경량 온톨로지 생성 연구

한동일* · 권혁인** · 백선경***

A Study for the Generation of the Lightweight Ontologies

Dong-IL Han* · Hyeong-In, Kwon** · Sun-Kyoung, Baek***

■ Abstract ■

This paper illustrates the application of co-occurrence theory to generate lightweight ontologies semi-automatically. The proposed model includes three steps of a (Semi-) Automatic creation of Ontology; (they are conceptually named as) the Syntactic-based Ontology, the Semantic-based Ontology and the Ontology Refinement. Each of these three steps are designed to interactively work together, so as to generate Lightweight Ontologies. The Syntactic-based Ontology step includes generating Association words using co-occurrence in web documents. The Semantic-based Ontology step includes the Alignment large Association words with small Ontology, through the process of semantic relations by contextual terms. Finally, the Ontology Refinement step includes the domain expert to refine the lightweight Ontologies. We also conducted a case study to generate lightweight ontologies in specific domains(news domain).

In this paper, we found two directions including (1) employment co-occurrence theory to generate Syntactic-based Ontology automatically and (2) Alignment large Association words with small Ontology to generate lightweight ontologies semi-automatically. So far as the design and the generation of big Ontology is concerned, the proposed research will offer useful implications to the researchers and practitioners so as to improve the research level to the commercial use.

Keyword : Semantic Web, Ontology, Ontology Alignment

논문투고일 : 2008년 10월 20일 논문수정완료일 : 2008년 12월 06일 논문게재확정일 : 2008년 12월 08일

* KT 미래기술연구소 책임연구원
** 중앙대학교 상경학부 교수
*** KT 미래기술연구소 연구원

1. 서 론

시맨틱 웹이란 웹상에 존재하는 정보를 사람뿐만 아니라 기계(컴퓨터)가 의미를 파악하고 사용자의 요구에 적합한 결과를 제공하며, 사람과 기계 또는 기계와 기계 상호 간에 협업을 원활히 수행함으로써, 사람을 대신하여 자동적인 서비스가 가능한 웹을 말한다[3]. 즉, 시맨틱 웹은 컴퓨터가 정보 자원의 의미를 이해하고, 자동화하고, 통합하고, 재사용할 수 있는 차세대 웹 기술이며, 특히 온톨로지는 사람과 기계 또는 기계 상호간 매개 역할을 하는 시맨틱 웹의 핵심 기술이다[13~15]. 온톨로지는 공유된 개념화에 대한 형식적인 명세 체계로서, 도메인 어휘의 의미 정보를 제공한다[11]. 온톨로지는 일종의 지식 표현으로 이를 통해 컴퓨터는 표현된 개념들을 이해하고 추론을 통한 지식처리를 할 수 있다. 추론 등의 처리를 위해서는 온톨로지의 공리(Axiom)와 규칙(Rule) 체계가 필요하다.

한편, 시맨틱 웹 서비스 제공을 위해 지나치게 지식 표현이 상세화 된 온톨로지(Heavy Ontology)를 구축한다면, 실제 애플리케이션 적용시 뛰어난 가치를 제공할 수 없다. 이러한 현상의 주요 원인은 온톨로지에 표현된 지식을 포착(Capture)하기가 매우 어렵고 복잡하여 추론의 성능이 저하되기 때문이다. 현실적으로 표현력을 최소화한 온톨로지(Lightweight Ontology, Little Semantics)와 계산상 복잡한 온톨로지(Heavy Ontology)는 상충관계에 있다. 단순한 온톨로지(Lightweight Ontology, Little Semantics)는 장기간 사용이 가능하여, 시맨틱 추론의 부담이 없어 성능의 제약도 없다는 것이 대세이다[17].

현재까지 이러한 온톨로지는 특정 분야의 전문가들에 의한 수작업에 의존하거나, 온톨로지 스키마(Schema) 구조를 참조하여 인스턴스(Instance)를 자동으로 추출한 후 다시 수동으로 온톨로지를 보완하는 방법에 의존하고 있다. 이러한 이유로 온톨로지는 시맨틱 웹의 핵심 구성 요소임에도 불구하고, 현재까지도 상용화 서비스를 위한 대규모 온톨로지를 성공리에 적용하는 사례가 미흡하다. 더욱이 산재하고 있는 다양한 형태의 온톨로지를 통합(Integration)하여 새로운 시맨틱 웹 서비스 제공을 위해서는 시맨틱 웹의 특성상 이기종(Heterogeneous) 분산(Distributed) 환경이기 때문에 필연적으로 수반되는 온톨로지 매칭 방법이 요구되고 있다.

따라서 본 연구에서는 현실적으로 표현력을 최소화하면서 넓은 범위의 지식을 포함한 대규모 경량 온톨로지 자동 구축 방법과 다양한 이기종 분산 환경에서의 온톨로지들을 통합할 수 있는 온톨로지 매칭 방안을 제시하고자 한다. 아울러 제안한 방안을 토대로 뉴스 도메인에 적용한 사례연구도 제공한다. 본 논문의 제 2장에서는 관련 연구를 제시하고, 이를 토대로 경량 온톨로지 구축 방안을 제 3장에서 나열한다. 제 4장에는 제안된 온톨로지 구축 방안의 적용 사례를 열거하고, 마지막으로 제 5장에서는 결론 순으로 살펴보고자 한다.

2. 관련 연구

웹과 같은 이기종 분산 환경에서 지식을 공유하고 재사용하는 방안에 대한 연구가 진행되고 있으나, 지식 서비스를 위한 지식의 공유와 재사용은 지식 시스템에서 매우 어려운 연구과제 중에 하나로 인식되고 있다[5].

따라서 기존 웹의 문제점을 해결하기 위해 시맨틱 웹을 고려하고 있다. 웹은 다양한 형태의 지식을 포함하고 있다. 그러나 현재까지 웹상에 존재하는 지식은 조각화 되어 있으며, 기계가 상호 연결하기에 매우 어려운 구조적 결함을 가지고 있다. 정보의 조각화, 프로세스와 애플리케이션의 기능화는 시맨틱 웹 기술을 활용하면 인간과 기계가 동시에 의미를 공유하고, 지식이 연결될 수 있도록 구성할 수 있다. 이러한 시맨틱 웹 기술은 기존 인터넷에 지식 공간(Knowledge Space) 즉, 온톨로지를 제공하여 인간과 기계가 연결, 진화, 공유

가능하며, 예기치 못할 정도의 규모와 새로운 방식으로 지식을 이용할 수 있도록 지원한[2, 6]. 이러한 이유 등으로 인해 특정 형태의 목표를 달성하기 위해 온톨로지로 대표되는 지식 공간 또는 지식 계층(Knowledge Layer)을 도메인 지식 응용과 관리를 지원할 수 있는 인프라 역할로 기대되고 있다[1, 7, 8].

그러나 온톨로지의 중요성에도 불구하고 온톨로지는 완전하게 형식적(Formal)으로만 구성할 수 없다. 반면에 반형식적인 온톨로지는 온톨로지가 부분적으로 불완전한 지식, 예를 들어 불일치성 혹은 제약조건 위반 등의 형태로 구축된다. 즉, 다양한 소스로부터 지식을 추출하고 통합하며, 수많은 작업자에 의해 온톨로지가 구축되므로 이러한 반형식적 상태가 불가피하다[12].

한편, 온톨로지 자동생성과 관련된 기존 연구의 대부분은 주어진 온톨로지 스키마에 인스턴스를 자동생성하거나, 특정 영역에 해당하는 데이터를 분석하는 기법인 데이터마이닝 등을 통해 온톨로지를 생성하는 연구가 추진되고 있다.

위와 같은 연구는 크게 온톨로지 생성의 자동화 연구[14], [14]와 기존 온톨로지의 매칭 연구[10, 16, 18]로 구분되어 진행되고 있다.

우선 온톨로지 생성의 자동화 주요 연구를 살펴보면 아래와 같다. 시맨틱 웹 어노테이션을 위한 웹 문서의 자동 의미정보 추출 연구에서는 비정형화된 웹 문서로부터 정형화된 온톨로지의 인스턴스를 자동으로 추출함으로써 대용량 웹의 의미화 및 자동화 작업을 가속화하려는 연구이다. 이러한 연구에서는 반드시 사용자의 학습 데이터를 생성한 후, 기계학습 방법인 SVM(Support Vector Machine)과 베이지안(Bayesian) 분류기 등을 통해 인스턴스를 생성하려는 시도를 하고 있으나, 학습 데이터 범위 및 실험데이터가 소규모로 한정되어 있어 인스턴스 추출 방식의 한계점을 내재하고 있다. 또한 비구조 웹 문서로부터 온톨로지 인스턴스를 자동으로 추출하는 연구는 온톨로지가 존재한다는 전제하에 정보추출을 통한 방식으로 정보추출 단계에서 정보 추출 규칙인 문 패턴구조를 이용, 인스턴스를 제한적으로 추출하는 방식이다. 또 다른 관점의 연구인 문서로부터 개념 간의 관계추출을 통한 온톨로지 자동 구축 연구는 온톨로지 내에 존재하는 개념과 개념 사이의 관계를 자동으로 추출하고 구문 패턴과 연관 패턴을 군집화하여 각 관계의 이름을 지정하는 방식이다. 이 방 식은 관계연관 정보를 통해 각 관계를 정의할 수 있는 이름 부여 등의 제약점을 가지고 있다. 위에서 나열한 연구들의 공통된 제약점은 제한된 학습 데이터나 패턴 등을 이용하기 때문에 대규모 온톨로지 생성에는 아직 미흡한 점이 많다.

또한, 기존 온톨로지들을 통합하여 새로운 온톨로지를 생성하는 온톨로지 매칭 방식에 대한 연구도 고려할 만하다. 그러나 불일치 또는 부정확 매칭의 보완과 온톨로지 진화를 통한 매칭 등이 아직 도전적인 과제로 존재하고 있다. 아울러 현실적으로 적용하기에 너무 복잡한 로직을 이용한 매칭 방법의 설계는 대규모 온톨로지 생성을 위한 비용 대비 효과 측면에서도 새로운 접근 방식이 요구된다.

[9]와 같은 연구에서는 검색에서 실용적으로 응용할 수 있는 기법인 공빈도 이론(Co-occurrence theory)을 이용하여 경량 온톨로지를 생성하고 생성된 온톨로지를 이용하여 검색 향상을 도모하였다. 그러나 위 연구에서는 온톨로지의 개념적 속성을 구체적으로 활용하지 못하였고, 매칭의 대상이 되는 객체들간의 다양한 체계에 대해 추상적인 접근을 시도하였다. 따라서 좀 더 응용에 잘 적용될 수 있고 개념적으로 명백한 대규모 온톨로지 생성 연구가 주요한 연구 이슈로 남아있다.

앞에서 기술한바와 같이 시맨틱 웹은 지식베이스 구축 방식에 있어 보편성, 분산성, 대용량 이라는 측면에서 차이점이 있다. 따라서 본 연구의 필요성은 아래와 같다.

첫째, 온톨로지의 대상은 대규모로 구축되어야 한다. 도메인 전문가가 수작업으로 구축하거나, 비용대비 효과성이 미흡한 복잡한 로직을 통한 자동

화 구축 방식으로는 부족하다. 실용적으로 응용가능한 온톨로지 구축 방식이 요구된다.

둘째, 온톨로지는 지속적으로 진화되어야 한다. 도메인이 변경되고, 정보가 변경되는 등 중앙집중적인 관리의 범위를 넘어 수시로 변경되어야 한다. 온톨로지는 도메인 전문가가 한 번 작업하는 정적인 과정이 아니고 지속적으로 관리되어야 한다.

셋째, 최적의 온톨로지(Right Ontology)는 사실상 불가능하다. 단지, 다양한 관점과 목적의 온톨로지들이 존재할 뿐이다. 따라서 다양한 형태의 온톨로지들을 용도에 따라 매칭할 수 있는 기법이 요구된다.

따라서 본 연구에서는 보편성, 분산성, 대용량을 추구할 수 있는 경량 온톨로지 생성 기법에 대해 기술하고자 한다.

3. 경량 온톨로지 생성 방법

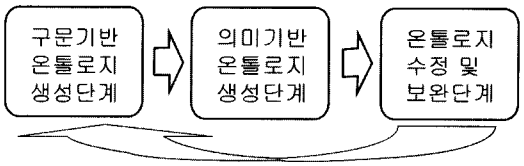
기존 소규모로 조각화 되어 있는 온톨로지를 하나의 거대한 온톨로지로 통합하는 온톨로지 매칭 방법은 웹과 같은 시맨틱 웹의 특성상 이기종(Heterogeneous) 분산(Distributed) 환경이기 때문에 필연적으로 수반되는 문제이다. 온톨로지는 서로 다른 의미들 간의 공통되는 의미적 이해와 상호호환성(Interoperability)을 지원하기 위해 사용된다. 생성되는 온톨로지 자체도 각각 이기종 환경을 포함하고 있다. 그러므로 온톨로지 매칭은 서로 다른 온톨로지간의 의미적 관계성을 발견하는 방식으로 진행되고 있다. 그러나 현재까지 온톨로지 구축 과정이 매우 어렵고, 각 분여별 전문가에 의존하여 구축되고 있어, 소수의 온톨로지간의 매칭 방식이 제안되고 있다.

온톨로지 매칭의 근본적인 제약은 매칭의 대상이 반드시 온톨로지라는 가정에서 수행되어왔다. 온톨로지간의 매칭은 도메인의 동질성, 구조 변경 가능성, 새로운 온톨로지 생성 등의 요인에 따라 언어 자원에 의한 기법, 확률적 추론 기법, 다양한 접근 방식의 통합 기법 등을 활용해서 매칭을 수행하고 있다.

본 논문에서 제안하는 경량 온톨로지 생성 방법은 구문기반 온톨로지를 의미기반 온톨로지로 매칭하여 대규모 온톨로지를 생성하는 방식이다. 따라서 매칭의 대상이 모두 온톨로지일 필요성은 없다.

3.1 경량 온톨로지의 단계별 생성

제안하는 경량 온톨로지 생성 과정은 크게 3 단계 즉, 구문기반 온톨로지 생성, 의미기반 온톨로지 생성, 온톨로지 수정/보완으로 구성되며, 이러한 과정은 일정한 주기로 반복적으로 구축될 수 있다. [그림 1]에서 기술한바와 같이 각 단계는 세부 단계를 포함하고 있다.



[그림 1] 경량 온톨로지 생성 단계

본 연구에서는 온톨로지 생성에 소요되는 인력 및 온톨로지 구축 시간을 최대한 단축시키기 위해 구문기반(Syntactic-based) 온톨로지 생성과 의미기반(Semantic-based) 온톨로지 생성의 단계별 과정을 통하여 연관어 기반 온톨로지 생성 과정으로 구성되어 있다. 웹 문서상에 존재하는 용어들(Terms) 사이에 포함된 구문적인 관계를 파악한 후 온톨로지에 매칭하여 의미 체계를 생성한다. 각 단계별 세부 설명은 아래와 같다.

첫 번째 단계인 구문기반 온톨로지 생성 과정은, 웹 상에 존재하는 다양한 문서들로 이루어진 웹 문서를 수집하여, 일정한 분류 단위 또는 도메인 등을 고려한 분야별 웹 문서를 대상으로 구문기반 온톨로지를 생성한다. 우선 연관어의 대상 추출을 위해서는 형태소 분석기, 앤그램(nGram) 방식 등을 이용하여 연관어 핵심 대상을 선택한다. 특히 이 과정에서는 불용어 및 중복어를 제거하여 순수

한 대상 연관어들 만을 추출하도록 한다. 다음으로 연관어들 간의 연관도 측정을 위해 MI(Mutual Information), TF(Term Frequency : 문서 내에서의 단어 출현 횟수)/IDF(Inversed Document Frequency : 전체 문서 중 단어가 출현한 문서의 개수의 역수), C-Value/NC-Value 등의 알고리즘 중 연관 관계를 가장 잘 측정할 수 있는 방법을 선택하다. 웹 문서를 대상으로 연관도를 측정하는 이유는 웹 문서상의 용어들 사이에 발행할 수 있는 관계인 본질적인 특성에 기인한다. 마지막으로 측정된 연관도를 임계치 조건을 만족하도록 정규화 한 후에 연관어 간의 관계들을 노드-아크(Node-Arc)를 갖도록 그래프 이론에 적용하여 연관어 그래프를 생성하는 과정을 포함한다.

두 번째 단계인 의미기반 온톨로지 생성은, 첫 번째 단계에서 생성된 구문기반 온톨로지를 온톨로지에 매칭하여 전체적으로 하나의 거대한 온톨로지(Big Ontology)를 생성한다. 첫 번째 단계에서 생성된 구문기반 온톨로지는 문서 관점에서는 의미 있는 관계(연관어 대상 간의 연결)로 판단할 수 있겠지만, 온톨로지 관점에서는 의미 체계가 없고 관계의 의미가 부여되지 않은 단순한 그래프이다. 따라서 시맨틱 웹을 위한 온톨로지 관점에서 인간과 기계가 이해 가능한 구문적인 관계와 의미적인 관계를 동시에 고려할 수 있도록 구문기반 온톨로지를 온톨로지에 매칭하여 전체적으로 거대한 의미기반 온톨로지를 생성한다. 대부분의 경우 온톨로지는 구문기반 온톨로지를 포함하는 개념체계이므로 소규모이고, 문서기반의 연관 어휘들은 대규모이다. 따라서 연관어 그래프 내의 용어들은 대부분 온톨로지 내의 개념체계 내의 용어에 속하는 속성을 활용하여 구문기반 온톨로지를 온톨로지에 매칭하여 대규모 의미기반 온톨로지를 생성한다.

세 번째 단계인 두 번째 단계를 통해 생성된 거대한 온톨로지를 개별 도메인 전문가가 수정/보완토록 하여 의미 기반 온톨로지를 생성한다. 구문기반 온톨로지를 온톨로지와 매칭을 통해 거대한 온톨로지로 생성되었으나, 도메인 전문가로부터의 수정 및 보완을 통해 최종적으로 특정 응용에 맞도록 정제될 수 있다.

위와 같은 과정으로 생성된 온톨로지는 온톨로지 저장소에 트리플(Subject, Predicate, Objects로 이루어진 SPO 형태의 단위) 형태로 저장되어, 시맨틱 웹 응용 애플리케이션 이용자 또는 시맨틱 웹 응용을 위한 소프트웨어 에이전트가 이용할 수 있는 상태가 된다.

3.2 구문기반 온톨로지와 의미기반 온톨로지의 매칭

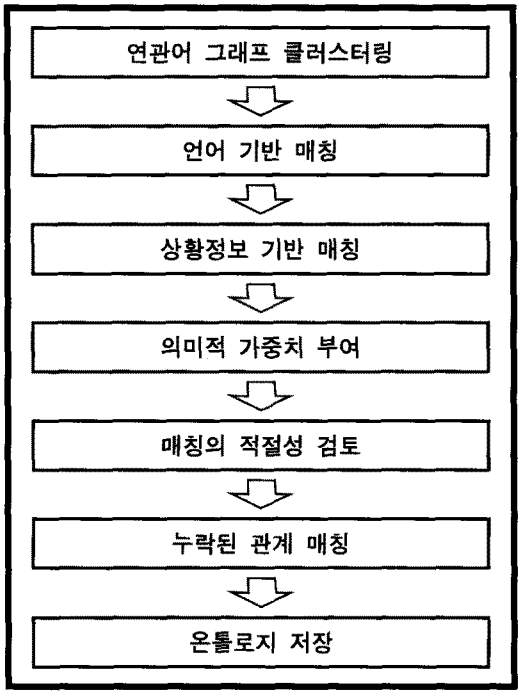
제안하는 구문기반 온톨로지를 의미기반 온톨로지로 생성하기 위한 연관어 그래프와 온톨로지의 매칭 방식은 기존 온톨로지간의 매칭 방식과 차이점이 존재한다. 우선 기존 온톨로지간의 매칭 방식은 매칭의 대상이 반드시 온톨로지로 제한하였다. 그러나 연관어 그래프와 온톨로지 간의 매칭에서는 온톨로지의 특성을 일부 가지고 있지 않으며, 설사 온톨로지와 비슷한 구조를 가지고 있더라도 온톨로지간의 매칭 기법을 그대로 활용할 수는 없는 상황이다. 좀 더 상세한 차이점은 아래와 같다.

첫째, 온톨로지간의 매칭은 각각의 온톨로지들이 의미적(Semantic) 관계도이지만, 연관어 그래프는 구문적(Syntactic) 관계도이다. 그러므로 연관어 그래프와 온톨로지 간의 매칭은 구문적 관계도와 의미적 관계도의 매칭 문제이다.

둘째, 온톨로지간의 매칭 대상이 되는 개별 온톨로지는 온톨로지 특성(예 : 개념, 속성, 관계, 제약조건, 공리, 인스턴스)을 포함하고 있지만, 연관어 그래프와 온톨로지 매칭에서의 연관어 그래프는 단지 단어(Term)들 간의 동시 발생 정도를 토대로 구성된 관계도이므로 온톨로지 특성을 보유하고 있지 않다. 그러므로 연관어 그래프와 온톨로지 간의 매칭은 온톨로지와 비온톨로지 간의 매칭 문제이다.

셋째, 온톨로지간의 매칭 문제에서는 모든 생성 온톨로지들이 완벽하게 의미적 관계(예 : 상하위 관계, 형제 관계 등)를 형성하고 있지만, 연관어 그래프와 온톨로지 매칭 문제에서는 연관어 대상들이 구문적 관계이므로 매칭 후 결과가 부분적으로만 의미적 관계를 형성하며, 부분적으로는 구문적 관계로 남아 있는 문제점을 내재하고 있다.

따라서 본 절에서는 인간의 개입 시간을 최소화하면서도 온톨로지 단축시간을 최대한 단축하기 위하여, 구문기반(Syntactic-based) 연관어 그래프와 의미기반(Semantic-based) 온톨로지간 매칭의 단계별 과정을 통하여 연관어 그래프와 온톨로지 간의 매칭을 기술하고자 하며, 앞 절에서 의미기반 온톨로지 생성과정을 좀 더 상세하게 설명하고자 한다. 이러한 연관어와 온톨로지 매칭은 웹 문서상에 존재하는 용어들 사이에 포함된 구문적인 관계를 포함하는 연관어 그래프를 클러스터링 한 후, 온톨로지를 포함하는 온톨로지와 언어기반 매칭, 상황정보 기반 매칭, 가중치 부여, 매칭의 적절성 검토 및 누락된 관계 매칭 과정을 포함한다. [그림 2]에서 기술한바와 같이 각 단계는 세부 단계를 제시하고 있다.



[그림 2] 연관어와 온톨로지 매칭

본 연구에서는 제안하는 연관어 그래프와 온톨로지 매칭 방안은 웹 문서상에 존재하는 수많은 용어들의 구문적 관계인 연관어 그래프를 이용하여 온톨로지에 매칭을 위하여, 연관어 그래프를 클러스터링하고, 언어기반 매칭 기법으로 1 : N의 관계를 형성한 후, 상황정보 기반 매칭 기법을 이용하여 1 : N의 관계를 의미적으로 분해한 1 : 1 관계로 만들고, 새롭게 형성된 연관어 그래프와 온톨로지 간의 관계에 가중치를 부여하는 과정으로 진행된다. 또한 연관어 그래프의 용어들이 일정하게 온톨로지와 매칭 되었는지 판단할 수 있도록 연관어 그래프 클러스터링 단위로 클러스터에 포함된 용어 중에 일부가 온톨로지에 일정하게 매칭 되었는지 검토하는 과정 및 연관어 그래프 클러스터링 내 용어들이 적어도 한 개의 용어도 온톨로지에 매칭되지 않았을 경우, 온톨로지 수정/보완 또는 누락된 연관어 클러스터 내의 일정 용어를 기 구축된 온톨로지에 인위적 매칭을 하는 과정도 포함하고 있다. [그림 2]에 기술되었듯이 각 단계별 과정은 아래와 같다.

첫째, 연관어 그래프를 클러스터링 한다. 우선 연관어 그래프를 클러스터링 하는 이유는 매칭의 대상이 되는 온톨로지 용어(개념, 속성, 관계, 인스턴스 등)와 연관어 그래프 용어(예 : 상호정보량 기반의 용어간 연결의 대상 용어들) 간의 숫자의 차이를 해결하기 위해서이다. 예를 들어 온톨로지의 용어 개수와 연관어 그래프의 용어 개수의 비율이 1 : 100 또는 1 : 1000 등으로 구성되어 있을 경우, 온톨로지의 용어가 연관어 그래프에 일정한 비율로 매칭 되었는지 판단하려고 클러스터를 수행한다. 연관어 그래프 구성 요소인 용어들은 전문가들에 의해서 생성된 온톨로지에 비해 다양한 용어를 포함하고 있고, 신조어, 복합어, 최신 유행어 들을 포함하거나 일정한 기간 동안만 사용되는 용어들이어서 객관적 관점에서 만들어진 온톨로지

용어와 매칭 되지 못하는 경우가 발생할 수 있다. 따라서 연관어 그래프의 연관도에 최신 용어들만으로 구성되어 있는 부분이 있고, 이 영역이 클러스터링 되어 있다면 이러한 영역은 온톨로지에 인위적인 매칭을 하기 위해 연관어 그래프의 클러스터링 단위로 온톨로지와 매칭 되었는지를 판단하기 위해 연관어 그래프를 클러스터링 한다. 연관어 그래프는 연관계수를 이용하여 HCS(Highly Connected Sub-graph) 또는 Chamelon 등의 방법을 통해 클러스터링을 수행할 수 있다.

둘째, 언어기반 매칭에서는 연관어의 용어와 온톨로지의 용어들을 언어적 관점에서만 매칭하는 기법이다. 크게 문자열 기반(String-based) 단계와 언어자원(Linguistic Resource) 단계를 포함한다. 우선 문자열 기반 단계에서는 유사한 용어는 유사한 이름 혹은 표현을 사용한다고 가정한다. 예를 들어 온톨로지의 컨셉, 인스턴스 등에서의 표현과 연관어 그래프에서의 연관어 대상 용어간의 문자열을 비교한다. 이 경우에는 정확하게 매칭되는 경우를 1 : N 관계로 파악한다. 언어자원 단계에서는 언어적 자원(Linguistic Resource)을 활용하여 용어의 의미와 용어간 관계를 파악한다. 언어적 관계어(유의어, 동의어, 약어 등)를 참조하여, 온톨로지 컨셉, 인스턴스의 표현과 연관어의 대상이 되는 용어간의 매칭 관계를 1 : N으로 파악한다.

셋째, 상황정보 기반 매칭에서는 연관어의 용어와 온톨로지의 용어들을 의미적(Semantic) 기반으로 매칭하는 단계이다. 상이한 상황정보는 상이한 구조(Structure), 속성(Property), 관계(Relation)로 표현하므로 온톨로지 주변 컨셉 및 속성 등과 연관어 그래프의 1차적, 2차적 연관 용어를 상황정보로 활용하여 매칭을 수행한다. 우선 온톨로지의 상황정보로는 대상 용어들의 상위 관계(Super Class) 컨셉, 형제 관계(Sibling) 컨셉, 대상 용어의 속성(Property)과 객체(ObjectProperty) 및 인스턴스에 해당하는 용어인 경우는 최상위 컨셉의 관련 정보를 대상으로 한다. 반면에 연관어 그래프의 상황정보로는 연관어의 대상 용어의 1차적인 연관어 중에 한 개 및 선택된 1차적 연관어의 연관어들을 상황정보로 활용한다. 이러한 연관어 그래프와 온톨로지의 매칭시 가장 높은 유사도를 갖는 매칭을 선택하고, 나머지 매칭된 연결(Link)을 해제하여 1 : 1의 관계로 매칭을 수행한다. 이 단계에서 가장 높은 유사도 매칭의 선택을 위해서는 Dice Coefficient Function 등을 이용할 수 있다. 예를 들어, 온톨로지 특정 용어의 상황정보와 연관어 그래프의 특정 용어에 해당하는 상황정보의 합집합을 분모로 하고, 온톨로지의 특정 용어의 상황정보와 연관어 그래프의 특정 용어에 해당하는 상황정보의 교집합을 분자로 하여 가장 유사도가 높은 매칭 관계만을 선택한다. 상황정보 기반 매칭 기법의 시맨틱 매칭의 다음 단계로 제약조건기반(Constraint-based) 단계이다. 이 단계에서는 상이한 상황정보는 상이한 구조, 속성, 관계로 표현하기 때문에 매칭의 연결(Link)이 존재하는 용어(x)와 온톨로지 속성(y)이 일치하면서 용어(x)의 1차/2차 연관어(용어 (x)의 상황정보)가 온톨로지 속성(y)의 상황정보(속성 (y)의 Domain과 Range)까지 일치한다면 해당 용어(x)와 온톨로지 속성(y)는 거의 같은 대상을 간주하는 매칭 관계이다.

넷째, 의미적 가중치 부여 단계에서는 1 : 1 관계로 매칭 된 용어들 간의 관계에 의미적 가중치를 부여하는 단계를 포함한다. 전 단계를 통해 1 : 1 관계로 분해된 각각의 연결(Link)은 새롭게 생성된 관계/호(Relation/Arc)로써 온톨로지를 그래프로 간주하고 다양한 그래프 탐색 알고리즘 적용을 위한 기존 호 가중치에 상응하는(Correspondent) 연결 가중치를 부여할 필요가 있다. 이 경우 연결의 신뢰성을 고려한 가중치 부여 값이 요구된다. 기본적인 호 가중치 부여 기준은 [0, 1]사이에 부여하되, 상위클래스(Superclass) 관계 보다는 높은 값을 속성(Property)보다는 낮은 값의 부여이다. 왜냐하면 온톨로지는 전문가의 도메인 지식을 통해 구축된 개념, 속성, 인스턴스 등을 포함하고 있고, 이러한 전문 지식의 표현 중에서 상위 개념(Superclass)관계와 속성(Property)은 하위 개념

(subclass) 관계, 인스턴스(Instance), 연관어 용어들(Term) 보다는 개념적 수준의 용어라고 판단되기 때문이다. 그러므로 상위 클래스 관계 < 연결(Link) < 속성 < 하위클래스 관계 순으로 가중치를 부여한다.

다섯 번째, 매칭의 적절성 검토는 첫 번째 단계인 연관어 그래프 클러스터링 단계에서 클러스터링 된 단위, 즉 연관어 그래프 중 특정 용어로부터 Nth 차원적으로 연결되어 있는 연관어들로 구성된 연관어 그래프 클러스터링 단위 내에서 하나의 용어라도 온톨로지에 매칭 되었는지를 파악하는 단계이다. 예를 들어, 신조어와 연관되어 있는 연관어들로 구성된 연관어 그래프 클러스터들의 용어는 도메인 전문가의 지식에 기반하여 구축된 온톨로지에는 포함되어 있지 않을 수 있으며, 이러한 연관어 그래프 내 클러스터를 파악하는 단계이다.

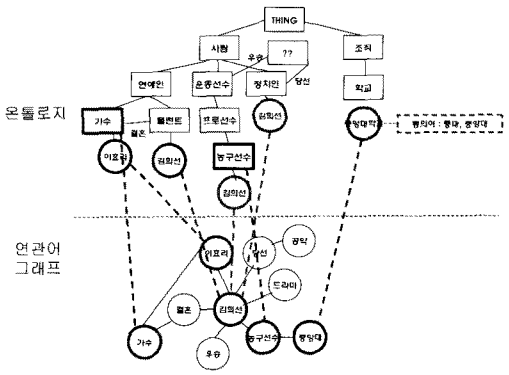
여섯 번째, 누락된 관계 매칭에서는 다섯 번째 단계에서 매칭의 적절성 검토에서 온톨로지에 매칭 되지 않았던 연관어 그래프 내 클러스터 되어 있는 용어들 중의 대표어를 온톨로지에 인위적으로 연결하는 단계이다. 왜냐하면 연관어 그래프는 구문적인 그래프이고 온톨로지는 의미적 그래프인데 상호 매칭을 통해 하나의 대규모 온톨로지를 구성하려면, 연관어 그래프의 용어들이 일정한 분포로 온톨로지의 용어에 매칭 되어야 하기 때문이다.

마지막으로, 연관어 그래프와 온톨로지 매칭을 통해 구축된 거대한 온톨로지는 온톨로지 저장소에 트리플 형태로 저장되어 이용자 또는 시맨틱 웹 애플리케이션이 이용할 수 있다.

4. 구현 사례

본 장에서는 제 3장에서 제시한 경량 온톨로지의 단계별 생성 과정을 실제 사례를 통해 예증해 보이고자 한다.

먼저 구문기반 온톨로지 생성과정은 제 3장에서 상세하게 설명하였으므로 의미기반 온톨로지 생성과정만을 중심으로 예증해 보이겠다. [그림 3]은

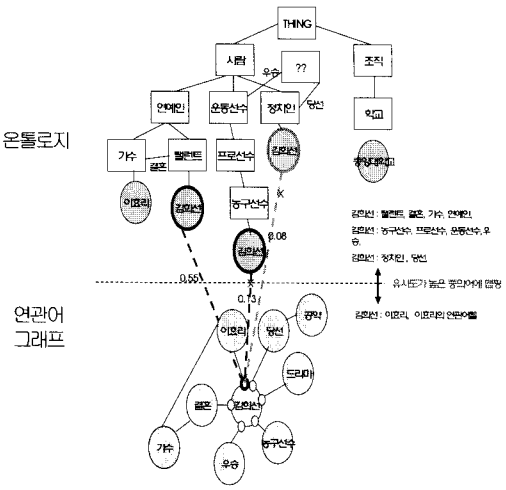


[그림 3] 온톨로지와 연관어 그래프 영역

실제 대규모 온톨로지 구축 프로젝트의 일부분을 발췌한 내용이다. 온톨로지 영역과 연관어 그래프 영역으로 구분되어 있는 부분들을 매칭 하여 하나의 대규모 온톨로지를 구축하고자 한다. 연관어 그래프 영역의 '김희선'이라는 용어의 경우 온톨로지 영역의 '탤런트 김희선', '농구선수 김희선', '정치인 김희선'과 1 : N의 관계로 매칭되어 연결(Link)된다. 또한 연관어 그래프의 '농구선수'는 온톨로지의 프로선수라는 클래스의 하위 클래스인 '농수선수'로 매칭 되고, 연관어 그래프의 '중앙대'는 온톨로지의 학교라는 인스턴스인 '중앙대학교'의 동의어 '중대', '중앙대'에 매칭 된다. 특히 1 : N 관계로 매칭 된 '김희선'의 경우는 온톨로지의 어떤 클래스의 인스턴스에 해당하는지 분리해야 한다. 그러나 언어기반 연관어와 온톨로지 매칭 단계에서는 3장에서 상술한 바와 같이 문자열 기반(String-based) 단계와 언어자원(Linguistic Resource) 단계를 거치면서 정확하게 매칭되거나 언어적 관계(유의어, 동의어, 약어 등)에 의미적 유사도로 매칭 될 경우 1 : N 관계로 연결(Link)하며 매칭 결과 값은 '0' 또는 '1'로 구분한다.

다음으로 [그림 4]는 상황정보기반 연관어와 온톨로지 매칭 예시도 이다. 온톨로지 영역과 연관어 그래프 영역으로 분리되어 있으나, [그림 4]에서 알 수 있듯이 언어기반 연관어와 온톨로지 매칭된 결과를 상황정보를 토대로 분리하고자 한다. 왜냐하면 온톨로지에 특정 용어(개념, 속성, 인스

턴스 등) 값을 찾고자 할 때 연관어 그래프와 여러 개가 매칭되어 있지만 연관어 그래프의 용어는 온톨로지의 특정 용어와 관계(Relation)이지 모두와의 관계를 의미하지 않기 때문이다. 예를 들어, [그림 4]의 연관어 그래프 용어인 '김희선'은 언어기반 연관어와 온톨로지 매칭 결과로는 1 : 3의 관계로 연결되어 있으나, 온톨로지의 상황정보와 연관어 그래프의 상황정보를 통해 의미적(Semantic)으로 분리할 수 있다.

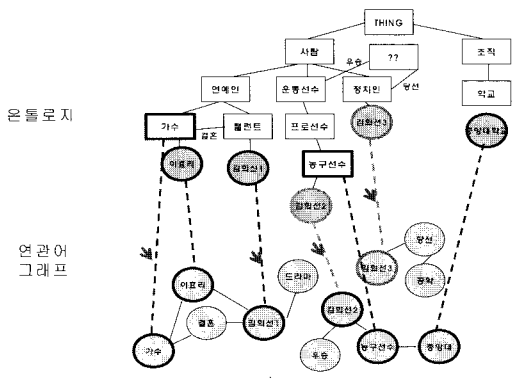


[그림 4] 상황정보기반 연관어와 온톨로지 매칭 예시도

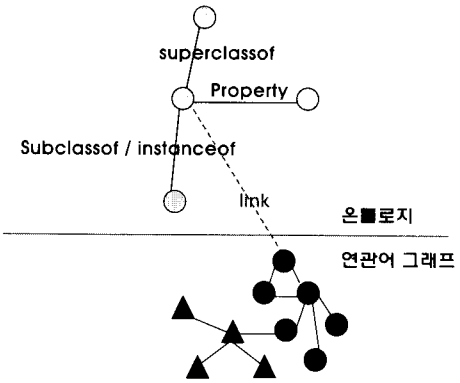
온톨로지 상황정보는 대상 용어의 상위(Super-class) 컨셉, 형제 컨셉(Sibling), 관계(Property, Object-Property), 인스턴스의 상위 컨셉 등을 활용하고, 연관어 그래프 용어의 상황정보는 1차적 연관어 중 한 개와 선택된 1차적 연관어의 연관어들을 활용할 수 있다

[그림 4]에서 연관어 그래프의 용어 '김희선'은 1차적으로 '이효리'와 선택된 1차적 연관어의 연관어인 '가수', '결혼' 등을 상황정보로 활용할 수 있다. 반면에 온톨로지의 용어 중에 하나인 '탤런트 김희선'의 상황정보는 '김희선'이라는 인스턴스의 컨셉인 '탤런트', 탤런트의 속성인 '결혼', 탤런트와 결혼 관계인 컨셉인 '가수', 탤런트의 상위 컨셉인 '연예인' 등을 온톨로지의 상황정보로 활용할 수 있다. 결과적으로 연관어 그래프의 용어인 '김희선'의 일부 의미인 온톨로지의 탤런트 '김희선'과 연결이 되고 연관어 그래프의 용어인 또 다른 '김희선'의 경우도 유사도가 높은 순서로 매칭 할 수 있다. 연관어 그래프의 다른 '김희선'의 경우에도 김희선-당선-공약의 경우와 온톨로지의 김희선-정치인-당선 등으로 반복하면서 유사도가 높은 순으로 매칭이 가능하다. 매칭 방식에 대해서는 Dice Coefficient Function 등을 활용할 수 있다.

다음으로 [그림 5]에서 온톨로지와 연관어 그래프의 분리된 예시도이다. 온톨로지 영역과 연관어 그래프 영역으로 분리되어 있으나, [그림 4]의 설명에서도 언급하였듯이 상황정보기반 연관어와 온톨로지 매칭된 결과를 토대로 각각의 용어를 의미적(semantic)으로 상세하게 분리한 후, 각각의 연결된 관계가 의미적으로 거의 일치하는지를 추가 확인하는 과정이다. 전술한 단계를 통해 매칭된 경우 각각의 연결(Link)의 대상이 되는 연관어 그래프의 용어와 온톨로지의 용어간 의미적으로 거의 일치하는지를 파악하는 방식이다. 예를 들어, 연관어 그래프의 용어인 '가수'와 온톨로지의 용어 '가수'는 연관어 그래프의 가수 상황정보에서 '결혼'이라는 1차적인 연관어가 있고, 온톨로지의 용어 '가수'라는 컨셉의 속성(Property)로 '결혼'이 있기 때문에 두 용어는 거의 일치한다고 판단할 수 있다.



[그림 5] 온톨로지와 연관어의 매칭된 예시도



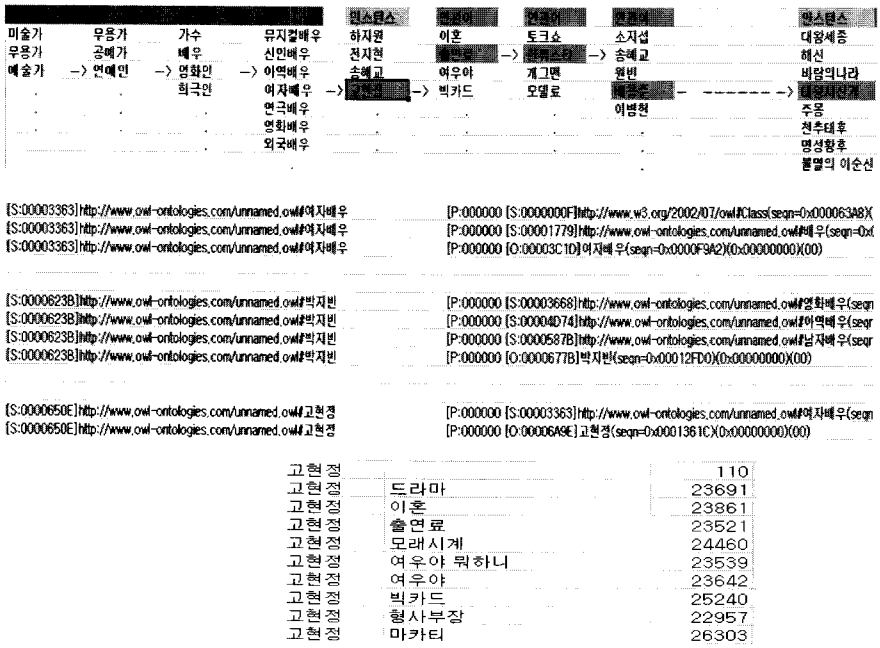
[그림 6] 온톨로지와 연관도 그래프의 연결 예시도

마지막으로 [그림 6]에서는 온톨로지와 연관어 그래프의 연결 예시도이다. 전술한 단계를 통해 온톨로지와 연관어 그래프가 1 : 1관계로 매칭되어 거대한 온톨로지로 활용하기 위해서는 연결 부분에 일정한 의미적 가중치를 부여하여 하나의 동질성 있는 온톨로지로 생성하고자 한다. 기본적으로 의미적 가중치를 부여하는 방식은 연결(Link) 가중치가 온톨로지의 속성(Property) 가중치보다 낮게 부여한다(상위(Superclass) 관계 < 연결(Link) < 속성(Property) <하위(subclass) 관계). 또한 상황정보기반 기법 중에 하나인 제약조건 기반(도 2의 상황정보 기반 매칭 설명 참조) 기법 체크 결과를 반영하여 연결(Link) 가중치를 구분하며, 유효한 값으로 체크된 연결의 경우 온톨로지의 속성 가중치와 대응한 값을 부여한다. 이 경우라도 온톨로지의 속성 최대 가중치 값보다는 낮게 부여하며, 문서를 가진 속성의 가중치 값보다는 낮게 부여한다. 만약 온톨로지의 속성 값이 존재하지 않는 경우는 아래 2)의 경우처럼 가중치를 부여하되, 특히 연결사례 2의 경우는 보정계수 값을 추가해야 한다.

[그림 6]의 연결(Link) 가중치 부여 함수는 아래와 같다.

4.1 온톨로지 속성이 한 개 이상 존재하는 경우

- 연결(Link) 사례 1 : 상황정보의 제약조건 기반 체크결과 유효한 경우
  - [min] $P_w < L_w <$ [*min*] ([*max*] P_w, [*min*] P_d)



[그림 7] 온톨로지 스키마와 연관어와의 관계 및 트리플 저장 구조

- 연결(Link) 사례 2 : 상황정보 제역조건 기반 체크결과 유효하지 않은 경우
  - 상위 < L_w < [*min*] P_{w*}

    L_w : link 가중치,

    P_w : Property 가중치,

    P_d : 문서를 가진 Property 가중치

4.2 온톨로지 속성이 존재하지 않는 경우

- 연결(Link) 사례 1 : 상황정보의 제약조건 기반 체크결과 유효한 경우
  - 상위 < L_w < 하위
- 연결(Link) 사례 2 : 상황정보 제역조건 기반 체크결과 유효하지 않은 경우
  - 상위 < L_w^* 보정계수< 하위

[그림 7]은 연관어 기반 경량 온톨로지 생성의 결과물 예시로서 온톨로지 스키마와 연관어가 연결된 결과이므로 트리플 형태로 저장될 수 있음을 알 수 있다. 스키마에 연예인 > 배우 > 여자배우 > 고현정으로 구성되어진 온톨로지 스키마와 인스턴스 예시가 상단에 표시되어 있고, 이러한 개념 체계가 온톨로지 스키마의 트리플 구조가 트리플 형태로 저장되는 예시(중간)와 인스턴스로 존재하는 고현정의 대상 연관어 예시(하단)를 확인할 수 있다.

전술한 바와 같이, 본 논문에서는 연관어 그래프와 온톨로지 간의 매칭으로 인해, 기존의 인간의 개입이 많고, 구축 시간이 많이 소요되는 방식보다 시맨틱 웹 응용이 가능한 실생활에서의 가치를 더할 것으로 기대된다.

5. 결 론

본 논문에서는 웹상의 데이터의 의미를 처리할 수 있는 시맨틱 웹 기반 구조를 구성하기 위해 의미기반 온톨로지와 구문기반 온톨로지로 정의한 연관어 그래프 간의 매칭 기법을 적용한 경량 온톨로지 생성에 관한 단계별 과정을 살펴보았다. 또한 기존 수작업 형태의 온톨로지 생성 부분을 (반)자동적으로 발전시키고 특정 도메인이 아닌 넓은 도메인의 지식을 포함하는 대규모 온톨로지를 생성하여 기존의 취약점을 해결하였다.

따라서 본 연구는 학문적인 측면에서 주어인 온톨로지 스키마에 인스턴스를 자동으로 생성하거나, 특정 영역에 해당하는 데이터를 분석하는 기법을 통해 온톨로지를 정적으로 구축하는 방식과는 차별성이 있다. 우선 온톨로지는 진화할 수 있다는 가정으로, 수시로 변화할 수 있는 구문기반 온톨로지를 의미기반 온톨로지의 선결 작업 대상으로 포함 하였다. 또한 온톨로지 병합 대상을 온톨로지와 비온톨로지간의 매칭 기법으로 통해 보편적이고 대용량의 온톨로지 구축을 추구하였다.

또한 실무적인 측면에서는 기존 소규모로 구축된 온톨로지로 인한 이용의 제한성과 너무 복잡한 로직 등을 통한 구축과정의 어려움을 해소하고자 하였다. 따라서 거대한 지식을 하나의 범용적인 시스템으로 상호 연결 할 수 있도록 하여 다양한 지식 서비스 제공에 용이토록 기여하였다.

그러나 본 논문에서 제안하는 방법으로 생성된 온톨로지는 모든 개념과 속성 및 관계 체계를 활용하지 않아 복잡한 온톨로지 추론이 요구되는 시맨틱 응용 분야에 적용할 수 없는 부족한 점이 있어 본 논문에서 제시된 과정을 통해 생성된 경량 온톨로지를 보완하는 추가 작업이 필요하다.

향후, 경량 온톨로지를 보완하여 복잡한 서비스 처리를 위해 온톨로지 추론과 시맨틱 그래프 알고리즘을 적용하고, 더 나아가 개념들의 제약사항을 고려할 수 있는 서비스 개발에 관한 연구를 진행할 계획이다.

참 고 문 헌

[1] 김왕석, 변영태, "Enterprise Architecture를 위한 시맨틱 웹 기반의 온톨로지 설계 및 구현", 「한국IT서비스학회지」, 제7권, 제3호(2008),

pp.239-252.

[2] 한동일, 권혁인, 최호준, "시맨틱 검색 시스템의 구현과 평가에 관한 연구", 「한국IT서비스학회지」, 제7권, 제3호(2008), pp.253-269.

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

[4] Cannataro, M., and C. Comito, "A Data Mining Ontology for Grid Programming", *1st Workshop on Semantic in Peer-to-Peer and Grid Computing at the Twelfth International World Wide Web Conference*, 2003.

[5] Chen, H., and Z. Wu, "OKSA : an Open Knowledge Service Architecture for Building Large Scale Knowledge System in Semantic Web, Systems, Man and Cybernetics", *IEEE International Conference on 5-8,* Vol.5(2003), pp.4858-4863.

[6] Davis, M., "Industry Roadmap to Web 3.0 and Multibillion Dollar Market Opportunities", *Project10X's Semantic Wave 2008 Report.* 2008.

[7] De Roure, D., N. Baker, N. R. Jennings, and N. R. Shadbolt, "The Evolution of the Grid", *Technical Report of the National e-Science Centre*, UKeS-2002-02, 2002.

[8] De Roure, D., N. Nicholas, N. R. Jennings, and N.R. Shadbolt, "The Semantic Grid : A Future e-Science Infrastructure", *Grid Computing : Making the Global Infrastructure a Reality*, 2003.

[9] Ding, Y., and R. Engels, "IR and AI : using co-occurrence theory to generate lightweight ontologies", *Database and Expert Systems Applications, Proceedings. 12th International Workshop*, 2001, pp.961-965.

[10] Doan, A., J. Madhavan, P. Domingos, and A. Y. Halevy, "Ontology Matching : A Machine Learning Approach", *Handbook on Ontologies*, 2004, pp.385-404.

[11] Gomez-Perez, A. and O. Corcho, "Ontology Language for the Semantic Web", *IEEE,* Vol.17(2002), pp.54-60.

[12] Gruber, T., "It Is What It Does : The Pragmatics of Ontology, invited talk at Sharing the Knowledge", *International CIDOC CRM Symposium*, 2003, Washington, DC.

[13] Han, D. I., S. B. Ha, and H. J. Choi, "Fox Service : An Implementation Case of Ontology-based Search Agent in Mobile Environments", *Mobile Data Management*, 7th International Conference, 2005, pp.85-85.

[14] Han, J., and K. C. Chang, "Data Mining for Web Intelligence", *IEEE Computer*, Vol.35, No.11(2002), pp.64-70.

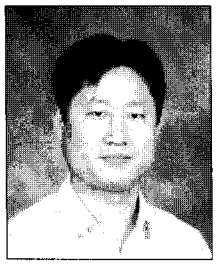
[15] Liu, J., N. Zhong, Y. Yao, and Z. W. Ras, "The Wisdom Web : New Challenges for Web Intelligence(WI)", *Journal of Intelligent Information Systems,* Vol.20, No.1(2003), pp.5-9.

[16] Noy, N., "Ontology Mapping and Alignment", *SSSW-2005*, 2005.

[17] Sheth, A. "From Semantic Search and Integration to Analytics", *Dagstuhl Seminar on Semantic Interoperability and Integration*, 2004.

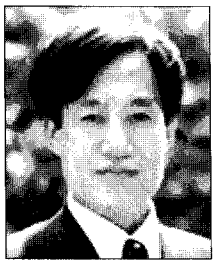
[18] Shvaiko, P. and J. Euzenat, "Tutorial on Ontology Mathcing", *SWAP-2006*, Pica, Italy, 2006.

◈ 저 자 소 개 ◈



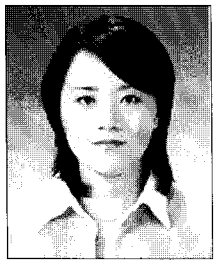
한 동 일 (dihan@kt.com)

중앙대학교에서 경영학 석사를 하고, 동 대학에서 경영정보전공 박사를 받았다. 현재 KT 미래기술연구소에서 시맨틱 웹 분야를 연구 중이다. 주요 관심분야는 시맨틱 웹, 웹2.0, 시맨틱 웹 서비스 분야 등이다.



권 혁 인 (hikwon@cau.ac.kr)

중앙대학교에서 컴퓨터공학 석사를 하고, 파리 제6대학에서 전자계산학 박사를 받았다. 현재 중앙대학교 상경학부 교수로 재직 중이며 주요 연구관심 분야로 서비스사이언스, 비즈니스 모델, 게임 켄텐츠 등이 있다. 또한 한국데이터베이스학회, IT서비스학회 등에 논문을 게재하였으며, 디지털콘텐츠 생산 및 유통 기반 구축사업, 게임 산업 인력양성사업 등 다수의 프로젝트 연구책임자로 연구를 수행하였다.



백 선 경 (sk100@kt.com)

조선대학교에서 정보컴퓨터교육학 석사를 하고, 동 대학에서 전자계산학 박사를 받았다. 현재 KT 미래기술연구소에서 시맨틱 웹 분야를 연구 중이다. 주요 관심분야는 시맨틱 웹, 시맨틱 정보 검색, 감성정보처리, HCI 분야 등이다.