 코드 길이(F)	1로 설정된 비트수(m)	평균 엘리먼트 수 /문서	XML문서 파일의 수
1,000	10	5	53	19
10,000	14	7	210	48
100,000	17	8	300	334
500,000	19	9	400	1,250
1,000,000	20	10	500	2,000
사용된 경로-지향 질의어의 유형		/bookinfo/writer/name		
		/bookinfo/title[@type='computer']		
		/bookinfo/writer[name$contains$'Chang']		

그림 5는 경로-지향 질의처리 시스템의 초기화면을 나타낸 것이다.

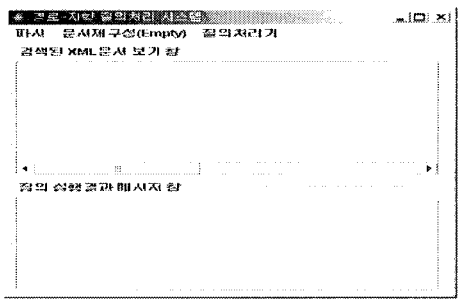

그림 5. 경로-지향 질의처리 시스템의 초기화면
Fig. 5. Initial screen of query processing system

그림 6은 파서에 의한 XML문서의 파싱 및 파싱 결과가 데이터베이스에 저장되는 화면이다.

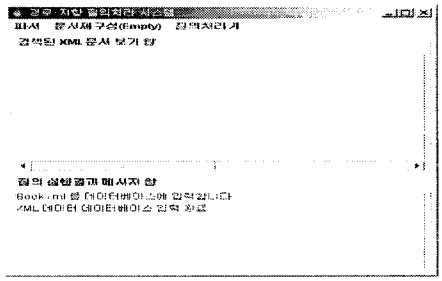

그림 6. 파서에 의한 XML문서 파싱
Fig. 6. An XML document parsing by parser

그림 7과 그림 8은 질의처리기에 의해 술어를 포함하는 경로-지향 질의어 입력과 처리 결과를 나타낸 화면을 보여 준 것이다.

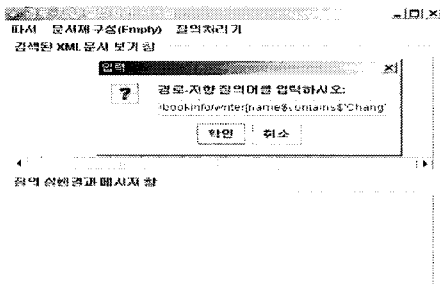

그림 7. 경로-지향 질의 입력화면
Fig. 7. Input screen of Path-oriented query

그림 8은 검색결과 질의조건을 만족하는 문서가 3개이며 그중 문서번호(DocID)가 1에 해당하는 문서를 문서재구성기를 이용하여 재구성하여 보여준 것이다.

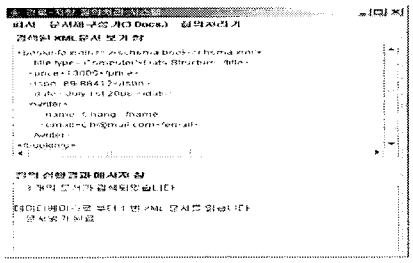

그림 8. 질의처리 결과화면
Fig. 8. Result screen of query processing

그림 9는 검색된 XML문서들 중에서 문서재구성기를 이용하여 검색된 문서들 중에서 사용자가 원하는 문서를 사용자의 로컬 컴퓨터에 저장하는 화면이다.

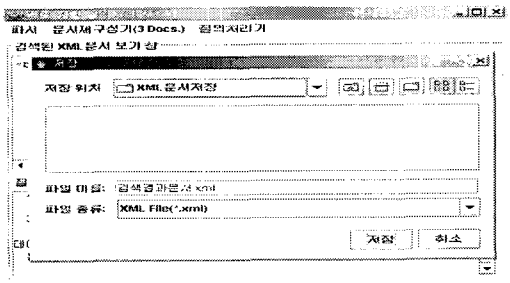

그림 9. 검색된 XML문서를 로컬 컴퓨터에 저장하는 화면
Fig. 9. Saving screen of searched XML document on local computer

그림 10은 검색된 문서들 중에서 사용자측의 로컬 컴퓨터에 저장된 XML문서를 브라우저를 통해 보여준 것이다.

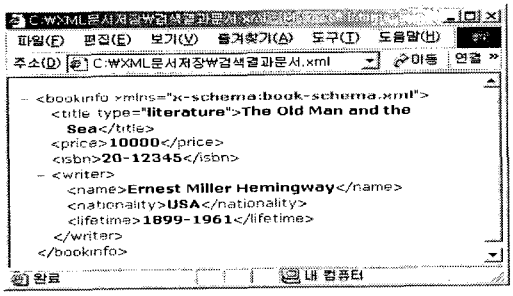

그림 10. 로컬 컴퓨터에 저장된 XML문서
Fig. 10. Saved XML document on local computer

기존의 논문은 인덱싱 성능개선에만 중점을 두고 있는 반면에 본 논문은 기존의 기법과 동등한 인덱싱 성능을 유지함은 물론 사용자에게 편리하고 쉬운 XML문서 검색 인터페이스를 함께 제공하고 있다.

V. 결 론

본 논문에서는 더욱 그 활용범위가 확대되고 있는 XML문서에 대한 경로-지향 질의어를 효율적으로 처리하기 위한 경로-지향 질의처리 시스템을 제안하고 이에 대한 설계 및 구현을 하였다. 또한 관계형 데이터베이스에 저장된 XML문서를 대상으로 하는 경로-지향 질의어의 평가를 효율적으로 수행하기 위해 경로서명 파일과 이진 트라이구조를 결합한 인덱싱 구조를 제안하였다. 시스템의 구현 결과 사용자에게는 경로-지향 질의를 이용한 보다 편리한 XML문서 검색 기능 제공이 가능하였다.

참고문헌

[1] Y. Chen and G. Huck, "Path signature: A Way to Speed up Evaluation of Path-oriented Queries in Document Databases", WISE2000, pp. 240-244, 2000.

[2] W. B. Frakes and R. Baeza-Yates, "Information Retrieval:Data Structures and Algorithms", Prentice Hall PTR; Facsimile edition, 1992

[3] W3C, "XML Path Language(XPath) 2.0", http://www.w3.org/2003/08/DIFF-xpath20, 2003.

[4] W3C, "Extensible Markup Language (XML)", http://www.w3.org/XML/, 1998.

[5] W3C, "XML Query (XQuery) Requirements", http://www.w3.org/TR/2003/WD-xquery-requir-ements-20031112, 2003.

[6] A. Deutch and M. Fernandez and D.Forescu and A. Levy and D. Suciu, "XML-QL : A Query Languagefor XML", http://www.w3.org/TR/NOTE-xml-ql, Aug. 1998.

[7] D. Eastlake and J. Reagle and D. Solo and W3C Recommendation, "XML-Signature Syntax and Processing", http://www.w3.org/TR/2002/REC-xmldsig-core- 20020212, Feb. 2002.

[8] W3C, "Document Object Model (DOM)", http://www.w3.org/DOM/, 2002.

[9] H. M. Deitel and P. J. Deitel and T. R. Nieto and T. Lin, P. Sadhu, "XML How TO PROGRAM", Prentice Hall, 2000.

[10] 박희숙, 조우현, "단축-경로와 확장성 해싱 기법을 이용한 경로-지향 질의의 평가속도 개선 방법", 정보처리학회논문지, 제11-D권, 제7호, pp. 1409-1416, 2004.

[11] 박희숙, 조우현, "XML 데이터베이스에서 경로-지향 질의처리를 위한 병렬 매치 방법", 정보과학회논문지, 데이터베이스 32권, 제5호, pp. 558-566, 2005.

저자소개

박 희 숙(Hee-Sook Park)

1995년 한국방송대학교전자계산학과(이학사)
1998년 경남대학교 교육대학원 전자계산교육전공(교육학석사)
2006년 부경대학교 대학원 컴퓨터공학과(공학박사)
※관심분야 : 객체지향 데이터베이스, 공간 데이터베이스, 데이터베이스 인덱싱 성능개선 문제

박 주 현(Ju-Hyun Park)

2000년 경남대학교 컴퓨터공학과 (공학사)
2002년 경남대학교 대학원 컴퓨터공학과 (공학석사)
2007년 부경대학교 대학원 컴퓨터공학과(박사수료)
※관심분야 : 공간 데이터베이스, 객체지향 데이터베이스

조 우 현(Woo-Hyun Cho)

1985년 경북대학교 전자공학과 전산공학전공 (공학사)
1988년 경북대학교 대학원 전자공학과 전산공학전공(공학석사)
1998년 경북대학교 대학원 전자공학과 전산공학전공 (공학박사)
1989년-현재 부경대학교 공과대학 전자컴퓨터정보통신공학부 교수
※관심분야 : 지식의 표현과 병렬처리, 멀티미디어 데이터베이스관리시스템, 객체지향데이터베이스