

상품에 대한 공급자 검색 문제 해결하기 위한 지능형 상품 에이전트 개발 채상용¹, 김경필², 김우주³, 김창욱⁴

^{1,3} Department of Information Industrial Engineering College of Engineering, Yonsei University
134, Shinchonl-Dong, Soedaemun-Gu, Seoul, 120-749, Korea
Tel: +82-2-2123-7754, Fax: +82-2-364-7807, E-mail: slashap@yonsei.ac.kr, wkim@yonsei.ac.kr

^{2,4} Department of Information Industrial Engineering College of Engineering, Yonsei University
134, Shinchonl-Dong, Soedaemun-Gu, Seoul, 120-749, Korea
Tel: +82-2-2123-, Fax: +82-2-2260-8824, E-mail: feel6788@hotmail.com, kimco@yonsei.ac.kr

^{1,3} 연세대학교 정보산업공학과
서울 서대문 신촌동 134, 120-749
Tel: +82-2-2123-7754, Fax: +82-2-364-7807, E-mail: slashap@yonsei.ac.kr, wkim@yonsei.ac.kr

^{2,4} 연세대학교 정보산업공학과
서울 서대문 신촌동 134, 120-749
Tel: +82-2-2123-4010, Fax: +82-2-2260-7807, E-mail: feel6788@hotmail.com, kimco@yonsei.ac.kr

Abstract

인터넷상에 존재하는 수 많은 웹 페이지들에는 정형화되지 않은 각종 정보들이 이종의 형태로 산재되어 있다. 현재의 검색 기술을 통하여 필요한 정보를 찾아내는 것은 시간과 비용이 많이 소요되는 비효율적인 방법으로 이뤄지고 있다. 이러한 상황에서 사용자가 원하는 정보를 검색 및 추출해내어 정형화시키는 것은 매우 중요하다.

전자상거래의 폭발적 성장에도 불구하고 전자상거래 표준 활용 및 적용이 미비하여 e-Procurement, e-Marketplace, on-Line Shopping Mall 등에서 소비자가 원하는 상품 정보를 손쉽게 획득하지 못하고 있다. 이는 공급자에게는 보다 많은 매출의 기회를 구매자에게는 보다 좋은 자재 및 상품을 저렴한 가격에 소싱 할 수 있는 기회를 제공하지 못하는 문제점이 발생한다.

본 연구에서 제안하고자 하는 지능형 상품 에이전트는 소비자가 구매하고자 하는 특정 상품에 대한 공급자 검색 문제를 해결하기 위하여, 시스템 내부 정보의 확장 및 지식화 뿐만 아니라 웹 상의 다양한 상품 정보를 자동적으로 수집 및 가공하여 저장하는 역할을 수행한다. 이러한 연구를 위해서 사용한 기술은 우선 database의 schema를 읽어 들일 수 있는 DB schema reader, 인터넷 웹 페이지(웹 문서)를 방문해서 다양한 정보들의 URL을 수집하는 일을 하는 Meta Search Engine과 Focused Crawler, 그리고 다른 형태의 데이터 구조를 특정 목적에

따라 표준화된 형태로 바꾸는 Wrapper가 있다. 이러한 기술들을 연동하여 필요한 정보들을 추출 공급자 검색 문제를 해결하고자 하는 것이 연구의 목적이다. 정보추출은 사용자의 관심사에 적합한 문서들로부터 어떤 구체적인 사실이나 관계를 정확히 추출하는 작업을 가리킨다.

Keywords

E-Commerce, E-Business, Information Extraction, Information Integration, Semantic Web

서 론

현재의 전자상거래의 폭발적 성장에도 불구하고 전자상거래 표준 활용 및 적용이 미비하여 e-Procurement, e-Marketplace, on-Line Shopping Mall 등에서 소비자가 원하는 상품 및 기업 정보를 손쉽게 획득하지 못하는 문제가 발생하고 있다. 그 결과로 공급자에게는 보다 많은 매출의 기회를 제공하지 못하고 있으며, 구매자에게도 보다 좋은 자재 및 상품을 저렴한 가격에 소싱하지 못하는 문제점이 발생하고 있다.

소비자가 구매하고자 하는 특정 상품에 대한 공급자 검색 문제를 해결하기 위하여 본 연구에서 제안하고자 하는 지능형 상품 및 기업 소싱 에이전트는 웹 상의 다양한 상품 정보 및 기업

정보를 수집 및 가공하여 저장하는 역할을 수행하여 보다 정확하고 신속하게 상품 및 기업 정보를 사용자에게 제공한다. 뿐만 아니라 상품 및 기업 정보와 시스템 내부 정보를 확장하여 정보를 지식화하는 기능을 수행한다.

현재 e-Business는 지속적으로 수행되고 있으며, 기업간 협업 및 전자상거래 기반이 구축되어 국가 산업의 e-Business화 체제가 적용, 확립되고 있다. 이러한 시점에서 제반 자원의 효율성 도모하고, 기업 경쟁력 및 국가 경쟁력 강화를 위한 핵심 전략으로 e-Business의 성공적 정착과 확산을 위한 체계적인 실천 계획이 필요하다는 점에서 지능형 상품 및 기업 소싱 에이전트의 기술을 개발하는 것은 산업적, 경제적, 기술적 이유에서 중요하다고 할 수 있다.

산업적 중요성은 현재 기업 내부 업무 효율 향상을 목적으로 하는 내부 업무 정보화 단계에서 비즈니스 관계에 있는 협력업체와의 협업을 강조하는 조직 정보화 단계로써, 기업간 협업을 위한 국가 기반 체제의 중요성이 날로 증대되고 있다. 또한 국내외 전자상거래 규모의 폭발적인 증가와 전체 산업구조의 재편에도 불구하고, 국내 전자상거래를 위한 상품, 서비스, 기업 콘텐츠 정보 표준 활용이 상당히 미흡한 수준에 머무르고 있어, 전자상거래를 위한 상품 및 기업 정보 검색, 공유의 한계를 증가시켜, e-Business를 이용한 고객 및 기업들의 부가가치 창출에 어려움을 더해가고 있다. 그리고, 경제적 중요성은 새로운 사업기회를 제공함으로써 과거에는 존재하지 않았던 사업이나 상품의 정보 제공 및 획득을 용이하게 하고, 기업 활동에서 발생하는 모든 업무 등이 웹을 기반으로 한 개방형 네트워크를 통하여 통합됨으로써, 기업 통합 및 유통 경로의 단축, 정보 수집 비용 및 상품 생산 비용의 감소, 고객뿐 아니라 제조 및 유통업자에게까지 거래 및 업무의 효율성과 경쟁력을 증가시킬 수 있다.

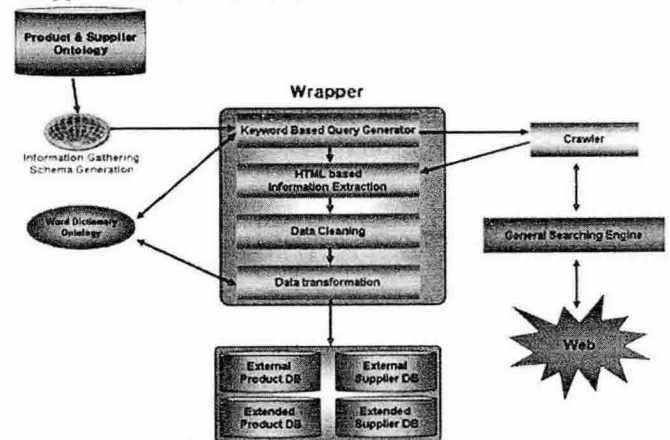
기술적 중요성은 미래 지향적인 의미 웹 언어를 이용하여 온라인상에 존재하는 기업 및 상품 정보를 표현하기 위한 스키마를 제작함으로써, 정보의 확장 및 재사용을 용이하게 하는 것과 비정형화로 구축된 웹에서 정보를 추출하기 위하여 기존의 키워드 기반 정보 추출 에이전트와는 다른 온톨로지 기반 정보 추출 에이전트를 구축하여, 현재의 웹과 의미 웹의 상호 보완적인 발전 가능성을 제시하며, 이질적인 형태를 가진 정보들을 정보 통합 기법으로 재가공하여 표준화된 스키마에 적합하도록 변형하는 방법론을 제시함으로써 정보의 재사용성을 증대시킬 수 있다.

정보추출의 개요

정보추출은 한 문서에서 그 문서의 중심적 의미 또는 추출하려는 사람이 추구하는 의미를 포함하는 특정 구성요소를 인식하여 추출하는 작업을 일컫는다. 우리가 추구하는 정보추출의 예로는 상품정보를 제공하는 쇼핑몰의 웹 문서로부터 가격을 비롯한 화면크기, 전압 등과 같은 상품 정보를 선별하는 것을 들 수 있다.

정보추출의 대상이 되는 정보소스에 존재하는 문서들은 크게 비구조화(unstructured) 문서, 구조화(structured) 문서, 그리고 준 구조화(semi-structured) 문서의 세 가지 형태로 나누어진다. 비구조화 문서는 신문기사와 같이 일반 자연어 문장으로만 구성되어 있고 어떤 일정한 형식이 없는 텍스트 문서를 말한다. 비구조화 문서로부터의 정보추출은 구문분석과 같은 자연어처리를 필요로 하기 때문에 매우 어렵다. 구조화 문서는 실제 데이터와 그 데이터가 나타내는 의미를 메타 데이터를 통해 명시적으로 표현하는 문서 형태이다. 테이블이 하나의 예가 될 수 있으며, 이때 테이블 헤더는 그 열이 나타내는 의미를 기술하는 메타 데이터라고 할 수 있다. 우리가 추구하려는 정보추출은 특수문자를 기준으로 정보를 추출하는 준 구조화 문서에서의 정보추출과 테이블에서 정보를 추출하는 구조화 문서에서의 정보추출이라고 할 수 있겠다.

특히 웹으로부터 특정 요소를 추출하려면 크게 두 가지 기능이 필요한데, 관련 정보를 포함하는 url을 수집하는 Crawler의 기능과, 해당 url의 html 문서에서 특정 요소를 특정 형태로 추출하는 Wrapper기능이 그것이다.



<그림 1 전체 시스템의 구조>

지능형 상품 및 기업 소싱 에이전트는 위의 그림과 같은 아키텍처를 갖는다.

주요 구성요소

- Crawler

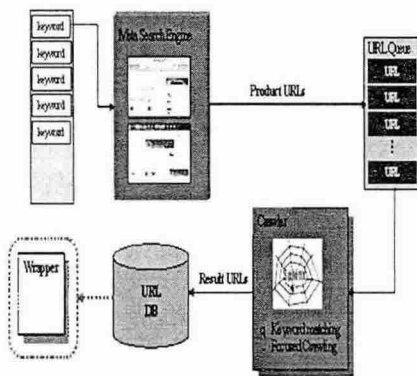
Meta Search Engine이란 크롤러(crawler)와

같은 정보수집 프로그램을 이용하여 대량으로 인터넷 정보를 수집하여 이를 데이터화하고, 이용자들이 편리하게 정보를 검색하여 이용할 수 있게 안내하는 시스템을 말한다. Meta Search Engine은 한번의 Keyword 입력으로 여러 검색엔진을 통한 검색 결과를 가져올 수 있다는 장점이 있다.

Crawler는 인터넷에 웹 페이지(웹 문서)를 방문해서 수집하는 일을 한다. 이 때 한 페이지만 방문하고 끝나는 것이 아니라 그 페이지에 링크되어 있는 또 다른 페이지를 차례로 방문한다. 여기서 Focused Crawler는 한 단계 더 발전하여 해당 HTML문서에 해당 Keyword가 있으면 Crawling을 지속하고 Keyword가 없으면 그 페이지의 하위 링크들은 더 이상 Crawling하지 않는 Intelligent한 Crawler를 말한다.

이러한 Meta Search Engine과 Focused Crawler의 장점을 활용하여 Crawler 모듈은 메인 시스템으로부터 Keyword Set을 입력 받아 Meta Search Engine을 통하여 URL Set을 받아오게 된다. 또한 이 URL Set에 있는 각 URL을 Seed로 하는 Focused Crawler는 Keyword가 존재하는 모든 웹 문서의 URL들을 Crawling한다. Crawling한 결과인 URL들을 상품 속성 DB의 정보를 획득하기 위하여 Wrapper에게 URL Queue 개념으로 URL DB Table에 저장하는 시스템이 Crawler 모듈이다.

구조

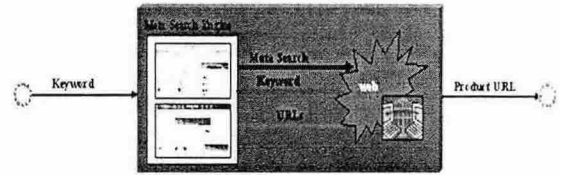


<그림 2 Crawler의 구조>

Crawler는 특정 상품의 product name을 가지고 해당 상품에 대한 url을 수집한다. 수집한 url은 URL DB Table에 저장되며, Wrapper는 URL DB Table의 url을 가지고 상품정보를 수집한다. 우리가 설계한 정보추출 시스템은 URL DB Table을 중앙에 두고 양쪽에 Wrapper와 Crawler가 독립적으로 작동하는 매우 유연한 구조를 장점으로 가진다.

상세기능

- Meta Search

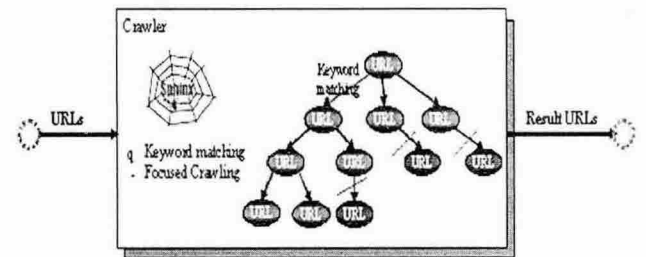


<그림 3 Meta Search>

메인 시스템으로부터 Keyword를 입력 받아 상용화된 검색 포털 사이트를 통해 그 Keyword에 해당하는 웹 문서를 검색하여 URL Set을 가져온다. Meta Search의 용이한 확장을 위하여 각 상용화된 검색 포털 사이트 별 URL Crawl Pattern 데이터를 데이터베이스에 유지한다.

차후 검색 포털 사이트 확장 시, 데이터베이스에 URL Crawl Pattern을 등록하는 과정만으로 Meta Search를 확장을 용이하게 할 수 있다.

- Focused Crawler



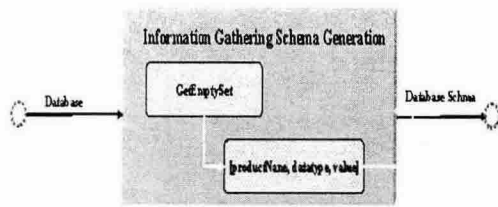
<그림 4 Focused Crawler>

Meta Search 모듈로부터 URL과 Keyword를 입력받아 해당 URL의 Link를 주어진 조건에 따라 Crawling한다. 검색 결과로 도출된 해당 URL 문서에 입력받은 Keyword가 존재한다면 해당 URL 문서의 하위 하이퍼링크로 계속해서 Crawling을 하게 되고, Keyword가 존재하지 않는다면 그 하위의 하이퍼링크로는 더 이상 Crawling하지 않는다. Crawl된 url들을 상품속성DB의 Field를 채우기 위한 모듈인 Wrapper가 사용할 수 있도록 임시저장소인 URL DB Table에 적재한다.

- DB Schema Reader

정보의 추출을 위하여 각 상황에 필요한 상품의 Schema를 생성하는 기능을 가진 것이 Data Base Schema Reader이다. 이는 database내에 있는 속성들의 datatype들이 정의 되어 있는 것과 같이 실제 web에 있는 정보를 추출한 후 database에 insert할 때 web에서 찾은 정보를 database내에 있는 datatype과 일치하게 해주는 것이다.

구조



<그림 5 DataBase Schema Reader>

실제 web에 있는 정보들을 추출하여 실제 database에 datatype에 맞게 입력을 하기 위한 모듈이다. 이는 database내에 있는 productName과 비어있는 record의 값들에 대한 datatype을 database Schema로부터 알아 내어 실제 정보를 입력할 때 database내에 있는 Schema와 맞는 data형태를 되돌려 주는 역할을 한다.

- Wrapper

Wrapper란, 어떤 데이터 구조를 특정 목적에 따라 다른 형태의 데이터 구조로 바꾸는 것이다. CMU의 Sandeep Pandey은 wrapper를 크게 DB wrapper와 web wrapper로 분류했는데, DB wrapper는 DB의 구조를 변형하는 역할을 하고, web wrapper는 웹상에서 문서 검색, 정보 추출, 구조적 재조합 같은 역할을 한다.

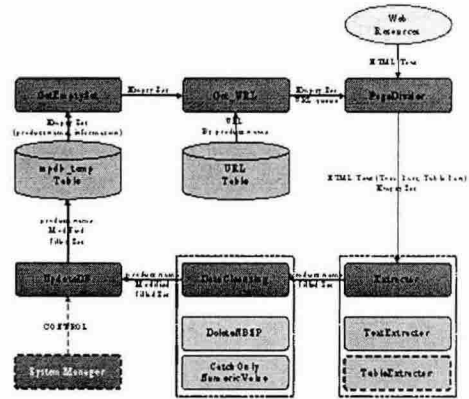
Wrapper는 일반적으로 그 목적에 의해 다음과 같은 다섯 가지 특징을 가진다. 첫째, 디자인적 측면에서 모듈단위의 구조를 가져야 하고, 둘째, 객체지향 환경에서 추구하는 바와 동일하게 쉬운 사용과 재사용이 가능한 구조를 가지며, 셋째, tree구조 혹은 웹상에서 link를 따라 정보를 추출할 수 있는 다계층 추출구조를 가지고, 넷째, 추출할 정보에 대한 명확한 의미를 정의할 수 있어야 하며, 마지막으로 추출된 결과에 대한 쿼리를 수행할 수 있는 구조여야 한다.

초기 Wrapper에 대한 연구는 특정 도메인에서 정보를 추출하기 위해 규칙을 수동으로 생성하였다. 수동 규칙 생성 방법은 사람이 직접 손으로 수행하는 작업이기 때문에 시간이 많이 걸리고 새로운 정보에 대해서 확장성과 유연성이 결여되는 문제점이 있었다. 최근에는 이러한 문제점을 보완하고자 정보추출 규칙의 수동생성에 따른 문제점을 최소화하기 위해, 최소한의 사용자 입력을 받거나 또는 wrapper를 학습시키는 연구들이 진행 중에 있다.

Web wrapper의 중요한 기능 중의 하나인 정보추출은 한 웹 문서에서 필요한 의미를 나타내는 특정 구성요소를 인식하여 추출하는 작업이다. 기존의 정보추출 시스템은 대부분 정보추출 규칙을 수동으로 구성하여 적용하였기 때문에 추출의 정확성은 높지만 유연성, 확장성, 효율성의 측면에서 문제점이 발생한다. 이에 반해 우리가 개발한 wrapper의 장점은 표준 전자 카탈로그

스키마를 이용하여 인터넷 상의 정보를 자동으로 수집, 가공한다. 그 구조를 살펴보면, crawler가 DB에 쌓아 주는 각 상품마다의 URL을 받아와서 DB에 상품 속성값의 record가 비어있는 것을 찾아와 그에 대한 정보를 추출하여 빈 record에 저장을 하는 형태로 이루어져 있다.

본연구의 Wrapper 구조

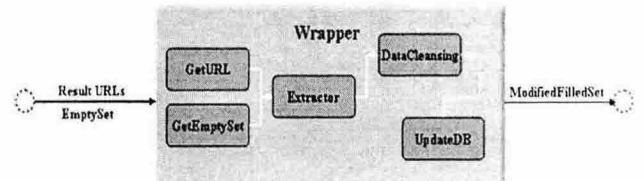


<그림 6 Wrapper의 구조>

이상의 구조를 살펴보면, Crawler가 DB에 쌓아 주는 각 상품마다의 URL을 받아와서 DB에 상품 속성값의 record가 비어 있는 것을 찾아와 그에 대한 정보를 추출하여 빈 record에 저장을 하는 구조로 이루어져 있다.

상세기능

- Wrapping



<그림 7 Wrapping>

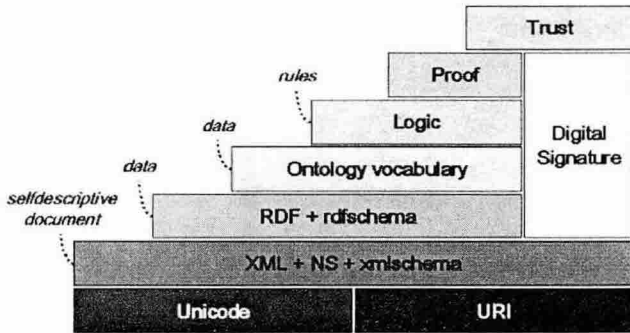
우선 상품정보 DB에 비어 있는 record값을 인식한 후 해당 상품의 url들의 정보를 URL DB Table에서 받아온다. 다음으로는 표 부분을 인식하고 html 문서에서 공백으로 인식이 되는 공백기호들을 삭제하고, extractor에서 전체 상품정보를 추출한 후, DB 스키마에 맞게 정보 유형을 변환하고 마지막으로 DB에 정보를 갱신하는 기능을 가진다.

기술적인 차별화

본 연구에서 기술적인 차별성을 두기 위해서 시맨틱 웹 에서 주요 근간을 이루는 온톨로지를 활용하였다.

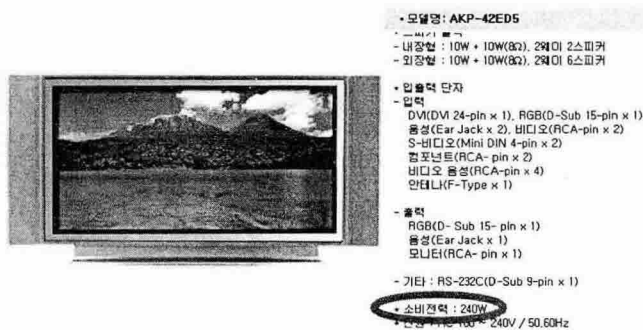
오늘날 웹 자원이 폭발적으로 증가함에 따라, 정보의 양이 인간이 처리할 수 있는 한계를 초과함으로써, 인간 대신 기계가 웹 자원을 스스로 처리할 수 있도록 웹에 의미를 부여하는 시맨틱 웹 분야 연구에 대한 수요도 기하급수적으로 증가하고 있다.

시맨틱 웹은 아래 그림에서 보듯이 일곱 계층으로 구분되는데, 현재 시맨틱 웹의 추세는 2단계인 발전과 전개 단계(계층 4 - Ontology vocabulary의 활용)이다.



<그림 8 The Semantic Web Layers>

본 연구에서는 이러한 Ontology vocabulary를 Wrapper에서 활용함으로써 보다 정확한 정보추출을 가능하게 하였다. 그 예시는 다음과 같다.



<그림 9 쇼퍼몰의 예시>

위 그림과 같은 상품을 인터넷상에 존재하는 쇼퍼몰등에서 html Wrapping을 할 때 아래와 같은 온톨로지를 구축하여 검색에 사용 하여 보다 정확한 정보검색을 하게 한다.

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:j.0="http://protege.stanford.edu/plugins/owl/protege#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns="http://www.owl-ontologies.com/unnamed.owl#"
  xml:base="http://www.owl-ontologies.com/unnamed.owl">
  <owl:Ontology rdf:about="">
    <owl:imports
      rdf:resource="http://protege.stanford.edu/plugins/owl/protege"/>
    </owl:Ontology>
    .....
    <owl:Class rdf:ID="Television">
      <rdfs:subClassOf>
        <owl:Restriction>
          <owl:has Value>
```

```
<PowerConsumptionUnit rdf:ID="Watt">
  <unit
    rdf:datatype="http://www.w3.org/2001/XMLSchema#string">W</unit>
  </PowerConsumptionUnit>
  <owl:hasValue>
    <owl:onProperty>
      <owl:FunctionalProperty
        rdf:ID="hasPowerConsumptionUnit"/>
      </owl:onProperty>
    </owl:Restriction>
  </rdfs:subClassOf>
  .....
```

<그림 10 온톨로지 예시>

위의 온톨로지를 사용함으로써 위의 TV 모델명에서 속성값이 소비전력인 것에서 value값이 Watt로 나와 있어야지 보통의 html Wrapping에서는 정확한 값으로 인식을 한다. 그러나 본 연구에서 기술적인 차별성을 두고자 한 것은 온톨로지를 활용함으로써 Watt값이 unit으로 String값인 W를 가지고 있으므로 Wrapper가 정확한 추출을 원하는 Watt를 W로도 인식을 하고 정확한 value값을 추출해 낼 수 있는 것이다. 즉, 속성값 추출의 시점에서 추출의 대안을 평가하고, 속성의 단위를 확인하며, 속성의 도메인을 확인하고, 유사 속성등을 확인하는 작업을 하는 것이다. 이러한 온톨로지 vocablury가 모든 Wrapping하는 상품에 관하여 구축이 되어 있으면 보다 정확한 정보추출이 가능하게 될 것이다.

기대 효과 및 활용 방안

지능형 상품 소싱 에이전트 개발은 기업, 파트너, 공급자, 고객들로 분리된 비즈니스 환경을 통합함으로써 향상된 e-Business 수행을 위한 통합 환경을 제공함으로써 전자상거래 비즈니스 모델을 변화시킬 것으로 기대가 된다. 뿐만 아니라 소싱 에이전트 개발 적용은 자동화된 비즈니스를 수행하기를 원하는 기업의 거래 파트너, 고객, 공급자 등을 통합하여 고객이 원하는 상품 및 기업 정보를 자동적으로 연결해 줌으로써 탐색 비용이 크게 감소하고, 이를 통해 향상된 e-Business를 수행할 수 있게 한다.

- 웹 서비스를 통한 자동화된 비즈니스 모델 변화
- 기업의 거래 파트너, 고객, 공급자 통합 환경
- 상품 및 기업 정보 자동적 연결
- 개별 웹 서비스를 연결하는 브로커 등장으로 다이나믹한 e-Business 환경 조성

이에 대한 활용방안으로서는, 현재 e-Business로 거래되고 있는 상품 및 공급자 정보를 인터넷 통합 검색을 통해 검색하고, 이를 UNSPSC 등 국내외 표준 분류체계에 따라 상품분류, 식별표준, 속성표준 스키마를 정의하여 상품 및 공급자 정보의 표준 전자카탈로그 데이터베이스를 구축한다. 이를 기반으로 기존에 구축된 e-Procurement, e-

Marketplace, 온라인 쇼핑물과 연계하여 고객이 원하는 상품 및 공급자 정보를 폭넓고 정확하게 검색하여 비교 평가 할 수 있도록 서비스를 제공한다.

또한 기업 내부 또는 개별 구축 운영되는 e-Procurement, e-Marketplace 등에 이미 존재하는 상품정보, 공급자정보 등을 이용하여 상품과 공급자 사이의 지식을 추출하여 표준 전자카탈로그 형식으로 데이터베이스를 구축하고 이를 통해 고객이 다양한 정보를 쉽게 검색할 수 있도록 활용한다.

- 상품/공급자 정보에 대한 통합 검색 및 수집
 - e-Procurement, e-Marketplace, 온라인 쇼핑물 등 인터넷상에 거래되고 있는 상품과 공급자에 대한 정보 획득
 - 수집된 정보에 대한 데이터 Cleansing 작업을 통한 분석
 - 데이터 분석을 통해 상품/공급자 정보의 보강
- e-Procurement, e-Marketplace, 온라인 쇼핑물 등과 연계
 - e-Procurement, e-Marketplace, 온라인 쇼핑물 등에 개발 구축된 상품/공급자 정보 수집 서비스와 연계하여 제공
 - 정보 수집 서비스 제공을 통한 시장 진입 조기화 & 가시화

결론 및 Future Work

본 연구에서는 현재 계획되어 있는 Wrapper, Crawler 개발이 되어 있으며 이 두 가지의 프로세스를 연동하여 상품의 정보추출에 관하여 지능형 상품 소싱 에이전트를 개발하여 전자상거래에 있어 사용자의 만족도를 높이려고 하였다.

향후 추진계획으로는, 의미 웹을 이용하여 추출된 정보를 표현함으로써 기존의 데이터베이스를 이용하여 데이터를 저장하는 것과는 다르게 하나의 사실로써 정보를 저장하고 기업 및 상품 정보 등은 정보검색에서 지식으로 사용할 수 있게 할 것을 더 연구 하고자 한다.

참고문헌

[1] Tim Berners-Lee, James Hendler, Ora Lassila. (2001). "The Semantic Web" *Scientific American*. 2001.
 [2] David W. Embley, Cui Tao, Stephen W. Liddle (2002) "Automatically Extracting Ontologically Specified Data from HTML Tables with Unknown

Structure"

[3] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogeneous information sources. In *IPSJ Conference*, pages 7-18, Tokyo, Japan, October 1994.
 [4] Chen, Tsai, Tsai (1999) "Mining Tables from Large Scale HTML Texts"
 [5] Lira, C.Y. and Koo, J.L.T. (1999) "Learning to Recognize Tables in Free Text," *Proceedings of the 37th Annual Meeting of ACL*, 1999, pp. 443-450.
 [6] Jean Robert Gruser, Louiqa Raschid, Maria Esther Vidal, Laura Bright. "Wrapper Generation for Web Accessible Data Sources," *coopis*, p. 14, Third International Conference of Cooperative Information Systems
 [7] Selberg, E. and O. Etzioni (1997). "The MetaCrawler Architecture for Resource Aggregation on the Web," *IEEE Expert*, Vol. 12, pp. 11-14.
 [8] Howe, A. E. and D. Dreilinger (1997). "Savvy Search: A Metasearch Engine that Learns which Search Engines to Query," *AI Magazine*, Vol. 18, pp. 219-25
 [9] Soumen Chakrabarti, Martin van den Berg, Byron Dom (1999). "Focused crawling: a new approach to topic-specific Web resource discovery," Elsevier Science B. V
 [10] Robert C. Miller and Krishna Bharat (1998) "SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers," WWW7, Brisbane Australia