시맨틱 웹 기반의 고객 정보 검색 시스템의 설계 및 구현

황 정 희[†] · 구 미 숙[††] · 이 현 아[†††] · 류 근 호[††††]

요 약

시맨틱 웹 기반의 서비스를 제공하기 위해서는 특정 도메인 지식에 대한 명시적인 명세화 및 지식의 개념과 개념과의 관계를 정형화하는 온톨로지를 통해서 이루어진다. 그러므로 특정 도메인 지식의 명세화와 정형화를 위해서는 관심 있는 도메인내의 지식을 개념화하는 온톨로지의 생성이 필수적이다. 이 논문에서는 택배 마케팅의 잠재 고객에 대한 정보를 검색하기 위해, 시맨틱 웹 기반의 온톨로지 생성을 위한 구체적인 도메인을 설계하고, 생성된 온톨로지를 이용하는 정보 검색 방법을 제안하였다. 제안된 온톨로지를 기반으로 고객의 정보를 수집하는 데이터 검색 로봇을 구현 하였다. 아울러 온톨로지의 생성과 데이터 검색 로봇이 데이터 검색을 정확하게 수행함을 확인 하였다.

키워드 : 온톨로지, 시맨틱 웹, 검색 에이전트, XTM

Design and Implementation of Customer Information Retrieval System based on Semantic Web

Jeong Hee Hwang[†] · Mi Sug Gu[††] · Hyun Ah Lee[†††] · Keun Ho Ryu[††††]

ABSTRACT

Ontology specifies the knowledge in a specific domain and defines the concepts of knowledge and the relationships between concepts. It is possible to provide the service based on the semantic web through the ontology. Therefore, to specify and define the knowledge in a specific domain, it is required to generate the ontology which conceptualizes the knowledge. Accordingly, to search the information of potential customers for home-delivery marketing of post office, we design the specific domain to generate the ontology based on the semantic web in this paper. And we propose how to retrieve the information, using the generated ontology. We implement the data search robot which collects the information based on the generated ontology. Also, we confirm that the ontology and the search robot perform the information retrieval exactly.

Key Words : Ontology, Semantic Web, Search Agent, XML Topic Maps

1. 서 론

현재의 웹 검색엔진은 단어의 빈도수나 어휘 정보를 이용하여 문서의 유사도를 측정하고 순위를 부여하는 방식을 이용하고 있어서, 사용자는 유용한 정보를 찾기 위해 많은 시간을 낭비하게 된다. 이러한 문제점을 개선하기 위해 최근 차세대 웹으로 시맨틱 웹이 등장하였다. 시맨틱 웹은 정보의 의미를 개념으로 정의하고 개념간의 관계성을 명시화하여 웹 문서에 의미 정보를 덧붙이고, 소프트웨어 에이전트가 이 의미정보를 자동으로 검색하여 정보를 제공한다[1].

※ 이 논문은 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음.
† 정 회 원 : 남서울대학교 컴퓨터학과 전임강사
†† 준 회 원 : 충북대학교 대학원 전자계산학과 박사과정
††† 정 회 원 : 한국전자 통신연구원 우정기술센터 연구원
†††† 종신회원 : 충북대학교 전기전자 컴퓨터공학부 교수
논문접수 : 2005년 5월 25일, 심사완료 : 2006년 3월 3일

시맨틱 웹 기반의 서비스 제공은 특정 도메인 지식에 대한 명시적인 명세화 및 지식의 개념과 개념과의 관계를 정형화하는 온톨로지를 통해서 이루어진다. 온톨로지는 어느 특정 도메인에 관련된 단어들을 계층적 구조로 표현하고, 추가적으로 이를 확장할 수 있는 추론 규칙을 포함한다. 온톨로지의 역할 중 하나는 서로 다른 데이터베이스가 같은 개념에 대해서 서로 다른 단어나 식별 자를 사용할 경우에 이를 해결해 주는 것이다. 그러므로 두 데이터베이스의 정보를 비교, 통합 할 때 온톨로지를 통해서 단어의 의미에 대한 비교가 이루어지므로 온톨로지의 생성은 시맨틱 웹 기반의 검색에서는 기본적이고 필수적이다.

이 논문에서 택배 마케팅을 위한 관련 정보들을 온톨로지의 도메인으로 한다. 택배 마케팅은 택배를 주 업무로 하는 택배회사나 우체국에서 수익 창출을 위한 마케팅 전략중의 하나로, 기존의 웹에 있는 쇼핑 사이트들의 정보를 추출하

여, 이러한 정보를 기반으로 그 쇼핑 사이트에 대한 택배 마케팅을 할 수 있다. 쇼핑 사이트가 취급하는 상품과 그 회사가 위치하고 있는 지역과 전화번호 이 메일(e-mail) 주소 등의 정보를 입수하여 마케팅을 할 수 있게 되는 것이다. 이 논문에서 잠재 고객이라고 명칭을 한 대상은 이와 같은 쇼핑 사이트들을 의미한다.

그러므로 이 논문에서 택배 마케팅의 잠재 고객에 대한 정보를 검색하기 위해, 시맨틱 웹 기반의 온톨로지 생성을 위한 구체적인 도메인을 설계하고, 생성된 온톨로지를 이용하는 정보 검색 방법을 제안한다. 즉, 택배 마케팅을 위한 시맨틱 웹 기반의 지능적인 정보 검색을 위해, 웹에 등록되어 있는 인터넷 쇼핑 사이트와 관련된 상점의 정보 및 취급 상품 등에 관한 구체적인 잠재 고객 도메인의 정보를 수집할 수 있는 데이터 검색 로봇[2]을 구현하였다. 아울러 온톨로지를 기반으로 하는 검색 시스템을 설계하고 구현하여 온톨로지의 생성과 온톨로지를 이용한 정보 검색의 수행을 보인다.

이 논문의 구성은 다음과 같다. 제 2장에서 온톨로지 기반의 정보 검색 기술과 기존연구 및 온톨로지 언어에 대하여 기술한다. 그리고 제 3장에서는 시맨틱 웹 기반의 고객 정보 탐색 시스템을 위한 온톨로지 생성의 도메인 범위 및 설계 내용에 대하여 설명한다. 제 4장에서는 고객 정보 탐색 시스템을 구현하기 위한 온톨로지와 에이전트의 설계 및 온톨로지 기반의 정보 검색 시스템의 전체 프레임 워크의 설계 내용을 기술한다. 그리고 제 5장에서는 온톨로지를 이용하는 검색 시스템의 구현 및 검색 결과를 기술하고 결론을 맺는다.

2. 시맨틱 웹 기반의 고객 정보 탐색

시맨틱 웹은 기존의 웹과 완전히 구별되는 새로운 웹의 개념이 아니라 현재 웹을 확장하여 웹에 있는 정보에 잘 정의된 의미를 부여하고 이를 통해 컴퓨터와 사람이 협동적으로 작업을 수행할 수 있도록 하는 패러다임이다[1,2]. 이 장에서는 온톨로지 기반의 정보 검색에 대한 기존연구 및 온톨로지 언어에 대한 비교 설명을 한다.

2.1 온톨로지 기반의 정보 검색

시맨틱 웹은 웹 상의 정보에 잘 정의된 의미를 부여함으로써 사람뿐만 아니라 컴퓨터도 쉽게 문서의 의미를 해석할 수 있도록 하여 컴퓨터를 이용한 정보의 검색, 해석 및 통합 등의 업무를 자동화하기 위한 목적으로 제안되었다. 일반적으로 시맨틱 웹은 이러한 지식의 정의와 관련된 온톨로지 연구와, 웹 자원을 서술하기 위한 RDF 및 RDFS와 같은 연구, 그리고 이를 활용하기 위한 자동화 된 자율적 프로그램인 에이전트에 관한 연구 등을 포함한다[3, 4].

시맨틱 웹의 핵심 개념인 온톨로지는 "개념화의 규정"(specification of a conceptualization)을 말한다. 즉, 해당 영역의 개념들과 이들 개념간의 상호 관계를 정의하는 것을 온톨로지라고 한다[5, 6]. 개념화란 대상으로 삼고 있는 세계에서 일어나는 현상에 연관된 개념들을 파악하기 위한 추상적 모델을 말한다. 명시적이란 것은 개념의 사용 유형과 사용된 유형의 제약조건이 명시적이란 것을 의미한다. 그리고 형식적이란 것은 기계가 읽을 수 있어야 한다는 것을 말하고, 공유는 온톨로지가 표현하는 개념이 개별적이 아닌 해당 그룹 구성원간에 합의된 지식에 바탕을 두고 있다는 것을 의미한다.

온톨로지를 사용하면 어떠한 효용이 있는지에 대해, 예를 들어 구글(Google)과 야후(Yahoo) 검색 엔진을 비교해 볼 수 있다. 구글은 기본적으로 사용자가 제시한 주제어를 가지고 이 주제를 포함한 웹 문서를 기계적으로 검색하여 제시한다. 사용자의 "의도"를 반영하기보다 문서의 외형적 특징에 의존한 기계적 계산 결과인 경우가 많다. 반면 야후는 사람이 미리 정의한 주제들과 이들을 상속 관계나 부분-전체 관계를 써서 계층적으로 분류한 주제 계통을 가지고 웹 문서들을 이 주제에 맞게 분류하고 이들 간의 연관 관계를 미리 정의하여 놓았다. 검색의 예를 들어 설명하면, 여행 정보를 찾고자 하는 경우, 야후는 "1. 여가생활과 스포츠>여행, 관광 2. 비즈니스와 경제>기업간거래(B2B)>여행,관광 3. 엔터테인먼트>음악>음악감상실>테마별감상>여행 4. 비즈니스와 경제>취업, 채용 >회사별 >여행, 교통"과 같은 4개의 카테고리를 제시하고 각 카테고리에 하위 카테고리를 계층적으로 두어 해당 정보를 찾아갈 수 있게 한다. 반면에 "여행"이라는 검색어를 가지고 구글에서 찾게 되면 2,130,000개의 관련 문서를 제시한다. 검색 목적에 따라 구글과 같은 검색이 유용한 경우도 많지만 특정 의도에 적합한 문서를 찾고자 할 때는 야후와 같은 주제 분류가 유용할 것이다. 이와 같이 야후의 계층적 카테고리는 계층적으로 정의한 온톨로지를 기반으로 문서를 분류하고 검색하는 것과 유사하다[25].

이와 같이 시맨틱 웹에서의 온톨로지 역할이 증가함에 따라 현재의 웹 구조에 시맨틱 웹 기술을 결합한 시맨틱 웹 기반의 검색 시스템이 국내외적으로 중요한 정보를 빠르게 찾고, 정보 검색에 정확도를 향상시킬 수 있다는 점에서 중요한 기술로 자리 잡아 가고 있다. 또한 온톨로지 기반의 정보 검색은 온톨로지에 정의된 개념과 규칙을 활용하여 검색 향상을 위한 추론 규칙을 이용하기 때문에 단순히 사용자의 질의와 일치하는 문서만 보여주는 것이 아니라 사용자의 질의에 대한 의미 분석을 통해 그와 관련된 정보를 온톨로지에 표현된 관계에 따라 적절한 질의의 수정도 가능하다.

최호섭외[2004]에서는 시맨틱 웹 기반 검색 시스템 구조를 제시하였다. 이 시스템 구조는 서브시스템인 검색 엔진과 온톨로지 시스템으로 구성된다[25]. 그리고 Pepper[2000]에서는 토픽맵 모델의 세가지 핵심 요소인 토픽(Topic), 어커런스(Occurrence), 어소시에이션(Association)등에 대한 설명과 관계에 대해서 "opera"에 대한 토픽맵을 예를 들어 설명하고 있다[17]. 그리고 XTM을 이용한 온톨로지 구축의 또다른 예로써, 정호영외[2003]에서는 토픽맵을 이용하여 교

수, 학생, 논문간의 관계를 정의하여 온톨로지를 작성하여, 이 온톨로지를 이용하여 검색을 좀 더 용이하게 하는 지식맵을 구현하였다[23]. 그리고 김정민외[2004]에서는 토픽맵 모델을 기반으로 온톨로지를 생성, 저장, 검색하는 온톨로지 관리 시스템인 K-Box를 구현하였다[24]. Lin[2002]는 불경의 버전에 대한 작성자, 버전, 시대, 장소간의 관계를 정의하여 토픽 맵 온톨로지를 구축하여 검색시스템을 구현하였다[22]. 기존연구의 토픽맵을 이용한 온톨로지를 구축한 검색 시스템은 기존에 존재하고 있는 지식이나 데이터를 이용한 검색 시스템이다. 기존 연구는 융통성이 결여되어 새로운 지식이 생성되면 실시간으로 반영하지 못하는 문제를 갖는다. 따라서 우리는 기존에 제시되지 못한 지식의 실시간 반영의 융통성을 제공하기 위하여 정보 검색 에이전트를 사용하였고 웹에 있는 데이터를 실시간으로 가져와서 그 데이터를 기반으로 하여 구축한 온톨로지를 사용한다.

2.2 온톨로지 언어

RDF(Resource Description Framework)는 W3C의 가장 기본적 시맨틱 웹 언어로서 웹에 있는 자원(resource)에 관한 메타 정보를 표현하기 위한 언어이다[7,8]. 특히 웹 자원을 표현하는 데 기본이 되는 제목, 저자, 저작권과 같은 웹 문서에 관한 메타 데이터를 표현할 목적으로 개발되었으나, 웹 자원의 개념을 웹상에서 다른 것과 구별하여 식별할 수 있는 대상으로 일반화하면 다른 목적으로도 RDF를 활용할 수 있다.

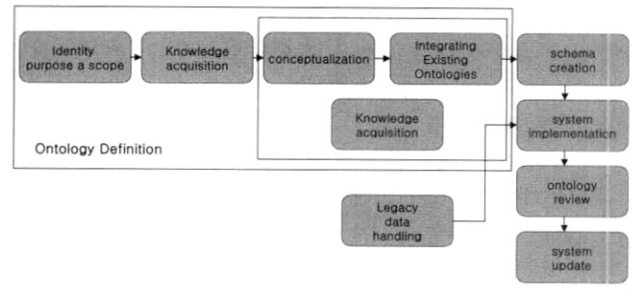
Ontology Inference Layer(OIL)는 On-To-Knowledge 프로젝트를 수행한 IST(Information Society Technologies)에 의해 개발되었다[9]. OIL은 온톨로지를 위한 웹에서 기계가 접근 가능한 온톨로지를 표현하며, 언어적인 측면과 XML 스키마, RDF 스키마 등과의 연계연구가 많이 이루어져 OWL까지 발전하는 기반이 되었다.

DAML_OIL은 웹 온톨로지 언어로서 DAML(DARPA Agent Markup Language) 프로그램의 DAML ONT와 주로 유럽에서 개발된 OIL의 결합을 통하여 만들어졌다[10-13]. 관심영역(domain)의 구조를 서술하기 위한 목적을 갖는데, 온톨로지는 클래스와 속성의 성격을 서술한 공리의 집합으로 구성된다. 그러므로 DAML+OIL은 기본적으로 표현력을 강조한 Description Logic이며, 클래스들은 URI로 지칭되는 이름이나, 클래스 수식을 만들기 위한 다양한 구성자(constructor)로 표현된 수식을 나타낸다.

DAML+OIL에서 사용 가능한 공리는 subClassOf, sameClassAs, subPropertyOf, samePropertyAs, disjointWith, inverserOf, transitiveProperty, uniqueProperty등이 있다.

토픽맵(Topic Map)은 XTM(XML Topic Maps)으로 표현하는데, 주제 중심으로 개념을 명세 화하고 개념들 간의 연관 관계를 정의하는 모델로서 ISO에서 표준으로 제안하고 있다[14-17]. RDF가 자원 중심인데 반해 토픽맵은 주제 중심이다. RDF는 URL로 접근 가능한 자원들에 대해 메타 데이터를 생성하기 위한 모델을 제시하고, 토픽맵은 개념이나 사물에 대하여 정형화된 명세를 생성하기 위한 모델을 제시한다. 토픽맵은 초기에는 전자 색인을 위한 데이터 모델로 고안되었으나 현재에는 지식 관리 시스템의 지식맵, 컨텐츠 관리 시스템의 컨텐츠 맵 그리고 시맨틱 웹의 온톨로지 등의 데이터 모델로 사용되고 있다.

토픽맵 모델의 핵심 요소는 토픽(Topic), 어커런스(Occurrence), 어소시에이션(Association)으로 볼 수 있다. 토픽은 객체 지향 모델에서의 클래스나 객체에 해당하는 것으로 표현하고자 하는 대상을 가리키고, 어커런스는 토픽에 종속되는 개념으로 해당 토픽에 대한 실제 내용이 남겨있는 자원의 주소(URI)나 지식 데이터 자체를 가리킨다. 그리고 토픽들 사이에는 토픽들 간의 연관성을 표현하는 어소시에이션이 있다. 온톨로지가 정의되면 용어를 토픽으로, 분류를 토픽 타입과 토픽간의 관계로, 연관 관계를 토픽들 사이의 어소시에이션으로 매핑하여 초기 토픽맵을 생성한다. (그림 1)은 온톨로지 정의 및 토픽맵 데이터 모델을 생성하는 과정에 대해 보여주고 있다[24].



(그림 1) 온톨로지 정의 및 데이터 모델 생성 과정

3. 고객 정보 탐색 시스템의 온톨로지 정의 및 설계

지식을 사람이나 기계들 사이에 공유하기 위해서는 상호간에 이해 가능한 용어들로 개념을 정의한 다음 표준화된 모델과 형식화된 언어로 표현해야 한다. 시맨틱 웹 기반의 고객 정보 탐색 시스템에서도 해당 도메인에 대한 특정 지식을 포함하는 온톨로지의 설계가 필요하다. 이 장에서는 고객 정보 탐색 시스템의 기반이 되는 온톨로지의 설계의 지식 도메인 범위의 설계 및 XTM으로 작성한 예를 보인다.

이 논문에서는 온톨로지를 구축하기 위해서 토픽맵을 사용하였다. W3C의 RDF가 웹상에 존재하는 자원 즉, 웹 페이지들 사이의 연결에 초점을 맞추고 있는 것에 반해 토픽맵은 형상화해야 하는 토픽의 대상이 웹 페이지, 그림, 전자문서와 같은 주소를 갖는 객체뿐만 아니라 철학, 역사, 이성, 도덕 등과 같은 추상적 개념의 지식들도 표현이 가능하다는 특징이 있기 때문이다. 특정 문서의 토픽은 그 문서의 작성자가 나타내고자 하는 주제를 표현할 수 있는 단어들로 구성되어 있다.

토픽맵 기반의 온톨로지 구축을 위해서 필요로 하는 온톨로지 설계의 기반이 되고 있는 세 가지의 토픽맵 모델에 대해 기술한다.

첫째는 토픽으로써, 이 논문에서는 마케팅을 위한 잠재고객, 즉 웹상의 홈쇼핑 사이트들에 대한 정보를 검색하기 위한 목적으로 온톨로지를 생성하였다. 이에 따라 토픽맵을 이용한 온톨로지는 세 가지의 토픽 타입 즉, 회사의 기본정보(company information), 지역(region), 품목(items)에 대하여 설계하고 이를 기반으로 구축하였다. 여기서 지역은 홈쇼핑 사이트를 운영하는 업체가 실제 위치하고 있는 지역을 의미하고, 품목은 홈 쇼핑 사이트에서 취급하고 있는 상품의 품목들을 의미한다. 그리고 회사 정보는 홈 쇼핑 사이트를 운영하는 업체의 기본적인 정보(전화번호, URL, e-mail 주소 등)들을 의미한다.

지역이라는 토픽타입에 대한 도메인은 홈쇼핑 사이트의 운영 업체가 위치한 지역을 계층 구조로 구분하여, 지역이라는 토픽을 루트로 하고 그 서브 노드에 각 지역 및 도시명을 계층적 구조의 토픽으로 배치한다. 그리고 품목이라는 토픽타입에 대해서는 전체 품목을 토픽의 루트로 하여 그 하부에 각각 개별 상품명을 계층적으로 배치한다. 이러한 지역 및 품목에 대한 토픽의 계층적 분류는 데이터 마이닝의 클러스터링, 연관 규칙 및 분류 규칙 등을 이용하여 관심 있는 속성에 대한 규칙을 유도하여 자동화 할 수 있다. 그리고 토픽타입 '회사정보'는 홈 쇼핑 사이트의 실제 주소, 전화번호, 취급 상품, URL 등의 기본 정보를 포함하므로 다른 토픽 타입 '지역' 및 '품목'과 연관 관계가 구성된다.

두 번째는 어커런스인데, 각 토픽은 자신이 참조하는 자원(resource)과 연결 될 수 있다. 예를 들어 홈 쇼핑 사이트의 회사 정보를 알아보기 위해서는 그러한 정보를 가지고 있는 URL인 http://dblab.chungbuk.ac.kr/~information을 참조 할 수 있다. 그리고 홈쇼핑 사이트가 위치한 실제 주소에 대한 내용은 http://dblab.chungbuk.ac.kr/~regions을 참조 하면 된다. 이러한 연결 정보를 어커런스라고 하고, 여러 형태가 있다. 문서파일, 이미지 파일, 비디오 파일, 데이터베이스 내 특정 레코드 등으로 나타낼 수 있다. 그리고 이러한 어커런스에 의해 참조되는 자원들은 토픽맵과 별도로 저장된다. 토픽맵 내의 토픽들은 자신의 자원을 가리키기 위해 HyTime이나 Xpointer와 같은 기법들을 사용한다.

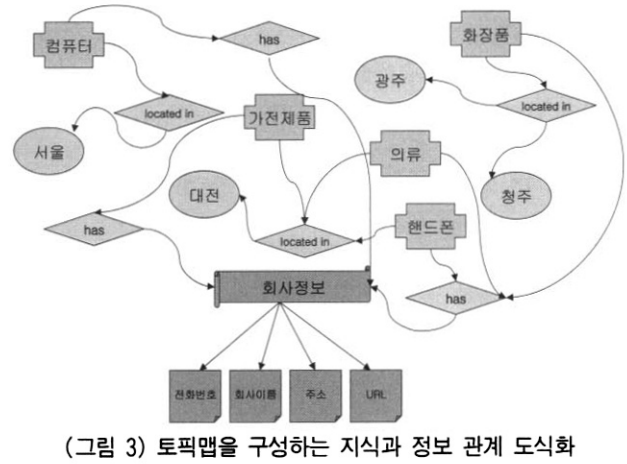
세 번째는 어소시에이션으로, 토픽맵 표준안 스펙에는 토픽들 간의 연관관계를 정의 할 수 있는 방법을 제공한다. 토픽 연관관계는 둘 이상의 토픽들 사이의 상하관계가 아닌 의미적인 관계를 정의하는 관계이다. 이것에 대한 표현 예는 아래와 같다.

Home Shopping Site of Computer *is located in* Seoul.
Home Shopping Site of Television *has* the information of the company.

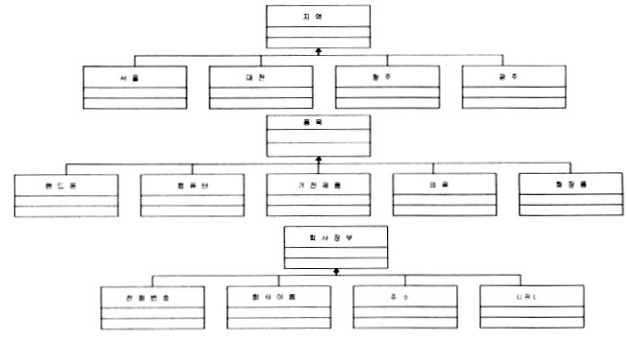
위의 예에서 computer와 Seoul은 토픽인데, 이 두 토픽사이에는 is located in 이라는 연관관계가 존재한다. 이와 같이 토픽맵의 토픽들은 서로 독립적인 객체이면서 동시에 특정 타입의 연관관계로 연결된 링크를 가진다. 토픽들이 토픽타입이라는 것으로 분류가 되는 것처럼 연관관계도 토픽의 연관관계 타입으로 분류된다. 즉, 위 예에서 'is located in', 'has', 'a member of' 등이 연관관계 타입이다. 토픽 타입과 연관관계 타입은 지식 및 정보 표현, 분류, 구조화를 위한 토픽맵의 중요한 기능이다. 그리고 Computer는 product_item이라는 roleSpec과, 회사의 전화번호와 주소, URL, 회사 이름 등의 회사의 정보인 information은 company_info

```
<!-- Association Type -->
    <topic id="has">
        <baseName>
            <baseNameString>AT: Has</baseNameString>
        </baseName>
    </topic>
    <!-- Role Spec -->
    <topic id="product_item">
        <baseName>
            <baseNameString>RoleSpec: product_item</baseNameString>
        </baseName>
    </topic>
    <topic id="company_info">
        <baseName>
            <baseNameString>RoleSpec: company_info</baseNameString>
        </baseName>
    </topic>
    <association id="has-information_tv">
        <instanceOf>
            <topicRef xlink:href="#has"/>
        </instanceOf>
        <member>
            <roleSpec>
                <topicRef xlink:href="#product_item"/>
            </roleSpec>
            <topicRef xlink:href="#tv"/>
        </member>
        <member>
            <roleSpec>
                <topicRef xlink:href="#company_info"/>
            </roleSpec>
            <topicRef xlink:href="#information_tv"/>
        </member>
    </association>
```

(그림 2) 어소시에이션 타입의 XTM



(그림 3) 토픽맵을 구성하는 지식과 정보 관계 도식화



(그림 4) 토픽에 대한 클래스 구조

라는 roleSpec을 가지고 있다. 아래 (그림 2)는 'has'라는 연관관계 타입을 표현하고 있는 association관계를 보여주는 예이다.

그리고 (그림 3)과 (그림 4)는 검색 시스템의 기반이 되고 있는 온톨로지 생성을 위한 토픽과 토픽간의 연관관계를 도식화한 것이다.

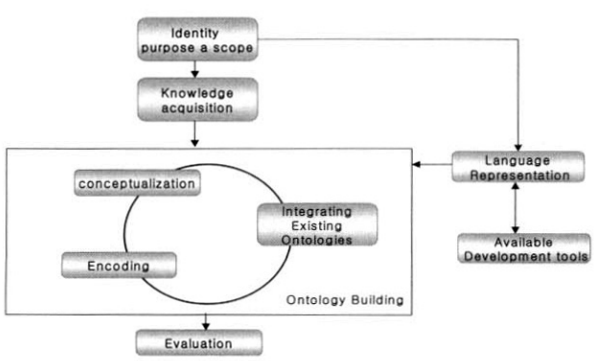
4. 온톨로지 기반의 검색 시스템 설계

이 장에서는 시맨틱 웹 기반의 고객 정보 탐색 시스템을 위한 온톨로지의 생성, 검색 에이전트 및 온톨로지를 이용하는 검색 시스템의 설계 내용을 기술한다.

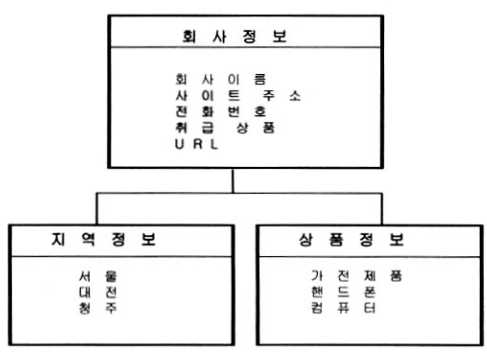
4.1 온톨로지 설계

온톨로지를 자동 생성하기 위해서는 먼저 관심 있는 도메인에 대한 온톨로지의 설계가 필요하다. (그림 5)는 온톨로지 구축을 위한 흐름 도를 보여주며, (그림 6)은 웹상의 쇼핑 사이트에 대한 기본 정보를 기준으로 온톨로지화 할 수 있는 지역 정보와 상품 정보를 정의하고 그에 대한 개념을 체계화한 클래스 구조를 보여준다.

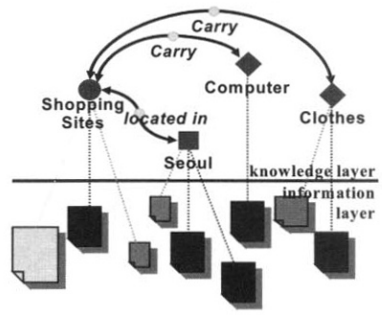
쇼핑 사이트의 회사 정보는 지역정보와 상품 정보로 나뉘게 되고 이를 토픽맵으로 구성하면 (그림 7)과 같은 형태의 지식 맵이 생성된다. 그리고 이 정보는 사용자가 특정 상품을 쇼핑할 때나 검색할 때 더 정확한 지식을 제공할 수 있는 기반이 된다.



(그림 5) 온톨로지 구축 흐름도



(그림 6) 쇼핑 사이트의 기본 정보 검색을 위한 온톨로지 도메인 설계



(그림 7) 쇼핑 사이트 지식맵

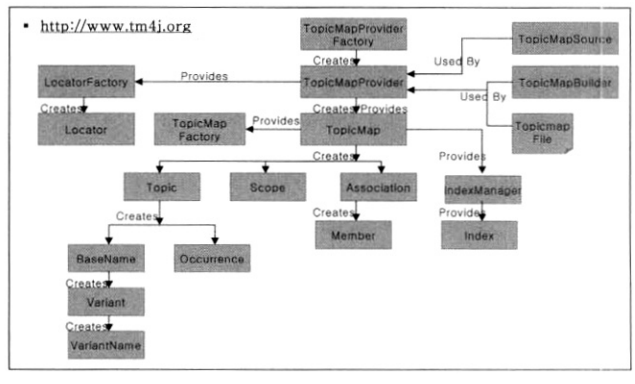
온톨로지를 생성하기 위한 도메인이 설정되면, 이에 따라 토픽맵에서 필요한 기본 요소인 토픽, 어소시에이션, 어커런스 등의 관계를 명시해야 한다. 이러한 토픽 관계를 명세하기 위해서 자동화된 온톨로지 생성 프로그램에 의해 토픽맵의 소스인 XTM으로 작성하였다. 작성된 XTM에 구문 오류는 온토피아(Ontopia)사의 옴니게이터(Omnigator)[18]를 이용하였다.

구문에 에러가 없는 XTM은 공개 소스 엔진인 TM4J[19]에 의해 파싱을 하는데, 파싱 과정에서는 온톨로지의 무결성을 보장하기 위해 온톨로지의 제약 조건을 사전에 검사한다. 즉, 어소시에이션의 카디날리티(cardinality), 상위 토픽이 가질 수 있는 하위 토픽의 종류, 어커런스 타입, 어소시에이션의 롤(role) 등의 제약 조건을 검증하기 위해 토픽맵 스키마를 참조한다.

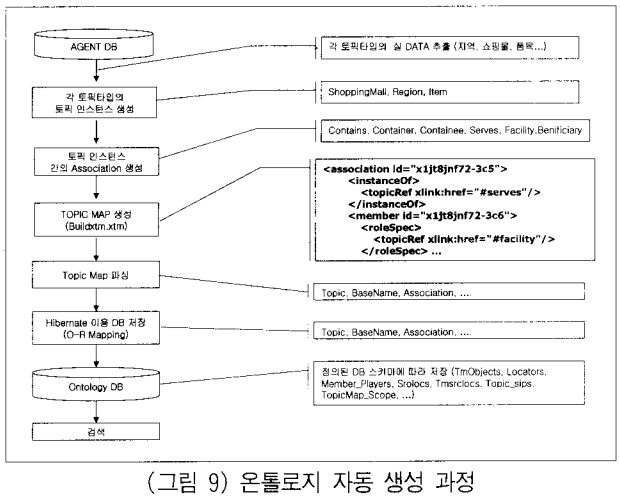
아래 (그림 8)은 TM4J를 구성하는 기본적인 아키텍처의 구조이고, 주요 클래스에 대한 역할을 요약하면 다음과 같다.

토픽맵 프로바이더 팩토리(TopicMapProviderFactory)는 토픽맵의 생성과 이를 저장하기 위해 여러 타입의 back-end를 위한 개별적인 구현을 가능하게 하는 추상 클래스이고, 토픽맵 프로바이더는 저장 구조에 연결시켜주는 역할을 한다.

토픽맵 소스는 토픽맵을 생성하는 과정과 데이터의 조합에 대한 추상화된 표현이다. 그리고 토픽맵은 그 상위 클래스인 토픽맵 프로바이더에 의해서 관리되는 저장 장치에 있는 하나의 토픽맵을 표현한다. 그리고 로케이터 팩토리(LocatorFactory)는 리소스의 주소를 표현하는 객체를 만드



(그림 8) TM4J의 계층적 클래스도

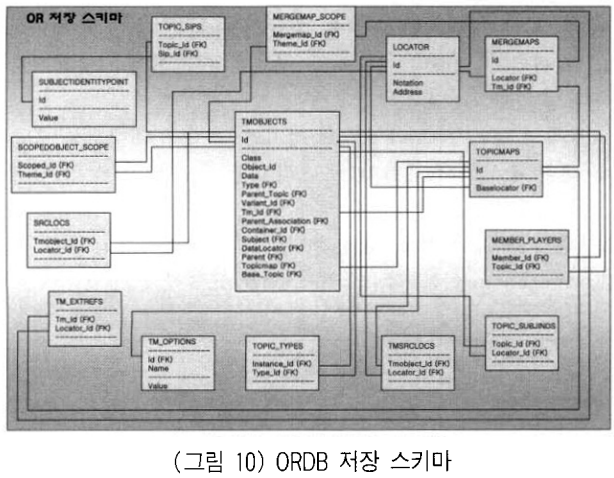


(그림 9) 온톨로지 자동 생성 과정

〈표 1〉 토픽맵 자동생성 예 : Build.xtm

```
<association id="x1jt8jnf72-3b1">
    <instanceOf>
      <topicRef xlink:href="#serves"/>
    </instanceOf>
    <member id="x1jt8jnf72-3b2">
      <roleSpec>
        <topicRef xlink:href="#facility"/>
      </roleSpec>
      <topicRef xlink:href="#아동도서-enjoy Big - KTmall"/>
    </member>
    <member id="x1jt8jnf72-3b3">
      <roleSpec>
        <topicRef xlink:href="#benificiary"/>
      </roleSpec>
      <topicRef xlink:href="#enjoy Big - KTmall"/>
    </member>
  </association>

...
```



(그림 10) ORDB 저장 스키마

는 매소드를 제공한다. 또한 토픽맵 프로바이더 인터페이스와 토픽맵 인터페이스의 getLocatorFactory() 매소드를 호출하여 검색이 가능하다.

이와 같은 과정을 통해서 생성되는 온톨로지의 생성은 개념과 개념간의 관계를 설계를 통해 규정하고 이를 프로그램화하였다. (그림 9)는 온톨로지의 TM4J 엔진을 통해서 생성되는 과정을 그림으로 나타낸 것이다.

이 논문에서는 에이전트를 통해 웹으로부터 발견한 데이터를 기반으로 데이터를 분류하고 이를 개념 관계에 의해 자동으로 온톨로지를 구성하며, 웹에서 발견한 지식의 리소스 정보를 토픽맵의 토픽, 어소시에이션, 어커런스에 정의하고 이를 통해 검색할 수 있도록 하였다.

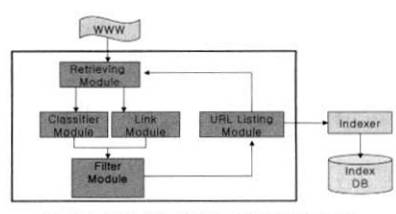
아래 <표 1>은 자동으로 생성된 온톨로지 소스의 일부인 build.xtm을 보여주며, (그림 10)은 토픽맵의 저장을 위한 객체 관계형의 저장스키마 구조를 설명하는 그림이다.

4.2 웹 검색 에이전트 설계

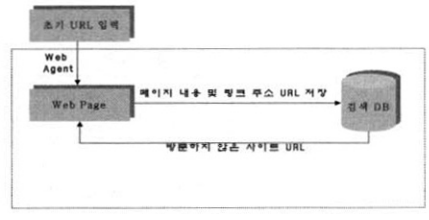
웹 검색 엔진은 웹에서 정보를 편리하게 찾을 수 있는 링크 정보를 저장하고 있는 데이터베이스이다. 에이전트에 의한 데이터베이스 구축 방법은 웹 서버를 돌아다니며 정보를 수집하고, 분류, 색인 화하여 데이터베이스에 저장시키는 작업을 수행하는 소프트웨어인 웹 로봇(Robot Agent)을 통해 이루어진다.

이 논문에서 웹 검색 에이전트는 스스로 웹 사이트의 링크를 따라 돌며, 웹 사이트의 정보를 데이터베이스에 저장하는 가상로봇으로써, 인수로 웹 사이트의 주소를 초기 치로 입력받아 검색을 시작한다. 그리고 쇼핑 사이트만을 필터링 하기 위해서 사이트 내에 '쇼핑'과 관련된 단어나 문서 등이 있는지 또는 취급 품목의 정보를 포함하는 단어나 문서가 있는지 등과 같이 일반적으로 쇼핑 사이트에서 볼 수 있는 문서들을 체크하여 쇼핑사이트라는 것을 판단하도록 하였다. 검색 된 쇼핑 사이트에서는 쇼핑 사이트의 자체 URL 및 링크된 사이트의 URL, Title, 취급 품목 및 쇼핑사이트를 운영하는 회사의 실제 주소 등의 정보를 추출하였다. (그림 11)은 웹 검색 에이전트의 구조에 대한 설계도이다.

웹 검색 에이전트의 구성 모듈은 웹을 검색하는 검색 모듈, 검색된 정보는 분리 모듈과 링크 정보를 처리하는 모듈로 나누어지고, 분리된 정보에서 검색 목적에 맞는 필요 정보만을 다시 걸러내는 필터모듈로 구성하였다. 그리고 (그림 12)는 에이전트에 의해 검색 과정에서 처리되는 데이터의 흐름을 보여준다.



(그림 11) 웹 검색 에이전트 구조



(그림 12) 에이전트 동작 흐름도

〈표 2〉 사이트 정보 테이블 스키마

테이블 명세서						
테이블 ID	page					
NO	컬럼 ID	컬럼명	Type	Length	NULL	Key
1	p_no	page 번호	int			P.K
2	site	site 주소	varchar2	256		
3	title	site 이름	varchar2	256		
4	visited	방문 기록	varchar2	1	N.N	

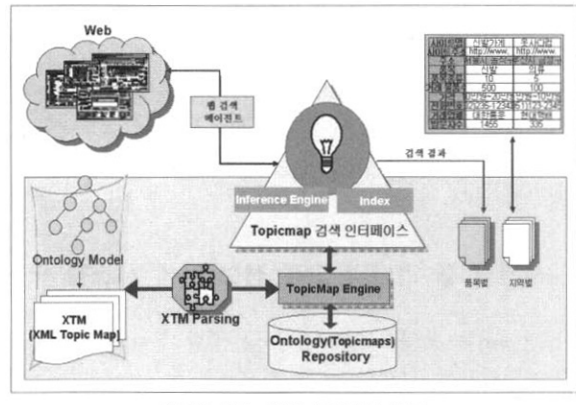
〈표 3〉 품목 정보 테이블 스키마

테이블 명세서						
테이블 ID	category					
NO	컬럼 ID	컬럼명	Type	Length	NULL	Key
1	c_no	category	int			P.K F.K
2	item	번호 취급 품목	varchar2	50		P.K

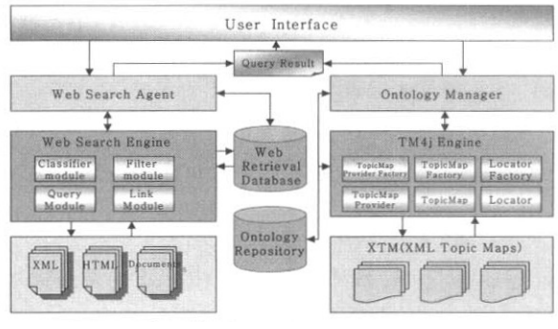
그리고 웹 검색 에이전트로부터 가져온 데이터 저장을 위한 데이터베이스 설계를 다음과 같이 하였다. 웹 검색 에이전트에 의한 검색 결과인 웹 사이트의 URL 주소와 Title 그리고 웹 쇼핑 사이트에서 취급하는 품목 및 실제 주소, 사이트에 링크되어 있는 주소는 데이터베이스에 저장하여, 4개 테이블 즉, 사이트의 URL를 저장하는 테이블 page, 각 사이트에서 취급하는 품목을 저장하는 품목 테이블 category, 실제 주소가 이미지로 구성되어 있는 이미지 정보 테이블 image_address, 실제 주소 정보를 포함하는 텍스트 정보를 포함하는 텍스트 테이블 text_address의 스키마를 생성하였으며, 다음 <표 2>와 <표 3>은 그 일부분을 보여준다.

4.3 온톨로지 기반의 검색 시스템 설계

쇼핑 사이트의 정보 추출을 위한 온톨로지 기반의 정보 검색 시스템의 설계에 대해서 기술한다. (그림 13)은 온톨로지 기반의 정보 검색 시스템의 전체 프레임워크를 보여준다. 제안하는 검색 시스템은 웹으로부터 웹 검색 에이전트를 이용하여 정보를 추출하고 추출된 정보는 온톨로지를 기반으로 하여 더 정확한 검색 결과를 도출해 내는 것을 목적으로 한다.



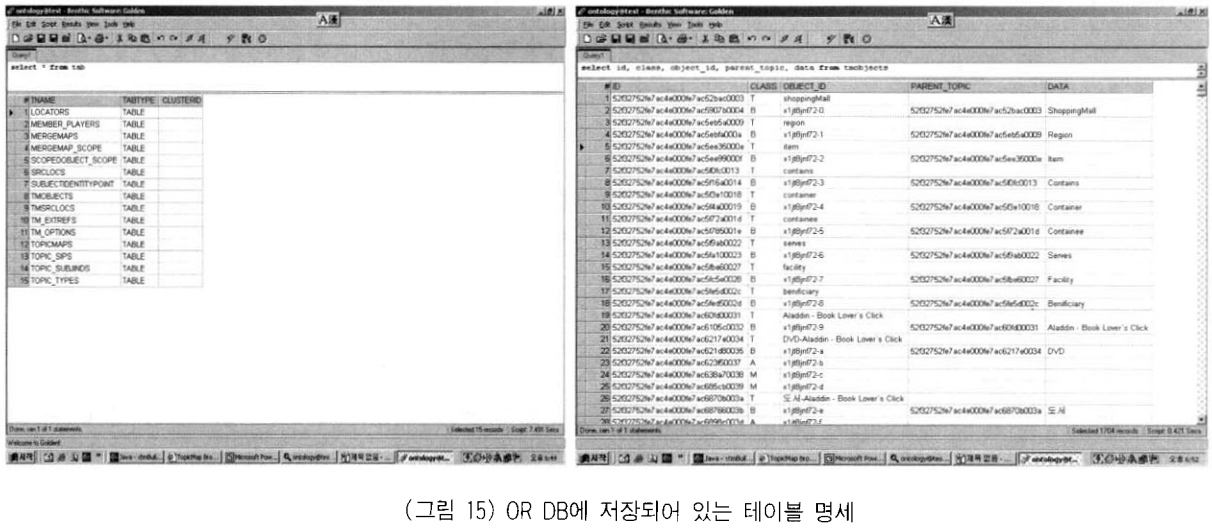
(그림 13) 전체 프레임 워크



(그림 14) 상세 모듈 구성도

이에 대한 세부 모듈의 구성은 (그림 14)에서 상세하게 보여주고 있으며, 웹 검색 에이전트 모듈과 온톨로지 생성 및 관리를 위한 모듈로 나눌 수 있다. 온톨로지 기반의 검색 시스템은 사용자 인터페이스, 온톨로지 저장소, 온톨로지 생성 엔진 및 웹 서버 엔진으로 구성되며, 이 논문에서는 온톨로지 구축을 위해 TM4J의 공개 소스 엔진을 이용하여 토픽맵을 생성 및 관리하고, 온톨로지를 자동 생성하는데 이용하였다.

이 논문에서 제안한 검색 시스템에서는 사용자가 사용자



(그림 15) OR DB에 저장되어 있는 테이블 명세

인터페이스를 통해 탐색하고자 하는 고객 정보를 지역과 상품으로 나누어 질의할 수 있다. 그리고 질의 내용은 온톨로지 엔진은 이미 웹에서 에이전트를 통해 검출한 기본 정보를 검색 데이터베이스에 저장하고 있으므로, 온톨로지 저장소의 정보를 이용하여 사용자가 원하는 검색 결과의 데이터를 제공해 준다.

5. 시맨틱 웹 기반의 고객 정보 검색 시스템 구현

이 장에서는 온톨로지의 생성 및 구현 결과를 보이고, 웹 검색 에이전트의 구현을 통해 웹으로부터 검색하여 저장한 결과 내용에 대해 설명한다. 또한 에이전트의 검색 결과에 대해 온톨로지 정보를 이용한 검색 결과의 예를 보인다.

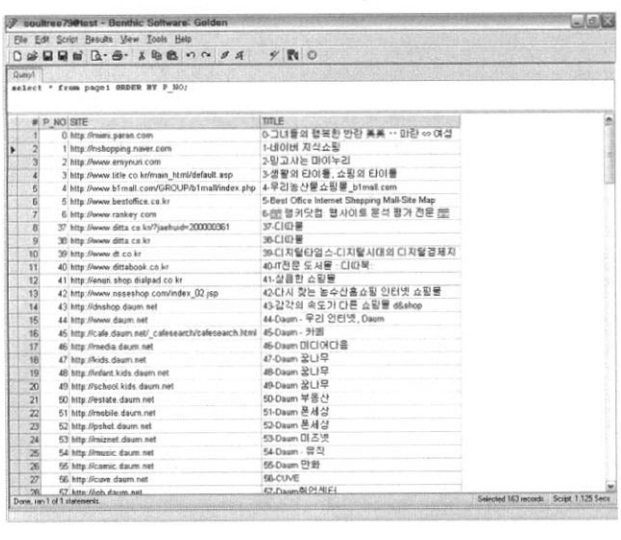
4장에서 기술한 온톨로지의 자동 생성 설계에 따라 자동으로 생성된 토픽맵의 정보는 TM4J의 하이버네이트(hibernate)[20]의 저장 방식에 따라 객체 관계형 데이터 베이스에 자동으로 테이블을 생성하고 각 테이블의 정보를 저장하게 된다. (그림 15)는 자동으로 생성된 테이블의 정보를 보여준다.

5.1 검색 에이전트 구현

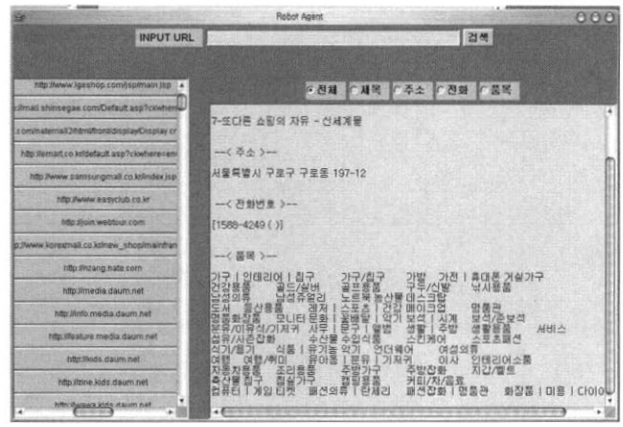
웹 쇼핑사이트를 검색하기 위한 URL의 초기화는 포탈 쇼핑몰 14개, 전문 쇼핑몰 14개로 하여 웹 검색 에이전트에

〈표 4〉 전문 쇼핑몰 초기 URL

NO	URL	싸이트명	주소	비고
1	http://www.caraudioplaza.com/	//카오디오플라자	(I)	
2	http://www.honamfishing.co.kr/	//호남낚시	(T)	
3	http://www.computer100.co.kr/	//컴퓨터백화점	(I)	
4	http://www.notemaul.co.kr/	//노트마을	(T)	
5	http://www.thepharos.co.kr/	//골프천구	(I)	
6	http://www.golfabc.co.kr/	//골프 ABC	(T)	
7	http://www.gajunmart.co.kr/	//가전마트	(T)	
8	http://www.nextday.co.kr/	//넥스트데이 화장품	(T)	
9	http://www.bestoffice.co.kr/	//베스트오피스	(T)	
10	http://www.ilovehandphone.com/html	//아이러브핸드폰	(I)	
11	http://www.sammart.co.kr/	//삼마트	(T)	
12	http://www.ljhmall.co.kr	//이지할	(I)	
13	http://www.twoprice.com/	//투프라이스	(T)	
14	http://www.wellports.com/	//웰포츠	(T)	



(그림 16) TITLE 추출 저장 결과



(그림 17) 에이전트의 검색 결과의 예(search.java)

의한 검색을 수행하였다. 표 4는 검색 결과를 저장한 내용을 보여주고 있다.

다음 (그림 16)은 PAGE 테이블에 저장되어 있는 쇼핑 사이트에 대한 TITLE의 정보를 추출하여 저장한 결과이며, (그림 17)은 에이전트를 수행하는 search.java 실행 화면이다.

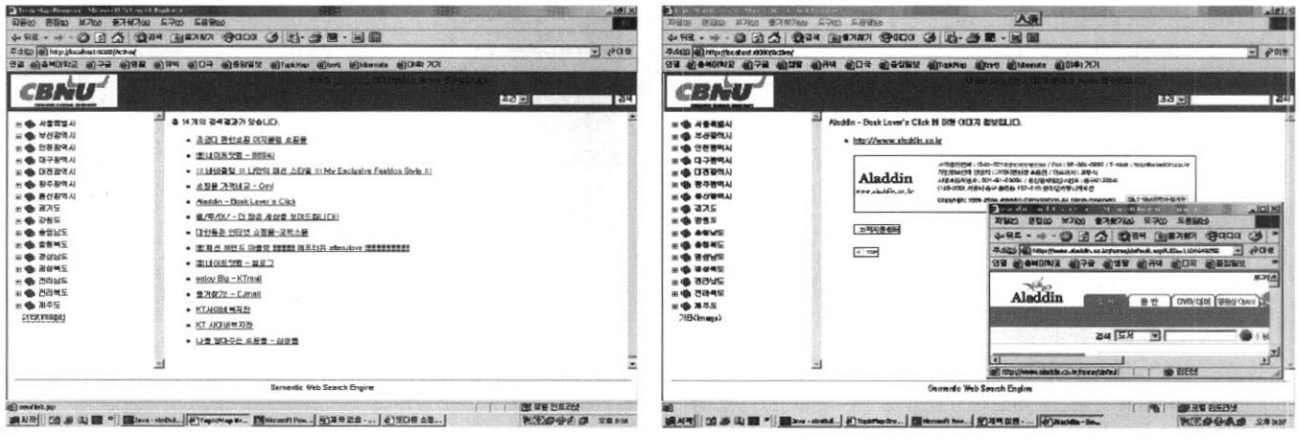
(그림 17)의 오른쪽에 보이는 것이 검색한 결과의 쇼핑 사이트의 URL 정보이고, 왼쪽의 화면은 검색한 결과의 특정사이트에 대한 전화번호, 쇼핑 사이트의 제목 및 취급 품목을 보고자 할 때 오른쪽 화면의 특정사이트를 클릭한 상태에서 해당 항목을 선택하면 기본 정보의 전체 정보 및 특정 정보를 볼 수 있다. 그리고 그림의 상단에 있는 INPUT URL은 검색하고자 하는 사이트의 초기화를 설정할 수 있는 것으로, 초기화하고자 하는 사이트를 직접 입력하면 에이전트는 그에 대한 초기 사이트를 이용하여 사이트에 링크가 있는 여러 사이트의 정보를 탐색하게 된다.

에이전트의 수행에 의한 사이트의 정보 수집에서 취급 항목과 주소 부분이 이미지로 처리되어 있는 경우에 대해 취급 항목의 경우 하위 디렉터리에서 취급 품목을 추출하였고, 주소 이미지 부분은 제공되는 소스에서 어떤 이미지가 주소를 포함하는지 불분명하여 대략적으로 하위 이미지 중 세 개의 이미지에 주소를 포함하는 경우가 많으므로 이를 모두 데이터베이스에 저장하였다.

5.2 온톨로지 생성 및 온톨로지를 이용한 정보 검색 수행

이 논문에서는 택배 마케팅을 위한 잠재 고객의 정보를 검색하기 위해 세 가지의 온톨로지 즉, 쇼핑 사이트의 기본 정보(company information), 지역(region), 품목(items)을 설계하였으며, 이를 이용하여 세 가지의 온톨로지 즉, 회사정보 온톨로지, 지역 온톨로지, 상품 품목 온톨로지를 TM4J 엔진을 이용하여 자동으로 생성하였다. (그림 18)은 이와 같은 온톨로지 중에서 지역 온톨로지를 이용한 검색한 결과이다.

위 (그림 18)은 지역 온톨로지를 이용한 정보 검색 시스템을 웹 화면으로 보여준 그림이며, 지역에 대한 계층구조를 생성하여 온톨로지 구조를 만들고, 웹 페이지의 화면의 왼쪽 편에 메뉴로 사용하였으며, 이와 같은 메뉴를 이용하여 계층구조를 가지고 따라서 클릭하면서 구체적으로 사용



(그림 18) 지역 온톨로지 기반의 검색 결과

자가 원하는 지역을 선택할 수 있다. 이렇게 선택된 지역에 있는 홈쇼핑 사이트가 화면의 오른쪽에 디스플레이가 되면 사용자는 자신이 원하는 정보를 가지고 있는 사이트를 선택하는 구조를 가지고 있다.

그리고 상품 온톨로지는 종합 쇼핑몰에서 일반적으로 사용하는 품목 명을 기반으로 계층적인 품목 명으로 구성된 온톨로지 구조를 가지고 있으며, 지역 온톨로지 에서처럼 왼쪽 메뉴의 상품의 계층구조를 따라 가면서 사용자가 원하는 해당 품목을 클릭하면 그 상품을 취급하는 쇼핑 사이트의 정보를 오른쪽에 디스플레이 해주는 결과를 볼 수 있다.

이와 같은 온톨로지를 이용한 정보 검색 시스템은 기존의 구글(Google)과 야후(Yahoo)와 같은 검색 시스템이 갖고 있는 키워드 검색 방식의 문제점으로 지적되고 있는 사용자가 원하는 정보와 관계없는 다양한 정보를 제시해 준다는 단점을 극복할 수 있는 특징이 있다. 즉 개념간의 관계를 정의하고 그 개념들을 이용하여 구축된 온톨로지를 이용하므로, 사용자가 원하는 정보와 무관한 정보를 걸러 내주고, 정확하게 사용자가 원하는 정보를 제시 해 줄 수 있다. 그리고 기존 연구[22, 24]의 온톨로지를 이용한 검색 시스템과의 차이점은 기존 연구에서는 기존에 확립되어 있는 데이터를 이용한 검색 시스템이었는데 이 논문에서 구현한 정보 검색 시스템은 웹 에이전트를 이용하여 웹에 있는 데이터를 실시간으로 받아들여서 이용하는 검색 시스템이라는 점이다. 그러므로 이 논문에서 구현한 온톨로지를 이용한 검색 시스템은 기존의 웹 검색 시스템과는 달리 시맨틱적인 요소를 추가하였으며, 기존의 온톨로지를 이용한 검색 시스템과는 다르게 실시간으로 웹 데이터의 처리에 의한 온톨로지 생성이 가능하다는 특징이 있다.

6. 결 론

이 논문에서는 시맨틱 웹 기반의 택배 마케팅에 필요한 잠재 고객의 정보를 검색하기 위해서구체적인 지식 도메인을 대상으로 하는 온톨로지를 자동 생성하였다. 그리고 생성된 온톨로지를 기반으로 웹으로부터 사용자가 필요로 하는 정보를 검색하기 위해서 웹 로봇을 구현하였다. 즉, 잠재 고객의 검색을 위한 지식 도메인의 정의 및 지식을 정형화하기 위한 온톨로지와 웹으로부터 고객의 정보를 추출하기 위한 에이전트를 설계 및 구현하였다. 아울러 택배 마케팅에서 온톨로지를 생성하는 것과 온톨로지를 이용하여 정보 검색의 수행이 효과적으로 실행됨을 보여주었다.

제안한 온톨로지의 생성 및 이를 이용하는 정보 검색 시스템은 잠재 고객에 대한 신뢰성 있는 정보를 효율적으로 검색할 수 있는 기반이 되고, 온톨로지의 생성은 정보 지식의 공유와 관련된 도메인에서의 일관성을 고려하므로 향후 웹에서 고객의 효율적인 쇼핑을 위한 상품 추천, 마케팅 전략 및 택배 서비스의 질적 향상을 도모할 수 있는 기반이 된다. 아울러 이 논문에서 제안하는 온톨로지 생성과정을 향상시키고 기존 연구와의 실험 결과에 대한 체계적인 비교 및 분석을 위해 적극적인 연구가 필요하다.

참 고 문 헌

[1] T.Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web," Scienctific American, 2001.
[2] J. Hendler, "Agents and the Semantic Web," IEEE Intelligent Systems, Vol.16, No.2, March/April, 2001.
[3] S. Decker, S. Melink, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, I. Horrocks, "The Semantic Web: the roles of XML and RDF," IEEE Internet Computing, Vol. 4, No.5, pp.63-73, 2000.
[4] M. Klein, "XML, RDF, and Relatives," IEEE Intelligent System, Vol.16, No.2. March/April, 2001.
[5] A.Maedche, S.Staab, "Semi-Automatic Engineering of Ontologies from Text", Institute AIFB, University of Karlsruhe, Germany, 2000.
[6] A. Gomez-Perez, O. Corcho, "Ontology Languages for the Smantic Web," IEEE Intelligent Systems, Vol.17, No.1, 2002.
[7] W3C. RDF website. http://www.w3c.org/RDF
[8] W3C. RDFS website. http://www.w3c.org/TR/rdf-schema
[9] On-To-Knowledge project consortium, On-To-Knowledge project website,http://www.ontoknowledge.org
[10] D. McGuiness, R. Files, J. Hendler, L. Stein, "DAML+OIL:an

ontology language for the semantic web," IEEE Intelligent Systems, Vol.16, No.2, pp.72-80, 2002.
[11] Ian horrocks, "DAML+OIL:A Description Logic for the Semantic Web," IEEE Bulletin of the Technical Committee on Data Engineering Vol.25, No.1, 2002.
[12] Mike Dean et at. (Eds.), "Web Ontology Language (OWL) Reference Version 1.0," W3C Working Draft 12 November, 2002.
[13] K. Michael, et al., "Web Ontology Language(0WL) Guide Version 1.0, "W3C Working Draft 4 November, 2002.
[14] International Organization for Standardization, ISO/IEC 13250, Information Technology SGML Applications-Topic Map (ISO,Geneva 2000).
[15] M. Biesunski, M. Bryan, S. Newcomb, ISO/IEC 13250 TopicMaps.
[16] S. Pepper, G. Moore, "XML Topic Maps(XTM) 1.0", TopicMpas.org.
[17] Steve Pepper, "The TAO of Topic Maps," XML Conference & Exposition, 2000.
[18] http://www.ontopia.net
[19] http://www.tm4j.org
[20] http://www.hibernate.org
[21] A. Maedche, S. Staab, "Semi-Automatic Engineering of Ontologies from Text," Institute AIFB, University of Karlsruhe, Germany, 2000.
[22] Koung-lung Lin, Yen-jen Oyang, "Knowledge Management for a Buddhism Digital Archive with Topic Map," ICDAT 2002.
[23] 정호영, 김정민, 정준원, 김형주, "XTM 기반의 지식맵", 데이터베이스연구회지, 2003.
[24] 김정민, 박철만, 정준원, 이한준, 민경섭, 김형주, "K-Box : 토픽맵 기반의 온톨로지 관리 시스템", 한국 정보 과학회 논문지 C, Vol.10, No.1, 2004. 2.
[25] 최호섭, 옥철영, "정보 검색 시스템과 온톨로지", 한국 정보 과학회지 제22권 제4호, 2004. 4.

황 정 희

e-mail : jhhwang@dblab.chungbuk.ac.kr
1991년 충북대학교 전산통계학과(이학사)
2001년 충북대학교 전자계산학과(이학석사)
2005년 충북대학교 전자계산학과(이학박사)
2001년 8월~2006년 2월 정우시스템(주) 연구소장
현 재 남서울대학교 컴퓨터학과 전임강사
관심분야 : XML, 데이터 마이닝, 능동 데이터베이스, 유비쿼터스 컴퓨팅, 시공간 데이터베이스

구 미 숙

e-mail : gumisug@dblab.chungbuk.ac.kr
1986년 충남대학교 영어영문학과(문학사)
2004년 충북대학교 전자계산학과(이학석사)
2006년~현재 충북대학교 전자계산학과 박사과정
관심분야 : XML, 시공간 데이터베이스, 유비쿼터스 컴퓨팅, 바이오 인포매틱스, 데이터 마이닝

이 현 아

e-mail : halee@etri.re.kr
1999년 충북대학교 전기전자공학부(공학사)
2002년 충북대학교 컴퓨터과학과(이학석사)
2006년~현재 한국전자 통신연구원 우정기술센터 연구원
관심분야 : 시공간데이터베이스, 객체지향데이터베이스, GIS, 시맨택웹, RFID

류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr
1976년 숭실대학교 전산학과(이학사)
1980년 연세대학교 전산전공(공학석사)
1988년 연세대학교 전산전공(공학박사)
1976년~1986년 육군군수 지원사 전산실(ROTC 장교), 한국전자통신 연구원(연구원), 한국방송통신대 전산학과(조교수) 근무
1989년~1991년 Univ. of Arizona Research Staff(TempIS 연구원, Temporal DB)
1986년~현재 충북대학교 전기전자 컴퓨터공학부 교수
관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 지식기반 정보검색 시스템, 유비쿼터스컴퓨팅 및 스트림데이터처리, 데이터 마이닝, 데이터베이스 보안, 바이오 인포메틱스