韓國컴퓨터情報學會 論文誌
第14卷 第2號, 2009. 2.

2009-14-2-4-3

OWL 온톨로지를 기반으로 하는 논문 검색 시스템에 관한 연구

선 복 근*, 위 다 현**, 한 광 록***

A Study on Paper Retrieval System based on OWL Ontology

Bok-Keun Sun*, Da-Hyun We**, Kwang-Rok Han***

요 약

기존의 논문 검색은 키워드 기반 검색이고, 발간된 자료의 양이 방대해 짐에 따라 사용자가 원하는 정보를 검색하는데 어려움이 가중되고 있다. 사용자의 의도에 맞는 정보를 검색하기 위해서는 인터넷 환경에서 웹 문서 자원 사이의 의미 정보를 온톨로지로 표현하고, 이 온톨로지를 컴퓨터가 이해할 수 있게 하는 시맨틱 웹의 도입이 필요하다.

따라서 본 논문에서는 OWL 온톨로지 기반의 추론을 통한 논문 정보 검색시스템에 대하여 논한다. 시맨틱 웹의 새로운 온톨로지 언어로 부상한 OWL 기반의 논문 온톨로지를 구축하고, 논문 속성들 간의 다양한 상관관계를 서술논리 쿼리로 작성한다. 검색시스템은 이 쿼리를 기반으로 논문 온톨로지에 대하여 추론함으로써 지능적인 정보검색이 가능하도록 하였다. 마지막으로 기존 논문 검색 방법과 본 논문의 실험 결과를 비교하였다.

Abstract

The conventional paper retrieval is the keyword-based search and as a huge amount of data be published, this search becomes more difficult in retrieving information that user want to find. In order to search for information to the user's intent, we need to introduce semantic Web that represents semantics of Web document resources on the Internet environment as ontology and enables the computer to understand this ontology.

Therefore, we describe a paper retrieval system through OWL(Ontology Web Language) ontology-based reason in this paper. We build the paper ontology based on OWL which is new popular ontology language for semantic Web and represent the correlation among diverse paper properties as the DL(description logic) query, and then this system infers the correct results from the paper ontology by using the DL query and makes it possible to retrieve information intelligently. Finally, we compared our experimental result with the conventional retrieval.

▸ Keyword : 온톨로지(Ontology), 웹 온톨로지언어(OWL(Ontology Web Language), 서술논리(Description Logic), 추론(Reason), 논문검색(Paper Retrieval)

• 제1저자 : 선복근
• 투고일 : 2008. 12. 1, 심사일 : 2009. 1. 6, 게재확정일 : 2009. 2. 23.
* 호서대학교 공학교육혁신센터 전임강사 ** 호서대학교 메카트로닉스공학과 석사과정
*** 호서대학교 메카트로닉스공학과 교수
※ 이 논문은 2008년도 호서대학교의 재원으로 학술연구비 지원을 받아 수행된 연구임"(과제20080135)

Ⅰ. 서 론

하이퍼텍스트에 기반을 둔 기존의 웹은 현재 수많은 기관과 커뮤니티와 개인들이 서로 다른 목적으로 생성한 방대한 양의 정보가 기존의 검색 방식으로는 감당할 수 없을 정도가 되어 불편함을 초래하고 있다. 기존의 웹 검색은 단어의 빈도수와 어휘정보를 이용하여 문서의 유사도를 측정하고 순위를 매기기 때문에 사용자의 질의와 관계없는 많은 문서를 결과로 가져올 수 있고, 유의어, 동음이의어 등의 문제를 해결하지 못하며, 상업 및 학술정보 같은 정보의 특징을 분별해낼 수 없다. 이로 인해 사용자는 불필요한 정보를 걸러내느라 시간을 낭비해야 한다. 또한, 지금까지의 웹 기술은 사람과 컴퓨터간의 정보교환에만 치중하여 컴퓨터와 웹 데이터간의 의미적 해석 및 처리의 문제를 가지고 있다. 시맨틱 웹(SemanticWeb)은 기존의 웹을 확장한 형태로, 메타데이터의 개념을 적용해 정보에 정형화된 의미를 부여하여 소프트웨어 에이전트가 이 의미정보를 자동으로 추출할 수 있는 환경을 제공함으로써 정보의 확장 및 공유를 가능하게 한다. 온톨로지(ontology)는 사람과 컴퓨터 간의 공유되는 지식을 개념적으로 표현 한 것으로써, 지식의 개념들간의 구조와 여타 관계 및 그들의 제약을 표현하는 일종의 개념형 데이터베이스이다. 특히, 시맨틱 웹의 온톨로지는 기존의 온톨로지를 웹을 통해 확장시키는 방법을 제시하고 있다[1-5].

기존의 검색 엔진에서 키워드나 제목으로 검색할 경우, 학술정보를 이외의 분야에서도 검색결과를 출력한다. 학술정보 전용 검색엔진을 이용할 경우에는 해당 데이터베이스에 존재할 경우에만 원하는 정보를 출력하게 되며 여러 분야에 해당하는 학술정보를 출력하기 때문에 사용자의 2차 검증이 필수적이다. 상세검색을 통해 학술분야 혹은 학회나 학회지에 대한 논문 정보 통해 검색하더라도 저마다의 카테고리로 분류되어있어 통합적인 검색이 이루어지지 않는 실정이다. 이와 같은 단점을 극복하기 위하여 본 논문에서는 논문에 대한 OWL(Ontology Web Lanuage) 온톨로지를 설계하였다. 이를 기반으로 웹에 산재한 논문들에 대한 요약정보를 XML 형태의 논문 주석 정보 데이터로 정의하여, 사용자 질의에 대한 OWL 온톨로지를 생성하여 추론을 통하여 논문 정보를 검색 가능한 시스템을 제안한다. 사용자로부터 질의를 입력받았을 때, 구축된 OWL 온토로지와 추론 기법을 적용하여, 사용자가 찾고자 하는 검색 결과를 보다 정확하게 제공하고자 한다.

Ⅱ. 관련 연구

시맨틱 웹은 웹의 창시자인 팀 버너스리에 의해 1998년에 제안되었으며 각종 회의와 연구를 통해 지속적으로 관련규격과 기술이 개발되고 있다[5,6].

시맨틱 웹은 현재의 인터넷과 같은 분산 환경에서 웹 문서, 각종 파일, 서비스 등의 자원에 대한 정보와 자원 사이의 관계인 의미 정보(Semantics)를 컴퓨터가 처리할 수 있는 온톨로지 형태로 표현하고, 이를 자동화된 컴퓨터가 처리하도록 하는 프레임워크이자 기술이다. 이를 위해선 컴퓨터가 이해할 수 있는 언어로 웹을 구성해야 한다[7-12]. RDF(Resource Development Framework), 온톨로지, OWL 등은 이를 위해 논의되고 있는 기술이다. 현재 시맨틱 웹은 W3C(World Wide Web Consortium)를 중심으로 메타데이터(meta data)를 통해 정보의 의미를 이해하고 처리하도록 하는 자원 설명 기술과 지식 설명 기술이 결합된 연구 방향이 한 줄기를 이루고 있다. W3C는 RDF 기반의 온톨로지 기술에 관심을 가지고 있다[13-17].

다른 한 연구는 ISO를 중심으로 XTM언어를 이용한 정보와 지식의 분산 처리, 통합 관리 등에 중점을 두는 방향이다. ISO쪽은 토픽 맵(Topic Maps) 기술에 중점을 두고 있다. 토픽맵은 ISO/IEC 13250 표준으로 XML 기반의 XTM(XML Topic Maps)이라는 언어를 사용해 정보와 지식의 분산 관리를 지원한다. 이 중에서 현재 가장 활발하게 논의되고 있고 실제로 사용하고 있는 기술은 메타데이터를 이용한 자동화 처리 부분이다[18].

시맨틱 웹의 온톨로지 표준화는 W3C를 중심으로 진행되고 있다. W3C는 2001년 웹 온톨로지 워킹그룹을 구성하여 표준화 및 기술 개발을 진행하고 있으며, OWL 개발도 진행중이다. 웹 온톨로지 워킹 그룹은 2002년에 OWL 초안 Version 1.0을 발표했는데 필요사항만 언급하고 구체적 표준은 제시하지 않았다. 다만 온톨로지를 위해서 메타 데이터가 필수 요소라는 것은 분명하게 밝혔다.

OWL은 시맨틱 웹의 온톨로지 구축에 사용하기 위해 만든 웹언어이고 이 OWL이 개발되는 이유는 초기의 마크업 언어인 CycL(CYCR Language), KIF(Knowledge Interchange Format), Ontolingua 등이 온톨로지 표현에 적합하지 않기 때문이다. 대표적인 OWL로 OIL(Ontology Inference Layer), DAML(DARPA Agent Markup Language), SHOE(Simple HTML Ontology Extensions), N3(Notation 3) 등이 있다. OIL과 DAML은 RDF와 RDF

스키마를 기반으로 연구가 진행되고 있는 마크업 언어다. 현재 유럽 연합의 IST(Information Society Technologies)를 중심으로 연구 진행 중에 있으며, OIL에 대한 백서를 발표한 상태이다. DAML은 DARPA의 지원을 받아 연구한 언어로 심부름꾼 프로그램 사이의 통신에는 유용하나 초기에는 온톨로지를 표현하지 못하는 문제점이 있었다. DAML은 온톨로지를 표현하기 위해 OIL을 포함한 DAML+OIL로 연구가 진행되고 있다. SHOE는 메릴랜드 대학을 중심으로 연구가 되고 있는 언어로, HTML에서 온톨로지 표현이 가능한 언어다. N3는 사용자 중심으로 설계된 언어로 RDF와 함께 서로 보완 역할을 수행하도록 만들었다.

시맨틱 웹의 미래를 한 번에 그리기란 쉽지 않다. 따라서 이제 몇 년에 불과한 시맨틱 웹에 대한 논의는 지금도 계속 진행되고 있고, 2005년 3월에 시맨틱 기술 컨퍼런스가 열려 좀 더 발전된 논의가 이루어졌지만 여전히 시맨틱 웹은 진행형이며 앞으로도 당분간 계속 연구되고 논의될 것이다. 시맨틱 웹은 워낙 광범위한 분야이기 때문에 한 번에 그 모습을 드러내지 않고 분야 별로 조금씩 모습을 드러내고 있다.

국회 전자 도서관은 일반 검색, 상세 검색, 고급 검색의 3가지 서비스를 제공한다. 사용자가 자료를 구분하고 항목을 선택하는 등 효율적 검색이 이루어질 수 있으나 단순한 키워드 기반 검색이므로 사용자의 의도와는 전혀 무관한 정보를 제공해 주는 경우가 발생한다. KSI 학술 논문 정보 시스템은 사용자의 편의를 위해 단어 기반 검색 서비스와 분류 시스템을 제공하지만 의미 기반이 아닌 검색 범위를 제한하는 방식을 제공하고 있기 때문에 사용자가 원하는 결과를 찾기 위해 여러 번 검색을 시도해야 한다. 기존 학술 자료 검색 방법을 보완하기 위해 시맨틱 웹 기술이 도입하여 논문 관련 정보를 온톨로지로 구축하는 검색 모델이 제안되고 있다[18].

본 논문은 온톨로지 기반 논문 정보 검색 시스템 설계에 대해 기술한 논문을 기반으로 시스템을 구축하고 그 효율성을 검증했다[19].

III. OWL 기반의 논문 검색 시스템

3.1 논문 검색용 온톨로지의 설계

3.1.1 온톨로지 클래스

그림 1은 웹온톨로지 기반의 논문 정보 검색 시스템 내에서 사용될 정보들을 표현하기 위해서 구축할 온톨로지의 전반적인 구조를 보여준다. 온톨로지는 계층적 구조를 가지고 있으며, 실선은 하위 클래스 개념을 표현하고 있다. 점선은 클래스간의 관계를 표현한 것으로 요소에 해당하는 객체 속성(Object Properties)을 표현하고 있다. 'Paper'는 최상위 클래스로서 'Domain' 클래스와 논문 정보를 묘사하는

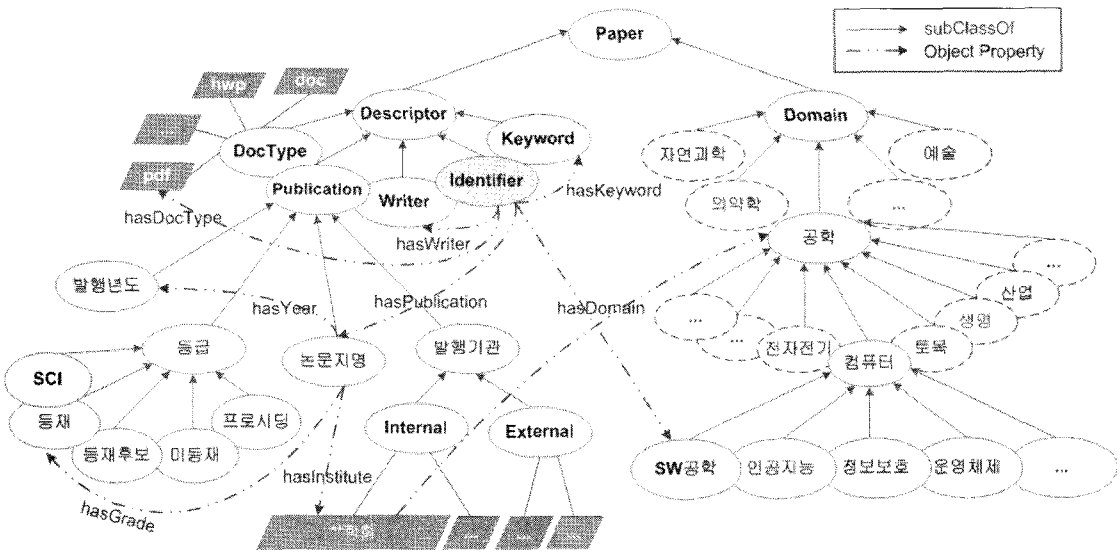


그림 1. 계층적인 온톨로지 구조
Fig. 1. Structure of hierarchical ontology

'Descriptor' 클래스를 하위 클래스로 갖는다. 'Publication' 클래스는 논문지의 속성을 상세히 기술함으로써 발행된 논문지의 등급과 국내외 학회정보를 이용한 검색이 가능하도록 설계하였다. 최상위 계층인 'Paper' 클래스의 하위 클래스인 'Domain'과 'Descriptor'에 대한 상세한 기술을 표 1에 나타내었다.

표 1. Paper Descriptor 개념 온톨로지
Table 1. Conceptual ontology of Paper Descriptor

클래스			상세
Domain			- 학술 분야별 분류 - 학회, 키워드, 논문 분류에 사용 - 학진 연구분야 분류표 참조
Descriptor	Identifier		- 논문 고유 번호 - 타이틀을 대신하는 DB key
	DocType		- 논문데이터의 파일 종류 - PDF, DOC, HWP, PPT 4개의 속성으로 분류
	Keyword		- 논문 검색 키워드
	Writer		- 저자
	Publication	Name	- 논문지명
		Year	- 발행년도
		Grade	- 논문지 등급 - 학진 논문 등재지 리스트 참조 - SCI, SCIE, Record(등재) UnRecord(미등재) Candidate(등재후보) Proceeding
		Publisher	- 발행학회
		Lang-uage	- 국내, 국외 논문 구분 - External, Internal 두 개의 하위클래스로 구성

3.1.2 객체 속성 정의

OWL 온톨로지의 객체 속성은 클래스들 사이의 관계를 기술하기 위한 것으로 본 논문의 OWL 온톨로지에서 정의한 객체 속성들은 표 2와 같다.

그림 2는 그림 1의 온톨로지 클래스 중에서 Identifier와 관계있는 클래스와 객체 속성들의 일부를 나타낸다. 이 Identifier 클래스는 논문 온톨로지의 핵심이 되는 클래스이며 hasKeyword 등의 객체 속성에 의해 다른 클래스들과 관계를 맺는다.

3.2. 논문정보 검색 시스템

3.2.1 데이터 구성

3.2.1.1 원본 논문 정보 데이터

웹에 산재된 기존의 논문 데이터에 대한 요약 정보를 원본 논문 정보 데이터라고 하며, HTML 형식의 데이터이고, 본 논문에서는 현재 학술 데이터베이스 검색 사이트에서 검색한 논문에 대한 요약 정보를 이용한다.

표 2. 객체 속성의 정의
Table 2. Definition of object properties

객체속성	관계 기술
hasDomain	Identifier 클래스는 Domain 클래스의 구성원들 중 한 개 이상을 구성원으로 갖는다.
hasKeyword	Identifier 클래스는 Keyword 클래스의 구성원들을 구성원으로 갖는다.
hasWriter	Identifier 클래스는 Writer 클래스의 구성원들을 구성원으로 갖는다.
hasPublication	Identifier 클래스는 Name 클래스의 구성원들 중 한 개만 구성원으로 갖는다.
hasDocType	Identifier클래스는 DocType 클래스의 구성원들을 구성원으로 갖는다.
hasGrade	Name클래스는 Grade 클래스의 구성원들을 구성원으로 갖는다.
hasPublisher	Name클래스는 Publisher 클래스의 구성원들 중 한 개만 구성원으로 갖는다.

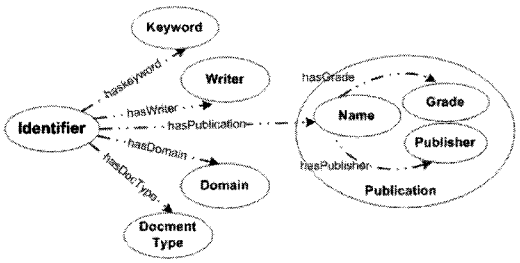


그림 2. Paper Descriptor 온톨로지의 객체속성
Fig. 2. Object properties of Paper Descriptor ontology

이 요약 정보에는 일반적으로 사용자들이 논문을 검색할 때 필요한 모든 정보를 포함하고 있기 때문에 논문 온톨로지를 구성하는데 필요한 요소를 모두 얻을 수 있다. 그림 3은 원본 논문 정보의 요약 정보에 대한 HTML 데이터를 나타낸다.

3.2.1.2 논문 주석 정보 데이터

HTML 형식의 논문 용약 정보 데이터로부터 개념 온톨로

지를 참조하여 XML 형식의 메타 데이터로 변환한 것을 논문 주석 정보 데이터라고 한다.

그림 4는 그림 3의 HTML 형식의 원본 논문의 요약 정보를 XML 형식의 논문 주석 정보 데이터로 변환한 것이다. 이 논문 주석 정보 데이터의 XML 태그는 더블린코어를 기본으로 시스템에서 필요한 태그들을 추가해 정의하였다.

```
<table>
  <tr>
    <td>제목</td>
    <td>데이터베이스 시스템에 기반한 효율적인 OWL
저장시스템 설계 및 성능분석</td>
  </tr><tr>
    <td>제목(영문)</td>
    <td>The Design and Performance Analysis of an Effective
OWL Storage System Based on the DBMS</td>
  </tr><tr>
    <td>저자</td>
    <td>조성환 ( Seong Hwan Cho ) , 김성식 ( Seong Sik
Kim ) , 김태영 ( Tae Young Kim )</td>
  </tr><tr>
    <td>발행년도</td>
    <td>2008</td>
  </tr><tr>
    <td>발행기관</td>
    <td>한국컴퓨터교육학회</td>
  </tr><tr>
    <td>발행정보</td>
    <td>컴퓨터교육학회논문지, Vol.11, No.5,
&amp;nbsp;&amp;nbsp;Startpage 77, Endpage 88, Totalpage 12</td>
  </tr><tr>
    <td>주제키워드</td>
    <td>온톨로지, 시멘틱 웹, 추론, OWL, Ontology, Semantic
Web, Inference</td>
  </tr>
</table>
```

그림 3. 원본 논문 정보 데이터 (HTML)
Fig. 3. Information data of resource paper(HTML)

```
<?xml version="1.0" encoding="utf-8"?>
<annotation>
  <identifier>Comp_1</identifier>
  <type>컴퓨터공학</type>
  <format>pdf</format>
  <source>http:///abcfg</source>
  <title>
    <title_kor>데이터베이스 시스템에 기반한 효율적인 OWL
저장시스템 설계 및 성능분석</title_kor>
    <title_eng>The Design and Performance Analysis of an
Effective OWL Storage System Based on the
DBMS</title_eng>
  </title>
  <subject>
    <keyword>온톨로지</keyword>
    <keyword>시멘틱 웹</keyword>
    <keyword>추론</keyword>
    <keyword>OWL</keyword>
    <keyword>Ontology</keyword>
    <keyword>Semantic Web</keyword>
    <keyword>Inference</keyword>
  </subject>
  <creator>
    <writer>조성환</writer>
    <writer> 김성식</writer>
    <writer>김태영</writer>
  </creator>
  <contributor>
    <name>컴퓨터교육학회논문지</name>
    <vol>11</vol>
    <no>5</no>
    <publisher> 한국컴퓨터교육학회</publisher>
    <date>2008</date>
    <grade>Record</grade>
    <language>KOR</language>
  </contributor>
  <description>생략</description>
</annotation>
```

그림 4. 논문 주석 정보 데이터(XML)
Fig. 4. Information data of paper annotation(XML)

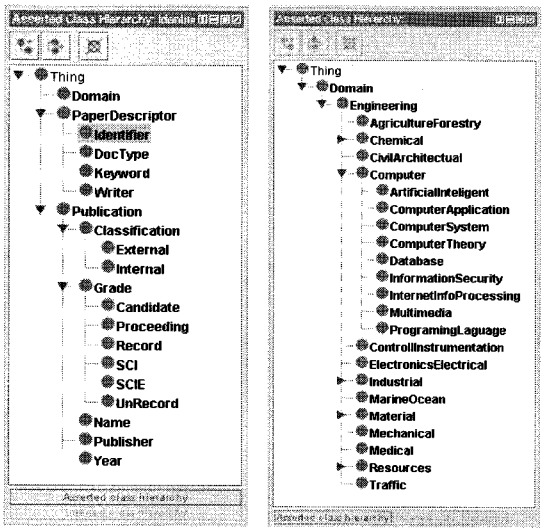
3.2.1.3 개념 온톨로지

개념 온톨로지는 본 논문에서 제안한 시스템에서 사용하는 데이터를 일정한 패턴으로 정형화하기 위한 기본 정보를 포함하고 있는 OWL 온톨로지이다.

논문 주석 정보 데이터들로부터 OWL 온톨로지를 구축하기 위하여 논문의 특정 주제가 가질 수 있는 클래스들과 클래스들 사이의 관계인 객체 속성을 기술하여 논문 온톨로지의 서술논리(Description Logic)를 제공하기 위한 것이다. 그림 5는 Protege4.0을 이용하여 작성한 개념 온톨로지의 클래스 계층을 나타낸다. 개념 온톨로지는 논문 정보에 대한 클래스와 객체 속성을 정의하고 클래스들의 서술논리를 기술한 Paper Descriptor 온톨로지(그림 5.a)와 학술분야를 분류한 Domain 온톨로지(그림 5.b)로 나누어 작성 한다. 이 개념 온톨로지는 논문 주석 정보 데이터를 작성할 때와 논문 온톨로지를 작성할 때에 참조된다.

3.2.1.4 URI 데이터베이스

URI 데이터베이스는 추론엔진에서 검색한 결과로 반환하는 Identifier 클래스의 인스턴스들을 키로 하여 웹 상에 존재하는 실제 논문의 링크 URI을 저장하기 위하여 사용된다. 원본 논문 정보 정보 데이터를 가공하면서 논문의 고유한 Identifier를 부여하고, 키워드와 제목 그리고 논문의 URI를 기록한다.



a. Paper descriptor b. Domain
그림 5. 개념 온톨로지의 클래스 계층
Fig. 5. Class hierarchy of conceptual ontology

3.2.2 시스템 구조

그림 6은 본 논문에서 제안한 시스템의 구조를 나타낸다. 시스템은 크게 웹브라우저를 이용한 사용자 인터페이스와 전처리기, 사용자 질의어를 이용하여 논문 온톨로지를 생성하고 추론하는 등 검색의 중심이 되는 검색기 등으로 구성한다.

3.2.2.1 사용자 인터페이스

사용자는 웹 브라우저를 이용하여 논문 검색 시스템을 이용할 수 있다. 내부적으로는 논문의 키워드를 비롯하여, 논문의 분류 및 등급, 국내외 발행 등 다양한 조건을 입력 받고, 검색 결과를 확인할 수 있도록 설계하였다. 사용자 인터페이스를 통해 입력받은 데이터를 검색기의 질의 관리기로 전달하는 일과 추론 엔진으로부터 추론 결과를 전달받아 웹 상에서 원본 논문을 다운로드할 수 있다.

3.2.2.2 전처리기

웹에 산재한 HTML 형식의 논문 요약 정보인 원본 논문 정보 데이터를 수집하여 저장하고, OWL 온톨로지 생성을 위한 메타 데이터가 되는 XML 형식의 논문 주석 정보 데이터를 생성하고 저장한다.

3.2.2.3 검색기

① 질의어 관리기

사용자 인터페이스로부터 전달받은 질의어들을 관리한다.

- 질의어들 중에서 제 1 키워드만을 추출하여 키워드 검색을 함으로써 1차로 이 키워드와 관련된 논문들의 주석 정보 데이터를 검색해낸다.
- DL(Description Logic) 쿼리를 생성하기 위하여 제 1 키워드를 제외한 나머지 질의어들을 추출하여 DL 쿼리 생성기에 전달한다.

② OWL 온톨로지 생성기

개념 온톨로지를 참조하여 키워드 검색에 의해 1차로 검색된 논문들의 주석 정보 데이터들에 대응하는 OWL 온톨로지를 생성한다.

③ DL 쿼리 생성기

질의어 관리기에서 전달받은 질의어들 중에서 제 1 키워드

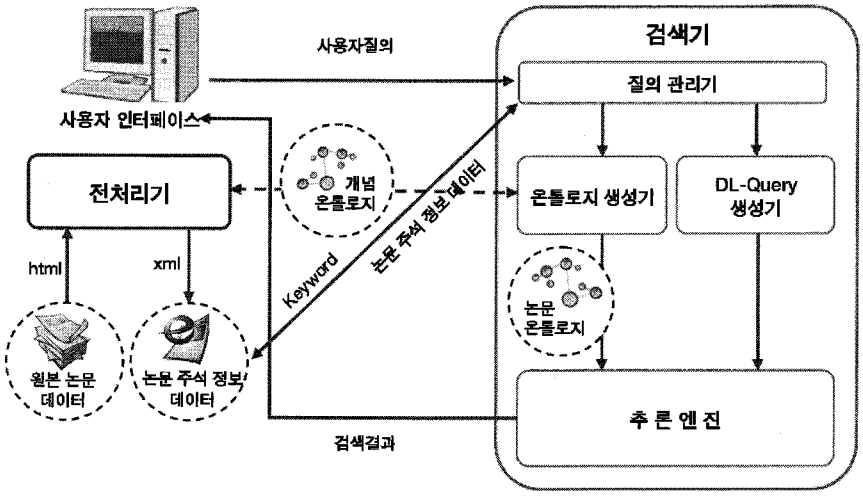


그림 6. 논문 검색 시스템 구조
Fig. 6. Structure of paper retrieval system

를 제외한 나머지 질의어들을 대상으로 추론을 위한 DL 쿼리를 생성하여 추론 엔진에 전달한다.

④ 추론엔진

온톨로지 생성기에서 작성된 논문의 OWL 온톨로지를 로드하고, 생성된 DL 쿼리를 기반으로 OWL 온톨로지에 대한 추론을 한다. 추론 결과로 해당하는 논문의 Identifier를 반환한다. 본 논문의 시스템은 Pallet 추론 엔진을 사용한다 [19].

3.3 OWL 온톨로지 기반 검색 프로세스

3.3.1 논문 주석 정보 데이터 생성

3.2 절의 그림 6에서 기술한 바와 같이 전처리기는 XML 형식의 메타 데이터인 논문 주석 정보 데이터를 생성한다. 전처리기는 다양한 HTML 형태의 원본 논문 정보 데이터를 파싱 한 후 필요한 문장 및 어휘를 추출해내고 유의어 처리를 하여 XML 형식으로 논문 주석 정보 데이터를 생성해낸다. 이때 논문 주석 정보 데이터의 태그는 3.2.1절에서 기술한 바와 같이 더블린 코어 태그와 개념 온톨로지의 클래스 정보를 참조하여 논문 검색에 필요한 일부 태그를 추가하여 표 3과 같이 정의하였다.

표3에서 주석정보 데이터의 태그와 개념 온톨로지 클레스명이 일 대 일로 대응하지 않는 것은 실제 추론에 이용되지 않는 태그가 제외되었기 때문이다.

3.3.2 OWL 논문 온톨로지의 생성

그림 6에 나타낸 바와 같이 OWL 온톨로지 생성기는 검색기의 질의어 관리기에서 제 1 키워드에 대하여 1 차 검색하여 반환되는 XML 형식의 논문 주석 정보 데이터를 대상으로 개념 온톨로지를 참조하여 최종의 OWL 논문 온톨로지를 생성한다. 이때 논문 주석 정보의 XML 태그와 OWL 논문 온톨로지 클래스간의 관계는 3.3.1 절의 표 3을 기반으로 하였다. 또한 논문 온톨로지에서 각 클래스는 다른 클래스와의 관계를 표현하는 속성을 정의하고 또 해당 클래스를 구성하는 인스턴스들의 제약사항들(restrictions)을 기술해야 한다. 이것을 클래스 서술(Class Description)이라 하고, 그림 7은 Protege4.0으로 작성한 논문 온톨로지의 Identifier 클래스에 대한 클래스 서술의 예를 보여준다.

표 3. 주석정보 데이터의 태그와 개념 온톨로지의 클래스명
Table 3. Tags of annotation data and class name of conceptual ontology

태그		entity	개념온톨로지 클래스
title	kor	제목	-
	eng	제목 (영문)	
identifier		인스턴스 키	Identifier
subject		키워드	Keyword
creator		저자	Writer
contributor	name	논문지정보	Name
	vol	권	-
	no	호	-
publisher		발행학회정보	Publisher
date		발행년도	Year
grade		논문지등급	Grade
language		기술언어	Classificaion
type		논문분야	Domain
source		원본파일 링크	-
description		초록	-
format		논문파일 형태	DocType

3.3.3 DL 쿼리 생성

DL 쿼리는 사용자가 입력한 질의어들 중에서 제 1 키워드를 제외한 나머지 질의어들을 가지고 생성된다. 따라서 DL 쿼리 생성기는 질의 관리기로부터 그림 8와 같은 XML형식의 질의어를 질의어 관리기로부터 전달받는다. DL 쿼리 생성기는 입력된 질의어들로부터 추론엔진에 입력될 DL 쿼리를 생성해낸다.

그림 9는 그림 8의 질의어들로부터 생성한 DL 쿼리를 나타낸다. 그림 8의 사용자 질의어들 중에서 제 1 키워드인 '온톨로지'는 3.2 절에서 설명한 것과 같이 질의어 관리기에서 키워드 기반 검색을 위하여 사용되었고 그 결과, 1차로 XML 형식의 논문 주석 정보 데이터들을 검색하하기 위하여 이미 사용하였기 때문에 그림 8의 DL 쿼리에는 제 1 키워드인 '온톨로지'는 제외된 것을 보여주고 있다.

● Identifier
hasDomain *some* ArtificialIntelligent
hasDocType *value* PDF
hasDomain *some* 인공지능시스템및응용
hasKeyword *value* Pulse_Diagnosis_Data
hasKeyword *value* Service_of_Diagnosis
hasKeyword *value* ontokogy
hasKeyword *value* 정보처리학회논문집B
haswriter *value* 한준수

그림 7. Identifier 클래스의 클래스 서술
Fig. 7. Class description of Identifier class

```
<?xml version="1.0" encoding="utf-8"?>
<query>
  <subject>
    <keyword_1>온톨로지</keword_1>
    <keyword_2>시맨틱웹</keword_2>
  </subject>
  <date>2008</date>
  <grade>Record</grade>
  <type>Computer</type>
</query>
```

그림 8. 사용자 질의어의 XML 표현
Fig. 8. XML expression of user query

hasKeyword value 시맨틱웹 and hasYear some 2008 and hasDomain some Computer and hasPublication some (hasGrade value Record)

그림 9. 생성된 DL 쿼리
Fig. 9. Generated DL query

3.3.4 OWL 온톨로지 추론

추론 엔진은 OWL 형식의 온톨로지를 로드해 일관성 검사를 수행한 후 클래스의 계층구조를 추론하고 마지막으로 클래스에 대하여 기술된 객체속성을 추론한다. 1차적으로 제 1 키워드를 가지고 키워드 검색한 논문 주석 정보 데이터들로부터 생성한 OWL 논문 온톨로지를 대상으로 DL 쿼리에 대한 2차 추론을 하여 사용자 질의에 부합하는 논문의 Identifier를 결과 처리기에 반환한다.

그림 10은 '온톨로지'라는 제 1 키워드에 대하여 1 차로 검색한 논문들의 OWL 온톨로지를 추론엔진에 의해 추론하여 분류된 클래스들의 계층 구조를 나타낸다.

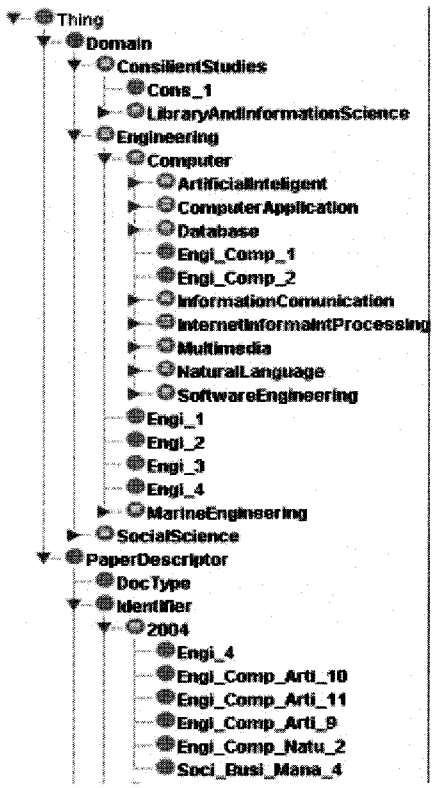


그림 10. 추론된 OWL 온톨로지의 계층 구조
Fig. 10. Hierarchy of inferred OWL ontology

3.3.5 검색 데이터 흐름

본 절에서 지금까지 설명한 시스템의 데이터 흐름을 종합하여 기술한다. 그림 11은 검색데이터의 종합적인 흐름을 보여주고 있으며, 이 흐름을 크게 분류하면 전처리과정과 사용자 질의에 대한 검색 처리과정으로 분류할 수 있다. 그림 11에서 점선은 온톨로지 또는 데이터베이스를 참조하는 것을 표현하고 있으며, 실선은 실제 데이터들의 흐름을 보인 것이다. 자원 데이터베이스에는 웹에 산재된 데이터로부터 수집하여 가공한 HTML 형태의 원본 논문 정보 데이터가 저장된다. 원본 논문 데이터가 1-5까지 전처리 과정을 거치면 논문 주석 정보 데이터로서 생성된 XML 문서와 함께 URI 데이터베이스에 등록된다. A-K는 사용자 질의를 처리하는 과정을 나

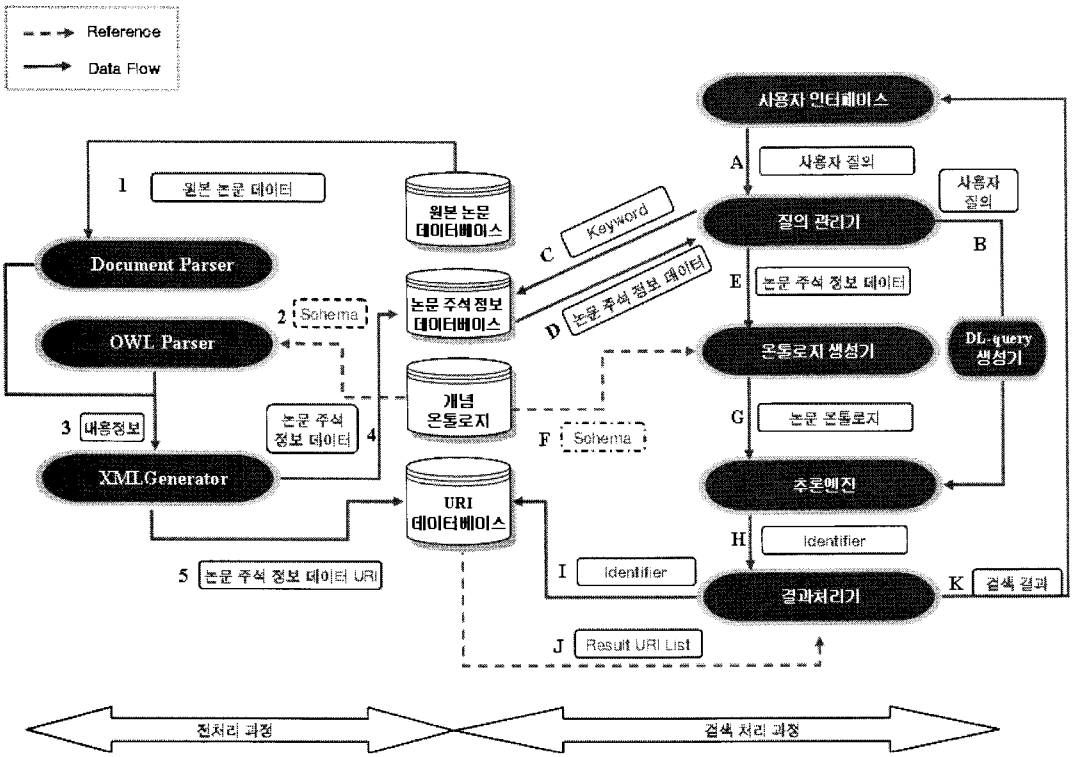


그림 11. 온톨로지 기반의 논문 검색 시스템 데이터 처리 과정
Fig. 11. Data process of paper retrieval system based on ontology

타낸 것이다. 사용자 인터페이스에서 사용자 질의를 처리하여 제 1 키워드와 나머지 질의어들을 분류한다.

질의어 관리기에서 제 1 키워드에 의하여 질의한 검색 결과물인 XML 형식의 논문 주석정보 데이터들이 온톨로지 생성기에 전달되고, 나머지 질의어들은 DL 쿼리를 생성하여 추론 엔진에 전달된다. 온톨로지 생성기에서는 XML 메타 데이터에 대응하는 OWL 논문 온톨로지를 생성한 후 추론 엔진에 전달한다. 추론엔진은 OWL 논문 온톨로지에 대하여 DL 쿼리를 기반으로 하여 추론을 한다. 추론 결과로 반환되는 OWL 논문 온톨로지 구성 클래스인 Identifier가 결과 처리기에 전달되고, 이 Identifier를 참조하여 URI 데이터베이스에 링크된 최종 논문들을 검색 결과로 사용자에게 제공한다.

IV. 실험 및 평가

본 장에서는 구현된 시스템을 평가하기 위해 다음과 같은 방법으로 실험한다. 논문을 검색하기 위하여 키워드와 학술분야를 비롯한 복합적인 검색조건을 적용한 사용자 질의를 바탕으로 반환된 결과가 사용자 질의에 부합하는가에 대한 정확도를 측정한다. 또한 동일한 데이터를 바탕으로 다른 논문 검색 사이트와의 검색 결과를 비교한다. 검색결과의 성능평가를 위해 정확도(precision)와 재현율(recall)을 측정하였다. 본 논문에서 정확도 ρ와 재현율) γ은 식 (1)과 같이 계산하였다.

$$\rho=\eta/(\eta+\epsilon),\ \gamma=\eta/(\eta+\vartheta) \quad \cdots\cdots (1)$$

식에서 η는 검색된 논문들 중 올바른 논문 개수, ϵ은 검색된 된 논문들 중 올바르지 않은 논문 개수, ϑ는 올바르나 검색되어 지지 않은 논문 개수를 의미한다.

본 논문의 원본 논문 데이터 수집은 논문 검색 사이트 'KSI KISS'에서 제공하는 논문 정보 중 100 건이 이용되었다. 데이터 수집에 사용된 원본 논문 정보에는 키워드, 제목, 초록, 날짜, 저자, 발행정보 등이 포함되어 있고 전처리 과정을 통해 데이터를 수집 및 저장하였다. 사용자 질의어들 중에서 제 1 키워드를 대상으로 OWL 온톨지를 생성하고, 나머지 질의어들에 대해서는 DL 쿼리를 작성하여 생성된 OWL 온

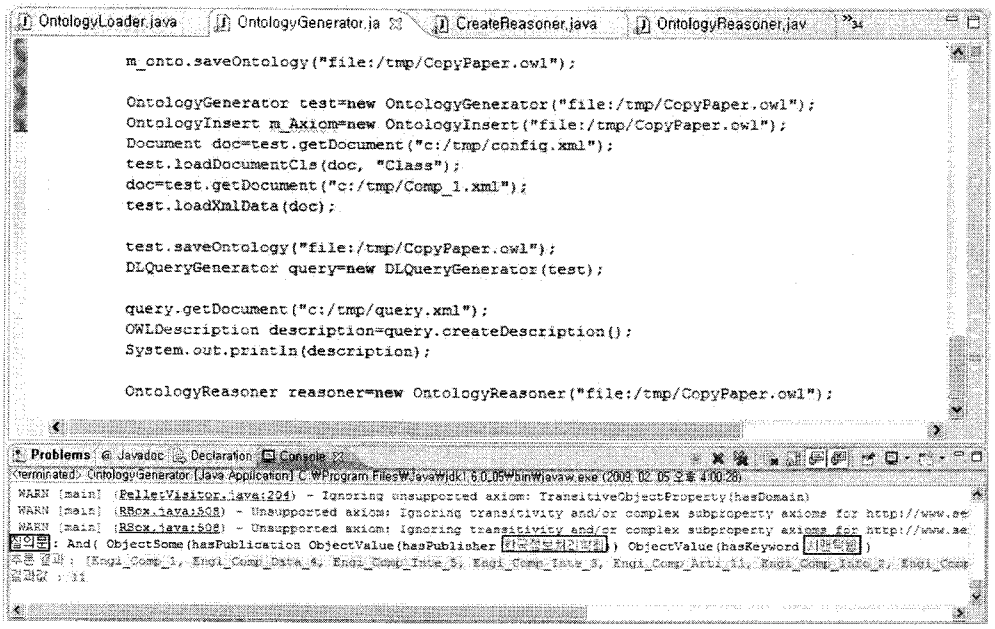


그림 12. OWL 온톨로지 추론 프로그램
Fig. 12. OWL Ontology Reasoning Program

톨로지에 대한 추론을 하여 결과를 얻는 방식으로 실험을 진행하였으며 실행 결과는 다음과 같다.

표 4. 실험 결과
Table 4. Experimental result
(2 개의 질의어 실험)

사용자 질의	키워드1+키워드2		키워드1+발행년도		키워드1+발행정보	
	A	B	A	B	A	B
η	9	17	31	40	16	16
ϑ	11	13	19	10	7	6
ε	14	6	13	13	16	14
Precision(%)	39	73	68	75	50	53
Recall(%)	45	56	62	80	69	72

그림 12는 java 기반으로 개발한 온톨로지 추론 프로그램으로 콘솔창을 통해 검색 조건과 결과를 확인 할 수 있다. 제1키워드 기반으로 생성된 OWL 온톨로지를 대상으로 제 2키워드와 발행정보를 추론한 것이다. 추론 결과로 논문의 Identifier를 결과값으로 검색결과물의 개수를 출력했다. 본 논문에서 제안한 검색시스템의 결과와 기존 논문 검색 사이트에서 동일한 조건으로 검색한 결과와 비교분석하여 표 4에 나타냈다. 두 검색결과의 대조군은 동일한 100건의 데이터를 대상으로 사용자가 해당질의를 통해 검색하고자 했던 논문 리스트이다.

(3 개의 질의어 실험)

사용자 질의	키워드1+키워드2+발행정보		키워드1+키워드2+학술분야	
	A	B	A	B
η	5	10	7	14
ϑ	13	7	13	6
ε	4	1	4	5
Precision(%)	55	90	63	73
Recall(%)	27	58	35	70

실험1은 제 1 키워드와 다른 질의어 하나씩을 조합한 3가지 경우에 대한 검색 결과를 일반 논문 검색 사이트와 비교 측정하였다. 실험2는 제 1 키워드와 다른 질의어 2 개씩을 조

합한 2 가지 경우에 대하여 검색한 결과를 비교하였다. 표 4에서 A는 일반 논문 검색 사이트의 검색결과이며, B는 본 논문에서 제안한 OWL 온톨로지 기반의 추론 결과이다.

실험 결과 본 시스템에서 제안하는 검색방법을 사용했을 경우, 사용자 질의어가 2개일 때 평균 정확도는 67% 재현율은 69%를 나타내었고, 사용자 질의가 3 개일 때 평균 정확도는 81.5% 재현율은 64%였다. 반면, 일반 논문 검색 사이트에서 동일한 질의로 검색한 결과 사용자 질의어가 2 개일 때 평균 정확도는 52% 재현율은 59%였으며, 질의어가 3 개일 때 평균 정확도는 59% 재현율은 32%의 수치를 나타내었다. 따라서 시험 결과에서 알 수 있는 바와 같이 OWL 기반으로 논문의 온톨로지를 작성하고 2차 적인 추론함으로써 검색 효율을 높일 수 있다는 것을 확인하였다.

V. 결론

본 논문에서는 기존의 논문 검색 서비스가 갖는 단점을 극복하기 위하여 시맨틱 웹의 새로운 온톨로지 언어로 부상한 OWL 기반의 논문 온톨로지를 구축하고, 이 온톨로지에 대한 추론 기능을 부여하여 사용자의 의도에 적합한 논문을 검색해 낼 수 있는 OWL 온톨로지 기반의 논문정보 검색시스템에 대하여 기술하였다. 이를 위하여 웹에 존재하는 논문들의 요약정보를 XML 형식의 메타 데이터인 논문 주석 정보 데이터로 작성하고, 이것을 기반으로 OWL 논문 온톨로지를 생성한 후에 사용자 질의어들로부터 DL(Description Logic) 쿼리를 작성하여 이를 근거로 논문 온톨로지에 대하여 추론함으로써 지능적인 정보 검색이 가능하도록 하였다. 본 논문의 온톨로지 저작도구로는 Protege 4.0을 사용하였고, 추론 엔진으로 Pellet 1.5를 이용하였다.

논문의 검증을 위하여 논문 검색 사이트 KSI KISS에서 제공하는 논문 100 여건을 대상으로 검색 사이트에서 검색한 결과와 본 논문의 온톨로지 기반의 검색을 수행한 결과를 비교하였으며, 결과적으로 본 논문의 시스템이 사용자의 질의에 대해 보다 정확한 검색 결과를 제공하는 것을 확인하였다.

차후에 더욱 정확한 논문에 대한 검색결과를 얻고 사용자 만족도를 향상시키기 위해서는 보다 정형화된 온톨로지의 구축과 대용량 온톨로지를 대상으로 하는 추론엔진에 대한 연구가 필요할 것으로 생각된다.

참고문헌

[1] 최중민, 조성정, 김진형, 이재호, 양정진, 김인철, 강민구, 박영택, "특집 : 시맨틱 웹," 정보과학회지, 제 21권, 제 3호, 3-50쪽, 2003년 5월.

[2] 정도헌, "시맨틱웹을 위한 온톨로지 언어와 구현사례 연구," 정보관리연구, 제 34권, 제 3호, 87-109쪽, 2003년 9월.

[3] 김이란, "온톨로지의 의미정보를 이용한 RDF 스키마 생성에 관한 연구," 석사학위논문, 연세대학교 대학원 문헌정보학과, 2001년 2월.

[4] 김중태, "*웹 2.0 시대의 기회, 시맨틱웹*," 디지털미디어리서치, 2005년.

[5] 오삼균, "시맨틱 웹 기술과 활용방안," 정보관리학회지, 제 19권, 제 4호, 297-319쪽, 2002년 12월.

[6] Berners-Lee. T. and J. Hendler. and O. Lassila. "*The Semantic Web,*" Scientific American, May. 2001.

[7] 서은석, 최용준, 박영택, "OWL-DL 기반의 대용량 ABox 추론 기법," 한국정보과학회 추계학술발표논문집, 제 33권, 제 2(B)호, 352-356쪽, 2006년 10월.

[8] 정성무외 6인, "*KEM 고도화를 위한 온톨로지 기반 시맨틱 웹 연구*," 한국교육학술정보원, 연구보고서, 2006년 10월.

[9] 서은석, 최용준, 박영택, "대용량 ABox에서 서술논리 SHIQ(D) 추론," 정보과학회 논문지:소프트웨어 및 응용, 제 34권, 제 6호, 530-539쪽, 2007년 6월.

[10] Hyunjang Kong, Myunggwon Hwang, Pankoo Kim, "Design of the Automatic Ontology Building System about the Specific Domain Knowledge," Proceeding of ICA0T2006, pp.1405-1408, May. 2006.

[11] Tobias Horney, "Design of an ontological knowledge structure for a query language for multiple data sources," Swedish Defence Research Agency, Scientific repor.t FOI-R-0498—SE, Mar. 2002.

[12] Zhanjun Li, Victor Raskin, Karthik Ramani, "A Methodology of Engineering Ontology development for Information Retrieval," Proceeding of International Conference on Engineering design ICED'07, pp. 1- 12, Aug. 2007.

[14] Smith. M. K. and C. Welty, and D. L. McGuinness "Web Ontology Language (OWL) Guide Version 1.0," W3C Candidate Recommendation, http://www.w3.org/TR/owl-guide, 18 Aug. 2003.
[15] Deborah L. McGuinness, Frank van Harmelen, "OWL Web Ontology Language Overview," W3C Recommendation,http://www.w3.org/TR/2004/REC-owl-features-20040210/ ,10 Feb. 2004.
[16] Peter F. Patel-Schneider, Patrick Hayes, and Ian Horrocks, "OWL Web Ontology Language Semantics and Abstract Syntax," W3C Recommendation, http://www.w3.org/TR/2004/REC -owl-semantics-20040210/,.10 Feb.. 2004.
[17] Dan Brickley and R.V. Guha, "RDF Vocabulary Description Language 1.0: RDF Schema , Editors," W3C Recommendation. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/, 10 Feb. 2004.
[18] 박현철, "온톨로지 기반 가중치 부여 논문 검색 모델," 한국정보과학회, 가을 학술발표논문집, 제 34권. 제 2(C)호, 328-330쪽, 2007년 10월.
[19] 위다현, 강현민, 한광록, "온톨로지 기반 논문 검색 시스템 설계," 대한전자공학회 하계 학술대회, 2008년 6월.
[20] Natalya F. Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W. Fergerson, Mark A. Musen, "Creating Semantic Web Contents with Protégé-2000," IEEE INTELLIGENT SYSTEMS., pp. 60-71, 2001.
[21] Evren Sirin. and Bijan Parasia. and Bernardo Cuenca Grau, "Pellat: A Practical OWL-DL Reasoner," Web Semantics: Science, Services and Agents on the World Wide Web, Vol 5. Issue 2, 2007.

저 자 소 개

선 복 근

2006년 2월 : 호서대학교 컴퓨터공학과 공학박사
2008년 ~ 현재 : 호서대학교 공학교육혁신센터 전임강사
관심분야 : HCI, 시멘틱웹,정보검색

위 다 현

2009년 2월 : 호서대학교 메카트로닉스공학과 석사
관심분야 : HCI, 웹 프로그래밍, 시맨틱웹, 온톨로지, 정보검색

한 광 록

1989년 8월 : 인하대대학교 정보공학박사
1991년 ~현재 : 호서대학교 교수
관심분야 : 정보검색, HCI, 멀티미디어, 시멘틱웹