규칙기반 온톨로지 인스턴스 자동생성

윤현주* • 변상용** • 김장형** • 변영철**

Ontology Instance Generation using Rule-based Approach

Hyun-Ju Youn*, Sang-Yong Byun**, Jang-Hyung Kim** and Yung-Cheol Byun**

ABSTRACT

In this paper, we propose a method generating of instance ontology from web pages for the lodge. First, we analyze non-structured web documents and generate information extraction rules in order to extract specific informations from web sources. And we extract the informations from various web pages related to lodge using generated rules. And then ontology instances are automatically generated through adding OWL semantic information to the extracted informations. Our method is able to cut down on time and cost than a method which generates various ontology instances manually. Also agents can search the information corresponding to semantically what the user needs in the various web pages using numerous vocabulary having same meaning and provide good quality of informations to user.

Key Words : Rule-based approach, ontology, instance generation, information retrieval

Ⅰ. 서 론

HTML은 브라우저에 디스플레이 또는 레이아웃을 중심으로 하는 표현 중심의 기술로서 문서의 내용과 의미를 나타내는 시맨틱 정보를 구조적으로 표현하기는 쉽지 않다. 현재의 웹은 정보의 양이 증가함으로써 사용자의 요구에 맞는 정보를 찾을 확률이 높아진다는 좋은 점이 있지만 정보의 다량화로 인해 사용자가 정보를 검색할 때 더 많은 시간과 노력을 소비해야 한다는 단점이 있다[1].

이러한 문제점을 해결하기 위한 시맨틱 웹은 웹상의 데이터의 의미를 인간이 아닌 기계가 이해하고 처리할 수 있도록 하는 기술로서 기존의 웹과 구분되는 것이 아니라 기존의 웹에 의미 정보를 부여하는 것이다[2]. HTML 문서에서 데이터를 추출할 때의 문제점은 앞에서도 언급했듯이 HTML 문서의 목적이 사람에게 시각적으로 정보를 보여주는 데 있다는 점이다. 그리고 정보를 제공하는 사이트 및 문서 작성자의 스타일에 따라서 사용하는 단어와 레이아웃이 서로 다르다. 원하는 정보를 추출하기 위해 화면 구성 관련 내용, image, formatting 등의 정보를 제거해야 하며 의미는 같지만 어휘를 서로 다르게 사용하는 여러 정보 소스로부터 필요한 정보 추출을 수행하기에는 어려움이

* 제주대학교 대학원 컴퓨터공학과
Dept. of Computer Engineering, Cheju Nat'l Univ.
** 제주대학교 통신컴퓨터공학부, 첨단기술 연구소
Faculty of Telecommunication and Computer Eng. ResearchInstitute of Advanced Technology, Cheju Nat'l Univ.

많다[3].

본 논문에서는 비 구조화된 웹 문서를 분석하여 여러 웹 페이지에 적용 가능한 공통 규칙을 생성하고 그에 기반하여 숙박 관련 여러 HTML 문서로부터 원하는 정보를 추출한다. 추출된 정보에 시맨틱 정보를 추가하여 온톨로지 인스턴스를 자동으로 생성하는 방법에 대하여 제안한다. 생성된 온톨로지 인스턴스를 이용하여 검색 엔진은 의미 기반 검색이 가능하게 되며 사용자가 원하는 정보와 의미가 같은 정보를 제공하게 되므로 사용자에게 양질의 정보를 제공할 수 있다.

본 논문의 구성은 다음과 같다. Ⅱ절에서는 시맨틱 웹과 온톨로지 그리고 규칙기반 정보 추출 관련 연구들과 온톨로지 생성 관련 연구들에 대해 논하고, Ⅲ절에서는 전체 시스템 흐름과 숙박 정보 온톨로지 구성에 관해 설명을 한다. Ⅳ절에서는 비 구조화된 문서에서 정보를 추출하기 위한 규칙과 온톨로지 인스턴스 자동 생성에 관해 설명하고 Ⅴ절에서는 실제 웹 페이지를 테스트한 결과에 대해 논한다. 끝으로 Ⅵ절에서는 결론 및 향후 연구 방향을 기술한다.

Ⅱ. 관련 연구

웹의 발달로 인해 정보의 양이 증가함으로써 사용자가 정보를 검색할 때 더 많은 시간과 노력을 소비해야 한다는 문제를 해결하기 위해 Tim Berners-Lee를 비롯한 연구자와 학자들이 "시맨틱 웹"이라는 차세대 웹 개념을 제안하고 있다. 이는 문서의 의미를 명백하고 기계가 이해할 수 있는 형태로 표현하여 컴퓨터가 웹 자원들을 효율적으로 관리할 수 있게 하려는 것이다.

시맨틱 웹의 핵심 기술이자 핵심 구성 요소라고 할 수 있는 온톨로지는 단순히 용어(term)의 체계적 구조화가 아니라 특정한 영역의 개념에 대한 정의와 관계, 그리고 개념이 가지는 특수한 속성들로 이루어진 집합체이며 기계와 기계 사이에도 형식적 모델을 통해 원활하게 의사소통이 가능하도록 하는 의미적인 구조를 가진다. 이로 인해 온톨로지는 WWW에서 적용되는 시맨틱 웹을 가능하게 하며 응용프로그램 사이에서 웹 기반 지식 처리, 공유, 재사용 하는 것을 가능하게 하는 중요한 역할을 한다[4,5].

온톨로지는 초기에는 단순히 지식 공유와 재사용을 위해서 개발되었으나 점차 온톨로지의 개념이 지능적인 정보 통합, 협동 정보 시스템, 정보 검색, 전자 상거래, 지식 관리 같은 분야로 확대되고 있으며 온톨로지 기반 정보검색 기술은 중요한 정보가 있는 자원을 빠르게 찾아 사용할 수 있다는 점과 자원을 찾는 정확도를 향상시킬 수 있다는 점에서 중요한 기술로 자리 잡아 가고 있다.[6]

또한 검색엔진이 온톨로지에 정의된 개념과 규칙을 활용하면서, 검색 향상을 위해 추론 규칙을 이용하기 때문에, 단순히 사용자의 질의와 일치되는 문서만 보여주는 것이 아니라 사용자의 질의의 의미를 분석하여 그와 관련된 정보를 온톨로지에 표현된 관계에 따라 다시 질의를 적절하게 바꿀 수도 있게 한다[7,8].

본 논문에서는 시맨틱 웹과 온톨로지의 장점을 활용하여 여행과 관련된 도메인으로 범위를 한정, 온톨로지 언어인 OWL McGuinness[9,10]을 사용하여 숙박 정보 온톨로지를 구축한다. 또한 온톨로지 인스턴스를 자동으로 생성함으로써 온톨로지 기반의 의미 검색이 가능하게 되어 사용자에게 효율적인 검색이 가능하도록 한다.

규칙을 기반으로 정보를 식별 및 추출하는 연구들이 많이 있다. 논문 이 등[11]은 출판물에 따라 다양한 종류의 포맷팅 방식이 사용되는 논문 영상으로부터 지식베이스를 구성하는 지식규칙에 기반하여 정형화된 데이터를 추출한다. 이 경우 문서 영상으로부터 이미지, 드로잉, 테이블 등의 비 텍스트 객체와 텍스트 객체들을 규칙에 기반하여 영역 분할 및 식별을 한다. 규칙에 기반하여 정교하게 영역 분할 및 식별을 하지만 특정 정보를 추출하지는 않고 있다

논문 서 등[12]은 준 구조화된 웹 문서에서 각 사이트에 규칙 파일 하나씩을 자동 생성하여 정보를 추출한다. 하지만 라벨 문서에 한정되어 있어서 정보의 추출은 준 구조화된 웹 문서의 테이블이나 리스트 등의 구조화된 부분에서만 추출 가능하며 여러 사이트를 동시에 검색해야 하기 때문에 네트워크 속도에 영향을 많이 받는 단점이 있다. 또한 XML을 사용하여 다른 도메인에도 적용 가능하다는 장점이 있는 반면 RDF나 OWL에 비해 관계 표현에 제약이 많다.

제안하는 방법은 비 구조화된 웹 문서로부터 원하는 정보를 추출하기 위해 문서를 분석하여 의

미는 같으나 다른 어휘들을 사용하는 여러 웹 문서에 적용할 수 있는 공통 규칙을 생성한다. 그 규칙을 기반으로 숙박 관련 여러 웹 페이지들로부터 원하는 정보를 추출한다.

온톨로지 반자동 생성에 관한 논문 임 등[13]은 약품 매뉴얼에 있는 텍스트들을 대상으로 전문용어의 처리에 의한 도메인 온톨로지를 반자동으로 구축한다. 이는 텍스트들이 전문용어인 준 정형적인 데이터를 이용한 온톨로지 반자동 생성으로 생성된 온톨로지는 개념들과 그들의 관계만 표현하고 있다.

본 논문에서는 규칙에 기반하여 웹 문서에서 추출한 정보에 시맨틱 언어인 OWL을 사용하여 숙박 온톨로지 구조에 적합하게 시맨틱 정보를 추가하여 온톨로지 인스턴스를 자동으로 생성한다. 이렇게 함으로써 수 많은 온톨로지 인스턴스들을 수동으로 생성 시 드는 시간과 비용을 줄일 수 있으며 검색 엔진은 찾으려고 하는 키워드와 의미적으로 유사하지만 구조적으로 다른 어휘들을 가진 여러 웹 문서에서 원하는 정보를 찾아줄 수 있어 사용자에게 양질의 정보를 제공할 수 있다.

III. 시스템 흐름도와 숙박 온톨로지

제안하는 방법이 실생활에 적용할 수 있음을 보이기 위하여 여행과 관련된 도메인으로 범위를 한정하여 제주도 숙박 정보에 대한 온톨로지를 온톨로지 언어인 OWL을 이용하여 생성, 여행 정보 중 숙박 정보를 담고 있는 HTML 웹 페이지를 테스트에 이용하였다. Fig. 1은 전체 시스템 흐름도이다.

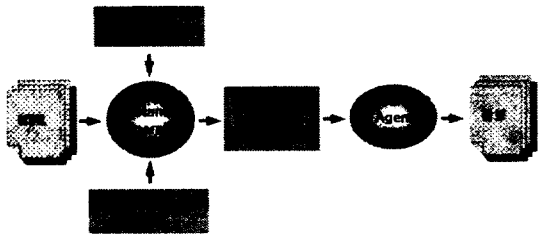

Fig. 1 System architecture.

비 구조화된 텍스트로 이루어진 웹 문서에서 원하는 정보를 추출하기 위해 문서를 분석하여 의미는 같으나 서로 다른 어휘들을 사용하는 여러 웹 문서에 적용 가능한 공통 규칙을 생성한다. 생성된 규칙에 의해 숙박 관련 여러 웹 문서로부터 원하는 정보를 추출한다. 추출된 정보에 숙박 온톨로지 구조에 적합하게 OWL을 사용하여 시맨틱 정보를 추가하여 온톨로지 인스턴스를 자동으로 생성한다. 사용자가 검색을 요청하였을 때 검색 Agent는 Jena API를 이용하여 온톨로지 기반의 의미적 정보 검색이 가능하여 사용자에게 양질의 정보를 빠르고 정확하게 제공할 수 있다. Fig. 2는 숙박 정보 온톨로지의 구조를 그림으로 표현한 것이다.

숙박 클래스와 그 하위 클래스인 호텔, 펜션, 콘도 클래스가 있으며 호텔, 펜션, 콘도 등에 모두 존재하는 객실 클래스로 구성된다. 객실 클래스는 하위 클래스로 객실의 전망과 객실의 종류를 나타내는 RoomView 클래스와 RoomKind 클래스를 가진다. 호텔, 펜션, 콘도 클래스와 객실 클래스 요소 간의 관계를 객체형 속성인 'hasRoom'으로 표현하고 있으며 객실과 객실전망, 객실형태와의 관계를 객체형 속성인 'hasView'와 'hasKind'로 표현하고 있다. 그리고 숙박 클래스 즉, 호텔, 펜션, 콘도 클래스들은 클래스 요소가 취해야 하는 데이터의 형식과 값을 기술하기 위해 몇 가지 데이터형 속성들로 구성된다.

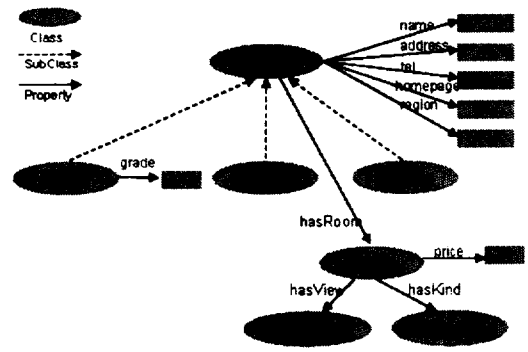

Fig. 2 Lodge ontology architecture.

Fig. 3은 실제로 OWL을 이용하여 작성한 숙박 온톨로지 구조의 일부분을 보여준다. 온톨로지 구조도에 보이는 것처럼 클래스들과 속성들이 정의되어 있으며 객실의 종류와 같은 의미를 갖는 어휘

들을 OWL을 사용하여 정의하고 있다. 같은 의미이지만 서로 다르게 사용하는 어휘들을 Fig. 3과 같이 정의해 주면 같은 의미로 이해하여 정보를 검색할 수 있다.

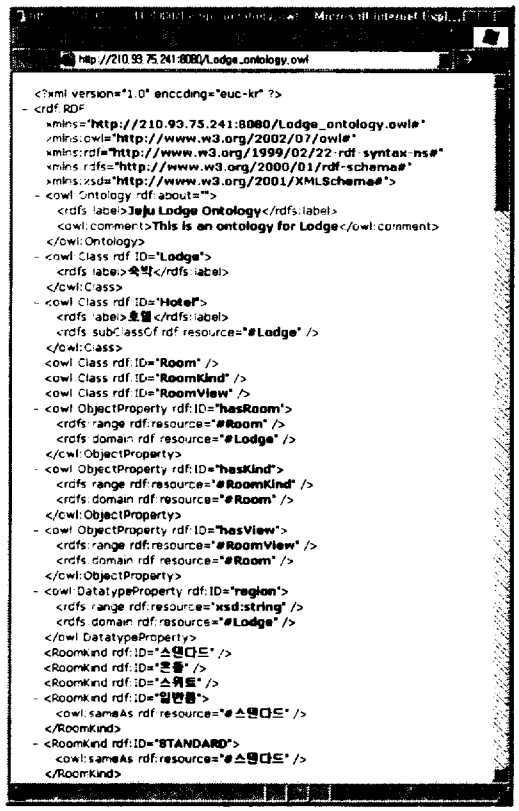

Fig. 3 Example of lodge ontology.

Ⅳ. 규칙기반 온톨로지 인스턴스 자동생성

제주도 숙박 관련 여러 웹 문서들이 의미는 같으나 각기 다른 어휘들을 사용하기 때문에 이들로부터 원하는 정보를 추출하기 위해서 숙박 관련 비구조화된 웹 문서인 호텔의 HTML 문서를 입력으로 받아 객실명, 객실가격, 기타 등의 숙박 정보를 추출하기 위해 문서를 분석하여 규칙을 생성한다. 본 연구에서 사용한 호텔 문서 관련 규칙의 예는 Fig. 4와 같다.

<RoomName>

(RoomKind SameAs in LodgeOntology) AND
if (+ RoomKind +)
then RoomName

if (postfix is "객실 ") OR
(postfix is "룸 ")
then RoomName

if RoomKind behind locate Price
then RoomName

if RoomKind behind locate RoomKind
then first-RoomKind is RoomName

<Price>

if postfix of digit is "원 "
then Price

if postfix of digit is "Won"
then Price

if prefix of digit is "\"
then Price

if digit > 100000
then Price

if (+ Price +) or (PriceColor is not black)
then Price

Fig. 4 Example of rules.

- 객실은 일반적으로 '일반룸', '스위트룸', '온돌룸' 등의 객실 종류를 가지며, 객실명으로 객실을 설명하는 경우가 아니면 굵은 글씨체를 사용하기도 하며 접미사로 '객실' 또는 '룸'을 두어 객실명 임을 명확히 하는 경우도 있다. 또한 가격 정보 앞에 위치하며 객실전망이 아닌 경우 객실명을 의미하는 등의 규칙을 갖는다.
- 가격은 접미사가 '원'또는 'Won'이며 접두사로 '\'을 가지며 굵은 글씨체 또는 글자색을 검정색이 아닌 다른색으로 표현하는 등의 규칙을 갖는다.
- 기타 사항으로 객실전망은 접미사가 '전망'이며 '산', '바다' 등의 단어와 결합하여 객실전망 정보를 나타내며 객실전망 뒤에는 어떠한 조사도

결합되지 않으며 하나 또는 하나 이상의 공백이 위치하는 등의 규칙을 갖는다.

- 영역을 제주로 한정할 경우 주소는 '제주도'로 시작한다.
- 위치하는 지역은 주소의 두 번째 단어에 의해 알 수 있으며, 본 논문에서는 지역을 제주시, 서귀포시, 북제주군, 남제주군으로 구분하고 있다.
- 전화번호는 '(Tel)' 또는 'TEL'로 시작하며 3~4개의 숫자와 -(하이픈), 그리고 4개의 숫자가 결합한 형태를 갖는다.

이와 같은 비 구조화된 여러 웹 문서에 적용 가능한 공통 규칙에 기반하여 숙박 관련 여러 웹 문서로부터 원하는 정보를 추출한다. 추출된 정보에 OWL을 사용하여 숙박 온톨로지에 적합한 시맨틱 정보를 추가하여 온톨로지 인스턴스 파일을 자동으로 생성한다.

Fig. 5는 자동 생성된 온톨로지 인스턴스의 예로 숙박 온톨로지에 정의되어 있는 호텔 클래스의 인스턴스이다. id가 silla1이며 이름은 '제주 신라 호텔'이고 '주소'와 '전화번호', '홈페이지 주소', '지역', '등급' 등이 기술되어 있으며 객실 'room001'을 가진다. 'room001'은 숙박 정보 온톨로지에 정의되어 있는 객실 클래스의 인스턴스로 객실 종류로는 '럭셔리', 객실 전망으로는 '산전망', 객실 가격은 365,000원 임이 기술되어 있다. 이는 실제 존재하는 호텔 인스턴스의 예로 신라호텔의 웹 페이지를 입력으로 받아 규칙에 기반하여 silla1이라는 실제 산전망의 가격이 365,000원인 럭셔리 룸이 온톨로지 인스턴스로 자동 생성된 예이다.

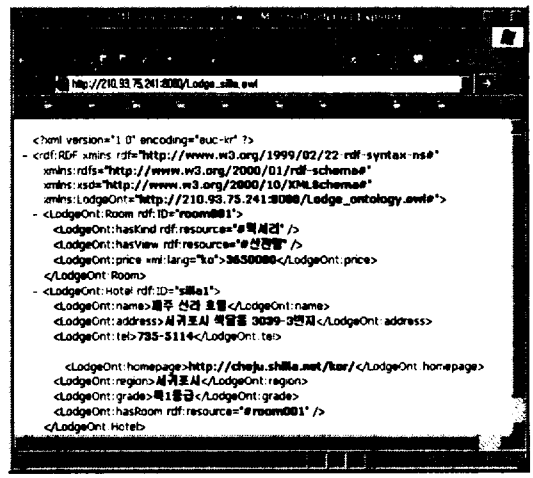

Fig. 5 Ontology instance.

여러 웹 문서에 공통된 하나의 규칙을 적용하여 온톨로지 인스턴스 자동 생성이 가능하다. 검색 엔진은 생성된 인스턴스를 이용함으로써 키워드는 다르더라도 동일한 의미를 갖는 다양한 키워드에 대한 효율적인 검색이 가능하다.

V. 구현 및 테스트

온톨로지의 생성은 Jena API를 이용하여 OWL 온톨로지를 생성하는 프로그램을 구현하였다. 온톨로지 인스턴스를 생성할 때 규칙을 기반으로 Fig. 3의 숙박 온톨로지에 정의된 구조에 적합하게 시맨틱 정보를 추가하여 자동으로 생성한다.

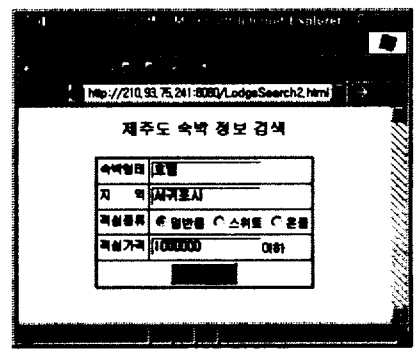

Fig. 6 User interface.

Fig. 6은 사용자 정보 검색 화면의 예로서, 사용자가 "제주도 호텔 중 서귀포시에 있으며 객실 가격이 1,000,000원 이하인 일반 객실은?" 이라는 질의를 요청하는 모습이다.

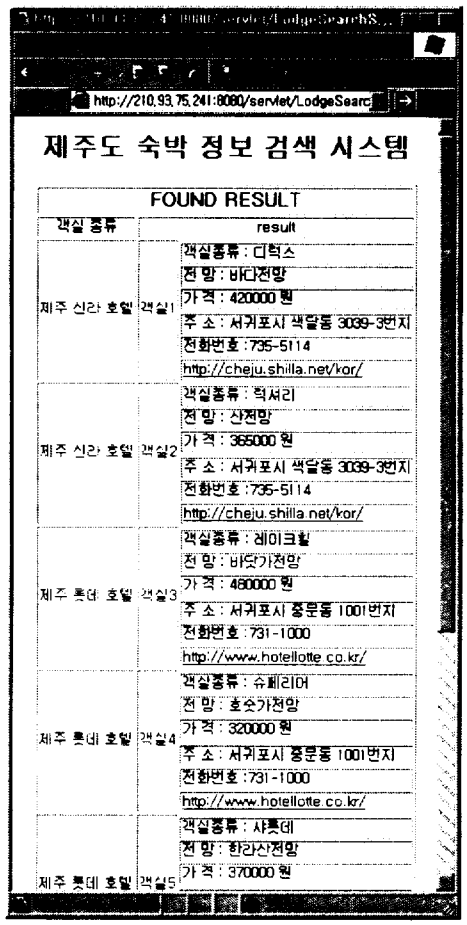

Fig. 7 Ontology-based search result.

Fig. 7은 질의에 대한 검색 결과를 보여주는 예로서, 온톨로지 기반의 의미적 정보 검색 결과를 보여준다. 이 결과는 동일한 의미의 키워드에 대한 검색이 가능함을 보여준다. 예를 들어, 숙박 정보에서 객실의 형태 중 고급방을 의미하는 스위트룸이나 한국식 온돌방이 아닌 일반룸 즉, 스탠다드룸과 디럭스, 럭셔리, 슈페리어 등을 같은 의미로 이해하여 정보를 검색하고 있으며 마찬가지로 객실전망으로 바다전망, 바닷가전망 등을 같은 의미로 이해하여 정보를 검색하고 있다.

Ⅵ. 결론 및 향후 연구 방향

본 논문에서는 비 구조화된 여러 웹 문서를 분석하여 사용자에게 원하는 정보를 제공하기 위하여 다양한 숙박 관련 웹 문서에 적용 가능한 규칙을 생성, 그 규칙에 기반하여 정보를 추출하고 추출된 정보에 숙박 온톨로지 구조에 적합한 시맨틱 정보를 추가하여 온톨로지 인스턴스를 자동으로 구축하는 방법에 대하여 제안하였다.

기존 웹 페이지에 대한 온톨로지 인스턴스를 수작업이 아닌 자동으로 구축함으로써 시간과 노력을 줄일 수 있으며 생성된 온톨로지 인스턴스를 기반으로 정보를 검색함으로써 보다 양질의 정보를 검색할 수 있다는 장점을 얻을 수 있다. 실제 응용이 가능함을 보이기 위하여 제주도 관광관련 숙박 정보를 대상으로 테스트한 결과 온톨로지 인스턴스를 자동 생성하여 이용함으로써 효율적인 검색이 가능하여 사용자에게 양질의 정보를 제공할 수 있었다. 한편, 웹 페이지에 원하는 정보가 이미지로 되어 있을 경우, 가령 주소, 전화번호, URL 등이 정보를 제공하는 사이트 작성자의 스타일에 따라서 하나의 이미지로 작성되어 있을 경우에는 정보를 추출하는데 실패하였다.

향후 온톨로지 인스턴스만이 아닌 온톨로지를 자동 생성하는 방안과 사용자에게 검색 결과를 비교하여 더 나은 정보를 추천해 줄 수 있는 방안에 대하여 연구할 계획이다.

참고문헌

1) 양정진, 2003, 시맨틱 웹에서의 온톨로지 공학, 정보과학회지, 제21권, 제3호, pp. 28-35.
2) 신효필, 2004, 지식기반(Knowledge Base)으로서의 온톨로지(Ontology)와 시맨틱 웹(Semantic Web) , 정보처리학회지 VOL. 11NO. 01pp. 0064~0075.
3) Latifur Kahn, Feng Luo, 2002, Ontology Construction for Information Selection, Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence, pp. 122-127.

4) 최호섭, 옥철영, 2004, 정보검색 시스템과 온톨로지, 정보과학회지, 제22권, 제4호, pp.62-71.
5) 박재홍, 임유정, 김도완, 박찬규, 조현규, 2002, Semantic Web 환경에서의 자원발견, 정보처리학회2004년추계학술대회, pp.0713-0716.
6) 김재훈, 2004, 정보추출의 기술 현황, 정보과학회지, pp.0035-0046.
7) 하상범, 박영택, 2004, 온톨로지를 통한 추론형 시멘틱 검색 시스템에 관한 연구, 정보과학회 2004년춘계학술대회, 제31권, 제1호, pp.0625-0627.
8) 정은경, 김영민, 변영철, 이상준, 2003, 온톨로지 기반의 정보검색, 정보과학회2003년추계학술대회, pp.0121-0123.
9) Deborah L. McGuinness, Frank van Harmelen, 2004, OWL Web Ontology Language Overview, http://www.w3.org/TR/owl-features/.
10) Michael K. Smith, 2004, OWL Web Ontology Language Guide, http://www.w3.org/TR/owl-guide/.
11) Kyong-Ho Lee, Yoon-Chul Choy, Sung-Bae Cho, 2000, Geometric and Logical Structure Analysis of Document Images : Knowledge-based and Syntactic Approaches, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1224-1240.
12) H. Seo, J. Yang, and J. Choi, 2001, Building Intelligent Systems for Mining Information Extraction Rules from Web Pages by Using Domain Knowledge, IEEE International Symposium on Industrial Electronics (IEEE-ISIE 2001), pp. 322-327.
13) 임수연, 송무회, 이상조, 2004, 전문용어의 처리에 의한 도메인 온톨로지의 구축, 정보과학회 논문지 B, VOL. 31No. 03pp. 0353~0360.