퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 3, pp. 291-296

도메인지식의 계층화를 통한 온톨로지 인스턴스의 속성정보 추출

An Extraction of Property of Ontology Instance Using Stratification of Domain Knowledge

장문수* · +강선미**

Moon-Soo Chang and Sun-Mee Kang

* 서경대학교 소프트웨어학과

** 서경대학교 컴퓨터과학과

요 약

최근에 여러 분야에서 구축되고 있는 온톨로지는 기계가 이해할 수 있는 지식을 축적하는 것을 목표로 하고 있다. 기계가 온톨로지를 이용하여 정보의 관리 및 해석을 스스로 하는 것이 가능할 것으로 본다. 본 논문에서는 온톨로지의 인스턴스를 구성하는 속성을 기존 웹 문서의 구조정보로부터 추출하는 알고리즘을 제안하였다. 특히, 속성 정보로 구성하는 도메인 지식을 계층화함으로써 속성 추출 알고리즘을 개선하고, 추출 결과의 품질을 향상시킨다. 2만 문서를 대상으로 제안된 알고리즘을 적용한 결과 약 83%의 신뢰도의 속성 정보를 추출할 수 있었다.

키워드 : 온톨로지, 인스턴스, 도메인 지식, 속성, 시맨틱 웹

Abstract

The ontology has been used widely in recent years with its aim to accumulate knowledge that machine can comprehend. We believe that machine can manage and analyze information on its own using the ontology. In this paper, we propose an algorithm that allows us to extract properties of ontology instances from structured information already existing in web documents. In particular, by stratification of the domain knowledge that is composed of property information, we were able to make the algorithm better and improve the quality of extraction results. In our experiments with 20 thousands targeted documents, we were able to extract property information with 83% confidence.

Key Words : Ontology, Instance, Domain knowledge, Property, Semantic web

1. 서 론

최근의 웹정보 사회는 수억 페이지 이상 축적된 웹문서와 언론사나 기타 정보제공회사로부터 하루에도 수없이 쏟아지는 다이나믹 웹문서, 그리고 수많은 사용자 스스로 생산해내는 블로그 문서 등, 개인이 처리하기에는 불가능에 가까울 정도의 정보의 홍수 사회가 되었다. 정보 검색은 사용자에게 필요한 정보를 찾아주는(searching) 시대에서 사용자가 원하는 정보를 걸러주는(filtering) 시대가 되었다. 정보 필터링 기술과 정보 추출(information extraction) 기술이 점차로 발전하고 있지만 정보의 양이 확대되는 속도를 따라가지 못하고 있는 실정이다.

이러한 상황의 원인으로 정보량의 폭발적인 증가 외에 현재의 HTML 기반의 웹환경을 들 수 있다. HTML은 정보의 가시적인 표현만 기술할 뿐으로, 정보의 의미를 구조화 혹은 체계화하는 기능이 없기 때문에 웹문서로 표현된 정보의 해석이나 활용은 순전히 사람에 의해 수행될 수밖에 없다.

이런 상황을 해결하는 방법으로 W3C에서는 차세대 웹환경으로 기계가 이해할 수 있는 의미 구조를 가진 시맨틱 웹을 제안하고 있다[1]. 그리고 이 의미 구조를 표현하는 XML 기반의 온톨로지를 또한 제안하고 있다.

시맨틱 웹이 기존 웹을 대체하기 위해서는 컨텐츠 생성의 용이성, 온톨로지의 공유, 기존 웹기술과의 공존, 기존 웹정보의 전이 등의 과제들이 해결되어야 한다. 기존의 수많은 웹정보를 시맨틱 웹 환경으로 이식하는 문제나 모든 서비스에 활용가능한 온톨로지의 구축 문제는 기존 웹정보를 활용하지 않고는 해결할 수 없다. 온톨로지를 구축하기 위해서는 온톨로지를 구성하는 개념을 정의해야 하는데, 소수의 전문가에 의해서 모든 사물을 정확하고 세밀하게 개념화할 수는 없다.

예를 들어, 휴대폰을 개념화하기 위해서는 휴대폰의 속성을 정의해야 하고, 이것을 휴대폰 관련 서비스에 사용하기 위하여 휴대폰 속성("제품명", "색상" 등)에 대한 속성값

접수일자 : 2007년 5월 18일
완료일자 : 2007년 5월 31일
감사의 글 : 본 논문은 정통부 및 정보통신연구진흥원의 정보통신 선도기반 기술개발사업의 연구결과로 수행되었습니다.
+ : 교신저자

('SCH-B630", "실버" 등)을 구축해야 한다. 여기에는 각각에 대한 문제점이 존재하는데, 하나는 휴대폰에 대한 모든 속성을 인위적으로 정의할 수 없다는 것이고, 또 하나는 현재 웹 환경에서는 시맨틱 웹의 형태로 제공되는 정보가 없기 때문에 속성에 대한 속성값을 의미적으로 추출할 수 없다는 것이다. 따라서 온톨로지를 실제 서비스에서 활용하는 온톨로지 인스턴스의 속성정보를 기존 웹정보로부터 자동으로 추출하는 방법에 대한 연구가 요구될 수 밖에 없다.

본 논문에서는 기존 웹문서 중에서 <table>과 같이 구조화된 정보로부터 온톨로지 인스턴스 정보를 추출하는 알고리즘을 제안한다. 제안하는 알고리즘에서는 온톨로지의 속성 정보로 구성되는 도메인 지식을 사용하는데, 이 도메인 지식을 계층화함으로써 속성 정보 추출 알고리즘을 개선하고자 한다.

2장에서는 관련 연구로 온톨로지와 정보추출에 대해서 설명하고, 3장에서는 본 논문에서 제안하는 웹문서의 구조정보를 추출하는 알고리즘을 제시한다. 4장에서는 제안하는 방법에 대한 실험과 실험결과에 대한 분석을 기술한다.

2. 관련 연구

2.1 온톨로지 연구

온톨로지는 한 사회에서 통용되는 개념과 그 개념들 간의 관계를 정의함으로써 그 사회에서 축적되는 지식을 표현하고 활용하는 도구로서 정의되어 왔으며, 최근에는 컴퓨터를 이용한 지식의 축적과 활용의 도구로서 2000년대 이후로 많은 연구가 진행되고 있는 분야이다. 온톨로지는 개념들의 집합으로서, 하나의 개념은 개념의 정의와 그 개념의 특징을 표현하는 속성, 다른 개념과의 관계를 나타내는 개념관계로 표현된다. 개념의 정의는 사람을 위한 표현이고, 기계는 속성과 관계를 통하여 개념을 이해한다. 기계가 온톨로지를 이용하여 문서로부터 의미를 직접 이해함으로써 사람의 판단이나 추가적인 정보 입력없이 고차원의 서비스가 가능하게 된다. 따라서, 온톨로지 관련 연구는 속성과 관계의 구축방법에 관한 연구가 활발히 진행되고 있다[2-5].

온톨로지는 시맨틱 웹의 표준을 정하는 W3C에서 권고하는 OWL로 표기하는 경우가 많으며, 구축 도구로는 자바기반의 Protege 등이 있다. 온톨로지의 상위 개념에 대한 구축은 SUMO[1]를 비롯하여 국내외적으로 많은 연구가 이루어져 왔다[2,4,6]. 개념에 관한 연구는 기존 시소러스나 WordNet과 같은 개념망 연구에서 많은 결과물을 내고 있었기 때문에, 국내에서는 이러한 연구로부터 온톨로지 연구가 발전하여 계승되는 경향도 있다[7-8].

최근 국내에서는 국가 정보 인프라 구축의 하나로 온톨로지 구축 사업을 진행하고 있다. 특히, 국가 IT 온톨로지 구축 사업[2]은 온톨로지의 개념의 정립과 함께 인스턴스의 구축, 서비스 모델 개발을 동시에 진행함으로써, 방대한 IT분야의 지식을 온톨로지로 체계화함과 동시에 차세대 웹 시장의 선점을 목표로 하고 있다.

2.2 정보 추출 연구

웹문서로부터 정보를 추출하는 기술은 웹정보의 양이 방대해지고 푸쉬 서비스(push service)[2]에 대한 요구가 커져감에 따라 발전하기 시작하였다. 기존 텍스트 문서에서는 그 요구가 크지 않았던 것에 비하여 웹문서에 대한 정보 추출의 요구가 커진 것은 정보의 디지털화, 방대한 정보량과 함께 규격화된 정보의 양이 급속도로 늘고 있는 것이 그 원인으로 볼 수 있다.

정보 추출 기술은 크게 두가지 분야로 나눌 수 있는데, 하나는 언어처리 기술을 활용한 문장 분석을 통한 정보 추출이고, 나머지 하나는 정보 패턴 분석을 통한 정보 추출이다[9]. 전자의 기술은 기존 텍스트 기반의 전통적인 기술로 품사나 어휘 정보 등의 언어 자질에 대한 연구가 요구되어 복잡한 시스템 구성을 필요로 한다. 후자는 웹문서와 같이 구조적인 정보가 포함된 문서에 대해서 형태적 혹은 통사적(syntactic) 정보를 활용하여 비교적 간단한 알고리즘으로 구현이 가능하다. 이러한 알고리즘을 래퍼(wrapper)라고 하며 웹정보 추출에 많이 사용되고 있다.

래퍼 기술의 문제점은 추출 대상인 웹문서가 항상 변경될 가능성이 있다는 데에 있다. 문서가 변경되면 래퍼에 적용한 패턴이 달라지게 되고 따라서 래퍼는 제기능을 잃게 되어 새로운 래퍼를 개발해야 한다. 또한, 정보를 제공하는 웹사이트마다 같은 종류의 정보에 대해서 다른 패턴으로 문서를 구성하고 있기 때문에 추출 대상마다 각각의 래퍼 패턴을 구비해야하는 문제점도 있다[9]. 이것을 해결하는 방법으로 기계학습 기술이 도입되어 패턴에 대한 통계적인 정보를 이용하여 보편적인 래퍼를 구성하는 방법들이 제안되었다[10-12]. 그러나 이 방법은 강인한 래퍼가 구현되기는 하나 학습문서의 정보량에 따라 결과에 대한 신뢰도가 기존 방법에 비하여 떨어질 수 있다는 문제점을 안고 있다.

3. 웹문서의 구조정보 추출

온톨로지가 시맨틱 웹 환경에서 실제 서비스에 활용되기 위해서는 온톨로지의 인스턴스 정보가 웹 상에 존재해야 한다. 또한, 온톨로지의 개념을 구성하는 요소로서 개념의 특징을 표현하는 속성(property)은 그 개념에 대한 완전한 이해를 통해 정의될 수 있다. 이러한 요구는 HTML로 표현된 기존 웹으로부터 정보를 추출함으로써 해결될 수 있다. 특히, IT분야의 웹정보는 상품을 비롯하여 많은 관련 문서가 구조화된 형태로 제공되고 있기 때문에 비교적 손쉽게 정보 추출이 가능하다.

본 논문에서는 의미 태깅이 되어 있지 않은 HTML문서에서 구조화가 되어 있거나 일부 구조화가 이루어져 있는 정보에 대해서 도메인 지식을 활용하는 래퍼 기능으로 정보를 추출하는 알고리즘을 제시한다. 그리고, 구조화된 정보에 기술되어 있는 내용이 대부분 인스턴스의 속성 정보라는 점에 착안하여 도메인 지식을 속성 정보로 구성하고, 지식의 활용도를 높이기 위하여 이것을 계층화한다.

3.1 구조정보 추출 과정

HTML문서에서는 <table> 태그를 이용하여 정보를 구조화한다. 현재 대부분의 포털 사이트를 비롯하여 상품 소개 정보나 프로야구 경기 결과 등과 같이 수없이 많이 생성되는

1) Suggested Upper Merged Ontology (SUMO): Teknowledge사(http://ontology.teknowledge.com)에서 개발 중인 상위 온톨로지로서 685가의 클래스로 구성되어 있다.

2) 푸쉬 정보 서비스는 사용자가 요구하는 정보만을 골라서 자동으로 일정한 시간 간격으로 사용자에게 배달하는 서비스를 말하는 것으로, 사용자가 정보를 찾아다니는 기존의 정보 서비스와 반대되는 개념이다.

유사한 내용의 정보를 표현하기 위하여 정보제공 사이트는 <table> 태그를 이용하여 표현하는 것이 일반화되어 있다. 본 논문에서는 HTML문서에서 테이블 구조를 찾아내어 원하는 정보 형태를 파악하고 구조에 맞추어 정보를 추출한다. 그림 1은 구조정보를 추출하는 과정을 나타내고 있다.

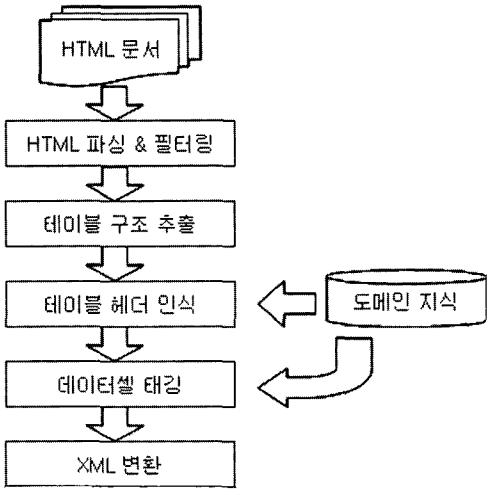


그림 1. 구조정보 추출과정.
Fig.1. Extraction flow for structured information.

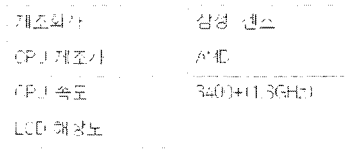
HTML 입력문서로부터 필요한 정보를 추출하는 과정에서 우선 문서의 구조를 파악하기 위하여 HTML 파싱(parsing)을 수행한다. HTML 문서의 파싱 결과로부터 HTML 태그의 기능을 파악할 수 있으므로 테이블 정보만 추출할 수 있게 된다. 본 논문에서는 필요한 온톨로지 인스턴스 관련 정보를 제외한 부분, 즉 이미지 정보나 텍스트 정렬 및 디스플레이를 위한 태그, 그리고 텍스트의 내용과 관련 없는 자바스크립트 코드 등을 정보추출 이전에 필터링한다.

또한, 정보 추출 시에 텍스트 처리 오류를 유발시키는 특수문자들을 사전에 처리가 가능한 일반 문자로 치환하거나 치환이 불가능한 것은 제거한다. 이러한 과정을 묶어서 정보추출 전처리라고 한다. 전처리는 정보 추출에 필요한 테이블 부분을 제외한 모든 저해요소를 처리하는 과정이다.

전처리를 통해 정제된 HTML 문서로부터 필요한 테이블을 찾아서 분리하기 위해 테이블 구조를 추출한다. 그리고 추출된 테이블이 어떤 형태를 가지는지 분석한다. 테이블의 형태는 테이블의 헤더의 위치에 따라 상위 헤더 타입과 좌측 헤더 타입으로 나눌 수 있다. 테이블의 헤더 정보는 인스턴스의 속성에 해당하고 헤더에 따라오는 나머지 데이터는 인스턴스의 속성값에 해당한다. 따라서 헤더의 위치를 정확히 파악해야만 속성과 속성값을 추출할 수 있다.

본 논문에서는 헤더를 인식하기 위하여 테이블의 각 셀(<td>...</td>)들을 온톨로지 인스턴스의 속성으로 구성된 도메인 지식과 비교하여 속성이 있는 셀들을 구별한다. 그림 2는 헤더 타입에 따른 테이블의 형태를 비교하고 있다. 하나의 상품이나 회사와 같은 인스턴스의 사양에 대한 테이블은 좌측 헤더 타입이 대부분이고, 여러 인스턴스들의 정보를 나열하는 리스트 형식의 테이블은 상위 헤더 타입으로 구성되어 있다.

정보 추출 타깃 테이블에는 속성인 헤더와 함께 각각의 속성값(value)이 기록되어 있다. 본 논문에서는 온톨로지의 속성을 정의하는 것과 온톨로지 활용 서비스를 위한 기존 웹 정보의 전이를 동시에 목표로 하고 있기 때문에 속성과 속성값 모두를 추출대상으로 한다. 그래서 이들을 묶어서 (속성, 속성값) 형태의 데이터 셀로 정의한다. 데이터 셀 태깅 모듈에서는 이 결합 형태를 추출하여 저장하는데, 이 과정에서 속성과 속성값으로 추출되지 않는 부분은 모두 제거되고 필요한 정보만 남게 된다.



(a) 좌측 헤드 타입 테이블
(a) Left-header type table

품목↑	등록	업체	최저가
☐ 삼성 센스 NT-G10/MS340	3월10일	35	842,000
☐ 삼성 센스 NT-G15A/M171	7월03일	83	911,000
☐ 삼성 센스 NT-G15A/Y160	7월03일	77	1,022,000

(b) 상위 헤드 타입 테이블
(b) Top-header type table

그림 2. 헤드 타입에 따른 테이블 형태 비교.
Fig. 2. Comparison of tables by header type.

마지막으로, 시맨틱 웹과 온톨로지는 XML을 기반으로 하고 있기 때문에, 추출된 속성 정보를 XML 형태로 변환한다. 본 논문에서는 데이터 셀들을 모두 추출하여 임의의 형태로 태깅한 후, 간단한 XML 태그를 정의하여 데이터 셀 리스트를 XML 형태로 변환한다. 그림 3은 XML로 변환된 속성 추출 결과물을 나타내고 있다.

```
<?xml version="1.0" encoding="euc-kr" ?>
- <coreonto-instance>
  - <Item domain="차세대PC" filepath="D:\ETRI\Data\정리(10.31)\HTML\I-차세대PC
      <property name="MODEL">R65/W183</property>
      <property name="CPU">Intel Core Duo T2400(1.83GHz)</property>
      <property name="RAM">1GB DDR2 667 SDRAM(최대 2GB)</property>
      <property name="L2 CACHE">2MB</property>
      <property name="LCD">15인치 TFT</property>
      <property name="RESOLUTION">SXGA+(1,400 x 1,050)</property>
      <property name="VGA">nVIDIA GeForce 7400</property>
      <property name="VRAM">128MB(터보캐쉬 256MB)</property>
      <property name="AUDIO">HD Audio, 스테레오 스피커</property>
      <property name="HDD">80GB(S-ATA, 5,400rpm)</property>
      <property name="OPTICAL DRIVE">DVD + CD-RW 콤보</property>
      <property name="POINTING DEVICE">터치패드</property>
      <property name="MODEM">56Kbps</property>
      <property name="LAN">10/100Mbps Ethernet</property>
      <property name="WIRELESS LAN">802.11a/b/g, 블루투스, IrDA</property>
      <property name="SLOT">Type II x 1 PCMCIA, Express Card x 1, 6-in-1 멀티 카
```

그림 3. XML로 표현된 속성 추출 결과물.
Fig. 3. Extraction result of property formed by XML.

3.2 도메인 지식의 활용

본 논문에서는 웹문서의 테이블 타입을 판단하거나 데이터 셀을 태깅할 때 도메인 지식을 사용한다. 대부분의 테이블은 헤더를 가지고 있는데, 이 헤더는 그 아래에 나오는 데이터의 종류나 특징을 나타내는데 사용된다. 온톨로지의 인스턴스 정보를 가지고 있는 테이블은 그 인스턴스의 속성에 관한 정보를 표현하는 경우가 일반적이므로 테이블의 헤더는 인스턴스의 속성이 될 가능성이 매우 높다. 본 논문에서는 테이블 구조나 정보추출을 위하여 헤더인식에 사용하는 도메인 지식을 구성하기 위하여 인스턴스의 속성 후보를 사용한다. 즉, 도메인 지식은 온톨로지 인스턴스를 구축하고자 하는 대상 분야에서 의미를 가지는 용어들을 수집하여 지식베이스

형태로 저장하게 된다.

3.2.1 도메인의 분류

온톨로지의 속성은 "제품명"과 같이 많은 클래스에서 공통적으로 존재하는 속성도 있지만, 각각의 클래스마다 고유의 속성이 존재하는 경우도 많다. 특히 같은 도메인이라도 IT분야와 같이 범위가 광범위할 경우에는 각각의 온톨로지 클래스의 특징이 다른 경우가 많다. 따라서 하나의 도메인 지식으로 도메인에 속하는 모든 온톨로지 클래스나 인스턴스에 적용할 경우 오류를 유발할 가능성이 매우 높다. 본 논문에서는 도메인을 여러 가지 방향으로 세분화하여 이를 토대로 도메인 지식을 계층화함으로써 이런 문제를 해결하고자 한다.

본 논문에서는 도메인 지식의 계층을 나누기 위하여 다음 세가지 분류 기준을 제시한다.

IT839에서 제시하는 IT 11개 분야

IT온톨로지 클래스의 종류

상품 분류 카테고리

IT 11개 분야는 차세대 이동통신, RFID/USN, 디지털 TV/방송, 차세대PC 등 IT839에서 제시하는 유망 IT분야를 말한다. IT 온톨로지 클래스의 종류는 정보통신부의 국가 IT 온톨로지 인프라 기술개발 과제[2]에서 1차적으로 구축한 IT 분야 온톨로지의 클래스를 분석하여 클래스의 유형을 나눈 것으로서, 상품, 인물, 회사, 도서, 논문, 장비/부품, 기술 등이 있다. 여기서 IT분야의 특성에 따라 구조화된 웹정보의 대부분은 IT 상품으로 분류된다. 따라서 본 논문에서는 다양한 IT 상품을 분류하기 위하여 가격비교 웹사이트에서 제공하고 있는 상품 카테고리를 도메인 지식의 분류 기준으로 사용한다.

다양하게 수집한 웹문서를 앞에서 제시한 세 가지 분류 기준으로 나누고 해당 문서로부터 추출되는 키워드를 속성 후보로 도메인 지식에 포함시킨다.

3.2.2 속성 후보의 빈도수

HTML문서에는 의미정보가 없기 때문에 테이블 내에 속성과 속성값이 혼재되어 나타난다. 기본적으로는 테이블의 행과 열의 구분으로 속성과 속성값을 구별한다. 그러나 복잡한 테이블 구조에서는 테이블 헤더가 두 개 이상의 행이나 열로 구성되기도 하고, 테이블의 일부가 두 개 이상의 셀이 병합되어 나타나기도 하기 때문에 행과 열 정보만으로는 속성 영역을 정확하게 구분할 수 없다. 본 논문에서는 이러한 구조정보와 함께 속성 후보의 빈도수를 이용하여 속성을 추출한다. 특정 도메인에서 나타나는 속성은 웹사이트에 관계없이 대체로 비슷하다. 따라서, 테이블에 나타나는 모든 정보의 빈도수를 측정하여 고빈도의 정보를 속성후보로 선정한다.

그리고, 구조정보 추출과정 중 데이터셀 태깅에서 속성값을 특정 속성과 매칭시키기 위해서는 신뢰성이 있는 속성명이 필요하다. 빈도수에 의해 선정된 속성후보에는 정보를 제공하는 웹사이트에 따라서 같은 속성이면서 다른 이름을 사용하는 경우가 상당수 포함된다. 본 논문에서는 유사속성을 묶어주고 최빈도수의 속성후보를 대표속성으로 선정하여 속성 태깅에 사용한다.

그림 4는 빈도수를 이용한 속성 후보 추출 알고리즘을 나타낸다.

1. 도메인의 세부계층별로 후보문서 선정
2. 후보문서의 테이블 정보의 빈도수 측정
3. 유사한 속성후보 그룹화
4. 유사 속성 중 최빈도 속성을 대표속성으로 선정

그림 4. 속성 선정 알고리즘.
Fig. 4. Property selection algorithm.

3.2.3 도메인 지식의 구조

본 논문에서 제안하는 구조정보 추출과정에서 도메인 지식은 여러 부분에서 활용되고 있으며 가장 중요한 목적은 웹정보로부터 온톨로지의 속성을 태깅함으로써 온톨로지 인스턴스 생성을 유도하는 것이다. 이러한 목적들에 맞게 도메인 지식을 활용하기 위하여 본 논문에서는 그림 5와 같은 구조로 도메인 지식을 구성한다.

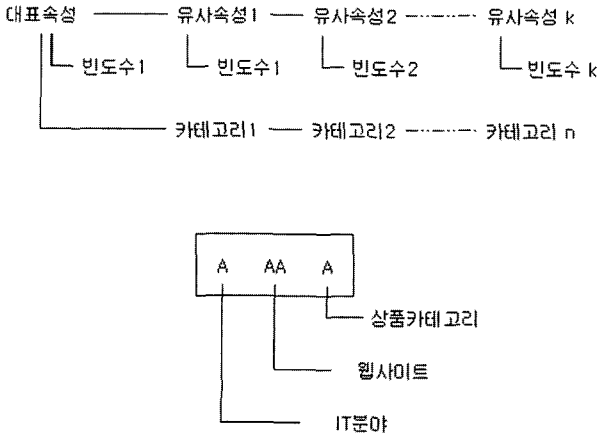


그림 5. 도메인 지식의 구조도.
Fig. 5. Structure diagram of domain knowledge.

유사속성 리스트와 대표속성을 링크함으로써 데이터 셀 태깅에서 속성태깅은 도메인 지식의 유사속성 검색을 통하여 대표속성으로 태깅이 이루어진다. 카테고리는 IT분야, 웹사이트, 상품카테고리를 가지고 네 개의 영문자로 코딩되어 있다. 카테고리 리스트는 대표속성 및 유사속성이 출현하는 웹문서의 카테고리를 연결하고 있다. 이것은 문서의 세부 도메인이 정해지면 해당 도메인에 속하는 속성만 참조할 수 있도록 한다. 즉, 휴대폰에 관한 문서는 휴대폰과 관련된 카테고리가 연결된 속성만 도메인 지식으로 참조되어 테이블 헤더 인식 및 속성 태깅에 사용된다. 도메인의 범위를 좁혀서 차세대 이동통신 분야에서 다나와 웹사이트에서 제공하는 문서의 속성만을 원할 경우에는 해당 카테고리의 속성만 대상이 될 수 있다.

4. 실험 및 분석

4.1 실험 배경

온톨로지 인스턴스 구축에 사용되는 속성 정보를 추출하기 위한 실험대상으로 본 논문에서는 IT839에서 제시하는 IT분야 문서를 검토하여 문서의 양이 충분하고 구조정보가 비교적 안정적으로 제공되는 7개 분야에 대해서 10개의 웹사이트로부터 웹문서를 수집하였다. 실험을 위한 도메인의 범

위를 줄이기 위하여 IT분야 온톨로지 클래스의 종류 중에서 상품에 관한 클래스만을 대상으로 하였다.

문서 수집은 카테고리별로 이루어졌으며 수집된 문서수는 카테고리 중복을 포함하여 2만 문서이다. 수집된 문서에 대하여 본 논문에서 제안하는 구조정보 추출 알고리즘을 수행한 결과 18,000개 정도의 문서가 처리되어 90%의 처리 성공률을 나타내었다. 오류로 인하여 처리가 안된 문서는 문서 자체의 HTML문법 오류와 처리모듈에서 지원하지 못하는 테이블 구조로 인하여 결과를 내지 못한 문서들이었다. 그러나, 이 결과는 처리 대상을 선정할 때 처리가능한 범위의 웹사이트로 제한하여 문서를 수집하였을 뿐만 아니라, 한편으로 처리 모듈의 개선을 통하여 성공률을 올릴 수 있기 때문에 본 논문에 있어서는 큰 의미를 가지지 않는다.

4.2 속성 추출 신뢰도 실험

본 논문에서는 도메인 지식의 활용에 따른 속성 추출 결과를 알아보기 위하여 처리가 완료된 문서 중에서 무작위로 문서를 선정하여 미리 속성 추출 교육을 받은 숙련자에 의해 수작업으로 속성 추출 정확도를 측정하였다. 측정 방법은 원문서에 존재하는 속성이 추출 실험 후 결과 파일에 나타나지 않는 경우를 오류로 정의하고, 총 속성의 개수에서 추출 성공한 속성개수의 비율로 속성 추출 신뢰도로 표시하였다. 실험은 15개 카테고리에 대해서 10개씩 문서를 선정하여 실시하였으며, 총 1573개 속성이 출현하여 272개의 속성이 추출되지 못하였다. 그 결과로 표 1에 나타난 바와 같이 평균 83%의 신뢰도를 얻을 수 있었다.

표 1. 속성 추출 실험 결과.
Table 1. The result of property extraction experiment.

사이트 정보	속성수	속성 오류	신뢰도
다나와네트워크	49	1	98%
다나와무선랜	58	15	74%
다나와셋탑박스	90	57	37%
다나와영상장비	70	22	69%
다나와인터넷폰	59	17	71%
다나와캠코더	106	15	86%
다나와DMB	124	67	46%
다나와DTV	180	33	82%
다나와Mobile	205	11	95%
디스플레이뱅크	113	3	97%
마이마진인터넷폰	62	14	77%
마이마진휴대폰	150	8	95%
뭉클	80	0	100%
세티즌	101	2	98%
투데이스PPC	126	7	94%
전체	1573	272	83%

속성 오류는 카테고리 별로 일정하지 않은 분포로 나타났다. 뭉클 사이트[13]의 경우에 오류가 하나도 없는 반면, 다나와 사이트[14]의 셋탑박스 카테고리에서는 90개중 57개의 속성이 추출되지 못하여 신뢰도가 40%에도 미치지 못하였다. 같은 다나와 사이트에서도 네트워크 카테고리에서는 단 한 개의 오류만 발생하여 한 사이트 내에서도 카테고리에 따라 신뢰도가 다르게 나왔다.

오류를 분석해 본 결과, 오류가 적은 페이지의 정보는 해당 사이트에서 수집하여 데이터베이스화한 정보인데 반하여, 오류를 많이 발생시키는 카테고리의 문서들은 개별 상품 제조사에서 제공하는 정보를 그대로 테이블에 포함하는 경우가 많았다. 이런 경우 제조사 특유의 속성이 상당수 포함되어 속성의 빈도수가 매우 낮게 나와 도메인 지식에서 속성으로 등록되지 못하게 됨으로써 추출이 불가능하게 되었다. 따라서 이러한 통계상의 오류를 보완하는 알고리즘에 대한 후속 연구가 필요함을 알 수 있었다.

5. 결 론

시맨틱 웹에서 기계가 이해할 수 있는 의미표현 등의 목적으로 온톨로지가 여러 분야에서 구축되고 있다. 본 논문에서는 온톨로지의 인스턴스를 구성하는 속성 정보를 기존 웹문서의 구조정보로부터 추출하고자 하였다. 이를 위하여 우리는 속성 정보로 구성된 도메인 지식을 계층화하여 구성하였으며, 이것을 활용하여 웹문서의 구조정보의 형태를 파악하고 정확한 속성과 속성값을 추출하고자 하였다. 제안한 아이디어를 검증하기 위하여 본 논문에서는 2만여 개의 웹문서를 수집하여 속성을 추출하는 실험을 수행하였다. 그 중 일부 결과물을 정밀하게 수작업으로 확인한 결과 약 83%의 신뢰도로 속성 정보를 추출할 수 있었다. 실험 대상 분야가 급변하고 있는 IT분야이고 웹사이트에서 제공하는 구조 데이터의 신뢰성이 다소 떨어지기 때문에 90%이상의 좋은 결과를 얻지는 못하였다.

따라서 향후 개선할 점으로는 적은 빈도수로 출현하는 속성을 도메인 지식으로 추출하는 알고리즘를 개발해야 하고, HTML의 문법적인 오류나 javascript 등의 부가적인 코드에 영향을 받지 않는 알고리즘을 개발해야 한다. 또한, 본 논문에서는 비교적 명확한 테이블 구조를 대상으로 실험을 하였으나, 많은 사이트들이 HTML문서에서 비구조적인 서술형태의 정보를 테이블 구조에 포함시키고 있다. 따라서 속성 추출의 대상 문서의 범위를 넓히기 위해서는 비구조적인 형태의 정보로부터 속성 정보를 추출하는 알고리즘이 보완될 필요가 있다.

참 고 문 헌

[1] 김중태, 시맨틱 웹, 디지털미디어리서치, 2006.
[2] 김재호, 신지애, 최기선, "국가 IT 온톨로지 구축", 한국정보과학회 가을 학술발표논문집, 제33권, 2(B)호, pp. 16-19, 2006.
[3] 구미숙, 황정희 ,류근호, 홍장의, "데이터마이닝 기법을 이용한 XML 문서의 온톨로지 반자동 생성", 정보처리학회논문지D, 제13권, 3호, pp. 299-308, 2006.
[4] 조이현, 박대원, 박동훈, 문홍구, 권혁철, "비전문가에 의한 상하위 관계 중심의 온톨로지 공동구축 방법", 한국지능정보시스템학회 2006년 추계학술대회 논문집, pp. 87-91, 2006.
[5] 최정화, 박영택, "의미 중의성을 고려한 온톨로지 기반 메타데이터의 자동 생성", 정보과학회 논문지:

소프트웨어 및 응용, 제33권, 11호, pp. 986-998, 2006.
[6] 강인수, 정한민, 이승우, 김평, 성원경, "국가과학기술 R&D 기반정보 온톨로지", 한국콘텐츠학회 2006년 춘계종합학술대회 논문집, 제4권, 1호, pp. 231-234, 2006.
[7] 최호섭, 임지희, 배영준, 최수일, 옥철영, "온톨로지 구축 방법과 사례", 정보과학회지 제24권, 4호, pp. 31-44, 2006.
[8] 한성국, 이현실, "시소러스를 활용한 온톨로지 구축 방안 연구", 한국비블리아학회지, 제17권, 1호, pp. 285-303, 2006.
[9] 최중민, "인터넷 정보 추출 에이전트", 한국정보과학회지, 제18권, 5호, pp. 48-53, 2000.
[10] 서희경, 양재영, 최중민, "준구조화된 정보소스에 대한 지식기반의 Wrapper 학습 에이전트", 정보과학회논문지: 소프트웨어 및 응용, 제29권, 1호, pp. 42-52, 202.
[11] 정창후, 이민호, 주원균, 맹성현, "웹페이지에서 레이블이 없는 텍스트 인식을 위한 확률 모델", 한국정보과학회 2003년도 가을 학술발표논문집, 제30권, 2(I)호, pp. 163-165, 2003.
[12] 정창후, 서정현, 류범종, 맹성현, "도메인 지식을 이용한 랩퍼에서 규칙 생성 정확도 향상", 한국정보과학회 2003년도 봄 학술발표논문집, 제30권, 1(A)호, pp. 662-664, 2003.
[13] 뭉클, http://www.muncle.com/
[14] 다나와, http://www.danawa.com/

저 자 소 개

장문수(Moon-soo Chang)
1992년 : 고려대학교 전자전산공학과 학사.
1994년 : 동 대학원 전자공학과 석사
2001년 : 동경공업대학 지능시스템전공 박사
2000년~2003년 : 한국전자통신연구원 선임연구원
2003년~현재 : 서경대학교 소프트웨어학과 전임강사
관심분야 : 언어이해, 대화처리, 지능시스템, 정보검색, 온톨로지
E-mail : cosmos@skuniv.ac.kr

강선미 (Sun-mee Kang)
1981년 : 고려대학교 전자공학과 학사
1988년 : 에얼랑겐-뉘렌베르그 대학교 전기전자공학과 Diplom
1992년 : 고려대학교 대학원 전자공학과 박사
1992년~1994년 : 고려대학교 정보통신기술공동연구소 연구조교수
1994년~1997년 : 고려대학교 산업대학원 객원조교수
1997년~현재 : 서경대학교 컴퓨터과학과 조교수

관심분야 : 음성처리, 패턴인식, 온톨로지
E-mail : smkang@skuniv.ac.kr