특집

지식기반(Knowledge Base)으로서의 온톨로지(Ontology)와 시멘틱 웹(Semantic Web)

신 효 필*

목 차

1. 서 론

90년대부터 인공지능(Artificial Intelligence)의 지식공학(knowledge engineering) 분야에서 온톨로지(Ontology)가 지식의 공유(sharing)와 재사용(reuse)관점에서 활발하게 사용되기 시작했다. 현재 온톨로지는 이런 지식공학 외에 에이전트에 기반한 소프트웨어 공학이나 전자상거래 등 여러 분야에 널리 퍼져 사용되고 있다. 그러나 그 적용범위의 다양함과 실체의 불분명함으로 인해 그 사용이 혼란스러운 것도 사실이다.

인공지능 관점에서 본다면 지식표현(knowledge representation)이 인공지능의 한 영역으로 발전하면서 이와 관련된 여러 이론적인 방법론이나 실제 시스템들이 개발되었다. 전문가 시스템(expert system)이나 지식기반(knowledge based) 시스템에서 지식이란 문제를 해결하는 것이 필수적인 요소라고 한다면, 이 지식을 어떻게 표현할 수 있느냐가 고려되어야 한다. 이런 지식표현은 여러 영역에 걸치는 분야라고 할 수 있는데, 크게 논리(logic), 온톨로지(ontology), 그리고 계산(computation)의 세 주제와 관련을 맺는다[1]. 논리에 의해 추론의 형식적 구조와 규칙들이 가능해지고, 온톨로지는 그 적용 범위에 존재하는 대상을 규정하며, 계산에 의해 이 지식기반이 순수한 철학과 구별되어 응용될 수 있게 한다. 논리 없이는 어떤 진술이 잉여적인지 모순적인지 알 수 없으며, 온톨로지 없이는 용어와 심벌들이 잘못 정의되거나 혼동스럽게 된다. 그리고 계산없이 온톨로지와 논리가 컴퓨터 프로그램으로 구현될 수 없다. 따라서 지식기반은 논리와 온톨로지를 어떤 도메인에서 계산가능한 모형으로 구축하는 것이라 할 수 있다.

이런 연장선으로, 지식표현의 근간이 되는 온톨로지에 대한 전산적 연구가 근래 많이 행해지고 있다. 이 글에서는 어느 때보다 많은 관심을 끌고 있는 온톨로지에 대해 그 배경과 정의, 그리고 더 나아가 실제 시스템으로 구현되어 있는 온톨로지에 대해 살펴본다. 또한 이 온톨로지는 여러 분야에서 활발히 사용되고 있기 때문에 이 온톨로지가 기본 자료로 이용되고 있는 소위 시멘틱 웹(semantic web)에서의 그 기능과 구축에 대해서도 살펴보도록 한다.

* 서울대학교 인문대학 언어학과 조교수

2. 온톨로지(Ontology)

2.1 배경과 정의

온톨로지라는 말은 희랍어 'ontos(being)'와 'logos(word)'에 기인한다. 이는 원래 철학, 특히 형이상학의 한 분야로, 이 세계에 존재(being)하는 것들의 종류, 그 본성과 관계 등에 대한 연구나 학문을 지칭하는 말이다. 이것이 철학에 도입된 것은 19세기의 독일 철학자들에 의해서였는데, 존재론을 자연과학에서 다양하게 언급하는 존재들과 구분하기 위해서였다. 철학적 관점에서 본다면 온톨로지는 세상의 어떤 관점을 설명하는 분류체계를 제공하는 것이라 할 수 있다.

이 용어가 전산학 문헌에서 처음으로 등장한 것은 1967년 S.H. Mealy에 의해서라고 여겨진다[1]. 이후 이 온톨로지는 지난 수십년간 특히 인공지능 분야에서 지식표현(knowledge representation), 지식공학 관점에서 다양하게 사용되고 있다. 이런 관점에서 가장 널리 알려진 정의는 Gruber[2]의, "온톨로지란 공유된 개념화(shared conceptualization)의 형식적이고 분명한 명세"라는 정의다. 여기서 개념화란 어떤 목적으로 표현하고자 하는 대상을 추상화하고 단순화시킨 것이다. 형식적이란 규정된 용어들과 그들 사이의 관계를 컴퓨터가 이해할 수 있는 방법으로 표시하는 것이다[3].

한편 다른 지식 공학적 관점에서, "온톨로지란 어떤 특정 영역에서 존재하거나 존재할 수 있는 대상들의 범주(category)에 관한 연구"라고 정의된다[1]. 따라서 이런 연구의 결과인 온톨로지는 가능한 세계의 대상의 목록(catalogue)이라고 할 수 있다. 이런 관점에서 본다면 일반적으로 온톨로지는 대상을 범주화하여 명칭을 붙이거나 개념화된 구조를 총칭하는 의미로 사용되며 각각의 범주명은 개념이나 범주로 언급된다.

그러나 온톨로지에 관한 어떤 보편적인 정의도 없다는 점이 지적되어야 한다. 그 이유 중의 하나는 앞에서도 잠시 언급한 대로 그 사용의 광범위함 때문이다. 온톨로지의 개념은 인간과 컴퓨터 시스템 사이의 의사소통을 위해서 적용되기도 하고, 지식의 조직과 재사용(reuse) 관점에서, 그리고 추론(inference)의 관점에서 적용되는 등 다양한 양상을 보인다.

2.2 온톨로지의 형식적 정의

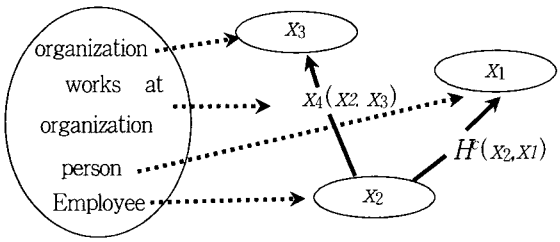
온톨로지는 형식적으로 다음과 같이 정의되기도 한다[4]. 온톨로지는 개념(concepts), 관계(relations), 계층(hierarchy), 그리고 함수(function)를 축으로 다음과 같이 5 가지로 이루어진다.

$O := \{C, R, H^C, rel, A^O\}$

1. C 는 개념을, R 은 관계를 나타낸다.
2. H^C, 는 계층적 관계(concept hierarchy)를 나타낸다. HC, 는 $H^C \subseteq C \times C$ 의 직접적 관계를 나타내며 이는 개념 계층 또는 분류(taxonomy)가 된다. $H^C(C_1, C_2)$ 는 C_1이 C_2의 하위개념임을 나타낸다.
3. 함수(function):
   rel : $R \rightarrow C \times C$ 는 개념을 비분류적(non-taxonomically)으로 관련시킨다.
   dom: $R \rightarrow C$ 와 $dom(R) := \Pi_1(rel(R))$ 은 R의 정의역(domain)을 나타낸다
   range: $R \rightarrow C$ 와 $range(R) := \Pi_2(rel(R))$은 치역(range)를 나타낸다.
   $rel(R) = (C_1, C_2)$ 는 $R(C_1, C_2)$ 와 같다.
4. 온톨로지 공리(A set of ontology axioms) A^O: 적절한 논리언어로 표시된다. 일차 논리(first order logic)가 그 예이다.

다음은 이 정의에 따른 예이다[4]: $C = \{x_1, x_2,$

x_3}, R := {x_4}, 계층적 관계 $H^C(x_1, x_2)$, 비분류적 관계 $x_4(x_2, x_3)$, 어휘부 L^C={"Person", "Employee", "Organization"}, L^R={"works at organization"}, 함수 F("Person") = x_1, F("Employee") = x_2 , F("Organization") = x_1, ("Works at organization") = x_4



(그림 1) Maedche[4]에 의한 온톨로지의 예

2.3 온톨로지의 구조

앞의 논의를 바탕으로 지식공학 관점에서 온톨로지는 어떤 범주나 개념들이 이 세상이나 어떤 특정 영역에 존재하는지, 어떤 속성을 지니고 있는지, 그리고 서로 어떻게 연결되어 있는지에 관한 정보를 지니고 있는 데이터베이스라고 할 수 있다. 이렇게 규정된 온톨로지를 그 목적에 따라 구축하는 데는 여러 고려할 사항이 있다. 구조적으로 온톨로지는 분류적인 계층구조(taxonomic hierarchies)로 여겨지기도 한다. 즉, 어떤 개념이나 클래스들을 계층적으로나 또는 격자(lattices)로 표시할 수 있다.

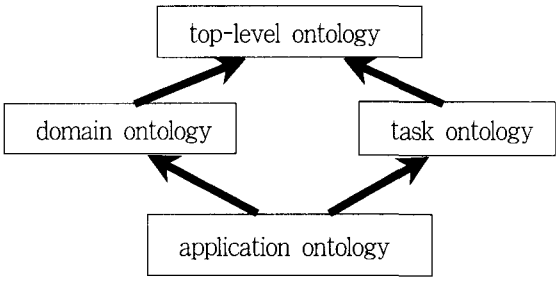
구조적인 면 외에 개념 규정면에서 이 세상이나 또는 특정 영역의 대상을 표시하기 위해 온톨로지 범주를 어떻게 설정할 수 있는가 하는 문제가 제기된다. 이는 전통적으로 철학적인, 형이상학적인 구분에서 뿐만 아니라 실제 온톨로지 시스템을 구축하는 지식공학에서도 중요한 사항이다. 논리적으로 그 근간이 되는 철학자들의 범주 구분도 아주 다양하게 나타난다. Heraclitus는 세상을 이분법적으로 구분하여 'physis'와 'logos'로부터 출발하고, Peirce는 Firstness(일항 관계), Secondness(이항 관계), Thirdness(삼항 관계)로, Whitehead는 Continuants(영속하는 대상), Occurrents(끊임없이 변하는 대상) 등으로 구분하기도 한다[1].

이렇게 철학적으로도 규정하기도 어려운 (상위) 범주이지만 인공지능에서 많은 온톨로지 체계에서는 'Thing', 'All' 등의 범주에서 그 하위범주로 개념을 규정해 나가고 있다. 실제로 Cyc 온톨로지는 최상위 범주 THING에서 IndividualObject, Intangible, RepresentedThing 등으로 확장해 나가고 있다. 이에 대해 Sowa[1]는 이런 구분의 논리적, 철학적 토대에 관해 의문을 제기하고 있다. 즉, 이런 분류 기준은 불분명하며 최상위 범주에 대한 종차부터 상호배타적이지 않다고 한다. 따라서 범주를 개념화하는 것은 온톨로지 구축작업에서 여전히 어렵고 논란이 되는 문제라 할 수 있다.

이렇게 온톨로지를 지식기반 관점에서 구축할 때 일반적으로 다음과 같은 사항이 고려되기도 한다[5]. 첫째, 어느 온톨로지도 단일하지 않다. 온톨로지란 발견되는 자연적 대상이 아니라 어떤 의도를 가지고 구축되는 인공물이기 때문에 어떤 작은 영역에서도 공통되는 구조가 있지 않다. 둘째, 온톨로지는 어떤 임무를 위해 만들어지는 특정성을 보인다. 따라서 자연언어처리(natural language processing)를 위해 구축된 온톨로지는 추론이나, 계획(planning) 등을 위한 온톨로지로는 부적합할 수 있다. 각각의 목적에 따라 다를 수 있다. 셋째, 온톨로지는 검색하거나 그 정보를 찾아보기 쉬운 구조로 되어 있어야 한다. 즉, 사용하기에 편리해야 한다. 넷째, 새로운 개념을 추가하거나 개념적 관계를 확장할 수 있도록 모듈화되어 있어야 한다. 다섯째 개념의 정밀성(granulity)이 고려되어야 한다. 다른 개념과 충분히 구분될 수 있도록 세밀한 개념이 되어야 한다. 마지막으로 잉여성(redundancy)

을 고려할 수 있다. 온톨로지로 개념을 분류하는 것은 필연적으로 잉여적일 수 밖에 없다. 개념을 다차원적으로 구분할 때 개념들이 종종 겹칠 수 있기 때문이다. 따라서 이런 잉여성을 제거하기 위해 개념을 재구조화는 것은 부질없는 일일 수도 있다.

이런 구조적인 측면에서의 어려움과 더불어, Guarino는[6] 온톨로지의 기본 구조를 다음과 같이 제안하고 상위 온톨로지(top-level ontology)의 필요성을 역설하고 있다.



(그림 2) 상위 온톨로지

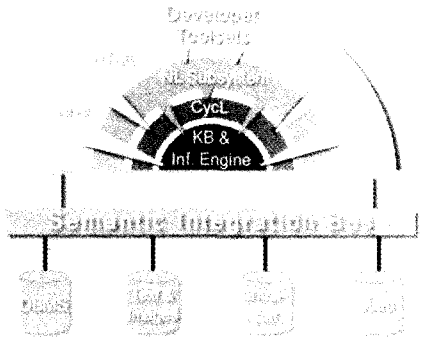
이 상위 온톨로지는 기존의 시간(time), 공간(space), 대상(object), 사건(event) 등과 같은 개념들로 어떤 특정 영역이나 문제와는 독립된 즉, 어느 세계에나 적용될 수 있는 기본적인 개념체계이다. 반면 영역(domain), 업무(task) 온톨로지는 어떤 특정 분야(의료, 자동차 등)이나 어떤 특정 업무(진단이나 영업)와 같은 곳에 적용될 수 있다. 응용 온톨로지는 가장 특정적인 온톨로지라 할 수 있다. 응용 온톨로지에서 개념들은 그 영역 각각의 대상들의 역할이 된다. 이렇게 상위 온톨로지를 구분하는 것은 지식의 공유나 재사용의 때문이다. 즉 서로 다른 목적에서는 같은 대상을 서로 다르게 분류하기 때문에 더 일반적인 체계인 상위 온톨로지에 각각의 응용시스템을 위치시켜 지식을 공유할 필요가 있다[4].

2.4 실제 구축된 온톨로지의 예

다양하게 정의되어 사용되는 온톨로지를 실제 시스템으로 구현한 예들을 살펴보는 것이 온톨로지를 이해하는데 도움이 될 수 있다. 대규모로 구축되어 있는 온톨로지로는 CYC, MIKRO-KOSMOS, GALEN, ENTERPRISE 등이 있으나 여기서는 CYC와 MIKROKOSMOS에 대해서 간략히 살펴보도록 한다.

2.4.1 CYC

Cyc 온톨로지는 1984년부터 Lenat and Guha에 의해 주도된 것으로, 인간의 기본적인 지식을 체계화하려는 목적으로 개발되었다[7]. 따라서 Cyc에서 개발하고 있는 지식기반은 어느 특정 분야보다는 방대한 인간의 기본 지식을 형식적으로 표시하려는 시도라고 할 수 있다. 사용되는 형식적인 기재로 CycL을 사용하며, 지식기반에는 CycL의 어휘를 이루는 개념(terms)들과 이들 사이를 연결하는 단언(assertion)들로 되어 있다. 다른 온톨로지와 달리 프레임을 사용하지 않는 특징을 보인다. 다음은 Cyc 지식기반시스템의 전체 개요다[7].



(그림 3) Cyc 지식 서버(Knowledge Server)

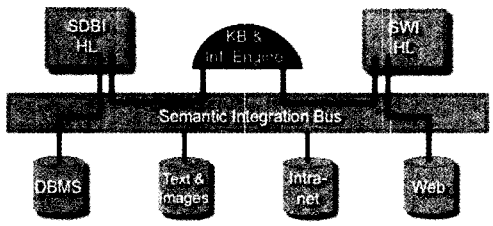
현재 Cyc 지식기반(Cyc KB)은 2만여 개의 개념들과 이 각각의 용어마다 직접 작성한 수십가지의 단언들로 되어 있다. 다음은 Cyc의 한 개념의 예다[8].

#Skin
DEF: A (piece of) skin serves as outer protective and tactile sensory covering for (part of) an animal's body. This is the collection of all pieces of skin. Some examples include #The GoldenFleece(representing an entire skin of an animal) and #BodyPartFn #YulBrynner #Scalp (representing a small portion of his skin).
ISA: Physiology #AnimalBodyPartType.
SEE ALSO: #SheetOfSomeStuff.

각 항목은 개념명으로 시작하고 다음에 이 개념의 의미를 명확히 하기 위한 영어로 주석을 달고 있다. ISA에 의해 이 개념이 속하는 더 광범위한 개념이 표시된다. 또한 이 개념의 하위요소(#BodyPartFn #YulBrynner #Scalp)와 상위요소(#TheGoldenFleece)가 명세될 수 있다. 그러나 프레임을 사용하지 않는 기술이기 때문에 그 정보의 기술이 엄밀하게 형식화되어 있는 것으로 보이지 않을 수도 있다. 다른 온톨로지 시스템에서와 마찬가지로 ISA에 의한 추론으로 개념들이 다양하게 서로 엮어진 계층구조나 격자를 이룰 수 있다. Cyc 지식기반은 개개의 지식을 서로 다른 관점이나 세밀성의 정도, 문화적 차이, 연령의 차이 등을 표시할 수 있는 더 미세한 격자로 표시할 수 있는 장치를 마련하고 있다.

이 Cyc의 지식기반은 (그림 3)에서처럼 첫째 자연언어처리, 둘째 의미통합버스(Semantic Inte-gration Bus), 그리고 개발자 툴셋(Developer toolset)과 연동되어 있다. 자연언어처리 관점에서 영어문장을 읽고 이해할 수 있는 시스템의 개발을 위해서는 Cyc에서 개발하고 있는 세상에 관한 지식, 상식이 필수적이라 생각한다. 이 자연언어처리 시스템은 어휘부(lexicon), 구문분석기(syntactic parser), 그리고 의미해석기(semantic interpreter)로 되어 있다. 또 Cyc 시스템은 지식기반을 확장하거나, 검색하거나 또는 추론을 하기 위한 장치로 여러 인터페이스 툴들도 개발하고 있다.

의미통합버스는 다양한 형태 - 구조화된 자료(데이터 베이스), 반구조화된 자료(웹 페이지), 구조화되지 않은 자료(문서화일) - 의 자료들을 사용가능한 지식으로 전환하기 위해 고안된 것이다. 자연언어처리기에 의해 문서가 읽혀지고 여기서 유용한 단언들을 도출해낸다. 이 단언들은 문서의 내용이나 주제를 나타내는 것으로 여겨지기 때문에 이 정보를 특정한 질문에 대한 답을 도출해내기 위해 사용한다. 다음은 Cyc에서 제시하는 의미통합버스의 구조다[7].



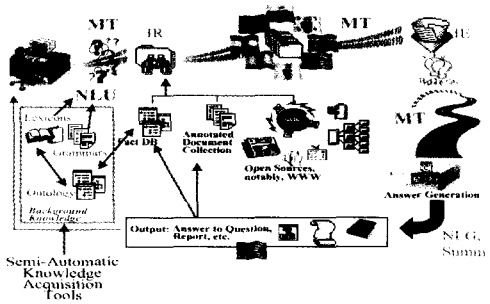
(그림 4) 의미 통합 버스(Semantic Integration Bus)

현재 이 Cyc 지식기반은 http://www.opencyc.org/에서 해당 시스템을 다운받아 사용할 수 있다.

2.4.2 미크로코스모스 온톨로지(The Mikrokosmos Ontology)

자연언어처리에서, 특히 기계번역(Machine Translation)과 관련해서 서로 상이한 언어들 사이에서 호환될 수 있는 개념적 장치로 '언어중립(interlingua)'적인 온톨로지의 개발에 관한 연구의 필요성이 제기되었다. Nirenburg[9]와 Mahesh[5]의 일련의 연구를 거쳐 자연언어의 텍스트의 의미를 표시하는(text meaning representation) 목적으로 개발된 것이 미국 뉴멕시코 주립대학, CRL(Computing Research Laboratory)의 미크로코스

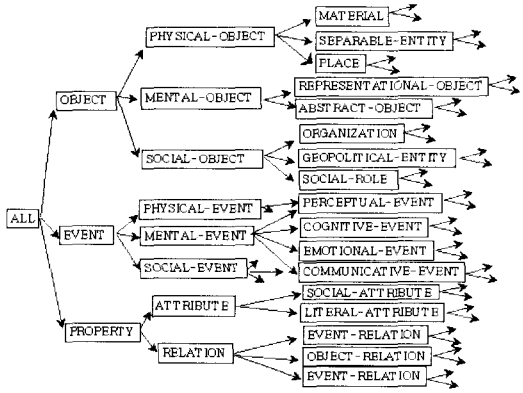
모스 온톨로지이다. 이 미크로코스모스 온톨로지는 다음 그림에서처럼 기계번역을 비롯한 여러 자연언어처리 시스템의 기본적인 지식자원(know-ledge resource)으로 사용되고 있다.



(그림 5) 지식자원으로서의 미크로코스모스 온톨로지

기계번역을 위한 언어중립적 의미표현(interlingual meaning representation)은 전자사전에서 단어의 의미를 기술하고 온톨로지에서 세상 지식을 표현함으로써 이루어진다. 따라서 온톨로지는 세상(또는 특정 영역)에 관한 지식의 구조체이며 기본적인 개념들과 계층적인 구조로 되어 있다. 이 온톨로지는 (1)서로 다른 언어들의 의미를 표현하고, (2)언어 중립적으로 자연언어 문장의 의미를 표시하고, (3)서로 다른 어휘 지식 기반 사이의 지식을 공유하는 기본 자료로 사용된다. 미크로코스모스에서 온톨로지는 첫째, 철학적 단위보다는 전산적 단위이며 둘째, 서로 연결된 개념(concept)들의 망이며 셋째, 개념들은 언어독립적(language-independent)인 것으로 설정된다. 이 개념을 기술하기 위한 메타언어로 현재 영어가 사용되고 있다.

구조적인 면을 살펴 보면, 가장 상위의 개념 'ALL'에서부터 세상의 지식을 대상(OBJECT), 사건(EVENT), 그리고 이 두 개념들이 지니고 있는 속성(PROPERTY)으로 구분하여 계층화한다. 다음은 미크로코스모스에서 설정하고 있는 상위 몇 개념부류들이다.



(그림 6) 미크로코스모스의 상위 몇 개념들

이 온톨로지는 Cyc 온톨로지와 달리 프레임(frame) 방식에 기초하고 있으며 ISA에 의해 개념들이 계층적으로 조직된다. 현재 이 온톨로지에는 대략 5,000여개의 개념이 규정되어 있으며, 35,000여의 서반아어와 20,000여의 중국어 어휘가 온톨로지와 사상되어 있다. 프레임은 어떤 개념을 구성하는 요소들의 속성(property)이 되는 슬롯(slot)으로 되어 있다. 이 속성이 어떤 종류의 값으로 되어 있는지를 명시하는 그 유형(facet)과 실제로 채워지는 값(filler)으로 개념의 지식이 구체화된다. 실제 개념들이 프레임으로 표시되는 예를 들어 설명하면 우선 '대상(object)'의 한 부류인 'TABLE'은 다음과 같은 개념구조로 표시될 수 있다.

〈표 1〉 TABLE의 개념 프레임

Concept	Slot	Facet	Filler(s)
TABLE	DEFINITION	VALUE	"a flat horizontal surface with legs"
	IS-A	VALUE	FURNITURE
	SUBCLASSES	VALUE	ALTAR, DESK, DINING-TABLE, NIGHT-TABLE
	AGE	VALUE	"(> 0)"
	COLOR	VALUE	blue, red, cyan, gray, green, magenda, orange, purple, tan, yellow
	CONTAINED-IN	DEFAULT	PLACE
	HAS-PARTS	SEM	FURNITURE-PART
	INSTRUMENT-OF	SEM	EVENT

프레임을 구성하는 요소들은 이 개념들의 속성을 표시하는 것으로, 온톨로지 체계에서 속성(PROPERTY) 개념들로만 표시된다. 즉, 개개의 슬롯은 특정 속성 프레임과 상응된다. 따라서 다른 두 개념체계인 대상(OBJECT)과 사건(EVENT)은 결코 프레임 슬롯으로 쓰일 수는 없으나, 그 슬롯을 채우는 값으로는 사용될 수 있다. <표 1>에서 IS-A, HAS-PARTS, AGE 등은 속성(PROPERTY)에 속하는 개념들이며, 그 값으로 채워지는 FURNITURE, FURNITURE-PART 등은 대상에 속하는 개념들이다. 따라서 온톨로지 체계에서 최상 삼분지 - 대상, 사건, 속성 -으로 구분하는 것은 세상지식을 대상과 사건으로 부류화하고 그 지식의 속성을 기술하기 위한 개념체계를 추가한 것으로 볼 수 있다.

슬롯은 다시 일반적인 슬롯(non-special)과 특수 슬롯(special slot)으로 구분된다. 이 관계는 다음과 같이 도식화될 수 있다.

<표 2> 슬롯의 종류

일반적 슬롯 (non special slot)	특성 (ATTRIBUTE)	값(value)	숫자, 수 범위, 문자(literal string), 대상(OBJECT), 사건(EVENT)
	관계 (RELATION)	값(value)	대상(OBJECT), 사건(EVENT)
		격-역할 요소 (Case-Role)	AGENT, THEME, INSTRUMENT, EXPERIENCER, BENEFICIARY, ACCOMPANIER, PURPOSE, LOCATION, SOURCE, DESTINATION, PATH
특수 슬롯 (special slot)	모든 개념에 적용되는 요소들	DEFINITION, IS-A, SUBCLASSES, INSTANCES	
	속성 (PROPERTY)에만 적용되는 개념들	DOMAIN, RANGE, INVERSE, MEASURED-IN	

일반적인 슬롯은 다시 'AGE', 'COLOR'와 같은 특성(ATTRIBUTE)과 HAS-PART와 같은 관계(RELATION)로 이루어진다. 이 '특성'은 어떤 대상의 속성을 기술하기 위해 쓰이는 것으로 그 값으로 '숫자'나 '수범위', '문자열(literal symbols)' 그리고 '대상', '사건'의 개념들이 사용된다. 따라서 '속성(PROPERTY)' 자체는 이 슬롯의 값으로 쓰일 수 없다. 관계(RELATION)는 '대상'이나 '속성'의 다양한 다른 개념과의 관계를 나타내기 위해 사용되는 것으로 주로 'HAS- ' 형태나 다른 관계를 나타내는 속성개념들로 되어 있다. 이 '관계'도 마찬가지로 그 채워지는 값은 '대상'이나 '사건'이어야 한다. 채워지는 값들이 '대상'이나 '사건' 개념들이라는 것은 이 값들로 인해 전체 개념체계가 서로 망으로 연결될 수 있음을 의미한다. 관계 속성에 해당하는 것으로 격-역할 슬롯(Case-Role Slot)이 있다. 이는 전통적으로 '의미역'으로 규정되는 것으로 '사건' 개념들의 다양한 관계를 명시하기 위해 사용된다. 따라서 이 격-역할 슬롯은 '사건' 개념들에만 나타나야 하며 '대상'이나 '속성' 개념에서는 사용되지 않는다.

모든 '관계'를 나타내는 개념은 그 반대-관계(INVERSE-RELATION)가 자동적으로 구축된다. 따라서 격-역할 관계인 AGENT는 AGENT-OF라는 그 반대의 관계를 갖게 되어 개념들이 양방향적으로 서로 연결될 수 있다. 앞의 <표 1>의 'TABLE'은 HAS-PARTS의 관계에 의해 FURNITURE-PART라는 개념을 값으로 갖는 것으로 표시되고 이 FURNITURE-PART는 다시 HAS-PARTS의 반대-관계인 PART-OF의 값으로 TABLE을 갖게 된다. 이는 채워지는 값들이 다른 개념과 서로 연결될 수 있는 망을 구성하게 됨을 의미한다.

특수 슬롯(special slot)은 모든 개념에 적용될 수 있는 'DEFINITION, IS-A, SUBCLASSES, INSTANCES' 와 몇 개념에만 적용되는 'DOMAIN, RANGE, INVERSE, MEASURED-IN' 등이 있다. 정의(DEFINITION)는 각 개념을 정의하여 기술하기 위한 것으로 온톨로지 구축관

점에서 참조용으로 사용된다. IS-A 슬롯은 모든 개념에 필수적인 것으로 해당 개념의 상위개념이 무엇인지 규정하며 SUBCLASSES는 반대로 하위 개념을 명시한다. 따라서 이 두 개념은 서로 반대-관계에 있는 것으로 생각할 수 있으며 계층적인 개념체계 구성과 상속이라는 점에서 중요한 기능을 한다. 각 개념의 실제 세계에서의 구체적인 실현으로 INSTANCE를 미크로코스모스 온톨로지에서는 구분하고 있는데 NATION이라는 개념에 대해 'FRANCE, KOREA' 등이 그 INSTANCE가 될 수 있다.

몇 개념에만 적용되는 관계로 DOMAIN, RANGE, IVNERSE, MEASURED-IN 등이 있다. 이 중에서 'DOMAIN'과 'RANGE'는 '속성' 개념에만 적용되며 이 개념이 어떤 영역(DOMAIN)에 쓰이며 어떤 범위(RANGE)로 적용되는지를 명시한다.

개념을 기술하는 각 요소는 그 취하는 값들의 더 미세한 구분을 하기 위한 장치로 슬롯의 유형(facet)이 어떤 종류인지를 명시한다. 현재 'VALUE, SEM, DEFAULT, SALIENCE, NOT'이 가능한 측면으로 설정되어 있다. 'VALUE'는 개개의 요소에 채워지는 값들이 실제의 값이라는 것을 명시하며 'SEM'은 선택제약으로 기능할 수 있는 개념들을 취할 때 사용된다. 'SEM, VALUE' 두 값 모두 하위 개념으로 상속된다. 'DEFAULT'는 취해지는 값이 전형적인 것임을 명시하기 위해 그리고 'SALIENCE'는 그 개념을 특징짓는 상대적으로 두드러진 값을 나타내기 위해 사용된다. 'NOT'은 특정 슬롯의 값을 상속되지 않게 막는 장치이며 따라서 그 값은 하위개념으로 상속되지 않는다.

이렇게 더 미세하게 규정되는 슬롯들을 채우는 값들은 앞에서 살펴본 대로 다른 개념이나, 숫자, 문자열 등이 될 수 있다. 이 채워지는 값들은 하나일 필요가 없으며 여러 값들이 가능하다. 따라서 이 개념에 의해 규정되는 관계들이 다양해지는 효과가 있다.

이렇게 구축된 온톨로지는 현재 기계번역, 정보검색(information retrieval), 질의어 응답시스템 등에서 사용되고 있으며, 이 온톨로지를 구축하고 유지하기 위한 여러 툴들이 개발되어 있다.

3. 온톨로지(Ontology)와 시멘틱 웹(Semantic Web)

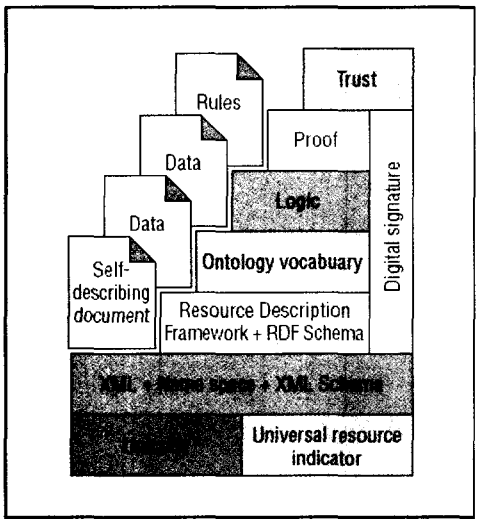
지금까지 살펴본대로 지식공학 측면에서의 온톨로지는 현재 컴퓨터 과학의 여러 분야에서 응용되어 사용되고 있다. 그 적용되는 분야로는 디지털 도서관, 멀티 데이터베이스 시스템, 지능형 에이전트, 기계학습, 데이터와 텍스트 마이닝, 인간과 컴퓨터 인터페이스 등 다양하다. 현재 이 온톨로지가 중심적인 지식기반으로 기능하는 응용시스템으로 대표적인 것으로 시멘틱 웹, 자연언어이해(natural language processing), 지식 처리(knolwedge management), 그리고 전자상거래(e-business) 등을 들 수 있다. 여기서는 현재 활발히 논의되고 있는 시멘틱 웹에서 온톨로지와 그 구축 및 유지에 대해 살펴보도록 한다.

3.1 시멘틱 웹(Semantic Web)

웹의 발달은 인터넷을 통하여 누구나 문서를 만들어 내고 쉽게 정보에 접근하게 하였다. 그러나 엄청난 양의 정보의 증가는 인간이 통제할 수 없을 정도로 방대해져서 사용자가 일일이 살펴보기란 불가능할 정도가 되었다. 이러한 문제를 해결하기 위해 Tim Berners-Lee를 비롯한 연구자와 학자들이 "시멘틱 웹"이라는 다음 세대의 웹 개념을 제안하고 있다. 이는 문서의 의미를 명백하고 기계가 이해할 수 있는 형태로 표현하여 컴퓨터가 웹 자원들을 효율적으로 관리할 수 있게 하려는 것이다.

만일 컴퓨터가 정보를 이해할 수 있다면 사용자가 필요로 하는 결과만을 찾아주는 의미기반 검색이 가능하고 사람과 기계, 기계와 기계 상호간에 연결을 원활히 수행할 수 있게 된다. 그러나 현재 대부분의 정보는 인간만이 이해할 수 있는 형태로 되어 있다. 따라서 온톨로지가 웹 내용에 필요한 의미를 제공하게 되고 소프트웨어 에이전트가 이를 이해하여 적절한 문맥의 정보만 도출해내게 할 수 있다. 이것이 시멘틱 웹의 기본적인 출발이라 할 수 있다.

이 시멘틱 웹은 현재 www 콘소시움(w3c)의 한 영역을 차지하며, 그 표준안이 계속 논의되고 있다.(http://www.w3c.org/2001/sw/) 시멘틱 웹은 Tim Berners-Lee의 제안과 같이 다음의 구조로 이루어진다.

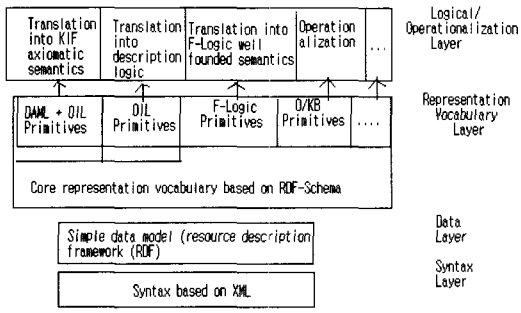


(그림 7) Tim Berners-Lee의 시멘틱 웹 'layer-cake'

서로 다른 웹 정보들에 의미를 부여하기 위해 시멘틱 웹에서 핵심이 되는 부분은 온톨로지라 할 수 있다. 이 온톨로지는 시멘틱 웹의 핵심 기술들 중 활발한 연구가 수행되는 분야로 이 온톨로지를 구축하고 통합하는 많은 표준안들과 기술들이 현재 개발되고 있다. 여기서는 시멘틱 웹에서 온톨로지의 역할과 그 언어로 사용되는 것들 그리고 온톨로지를 구축하고 개발하는 도구와 방법론에 대해 간략히 살펴보도록 한다.

3.2 온톨로지를 위한 계층적인 구조

시멘틱 웹에서는 (그림 7)과 관련하여 온톨로지를 위한 다음과 같은 계층적인 언어표현 구조를 상정한다.



(그림 8) Maedche[4]에 의한 계층적인 표현 구조

가장 하층을 이루고 있는 것은 시멘틱 웹의 구조적인 측면인 통사 층위로, XML이나 HTML을 사용하여 온톨로지와 지식기반을 단일하게 통사적으로 표상할 수 있게 한다. 그 다음 층인 자료 층위(data layer)는 RDF(Resource Description Frame-work)는 컴퓨터가 처리할 수 있는 메타데이터(metadata)로 표시 가능한 형식적 데이터와 구조(syntax)를 정의한다. RDF는 웹이나 문서 그리고 데이터베이스 등에 있는 어휘들을 그 의미를 규정할 수 있는 형태로 연결시켜 주게 되는데 이것이 바로 온톨로지라고 할 수 있다. 이 RDF 는 확장 가능한 타입 시스템을 정의하고 있는 것으로 개념 위계와 속성들에 대한 영역, 분야에 대한 제약을 정의하는 수단을 제공하고 있지만 표현력(expressive power)의 부족이 문제로 제기되었다. 즉, RDF-schema에서는 부정, 이접, 연접 등의 논리표현들이 제공되지 않아 온톨로지의 표현력을 제한하는 결과를 초래한다는 것이다.

그 다음 중간 단계로 표상 어휘 층위(representation vocabulary layer)가 있다. 이 층위는 온톨로지를 위한 서로 다른 기초적인 어휘들을 제공한다. RDF-schema에 의해 핵심적인 공통 어휘들이 규정된다면 특정 지식표현이나 각기 다른 어플리케이션을 위한 추가적인 어휘들도 정의될 수 있다. 이런 다른 표상어휘들로 OIL, DAML+OIL, DRDF(S) 등이 있다. OIL(the ontology inference layer)은 on-to-knowledge 라는 프로젝트로 유럽의 여러 나라들에 의해 지원되고 있다. OIL은 웹에서 온톨로지를 생성할 때 RDF의 부적절한 표현력을 보충하고 추론 등을 위한 형식의미론적 측면을 제공하기 위한 것이다[10]. 미국 DARPA(the Defense Advanced Research Projects Agency)는 w3c와 함께 DAML(DARPA Agent Markup Language)를 개발하고 있다. 이는 웹에서 에이전트들의 상호작용을 더 편리하게 하기 위해 RDF의 표현력을 증가시키려는 것이다. DAML은 그 첫번째 온톨로지 언어의 명세로 DAML-ONT를 2000년 10월에 발표했으나 그해 12월에 다시 DAML+OIL를 발표하여 DAML-ONT를 대치하였다. DAML+OIL의 의미형식은 KIF(the ANSI Knowledge Interchange Format)로 되어 있어 일차술어논리(first order predicate logic)로 표시가능하다. 현재는 새로운 온톨로지 언어로 OWL(Web Ontology Language)이 w3c에 의해 추천되고 있다. 이 OWL은 DAML+OIL 온톨로지 언어의 구축과 적용에서 알려진 교훈들을 바탕으로 DAML+OIL을 개정하려는 것으로 그 작업이 진행 중이다.

이외에 (그림 8)에는 명시되어 있지 않지만 온톨로지 언어로 제안되고 있는 것이 메릴랜드 주립대학에서 개발되고 있는 SHOE(the Simple HTML Ontology Extension)다. SHOE는 웹 페이지에 온톨로지에 기초한 지식 표현언어를 내장하려는 것이라 할 수 있다. 그 근본 철학은 만일 가장 중요한 정보가 구조적인 방법으로 제공될 수 있다면 인터넷 에이전트가 더 효율적으로 작업할 수 있다는 것이다. SHOE는 기존의 HTML을 지식에 기반한 태크 체계로 보충하고 각 웹 페이지를 하나나 그 이상의 온톨로지와 관련시켜 그 의미를 표시할 수 있게 한다. RDF와 비교할 때, SHOE는 그 연장선 상에 있다고 할 수 있으나 표현력에 있어서는 RDF보다 더 떨어진다고 할 수 있다.

마지막으로, 가장 상위의 논리와 가용화 층위(logical and operationalization)는 형식 의미론적 측면을 반영하는 것으로 중간 층위에서 도입된 표상 어휘들을 추론화하고 구체적으로 가용화할 수 있게 한다. 일례로 온톨로지 교환 언어(ontology interchange language)인 Ontolingua는 KIF에 기초한 의미기술을 사용하는 온토롤지를 가능하게 한다.

3.3 온토로지 개발 도구들

온톨로지를 개발함에 있어 태그를 여과하거나 인간이 이해할 수 있는 형태로 정보를 표시하기 위한 도구들이 필수적이다. 시멘틱 웹에서 온톨로지를 구축하기 위한 도구와 방법론에 대해 간략히 살펴보도록 하자.

3.3.1 Protege-2000

Protege-2000은 온톨로지를 편집하고 구축하기 위한 그래픽 툴이다[11]. 이것은 공개적인 자료구조를 바탕으로 개발되었으며 지식기반을 구축하기 위한 간편한 도구의 개발을 목표로 하고 있다. Protege-2000은 어느 특정 온톨로지를 따르지 않으며, 클래스(class)와 인스탄스(instance), 슬롯(slot) 등을 만들 수 있다. OIL 언어용 플러그 인(Plug in)이 있으며, DAML 용은 현재 개발중이다.

3.3.2 OilEd

OilEd는 DAML+OIL을 사용하여 사용자가 온톨로지를 구축할 수 있게 한다. 처음에는 OIL 언어의 사용을 예시하는 등의 단순한 편집기로 시작했기 때문에 온톨로지 통합이나 정렬 등의 환경을 제공하지 않는다. 여러 기업이나 연구소에서 교육, 연구용으로 많이 사용되고 있다.

3.3.3 OntoEdit

OntoEdit 은 w3c의 표준을 지원하는 환경이다. 즉, 온톨로지를 RDFS, XML 그리고 DAML+OIL 형태로 출력할 수 있게 한다. OntoEdit은 세 단계 - 요구(requirements), 정제(refinement) 그리고 평가(evaluation)- 로 되어 있다.

3.3.4 Chimaera

Chimaera 는 분산 온톨로지를 생성하고 유지하는 환경을 제공한다. 온톨로지를 편집할 수도 있지만 그 주된 기능은 여러 온톨로지를 결합하거나 진단하는데 사용된다. Chimaera는 사용자가 지식기반 자료를 불러와서 형식(format)을 달리하고 그 분류를 재조직하며 이름이 중복되는 것을 해결할 수 있게 한다.

지금까지 온톨로지 구축 도구들에 대해 간략히 살펴 보았다. 비록 이런 도구들이 있다고 해도 실제로 온톨로지를 개발해 나가는 데는 표준화된 과학적인 방법 보다는 여러 기술적인 측면이 많이 요구되고 있다. 따라서 실제 온톨로지 구축에서 얻어진 경험을 기초로 하여 여러 방법론들이 제시되기도 한다[10]. 그러나 개념을 획득하고 규정하며 그들간의 다양한 관계를 구축하는 것은 쉬운 작업이 아니며, 대부분 개발자들의 개인적인 경험에 의존하는 경우가 많다.

온톨로지 구축에 있어 가능한 방법론 중의 하나로 어휘부에 기초한 온톨로지 구축(lexicon based ontology construction)이 있다[10]. 이는 응용 언어의 개념에 기초하여 체계화된 개념 도출 방법을 설정하려는 시도로 언어의 확장된 어휘부(LEL: anguage extended lexicon)에 의해 이루어진다. 어휘부의 각각의 용어(term)는 두 유형의 기술로 이루어진다. 첫 유형은 개념(notion)으로 그 용어의 명시적 의미(denotation)를 나타내고, 두 번째 유형은 소위 행동적 반응(behavioral response)으로 그 용어의 함축적 의미(connotation)를 나타낸다. 이렇게 확장된 어휘부(LEL)를 구축하고 여기에 타당화(validation), 일관성(consistency) 점검을 통해 온톨로지가 구축될 수 있다. 그러나 이런 표준화된 방법은 모든 온톨로지 구축시스템에 적용되기 어려워 보인다. 따라서 개개의 온톨로지 개발에 특정적인 방법론들이 모색되고 있다. 이에 대한 앞으로의 많은 연구가 필요하다.

4. 결 론

지금까지 지식공학 관점에서 많이 쓰이고 있는 온톨로지의 개념에 대해 그 배경과 구조적/형식적 정의에서 출발하여 실제 구축된 시스템의 예들과 시멘틱 웹에서의 그 지위에 대해 간략히 살펴보았다. Karin[10]은 앞으로의 온톨로지의 전망과 관련하여 전통적인 의미에서의 온톨로지와 시멘틱 웹 관점에서 그 차이를 볼 수 있다고 한다. 전통적 의미에서 온톨로지는 어떤 특정 영역에서의 체계화된 지식을 엄밀하고 형식적으로 표시하는 것에 초점이 맞추어졌다. 따라서 그 체계가 사용 시 안정적이어야 한다. 물론 그 규모가 방대하여 구축에도 오랜 시간이 걸린다. 반면 웹과 관련하여서는 방대한 규모의 온톨로지가 아니라 작은 규모의 온톨로지들이 구축되고 서로 연결되고 있다. 즉, 개인들이 웹 페이지를 만들 듯이 이런 온톨로지를 만들고 있다. 더 나아가 웹에서 에이전트들이 서로 연결될 수 있는 응용 온톨로지(application ontology)의 필요성이 증대될 것이다. 이 경우 응용 온톨로지는

어떤 특정 영역의 온톨로지보다 더 제한적이지만 더 간단한 대상을 규정한다. 따라서 이런 응용 온톨로지를 위한 여러 기술적 준비가 필요할 수 있다.

지식 공학과 정보 처리에 있어 의미적 분석의 중요성은 온톨로지와 시맨틱 웹에서의 그 응용으로 잘 증명되고 있다. 현재 활발히 연구되고 있는 온톨로지와 그 개발 도구 및 방법론은 여러 다른 응용분야와 시스템으로 그 적용범위를 급속히 넓혀 가고 있다. 따라서 앞으로의 더 많은 연구와 투자가 필요하다.

참고문헌

[1] J.F. Sowa, Knowledge Representation, Brooks/Cole, 2000.

[2] T.R. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition 5, 1993.

[3] D. Fensel, Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Springer, 2001.

[4] A. Maedche, Ontology Learning For The Semantic Web, Kluwer Academic Press, 2002.

[5] K. Mahesh, Ontologies For Natural Language Processing, CRL Technical Report, 1995.

[6] N. Guarino, "Formal Ontology and Information Systems", In Proceedings of the FOIS'98, Formal Ontology in Information Systems, 1998.

[7] Information on CYC is available at http://www.cyc.com.

[8] B.C. Vickery, "Ontologies", Journal of Information Science, vol. 3, No.4, pp.277-286, 1997.

[9] S. Nirenburg et al., The Structure of Interlingua in TRANSLATOR. In S. Nirenburg ed., Machine Translation, Cambridge University Press, 1987.

[10] K.B. Karin and J.C.S Leite, "Ontology as a Requirements Engineering Product" downloadable from http://www-di.inf.puc-rio.br/~karin/tutorial.pdf.

[11] N. Noy, M. Sintek et al., "Creating Semantic Web Contents with Protege 2000", IEEE Intelligent Systems vol 16 No.2, 2001.

저자약력

신 효 필

1984-1988 서울대학교 언어학과 학사
1988-1990 서울대학교 언어학과 석사
1990-1994 서울대학교 언어학과 박사
1995-1997 Univ. of Missouri-Kansas City, 전산학 석사
1998.1-2001.1. CRL, New Mexico State Univ. USA. 연구원
2001.1-2001. 12. YY Technologies in Silicon Valley, 연구원
2001. 9 - 2003.2. 서울대학교 공과대학 전기공학부 계약조교수
2003. 3 - 현재 서울대학교 인문대학 언어학과 조교수