규칙기반 온톨로지 자동생성 및 검색

윤현주° 김영민 이상준 변영철
제주대학교 컴퓨터공학과
{baramdolii, mincando, sjlee, ycb}@cheju.ac.kr

Ontology Generation and Information Retrieval using Rule-based Approach

Hyunju Youn° Youngmin Kim Sangjoon Lee Yungcheol Byun
Dept. of Computer Engineering, Cheju National University

요 약

본 논문에서는 특정 도메인에 대한 온톨로지를 규칙에 기반하여 자동으로 생성하고, 이를 기반으로 원하는 정보를 추론을 통하여 효과적으로 검색하는 방법에 대해 제안한다. 제안하는 방법이 실생활에 적용할 수 있음을 보이기 위하여 여행과 관련된 정보중 숙박 정보를 담고 있는 HTML 웹 페이지를 테스트에 이용하였다. 웹 페이지에 표시되어 있는 숙박 정보에서 문서 구조 및 단어 측면에서의 규칙을 발견하고 이를 이용하여 온톨로지를 자동으로 생성한다. 숙박 정보 검색시 온톨로지에 정의된 관계를 이용하면 키워드는 다르더라도 동일한 의미를 갖는 다양한 키워드에 대한 효율적인 검색이 가능하다. 온톨로지 자동생성을 통하여 기존 웹 페이지에 온톨로지 추가시 드는 시간 및 비용을 줄일 수 있으며, 온톨로지 기반 검색 방법을 이용함으로써 사용자에게 보다 양질의 정보를 제공할 수 있다.

1. 서 론

HTML은 브라우저에 디스플레이 또는 레이아웃을 중심으로 하는 표현 중심의 기술로서 문서의 내용과 의미를 나타내는 시맨틱 정보를 구조적으로 표현하기는 쉽지 않다. 정보 검색에 있어서 기존의 데이터베이스 질의 시스템과 키워드 기반의 정보검색 시스템에서는 복잡한 질의문의 조합이나 여러 번의 재검색을 통하여 사용자의 요구에 맞는 결과를 찾아준다. 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. 그럼으로써 웹 상의 데이터가 의미있는 정보로 표현될 수 있는 방법을 제시한다[1].

본 논문에서는 특정 도메인에 대한(domain-specific) 온톨로지를 규칙에 기반하여 자동으로 생성하고, 이를 기반으로 원하는 정보를 추론을 통하여 효과적으로 검색하는 방법에 대해 제안한다. 기존의 웹에 추가되는 온톨로지를 규칙에 기반하여 자동으로 생성하고 추론을 통하여 효과적인 정보 검색이 가능함을 보인다. 온톨로지에 정의된 관계를 통하여 동일한 의미의 키워드에 대한 검색이 가능하다. 또한 크고 복잡한 도메인의 경우 수작업으로 온톨로지 구축 작업을 한다면 시간과 비용이 많이 들어감으로 규칙기반의 도메인 온톨로지의 자동 구축 방안을 이용함으로써 비용을 절감할 수 있다.

2. 관련 연구

온톨로지는 시맨틱 웹의 핵심 기술로 지식을 표현하고 추론하기 위해 사용된다. 표현 언어로는 RDF/RDFS, DAML+OIL, OWL등이 있다. W3C의 온톨로지 언어인 OWL은 온톨로지의 생성과 공유를 위한 시맨틱 마크업 언어로서 체계적인 온톨로지 구축을 지원할 수 있는 언어이다[2,3]. 이러한 언어로 구축된 온톨로지 기반의 정보 검색 기술은 자원을 빠르게 찾아 사용할 수 있다는 점과 자원을 찾는 정확도를 향상시킬 수 있다는 점, 또한 온톨로지에 정의된 개념과 개념간의 관계를 이용하여 사용자의 질의의 의미를 분석하고 동일한 의미를 갖는 키워드에 대한 검색이 가능하다는 장점들을 갖고 있다[4,5]. 기존의 연구된 논문에서는 하나의 클래스와 5개의 프라퍼티를 정의하여 온톨로지를 생성, 제한된 검색만이 가능하였다[6]. 본 연구에서는 이를 확장하여 온톨로지를 구축하고자 한다.

온톨로지를 생성하기 위한 저작 도구로는 대표적인 것으로 Protégé-2000, OntoEdit, Oiled 등이 있다. 저작도구들은 각기 다른 온톨로지 언어를 기반으로 하고 있으며, 서로 연동하기가 어렵고, 비용이 많이 들어간다[7,8,9]. 본 연구에서는 저작도구들의 문제점에 착안하여 비구조화된 텍스트로 이루어진 문서에서 원하는 정보를 추출하기 위해 사람이 문서를 분석하고 규칙을 생성하여 그 생성된 규칙에 의해 문서를 구조화하고 자동으로 온톨로지를 구축한다[10,11].

3. 제안하는 방법

3.1 온톨로지 자동생성

비구조화된 텍스트로 이루어진 HTML 웹 문서에서 원하는 정보를 추출하기 위해 문서를 분석, 규칙을 생성하여 생성된 규칙에 의해 문서의 구조를 인식하고 자동으로 온톨로지를 구축한다[12,13].

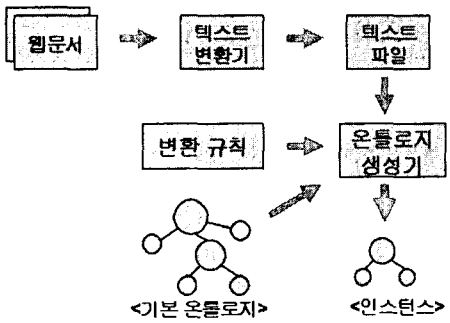

그림 1. 규칙기반 도메인 온톨로지 생성 흐름도

본 연구에서는 여행과 관련된 도메인으로 범위를 한정, 규칙기반 방법을 이용하여 비구조화된 문서에서 숙박 정보 온톨로지에 필요한 정보를 자동으로 추출한다.

기본 온톨로지에는 숙박정보 온톨로지의 클래스들과 속성들이 정의되어 있다. 이를 바탕으로 관련된 규칙과 파싱된 내용을 이용하여 HTML 웹 문서를 구조화된 문서로 변경한 후 온톨로지 생성기에 의해 인스턴스 파일을 자동으로 생성한다. 그림 1은 규칙기반 온톨로지 생성 흐름도이다.

제주 숙박 정보 중 비구조화된 문서인 호텔의 HTML 문서를 입력으로 받아 객실명, 객실가격, 기타등 정보를 추출하여 텍스트 파일을 생성한다. 본 연구에서 사용한 호텔 문서 관련 규칙의 예는 다음과 같다.

- ○ 객실명은 접미사로 '스위트', 접두사로 '온돌'을 가지며, 스탠다드형(또는 일반 객실)의 객실명으로는 '디럭스', '럭셔리', '스탠다드', '슈페리어' 등의 공통 규칙을 갖는다.
- ○ 가격은 접미사가 '원'또는 'Won'이며, 특1등급의 호텔인 경우 300,000원 이상의 가격을 갖는다.
- ○ 기타 사항으로 접미사가 '전망(뷰)'인 경우, '산전망' 과 '바다전망'으로 규칙을 갖는다.
- ○ 영역을 제주로 한정할 경우 주소는 '제주도'로 시작한다.
- ○ 위치하는 지역은 주소의 두 번째 단어에 의해 알 수 있으며, 전화번호는 '(Tel)' 또는 'TEL'로 시작하여 3, 4개의 숫자와 -(하이픈), 그리고 4개의 숫자가 결합한 형태를 갖는다.

이처럼 제안하는 방법이 실생활에 적용할 수 있음을 확인하기 위하여 제주도 숙박 정보에 대한 온톨로지를 OWL을 이용하여 생성한다. 생성된 온톨로지를 이용함으로써 의미적 정보검색이 가능하며 이로 인해 사용자에게 양질의 정보를 빠르고 정확하게 제공할 수 있다. 또한, 비구조화된 문서를 규칙에 기반하여 정보를 추출하고 구조화된 문서로 변경, 온톨로지를 반자동으로 구축하는 방안에 대하여 제안하고자 한다.

3.2 숙박 정보 온톨로지

그림 2는 숙박 정보 온톨로지의 구조를 그림으로 표현한 것이다. 호텔, 펜션, 콘도 등의 클래스 및 숙박 정보와 관련된 몇 가지 속성들로 구성된다. 그림 3은 실제로 OWL을 이용하여 작성한 숙박 정보 온톨로지 구조를 보여준다.

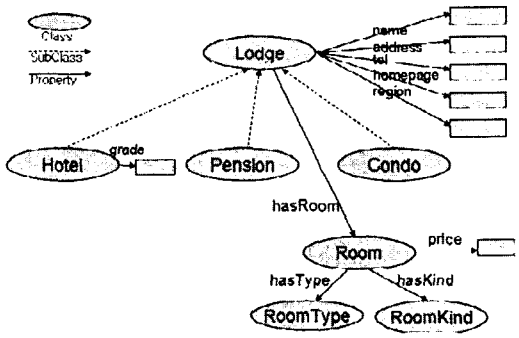

그림 2. 숙박 정보 온톨로지 구조도

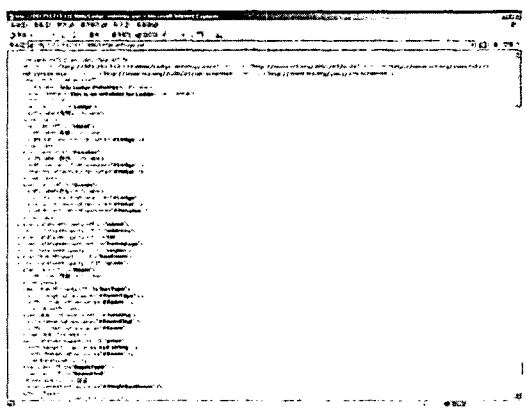

그림 3. 숙박 정보 온톨로지의 예

4. 구현 및 테스트

온톨로지의 생성은 Jena API를 이용하여 OWL 온톨로지를 생성하는 프로그램을 구현하였다. 온톨로지 인스턴스 생성시 규칙을 기반으로 그림 1의 기본 온톨로지에 정의된 구조로 자동으로 생성하였다.

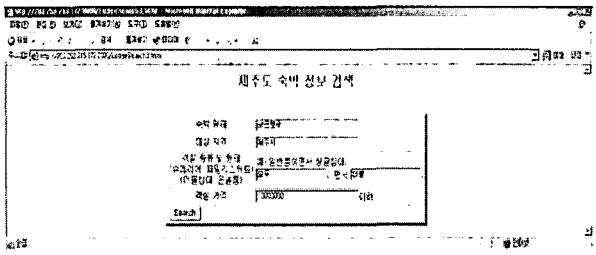

그림 4. 온톨로지를 이용한 정보 검색의 예

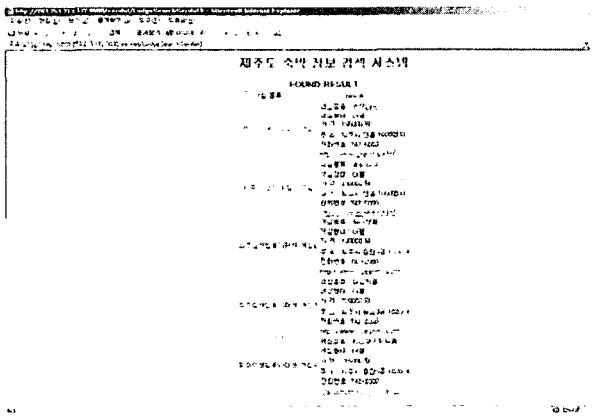

그림 5. 온톨로지 기반 검색 결과

그림 4는 정보 검색 화면의 예로서 사용자가 "제주도 숙박업소 중 제주시에 있으며 객실 가격이 1,000,000원이하의 더블 침대를 가진 객실은?" 이라는 질의를 요청하는 모습이다.

그림 5는 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 스탠다드, 일반, StandardRoom을 같은 의미로 이해하여 정보를 검색할 수 있다.

4. 결론

본 논문에서는 비구조화된 문서에서 규칙을 발견하고 이를 이용하여 기본 온톨로지에서 정의한 구조를 갖는 온톨로지 인스턴스를 자동으로 구축하는 방법에 대하여 제안하였다. 기존 웹 페이지에 대한 온톨로지를 수작업이 아닌 자동으로 구축함으로써 시간과 비용을 획기적으로 줄일 수 있다. 또한 생성된 온톨로지를 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 온톨로지를 자동 생성하여 이용함으로써 양질의 정보를 효율적으로 검색할 수 있음을 알 수 있었다.

5. 참고문헌

[1]최호섭, 옥철영, "정보검색 시스템과 온톨로지," 정보과학회지, 제22권, 제4호, pp.62-71, 2004.

[2]Deborah L. McGuinness, Frank van Harmelen, "OWL Web Ontology Language Overview," http://www.w3.org/TR/owl-features/, 2004.

[3]Michael K. Smith, "OWL Web Ontology Language Guide," http://www.w3.org/TR/owl-guide/, 2004.

[4]박재홍, 임유정, 김도완, 박찬규, 조현규, "Semantic Web 환경에서의 자원발견," 정보처리학회 2002년 추계학술대회, pp.0713-0716, 2002.

[5]하상범, 박영택, "온톨로지를 통한 추론형 시멘틱 검색 시스템에 관한 연구," 정보과학회 2004년 춘계학술대회, 제31권, 제1호, pp.0625-0627, 2004.

[6]정은경, 김영민, 변영철, 이상준, "온톨로지 기반의 정보검색," 정보과학회 2003년 추계학술대회, pp.0121-0123, 2003.

[7]Protégé-2000, http://protege.stanford.edu/index.html/.

[8]OilEd, http://oiled.man.ac.uk/.

[9]OntoEdit, http://www.ontoprise.de/home_en/.

[10]임수연, 송무희, 이상조, "전문용어의 처리에 의한 도메인 온톨로지의 구축," 한국정보과학회 논문지 B, 제31권, 제3호, pp.353-360, 2004.

[11]이현실, 이두영, "온톨로지 기반 한의학 처방 지식관리시스템 설계에 관한 연구," 한국정보관리학회지, pp.341-371, 2003.

[12]김재훈, "정보추출의 기술 현황," 정보과학회지, pp.0035-0046, 2004.

[13]노태길, 이상조, "규칙 기반의 기계학습을 통한 고유명사의 추출과 분류," 정보과학회 2000년 추계학술대회, pp.0170-0172, 2000.