시멘틱 검색시스템 구축을 위한 요구사항 분석 및 설계에 관한 연구*

A Study on Analysis of Requirements and Design of IR System for Semantic-based Information Retrieval

김 용(Yong Kim)**

초 록

웹 정보의 폭발적인 성장과 함께, 단순히 한 두 개의 키워드의 입력에 따른 검색은 너무 많은 검색결과를 가져오게 되기 때문에 전통적인 정보검색기법은 이용자들에게 있어서 만족할 수 없는 결과를 제공하고 있다. 본 연구에서는 정보에 대한 의미를 기반으로 정보검색의 질적 향상을 위한 기술의 개발을 목표로 하고 있다. 이를 위하여 시멘틱 웹 기술에서 요구되는 시멘틱 기반 검색에 대한 최근의 연구동향 및 기술을 분석하여 시멘틱 기반 검색시스템에서 요구사항을 파악하고, 지능형 검색시스템의 아키텍처, 시멘틱 검색 서비스 개발 과정과 핵심기술 등을 살펴보았다. 분석결과와 함께, 시멘틱 기반 정보검색 시스템의 전체적인 아키텍처에 대한 설계 및 요구사항을 제안하였다.

ABSTRACT

With the rapid expansion of web information, conventional information retrieval techniques are becoming inadequate for users and often result in disappointment, because a couple of simple keywords can easily produce information too much. This study aims at the development of Web information retrieval techniques based on semantics to improve the quality of understanding for information. To achieve the goal, this study analyzes technologies and current status of researches on semantic information retrieval. With the results which are requirements, system architecture and indexing method, this study proposes the system architecture of semantic-based information retrieval system.

키워드: 의미기반검색, 시멘틱검색, 색인, 이용자 인터페이스, 메타데이터, 시멘틱 웹
Semantic Information Retrieval, Indexing, User Interface, Metadata, Semantic Web

---

* 이 논문은 2011년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.
** 전북대학교 인문대학 문헌정보학과 조교수(yk9118@jbnu.ac.kr)
논문접수일자 : 2012년 2월 15일 논문심사일자 : 2012년 2월 24일 게재확정일자 : 2012년 3월 11일

1. 서 론

1.1 연구 배경

인터넷과 정보기술의 발전은 웹을 기반으로 하는 정보자원의 생산 및 유통에 있어서 새로운 전기를 제공하고 있다. 일반적으로 웹의 진화과정에서 차세대 웹은 크게 두 가지 측면으로서 인간에게 봉사하는 웹 서비스로서 경제적 관점의 웹 2.0과 웹의 창안자인 팀버너스리가 제안한 기술적 관점의 시멘틱 웹이 있다. 차세대 웹 관점에서 웹 2.0과 시멘틱 웹은 웹의 효율적인 활용을 통한 경제적 효과를 확보할 수 있는 웹과 지속적으로 새로운 기술을 수용하고 이를 발전시키는 동력으로서 기술적 측면에서의 웹을 동시에 추구하여야 하는 상호 보완적 관계에 있다고 할 수 있다. 변화하는 웹 환경에 있어서 특히 주목하여야 할 부분은 데이터의 양적증가라고 할 수 있다. 2011년 웹 공간에 축적된 데이터를 1800조 메가바이트, 즉 1.8제타바이트(ZB)에 이를 것으로 추산된다. 미국 의회도서관이 소장한 전체 장서의 데이터 1500만 메가바이트의 1200배에 이르는 양이라고 할 수 있다(IDC 2011). 이같이 폭발적으로 증가하는 정보환경에서 기존의 TF-IDF(Term Frequency Inverse Document Frequency)로 대변되는 용어의 출현빈도를 기반으로 한 문서와 질의어의 유사도로 정보자원을 검색하는 기술로서는 많은 한계점이 존재한다. 따라서 전통적인 검색기법을 발전시켜 정보의 의미를 파악하여 이용자의 검색필요성에 가장 적합한 정보를 찾아주기 위한 다양한 노력들이 시도되고 있다.

특히 웹과 같은 영역에서 정보검색은 다양한 분야의 정보들이 서로 연결되어 있는 상황에서 빠르고 정확하게 찾아주는 점에 초점을 맞추어 기술 개발이 집중적으로 이루어지고 있다. 이러한 웹을 둘러싼 환경적인 변화와 요구를 반영하여 웹 검색 분야에 있어서 다양한 시도가 나타나고 있으며 대표적으로 구글의 경우에 있어서 웹의 특성을 적극 반영하고 있는 페이지랭크(Page Rank) 기법을 도입하여 이용자 입장에서 높은 검색 효과를 얻는 서비스를 제공하고 있다(Henzinger 2000). 정보검색 기술은 이외에도 중복검색 결과의 제거, 메타 검색, 분산/통합 검색, 전문 검색 등 여러 이슈들에 대하여 활발한 연구가 이루어지고 있다.

다른 측면에서는 텍스트마이닝 기법을 활용한 주제별/장르별 문서 자동분류 및 군집화 기술을 정보검색에 활용하여 이용자 질의 의도에 적합한 검색 결과 제공에 있어서 다양한 연구가 진행되고 있다. 이 같은 정보검색 분야의 새로운 연구흐름은 전통적인 정보검색 기법으로는 폭발적으로 증가하는 정보자원의 효과적인 처리 및 이용자의 정보검색 요구를 적절하게 만족시킬 수 없기 때문이라고 할 수 있다. 따라서 전통적인 정보검색기법과 함께, 변화하는 정보환경에서 효과적으로 정보를 처리하고 이용자 요구를 수용하기 위한 새로운 검색시스템이 요구된다. 한편, Wissbrock(2004)는 이용자들이 검색을 수행하는 이유는 검색 이용자의 불완전한 지식을 완전한 상태로 바꾸는 지식 충족의 욕구 때문이라고 지적하였다. 그러나 웹 상에서 검색서비스를 이용하는 이용자들은 검색서비스를 이용하는데 따른 전문적인 지식의 부족과 편의성만을 추구하기 때문에 단순히 몇 개의 키워드만 입력하여 검색을 수행하는 행태를 보

여주고 있다. 이와 같은 현상은 인지적 구두쇠(cognitive miser)[1] 현상으로 설명할 수 있다. 즉 폭발적으로 증가하는 정보환경에서 원하는 정보를 찾기 위해 부담해야 할 인지적 노력을 최소화하고자 하는 이용자들의 일반적인 행동이라고 할 수 있다. 따라서 검색서비스를 제공하는 제공자의 입장에서는 변화하는 웹 환경에서 효과적으로 이용자의 정보요구를 만족하면서 이용자의 검색행위를 수용하기 위한 새로운 방법이 요구된다.

1.2 연구 필요성 및 범위

새로운 정보환경의 도래에 따른 시멘틱 웹 환경에서 전통적인 정보검색기법은 많은 한계점을 보여주고 있다. 따라서 시멘틱 웹에서의 정보검색을 위한 새로운 검색기법을 적용한 정보검색서비스는 필수적인 서비스분야라고 할 수 있다. 특히, 시멘틱 검색은 시멘틱 웹 기술의 대표적인 응용분야로서 시멘틱 웹 기술인 메타데이터 생성, 온톨로지 구축, 브라우징과 추론 등을 유기적으로 결합한 새로운 응용분야라고 할 수 있다. 그러나 시멘틱 웹이 기존 웹의 진화, 컴퓨터와 인간 간의 상호협력 관계의 증대 측면에서 높은 기대치가 있음에도 불구하고, 현재까지 상용화가 이루어진 시멘틱 웹 응용서비스의 부족, 시멘틱 검색시스템과 관련된 다양한 분야에 대한 깊이 있고 포괄적인 분야에 대한 연구의 미흡 등으로 실용적인 측면에서의 시멘틱 웹 기반 응용 사례가 미흡한 실정이다. 이는 기존의 많은 연구들이 시멘틱 웹에 대한 접근을 시스템과 개발자의 관점에서 접근함으로써 시멘틱 웹을 실질적으로 이용하는 이용자에 대한 이해의 부족과 함께, 시멘틱 웹에 대한 접근에 있어서 다양하고 포괄적인 측면에서 고려가 이루어지지 않고 지엽적이고 부분적으로 이루어졌기 때문이라고 할 수 있다.

그럼에도 불구하고 시멘틱 웹의 중요한 응용분야로서 시멘틱 검색은 시멘틱 기반 검색, 클러스터링 검색, 사회적/개인화 검색 등을 포함하는 차세대 검색기술의 주요 분야라고 할 수 있다. 이와 같은 검색들은 독립적으로 존재하는 기술들이 아니라 상호 밀접한 관계를 가지고 발전한다. 예를 들어, 시멘틱 기반 검색에 있어서 Hakia와 같은 시멘틱 기반 검색시스템은 특정 인물명을 검색 창에 입력하면 특정 인물을 추론하여 해당 키워드가 포함되지 않은 문서라 하더라도 의미적으로 연관된 정보를 검색결과로 제공한다. 또한 텍스트마이닝 기법을 적용한 클러스터링 검색과 사회적/개인화 검색 기법을 적용한 Clusty 시스템은 각각 통계적 분류, 태그 기반 키워드 검색 기술 등을 적용하여 특정 키워드를 입력하면 검색결과화면에 키워드와 통계적으로 유사한 주제를 군집화하여 제공하며, Collarity 시스템은 협업여과방식과 같은 기법을 활용하여 자신이 속한 커뮤니티의 사람들이

1) 1922년 미국 저널리스트인 Walter Lippmann에 의해 도입된 개념으로서 Wyer & Srull에 의하면 사람들은 인지적 구두쇠(cognitive miser)로 자신이 가지고 있는 인지적 자원(cognitive resource: 일종의 지식)을 사용하는데 매우 인색하다고 한다. 이런 이유로 제품을 평가할 때도 사람들은 제품간 차이를 일일이 파악하기보다는 가격과 브랜드와 같이 제품의 품질을 쉽게 파악할 수 있는 단서들을 주로 활용하는 전략을 쓴다. 최소한의 인지적 자원을 사용하여 이루어진 평가가 불완전할 수도 있겠지만 이때에도 소비자는 "하나를 보면 열을 안다"는 말로 자신의 선택을 합리화한다는 개념(Fiske and Taylor 1992).

자주 사용하는 키워드로 검색결과를 제공한다.

그러나 클러스터링 검색과 개인화 검색은 각각 클러스터링 기준의 모호성, 주관적인 개인화 성향을 해결해야 하는 과제를 안고 있으며 시멘틱 검색 방식은 기존 콘텐츠에서 메타데이터를 자동으로 생성 및 관리해야 하는 어려움이 존재한다. 이와 같은 구현의 어려움과 제한점에도 불구하고 정보검색 연구에 있어서 시멘틱 검색에 대한 연구의 중요성은 지속적으로 확대되면서 다양한 연구가 진행되고 있다. 이와 같은 이유는 시멘틱 기술을 활용한 시멘틱 검색시스템은 검색 이용자가 정확한 의미를 몰라도 사람의 말을 이해하는 것처럼 검색 문장의 의미를 파악해 의미에 맞는 대상을 찾아주는 방식이기 때문이다(Liu 2003). 검색에 투입되는 수고와 인지/심리적 부담감을 최소화하면서도 검색 결과는 정확하고, 검색의도와 의미적으로 연관된 정보를 검색해준다. 새로운 검색기법에서 시멘틱 기반 검색은 시멘틱 웹 기술 즉 온톨로지, 추론기술, 메타데이터 생성기술을 검색에 적용한 것으로서 시멘틱 기반 검색은 전통적인 키워드 기반의 통합검색과는 달리 수많은 정보 데이터를 분석하고 이용자 질의에 대하여 의미 정보에 기반한 정확도 높은 검색 결과를 제공할 수 있다는 점에서 앞으로 개인화 검색과 함께 차세대 검색 기술의 주요 연구분야로 자리 잡을 것으로 전망된다.

특히, 변화하는 웹 환경에서 웹 상의 정보자원의 효과적인 처리와 이용자의 정보요구를 만족할 수 있는 체계적이고 깊이 있는 연구가 요구된다. 따라서 본 연구에서는 시멘틱 웹상에서 적용이 가능한 시멘틱 검색시스템의 설계 및 구현을 목적으로 하고 있다.

그러나 시멘틱 검색 시스템의 실질적인 구현을 위해서는 온톨로지, 색인, 이용자 인터페이스, 웹 기반 응용서비스 등과 같은 다양한 분야에 대한 포괄적인 연구가 요구된다. 따라서 본 연구에서는 실질적으로 시멘틱 웹에 적용이 가능할 수 있는 시멘틱 검색시스템의 설계 및 구현을 위한 기본적인 요구사항을 기존의 선행 연구 등에 대한 분석을 통하여 파악하고 이를 기반으로 본 연구에서 제안하고자 하는 시멘틱 검색시스템에서 요구되는 기능별 모듈을 포함하는 시스템의 전체적인 구조를 설계하였다. 또한 온톨로지 기반의 메타데이터 생성 및 색인방법을 통하여 시멘틱 검색에서 요구되는 기능들을 수행할 수 있는 효과적인 색인방법을 제안하며 기존의 시멘틱 검색시스템의 단점으로 지적되고 있는 이용자 인터페이스를 개선할 수 있는 요구사항 및 기능에 대하여 제안하였다. 이와 같은 요구사항 및 개선방안을 기준으로 시멘틱 검색시스템에 대한 프로토타입을 구현하고 향후 연구방향을 검토해보고자 한다.

2. 이론적 배경

2.1 시멘틱 검색 개요

인터넷의 보편화와 방대한 웹 데이터로 인해 더 이상 기존 웹 검색 방식에 의존하지 않고 정보의 의미를 검색할 수 있는 시멘틱 검색에 대한 필요성이 증대되고 있다. 특히 Oddy의 연구(1997)에서는 이용자가 그들의 정보요구를 질의어 형태로 형식화할 때 정보요구가 정확하게 정의될 수 있다고 가정하지만, 많은 경우에 이

러한 가정은 옳지 않았다고 지적하고 있다. 비록 몇몇 사례의 경우에 이용자가 검색하고자 하는 대상을 알고 명시적으로 표현할 수 있으나 대부분의 경우는 그들의 정보요구를 명쾌하게 명시적으로 표현할 수 없다. 이용자는 불완전한 문제를 해결하기 위한 지식이 필요할 때마다 새로운 정보가 필요하며 이 같은 이용자들의 불완전한 지식을 완전하게 하려고 새로운 정보를 검색하게 된다. 이와 같은 이용자의 정보요구를 만족시켜주기 위한 중요한 수단으로서 정보검색시스템은 매우 효과적인 도구라고 할 수 있다. 그러나 Wissbrock(2004)은 정보검색시스템의 중요한 기능이 이용자의 정보요구의 충족임에도 불구하고, 대부분의 정보검색시스템들은 이를 중요하게 간주하지 않고 있다는 것을 지적하고 있다. 이와 같은 문제점은 현재와 같은 대용량의 정보환경에서 더욱 심각한 문제점으로 부각되고 있다.

기존 웹 검색이 텍스트에서 키워드 매칭 방식이라면, 시멘틱 검색에서는 객체검색방식이며 검색 결과도 단순 주소창이 아닌 객체의 개념, 속성, 요소 등을 포함한다(Guha, McCool, and Miller 2003). 이러한 시멘틱 웹 기반 검색은 검색 이용자가 웹 페이지에서 임의의 키워드를 찾을 때보다는 해당 키워드가 의미하는 하나 또는 그 이상의 개념을 찾고자 할 때 유용하다(Bangyong 2005). 바꾸어 말하면 웹 페이지와 연계된 메타데이터가 풍부할 때 시멘틱 검색의 장점이 부각된다.

그러나 이러한 장점에도 불구하고 시멘틱 검색은 검색 이용자로 하여금 찾고자 하는 해당 키워드가 속한 개념형태로 질의를 표현해야 하는 인지적 부담이 있다(Sure 2002; Albertoni, Bertone, and De Martino 2004). 이와 같이 시멘틱 검색은 시멘틱 웹 응용에 있어서 필수적으로 요구되는 분야라고 할 수 있다. 한편, 시멘틱 검색과 관련하여 Makela, Hyvonen, and Saarela (2006), Bonino et al.(2004), Albertoni, Bertone, and De Martino(2004)에 의해 시멘틱 검색 결과의 유용성, 정확성, 재현성, 검색 시간, 정보 시각화, 정보 요구 등의 관점에서 시멘틱 웹 검색 시스템에 대한 다양한 연구가 진행되어 왔다. 따라서 시멘틱 웹 환경에서 검색 이용자의 정보요구를 반영할 수 있는 시멘틱 검색 연구는 현재와 미래의 웹 환경에서 효과적인 웹 정보서비스를 위해서는 매우 중요한 연구분야라고 할 수 있다.

2.2 관련 연구

웹 2.0, 웹 3.0 시대의 도래와 함께, 웹을 둘러싼 환경적 변화는 이전과는 매우 다르게 진화하고 있다. 특히 전통적인 정보제공자 중심의 웹 환경에서 이용자의 적극적인 참여와 양방향적 상호작용으로의 진화는 웹이 단순한 정보의 저장소가 아닌 사회적 연결망과 집단지성과 같은 정보의 생산과 소비의 공동체의 중심에 있게 하였다. 이와 같은 환경적 변화에 따라 이용자 중심의 웹 검색서비스는 필수적인 웹 응용서비스라고 할 수 있다. 새로운 웹 환경의 도래에 따른 전통적인 검색방식의 제한점을 해결하기 위하여 시멘틱 검색 개념이 출현하게 되었다. 오늘날과 같이 웹의 활성화가 이루어지기 전에는 시멘틱 검색은 주로 자연어처리 기술을 기반으로 텍스트로부터 적절히 색인어를 추출하여 이를 통계적인 검색모델에 반영하는 연구가 대부분이라고 할 수 있다. 특히 90년대 중반에 와서는

TREC-5부터 NLP(Natural Language Processing), SIT(Special Interest Track)이 만들어지면서 본격적으로 자연어처리기술을 이용한 정보검색의 효과 향상을 위한 다양한 연구들이 진행되었다(장명길 외 2001). 그러나 웹을 둘러싼 정보기술이 발전하고 웹 상에 존재하는 정보가 폭발적으로 성장함에 따라 시멘틱 검색은 크게 웹 기반 검색기술과 온톨로지 기반 검색기술로 나뉘어 진행되고 있다.

먼저 웹 기반 검색기술은 크게 검색 서비스의 인덱싱 작업의 기반이 되는 페이지 평가 기술(Fujimura 2005), 이용자의 단순한 키워드 중심의 애매한 질의를 정량화하여 평가하는 기술(Qiu 2007) 등이 있다. 페이지 평가 기술은 이용자의 질의에 대하여 웹의 어떠한 페이지가 얼마나 신뢰성이 있는 좋은 페이지인가를 평가하는 것으로서 구글 검색엔진의 대표적인 기법인 PageRank 등이 대표적이라고 할 수 있다.

한편, 기존의 웹 기반 검색에 있어서 가장 큰 제한점은 웹 상에 존재하는 정보의 비구조적 속성이라고 할 수 있다. 특히, 웹에서 정보를 표현하는 표준언어인 HTML은 정보의 외적표현에 중심을 두고 있기 때문에 정보가 가지는 내용 및 구조에 대한 표현이 어렵다. 이와 같은 문제점을 해결하기 위한 방법으로서 XML과 같은 페이지 언어와 온톨로지라는 구조적 데이터 시스템이 제안 되었다. 이와 관련된 대표적인 검색시스템으로서 독일의 Sara Cohen의 XSEarch는 기존의 시멘틱 검색과는 다른 관점에서 XML을 기반으로 문서들을 검색한다. 즉, XML로 작성되어있는 문서와 문서의 계층적인 태그와 키워드를 색인하여 만든 목차 데이터베이스를 사용하여 질의문에 해당하는 문서의 유사도를 계산하여 순위가 높은 문서를 반환하는 방법을 사용한다(하상범, 박영택 2004).

온톨로지는 기존의 웹 데이터에 의미적 요소를 추가하기 위해 제안된 특화된 지식 데이터베이스 시스템으로 이것을 활용하면 좀 더 구조화된 데이터의 형태를 이룰 수 있을 뿐만 아니라 이를 기반으로 추론 또한 가능해지므로 시멘틱 웹의 핵심기술이라 할 수 있겠다(김정훈 외 2008). 한편, 시멘틱 검색에 있어서 가장 중요한 요소는 체계화된 온톨로지의 구축에 있다. 이와 관련하여 김학래와 김홍기(2007)는 개인의 전자문서를 효과적으로 관리하고 검색하기 위하여 전자문서의 메타데이터를 온톨로지 기반으로 생성하고 추론엔진을 이용하여 시멘틱 검색 기능을 제공하는 ONTALK 시스템을 제안하였다. 온톨로지를 구축하기 위한 표현 기술 연구는 크게 웹의 표준을 담당하고 있는 W3C를 중심으로 한 RDF(Resource Description Framework)와 ISO를 중심으로 하는 토픽맵(Topic Maps)기술로 나눌 수 있다.

토픽맵 기반의 다른 시멘틱 검색의 접근방법으로서 권창호(2009)는 기록물의 기술정보 메타데이타를 중심으로 정보자원을 구조화하여 이용자 질의의 접근점을 확장하고, 의미 있는 매칭을 통해 지식자원화된 검색결과값을 제공하기 위해 토픽맵 기반의 기록정보 검색시스템 구축을 시도하였다. 최근에는 사물의 형태, 색감, 질감 등의 사물의 직접적인 특징을 통하여 영상물에 대한 검색을 수행하는 단계에서 발전하여 사물에 대한 감성과 의미를 추상화된 시멘틱으로 표현하여 인간의 지각 및 감성에 기반한 감성기반 시멘틱에 대한 연구가 진행되고 있다(이준환, 박은종 2010).

3. 시멘틱 검색 핵심 분야

시멘틱 웹의 용용분야로서 시멘틱 검색은 다양한 핵심요소 및 요구사항을 필요로 한다. 시멘틱 검색은 전통적인 검색기법과는 달리 정보로서 객체에 대한 표현과 획득에 따른 관계성의 생성이 중요하다. 이와 같은 관점에서 시멘틱 검색은 세 가지 핵심기술요소로서 다양한 정보원으로부터 획득된 텍스트, 이미지, 멀티미디어 등과 같은 정보로부터 시멘틱 메타데이터를 생성하는 정보획득 분야, 획득된 정보로부터 온톨로지의 스키마와 인스턴스를 구축하고 구축되어진 온톨로지를 기반으로 질의확장 및 수정 등의 과정을 통하여 시멘틱 검색을 수행하는 영역으로서 온톨로지 구축을 통한 정보표현 분야, 그리고 이용자 관점에서 시멘틱 검색시스템과의 이용자와의 실질적인 상호소통의 게이트웨이 역할을 수행하는 이용자 인터페이스를 통한 정보이용분야로 구분할 수 있다. 이와 같은 관점에서 본 장에서는 시멘틱 웹 검색의 핵심요소로서 정보의 효과적인 획득, 온톨로지를 기반으로 습득된 정보의 관계성 등을 표현하는 정보표현 및 정보이용의 세 가지 분야로 구분하여 살펴보았다. 이를 기반으로 시멘틱웹 관련 문헌연구를 토대로 시멘틱 검색시스템의 설계 및 구현을 위한 요구사항을 도출하였다.

3.1 정보획득

정보획득은 다양한 정보원으로부터 획득된 텍스트, 이미지, 멀티미디어 등과 같은 정보로부터 시멘틱 메타데이터를 생성하는 분야라고 할 수 있다. 정보의 획득 관점에서 가장 기본적인 방법은 자연어처리, 통계적 기법, 기계학습 기법 등이 있다. 이러한 기술들은 다양한 정보로부터 의미 있는 메타데이터를 추출하기 위해 객체인식기법과 용어의 의미적 모호성을 해결하는 과정이라고 할 수 있다. 특히, 이와 같은 다양한 과정을 통하여 획득된 정보로 부터 시멘틱 검색 대상이 되는 메타데이터 생성이 현실적으로 불가능하다. 현실적으로는 대규모 메타데이터 생성과 시멘틱 정보의 추가 등이 가능하다 (Dill et al. 2003).

그러나 시멘틱 검색을 위한 메타데이터 생성은 매우 중요하면서 필수적인 과정이라고 할 수 있다. 이같은 과정에 있어서 Sheth(2004)는 새로이 생성되는 메타데이터의 양적 그리고 질적 측면에 있어서 서로 상충적인 관계가 있다고 지적하고 있다. 예를 들어, 데이터 생성 날짜, 문서 크기, 작성자 등과 같은 단순한 형태의 메타데이터 추출을 통하여 대규모 메타데이터 생성이 수행될 수 있으나, 회사, 본부, 산업, 비즈니스 등과 같은 획득된 정보의 내용과 주제에 대한 보다 구체적이고 의미적인 정보를 제공할 수 있는 메타데이터를 자동으로 추출하기 위해서는 보다 정밀하고 복잡한 기법들이 요구된다. 이와 같은 문제점은 단순히 전통적인 용어 출현빈도 등에 기반한 통계적 기법 또는 기계학습기법만을 통하여 해결될 수는 없다. 또한 다수의 생성된 메타데이터를 저장하고 관리할 수 있는 색인방법이 요구되고 있다. 바꾸어 말하면 하나의 정보객체를 표현하기 위하여 RDF와 RDF 스키마(RDF/RDFS)로 적용되며 이를 통하여 해당 객체를 표현하는 있어서 정보의 양이 증가하고 있고, 이와 같은 정보객체를 효율적으로 저장 및 검색할 수 있는 색인구조의 필요성이 높아지게

된다(Harth and Decker 2005). 특히 대용량의 데이터 환경에서는 검색의 효율성(efficiency)이 매우 중요한 고려요소가 되고 있으며 이를 해결하기 위하여 복잡하면서 대용량 데이터를 처리하기 위한 다양한 연구들이 진행되고 있다.

3.2 정보표현

정보표현(information representation) 분야는 온톨로지의 스키마와 인스턴스를 구축하고 구축되어진 온톨로지를 기반으로 질의확장 및 수정 등의 과정을 통하여 시멘틱 검색을 수행하는 영역이다. 온톨로지는 일정의 공유된 개념으로 컴퓨터와 인간이 동시에 이해할 수 있는 정보의 표현 형태이다. 현재의 웹은 링크를 통한 정보 제공으로 관련된 정보를 쉽게 찾을 수 있다는 장점이 있으나 수집 및 검색된 정보는 사람에 의해서 해석되고 정제되어야 한다는 점에서 단순한 정보 저장소 이상의 기능을 제공하지는 못하고 있다. 따라서 기존의 웹의 한계점을 극복하기 위해서 W3C에서는 표준으로 규정된 시맨틱 웹 온톨로지 언어(RDF, RDFS, OWL 등) 및 관련 기술에 대한 연구가 활발히 진행되고 있다. 특히 W3C가 OWL을 표준 온톨로지 언어로 지정함에 따라 대부분의 온톨로지 데이터들은 OWL로 기술될 전망이다(허선영, 김은경 2008).

온톨로지를 시멘틱 검색에 적용하기 위해서는 세 가지의 고려사항이 존재한다(Sheth 2004). 먼저 대부분의 실질적인 온톨로지는 반형식적(semi-formal)으로 기술되어야 한다. 이와 같이 온톨로지를 기술하는데 있어서 반구조화된 기준을 적용하는 것은 온톨로지가 정보를 표현하는데 있어서 부분적으로 불일치성, 제약조건 위반 등의 형태와 같이 불완전하게 구축되기 때문이다. 다양한 정보원으로부터 정보를 추출하고 통합하며, 통합된 정보를 기반으로 다양한 개발자에 온톨로지가 구축되므로 온톨로지 구축에 있어서 반형식적 표현은 불가피하다(Gruber 2003). 또한 구축될 온톨로지의 정보기술이 지나치게 세부적이고 구체적이면 실질적으로 응용에 있어서 수준 높은 가치를 제공하지 못한다. 이와 같은 현상의 주요 원인은 온톨로지에 표현된 정보를 포착하기가 매우 어렵기 때문이다. 따라서 추론과정에 따른 추론기능이 현격히 떨어질 수가 있는 제한점이 존재할 수 있다. 현실적으로 표현력을 최소화한 온톨로지와 연산에 따른 복잡한 온톨로지는 상호간의 상충적 관계에 있다. 표현력을 최소화한 온톨로지는 추가 및 수정에 지속적인 사용이 가능하며, 시멘틱 추론에 따른 연산상의 복잡성이 존재하지 않음으로 인하여 추론기능의 제약점이 약화될 수 있다. 그러나 상세화 및 정확성을 요구하는 시멘틱 웹 응용분야와 같은 경우에 있어서는 정보표현에 있어서 보다 구체적이고 세부적인 온톨로지가 적합한 경우도 존재할 수 있다. 마지막으로 시멘틱 웹 응용 분야가 정보의 획득 및 분석을 요구하는 응용분야인 검색 및 개인화 영역인 경우에 있어서 도메인과 응용업무에 보다 세부적으로 특화된 시멘틱 메타데이터가 요구된다.

3.3 정보이용

정보의 이용분야는 이용자 관점에서 시멘틱 검색시스템과의 이용자 인터페이스 영역이라고 할 수 있다. 구축된 온톨로지를 기반으로 시멘

틱 검색시스템에서 제공되는 추론기능, 연관어 검색 등의 다양한 기능을 제공하며 특히, 텍스트, 이미지, 동영상 등의 이종 정보를 검색하기 위하여 온톨로지와 메타데이터를 동시에 처리하는 시멘틱 질의처리는 매우 높은 수준의 가치를 이용자에게 제공할 수 있을 것이다. 현재까지 시멘틱 웹 검색에 대한 다양한 연구들이 수행되었으나 실질적인 상용화는 미흡하며 특히, 이용자 인터페이스에 대한 연구는 연구자의 실험결과를 확인해 보는 실험실 수준에 머물러 있어서 시멘틱 웹의 상용화에 있어서 제한점이 되고 있다. 특히, 기존의 시멘틱 검색 연구에서 제안하고 있는 이용자 인터페이스는 기존 검색시스템에서 제공하고 있는 이용자 인터페이스와 너무 상이하게 구성되어 있으며 검색방식에 있어서도 너무 복잡하여 익숙하지 못한 이용자로부터의 거부감을 초래할 수 있는 문제점이 있다. 예를 들어, 다양한 연구에서 제안하고 있는 시멘틱 검색시스템은 OWL 검색을 지원하며 SPO(Subject, Predicate, Object)형태의 질의문을 입력 받거나, RDQL(RDF Data Query Language) 질의 형태의 질의어 입력형식을 제공하고 검색을 수행한다. 또는 기존의 검색시스템의 결과를 개선시키는데 키워드 중심이 아닌 개념과 관계를 중심으로 구성하고 있다.

대표적인 온톨로지 기반의 시멘틱 검색시스템인 OntoWeb 경우는 시멘틱 웹 검색 인터페이스의 대표적인 사례로 브라우징 → 속성 선택 → 실제 값 입력 → 검색 수행의 순으로 단계별 검색을 수행한다. 이와 같은 문제점에 대하여 Albertoni 등(2004)의 연구에서는 시멘틱 검색을 웹 자원과 현실 세계의 객체들에 관한 의미를 명시적으로 표현하며, 검색 결과의 정확율과 재현률을 개선시키고 검색 이용자의 정보요구를 파악하고 이를 충족하려는 시도를 하고 있다. 그러나, 정보검색과정에 있어서 영향을 미치는 이용자의 행위요인으로 제한된 이용자의 지식, 정보 제공자와 정보 검색 이용자 간의 인식 차이, 이용자가 검색요구를 충족하지 못할 수 있다는 세 가지 우려를 언급하였다. 이러한 다양한 요인을 기반으로 기존의 검색시스템을 이용자와 시스템간 상호작용과 시각화를 강조하였다. 또한 기존 시멘틱 검색시스템들은 온톨로지 구조를 조회/수정할 수 있을 뿐 보다 지능화된 브라우징이 부족함을 언급하였다. 이와 같은 언급은 검색시스템과의 상호작용성을 중요한 인터페이스 요인으로 간주하였다. 또한 Makela 등(2006)은 시멘틱 검색과 이용자 간의 질의생성과 전체적인 이용자 인터페이스에 대한 구성의 관점에서의 연구를 수행하였다. 시멘틱 기술은 개념과 관계를 이용하여 지식을 그래프와 같은 수단을 통하여 시각화하여 표현하기에 적합하지만 검색 이용자가 이러한 지식을 검색하려는 질의생성이 어려움을 언급하면서 시멘틱 검색을 위한 사용의 용이성을 강조하였다. 현재까지 위의 연구에서와 같이 웹 상에서의 시멘틱 검색 인터페이스는 복잡하고, 검색 과정을 인식해야 하며, 기존 일반 웹 검색 인터페이스와 상당한 차이점이 존재하여 이용에 따른 적응 및 학습에 많은 시간이 요구될 수 있다.

4. 시멘틱 검색시스템 설계

현재까지 수행된 시멘틱 검색과 관련된 다양한 연구들을 통하여 많은 시멘틱 검색시스템들

이 제안되었다. 시멘틱 검색시스템은 시멘틱 웹에서 요구되는 주요한 세 가지 핵심요소로서 시멘틱 메타데이터 생성영역, 온톨로지 구축영역, 그리고 브라우징과 질의영역을 포함하고 있어야 한다. 따라서 본 연구에서 제안하고 있는 시멘틱 검색시스템의 기본구조는 크게 콘텐츠 계층, 시멘틱 계층, 서비스 계층으로 구성되어 있으며 각 계층별로 시멘틱 검색서비스 제공을 위하여 요구되는 기본적인 요구사항과 특징들을 포함하고 있다.

본 연구에서 제안하고 있는 시스템의 기본구조를 이루고 있는 각 계층은 시멘틱 검색의 세 부분야인 정보획득, 정보표현 및 정보이용 분야와 상호 연결하여 설명할 수 있다. 먼저 콘텐츠 계층은 정보획득, 시멘틱 계층은 정보표현, 그리고 서비스계층은 정보이용으로 구분할 수 있으며 각각의 계층들은 해당 기능을 수행하기 위하여 하위모듈로 구성되어 있다. 이와 같은 트리플 구조는 실질적인 시멘틱 검색기능을 제공하기 위하여 요구되는 요구사항을 만족하면서 기존의 시멘틱 검색시스템의 단점들을 보완할 수 있도록 설계되어졌다.

특히, 콘텐츠 계층에서는 획득된 정보에 대한 시멘틱 메타데이터의 자동 및 반자동 추출을 통해 텍스트 정보뿐만이 아니라 내용기반 검색을 통한 이미지검색이 가능하도록 설계되었다. 시멘틱 계층은 기존의 검색기능과 추론검색 기능을 혼합한 하이브리드 방식의 검색을 통하여 온톨로지 기반의 검색과 온톨로지에 대한 가중치를 적용하여 연관검색이 가능하다. 또한 서비스 계층은 이용자의 정보요구를 파악할 수 있도록 입력 방식을 기존 질의어 입력 인터페이스만을 제공하는 방식에서 탈피하여, 질의어 입력 인터페이스, 질의어 추천 인터페이스와의 동기화를 통해 구성하였다.

한편 이용자가 직접 입력하는 질의어, 추론, 질의어 추천 등의 과정을 통하여 검색된 검색결과는 상호관련성 있는 결과들끼리 클러스터링 방식으로 제공한다. 이를 위하여 제안하고 있는 시스템은 시멘틱 검색 기능을 제공하기 위하여 온톨로지 구축, 시멘틱 메타데이터 생성, 브라우징 및 질의 등의 과정을 거친다. 온톨로지 구축 단계에서는 Protégé라는 개발도구를 이용하여 온톨로지에 대한 설계와 인스턴스 생성 등의 기능을 수행한다. 다음으로 시멘틱 메타데이터 생성단계에 있어서 온톨로지 매핑 알고리즘과 온톨로지 기반 시멘틱 정보추출 알고리즘을 이용한다. 마지막으로 브라우징 및 질의 단계에서는 질의어 입력 인터페이스, 질의어 추천 인터페이스, 검색결과 인터페이스를 통해 시스템과 이용자와의 상호작용 과정으로 이용자의 보다 정확한 정보요구를 파악하여 검색 결과를 제공하도록 구성되어 있다.

4.1 요구사항 분석

현재까지 시멘틱 검색과 관련된 많은 연구 분석을 통하여 시멘틱 검색시스템, 특히, 시멘틱 웹에 적용될 수 있는 검색시스템에 대한 요구사항을 정리해 보면 다음과 같다. 먼저 시멘틱 웹 검색의 관점에서 이용자의 검색에 대한 요구 및 필요성을 반영하기 위한 검색시스템과 이용자의 상호작용성, 검색결과의 최신성, 재현성을 충분히 고려해야 한다. 예를 들어 이용자가 그들의 정보요구를 질의어의 형태로 형식화하는 과정에서 이용자 자신들의 정보요구를 정확하게

〈표 1〉 시멘틱 검색시스템 요구사항

기본 요구사항	상세 요구사항
이용자의 검색에 따른 정보요구를 반영	시스템과 이용자 간의 상호작용성, 검색된 결과의 최신성, 재현성
적절한 온톨로지의 형태/범위 결정	응용분야에 따른 시멘틱 메타데이터 수용 수준
기존 검색시스템과의 상이점 극복	이용자 인터페이스의 친근성과 유용성, 조작의 용이성

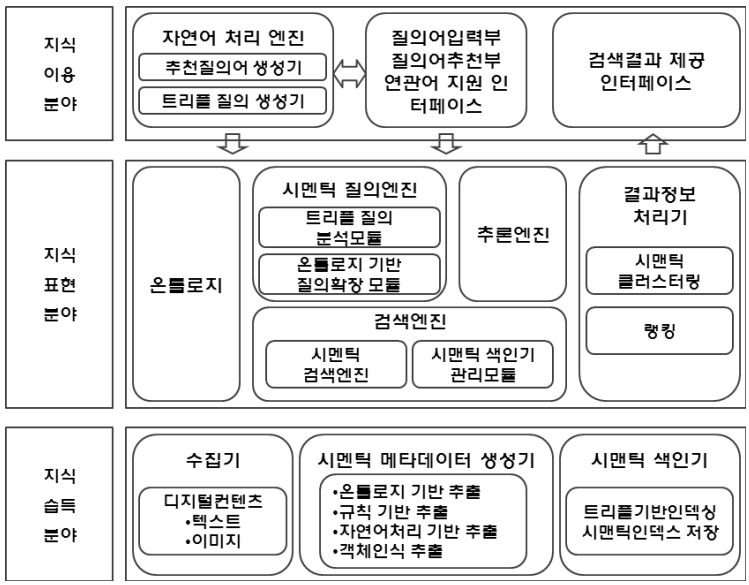
명시적으로 표현할 수 없으므로 검색시스템과 이용자 간의 충분한 상호작용을 지원할 수 있어야 하며 이용자의 적합성 판단 등과 같은 질의어에 대한 정제과정이 유기적으로 지원될 수 있도록 시스템을 설계하여야 한다. 이러한 상호작용과정은 이용자의 불완전한 지식을 완전한 상태로 유지할 수 있도록 지원한다. 또한 검색결과의 제공에 있어서는 검색 대상과 관련된 모든 정보를 제공할 수 있는 재현성에 대한 고려와 함께, 최신성 있는 정보에 대한 우선순위화 등의 기능이 제공되어야 한다.

둘째, 온톨로지의 형태와 범위는 응용분야에 따라 다양하게 구축될 수 있다. 복잡하지 않고 단순한 온톨로지를 요구하는 응용분야에 있어서는 온톨로지를 상대적으로 단순하게 구성한다. 그러나 시멘틱 검색 분야와 같이 복잡한 응용 분야에서는 온톨로지의 개념뿐만 아니라 관계를 처리해야 할 경우가 있으므로 적절한 규모와 범위를 설정하여 온톨로지를 구축하여야 한다. 또한 메타데이터 생성에 있어서도 단순 메타데이터보다는 의미를 포함하는 시멘틱 메타데이터 생성이 요구되는데, 이를 위해서는 자연어 처리 기술, 통계 기법, 기계 학습 기법 등과 같은 다양한 처리기법들을 통합적으로 사용할 수 있어야 한다.

마지막으로 이용자 인터페이스 관점에서 기존의 웹 검색과의 차이점을 극복할 수 있는 이용자 인터페이스의 친근성, 유용성, 조작의 용이성 등을 고려해야 한다. 기존 웹 검색과 너무 상이한 질의어 입력 인터페이스로 구성되거나, 검색방식이 너무 복잡하여 익숙하지 못한 이용자로부터의 거부감을 초래하거나, 조작의 어려움으로 인하여 사용방법의 습득에 많은 시간이 요구되는 이용자 인터페이스의 설계와 구성을 피하여야 한다. 이와 같은 관점에서 중요한 것은 이용자 인터페이스의 유용성을 확보할 수 있는 설계와 구성이 요구된다. 〈표 1〉에서는 위에서 언급하고 있는 시멘틱 검색을 위한 시스템 설계 및 구축에 따른 요구사항을 보여주고 있다.

4.2 시스템 아키텍처

본 연구에서 제안하고 있는 시멘틱 검색시스템은 위에서 설명하고 있는 바와 같이 크게 콘텐츠 계층, 시멘틱 계층, 서비스 계층으로 구성된다. 〈그림 1〉은 본 연구에서 제안하고 있는 시스템의 전체적인 구조를 보여주는 구성도이다. 우선 콘텐츠 계층은 시멘틱 정보수집기, 시멘틱 메타데이터 생성기, 시멘틱 색인기로 구성되어 있다. 시멘틱 정보수집기를 통하여 수집된 텍스트 및 이미지 정보는 시멘틱 메타데이터 생성기를 통해 메타데이터가 생성되고, 시멘틱 색인기에 의해 트리플 색인 구조로 저장된다. 특히 시멘틱 메타데이터 생성기에서는 텍스트 정



〈그림 1〉 제안시스템 구조

보에 대해서는 온톨로지 기반 추출, 규칙 기반 추출, 자연어 처리 기반 추출 등의 기법을 통합하는 텍스트 분석 방식을 활용하여 메타데이터를 생성한다. 이미지의 경우는 객체인식 이미지 처리 기술을 활용하여 색상, 질감 등과 같은 하위 수준의 메타데이터를 생성하거나 정보에 대한 내용 또는 주제 등과 같은 건물, 바다, 야경 등과 같은 상위 수준의 메타데이터를 생성하고, 이미지가 속한 주제 및 내용을 활용하여 최종 메타데이터를 생성한다.

다음은 시멘틱 계층으로서 해당 영역은 시멘틱 검색의 핵심이라고 할 수 있다. 콘텐츠 계층에서 생성된 콘텐츠를 대상으로 온톨로지가 추가 또는 갱신되고, 이러한 온톨로지를 기반으로 시멘틱 질의, 추론, 결과정보에 대한 처리가 가능하다. 제안하고 있는 검색시스템은 콘텐츠 계층의 시멘틱 색인을 기반으로 실질적인 검색과 관리가 수행된다. 특히 질의확장모듈은 이용자가 입력한 질의어를 기반으로 구축된 온톨로지를 참조하여 온톨로지에서 출현하는 용어로의 정규화과정을 통하여 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. 또한 결과 정보 처리기에서는 검색시스템을 통하여 가져온 검색결과를 이용자 인터페이스에 표시하기 전에 온톨로지의

가중치 기반 랭킹 또는 온톨로지 개념/속성에 따른 분류 등의 과정을 수행한다. 이와 같은 과정으로 통하여 검색된 결과중에서 연관성이 높은 결과물들은 군집화를 통하여 이용자에게 제공됨으로써 이용자에게 검색결과에 대한 평가와 검토에 높은 편의성을 제공할 수 있다.

마지막으로 서비스 계층은 이용자와 검색시스템간의 상호작용이 활발하게 진행되는 이용자 인터페이스 영역이라고 할 수 있다. 이와 같은 이용자 인터페이스 영역은 검색 입력부, 질의어 추천부를 포함하는 질의입력부와 자연어 처리시스템, 검색결과 제공 인터페이스로 구성되어 있다. 우선 질의입력부에서는 트리플 구조의 인터페이스로 구성되어 있으며 이와 같은 인터페이스를 통하여 이용자의 직접적인 질의어를 입력 받는 질의어입력창과 함께, 입력된 질의어에 대한 분석을 기반으로 제안시스템에서 제공하는 질의어 추천 인터페이스로 구성되어 있다. 이와 같은 기능별 인터페이스를 통하여 이용자의 정보요구 및 지원이 가능할 수 있도록 구성하였다. 또한 자연어 처리시스템에서는 입력된 질의문에 대해 시멘틱 계층의 검색시스템이 처리할 수 있도록 자연어를 트리플 질의어로 생성하는 역할을 수행한다.

4.3 온톨로지 구축 및 메타데이터 생성

제안된 시멘틱 검색시스템을 통하여 실질적인 검색 서비스를 제공하기 위해서는 크게 세단계의 과정을 거치게 된다. 먼저 검색서비스 제공을 위한 기반요소로서 초기 온톨로지를 구축하여야 한다. 이와 같은 과정을 거치게 되면 추가적인 온톨로지와 기존의 온톨로지에 대한 보완을 위하여 다양한 정보원으로부터 획득한 정보를 통하여 새로운 메타데이터를 생성하게 된다. 이와 같은 선처리 과정이 수행됨으로써 실질적인 시멘틱 검색 서비스 제공을 수행하게 된다. 본 연구에서는 서비스 제공의 대상을 뉴스기사로 한정하여 연구를 수행하였다.

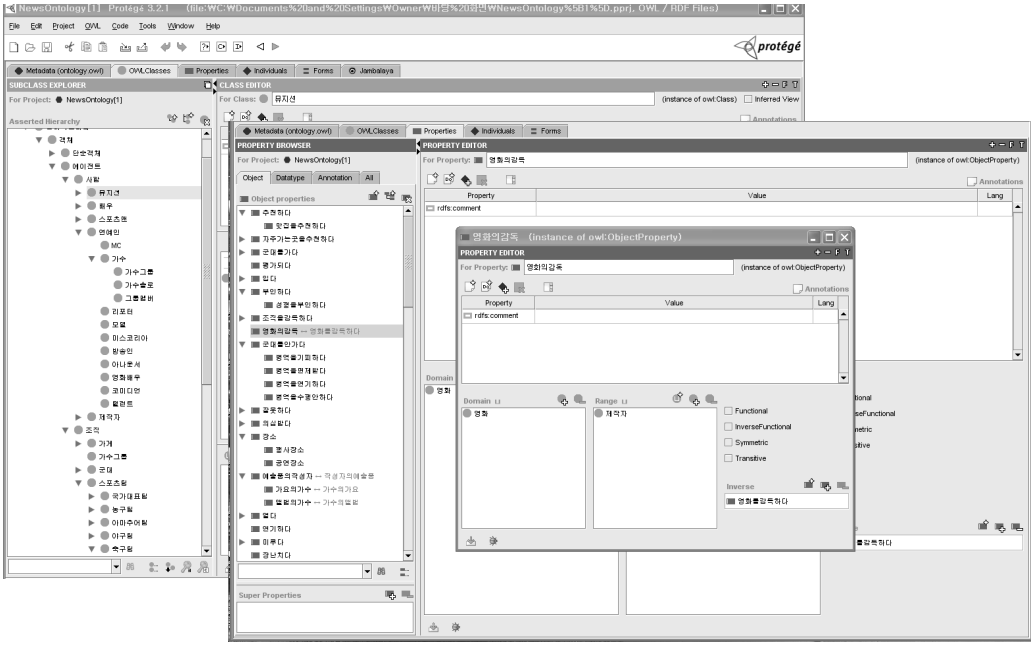
4.3.1 온톨로지 스키마 모델링 및 구축

온톨로지 기반의 검색서비스 제공을 위해서는 우선적으로 온톨로지 스키마를 생성하여야 한다. 온톨로지 스키마는 기존의 확보된 지식정보를 체계화하는 수단으로서 기존의 지식정보에 분류화 과정으로 통하여 다양한 클래스(class), 클래스를 특성화할 수 있는 속성(attribute)과 클래스와 클래스 또는 속성과 속성간의 관계(relationships)를 정의함으로써 해당 분야에서 공유될 수 있는 개념을 명시적으로 규정할 수 있다. 본 연구에서의 온톨로지 구축과정은 수집된 뉴스 원문기사로부터 의미단위의 용어를 추출하는 것이다. 이러한 과정에서 중요한 요소로서 정의될 수 있는 의미단위용어는 뉴스문서에 등장하는 단어를 의미하는 것으로서 여기에는 명사뿐만 아니라 동사, 형용사를 포함한다. 다음 과정으로는 추출된 의미단위용어 사이에 존재하는 동의어들에 대한 군집화과정이 수행되며 이러한 과정을 통하여 해당 군집에 대한 대표어가 정의된다. 각각의 의미단위용어들은 계층화의 형식을 가지면서 가장 적합한 개념, 속성 혹은 요소들로 온톨로지에 관계설정이 된다. 바꾸어 말하면 유사한 개념의 항목들을 대등관계로 군집화하고 이를 상위개념으로 재군집화를 수행한다. 또한, 이미 상하위관계가 존재해도 뉴스기사로부터 새로운 상하위관계가 발생하는 경

우에는 하나의 요소나 개념이 두 개 이상의 상위개념을 포함할 수 있도록 온톨로지에 수용한다. 이러한 과정을 통하여 구축된 온톨로지는 시멘틱 정보추출에 있어서 보다 높은 용이성을 제공하면서 시멘틱 검색이 가능할 수 있는 중요한 기반요소를 제공한다.

한편, 서비스 측면에 있어서 이용자의 정보요구를 반영하고 있는 질의어는 시멘틱 검색시스템에서 제공하는 추천질의어 생성에 있어서도 매우 중요한 역할을 수행한다. 또한 기계어 학습, 자연어 처리 등의 과정으로 통하여 수행되는 추론과정은 질의어 확장과 검색결과에 대한 이용자의 평가 및 지원을 위한 유사도가 높은 검색결과들에 대한 군집화 과정으로 통하여 제공되는 검색결과도 온톨로지를 기반으로 하고 있다. 특히, 뉴스 기사에 대해 단순 텍스트가 아닌 문장의 의미정보를 담고있는 트리플(SPO) 형태의 색인이 가능해진다. 본 연구에서는 온톨로지 모델링을 위한 도구로서 자바언어기반의 개방형 정보원의 통합 온톨로지 구축 프레임워크로 W3C에서 표준 온톨로지 언어로 권고한 OWL 및 RDF(S) 기반 온톨로지 설계에 적합하면서 다양한 플러그인을 확보하고 있으며, 여러 추론시스템과 연동 가능한 장점을 가지고 있는 Protégé를 사용하였다.

〈그림 2〉에서는 온톨로지 저작도구인 Protégé를 이용하여 구축한 뉴스 온톨로지의 일부를 보여주고 있다. 좌측 프레임에 나열된 내용을 살펴보면 “영화배우”, “탤런트”, “가수” 등이 “연예인”이라는 상위개념의 하위개념으로 정의되고 있으며 “연예인”은 “스포츠맨”, “제작자” 등과 함께 좀 더 일반화된 “사람”이라는 개념의 하위



〈그림 2〉 Protégé를 통하여 구축된 뉴스 온톨로지

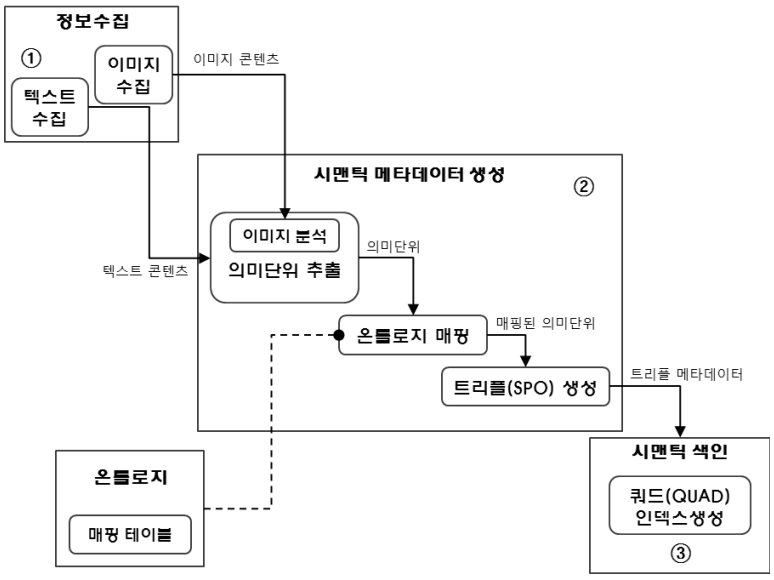
개념으로 정의되고 있음을 확인할 수 있다. 또한, "영화배우"라는 개념의 클래스는 "연예인"의 하위개념이면도 동시에 "배우"의 하위개념이기도 하다. 이와 같이 좌측의 프레임에서 볼 수 있는 클래스 탭(class tab)에서는 기본적으로 개념에 대한 상하위 계층을 구분하면서 생성등록 혹은 삭제 등을 용이하게 할 수 있는 기능을 제공한다.

한편 속성 등에 대한 개념을 제공하고 있는 속성 탭(property tab)에서는 개념과 요소간의 관계를 나타내는 객체속성(object property) 또는 특정값(value)을 가지는 데이터유형 속성(datatype property) 등을 정의할 수 있다.

〈그림 2〉에서 보면, "영화"라는 개념에서 "장르: string"라는 데이터유형 속성을 가지면서, "제작자"라는 개념클래스와 "영화를 감독하다"라는 객체속성 관계로 정의되어 있음을 알 수 있다. 또한 "창업하다", "출시하다" 등의 속성은 "운영하다"라는 상위 속성으로 서로 관계를 갖는 것과 같이 각각의 속성들은 계층구조를 가지도록 조직화할 수 있다. 구축된 온톨로지 스키마는 뉴스정보에 대한 모델링을 포함하여 뉴스정보에 포함된 이미지 데이터 및 뉴스 문서관리를 위한 기사분류, 기사제공자, 작성날짜 등과 같은 메타데이터에 대한 모델링을 포함한다.

4.3.2 메타데이터 생성 및 색인

시멘틱 검색을 위해서는 다양한 정보원으로부터 정보수집, 수집된 정보로부터 의미 있는 정보의 추출과 함께, 추출 및 생성된 정보에 대한 구조화 과정을 통하여 시멘틱 검색이 가능하다. 이와 같은 과정에 있어서 정보수집은 정보수집기를 통하여 수행되며 의미정보의 추출과정에서는 '시멘틱 메타데이터 생성'을 의미한다. 수집된 정보와 생성된 메타데이터는 구조화를 위한 색인과정을 거치게 된다. 〈그림 3〉에서는 이와 같은 일련의 과정을 단계별로 보여주고 있다.



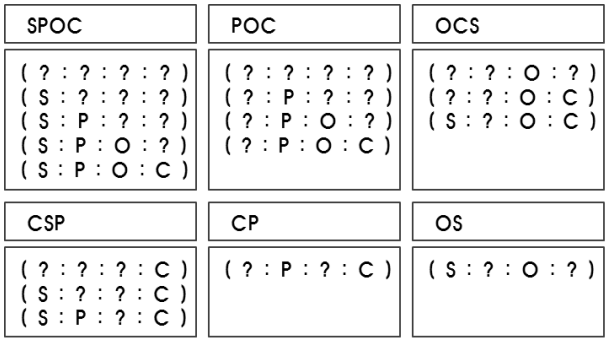
〈그림 3〉 메타데이터 생성 및 저장 과정

먼저 정보수집단계에서는 텍스트와 이미지로 되어 있는 콘텐츠를 수집하는 부분으로 뉴스기사를 검색하는데 있어서 구조화된 형식의 텍스트 콘텐츠를 수집하고 해당 콘텐츠와 관련된 이미지들을 자동으로 수집한다. 구조화된 형식은 뉴스기사 원본을 색인하기 위한 특정한 형식을 가진 텍스트 파일이다. 이를 통하여 정보수집에 있어서 구조화된 형식을 기반으로 HWP, PDF, DOC, XML, RSS 등의 다양한 형식의 정보를 수집할 수 있도록 모듈화 하여 텍스트 정보를 포함하여 멀티미디어 정보를 수집할 수 있도록 한다. 두 번째 단계로서 온톨로지 구축에 있어서 핵심이 되는 시멘틱 메타데이터 생성이 수행된다. 시멘틱 메타데이터 생성과정은 의미단위용어의 추출, 구축된 온톨로지에서의 속성, 클래스, 개념 등에 대한 관계성을 확보하는 매핑과정, 매핑과정을 통하여 의미정보를 생성하는 과정이 진행된다. 의미단위용어의 추출은 위에서 설명하고 있는 온톨로지 구축과정의 의미단위용어 추출과 동일하다. 이와 같은 과정을 통하여 뉴스의 원문기사로부터 의미단위용어를 추출한다. 그러나 이미지정보의 경우는 이미지 분석을 통하여 등장인물과 Architecture, Indoor, Terrain, Night, Snowscape, Sunset, Waterside의 일곱 개 분야의 이미지에 대한 의미들을 추출해 낸다.

둘째, 온톨로지 매핑은 전 단계에서 추출된 의미단위용어가 온톨로지내의 어떤 요소 또는 속성 등에 연관되는지를 결정하는 단계라고 할 수 있다. 개별적인 의미단위용어들은 해당 용어와 가장 적합한 개념, 속성 혹은 요소들에게 관계성을 갖도록 연결되게 되며 이를 통하여 의미단위용어가 가지고 있는 중의성문제를 해결할 수 있다. 마지막으로 트리플 생성은 온톨로지와 맵핑된 각각의 의미단위용어를 뉴스정보가 가진 정보로 표현하는 과정으로 주어부(subject), 서술부(predicate), 목적부(object)로 구성된 SPO 형식의 트리플구조로 표현된다. 이와 같은 과정을 통해 해당 뉴스정보는 기존의 태그방식이 아닌 온톨로지와 연결된 완전한 정보를 메타데이터로 갖게 된다.

시멘틱 색인단계는 이전 단계에서 생성된 메타데이터를 검색의 효율성을 위해 구조화하는 단계라고 할 수 있다. SPO구조의 트리플 형태인 메타데이터를 뉴스정보에 대한 출처가 포함된 쿼드(QUAD) 형태인 SPOC(subject, predicate, object, context)로 표현하면서 각 요소에 대하여 가능한 모든 조합을 추출하여 색인한다. 단순질의를 위해 요구되는 모든 접근유형에 대한 조합의 수는 총 16가지로 산출될 수 있으나 B+트리 색인방법을 활용하여 색인을 구성함으로써 Prefix 검색이 가능하다. 따라서 spoc, poc, ocs, csp, cp, os의 총 6개의 색인형식을 구성할 수 있으며 어떤 인덱스를 이용해 검색을 할지는 질의유형에 따라 적절한 색인형식을 선택하여 탐색하게 된다. 예를 들어 Predicate과 Context를 질의어로 Subject와 Object를 검색하는 (?:P:?:C) 형태의 질의 유형의 경우에 있어서 CP로 구성된 인덱스를 탐색하면 효율적으로 검색결과를 확보할 수 있다.

〈그림 4〉에서는 이용자의 질의유형에 대한 색인구조를 보여주고 있으며 이를 통하여 실질적인 검색에 있어서 보다 쉽게 검색결과를 확보할 수 있다. 〈그림 4〉에서 보여주고 있는 색인구조는 이용자의 다양한 질의형식에 대한 최소한의 색인구조를 보여주고 있으며 보다 다양하면



〈그림 4〉 색인구조

서 대용량의 데이터 처리 및 검색 성능의 향상을 위해서는 색인구조를 추가적으로 생성할 수 있다.

4.3.3 이용자 인터페이스

시멘틱 검색에서 이용자의 정보요구를 질의어의 형태로 입력 받으면서 한편으로 이용자의 질의어와 검색결과에 대한 분석을 통하여 질의어 추천 및 연관어 검색 기능 등을 제공하기 위해서 이용자 인터페이스는 매우 중요한 요소라고 할 수 있다. 본 연구에서는 이와 같은 이용자 인터페이스에 대한 요구조건을 충족하면서 기존의 시멘틱 검색시스템이 가지고 있는 이용에 따른 어려움과 복잡성을 극복하기 위하여 시멘틱 검색 인터페이스를 크게 세 부분으로 구성하였다. 세 가지의 인터페이스는 먼저 질의어 입력 인터페이스, 질의어 추천 인터페이스, 결과 제공 인터페이스로 구성되어 있다. 이를 다시 입력 부분과 출력 부분으로 구분하면 입력 부분은 질의어 입력 인터페이스, 질의어 추천 인터페이스로 구성되어 있어서 이용자는 질의어 입력 인터페이스를 통하여 질의어를 입력하면 추천 질의어 인터페이스를 통하여 이용자 질의어에 입력된 질의어를 기반으로 연관된 질의어를 추천 받는다. 따라서 이용자는 질의어 추천 인터페이스에서 자신의 질의를 선택하거나 다시 질의어 입력 인터페이스를 통하여 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 시멘틱 검색 인터페이스에 대한 개별 구성 기능은 다음과 같다.

우선 일반 질의어 입력 인터페이스는 기존 검색 인터페이스의 용도 및 이용방식에 있어서 유사하다. 그러나 기존의 질의어 입력 인터페이스와 차이점은 질의어 입력 인터페이스에 질의어가 입력되면 질의어 추천 인터페이스와 동기화되어 질의어 추천 인터페이스에는 해당 입력 값에 연동되는 온톨로지의 개념과 관계를 기반으로 검색시스템에서 제공하는 추천 질의어를 제시한다. 이와 같은 질의어 입력 인터페이스는 기존 검색 방식에 익숙한 검색 이용자에게 유용하고, 검색 행위의 비용 요인인 투입 수고와 인지/심리적 요인을 최소화할 수 있다. 또한 조작

이 용이하고 친근한 방식이므로 이용자의 관점에서는 기존 검색 방식과의 차이점을 느끼지 못할 뿐만 아니라 시스템에서 제공되는 연관성 있는 추천 질의어를 통하여 질의어의 확장 및 수정에 있어서 높은 효율성을 확보할 수 있다.

두 번째로 검색 이용자에게 의미적으로 연관된 질의어를 생성해주는 질의어 추천 인터페이스이다. 질의어 추천 인터페이스는 질의어 입력 인터페이스와 동기화되어 있어서 이용자가 질의어 입력 인터페이스를 통하여 질의어를 입력하게 되면 해당 질의어와 연관된 온톨로지의 개념과 관계를 기반으로 다양한 질의어를 추천할 수 있는 인터페이스라고 할 수 있다. 이러한 질의어 추천 인터페이스는 이용자에게 질의어를 표시할 뿐만 아니라, 이용자가 추천된 질의어 중 선택할 수 있는 역할을 수행한다. 질의어 추천 인터페이스는 검색행위의 관점에서 질의어에 대한 재입력의 수고를 최소화하면서 검색하고자 하는 대상에 대한 추천을 통해 이용자에게 인지/심리적 부담을 최소화할 수 있다. 또한 이용자 피드백에 대한 효과성을 보다 높일 수 있으며 시멘틱 검색이 효과적으로 검색할 수 있도록 지원할 뿐 만 아니라 추천 질의어를 수정, 삭제, 선택할 수 있도록 지원함으로서 이용자의 정보요구가 확보될 수 있기까지 이용자와 검색시스템 간의 상호작용을 지원한다. 특히 질의어 추천 인터페이스는 이용자가 질의어 입력 인터페이스를 통하여 입력된 최소한의 단어에 대하여 의미적으로 연관된 온톨로지 기반 질의어를 추천하므로 검색 결과의 재현율을 크게 향상시킬 수 있도록 지원한다. 이와 같은 과정은 이용자가 자신의 정보요구에 대한 명확한 인지가 없는 경우 또는 명시적으로 표현할 수 없을 경우에 보다 높은 효과를 얻을 수 있다.

마지막으로 검색결과를 제공하는 인터페이스는 질의어 입력 인터페이스, 질의어 추천 인터페이스에서 입력된 질의어의 결과를 제공하는 화면으로 이용자가 검색결과에 대한 해석 및 평가에 따른 어려움과 부담감을 최소화하기 위하여 연관성 높은 결과들을 군집화하여 표현함으로써 이용자의 검색결과에 대한 선택과 활용에 따른 효과성을 높일 수 있을 것이다.

5. 결 론

현재까지 시멘틱 웹에 대한 응용서비스 분야로서 시멘틱 검색에 대한 연구는 초기단계로서 현재 진행되는 연구분야에 있어서 시멘틱 웹의 구성 요소인 에이전트, 추론시스템, 온톨로지 등에 대한 기술적 검증과 프로토타입 구현 등 기술 중심 연구로 이루어지고 있는 상태이다. 본 연구에서는 실질적인 시멘틱 검색서비스를 위한 시멘틱 검색시스템에 대한 전체적인 시스템 구조, 메타데이터생성 및 색인구조 및 이용자 인터페이스에 대한 부분적인 구현과 제안을 하고 있다.

특히, 본 연구에서는 검색 이용자를 위한 시멘틱 기반 검색에 대한 요구사항을 도출하고, 이를 반영하는 시멘틱 검색시스템에 대한 전체적인 구조와 색인구조에 대한 설계와 구현을 수행하였다. 시멘틱 검색 관점에서 이용자의 정보요구를 반영할 수 있도록 시멘틱 검색시스템 설계에 있어서 이용자와 검색시스템 간의 상호작용성, 검색결과 제공에 있어서 이용자의 결과 해석 및 평가에 따른 편의성 등을 고려한 제안

을 하였다. 본 연구에서는 기존의 시멘틱 검색에 대한 연구가 주로 기술적인 측면에 중점을 두고 있는 반면에 이용자의 편의성 및 이용자와 시스템 간의 상호작용성에 대한 고려를 통한 요구사항 및 구조를 제안하고 있다는 점에 있어서 향후 추가적인 연구에 도움이 될 수 있을 것이다. 추가적인 연구로서는 본 연구에서 대상으로 하고 있는 뉴스정보로부터 추출된 의미단위로 구성된 지식베이스는 다양한 요소들을 포함하는 대규모(large-scale) 온톨로지를 생성해 낸다. 이와 같은 과정은 온톨로지 기반의 시멘틱 정보의 자동/반자동 추출 및 지능화된 검색 서비스 제공을 위하여 필수적인 요소라고 할 수 있다.

그러나 대규모 온톨로지를 기반으로 실제 응용서비스를 개발에 있어서 대용량 데이터에 대한 효과적인 관리에 따른 많은 문제점이 존재한다. 일부 온톨로지 관리도구에서 온톨로지 정보를 효과적으로 제공하기 위해 그래프 형태의 이차원적 화면을 이용자에게 제공하고 있으나, 실세계 정보를 다루는 응용분야에서는 개념 및 요소 등의 다양한 개체들 간의 관계가 복잡하여 이를 이해하는데 있어서 많은 어려움이 따른다. 따라서 지속적으로 이와 같은 문제점을 해결하기 위한 연구가 요구되며 모든 주제분야 및 대용량의 문서처리를 수행할 수 있는 대용량 정보환경에서의 실질적인 서비스 제공을 위한 연구가 지속적으로 요구된다.

참 고 문 헌

권창호. 2009. 토픽맵 기반의 기록정보 검색시스템 구축에 관한 연구. 『기록학연구』, 19: 57-102.

김정훈, 김건수, 이지형. 2008. 시멘틱 검색은 사회적공동체에 기반한다. 『한국인터넷정보학회 2008 춘계학술발표대회 발표논문집』, 9(1): 217-221.

김학래, 김홍기. 2007. 시멘틱 웹/온톨로지 기술을 이용한 개인용 전자문서 검색 시스템. 『한국전자거래학회지』, 12(1): 135-149.

이준환, 박은종. 2010. 감성 시멘틱을 이용한 영상검색. 『정보과학회지』, 28(8): 46-53.

장명길, 김현진, 장문수, 최재훈, 오효정, 이충희, 허정. 2001. 의미기반 정보검색. 『정보과학회지』, 19(10): 7-18.

하상범, 박영택. 2004. 온톨로지를 통한 추론형 시멘틱 검색 시스템에 관한 연구. 『한국정보과학회 2004년도 봄 학술대회 발표논문집』, 31(1: B): 625-627.

허선영, 김은경. 2008. 시멘틱 검색엔진 설계 및 구현. 『2008 한국컴퓨터종합학술대회 논문집』, 35(1): 331-335.

Albertoni, Ricardo, Alessio Bertone, and Monica De Martino. 2004. "Semantic Web and Information Visualization." *1st Italian Semantic Web Workshop*, Ancona, Italy.

Bangyong, Liang, Tang Jie, and Li Juanzi.

2005. "Association Search in Semantic Web: Search + Inference." *WWW 2005 Conference*, Chiba, Japan.

Bonio, Dario, Fulvio Ccorno, Laura Farinetti, and Alessio Bosca. 2004. "Ontology Driven Semantic Search." *WSEAS Transaction on Information Science and Application*, 6(1): 1597-1605.

Clusty. 〈http://clusty.com〉.

Collarity. 〈http://www.collarity.com〉.

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

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

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

Hakia. 〈http://www.hakia.com〉.

Henzinger, Monika. 2000. "Google Tutorial: Web Information Retrieval." *Tutorial on Web Information Retrieval at ICDE' 2000(16th International Conference on Data Engineering)*.

IDC. 2011. *The 2011 Digital Universe Study: Extracting Value from Chaos*. [cited 2012.2.12]. 〈http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm〉.

Liu, Jiming, Ning Zhong, Yiyu Yao, and Zbigniew W. Ras. 2003. "The Wisdom Web: New Challenges for Web Intelligence." *Journal of Intelligent Information Systems*, 20(1): 5-9.

Makela, Eetu, Eero Hyvonen, and Samppa Saarela. 2006. "Ontogator - A Semantic View-Based Search Engine Services for Web Applications." *5th International Semantic Web Conference 2006(ISWC 2006)*, Athens, GA, USA.

Oddy, R. 1997. "Information retrieval through man-machine dialogue." *Journal of Documentations*, 33(1): 1-14.

OntoWeb. 〈http://www.ontoweb.org〉.

Sheth, A. 2004. "From Semantic Search & Integration to Analytics." *Dagstuhl Seminar on Semantic Interoperability and Integration*, IBFI, Schloss Dagstuhl, Germany.

Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. 2003. "SemTag and SemSeeker: Bootstrapping the Semantic Web via automated semantic annotation." *Proceedings of the 12th International WWW Conference(WWW 2003)*, Budapest, Hungary.

Sure, Y. and V. Iosif. 2002. "First Results of a Semantic Web Technologies Evaluation." *DOA'02 Conference*, Karlsruhe, Germany.

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

Fiske, S. T. 1992. "Thinking is for doing: Portraits of social cognition from Daguerrotypes to Laserphoto." *Journal of Personality and Social Psychology*, 63: 877-889.

Fujimura, K., T. Inoue, and M. Sugisaki. 2005. "The EigenRumor Algorithm for Ranking Blogs." *WWW 2005 Workshop on the Weblogging Ecosystem*, Chiba, Japan.

Qiu, G., K. Liu, J. Bu, C. Chen, and Z. Kang. 2007. "Quantify Query Ambiguity using ODP Metadata." *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, Denmark.