

온톨로지를 적용한 상품정보 시스템 설계

박동훈^o 정성원, 박대원, 권혁철
부산대학교 컴퓨터공학과

dougu^o@hanmail.net, swjung@pusan.ac.kr, bluepepe@pusan.ac.kr, hckwon@pusan.ac.kr

Design for Product Information System Using Ontology

Donghun Park^o, Sungwon Jung, Daewon Park, Hyukchul Kwon
Dept of Computer Science and Engineering, Pusan National University

요 약

인터넷 쇼핑물에서는 다양한 제품이 존재하며 제조회사마다 제품 정보를 표현하는 방식이 다르며, 같은 제품이라도 쇼핑물마다 제품정보를 표현하는 방식이 다르다. 이런 환경에서 효율적인 제품정보의 수집과 제품정보의 검색이 필요하다. 쇼핑물의 입장에서 다양한 회사의 다양한 제품 정보 수집 시 정보 수집의 정확성과 효율을 높이고 고객에게는 좀 더 유연한 검색결과와 제공을 위한 상품 정보 시스템을 설계한다. 이를 위해서 첫째로, 제품정보에 대한 온톨로지를 구축하고, 온톨로지를 바탕으로 제품 정보를 추출하고, 데이터베이스화한다. 둘째로, 제품 정보에 대한 온톨로지를 이용하여 추론 기능을 이용한 검색 서비스를 적용한다. 본 논문은 MP3 제품에 대해 제품정보 수집과 검색을 위한 온톨로지 구축과 온톨로지를 이용한 정보추출, 추론 기술에 대해서 기술한다.

1. 서 론

인터넷에서는 수많은 쇼핑물이 존재하며 그 속에는 다양한 제품이 존재한다. 그리고 제품정보를 제공하는 곳에 따라 제품정보를 표현하는 방법도 다르다. 이런 환경에서 사용자가 원하는 제품을 찾는 데 불편함이 있다. 기존의 검색서비스는 키워드를 이용한 검색 서비스를 제공하고 있으며 사용자가 원하는 정보를 찾기 위해서는 검색창에서 키워드를 입력하여 검색하거나 검색할 제품의 카테고리를 찾아 들어가서 제공해주는 검색 키워드를 선택하여 원하는 정보를 찾게 된다. 이보다는 제품 정보의 의미처리에 기반을 둔 검색이 가능하면 사용자에게 좀 더 적합한 제품 검색 결과를 제공할 수 있을 것이다.

제품 검색 시 의미 기반의 검색을 지원하기 위해서 제품 정보 수집시 서로 다른 표현으로 된 제품정보를 다룰 수 있는 의미처리에 기반으로 하는 제품 정보 수집방법 연구와 제품정보검색에서는 키워드 검색뿐 아니라 사용자의 질의에 대한 적합한 검색결과를 제공할 수 있는 방법에 대한 연구가 필요하다.

서로 다른 방식으로 표현된 제품정보를 수집하기 위해서 다르게 표현된 제품정보의 의미를 이해할 수 있어야 한다. 이런 정보들 사이의 의미적인 처리를 가능하게 해주는 것이 온톨로지이다. 온톨로지는 정보를 표현하기 위한 개념을 정의하며, 정보의 공유를 가능하게 해준다.

온톨로지를 이용하여 서로 다르게 표현된 제품정보를 좀 더 효과적으로 수집을 할 수 있다. 또한, 온톨로지 기반의 추론이 가능하여 제품 정보 검색시 사용자의 질의에 적합한 결과를 얻을 수 있다.[1]

본 논문에서는 제품 정보의 수집과 검색을 위하여 시멘틱 웹 기술 기반의 온톨로지를 이용하여 도메인 온톨로지와 제품 온톨로지를 구축에 대하여 소개하고 제품 정보 추출과정에서 제품 온톨로지를 이용하여 제품 정보 테이블을 이용한 반자동 제품정보 수집방법과 사용자 질의에 대하여 보다, 정확한 결과를 얻기 위하여 온톨로지 기반의 추론을 통한 제품 검색 서비스를 제안한다.

2. 관련 연구

2.1 온톨로지 구축 방법

이론과 경험에 의해 온톨로지 구축방법으로 여러 가지 방법들이 제안되었으며 온톨로지를 구축하는 기본적인 과정은 목적확인, 개념화, 기호화, 기존 온톨로지 통합, 평가, 문서화와 같은 과정을 거쳐 온톨로지를 구축하게 된다. 본 논문에서는 도메인 온톨로지를 구축하고 구축한 도메인 온톨로지를 이용하여 새로운 온톨로지를 생성한다. 따라서 도메인 온톨로지를 구축하기 위한 대표적인 방법으로는 OTKM과 ONIONS 방법론이 있다.

OTKM(On-To-Knowledge Methodology for development and evolution of Ontology based Tool for knowledge Management)[2]가 있다. OTKM은 애플리케이션 지향의 온톨로지 개발에 중점 두는 방법론으로 타당성 연구(Feasibility study), 개시(Kickoff), 정제

*이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(지방연구중심대학육성사업/차세대융합기술연구사업단)

(Refinement), 평가(Evaluation), 유지보수(Maintenance) 5단계로 구성되어 있다. 타당성 연구는 온톨로지의 적용 범위를 정의하며, 개시를 통해 요구명세서와 초기 온톨로지를 구축한다. 정제는 초기 온톨로지를 검증하고 정제하여 공식온톨로지를 생성하고, 평가는 생성된 공식온톨로지를 분석하는 단계이다. 마지막으로 구축된 온톨로지를 유지보수하게 된다. 온톨로지 구축과정에서 정제, 평가, 유지보수 단계는 필요 시 반복적으로 수행하게 된다.

그리고 온톨로지 통합을 위한 방법론으로 ONIONS(Ontological Integration Of Navie Source)[2] 방법론은 다른 개념상의 온톨로지들의 통합 문제를 설명하기 위해 개발되었다. 주로 의학에서 도메인 온톨로지의 개념 통합을 지원하는 통합 온톨로지의 개발과 도서관에서 정보검색과 디지털 콘텐츠의 통합 등을 지원하기 위해 사용되었다. ONIONS 방법론은 도메인에서 유효한 전문용어 말뭉치(corpus)를 수집하고, 용어가 가진 개념을 구체화한다. 그리고 각 항목을 명세화한다. 명세화를 일반적인 온톨로지 라이브러리로 구성하여 도메인 온톨로지를 구현한다. 이 방법론은 전문 용어 수집을 통하여 온톨로지 구축이 가능하다.

2.2 정보 추출 기법

제품정보 검색 시 검색 결과의 정확성을 높이려면 제품정보 제공자가 제품 정보에 대해 상세하게 입력을 해줘야 한다. 하지만, 다양하고 많은 제품에 대해서 상세한 제품정보를 입력하기란 쉽지가 않다. 그래서 최근에는 제품정보 생성을 반자동으로 지원하는 연구가 확대되고 있으며 대표적인 반자동 정보 추출방법 중에는 온톨로지를 이용하여 정보를 추출하는 방법과 웹 문서의 의미 있는 테이블에서 정보를 추출하는 방법이 있다.

다양한 문서에서 정확한 정보 추출을 위해서 규칙을 사용하여 정보를 추출하게 된다. 이때 정보를 추출하는 규칙은 수동으로 생성하거나, 자동으로 생성하는 방법을 사용한다. 이런 규칙을 자동으로 생성하기 위하여 온톨로지를 이용한 정보추출 방법을 사용한다. 먼저, 정보 추출에 필요한 규칙의 자동화를 위해 정보추출에 필요한 도메인 지식을 이용한다. 이 도메인지식을 도메인 온톨로지 구성하게 된다. 도메인 지식에서 정보 추출을 위한 규칙을 도메인 온톨로지를 바탕으로 생성을 하게 된다. 이렇게 생성된 규칙을 이용하여 도메인과 관련된 문서에서 정보를 추출하게 된다. [3]

인터넷상에는 많은 양의 웹페이지가 존재하며 많은 웹페이지가 테이블형식의 문서 구조를 가진다. 그리고 웹페이지 속의 테이블은 정보를 가지고 있는 테이블과 웹페이지를 꾸미는 테이블로 구성되어 있으며 정보를 가지고 있는 테이블은 의미 있는 정보를 가지고 있으며, 구조화되어 정보를 추출하기 유용한 형태를 가진다. 웹 문서의 의미 있는 테이블에서 정보를 추출하는 방법[4]은 이 점에 착안해 웹페이지 속에 테이블 중 의미 있는 정보를 가지고 있는 테이블 구조를 학습하여 테이블 내에서 정보를 가진 테이블과 웹페이지를 꾸미는 테이블을 구분하고 정보를 가지고 있는 테이블의 정보를 추출을

한다.

3. 제품 정보 검색 시스템 구조

제안하는 제품정보 검색 시스템은 IT Device 관련 제품정보 검색을 위하여 반자동 제품정보 수집과 제품정보 검색을 지원하는 시스템을 만들며 실제 프로토타입으로 MP3 제품에 대한 제품정보 수집, 검색 시스템이다.

MP3 제품정보 검색을 위하여 온톨로지를 이용하여 상품 정보 테이블에서 정보를 수집하고 온톨로지와 온톨로지의 추론 기법을 바탕으로 제품정보 검색 서비스를 제공하려고 한다. 도메인 온톨로지를 이용하여 MP3 온톨로지를 구축하고 구축한 MP3 온톨로지는 테이블에서 제품 정보를 추출하고 제품 검색에 사용한다.

MP3 제품정보를 수집하고 검색하는 시스템을 구성하기 위하여 온톨로지 정의는 OWL 온톨로지[5][6]를 이용하였고, 온톨로지 데이터를 처리를 위하여 온톨로지 관련 API인 KAON2를 이용한다.[7] OWL-DL, SWRL을 이용한 추론기능을 제공하도록 한다. [그림1]은 MP3 제품 정보 검색 시스템의 구조이다.

시스템 구성은 크게 온톨로지, 정보추출 기능, 정보검색 기능 세 가지로 나눌 수 있으며, Domain 온톨로지, MP3 온톨로지는 제품정보 추출과 검색을 위한 온톨로지이며 Table Information Extraction Module, Ontology Management Module, MP3 Ontology Repository는 제품정보 추출과 추출한 제품정보를 온톨로지의 인스턴스로 저장하는 역할을 한다. Ontology Reasoning Module, Query Management Module은 제품정보 검색을 위한 추론과 질의 처리를 역할을 한다.

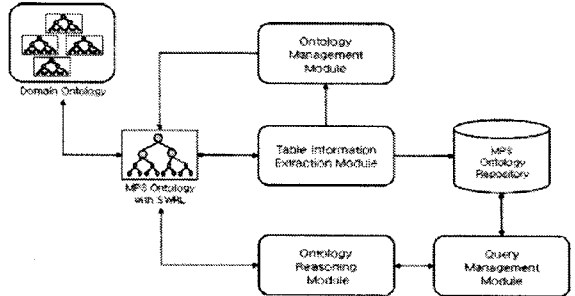


그림 1 MP3 상품 정보 시스템 구조도

3.1 IT Device에 대한 온톨로지 구축

IT Device 대한 온톨로지는 각각의 Device에 대한 온톨로지를 정의한 온톨로지이며 MP3, LCD TV, 디지털 카메라와 같은 Device 온톨로지와 MP3나 디지털 카메라에 들어갈 메모리, 배터리와 같은 Device의 구성요소에 대한 온톨로지를 정의한다. 그리고 Device의 다른 정보를 표현하기 위한 온톨로지를 구축한다. IT Device 대한 온톨로지를 이용하여 MP3 제품뿐 아니라 다른 제품 온톨로지를 구축할 때 재사용을 할 수 있다.

IT Device에 관하여 크게는 Audio Device, Video Device, Power Device, Storage Device,

Communication Device, Camera에 관한 온톨로지를 정의하였다. 그리고 IT Device에 관한 사양이나, 다른 정보를 표현하기 위해 Units, Manufacturer, File Format, Music 온톨로지를 정의하였다. 또한, Music, Units, Camera 온톨로지는 기존에 구축되어 있는 온톨로지를 활용한다.

표 1 정의한 Device 온톨로지

온톨로지	클래스 수(개)	프로퍼티 수(개)
Audio Device	59	26
	MP3, CDP 등과 같은 오디오 기기에 관하여 정의한 온톨로지	
Video Device	50	20
	TV, 모니터 등과 같은 디스플레이 기기를 정의한 온톨로지	
Storage Device	41	11
	메모리, 하드디스크와 같은 저장장치와 속성을 정의한 온톨로지	
Power Device	15	10
	배터리와 같은 전원 장치와 속성을 정의한 온톨로지	
File Format	6	7
	파일 포맷에 관하여 정의한 온톨로지	

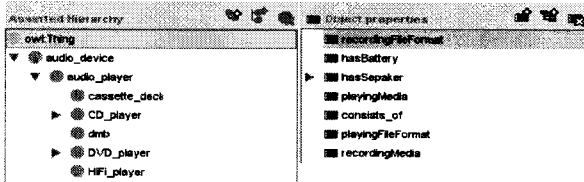


그림 2 Device 온톨로지의 예

3.2 MP3 온톨로지 구축과 인스턴스 생성

MP3 제품 정보의 검색과 추출을 위한 온톨로지를 구축하기 위해서 3.1에서 구축한 도메인 온톨로지를 기반으로 MP3 온톨로지를 구축한다. 그리고 MP3 제품정보를 추출하여 MP3 온톨로지의 인스턴스 정보를 추가하기 위하여 MP3 제조사와 인터넷 쇼핑몰의 제품 사양 테이블을 참고하였다.

3.2.1 MP3 온톨로지 구축

MP3 온톨로지를 구축하는 작업은 제품정보 테이블에서 제품 정보 추출시 기준이 되는 온톨로지이다. MP3 온톨로지를 구축하기 위해서 수동으로 구축한 여러 Device에 관한 온톨로지를 사용하였다. MP3 온톨로지의 클래스와 관계를 3.1에서 구축한 Device 온톨로지의 사용하여 MP3 온톨로지를 수동으로 구축하였다.

MP3 구축시 여러 Device 온톨로지서 MP3의 제품 사양, 기능, 구성하는 부품 등의 정보를 표현할 수 있는

클래스와 관계를 뽑아낸다. 그리고 뽑아낸 클래스와 관계를 이용하여 MP3 온톨로지를 정의한다.

[그림2]에서 MP3의 디스플레이를 표현하는 정보이며, [그림3]에서 LCD 사양을 나타내는 정보이다. MP3 온톨로지 구축시 필요한 온톨로지를 LCD 온톨로지를 통하여 정의를 할 수 있다. MP3의 디스플레이의 2.4", 262K Color, 320 x 240 은 LCD Panel의 Size, Resolution, Display Colors를 이용하여 정의할 수 있다. MP3의 FM 라디오에 대해서도 Radio에 대해 정의한 온톨로지를 이용하여 정의할 수 있다. [그림5]은 구축한 MP3 온톨로지의 예이다.

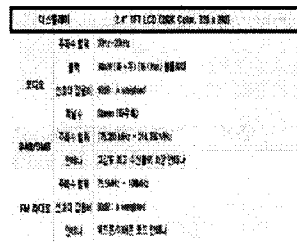


그림 3 MP3 사양테이블 중

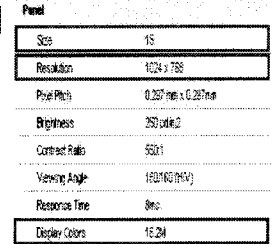


그림 4 LCD 사양테이블 중

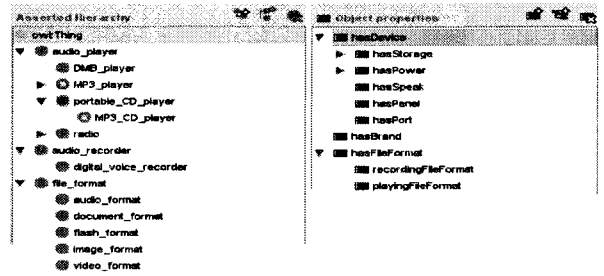


그림 5 MP3 온톨로지의 클래스와 오브젝트 프로퍼티

3.2.2 정보추출과 인스턴스 생성

정보추출을 위해서 Table Information Extraction Module과 Ontology Management Module을 사용한다. Table Information Extraction Module과 Ontology Management Module을 반복적으로 사용하여 정보추출에 사용한다.

Table Information Extraction Module은 MP3 온톨로지를 기반으로 수집된 제품정보 테이블에서 제품 정보를 추출하는 역할을 한다. 제품정보 테이블에서 MP3 제품 정보를 추출하는 과정은 다음과 같다.

- 첫째, 제품 정보가 있는 웹 페이지를 입력한다.
- 둘째, 웹 사이트의 테이블 정보 중에서 제품 정보를 가지고 있는 테이블을 먼저 구분해 낸다.
- 셋째, 테이블 속의 제품정보 구조와 MP3 제품 온톨로지와의 비교를 한다.
- 넷째, 온톨로지에 있는 정보가 테이블 정보에 있으면 추출한 제품정보를 MP3 온톨로지의 인스턴스로 생성한다.

생성된 인스턴스는 MP3 Ontology Repository에 저장
을 한다. MP3 Ontology Repository는 Table
Information Extraction Module에서 추출한 제품정보를
RDBMS에 저장할 수 있도록 온톨로지를 변환하고 저장
하는 역할을 한다.

3.2.3 온톨로지 수정

Table Information Extraction Module에서 제품정보
추출을 위해 MP3 온톨로지와 제품정보 테이블을 비교시
온톨로지에 테이블의 정보가 없을 때, MP3 온톨로지의
수정 후 제품정보를 추출할 수 있다. 이때 사용되는 모
듈이 Ontology Management Module이다.

Ontology Management Module이 온톨로지를 수정하
는 과정은 다음과 같다.

첫째, 테이블에서 추출한 MP3 제품정보 중 MP3 온톨
로지와 비교할 수 없는 정보는 Table Information
Extraction Module에서 온톨로지 수정을 위한 새로운 클
래스나 프로퍼티를 만들 수 있는 추천 정보를 제공한다.

둘째, 온톨로지 수정을 위한 추천 정보는 MP3 온톨로
지를 만들 때 사용한 도메인 온톨로지서 사용할 수 있
는 클래스와 프로퍼티를 검색한다.

셋째, 추천정보가 있으면 MP3 온톨로지의 클래스와
프로퍼티로 추가한다.

넷째, 추천 정보를 검색할 수 없으면 새로운 클래스나
프로퍼티를 직접 넣어준다.

다섯째, 수정된 MP3 온톨로지를 이용하여 다시 제품
정보 테이블에서 제품정보를 추출하고 인스턴스를 생성
한다.

3.3 정보 검색

정보검색을 위하여 Ontology Reasoning Module과
Query Management Module을 사용한다. Ontology
Reasoning Module은 온톨로지를 추론할 수 있는
SWRL(Semantic Web Rule Language)[8]을 지원하며
사용자가 입력한 질의에 대한 추론을 통하여 의미 기반
검색을 할 수 있게 해준다. 그리고 Query Management
Module은 사용자가 입력한 질의를 시스템이 처리할 수
있는 형태로 변환시켜주는 역할을 한다. 온톨로지의 추
론과 사용자 질의 처리를 통하여 정보검색 시 사용자의
질의에 대해 좀 더 적합한 결과를 줄 수 있다.

기존의 MP3 제조사의 홈페이지나 인터넷 쇼핑몰의 검
색서비스는 제조사별, 브랜드별, 출시일, 저장용량, 저장
방식, 부가기능, 가격별 검색 등을 지원하고 있으며 기존
의 검색 서비스에서는 사용자가 원하는 정보를 찾기 위
해서는 검색창에서 제품명이나 브랜드를 입력하여 검색
하거나 MP3 제품이 있는 카테고리를 찾아 들어가서 제
공해주는 검색 서비스에 맞추어 원하는 정보를 찾게 된
다.

기존의 검색서비스가 제공하는 키워드와 “기능이 다양
한 MP3”와 같은 자연어 질의를 처리하기 위해 Query
Management Module을 사용한다. Ontology Reasoning
Module은 SWRL(Semantic Web Rule Language)기반의

추론을 지원하여 새로운 사실을 추론할 수 있다.

다음 [표3]은 MP3 제품에 관련된 추론을 하기 위하여
추론규칙을 정의하였고, 추론규칙을 통하여 MP3제품과
제품의 속성을 추론할 수 있다. [표3]의 추론규칙①~③
까지는 제품속성에 대한 추론규칙이며, ④~⑦은 제품에
대한 추론규칙이다.

표 2 MP3 검색을 위한 SWRL 규칙

SWRL 규칙
① image_format(?file) ^ playingFileFormat(?mp3, ?file) → image_viewer(?mp3, ?file) - 이미지 뷰어 기능을 추론하기 위한 규칙
② document_format(?file) ^ playingFileFormat(?mp3, ?file) → text_viewer(?mp3, ?file) - 텍스트 뷰어(ebook)기능을 추론하기 위한 규칙
③ video_format(?file) ^ playingFileFormat(?mp3, ?file) → video_viewer(?mp3, ?file) - 동영상 기능을 추론하기 위한 규칙
④ MP3_player(?mp3) ^ text_viewer(?mp3, ?file) → mp3_player_with_ebook(?mp3) - ebook 기능을 가지는 mp3를 추론하기 위한 규칙
⑤ MP3_player(?mp3) ^ image_viewer(?mp3, ?file) → mp3_player_with_image(?mp3) - 이미지 뷰어 기능을 가지는 MP3를 추론하기 위한 규칙
⑥ MP3_player(?mp3) ^ image_viewer(?mp3, ?file1) ^ text_viewer(?mp3, ?file2) ^ video_viewer(?mp3, ?file3) → mp3_player_with_multi_function(?mp3) - 다기능 MP3를 추론하기 위한 규칙
⑦ MP3_player(?mp3) ^ video_viewer(?mp3, ?file) → mp3_player_with_video(?mp3) - 동영상 기능을 가지는 MP3를 추론하기 위한 규칙

3.4. 추론 검색 결과

기존의 검색에서는 제조사별, 브랜드별, 출시일, 저장
용량 등의 키워드 검색을 제공하고 있으며, “휴대하기
편한 MP3”, “기능이 다양한 MP3”와 같은 검색은 제대
로 지원해 주지 못하고 있다. 기존 검색에서는 “휴대하
기 편한 MP3”를 검색했을 때, MP3와 휴대용 충전기란
키워드가 포함된 검색결과가 나오며, “다기능 MP3”를
검색했을 때는 MP3와 기능이란 키워드가 있는 것이 검
색결과로 나왔다. “휴대하기 편한 MP3”란 질의를 했을
때, MP3의 크기가 작고 가벼우며, 튼튼한 MP3를 생각할
수 있으며, “기능이 다양한 MP3”란 질의를 했을 때는
음악재생 외에도, 사진을 보거나, 동영상을 볼 수 있는
MP3를 생각할 수 있다.

추론 검색을 통하여 기존 검색보다 나은 검색결과를
제공할 수 있다. 기능이 다양한 MP3는 여러 가지 재생
파일을 지원하고, 재생파일의 종류가 2가지 이상이라고
정의했을 때, MP3가 있고, 음악파일, 이미지파일, 동영
상파일, 텍스트파일을 지원한다는 규칙을 통하여 추론
후 MP3 제품을 검색할 수 있다. MP3 제품 중에서 [그림
6]는 다기능 MP3를 검색하기 위하여 다기능 MP3에 대
한 추론규칙인 MP3_player(?mp3) ^ image_viewer(?mp3,
?file1) ^ text_viewer(?mp3, ?file2) ^
video_viewer(?mp3, ?file3) →
mp3_player_with_multi_function(?mp3)를 이용하여 추론 후

SPARQL[9]을 통하여 검색한 결과이다.

Model_Name
AB
Mu-909
river_B20
river_E10
river_U10

<http://www.w3.org/Submission/SWRL/>
 [9] SPARQL website, <http://w3.org/TR/2007/CR-rdf-sparql-query-20076014>

그림 6 “다기능 MP3”를 추론 검색한 결과

4. 결론 및 향후 연구

본 논문은 다양한 제품정보 처리를 위하여 온톨로지를 이용하여 제품정보를 추출하는 방법과 수집된 제품정보를 바탕으로 제품 검색을 위한 추론 기술을 적용하여 제품정보의 검색과 수집의 처리 효율을 높이고자 한다. 제품정보의 추출시 웹 사이트의 제품정보 테이블이 페이지를 바탕하였다. 제품정보 테이블에서 제품정보를 표현하는 형식이 다르므로 제품 정보 추출시 온톨로지를 기반으로 하여 제품정보의 의미처리를 할 수 있도록 한다.

그리고 제품 정보 추출시 온톨로지의 수정기능을 추가하여 추출시 반영한다. 제품정보의 검색시 기존의 키워드 검색뿐 아니라 사용자의 질의를 확장을 통한 의미 기반의 검색을 지원하고자 한다. 본 연구는 제품정보 수집과 검색에 시맨틱 웹 기술을 적용하여 제품 정보 수집의 정확성 향상과 의미기반의 제품정보 검색이 검색결과의 적합성을 높여줄 것이라고 본다. 앞으로 제품정보 추출에 정확성을 높일 수 있도록 도메인 온톨로지를 보완하여 제품정보 추출시 온톨로지 수정부분의 자동화를 통하여 제품정보 추출의 자동화한다. 그리고 사용자 질의에 대한 자연어 처리나, 사용자가 질의를 쉽게 할 수 있는 인터페이스 연구가 필요하다.

참고문헌

[1] T.Berners-Lee, J. Hendler, O. Lassila, "The Semantic Web", Scientific American, 2001.
 [2] Asuncion Gomez-Perez, Mariano Fernandez-Lopez and Oscar Corcho, "Ontological Engineering", Springer, 2003
 [3] Burcu Yildiz, Silvia Miksch, "Motivating Ontology-Driven Information Extraction", Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007)
 [4] Sung-won Jung, Mi-young Kang, Hyuk-chul Kwon, "Constructing Domain Ontology Using Structural and Semantic Characteristics of Web-Table Head" LNAI 4570, pp. 665--674, 2007
 [5] W3C OWL website, <http://www.w3.org/TR/owl-features/>
 [6] W3C RDF website, <http://www.w3.org/RDF/>
 [7] KAON2 website, <http://kaon2.semanticweb.org/>
 [8] W3C SWRL website,