한국지능정보시스템학회논문지 제14권 제1호 2008년 3월(pp.67~84)

시맨틱 검색 시스템의 개념적 모형화와 그 구현에 대한 연구*

한동일
KT 미래기술연구소
(dihan@kt.co.kr)

권혁인
중앙대학교 상경학부
(hikwon@cau.ac.kr)

정학진
KT 미래기술연구소
(hjchong@kt.co.kr)

본 논문은 시맨틱 검색 시스템에 관한 포괄적인 개념적 모델 제안과 실질적인 구현 사례를 제시한다. 제안된 시맨틱 검색 시스템은 개념적으로 3계층의 아키텍처 지식획득 계층, 지식표현 계층, 지식이용 계층으로 구성하여 설계 및 구현되었다. 지식획득(Knowledge acquisition) 계층은 다양한 소스(Source)의 콘텐츠(텍스트, 이미지, 멀티미디어 등)로부터 시맨틱 메타데이터를 생성 및 저장하는 영역이다. 지식표현(Knowledge Representation) 계층은 온톨로지의 스키마와 인스턴스를 구축하고, 이러한 온톨로지 기반 질의 확장 등을 통해 시맨틱 검색을 처리하는 영역이다. 마지막으로 지식이용(Knowledge Utilization) 계층은 검색 이용자가 시맨틱 웹 언어 또는 온톨로지에 대한 지식이 없더라도 직관적으로 검색 질의(Query)를 입력하고 검색 결과를 확인할 수 있도록 구성하였다. 향후 제시된 시맨틱 검색 시스템은 기존 연구 수준의 시맨틱 검색 시스템을 상용화 수준으로 향상시킬 수 있는 계기가 될 것으로 기대된다.

논문접수일 : 2007년 02월 게재확정일 : 2007년 10월 교신저자 : 한동일

1. 서론

시맨틱 웹(Berners-Lee, 2001) 환경에서는 찾고자 하는 용어(term)의 복잡한 의미(Meaning)와 용어간의 관계(Relationship)를 온톨로지를 매개로 기계가 이해(Machine-understandable)할 수 있어서 인간의 개입을 최소화 한다. 그와 동시에 지능화된 에이전트(Autonomous Agent)의 복잡한 질의 해석을 통해 검색 결과를 제공하게 된다. 현재까지는 시맨틱 웹 기술이 잠재력을 완전하게 실현시키기 위해서는 좀 더 많은 연구가 필요하지만, 그 기술에 대한 미래는 매우 밝게 전망되고 있다(Passin, 2004).

그러나 현재의 웹 검색은 검색 이용자에게 키워드와 적중률(Hit ratio)을 기준으로 다량의 검색 결과를 제공하므로, 검색 이용자의 검색 요구에 좀 더 근접한 결과를 찾기 위해서는 관련 없는 검색 결과도 어쩔 수 없이 볼 수밖에 없는 상황이다. 다시 말하면, 현재의 웹 검색 방식은 검색 이용자의 니즈(Needs)를 충족시킬 만큼 정확하고 지능적인 검색

* .본 연구는 정보통신부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업[2005-S-083-02, 차세대 웹을 위한 시맨틱 서비스 에이전트 기술 개발]과 KT 지능형 검색서비스 기술개발 사업의 일환으로 수행하였음.

을 하지 못하고 있다(Ponnada and Shardal, 2007).

차세대 검색 엔진은 일반적으로 새로운 기술 즉, 시맨틱 검색, 클러스터링 검색, 사회적/개인화 검색 기술 등을 활용하는 향상된 검색 엔진을 통칭한다. 우선 시맨틱 검색은 소위 Web 3.0 기술로 일컬어지는 시맨틱 웹(Semantic Web) 기술인 온톨로지, 추론 기술, 메타데이터 생성 기술 등을 검색에 적용한 것이다. 그 예로, Hakia(www.hakia.com)와 같은 베타버전 검색 엔진은 특정 인물 명을 검색 창에 입력하면 특정 인물을 추론하여 해당 키워드가 포함되지 않은 문서라 하더라도 의미적으로 연관된 정보를 검색결과로 제공한다. 클러스터링 검색과 사회적/개인화 검색 기술은 각각 통계적 분류, 태그 기반 키워드 검색 기술 등을 적용하고 있다. 예를 들어, Clusty(www.clusty.com)와 같은 클러스터링 검색은 특정 키워드를 입력하면 검색결과 화면에 키워드와 통계적으로 유사한 주제를 그룹핑하여 제공하며, Collarity(www.collarity.com)와 같은 사회적/개인화 검색 엔진은 내가 속한 커뮤니티의 사람들이 자주 사용하는 키워드로 검색결과를 제공한다. 그러나 클러스터링 검색과 개인화 검색은 각각 클러스터링 기준의 모호성, 주관적인 개인화 성향을 해결해야 하는 과제로 안고 있다. 한편 시맨틱 검색 방식은 기존 콘텐츠에서 메타데이터를 자동으로 생성 및 온톨로지를 관리해야 하는 어려움이 있다.

일반적으로 웹 이용자들이 검색을 수행하는 이유는 검색 이용자의 불완전한 (Incomplete) 지식을 완전한 (Complete) 상태(Albertoni et al, 2004; Belkin, 1980)로 바꾸는 지식 충족의 욕구 때문이라고 한다. 그럼에도 불구하고, 현재의 웹 검색 이용자들은 약 10cm 크기의 직사각형 검색 창에 본능적으로 고급 검색보다는 한 단어의 키워드만 입력하는 행태를 보여주고 있다. 이는 인지적 구두쇠(Cognitive Miser) 현상으로, 정보 과부하(Information overload) 속에서 원하는 정보를 찾기 위해 부담해야 할 인지적 노력(Cognitive load)을 최소화하고자 하는 검색 이용자들의 행동으로 판단된다)(Belkin and Croft, 1992; Spink et al, 2001). 시맨틱 검색은 정보 검색 이론의 핵심적인 개념인 이러한 검색 이용자의 정보 니즈(Information Need)를 가장 잘 충족시킬 수 있다(Wissbrock, 2004). 왜냐하면, 시맨틱 기술을 활용한 의미기반 검색 엔진은 검색 이용자가 정확한 의미를 몰라도 사람의 말을 이해하는 것처럼 검색 문장의 의미를 파악해 의미에 맞는 대상을 찾아주는 방식이기 때문이다(Liu et al, 2003). 또한 시맨틱 검색은 검색에 투입되는 수고와 인지/심리적 부담감을 최소화하면서도 검색결과는 정확하고, 검색 의도와 의미적으로 연관된 정보를 제공하기 때문이기도 하다.

그러므로 본 연구에서는 기존 웹 검색과 시맨틱 웹 문헌 연구를 토대로 차세대 웹 검색 특히, 시맨틱 검색을 중심으로 포괄적인 개념적 모델을 제안하고, 실질적인 구현 사례를 제시하고자 한다. 제 2장에서는 관련 연구를 제시하고, 이를 토대로 제 3장, 제 4장에서 각각 시맨틱 검색 시스템의 개념적 모델을 도출하고 시맨틱 검색 시스템을 구성하였으며, 제 5장에서는 구현한 시맨틱 검색시스템에 대해 논의하고, 마지막으로 제 6장에서 결론을 맺도록 기술하였다.

2. 관련 연구

다양한 시맨틱 웹 응용 중에 하나인 시맨틱 검색은 3가지 핵심 기술 요소 즉, 시맨틱 메타데이터 생성, 온톨로지 구축, 그리고 브라우징과 질의가 매우 중요하다. 아래에서는 시맨틱 검색의 개요와 시맨틱 검색의 요소 기술을 포괄적인 관점에

서 분야별(지식획득, 지식표현, 지식이용)로 구분하여 살펴보고자 한다. 또한 기존 시맨틱 웹 관련 문헌연구를 토대로 시맨틱 검색 시스템의 요구사항을 도출하였다.

2.1 시맨틱 검색 개요

현재 인터넷의 보편화와 방대한 웹 데이터로 인해 더 이상 기존 웹 검색 방식에 의존하지 않고 정보의 의미를 검색할 수 있는 시맨틱 검색에 대한 필요성이 증대되고 있다. 기존 웹 검색이 텍스트(Texts)에서 키워드 매칭 방식이라면, 시맨틱 검색에서는 객체(Objects) 검색 방식이며 검색 결과도 단순 URL이 아닌 객체의 개념(Concepts), 속성(Properties), 개체(Instances)를 포함한다(Guha et al, 2003). 이러한 시맨틱 웹 기반 검색은 검색 이용자가 웹 페이지에서 임의의 키워드를 찾을 때 보다는 해당 키워드가 의미하는 하나 또는 그 이상의 개념(Concept)을 찾고자 할 때 유용하다(Bangyong et al, 2005 Richa, 2004). 더욱이 웹 페이지와 연계된 메타데이터가 풍부할 때는 시맨틱 검색의 장점이 더욱더 부각된다. 그러나 이러한 장점에도 불구하고 시맨틱 검색은 검색 이용자로 하여금 찾고자 하는 해당 키워드가 속한 개념(Concept) 형태로 질의를 표현해야 한다는 인지적 부담이 있다(Albertoni et al, 2004 Makela et al, 2006 Richa, 2004 Sure and Iosif, 2002).

일반적으로 정보 검색 시스템 성공의 핵심이 정보 니즈(Information Needs)의 충족임에도 불구하고, 대부분의 정보 검색 시스템들은 이를 중요하게 간주하지 않고 있다(Albertoni et al, 2004 Wissbrock, 2004). 검색 이용자가 그들의 정보 니즈를 질의(Query) 형태로 형식화할 때 정보 니즈가 정확하게 정의될 수 있다고 가정하지만, 많은 경우에 이러한 가정은 옳지 않았다(Morville, 2005 Oddy, 1977). 비록 몇몇 사례의 경우 검색 이용자가 검색하고자 하는 대상을 알고 명시적으로 표현할 수 있으나 대부분의 경우는 그들의 정보 니즈를 명쾌하게 명시적으로 표현할 수 없었다. 다시 말하면, 불안전한(Incomplete) 문제를 해결하기 위한 지식이 필요할 때마다 새로운 정보가 필요하고, 이러한 그들의 불완전한 지식을 완전하게(Complete)하려고 새로운 정보를 검색한다. 이러한 이유로 시맨틱 검색을 포함하는 모든 검색의 니즈는 정보 검색 이용자의 불완전한(Incomplete) 지식을 완전한(Complete)한 상태로 유지하고자 하는 것임을 알 수 있다(Belkin, 1980).

위와 같은 연구를 통해서도 알 수 있듯이 시맨틱 검색의 장점이 부각되기 위해서는 검색 이용자의 니즈(Needs)를 반영할 수 있고, 질의 표현의 인지적 부담을 최소화 할 수 있도록 구성해야 한다.

2.2 시맨틱 검색 시스템의 요소기술 연구

시맨틱 검색은 복잡한 검색 질의를 토대로 만족스러운 검색 결과를 제공하기 위해 웹 콘텐츠에 대한 의미 이해뿐만 아니라 논리적인 추론도 할 수 있다. 이러한 이유로 시맨틱 검색 시스템의 구조는 기존 단순 단어에 대한 인덱싱 방식이나, 클라이언트 계층과 서버 계층의 2계층 구조(Muddassar Iiyas et al, 2004)에서와 같은 단순한 개념적 아키텍처를 통한 단편적인 접근 방식에는 한계점이 있다. 그러므로 아래에서 시맨틱 검색의 요소 기술을 포괄적인 관점에서 지식획득 계층, 지식표현 계층, 지식이용 계층으로 구분하여 관련 연구를 살펴보도록 하겠다.

- **지식획득 연구**

지식획득(Knowledge acquisition) 분야는 다양한 소스(Source)의 콘텐츠(텍스트, 이미지, 멀티미

디어)로부터 시맨틱 메타데이터를 생성하는 영역이다. 지식의 획득 관점에서 가장 기본적인 시맨틱 기술은 자연어 처리(Natural Language Processing), 통계 기법(Statistics), 기계 학습 기법(Machine Learning Techniques) 등을 나열할 수 있다. 이러한 기술들은 다양한 콘텐츠로부터 의미 있는 시맨틱 메타데이터를 추출하기 위해 객체 인식(Named Entity Recognition)과 의미적 모호성 해결(Semantic Ambiguity Resolution) 방식으로 활용된다. 위와 같은 기술적 해결책이 없이는 다양한 콘텐츠로부터 시맨틱 검색 대상이 되는 시맨틱 메타데이터 생성이 매우 어렵다. 그럼에도 불구하고 대규모 메타데이터 생성과 시맨틱 어노테이션(annotation)이 가능(Dill et al, 2003)하나, 시맨틱 메타데이터 생성에서 생성되는 양과 질에는 상충관계(Trade-off)가 있다(Sheth, 2004). 예를 들면, 단순한 형태의 메타데이터 추출(데이터 생성 날짜, 문서 크기, 작성자 등)을 통한 대규모 메타데이터 생성이 수행될 수 있으나, 좀 더 의미적인 메타데이터 추출(회사, 본부, 산업, 비즈니스 등)은 완전 자동화하기에는 아직 미해결 이슈가 산재되어 있다.

지식획득 분야에서 중요한 또 다른 이슈는 다수의 생성된 메타데이터를 저장하고 관리할 수 있는 인덱스 기법이다. RDF와 RDF Schema(RDF/RDFS), OWL로 표현되는 콘텐츠의 양이 증가하고 있고, 이러한 콘텐츠를 효율적으로 구조화하여 저장, 검색할 수 있는 인덱스의 필요성이 높아지고 있기 때문이다(Harth and Decker, 2005). 현재까지 다양한 인덱스 구조가 제안되고 있으며 복잡한 구조에서의 성능과 대용량 처리 분야에서 개선하려는 연구들이 진행되고 있다.

- **지식표현 연구**

지식의 표현(Knowledge Representation) 분야는 온톨로지의 스키마와 인스턴스를 구축하고, 이러한 온톨로지 기반 질의 확장 등을 통해 시맨틱 검색을 처리하는 영역이다. 온톨로지는 일종의 공유된 개념으로 컴퓨터와 인간이 동시에 이해할 수 있는 지식의 표현 형태이다. 이러한 온톨로지를 시맨틱 검색에 적용하려면 크게 3가지 관점을 고려해야 한다(Sheth, 2004).

첫째, 대다수의 실제 온톨로지는 반형식적(semi-formal)으로 기술되어야 한다. 반형식적인 온톨로지는 온톨로지가 부분적으로 불완전한 지식 예를 들어, 불일치성, 제약조건 위반 등의 형태로 구축되기 때문이다. 즉, 다양한 소스로부터 지식을 추출하고 통합하며, 수많은 작업자에 의해 온톨로지가 구축되므로 이러한 반형식적 상태가 불가피하다(Gruber et al, 2003).

둘째, 구축될 온톨로지가 지나치게 표현이 상세화된다면(Heavy Ontology), 실제 애플리케이션 적용될 때 별다른 가치를 제공하지 못한다. 이러한 현상의 주요 원인은 온톨로지에 표현된 지식을 포착(capture)하기가 매우 어렵기 때문이다. 따라서 추론의 성능 저하가 우려된다. 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. 그러나 시맨틱 웹 응용 분야에 따라서는 복잡한 온톨로지(Heavy Ontology)가 적합한 경우도 많다.

마지막으로 시맨틱 웹 응용 분야가 검색, 탐색, 개인화 영역일 경우로 지식의 발견 및 지능적/분석적 애플리케이션에 적용할 때에는 도메인과 태스크 특화된 심도 깊은 시맨틱 메타데이터가 요구된다. 아울러 객체/개념(Entity/concept) 보다는

관계(Relation)의 로직 처리가 요구될 수 있다.

온톨로지를 시맨틱 검색에 적용하기 위한 위에서 언급한 3 가지 고려사항 외에도 최근에 멀티미디어 콘텐츠가 일반화되고 있는 추세를 반영하기 위해 주요 지식표현 방법에 대해 2가지 관점에서도 살펴보아야 한다(Sheth, 2004). 첫 번째는 기술기반(Description-based) 표현방법이다. 이 방식은 멀티미디어 콘텐츠를 일반적으로 창작자의 이름, 이미지 크기, 자막/서브타이틀, 키워드 기반으로 표현하는 방식이다. 또 다른 접근방식은 콘텐츠 기반(content-based) 표현방식으로 이미지 색상, 소리의 높낮이와 같은 고유의 특성에 따라 멀티미디어 콘텐츠를 표현하는 방식이다. 시맨틱 검색 관점에서는 두 가지 장점을 결합시킨 통합 방법으로 멀티미디어 표현을 취해야 좀더 고품질의 검색이 가능할 것으로 판단된다.

이러한 고려 요인에도 불구하고 지식표현 분야는 정보가 검색되기 전에 표현되어야 하고 지식표현의 질이 검색 성능에 직접적인 영향을 미친다는 이유 등으로 인해 시맨틱 검색의 본질이라고 할 수 있다.

- **지식이용 연구**

지식의 이용(Knowledge Utilization) 분야는 검색 이용자 관점에서 시맨틱 검색 시스템과의 인터페이스 영역이다. 시맨틱 검색에서 중요한 이유는 지식획득과 지식표현의 궁긍적인 목표는 정보 검색 이용자 요구를 만족시키는데 있기 때문이다. 이러한 이유로 이기종 콘텐츠를 검색하기 위해 온톨로지와 메타데이터를 동시에 활용하는 검색 이용자의 시맨틱 질의 처리는 고부가가치 분야라고 할 수 있다.

현재까지 시맨틱 검색의 상용화 사례는 미흡하며, 시맨틱 웹 검색 인터페이스에 대한 연구는 연구자의 실험결과를 확인해 보는 실험실 수준에 머물러 있어서 시맨틱 웹의 상용화에 많은 어려움이 산재되어 있다. 또한 실험실 수준의 검색 방식은 기존 웹 검색과 너무 상이한 검색 입력 창으로 구성되어 있고, 검색 방식이 너무 복잡하여 익숙하지 못한 이용자로부터의 거부감을 초래할 수 있다는 문제점이 있다. 예를 들어, OWL(Web Ontology Language) 검색을 지원하며 SPO(Subject, Predicate, Object)형태의 질의문을 입력받거나, RDQL(RDF Data Query Language) 질의 형태의 입력 폼을 제공하고 질의문을 입력 받아 검색을 수행하도록 구성되어 있다. 또는 기존의 검색엔진의 결과를 개선시키는데 키워드 중심이 아닌 개념(concept)과 관계(relation) 중심으로 입력 폼을 구성하였다. OntoWeb(www.ontoweb.org) 경우는 시맨틱 웹 검색 인터페이스의 대표적인 사례로 브라우징 → 속성 선택 → 실제 값 입력 → 검색 수행의 순으로 단계별 검색을 수행한다.

Albertoni의 연구(Albertoni et al, 2004)에서는 시맨틱 검색을 웹 자원과 현실 세계의 객체들(Objects)에 관한 의미를 명시적으로 표현하며, 검색 결과의 정확율과 재현율을 개선시키고 검색 이용자의 니즈를 파악하고 이를 충족하려는 시도를 추진하였다. 이러한 연구에서는 3가지 문제점을 언급하였다. 정보 검색 행위에 영향을 미치는 인간 행위 요인으로 제한된 정보 검색 이용자의 지식, 정보 제공자와 정보 검색 이용자간의 인식 차이, 그리고 검색 이용자가 검색 니즈를 충족하지 못할 수 있다는 점을 지적을 하였다. 이를 해결하기 위해 기존 검색 시스템을 검색 이용자와 검색 시스템간 상호작용과 시각화, 그래픽 한 상호작용, 그래픽 한 시각화를 강조하였다. 또한 기존 시맨틱 검색 시스템들은 온톨로지 구조를 조회/수정할 수 있을 뿐 지능적 브라우징이 부족함도 언급하였

다. 다시 말해, 검색 시스템과의 상호작용성을 중요한 인터페이스 요인으로 간주하였다.

Makela의 연구(Makela et al, 2006)에서도 시맨틱 검색과 검색 이용자간의 질의 생성과 화면 구성 관점의 연구를 수행하였다. 시맨틱 기술은 개념과 관계를 이용하여 지식을 그래프로 표현하기에 적합하지만 검색 이용자가 이러한 지식을 검색하려는 질의 생성이 어려움을 지적하였다. 이 연구에서는 시맨틱 UI(User Interfaces)를 제안하여 검색 이용자가 생성하기 어려운 질의 생성도 쉽게 구성할 수 있음을 비교 실험을 통해 제시하였다. 다시 말해, 시맨틱 검색을 위한 사용의 용이성을 강조하였다.

현재까지 위의 연구에서와 같이 시맨틱 웹 검색 인터페이스는 복잡하고, 검색 과정을 인식해야 하며, 기존 일반 웹 검색 인터페이스와 상당한 차이점이 존재하고 있어 검색 이용자가 학습하는데 많은 시간이 소요되리라고 판단된다.

2.3. 시맨틱 검색 시스템 요구사항 (Requirements)

기존 시맨틱 검색 문헌 연구를 토대로 지식획득, 지식표현, 지식이용 계층별로 시맨틱 검색 시스템의 주요 요구사항을 요약해 보면 아래와 같다.

첫째, 지식획득 계층에서의 고품질 시맨틱 메타데이터 생성 (반)자동화와 생성된 메타데이터의 구조화 방안이 요구된다.

웹 콘텐츠의 양이 방대하므로 검색 이용자의 니즈를 만족시킬 수 있는 고품질의 시맨틱 메타데이터 생성이 절실하다. 또한 생성된 메타데이터를 검색의 효율성을 위해 구조화하여 저장하는 과정이 요구된다. 현재까지 시맨틱 메타데이터 생성의 자동화 기법으로 자연어 처리(Natural Language Processing), 통계 기법(Statistics), 기계 학습 기법(Machine Learning Techniques) 등을 시도하고 있으나 구체적으로 적용된 사례가 부족하며 고품질의 의미 있는 메타데이터 생성은 더욱더 어려운 실정이다. 또한 시맨틱 메타데이터 저장 방식에 있어서도 이론적인 연구는 진행되고 있지만 검색 이용자의 질의에 적합한 상용화 수준의 인덱싱 방식에 대한 구현사례가 미흡하다.

둘째, 지식표현 계층의 온톨로지의 형태와 범위는 응용 분야에 따라 다르게 구축한다. 단순한 온톨로지만 필요로 하는 응용 분야에는 온톨로지를 상대적으로 단순하게 구성한다. 그러나 시맨틱 검색 분야와 같이 복잡한 응용 분야에서는 온톨로지의 개념뿐만 아니라 관계를 처리해야 할 경우가 있으므로 적절한 크기와 범위를 설정하여 온톨로지를 구축(Right Ontology)하여야 한다. 또한 메타데이터 생성에 있어서도 단순 메타데이터보다는 의미를 포함하는 시맨틱 메타데이터 생성이 요구되는데, 이를 위해서는 학제간 기법 즉, 자연어 처리 기술, 통계 기법, 기계 학습 기법 등을 통합적으로 사용할 수 있어야 한다.

마지막으로 지식이용 계층에서는 3가지 관점에서 살펴볼 수 있다. 우선 시맨틱 웹 검색 관점에서 검색의 니즈(Needs)를 반영할 수 있도록 시맨틱 웹 검색 시스템의 상호작용성을 고려해야 한다(한동일, 홍일유, 2007). 다음으로, 검색 이용자가 그들의 정보 니즈를 질의(Query) 형태로 형식화할 때 정보 니즈를 명쾌하게 명시적으로 표현할 수 없으므로 검색 시스템과 검색 이용자간의 상호작용 지원을 통해 검색 질의의 정제 과정이 유기적으로 지원될 수 있도록 함은 물론, 질의 생성의 어려움을 해소할 수 있도록 사용이 용이한 검색 시스템을 설계해야 한다. 이러한 상호작용 과정과 질의 생성 과정의 사용 용이성은 검색 이용자의 불완전한(Incomplete) 지식을 완전한(Complete)한 상태로 유지할 수 있도록 지원한다(Albertoni et

al, 2004 Belkin, 1980). 마지막으로 시맨틱 웹 인터페이스 관점에서 일반 웹 검색과의 차이점을 극복할 수 있도록 인터페이스의 친근성, 인터페이스의 유용성, 조작의 용이성 등을 고려해야 한다. 기존 웹 검색과 너무 상이한 검색 입력 인터페이스로 구성되거나, 검색 방식이 너무 복잡하면 익숙하지 못한 사용자로부터의 거부감을 초래할 수 있다. 또한 웹 검색 인터페이스가 너무 복잡하고 검색 과정을 인식해야 하므로 조작이 어렵거나, 학습에 많은 시간이 소요될 것으로 예상되어 유용성이 없어서는 안되므로 이에 대한 신중한 고려가 요구된다. 아래 <표 1>에 시맨틱 검색 시스템 요구사항을 정리하였다.

<표 1> 시맨틱 검색 시스템 요구사항

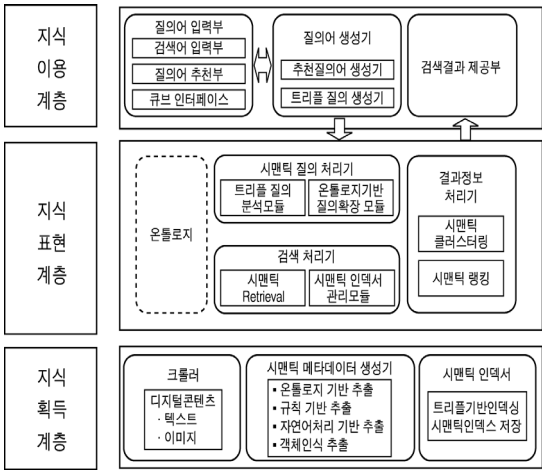
구 분	요구사항
지식획득 계층	• 시맨틱 메타데이터 생성 – 고품질의 의미 있는 메타데이터 생성의 (반)자동화 • 메타데이터 구조화 – 상용화 수준의 시맨틱 인덱서 구체화
지식표현 계층	• 온톨로지를 시맨틱 검색에 적용 – 적절한(Right) 온톨로지의 형태/범위 결정 – 관계(Relation) 중심의 로직 처리 • 멀티미디어 콘텐츠의 지식표현 – 기술 기반(Description-based)과 콘텐츠 기반(Content-based)의 메타데이터 기법 결합
지식이용 계층	• 검색의 니즈(Needs) 파악 용이 – 상호작용성을 고려한 I/F 설계 • 질의 생성의 어려움 해소 – 사용이 용이한 질의 생성 방식 • 복잡한 I/F 해소 – 친근하고 익숙한 I/F 설계

3. 시맨틱 검색 시스템의 개념 모델

본 장에서는 앞서 제시한 요구사항을 반영하는 시맨틱 검색 시스템을 개념적으로 구성하였다. 시맨틱 검색 시스템의 개념적 모델은 시맨틱 웹 최신의 기술과 기능을 포함한 구조이다.

3.1 제안하는 개념적 아키텍처

제안하는 시맨틱 검색 시스템의 개념 모델은 크게 지식획득 계층, 지식표현 계층, 지식이용계층으로 구성된다. [그림 1]은 전체적인 개념 모델의 아키텍처이다.



[그림 1] 시맨틱 검색 시스템의 개념적 아키텍처

[그림 1]의 아키텍처를 개략적으로 살펴보면, 우선 지식획득 계층은 웹 콘텐츠가 대상(Target)이며, “검색하고자 하는 대상이 어디에 있는가?”를 해결하기 위해 콘텐츠를 수집하고, 의미 있는 메타데이터를 추출하고, 트리플 형태로 저장하는 영역이다. 지식표현 계층은 온톨로지가 대상이며, “검색 결과를 어떻게 찾을까?”를 해결하기 위해 지식을 인간과 기계(소프트웨어)가 모두 이해할 수 있도록 온톨로지 형태로 표현한 영역이다. 즉, 온톨로지를 매개로 질의처리 및 결과처리를 수행하는 영역이다. 마지막으로 지식이용 계층은 검색 이용자가 대상이며, “무엇을 찾을까?”를 해결하기

위해 검색 이용자와 검색 시스템간의 상호작용이 활발하게 진행되는 영역이다.

3.2 개념 모델의 계층별 구성

• 지식획득 계층

지식획득 계층에서는 시맨틱 크롤러에서 수집된 콘텐츠(텍스트, 이미지)를 시맨틱 메타데이터 생성기를 통해 메타데이터를 생성하고, 시맨틱 인덱서에 트리플 인덱스 구조로 저장한다. 특히 시맨틱 메타데이터 생성기에서는 텍스트에 대해서는 온톨로지 기반 추출, 규칙 기반 추출, 자연어 처리 기반 추출 등의 기법을 통합하는 시맨틱 텍스트 분석 방식을 활용하여 고품질의 시맨틱 메타데이터를 생성한다.

반면 이미지/동영상의 경우는 객체인식 이미지/동영상처리 기술을 활용하여 하위 수준(Low Level)의 메타데이터(예 : 색상, 질감 등)를 생성은 물론이고 상위 수준(High Level)의 메타데이터(예 : 건물, 바다, 야경 등)도 생성하고, 이미지/동영상이 속한 주변정보(Context)까지 활용하여 기술 기반과 콘텐츠 기반의 메타데이터 생성 기법을 결합한다.

• 지식표현 계층

지식표현 계층에서는 지식획득 계층에서 생성된 메타데이터를 대상으로 온톨로지가 추가/갱신되고, 이러한 온톨로지를 기반으로 시맨틱 질의처리기, 결과정보 처리기가 운영될 수 있다. 검색처리기는 지식획득 계층의 시맨틱 인덱스를 대상으로 검색과 관리를 수행한다. 특히 질의 확장 모듈은 검색 이용자가 입력한 질의문을 토대로 온톨로지를 참조하여 온톨로지에서 사용하는 용어(Term)로 정규화(Normalization)하고 온톨로지의 구조에 따른 확장된 질의어를 생성하여 검색을 수행한다. 예를 들어, 검색 이용자가 질의문을 “스타들이 자주 가는 곳”이라고 입력하면 온톨로지를 참조 분석하여 스타의 상위 개념은 사람이고 스타의 하위 개념으로 연예인, 스포츠맨 등임을 알 수 있다. 또한 “자주 가는”, “들렀다” 등은 “자주 가다”가 대표 속성이고 상위 속성으로 “가다”, 하위 속성으로 “자주 가는 음식점” 이라는 사실을 온톨로지 구조를 통해 파악할 수 있다. 결국 이러한 온톨로지 구조를 기반으로 “연예인/스포츠맨(A, B, C, …)이 자주 가는 음식점/장소(A, B, C, …)”라는 확장된 질의가 생성되어 검색을 수행할 수 있다. 지식표현 계층의 또 다른 기능인 결과 정보 처리기에서는 검색 처리기를 통해 가져온 검색결과를 인터페이스에 표시하기 전에 온톨로지의 가중치 기반 랭킹 또는 온톨로지 개념/속성에 따른 분류 등의 과정을 통해 시맨틱 클러스터링을 수행한다. 결과적으로 지식표현 계층에서는 서비스에 적절한(Right) 온톨로지를 기반으로 다양한 시맨틱 질의 처리기를 통해 관계(Relation) 중심의 로직 처리가 가능하다.

• 지식이용 계층

지식이용 계층은 크게 질의어 입력부(검색어 입력부, 질의어 추천부, 큐브 인터페이스)와 질의어 생성기, 검색결과 제공부로 구성되어 있다. 우선 질의어 입력부에서는 트리플(Tripple) 인터페이스(검색어 입력부, 질의어 추천부, 큐브 인터페이스)로 구성되어 있어 검색 이용자의 정보 니즈를 파악 할 수 있도록 구성하였다. 또한 질의어 생성기에서는 입력된 질의어에 대해 지식표현 계층의 시맨틱 질의처리기와검색 처리기가 처리 할 수 있도록 입력된 질의어를 트리플 질의어로 생성하는 역할을 수행하도록 하여 질의 생성의 어려움을 해소할 수 있다. 검색 결과 제공부에서는 검색 결과의 해석이 용이하도록 속성(Property) 또는 관

계(Relation)를 중심으로 시맨틱 클러스터링 기법을 적용한 결과와 시맨틱 랭킹 기법을 적용하여 우선 순위에 따른 검색 결과를 제공하도록 구성하였다.

4. 시맨틱 검색 시스템의 구현

본 연구를 통해 개념 모델을 반영하여 시맨틱 기술을 활용한 실질적인 시맨틱 검색 시스템이 STARS(Semantic Technology bAsed Retrieval System)이며, 현재 프로토타입 개발을 완료한 상태이다.

STARS는 시맨틱 검색 서비스를 제공하기 위해 3단계(온톨로지 구축, 시맨틱 메타데이터 생성, 브라우징 및 질의)의 구축 과정을 거친다. 온톨로지 구축 단계에서는 Protégé라는 개발도구를 이용하여 온톨로지를 디자인, 인스턴스 생성 등을 통해 초기 온톨로지(Bootstrap Ontology)를 구축한다. 다음은 시맨틱 메타데이터 생성 단계로 온톨로지 매핑 알고리즘과 온톨로지/연관규칙/자연어처리/객체인식 기반 시맨틱 정보 추출 알고리즘을 이용한다. 마지막으로 브라우징 및 질의 단계에서는 트리플 인터페이스(Triple Interface : 검색어 입력 창, 질의어 추천 창, 큐브 인터페이스)를 통해 검색 이용자와의 상호작용 과정으로 검색 이용자의 정보 니즈를 파악하여 검색 결과를 제공하도록 구성하였다.

시맨틱 검색 서비스를 제공하기 위해서는 아래와 같은 3 단계의 과정을 통해 구현이 가능하다. 전체적으로 초기 온톨로지(Bootstrap Ontology)를 구축하고, 다양한 소스의 정보로부터 추가적인 메타데이터를 생성하여 온톨로지를 추가/보완 한 후, 시맨틱 검색 서비스 제공을 수행하는 단계로 요약된다. 단, 본 연구에서는 프로토타입 개발 단계였으므로 검색의 대상, 온톨로지와 메타데이터 범위를 뉴스 기사로 한정하였다. 아래는 주요 기능에 대한 구현 사례를 제시하였다. 또한 아래와 같이 구현한 의미에 대해 설명하고자 한다.

4.1 단계별 구현 방법

4.1.1 1단계 : 초기 온톨로지와 인스턴스 생성

• 온톨로지 스키마 모델링

온톨로지 기반의 응용 서비스 개발은 온톨로지 스키마를 생성하는 것으로부터 시작한다. 온톨로지 스키마는 지식체계의 진보된 형태로, 다양한 클래스(Class), 속성(Attribute), 관계(Relationships)를 정의함으로써 해당 분야 도메인에 대해 공유되어지는 개념을 명시적으로 규정하게 된다.

본 연구의 온톨로지 구축 프로세스는 뉴스 원문기사로부터 의미단위의 용어(Term)를 추출하는 과정으로부터 시작된다. 의미단위는 뉴스 문서에 등장하는 가능한 모든 키워드 (Vocabulary)를 말하며, 명사뿐만 아니라 동사, 형용사를 포함한다. 다음 단계로 추출된 의미 단위 용어 사이에 존재하는 동의어들이 그룹핑(Grouping)되어 대표어가 정의되며, 각각의 의미단위들이 가장 적합한 개념(Concept), 속성(Property) 혹은 인스턴스(Instance)로 온톨로지에 매핑(mapping) 된다. 이때, 개념 및 속성들 사이에 계층관계도 설정된다. 즉, 비슷한 개념의 항목들이 그룹화되어 보다 일반적인 상위 개념으로 묶여지게 된다. 또한, 이미 상하위관계가 존재해도 뉴스 기사로부터 새로운 상하위관계가 드러나는 경우, 하나의 인스턴스나 개념이 두 개 이상의 상위 개념을 가질 수 있도록 온톨로지에 수용한다.

이와 같이 온톨로지를 구축함으로써 시맨틱 정보추출이 한층 용이해 지고, 의미기반의 지능화된

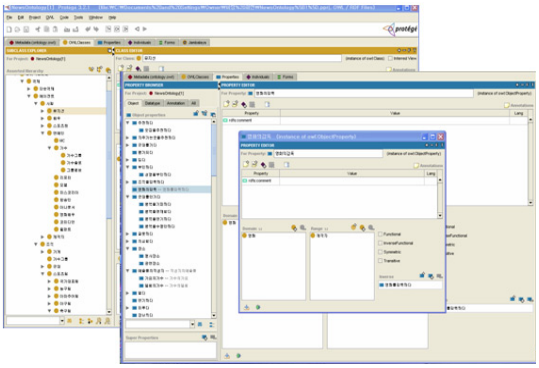
검색이 가능해진다. 서비스 측면에서 보면, 사용자 질의 의도가 반영되면서 실세계의 정보(real-world information)를 제공해줄 수 있는 추천 질의어 생성이 온톨로지를 기반으로 이루어진다. 추론을 통해 질의어를 확장하고, 시맨틱 클러스터링을 통해 효과적으로 검색결과를 제공하는 것도 온톨로지를 중심으로 이루어진다. 특히, 뉴스 기사에 대해 단순 텍스트(키워드 : keyword)가 아닌 문장의 의미정보를 담고 있는 트리플(SPO)형태의 인덱싱이 가능해진다.

온톨로지 모델링 도구로는 freeware인 Protégé를 사용했다. Protégé는 자바기반의 Open Source 통합 온톨로지 구축 프레임워크로 W3C에서 표준 온톨로지 언어로 권고한 OWL 및 RDF(S) 기반 온톨로지 설계에 적합하다. 또한, 다양한 플러그인을 확보하고 있으며, 여러 추론엔진과 연동 가능하다.

[그림 2]는 본 연구에서 온톨로지 저작도구인 프로테제(Protégé)를 이용해 구축한 뉴스 온톨로지의 일부이다. 좌측 클래스탭에 리스트된 내용을 살펴보면, [영화배우], [탤런트], [가수]등이 [연예인]이라는 상위 개념의 하위(sub)로 정의되고, [연예인]은 [스포츠맨], [제작자] 등과 함께 좀더 일반화된 [사람]이라는 개념의 하위 개념(sub-Classes)으로 정의되고 있는 것을 확인할 수 있다. 또한, [영화배우]개념은 [연예인]의 하위 개념이면서, [배우]의 하위개념이기도 하다. 이와 같이 클래스탭(Class Tab)에서는 기본적으로 개념을 상하위 계층에 맞춰 생성/등록 혹은 삭제할 수 있다. 속성탭(Property Tab)에서는 개념/인스턴스간의 관계를 나타내는Object Property 혹은 특정값(value)을 가지는 datatype Property등을 정의할 수 있다. [그림 2]에서 보면, [영화]개념은 <장르 : string>라는 datatype property를 가지면서, [제작자]개념과 <영화를 감독하다>라는 Object property 관계로 정의되어 있음을 알 수 있다. <창업하다>, <출시하다>속성이 <운영하다>라는 상위 속성으로 묶여있는 것과 같이 속성도 계층구조를 가지도록 조직화 할 수 있다.

뉴스 온톨로지 스키마는 뉴스 콘텐츠에 대한 지식 모델(Knowledge model) 뿐만 아니라 뉴스에 포함된 이미지 데이터 및 뉴스 문서관리를 위한 메타데이터(예, 기사분류, 기사제공자, 작성 날짜 등)에 대한 메타베이스 모델(Metabase model)도 포함한다.

이처럼 상향식(Bottom-up) 방식을 통해 구축된 뉴스 온톨로지는 명확성과 객관성(Clarity and Objectivity), 완결성(Completion), 일관성(Coherence), 데이터 밀접성(Distance Minimization), 서비스 유용성 등의 측면에서 폭넓은 경험과 지식을 보유한 도메인 전문가들로부터 평가를 통해 검증받게 된다.



[그림 2] Protégé 도구를 이용한 설계 화면

• 온톨로지 인스턴스(Ontology Population)

뉴스 기사로부터 추출된 의미단위로 구성된 지식베이스(KnowledgeBase)는 수많은 인스턴스를 포함하는 대규모(Large-scale) 온톨로지를 구성한다. 이는 온톨로지를 기반으로 시맨틱(semantic)

정보를 자동추출하고, 지능화된 뉴스 검색서비스를 제공하기 위해 필요한 요소이다. 그러나, 대규모 온톨로지를 기반으로 실제 응용서비스를 개발할 경우 고려해야 할 사항들이 있다. 우선 온톨로지 개발자 관점에서, 대용량 데이터 셋(Sets)에 대한 효과적인 관리가 문제이다. 일부 툴에서 온톨로지 정보를 효과적으로 제공하기 위해 그래프 형태의 뷰(View)를 사용자에게 제공하고 있으나, 실세계 정보를 다루는 응용에서는 개념 및 인스턴스가 매우 많고 개체간 관계가 복잡해 이해하는데 쉽지 않다. 향후 이러한 문제를 해결하기 위해 특정 영역만 부분적으로 브라우징할 수 있는 툴 개발이 고려되어야 할 것이다.

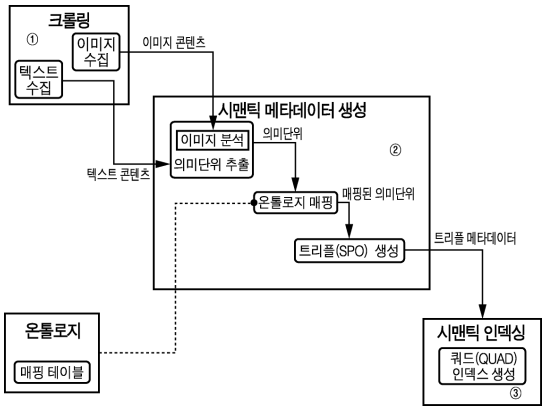
4.1.2 2단계 : 메타데이터 생성 및 인덱싱

본 연구는 콘텐츠의 수집, 의미 정보의 추출 및 구조화를 통해 시맨틱 웹 검색을 가능하게 하며 텍스트와 이미지를 그 대상으로 한다. 여기서는 콘텐츠의 수집을 '시맨틱 크롤링'으로, 의미 정보의 추출을 '시맨틱 메타데이터 생성'으로, 그 구조화를 '시맨틱 인덱싱'으로 정의하였다[그림 3].

먼저 시맨틱 크롤링 (①)은 텍스트와 이미지로 되어 있는 콘텐츠를 수집하는 부분으로 뉴스 콘텐츠를 검색하는데 용이한 SCD(Structured Crawl Document)포맷의 텍스트 콘텐츠를 수집하고, 해당 콘텐츠와 연관된 이미지들을 자동으로 수집한다. SCD는 데이터 원본을 색인하기 위한 특정한 형식을 가진 텍스트 파일이다.

향후 시맨틱 크롤링은 SCD포맷을 포함한 다양한 형태의 콘텐츠를 수집할 수 있도록 모듈화 하여 다양한 포맷(HWP, PDF, DOC, XML, RSS 등)의 텍스트 및 멀티미디어를 수집할 수 있도록 할 계획이다.

시맨틱 메타데이터 생성(②) 과정은 의미단위



[그림 3] 메타데이터 생성 및 저장 과정

의 추출, 온톨로지 매핑, 트리플 생성이라는 세 단계를 통해서 이루어 진다. 첫째, 의미단위의 추출은 앞서 설명한 온톨로지 구축 프로세스상의 의미단위 추출과 동일하다. 즉, 뉴스 원문기사로부터 의미단위의 용어(Term)를 추출한다. 단, 이미지 콘텐츠의 경우 이미지 분석 모듈을 통해 등장 인물과 7가지의 이미지 개념(Architecture, Indoor, Terrain, Night, Snowscape, Sunset, Waterside)과 같은 의미들을 추출해 낸다. 둘째, 온톨로지 매핑은 전 단계에서 추출된 의미단위가 온톨로지내의 어디에 매핑(mapping)되는지 결정하는 단계이다. 각 의미단위들은 가장 적합한 개념(Concept), 속성(Property) 혹은 인스턴스(Instance)로 온톨로지에 매핑 되며 이 과정에서 의미단위가 가지고 있는 중의성은 제거된다. 셋째, 트리플 생성은 온톨로지와 매핑된 각각의 의미단위를 콘텐츠가 가진 정보로 표현하는 과정으로 트리플(SPO-Subject, Predicate, Object)로 표현한다. 이 과정을 통해 해당 콘텐츠는 기존의 태그(Tag) 방식이 아닌 온톨로지와 연결된 완전한 정보를 메타데이터로 갖게 된다.

시맨틱 인덱싱 (③)은 앞서 생성된 메타데이터를 검색의 효율성을 위해 구조화하는 단계로서 트리

플 형태인 메타데이터를 콘텐츠 출처가 포함된 쿼드(QUAD) 형태인SPOC(Subject, Predicate, Object, Context)로 나타내고 가능한 모든 조합을 인덱싱한다. 조인(Join) 없는 단순 질의를 위해 필요한 Access Pattern의 조합의 수는 총 16가지이지만 B+ 트리로 인덱스를 구성할 경우 Prefix 검색이 가능하므로 SPOC, POC, OCS, CSP, CP, OS의 총 6개의 인덱스로 구성할 수 있다. 어떤 인덱스를 이용해 검색을 할지는 질의 유형에 따라 다르므로 질의 유형에 따라 탐색할 인덱스를 결정해야 한다. 예를 들어 (?:P:?:C) 형태의 질의 유형(Predicate과 Context를 질의어로 Subject와 Object를 검색 결과로)에는 CP로 구성된 인덱스를 탐색하면 쉽게 구할 수 있다. [그림 4] 질의 유형과 해당 인덱스를 나열한 것이다. 아래의 인덱스는 다양한 Access Pattern에 대한 최소한의 인덱스이며 검색의 성능 향상을 위해 추가 인덱스를 구성할 수도 있다.

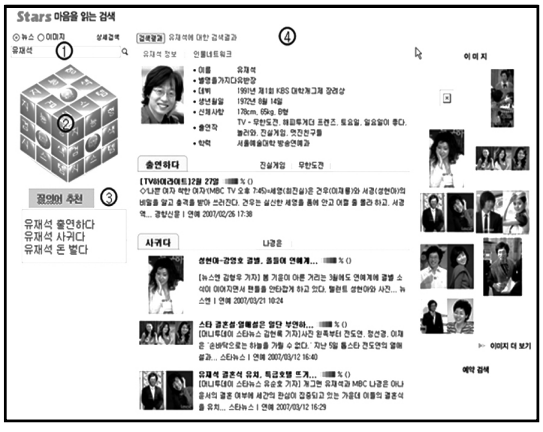
spoc	poc	ocs
(?:?:?:?) (s:?:?:?) (s:p:?:?) (s:p:o:?) (s:p:o:c)	(?:p:?:?) (?:p:o:?) (?:p:o:c)	(?:?:o:?) (?:?:o:c) (s:?:o:c)
csp	**cp**	**os**
(?:?:?:c) (s:?:?:c) (s:p:?:c)	(?:p:?:c)	(s:?:o:?)

[그림 4] 인덱스 구조

4.1.3 3단계 : 시맨틱 브라우징 및 질의

본 연구를 위하여 시맨틱 웹 검색에서의 행위 요인이 고려되고, 검색 니즈를 충족시키며, 기존 검색과의 차이를 극복할 수 있는 인터페이스의 프로토타입을 [그림 5] 같이 구현하였다. [그림 5]와 같이 제안한 시맨틱 웹 검색 인터페이스는 크게 네 부분 즉, 검색어 입력 창, 큐브 인터페이스, 질의어 추천 창, 결과 제공 창으로 구성되어 있다. 다시 입력 부분과 출력 부분으로 구분하면 입력 부분은 검색어 입력 창, 큐브 인터페이스, 질의어 추천 창으로 구성되어 있어서 검색 이용자는 검색어 입력 창 또는 큐브 인터페이스로 질의어를 입력하거나 선택하면, 추천 질의어 창에 입력 또는 선택된 값에 기반한 질의어를 추천 받는다. 아울러 시맨틱 웹 인터페이스는 검색 이용자의 입력 순서나 질의 형태에 따라 정방향(Forward) 방식과 역방향(Backward) 방식으로 나누어서 생각해 볼 수 있다. 정방향 방식은 이용자가 큐브의 면과 셀에 표시된 온톨로지 기반 대표 개념과 속성을 먼저 선택하여 질의어 추천을 받아서 검색을 수행하는 방식이고, 후방향 방식은 입력 창에서 구체적인 값을 입력하고 큐브 인터페이스가 해당 값의 개념과 속성을 큐브 면에 재배치함과 동시에 질의어 추천 창에 의미적으로 연관된 질의를 추천해주는 방식이다.

[그림 5]에서 볼 수 있듯이 시맨틱 브라우징과



[그림 5] 시맨틱 검색 화면

질의는 아래와 같은 개별 기능으로 구성되어 있다.

첫째, 일반 검색어 입력 창은 기존 검색 창과 용도 및 사용 방식이 유사하다. 다만, 일반 검색어 입력 창 (①)에서 특정 키워드를 입력하면 큐브 인터페이스와 질의어 추천 창과 동기화되어 큐브 인터페이스에서는 입력된 특정 키워드에 따른 온톨로지의 개념(Concept)과 관계(Relation)로 연동되고, 질의어 추천 창에는 해당 입력 값에 연동되는 온톨로지의 개념과 관계를 기반으로 추천 질의어를 제시한다.

둘째, 검색 이용자와의 상호작용성을 지원하는 큐브 인터페이스 (②)이다. 큐브의 각 면과 셀에 온톨로지로부터 대표 개념과 관계를 표시하여 검색 이용자가 면과 셀을 선택하면, 내부적으로 온톨로지 매핑(Mapping)을 통한 최적의 질의어를 추천한다. 큐브의 면과 셀에 표시되는 대표 개념과 속성은 검색 대상이 되는 원래 웹 문서의 메타데이터를 기반으로 구성된 온톨로지 구조에 의존적이다. 따라서 큐브의 각각의 면과 셀의 선택은 서로 다른 추천 질의어를 생성할 수 있으며 큐브의 면과 셀을 선택하기 위해 큐브를 회전 또는 선택하는 조작이 요구된다.

셋째, 검색 이용자에게 의미적으로 연관된 질의어를 생성해주는 질의어 추천 창(③)이다. 질의어 추천 창은 검색어 입력 창과 큐브 인터페이스와 동기화되어 있어, 검색어 입력 창 또는 큐브 인터페이스에서 입력 또는 선택할 경우, 해당 온톨로지의 개념과 관계를 기반으로 다양한 질의어를 추천할 수 있는 창이다. 이러한 질의어 추천 창은 이용자에게 질의어를 표시할 뿐만 아니라, 검색 이용자가 추천된 질의어 중 선택할 수 있는 역할을 수행한다.

마지막으로 검색 결과 제공 창(④)이다. 검색 결과 제공 창은 검색어 입력 창, 큐브 인터페이스, 질의어 추천 창에서 입력된 검색어의 결과를 제공하는 화면으로 이용자가 결과 해석에 따른 지루감과 부담감이 없도록 유사 검색 결과의 클러스터링, 최신 결과의 우선 제공 등을 원칙으로 구성하였다. 예를 들어, 검색결과를 검색 대상(Object)의 속성(Property) 또는 타 대상(Object)과의 관계(Relation)를 중심으로 클러스터링 하여 제공할 수 있다.

5. 논의

시맨틱 검색 시스템으로 구현된 STARS는 시맨틱 웹 검색의 3가지 핵심 기술 즉, 시맨틱 메타데이터 생성 영역, 온톨로지 구축 영역, 그리고 브라우징과 질의 영역을 포함하도록 구성되어 있다. 전체적으로 3계층(지식획득 계층, 지식표현 계층, 지식이용 계층)으로 구성되어 있으며 각 계층별로 독특한 특징들을 포함하고 있다.

우선 지식획득 계층은 온톨로지, 연관 규칙, 자연어 처리, 객체 인식 기술을 활용하여 고품질의 의미 있는 메타데이터 (반)자동 추출이 가능하고, 트리플 기반 인덱싱으로 시맨틱 메타데이터의 저장과 관리가 용이하다. 지식표현 계층은 기존검색과 추론검색을 혼합한(Hybrid) 검색을 통해 온톨로지 기반 검색과 온톨로지 가중치 적용으로 연관검색(Associate Search)이 가능하여 구축된 서비스 온톨로지를 참조하여 관계 중심의 질의 확장이 쉽다. 또한 이미지에 대한 객체인식 결과를 포함하는 이미지 온톨로지와 이미지의 주변 텍스트 예를 들어, 태그, 제목 등을 기반으로 구성된 서비스 온톨로지를 모두 활용하여 좀더 정확한 이미지 검색도 가능하다. 마지막으로 지식이용 계층은 검색 이용자의 니즈를 파악할 수 있도록 입력 방식을 기존 검색어 입력 창만 제공하는 방식에서 탈피하여, 검색어 입력 창, 질의어 추천 창, 그리고 큐브

인터페이스의 동기화를 통해 구성하였고, 검색 결과도 의미 기반의 클러스터링 방식으로 제공한다.

특히 시맨틱 인터페이스를 트리플(Triple Interface : 검색어 입력 창, 질의어 추천 창, 큐브 인터페이스)로 구성한 이유는 크게 3가지가 있다.

첫째, 정보 니즈(Information Need) 파악은 단 하나의 질의(Query)로는 명시적으로 표현될 수 없으므로 다중 인터페이스(Multi-interface)를 통해 해결하고자 한다(Belkin, 1980 Oddy, 1977; Wissbrock, 2004).

둘째, 시맨틱 웹이 다양한 질의들을 그래프 형태(Graph Patterns)로 명세화(Formalize) 하기에는 용이하나 검색 이용자에게는 질의를 명세화하기 어려우므로, 복잡한 질의 패턴을 직관적으로 해결하도록 시맨틱 인터페이스 지원이 요구된다(Albertoni et al, 2004 Athanasis et al, 2004 Catarci et al, 2004 Makela et al, 2006).

셋째, 대부분의 경우 검색 이용자는 검색 대상 분야의 전문가도 아니고 찾고자 하는 대상을 정확하게 알 수 없으므로, 검색 이용자의 검색 과정(Searching Process)을 지원해야 한다(Albertoni et al, 2004 Colucci et al, 2006; Makela et al, 2006).

결과적으로 구현된 시맨틱 검색 시스템을 이용하는 검색 이용자는 기존 방식과 동일한 검색어 입력 창을 통해서는 익숙하여 조작이 용이하고 친근한 방식이므로 검색 행위의 비용 요인인 투입 수고와 인지/심리적 요인을 최소화할 수 있다. 또한 다중 인터페이스로 새롭게 추가된 큐브 인터페이스 방식은 입력의 제한을 받는 모바일 환경에서 유용한 인터페이스 역할을 수행할 수 있을 것으로 판단된다(Han et al, 2005). 왜냐하면 큐브 인터페이스는 검색 이용자가 흥미를 유발하여 인지적/심리적 부담을 최소화 할 수 있고, 검색 이용자가 큐브의 면과 셀만을 조작하여 질의어를 생성하므로 검색의 행위 관점에서도 효과적이며 편리하다. 그러므로 검색 이용자가 검색의 니즈를 완성시키기까지 다양한 큐브의 선택과 수정 과정을 거치는데, 이러한 과정은 시스템과 이용자간의 상호작용성을 증진시켜 검색의 니즈 충족에 기여할 수 있다. 아울러 검색 이용자가 온톨로지에 대한 지식이 없다고 하더라도 직관적인 검색 질의를 입력할 수 있다. 마지막으로 질의어 추천 창은 검색 행위 관점에서 검색어 입력의 투입 수고를 최소화하며, 검색하고자 하는 대상에 대한 추천을 통해 검색 이용자에게 인지/심리적 부담을 최소화할 수 있을 것으로 판단된다. 시맨틱 검색이 효과적으로 검색을 수행될 수 있도록 지원할 뿐 만 아니라, 이용자의 검색 니즈를 충족할 수 있을 때까지 상호작용 즉, 추천 질의어를 수정, 삭제, 선택 할 수 있도록 지원한다. 특히 질의어 추천 창은 이용자가 검색어 입력 창과 큐브 인터페이스에서 입력 또는 선택한 한 두 단어와 의미적으로 연관된 온톨로지 기반 질의어를 추천하므로 검색 결과의 재현률을 크게 향상시킬 수 있도록 지원한다. 이러한 과정은 검색 이용자가 자신의 정보 니즈를 명시적으로 표현할 수 없어서 질의 생성의 어려움이 있을 경우 질의 생성의 용이성으로 인해 의미가 크다고 판단된다.

결국 검색 이용자는 질의어 추천 창에서 자신의 질의를 선택하거나 또 다시 검색어 입력 창 또는 큐브 인터페이스를 통해 자신이 원하는 질의어가 나올 때까지 반복하여 질의어를 생성할 수 있다. 출력 부분은 결과 제공 창으로 구성되어 있으며, 제공된 결과가 이용자가 원하는 결과가 아니라면 입력 부분을 재조작하여 또 다른 결과를 제공 받을 수 있다. 이러한 과정을 통해 검색 이용자 관점에서는 검색을 효율적이고 효과적으로 수행할 수 있을 것으로 판단된다.

6. 결론 및 향후 연구방향

현재까지 시맨틱 웹 연구는 상용화 관점에서는 아직 초기 단계이며, 진행되는 연구도 시맨틱 웹의 구성 요소인 에이전트, 추론 엔진, 온톨로지 등에 대한 기술적 검증과 프로토타입 구현 등 기술 중심 연구로 이루어지고 있는 상태이다. 이러한 상황에서 본 연구는 시맨틱 검색 시스템 기술 개발 과제의 일환으로 시맨틱 검색 시스템의 개념적 모형화와 구현을 목표로 "STARS" 프로토타입을 완성하였으며, 상용 서비스 제공을 위한 추가 기능 개발을 진행하고 있다. 본 연구는 웹 검색에서도 시맨틱 웹 특히, 시맨틱 검색 이용자에게 적합한 시맨틱 검색 시스템의 분야별 요구사항을 도출하였다. 다음으로 도출된 요구사항을 반영하는 시맨틱 검색 시스템의 개념적 모델을 3계층 구조 지식 획득, 지식 표현, 지식 이용으로 나눠 포괄적이며 체계적으로 접근하였다. 또한 제안된 개념 모델을 기반으로 시맨틱 검색 시스템 프로토타입 "STARS"를 개발하였다. 즉, 지식획득 관점에서 시맨틱 메타데이터 생성의 자동화를 위해 온톨로지 기반 추출, 규칙 기반 추출, 자연어 처리 기반 추출, 객체인식 추출 등의 기법을 사용하였고, 생성된 메타데이터를 구조화하기 위해 시맨틱 인덱서를 개발하였다. 지식표현 관점에서는 온톨로지의 형태와 범위는 시범 서비스 제공을 목표로 기존 하향식 방식을 탈피하여 상향식으로 구축하여 도메인의 적합성, 서비스의 유용성을 강조하였다. 지식이용 관점에서도 시맨틱 웹 인터페이스 관점에서 일반 웹 검색과의 차이점을 극복할 수 있도록 기존 검색 입력 창을 포함하는 인터페이스의 유용성과 조작의 용이성을 충분히 반영하였다. 더욱이 시맨틱 검색 관점에서 검색의 니즈를 반영할 수 있도록 시맨틱 검색 시스템 설계시 검색 이용자와 검색 시스템간 상호작용성을 트리플 인터페이스(검색어 입력창, 질의어 추천창, 큐브 인터페이스)를 통해 최대한 고려하였다.

그러나 논문에서는 아직까지 시맨틱 검색이 비교적 새로운 개념이고, 시맨틱 검색 시스템 관련 연구도 아직 초기단계에 있으므로, 제안된 개념 모델과 프로토타입 STARS가 기존의 검색 모델이나 시맨틱 검색 시스템과의 비교실험 등을 통한 객관적인 검증까지는 진행하지 못하였다.

향후 개발 완료된 시맨틱 검색 시스템의 효과성에 대한 평가를 통해 구현된 시맨틱 검색 시스템의 객관적 타당성을 확보할 계획이다. 이를 위해 시맨틱 검색 시스템 품질의 개념을 이론적으로 정립하고 검색의 효과성에 영향을 미치는 시맨틱 검색 시스템 품질평가 요인을 밝혀내는데 초점을 두고자 한다. 또한 이를 토대로 제안한 시맨틱 검색 시스템을 계량화하여 좀더 객관성 평가를 추진하고자 한다.

참고문헌

[1] 한동일, 홍일유, "검색 효과성에 영향을 미치는 시맨틱 검색 시스템 평가요인에 관한 실증적 연구", *2007 KMIS International Conference*, 2007, 52~57.

[2] Albertoni, R. et al., Semantic Web and Information Visualization, *1st Italian Semantic Web Workshop*, Ancona, Italy, 2004.

[3] Athanasis, N., et al., "Generation on the fly queries for the semantic web : The CIS-FORTH graphical RQL interface(CRQL)", *Proceedings of the Third International Semantic Web Conference*, (2004), 486~501.

[4] Bangyong, L. et al., "Association Search in Semantic Web : Search + Inference", *WWW 2005 Conference*, Chiba, Japan, 2005.

[5] Belkin, N., "Anomalous states of knowledge as a basis for information retrieval", *Canadian Journal of Information Science*, 5(1980), 133~143.

[6] Belkin, N. J., W. B. Croft, "Information filtering and information retrieval : two sides of the same coin?", *Commun.ACM*, vol.35(1992), 29~38.

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

[8] Bonino, D. et al., "Ontology Driven Semantic Search", *WSEAS Transaction on Information Science and Application*, Issue Vol.1 (2004).

[9] Catarci, T. et al., "An ontology based visual tool for query formulation support", *Proceedings of the 16th Eureopean Conference on Artificial Intelligence*, IOS Press(2004), 308~312.

[10] Chu, H., "Research in image indexing and retrieval as refleted in the literature", *Journal of the American Society for Information Science and Technology*, Vol.52, No.12(2001), 1011~1018.

[11] Clusty, http://clusty.com.

[12] Collarity, http://www.collarity.com.

[13] Colucci S., et al., "A semantic-based fully visual application for matchmaking and query refinement in B2C e-marketplaces", *ICEC '06*(2006), 14~16.

[14] Dill, et, al., "SemTag and SemSeeker : Bootstrapping the Semantic Web via automated semantic annotation", *Proceedings of the 12th International WWW Conference (WWW 2003)*, Budapest, Hungary, May 2003.

[15] Gruber, T., "It Is What It Does : The Pragmatics of Ontology, invited talk at Sharing the Knowledge", *International CIDOC CRM Symposium*, March 26-27, Washington, DC., 2003.

[16] Guha, R., R. McCool,E. Miller, "Semantic Search", WWW 2003 Conference, May 20-24, *ACM Press*, Budapest, Hungary, 2003.

[17] Hakia, http://www.hakia.com.

[18] Han, D., et al., "Fox Service : An Implementation Case of Ontology-based Search Agent in Mobile Environments", *The 7th International Conference on Mobile Data Management*, 2005.

[19] Harth, A., S. Decker, "Optimized Index Structures for Querying RDF from the Web", *Proceedings of the 3rd Latin American Web Congress(LA-WEB '2005)*, (2005), 71~80.

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

[21] Makela, E., et al., "Ontogator-A Semantic View-Based Search Engine Services for Web Applications", *5th International Semantic Web Conference 2006*, ISWC 2006, Athens, GA, USA, 2006.

[22] Morville, P., "Ambient Findability : What We Find Changes Who We become", *O'REILLY*, 2005.

[23] Mudassar Ilyas, Q., et al., "A Conceptual Architecture for Semantic Search

Engine", *Multitopic Conference, 2004. Proceedings of INMIC 2004, 8th International*, Vol.24, No.26(2004), 605~610.

[24] Oddy, R., "Information retrieval through man-machine dialogue", *Journal of Documentations*, Vol.33(1977), 1~14.

[25] OntoWeb, http://www.ontoweb.org.

[26] Passin, T. B., "Explorer's Guide to the semantic web", *Manning pubulications*, Canada(2004), 2~6.

[27] Ponnada, M., N. Sharda, "Model of a Semantic Web Search Engine for Multimedia Content Retrieval", *In the proceeding of 6th IEEE/ACIS International Conference on Computer and Information Science*, Melbourne, July 11~13, 2007.

[28] Richa, C., et al., "A Hybrid Approach for Searching in the Semantic Web", *WWW 2004 Conference*, New York, USA, 2004.

[29] Sheth, A., "From Semantic Search and Integration to Analytics", *Dagstuhl Seminar on Semantic Interoperability and Integration*, September, 19~24, 2004.

[30] Spink, A., et al., "Searching the Web : the public and their queries", *J.Am.Soc. Inf.Sci.Technol*, Vol.52(2001), 226~234.

[31] Sure, Y., V. Iosif, "First Results of a Semantic Web Technologies Evaluation", *DOA '02 Conference*, 2002.

[32] Wissbrock, F., "Information Need Assessment in Information Retrieval ; Beyond Lists and Queries", *27th German Conference on Artificial Intelligence*, KI2004, University of Ulm, Germany, 2004.

Abstract

A Study on the Conceptual Modeling and Implementation of a Semantic Search System

Dong-IL* · Hana Hyeong-In Kwonb** · Hak-Jin Chong***

This paper proposes a design and realization for the semantic search system. The proposed model includes three Architecture Layers of a Semantic Search System ; (they are conceptually named as) the Knowledge Acquisition, the Knowledge Representation and the Knowledge Utilization. Each of these three Layers are designed to interactively work together, so as to maximize the users' information needs. The Knowledge Acquisition Layer includes index and storage of Semantic Metadata from various source of web contents(eg : text, image, multimedia and so on). The Knowledge Representation Layer includes the ontology schema and instance, through the process of semantic search by ontology based query expansion. Finally, the Knowledge Utilization Layer includes the users to search query intuitively, and get its results without the users'knowledge of semantic web language or ontology. So far as the design and the realization of the semantic search site is concerned, the proposedsemantic search system will offer useful implications to the researchers and practitioners so as to improve the research level to the commercial use.

Key Words : Semantic Web, Semantic Search, Ontology

* KT Future Technology Laboratory
** Chung-Ang University Dept. of Business Administration College of Social Sciences
*** KT Future Technology Laboratory