시맨틱검색엔진의 성능평가에 관한 연구*

A Study on the Performance Evaluation of Semantic Retrieval Engines

노 영 희(Younghee Noh)**

초 록

본 연구에서는 유동성이 크고 데이터의 규모도 상당한 도서관에 일반화시켜 적용할 수 있는 지식베이스 및 검색엔진을 제안하였다. 이를 위해 총 세 개의 지식베이스(트리플 구조 온톨로지, 의미거리기반 의미망지식베이스, 키워드중심의 도치색인파일)를 구축하였고, 이의 성능을 측정하기 위해 각각 세 개의 검색엔진(추론규칙기반 제나검색엔진, 개념기반 검색엔진, 키워드기반 루씬검색엔진)을 구축하였다. 시스템 성능평가 결과, 종합적으로 개념기반 검색엔진이 가장 높은 성능을 보여주었고, 다음으로 온톨로지기반 제나검색엔진, 다음으로 일반 키워드 검색엔진 순으로 나타났다.

ABSTRACT

This study suggested knowledge base and search engine for the libraries that have the large-scaled data. For this purpose, 3 components of knowledge bases(triple ontology, concept-based knowledge base, inverted file) were constructed and 3 search engines(search engine JENA for rule-based reasoning, Concept-based search engine, keyword-based Lucene retrieval engine) were implemented to measure their performance. As a result, concept-based retrieval engine showed the best performance, followed by ontology-based Jena retrieval engine, and then by a normal keyword search engine.

키워드: 시맨틱 검색엔진, 개념기반 검색엔진, 키워드 검색엔진, 온톨로지, 성능평가, 제나, 루씬
Semantic Retrieval Engine, Concept-based Retrieval Engine, Keyword Retrieval Engine, Ontology, Performance Evaluation, Jena, Lucene

---

* 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 연구되었음 (KRF-2009-327-H00016).

** 건국대학교 인문과학대학 문헌정보학과 부교수(irs4u@kku.ac.kr)

논문접수일자 : 2011년 5월 19일 논문심사일자 : 2011년 6월 2일 게재확정일자 : 2011년 6월 14일

1. 서 론

1.1 연구의 배경 및 필요성

일반적인 웹의 목적은 사람 사이의 원활한 커뮤니케이션과 더불어 컴퓨터시스템들 간의 효과적인 커뮤니케이션이라 할 수 있다. 이 목적의 실현은 인터넷 자원을 소화할 수 있는 하부구조의 정립을 전제한 매우 복잡하고 시급한 과제라 할 수 있다. 단일 인터페이스로 모든 유형의 정보(텍스트, 음악, 동영상, 이미지 등)에 접근하는 방법과 용이성과 융통성을 그 특징으로 하는 HTML의 광범위한 수용으로 인터넷 정보량은 폭발적으로 증가했고 그 결과 전문색인 기반 검색엔진의 효율은 현격히 낮아졌다. 단순히 용어의 발생빈도수를 따라 정보를 차등 취급하는 방법은 웹 문서에 기술된 모든 내용에 거의 일률적인 중요성을 부여하는 것이며 이것은 전문색인 검색의 근본적인 단점이라고 할 수 있다.

이러한 현실의 한계를 극복하기 위해 자원 내용에 관한 요약정보로 검색기능을 강화할 수 있는 메타데이터 시스템 개발에 국제적인 관심이 집중되어 왔다. 그 중 특히 국제적 합의에 기반하여 제정된 더블린코어 메타데이터 스키마는 자원 유형과 무관하게 일반적으로 적용할 수 있는 국제적인 표준으로 이미 널리 알려져 있다. 그 밖에도 매체 유형의 특징을 반영한 메타데이터 스키마들이 다수 개발되어 구성요소가 단순한 더블린코어 메타데이터 스키마로 구현할 수 없는 상세 자원 기술에 쓰이고 있다.

모든 메타데이터가 RDF(Resource Description Framework)의 트리플(triple) 구조로 표현되며, 자원에 대한 기술이 주어-술어-목적어를 갖춘 논리적 단위로 저장되기 때문에 거대한 트리플 지식창고를 생성하게 되는 효과가 있다. 이 지식창고를 활용하여 논리적 추론이 가능한 새로운 유형의 질의에 답할 수 있고 트리플에 직접적으로 표현되어 있지 않은 암묵적 지식의 유추도 가능하게 된다.

그러나 이러한 RDF 스키마에는 서로 동일한 의미의 요소, 역관계, 통합, 상관성 등 메타데이터의 중요한 관계를 지원하는 능력이 결여 되어 있어서 시맨틱 웹 구성에 본격적인 웹 온톨로지 언어가 필요하다. RDF와 RDF 스키마를 기반으로 이 두 언어에 부족한 모델링 요소를 확장, 강화하여 개발된 언어가 DAML+OIL(DARPA Agent Markup Language+Ontology Inference Layer) 마크업 언어이다. 그러나 이 언어의 취약점은 용어간의 의미 혼동을 일으킬 수 있다는 것이며 이러한 단점을 보완하기 위해 이 언어와 거의 완벽한 호환성을 유지한 OWL이 제시된 것이다. 이 언어는 웹 문서 및 응용프로그램에 내재한 클래스와 속성들 간의 관계 정의 기능을 강화함으로써 DAML의 단점을 보완하도록 발전된 언어이다.

이처럼 시맨틱 웹은 정보를 지식화하여 정보의 효율적 검색, 통합, 재사용을 도모하는 새로운 기술로, 이를 효과적으로 표현하기 위해 다양한 기술이 개발되어 왔다.

그러나 현재 국내의 언어자원(시소러스, 사전, 온톨로지 등) 구축 및 연구사업은 특정 도메인(예를 들어 전산학분야, 의학분야 등)에 제한되어 구축되고 적용되는 실험적인 수준이 대부분이며, 다양한 분야의 구축과 적용이 가능하도록 새롭고 효율적인 언어자원 개발방안

은 매우 부족한 형편이다. 시소러스 및 온톨로지 구축사업은 수많은 전문가에 의한 수작업 구축방식에 의존하면서도 실제 개념을 정의하는 개별 전문가의 의견차이 또는 개념의 모호성으로 인해 정보검색은 물론 언어자원 구축의 실효를 거두는데 한계가 있었다. 또한 사업의 특성상 많은 분야별 전문가집단과 막대한 예산, 소요기간이 필요하기 때문에 전통적인 사업방식을 탈피하여 보다 효율적인 연구개발 방법이 필요하다. 따라서 효율적인 자원구축 기법으로 구축된 온톨로지를 기반으로 현실적인 의미기반 검색시스템을 개발하는 것이 필요하다 하겠다. 특히 동적환경의 다양한 주제분야의 지식정보를 의미망 형태로 효과적으로 구축하고 의미적 연관검색이 가능하도록 하기위해, 의미기반 연관검색(추론 알고리즘)에 관한 보다 깊이 있는 연구가 필요하다고 본다.

1.2 연구의 목적

본 연구에서는 기초 언어자원을 효과적으로 구축할 수 있는 방법론을 모색하고 그 성능을 평가한 1차 연구결과를 기반으로 추론에 기반한 의미검색 환경을 구현할 수 있도록 하고자 하였다(Noh 2011). 즉, 1차 연구에서 구축된 온톨로지는 정보관리학회지에 수록된 논문기사 189건이며, 누리미디어의 DBPia로부터 추출된 데이터이다. 이 데이터베이스는 논문기사에 대한 서지정보를 포함하여 목차정보와 초록정보, 그리고 저자의 소속정보 등이 메타데이터로 구축되어 있으며, 원문정보를 포함하고 있다. 이 때, 원문정보를 제외한 나머지 메타데이터만을 활용하여 서지온톨로지를 구축하였다. 온톨로지 작성 시의 편의를 위해 한자들은 모두 한글로 변환하였다. 원본 서지정보로부터 파서 및 필드분석기를 거쳐 정규화된 서지정보를 추출하고 이로부터 온톨로지를 생성하게 되는데, 개념 온톨로지는 수작업으로 구축하였고, 서지온톨로지는 자동으로 구축하여 각각 OWL 개념온톨로지와 OWL 서지온톨로지를 생성하였다. 시스템의 성능을 평가하기 위해 루씬(Lucene)이라는 검색엔진의 루씬자동색인기를 이용하여 전문검색용 색인파일을 별도로 구축하였다.

즉 1단계 연구에서 구축된 온톨로지를 기반으로 3개의 검색엔진을 개발하여 적용하고, 그 성능을 비교하고자 하였다. 본 연구단계에서는 개념기반 검색엔진과 이를 위한 의미망기반 지식베이스를 추가적으로 개발하였으며, 총 세 개의 검색엔진의 성능을 비교・평가하였다.

연구의 목적을 달성하기 위한 구체적인 세부목적은 다음과 같다.

첫째, 온톨로지를 구축한다. 온톨로지 구축 대상은 정보관리학회지 2007년부터 2009년까지의 3년간의 논문기사를 대상으로 하였으며, 구축방법은 온톨로지 구축도구를 이용한 수작업에 의한 구축방법과 알고리즘을 이용한 자동적인 구축방법으로 나뉜다.

둘째, 온톨로지 및 추론엔진의 성능을 비교・평가하고자 하였다. 성능평가를 위한 방법론은 다양하겠지만 본 연구에서는 가장 일반적인 성능평가 방법인 재현율과 정확률, 그리고 이 둘의 조합인 F_1척도를 사용하여 그 성능을 평가하고자 하였다.

셋째, 본 연구에서는 지금까지 연구되고 적용되어 온 온톨로지 구축방법을 이용하되, 구

축된 온톨로지로부터 이용자의 요구에 적합한 자료를 탐색하는 검색기법의 적용에 있어서, 개념기반 정보검색기법(Concept-based Information Retrieval Techniques)을 적용하였다. 개념기반 정보검색은 2000년대 초반에 연구되기 시작하였으며, 그 성능에 있어서 현재까지 연구되어 온 다른 어떤 검색기법보다 강력한 검색기법으로 평가되어 왔으나(Noh 2001), 사실상 그 개념을 이해하는데 이론이 너무 어려울 뿐만 아니라 개념기반 검색대상이 되는 의미망 지식베이스를 구축함에 있어 문헌수가 증가함에 따라 그 지식베이스 구축시간이 기하급수적으로 증가한다는 한계 때문에 10년 전에는 상용시스템에 적용하거나 일반화 시키지 못하고 있는 실정이었다. 그러나 최근에는 정보통신기술의 발전으로 하드웨어시스템의 성능이 향상되었음을 감안하여 제고할 필요가 있는 검색기법이라 판단된다. 이에 본 연구에서는 개념기반 정보검색기법을 시맨틱 웹 검색기법과 그 성능을 비교함으로써 실제 적용성 및 검색 성능의 향상을 도모하고자 하였다.

1.3 연구의 내용 및 방법

여러 가지 방법을 통해 온톨로지를 기반으로 한 시맨틱 웹 기반 검색이 가능하며 본 연구를 통해 구현하게 될 검색엔진은 개념 사이의 관계를 선언적으로 표현한 온톨로지 구조에 더해 절차적 관계를 표현함으로써 원하는 지식 항목의 검색을 하는 것이며, 이것은 논리적 추론 기술을 통해 가능한 것이다. 그 구체적인 연구 방법 및 연구 절차를 소개하면 아래와 같다.

첫째, 시맨틱 웹 및 추론기법 관련 이론들에 대해 지금까지 연구되어 온 것을 전반적으로 검토하였다.

둘째, 검색엔진의 성능을 평가하기 위해 온톨로지를 구축하였으며, 수작업과 자동으로 각각 온톨로지를 구축하고, 각각의 방법에 의해 구축된 온톨로지의 성능을 비교・평가하였다.

셋째, 구축된 온톨로지 및 지식베이스에 추론엔진을 적용하였다. 선행연구에서 개발한 다양한 추론엔진을 적용하고, 또한 개념기반 검색기법을 적용하였으며, 각각의 성능을 비교・분석하였다.

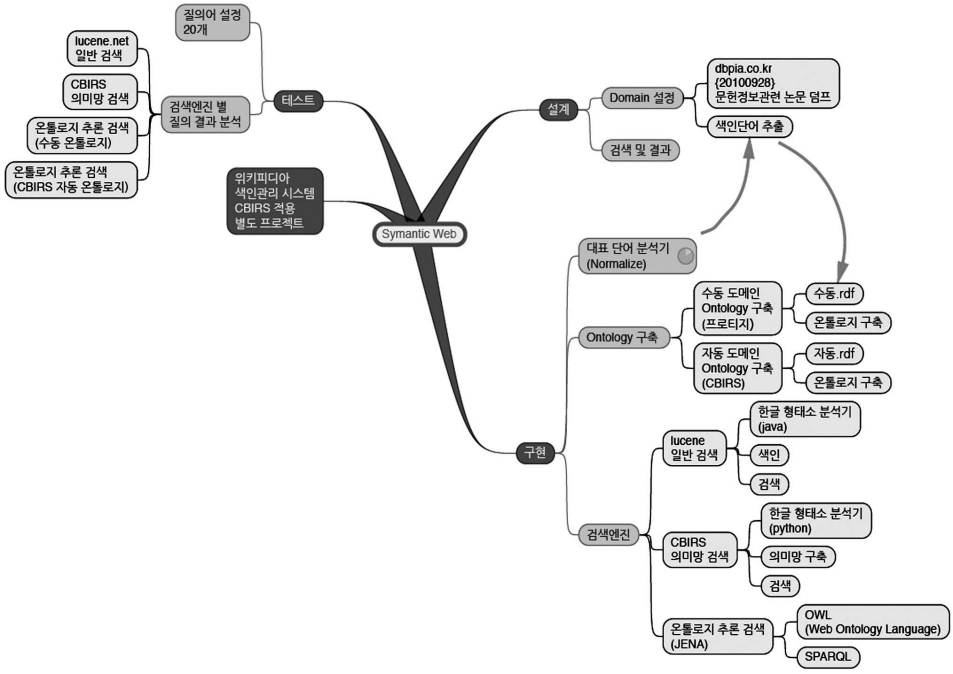
위의 연구내용 및 본 연구자가 구상한 마인드 맵은 〈그림 1〉과 같다.

2. 이론적 배경

2.1 선행연구

본 연구에서는 온톨로지에 추론엔진을 적용한 시맨틱검색관련 연구를 살펴보았으며, 시맨틱검색기법과 기존의 키워드검색기법의 성능을 평가한 연구들 중심으로 분석하였다.

국내에서 수행된 대부분의 논문이 온톨로지를 기반으로 한 시맨틱검색시스템의 구현에 초점을 두고 있고, 그 시스템 구성을 소개하는 논문이 다수인 것을 알 수 있다. 박종욱(2008)은 자연어 온톨로지기반 검색기술과 온톨로지를 활용하여 지능형 검색시스템을 만들고자 하였다. 이 온톨로지기반 검색시스템으로 복잡한 질의문을 수행하기 위해 온톨로지를 구축하고 메타데이터에 대한 표준화와 함께 명명화하여 사용자가 입력하는 자연어에 맞는 온톨로지기



〈그림 1〉 연구내용 구성 및 마인드맵

반 질의문으로 자동 생성하고 추론에이전트로 하여금 OWL에서 온톨로지 메타데이터 지식베이스의 추론을 통해 추출된 통계데이터를 재가공하여 사용자가 원하는 형태로 재생산하는 방식이다. 이 연구는 온톨로지를 개발하였으나 개발된 온톨로지의 성능을 평가하는 단계까지 가지는 못했다.

정은경 등(2003)의 연구도 시맨틱 웹 환경에서의 온톨로지기반 정보검색시스템을 개발하였다. 그러나 검색대상이 여행자정보라는 것과 다국어를 지원했다는 것 외에 온톨로지기반 검색시스템의 성능을 개선하기 위한 방법은 모색하지 못했다.

이재원 등(2010)은 사용자의 질의어, 선호도, 카탈로그 문서의 시맨틱을 도출하기 위해 분류지식베이스로부터 추출한 개념을 이용하였다. 도출된 개념을 이용하여, 사용자의 질의어 및 카탈로그 문서들 간의 색인어 불일치 문제를 해결한 시맨틱검색 모델을 제시하였다. 또한 사용자의 선호도 정보 역시 개념으로 표현함으로써, 협업적 필터링 기반 추천 알고리즘의 치명적인 단점인 희박성 문제를 해결하였다. 특히, 이전의 정보 추출 연구들이 검색 혹은 추천의 한 측면에 초점을 맞춘 모델을 제시한 반면, 이 연구는 검색 및 추천을 시맨틱 공간에서 수행할 수 있는 정보 추출 모델을 제시하였다.

김태환, 전호철, 최중민(2008)은 정보의 공유를 목적으로 하는 커뮤니티를 시맨틱 웹 서비스(Semantic Web Services) 기술을 이용하여 플랫폼에 상관없이 사용자 검색 질의와 가장 유사한 커뮤니티를 의미적으로 식별해 내고 커뮤니티 내의 정보 중 질의와 관련된 정보를

검색결과로 도출할 수 있는 구조를 제안하였다. 그러나 시스템간의 성능비교 등을 고려하지 않은 한계가 있다.

한동일, 권혁진, 정학진(2007)은 시맨틱검색 시스템에 관한 포괄적인 개념적 모델 제안과 실질적인 구현 사례를 제시한다. 제안된 시맨틱검색시스템은 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 온톨로지기반 질의 확장 등을 통해 시맨틱검색을 처리하는 영역이다. 마지막으로 지식이용(knowledge utilization) 계층은 검색 이용자가 시맨틱 웹 언어 또는 온톨로지에 대한 지식이 없더라도 직관적으로 검색 질의를 입력하고 검색 결과를 확인할 수 있도록 구성하였다.

그 외 연구로 박진석, 양기철, 오정진(2004)은 온톨로지기반 시맨틱검색시스템을 구현하였으며, 의미는 온톨로지 구축대상이 박물관 유물정보라는 것이나, 역시 성능평가 부분은 없다. 하상범, 한은영, 최호준(2005) 역시 OWL 기반의 SPARQL을 이용한 시맨틱검색시스템을 구현하였으며, 그 성능 평가는 하지 않았다.

기존의 키워드검색기법과의 성능차이를 비교한 논문은 소수이며, 온톨로지기반으로 구축된 시스템과 키워드검색기법으로 구축된 시스템과의 성능차이를 비교한 논문이 몇몇 있다. 강래구(2007)는 온톨로지를 상품 검색에 활용하기 위해 상품도메인 온톨로지를 구축하여 자연어 및 분류별 검색을 통해 얼마나 정확한 검색을 하게 되는지를 실험하였다. OWL로 온톨로지를 구축하고 온톨로지 추론알고리즘과 일반 키워드검색기법의 성능을 비교하였다. 그 결과 온톨로지기반 추론시스템이 키워드검색기법보다 성능이 높게 나타났다고 밝혔다. 그러나 연구에서 개발된 추론엔진에 대한 명확한 설명이 없어 신뢰도에 문제가 있다고 본다.

박지형 등(2007)은 전자부품 분야에서의 시맨틱검색 절차를 제안하였다. 검색절차는 시드검색, 시맨틱탐색, 순위화 3가지 단계로 구성되며, 여기에 'AND' 연산자 기능과 '중심주제어' 검색기능을 추가로 제안하였다. 또한 제안된 시맨틱검색을 제공하는 프로토타입 시스템을 구현하고, 시스템 평가를 위한 실험을 통해 분산되어 있는 다양한 웹 페이지에 대한 기존의 키워드검색보다 약 10% 이상의 성능 향상을 보였다고 하였다. 그러나 이 연구에서는 노드간의 의미적 관계를 측정하여 추론하지 못한 한계를 보였다.

김영민, 이상준(2003)의 경우, 공학 논문들을 대상으로 논문 제목의 구성 형태를 분석하고, 제목 내의 키워드들의 역할 정보들을 RDF 시맨틱으로 구성하여 논문검색에 이용하는 방법을 제안했다. XML형태의 시맨틱을 이용하여 논문 검색에 이용한 결과 키워들 만을 이용하는 기존 방법보다는 훨씬 검색자의 의도를 잘 반영하면서도 필요한 결과만을 얻을 수 있었다. 그러나 검색자의 의도를 반영했는지의 여부를 검색된 문헌의 수로 평가하고 일반적인 성능평가 기준을 적용하지 않아 객관성을 증명하기가 어렵다고 할 수 있다.

한편, 지금까지 개발된 온톨로지기반 검색엔진들의 성능을 비교한 연구가 있다. 심재문(2008)은 지금까지 개발된 추론엔진들의 성능을 평가하는 연구를 수행했다. 추론엔진들 성능을 정적평가, 동적평가, 전반적평가 등으로 구분하여 평가하였으며, 전반적으로 MINERVA가 가장 높은 성능을 보여 주는 것으로 나타났다. 그러나 기존에 개발된 검색엔진의 성능만 평가하였을 뿐 새로운 개선방안 모색은 하지 못한 논문이라 할 수 있다.

해외의 대부분의 연구가 온톨로지기반 검색시스템을 개발한 후 그 검색성능을 평가하는 단계까지 나아가는 것을 알 수 있으며, 주제분야나 그 기법면에서 매우 다양함을 알 수 있다. 예를 들어 단점을 보완하기 위해 온톨로지기반 시맨틱검색시스템을 개발한 경우로, 온톨로지 대상 질의어로 RDQL을 사용한 경우(Seaborne 2004), RQL을 사용한 경우(Karvounarakis et al. 2002), SPARQL을 사용한 경우(Prud'hommeaux & Seaborne 2006)가 있다. 또한 질의를 만족시키는 온톨로지 값을 Tuple 형태로 보여줌으로써 자연어 형태에 가깝게 결과를 주는 연구도 있다(Castells et al. 2004; Christophides et al. 2003; Maedche et al. 2003).

지금까지 수행된 연구들의 경우 시맨틱검색시스템을 개발하는데 주력하고 있거나 좀 더 나아간다면 키워드검색기법과의 성능차이를 비교하는 정도의 연구를 수행했다. 본 연구에서는 국내외적으로 그 성능이 검증되고 있는 제나기반 시맨틱검색시스템을 구현하고, 의미망기반 개념기반 검색시스템의 성능을 비교하고자 하였다.

원칙적으로 제나검색시스템은 검색대상이 온톨로지이고 개념기반 검색시스템의 검색대상은 노드간의 값이 유사도로 표현된 의미망 지식베이스이다. 동일한 주제분야를 대상으로 하여 20여개의 질의를 입력하고 재현율과 정확률의 평균으로 그 성능을 비교하였다. 추가적으로 루씬이라는 상용의 키워드검색시스템과의 성능도 비교하였다.

2.2 추론엔진

본 연구를 위해 검토된 추론엔진은 MINERVA, 제나를 비롯한 6개의 엔진이며, 이 중 가장 일반적으로 사용되고 있는 추론엔진은 제나로 분석되었으며, 제나는 본 연구에서 실험을 위해 구현 및 평가되었다.

2.2.1 MINERVA

MINERVA는 관계형 데이터베이스(RDB) 시스템 기반에서 개발된 고성능의 OWL 저장, 추론 및 쿼리 시스템이다. 이의 장점은 크게 두 가지로 언급할 수 있는데, 첫째는 반응 속도이며, 둘째는 규모성이다. 사용자가 원하는 검색을 수행하는 시점에 온톨로지 내용을 로드하여 추론하지 않고 사전에 로드하여 가능한 모든 추론을 수행하고 관계형 데이터베이스에 저장을 해 두는 접근법을 사용하기 때문에 규모성에 있어서 매우 우수하다고 볼 수 있다. 더욱이 관계형 데이터베이스에서 모든 추론 연산을 하기 때문에 MIERVA의 규모성은 메모리 기반의 추론엔진보다 우수하다. MINERVA는 현재 IBM의 Integrated Ontology Development Toolkit(IODT)의 한 요소로 제공되고 있다. 쿼리를 할 경우 SPARQL 언어의 부분집합이

라고 볼 수 있고 OWL-DL을 가능하게 하는 DLP(Description Logic Program)를 지원한다. MIERVA는 TBox를 위한 Programming (DLP)으로부터 번역된 논리 규칙집합도 지원한다(Lee et al. 2006; Ma et al. 2004; Zhou et al. 2006).

2.2.2 DLDB-OWL(HAWK)

DLDB-OWL는 데이터베이스 형의 추론엔진이며 DLDB-DL을 지원하고, HAWK은 그의 차기 버전이다. DLDB-OWL은 OWL 온톨로지를 파싱, 편집, 구동, 보관하는 것과 관련된 API를 제공하고 있다. 핵심 패키지는 온톨로지와 클래스나 속성 등과 같은 온톨로지 객체들의 일반 인터페이스를 정의할 수 있고, 온톨로지 모형을 구축하고 활용하는 API를 제공하게 된다. 한편 OWL패키지는 OWL언어로 된 온톨로지를 파싱하고 직렬화(serialization)하는 유틸리티를 제공한다(Pan 2006).

2.2.3 Pellet

Pellet은 Java 언어로 된 오픈 소스로서 OWL-DL 추론기이다. Pellet은 제나와 OWL API와 같이 사용할 수 있으며 DIG 인터페이스를 제공한다. Pellet API는 온톨로지의 일관성 점검, 택소노미의 분류, RDQL 언어로 된 쿼리 등의 지원을 하는 기능을 포함한다(Srin et al. 2007; Parsia et al. 2003; Wessel & Moller 2005).

2.2.4 RacerPro

RacerPro는 OWL 추론기로서 의미망을 위한 추론 서버이기도 하다. RacerPro의 원래 형태는 'Description Logics'으로 구성되어 있으며 OWL 기반의 의미망 온톨로지를 관리하기 위한 시스템으로 활용되고 있다. RacerPro는 대량의 자료를 관리할 수 있어 최적화된 검색 엔진을 가진 의미망 정보 저장소로도 볼 수 있다(Haarslev et al. 2004). Racer 추론기는 STS(Software, Technology & System)의 RacerPro1.9 프로그램을 세팅하면 자동으로 Protege와 연동된다. Protege에서 'Classifier Taxonomy'라는 버튼을 클릭하면 추론기가 실행되면서 오류제어 및 추론과정을 진행시키게 된다. 클래스 인스턴스 값을 채우고, 그와 관련된 추론이 이루어지면 기술하고자 하는 자원에 대한 표현들이 '.owl'이라는 파일 형식으로 출력된다.

2.2.5 Jena

Jena(제나)는 의미망 기반의 애플리케이션을 개발하기 위한 Java 프레임워크로서 RDF, RDFS, OWL 및 SPARQL 프로그래밍 환경을 제공하며, 규칙 기반의 추론 기관을 포함한다. 제나 프레임워크는 RDF API를 포함한다(Carroll et al. 2004; Kevin et al. 2003).

2.2.6 의미망기반 개념기반검색시스템 (CBIRS)

2000년대 초반에 본격적으로 연구되기 시작한 개념기반 정보검색 모형은 기존의 통계적 검색모형의 단점을 보완할 수 있는 차세대 검색모형으로 간주되었다. 개념기반 검색모형은 일반적으로 시스템에 입력되는 문헌 데이터베이스로부터 지식베이스를 자동으로 구축하고, 이 지식베이스를 대상으로 개념확장을 수행한

후 문헌 데이터베이스로부터 관련 정보를 검색한다. 따라서 개념기반 검색모형의 성능을 좌우하는 주요 요소는 지식베이스와 개념확장 알고리즘이라고 할 수 있다.

의미망 구조의 지식베이스를 기반으로 개념확장을 수행하는 알고리즘을 개념확장 알고리즘이라 하며, 개념확장 알고리즘으로는 bnb 알고리즘(branch-and-bound expansion activation algorithm)과 홉필드 넷 알고리즘(Hopfield net algorithm)이 사용되고 있다.

bnb 알고리즘은 적용되는 지식베이스에 따라 크게 경험적(heuristic) bnb 알고리즘과 순차적(sequencial) bnb 알고리즘으로 구분할 수 있는데, 전자는 주로 전통적인 시소러스에 적용되고 용어간의 관계 정의에 따라 개념확장을 수행한다. 후자는 의미망 구조의 문헌기반 지식베이스에 적용되어 지식베이스 내 용어간의 의미 거리에 따라 개념확장을 수행한다. 홉필드 넷 알고리즘은 신경망 구조의 지식베이스에 적용되는 개념확장 알고리즘이다.

개념기반 검색모형의 검색성능은 개념확장 대상이 되는 지식베이스에 따라서도 달라질 것이다. 개념확장 대상이 되는 의미망 구조의 지식베이스를 다양한 방법으로 구축할 수 있는데 문헌기반 지식베이스, 시소러스기반 지식베이스, 통합형 지식베이스, 그리고 동의어 처리형 지식베이스이다.

본 연구에서 개발한 순차적 bnb 알고리즘은 의미망 구조의 지식베이스에 적용되고, bnb 확장 활성화 탐색은 개념확장이 진행되는 동안 최단 경로를 찾기 위한 방법이며, 이용자가 제공한 용어에서 개념확장이 시작된다(Chen & Dhar 1991). 이용자가 제공한 초기 탐색어에는 1의 가중치가 부여되며, 다음으로 이 용어들과 직접적으로 관련이 있는 이웃하는 용어들을 탐색한다. 확장된 용어의 가중치는 이용자가 입력한 용어와의 링크 가중치를 기반으로 산출된다. 특정 기준치까지 용어들을 확장한 후에 확장된 용어와 이용자가 입력한 용어는 용어의 가중치순에 따라 우선순위 대기행렬(priority queue: $Q_{priority}$)에 저장된다.

의미망 구조의 지식베이스기반 bnb 알고리즘은 적절한 이용자 정의 상태에 도달할 때까지 확장 과정을 반복한다. 개념기반 정보검색에 채택된 알고리즘과의 상호작용 과정에서 이용자는 시스템에 적절한 확장용어 수(p)와 확장될 용어의 가중치(W_p)를 제공하도록 요청받는다. 이용자가 제시한 이 두 개의 변수는 bnb 반복 확장 과정의 중지 조건(stopping condition)으로 작용한다. 초기 탐색어는 처음에 동일한 가중치를 갖기 때문에 모두 활성화되며, 첫 번째 확장 후에 $Q_{priority}$는 내림차순으로 모든 초기 탐색어와 직접적으로 연결된 이웃 노드들을 찾아 그것의 W_p를 산출하게 된다. 다음으로 $Q_{priority}$에서 가장 상위의 용어에 대하여 개념확장을 하는 과정을 반복한 후, p번째까지의 노드를 구분해서 이용자가 제시한 두 조건을 모두 만족하는 용어만 최종 탐색문으로 선택된다.

이와 같이 이용자가 지정한 기준치는 시스템이 이용자의 초기 탐색어와 유사한 용어를 적어도 p개 생성할 것을 보장한다. 반복이 진행되는 동안 대기행렬에서 보다 높은 가중치를 갖는 용어들은 $Q_{priority}$에서 높은 순위에 놓이게 될 것이다. 시스템이 검색을 중지하는 시점은 출력 대기행렬의 노드가 p개 이상으로 구성되

어 있을 때, 또는 $Q_{priority}$에서 가장 높은 순위에 있는 노드의 가중치가 이용자가 제시한 기준치 즉, W_p보다 낮을 때이다.

3. 지식베이스 구축

3.1 온톨로지 구축

본 논문에서는 시맨틱검색환경을 구축하기 위해서 OWL 기반의 온톨로지를 구축하고 메타데이터를 생성하였다. OWL은 W3C에서 시맨틱 웹의 온톨로지 언어로 표준화된 언어로써 다양한 OWL공리를 제공한다. 이는 도메인 온톨로지를 객체와 속성의 관계로 유기적으로 정의할 수 있게 한다. 또한 OWL구문을 추론하기 위해서 사용되는 n-Triple형태는 기존의 추론엔진에서 추론에 용이한 구조로써 다양한 온톨로지기반의 추론을 가능하게 한다. 본 논문에서는 OWL Full, OWL DL, OWL Lite언어 중에서 OWL DL을 사용하며 이에 해당하는 공리를 사용하여 온톨로지를 구성하였다.

3.2 의미망지식베이스 구축

본 연구에서는 온톨로지기반 시맨틱검색엔진의 성능을 비교하기 위해 CBIRS를 구현하였으며, 이를 위해 문헌기반지식베이스를 구축하였다. 문헌기반 지식베이스는 네 개의 지식베이스 중 기본 지식베이스가 되는 것으로서, 실험 문헌 집단의 각 문헌에 출현한 용어를 자동으로 추출하고 추출된 용어들의 가중치를 산출한다. 또한 용어간의 유사도를 분석하여 의미망 구조의 지식베이스를 최종적으로 구축하게 되는데 그 과정을 구체적으로 살펴보면 다음과 같다.

먼저, 통계적 기법에 의해 용어와 용어간의 유사도를 산출하고, 이를 기반으로 의미망 구조의 지식베이스를 구축할 수 있다. 즉, 지식베이스를 구축하기 위해 각 문헌으로부터 용어를 추출하고 용어의 가중치를 산출한 다음 용어의 문헌 내 동시출현빈도를 기반으로 유사도를 산출하여 의미망으로 표현한다.

특정 문헌에 출현한 용어의 가중치를 산출하기 위한 수식은 다양하지만, 본 연구에서는 단어빈도와 역문헌 빈도를 각각 최대값으로 나누어 정규화시킨 수식을 사용하였다(Salton, Fox, & Wu 1983).

$$w_{ik} = \frac{tf_{ik}}{\max(tf_{ih})} \times \frac{idf_k}{\max(idf_h)}$$

tf_{ik} = 용어 k가 특정 문헌 i에서 출현한 빈도
$max(tf_{ih})$ = 특정 문헌에서 가장 높은 출현빈도를 갖는 단어의 빈도
idf_k = 용어 k의 역문헌 빈도
$max(idf_h)$ = 전체 문헌 데이터베이스에서 가장 높은 출현빈도를 갖는 단어의 역문헌 빈도

한편, 가중치가 부여된 두 용어간의 의미관계를 생성하기 위해서는 용어간의 유사도가 산출되어야 한다. 개념기반 검색을 위한 의미망 구조의 지식베이스를 구축하는데 사용되는 유사계수도 다양하지만, 본 연구에서는 코싸인 유사계수를 사용하여 용어간의 유사도를 산출하였다.

$$W(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ij} \times d_{ik}}{\sqrt{\sum_{i=1}^{n} d_{ij}^2 \times \sum_{i=1}^{n} d_{ik}^2}}$$

위 수식에서 $W(T_j, T_k)$는 용어 T_j와 용어 T_k간의 유사도 가중치를 나타내고, d_{ij}는 문헌 i에 출현한 용어 T_j의 가중치이며($0 \leq d_{ij} \leq 1$), d_{ik}는 문헌 i에 출현한 용어 T_k의 가중치이다($0 \leq d_{ik} \leq 1$).

4. 추론엔진의 성능평가

4.1 시스템 설계

본 연구에서는 총 세 개의 지식베이스(트리플구조 온톨로지, 의미거리기반 의미망지식베이스, 키워드중심의 도치색인파일)를 구축하였고, 이의 성능을 측정하기 위해 각각 세 개의 검색엔진(추론규칙 기반 제나검색엔진, 개념기반 검색엔진, 키워드기반 루씬검색엔진)을 구축하였다. 온톨로지의 성능 및 추론엔진의 성능을 평가하였으며, 온톨로지 개발, 추론엔진 적용 및 성능평가 비교과정을 간단히 그림으로 표현하면 〈그림 2〉와 같다.

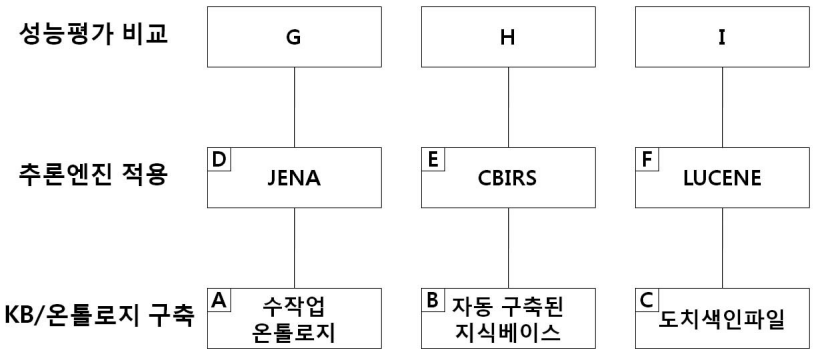
〈그림 2〉에서 보는 바와 같이 수작업으로 구축된 온톨로지(A)와 자동으로 구축된 의미망 지식베이스(B), 그리고 단순키워드검색을 위해 구축된 도치색인파일(C)로 기본적인 지식베이스를 구축한다. 온톨로지 A를 대상으로 제나 추론엔진(D)을 적용하고, 자동으로 구축된 의미망지식베이스에 개념기반 검색엔진(E)을 적용하며, 도치색인파일에는 루씬검색엔진(F)을 각각 적용한다. 이로부터 각각 G, H, I의 검색결과를 도출하여 그 성능을 재현율과 정확률로 평가한다.

지식베이스 및 추론엔진 또는 검색엔진의 성능을 평가하기 위해 개발된 질의어와 적합문헌 집단은 〈표 1〉과 같다.

〈표 1〉의 질문번호 1을 DL 쿼리로 생성한 예는 다음과 같다.

```
hasKeyword value 온톨로지 and
hasAuthorNameKor exactly 김수경 and
hasJournalNameKor exactly 정보관리학회지
```

추론엔진은 OWL 형식의 온톨로지를 로드해 일관성 검사를 수행한 후 클래스의 계층구



〈그림 2〉 온톨로지 개발, 추론엔진 적용 및 성능평가 비교과정

〈표 1〉 질의어와 적합문헌집단

번호	질의어	적합문헌
1	keyword - 온톨로지 author - 김수경 Journal name - 정보관리학회지	1124804, 978652, 898943
2	keyword - 자동분류 Journal name - 정보관리학회지	824144, 978655, 1256605, 937141, 824147, 853960
3	keyword - 저작권 year - 2007 Journal name - 정보관리학회지	937132, 824140, 824138, 824149
4	keyword - 이미지 검색 Journal name - 정보관리학회지	1013571, 898951, 898943, 1013570, 1319332
5	keyword - 도서관 마케팅 year - 2009 vol - 26 Journal name - 정보관리학회지	1256602, 1319341, 978651, 1211897 1256597, 1256587, 1168808
6	keyword - 기록물 keyword - 메타데이터 Journal name - 정보관리학회지	1124806, 1168801, 1168799, 853961 1319338, 1211894, 853962, 1256593
7	keyword - 직무분석 author - 안인자 Journal name - 정보관리학회지	978651, 1211892
8	keyword - 자동초록 Journal name - 정보관리학회지	1124813, 1256591, 937130, 1124808, 978646
9	keyword - 전문사서 Journal name - 정보관리학회지	978651, 1211892, 1256587, 1168805
10	keyword - 대학도서관 author - 노영희 Journal name - 정보관리학회지	937145, 1061190, 1211892
11	author - 이재윤 Journal name - 정보관리학회지	824144, 978653, 1124818, 898953, 937141, 937142 1124816, 898939
12	year - 2008 keyword - 공공도서관 Journal name - 정보관리학회지	1061189, 978647, 1061199, 1124803
13	keyword - 전자저널 Journal name - 정보관리학회지	1124810, 824148
14	keyword - 도서관 2.0 Journal name - 정보관리학회지	1256585, 1168802, 1168808, 1168803
15	keyword - web 2.0 Journal name - 정보관리학회지	1319331, 1256585, 824136, 1168802, 1211896 898947, 1061196, 1211893, 898943, 1319329, 978652 1319327
16	keyword - 국가지식정보 Journal name - 정보관리학회지	898950, 978650, 1319338, 1013578, 1256592
17	keyword - 커뮤니티 Journal name - 정보관리학회지	937128, 824143, 1211893
18	keyword - 오픈액세스 Journal name - 정보관리학회지	937132, 1319328, 1319335, 937128, 824140, 1168795
19	keyword - 학제성 keyword - 연구자 Journal name - 정보관리학회지	1168800, 1124818, 1211895, 898945
20	keyword - 디지털도서관 Journal name - 정보관리학회지	1168808, 937139, 1061188, 824146, 1168803, 1256597, 1256585, 1319328, 853959, 1061181, 1013572, 1124802, 853961, 937130, 1124808

조를 추론하고 클래스에 대하여 기술된 객체속성을 추론한다. 1차적으로 제 1키워드를 가지로 검색한 논문주석정보 데이터들로부터 생성한 OWL 논문온톨로지를 대상으로 DL 쿼리에 대한 2차 추론을 하여 사용자 질의에 부합하는 논문의 식별자(Article_ID)를 결과처리기에 반환한다.

개념기반검색의 경우 확장시에 확장되는 개념의 수와 확장되는 개념간의 유사도를 지정해 주어야 하는데, 본 연구에서는 확장되는 개념수의 기준을 8로 제한하였고, 유사도 기준은 0.5로 하였다.

4.2 시스템 구조

본 논문에서의 온톨로지기반의 시맨틱 추론의 목적은 온톨로지로부터 생성된 메타데이터에 대해 OWL에서 제공하는 공리를 사용하여 추론하고 이러한 추론된 사실을 바탕으로 사용자로부터 입력된 질의문으로 시맨틱검색을 수행하기 위함이다. 본 논문에서의 시맨틱검색은 'Description Logic' 기반의 온톨로지를 바탕으로 생성된 메타데이터에 대해 전방향 추론기법을 사용하여 검색하기 때문에 기존의 키워드검색이나 일반적인 질의응답시스템에서 찾기 힘든 정보에 대한 검색이 가능하다. 본 논문에서 구축된 시스템의 추론엔진 부분은 HP 연구소에서 개발된 제나 시맨틱 웹 라이브러리 중에서 일반적인 추론 기능을 제공하는 추론 모듈(Generic Rule Reasoner)을 활용한다. 제나는 RDF형태의 문서의 변환 및 추론에 용이한 모델을 제공하므로, OWL 형태의 온톨로지를 제나의 모델로 생성하고 일반적인 규칙 형태로 OWL 공리를 표현하여 온톨로지와 메타데이터를 융합하여 추론을 수행한다.

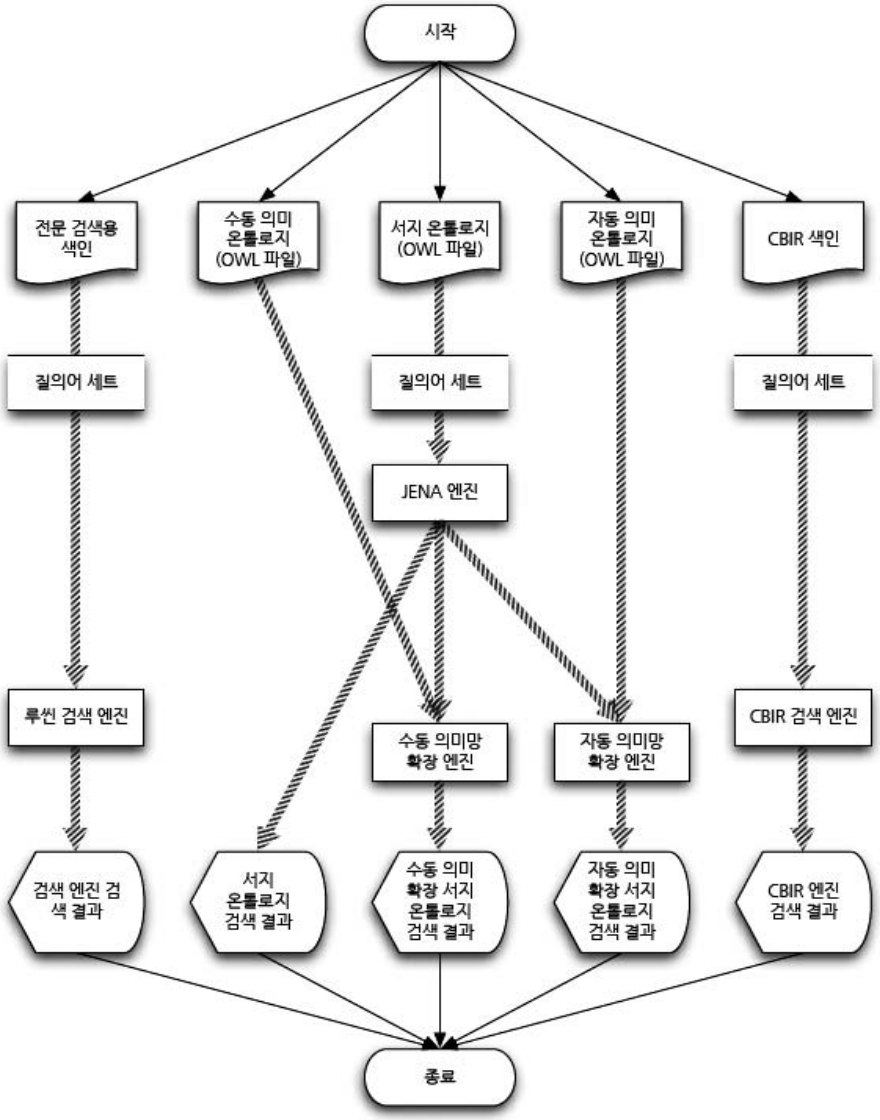
본 연구에서는 시스템 구현 단계가 크게 추론 및 검색에 이용되는 각종 색인 및 의미망 등을 구축하는 검색준비 단계와 실제 추론 및 검색을 하는 검색 단계로 나눠진다.

먼저 검색준비 단계로서, 이 단계에서 입력으로 필요한 것은 원본 서지정보이며, 이 정보를 가공하여 전문 검색용 색인, 서지온톨로지 OWL 파일, 자동 의미 온톨로지 OWL파일 및 CBIRS색인이 생성된다. 다만, 자동생성된 자동 의미 온톨로지와 비교를 위하여 수작업으로 작업한 수동 의미 온톨로지 OWL 파일이 검색단계에서 참조된다.

데이터의 흐름 관점에서 살펴보면 원본 서지정보는 정규화된 형식의 서지정보로 우선 만들어 지며, 이는 각각 도치색인 정보인 전문검색용 색인을 생성하기 위한 루씬색인생성기에 입력으로 활용된다. 또한 동시에 서지온톨로지 OWL 파일을 생성하기 위한 서지온톨로지 생성기의 입력으로 활용된다. 마찬가지로 정규화된 서지정보는 개념기반 정보검색(CBIR)을 위한 CBIR용 기초 데이터 생성기의 입력으로 활용된다. 이렇게 생성된 CBIR의 기초 데이터는 CBIR용 기초데이터 생성기에 이용되어 CBIR 색인 결과가 나온다. CBIR 색인 결과는 자동의미 온톨로지 OWL파일 생성을 위한 자동의미 온톨로지 생성기의 입력에 사용된다.

위의 결과로 생성된 자료들은 다음의 검색단계에서 각각 이용된다. 〈그림 3〉은 본 시스템에서의 추론방식 및 검색과정에 대한 개념도이다.

전문검색용 색인은 동일한 질의어 세트를 사



〈그림 3〉 전체적인 시스템 구성도

용하여 도치색인 검색엔진인 루씬검색엔진의 결과를 얻어낸다. 비슷한 방법으로 CBIR용 색인은 같은 질의어 세트를 사용하여 CBIR용 검색엔진의 검색 결과를 찾는다.

서지온톨로지인 경우 제나추론엔진만을 활용한 서지온톨로지 검색 결과를 얻는 경우와, 수동 의미망으로 질의어를 확장하여 검색에 활용한 수동 의미 확장 서지온톨로지 검색 결과, 그리고 자동 의미망으로 질의어를 확장하여 검색에 활용한 자동 의미 확장 서지온톨로지 검색 결과 등으로 나뉘어진다.

4.3 성능평가 결과

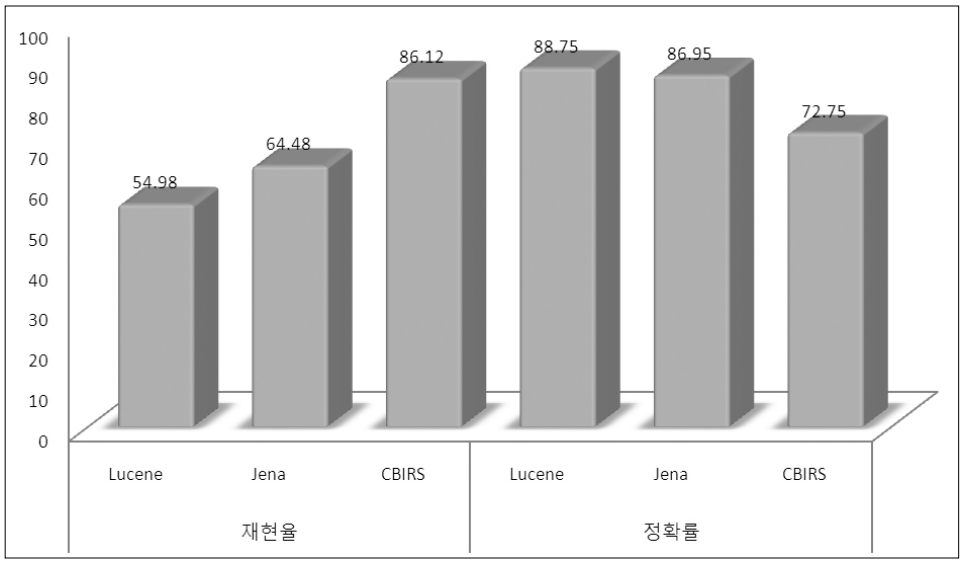
본 연구에서는 위에서 구축된 검색엔진의 성능을 가장 일반적인 평가방법으로 사용되고 있는 재현율과 정확률로 평가하였다. 세 개의 검색엔진의 성능을 평가한 결과는 다음 〈표 2〉, 〈그림 4〉와 같다.

먼저, 재현율에 있어서는 86.12%의 성능으로 개념기반검색기법이 가장 높은 성능을 보여준 것으로 나타났고, 다음으로 시맨틱검색이 64.48%로 높게 나타났으며, 키워드검색이 54.98%로 가장 낮게 나타났다. 이와 같은 재현율의 차이는 온톨로지 또는 의미망지식베이스를 기반으로 개념확장을 했느냐 하지 않았느냐의 차이로 보여진다.

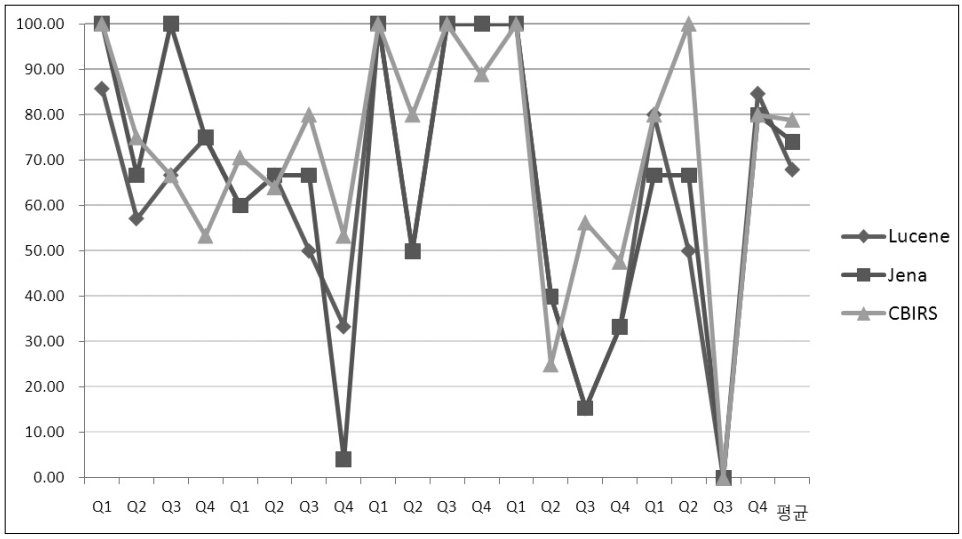
정확률에 있어서는 키워드검색과 시맨틱검색이 비슷한 성능을 보여주고 있고 CBIRS 검색은 가장 낮은 성능을 보여주고 있다. 이는 키워드검색의 경우 완전일치 되는 문헌만을 'AND' 조합에 의해 검색하기 때문에 매우 높은 정확률을 보여주는 것이고, 시맨틱검색은 온톨로지를 기반으로 비교적 정확한 개념확장을 하기 때문인 것을 알 수 있다. 반면에 CBIRS의 경우 확장되는 개념의 수 및 유사도를 기반으로 하며, 초기의

〈표 2〉 검색엔진의 성능평가 결과

Lucene			Jena			CBIRS		
재현율	정확률	F1척도	재현율	정확률	F1척도	재현율	정확률	F1척도
100	75	85.71	100	100	100.00	100	100	100.00
66.67	50	57.14	100	50	66.67	100	60	75.00
50	100	66.67	100	100	100.00	50	100	66.67
60	100	75.00	60	100	75.00	80	40	53.33
42.86	100	60.00	42.86	100	60.00	85.71	60	70.59
50	100	66.67	50	100	66.67	100	47.06	64.00
50	50	50.00	50	100	66.67	100	66.67	80.00
20	100	33.33	20	2.27	4.08	80	40	53.33
100	100	100.00	100	100	100.00	100	100	100.00
33.33	100	50.00	66.67	40	50.00	66.67	100	80.00
100	100	100.00	100	100	100.00	100	100	100.00
100	100	100.00	100	100	100.00	100	80	88.89
100	100	100.00	100	100	100.00	100	100	100.00
25	100	40.00	25	100	40.00	75	15	25.00
8.33	100	15.38	8.33	100	15.38	75	45	56.25
20	100	33.33	20	100	33.33	100	31.25	47.62
66.67	100	80.00	66.67	66.67	66.67	66.67	100	80.00
33.33	100	50.00	50	100	66.67	100	100	100.00
0	0	0.00	50	100	0.00	50	100	0.00
73.33	100	84.61	80	80	80.00	93.33	70	80.00
54.98	88.75	67.90	64.48	86.95	74.05	86.12	72.75	78.87



〈그림 4〉 검색엔진의 평균성능 비교(재현율, 정확률)



〈그림 5〉 검색엔진의 평균성능 비교(F1-measure)

개념이 1개에서 출발하지만 8개까지 확장될 수 있기 때문에, 확장된 용어 중에 관련 없는 개념이 일부 포함되어 정확률이 약간 떨어지는 것으로 분석되었다. 그러나 전반적으로 보았을 때는 개념기반검색의 성능이 약간 높은 것을 알 수 있다.

이는 F_1척도를 사용하여 확인할 수 있는데, 일반적으로 F_1척도를 이용할 경우 값이 클수록 좋은

검색결과가 된다. F_1척도의 수식은 다음과 같다.

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$$

〈그림 5〉에서 보는 바와 같이 루씬, 제나, CBIRS의 F값의 평균은 각각 67.90%, 74.05%, 78.87%로 나타났으며, CBIRS가 가장 높은 성능을 보여주었고, 루씬이 가장 낮은 성능을 보여주었다. 즉, CBIRS는 재현율에 있어서 가장 높은 성능을 보여 주었고, 정확률에 있어서는 가장 낮은 성능을 보여주었지만, 재현율과 정확률의 비중을 동일하게 하였을 때, 가장 높은 성능을 나타낸 것을 알 수 있다.

5. 논의 및 결론

본 연구는 지금까지 다양한 각도에서 연구되어온 온톨로지 및 추론엔진에 관한 연구를 일반화할 수 있는 방안을 제시하는데 있다. 특히 유동성이 크고 데이터의 규모도 상당한 도서관에 일반화 시켜 적용하는 것이 필요한데, 지금까지 이론적으로만 연구될 수밖에 없었던 문제점들을 발견하고 개선해 보고자 하였다.

이를 위해 실험을 위한 온톨로지를 구축하였으며, 온톨로지 구축대상은 정보관리학회지 2007년부터 2009년까지의 3년간의 논문기사를 대상으로 하였으며, 구축방법은 온톨로지 구축도구를 이용한 수작업에 의한 구축방법과 알고리즘을 이용한 자동적인 구축방법으로 나뉜다. 또한 온톨로지 및 추론엔진의 성능을 비교·평가하였다. 특히 본 연구에서는 지금까지 연구되고 적용되어 온 온톨로지 구축방법을 이용하되, 구축된 온톨로지로부터 이용자의 요구에 적합한 자료를 탐색하는 검색기법의 적용에 있어서, 개념기반 정보검색기법(Concept-based Information Retrieval Techniques)을 적용하였다. 즉, 본 연구에서는 개념기반 정보검색기법을 시맨틱 웹 검색기법인 제나와 그 성능을 비교함으로써 실제 적용성 및 검색 성능의 향상을 도모하고자 하였다.

본 연구에서는 실험을 위해 지식베이스와 검색엔진을 구축 및 개발하였다. 총 세개의 지식베이스(트리플구조 온톨로지, 의미거리기반 의미망지식베이스, 키워드중심의 도치색인파일)를 구축하였고, 이의 성능을 측정하기 위해 각각 세 개의 검색엔진(추론규칙 기반 제나검색엔진, 개념기반 검색엔진, 키워드기반 루씬검색엔진)을 구축하였다.

시스템 성능평가 결과 재현율에 있어서는 개념기반검색기법이 가장 높은 성능을 보여준 것으로 나타났고, 다음으로 시맨틱검색기법으로 나타났으며, 키워드검색기법이 가장 낮게 나타났다. 이와 같은 재현율의 차이는 온톨로지 또는 의미망지식베이스를 기반으로 개념확장을 했느냐 하지 않았느냐의 차이로 분석된다.

정확률에 있어서는 키워드검색과 시맨틱검색이 비슷한 성능을 보여주고 있고 CBIRS 검색은 가장 낮은 성능을 보여주고 있다. 이는 키워드검색의 경우 완전일치 되는 문헌만을 'AND' 조합에 의해 검색하기 때문에 매우 높은 정확률을 보여주는 것이고, 시맨틱검색은 온톨로지를 기반으로 비교적 정확한 개념확장을 하기 때문인 것을 알 수 있다. 전반적으로 보았을 때는 개념기반검색의 성능이 약간 높은 것을 알 수 있다.

또한, 세 검색엔진의 F값의 평균은 각각 67.90%, 74.05%, 78.87%로, 재현율과 정확률의 비중을 동일하게 하였을 때, CBIRS가 가장 높은 성능을 나타낸 것을 알 수 있다.

본 연구 결과는 유동적인 지식베이스 환경에서 높은 검색성능을 원하는 검색시스템에 적용할 수 있는 개념기반검색시스템의 성능을 온톨로지기반으로 구축된 시맨틱검색엔진인 제나와 비교하였으며, 그 성능에 있어 비교적 높은 것을 증명하였다. 따라서 동적환경에서 정보서비스를 제공하는 도서관, 정보센터, 전문연구기관 등의 서비스에의 범용적 적용가능성과 유용성을 높이는데 기여하게 될 것이다. 특히 서비스 대상 도메인이 변경되어 의미 검색 대상자원이 변경되었을 경우에도 기 구축된 언어자원의 의미망이 자동으로 최적화될 수 있도록 할 것이다. 또한 기존의 키워드 매칭방식의 검색기술이 제공하지 못했던 의미적 연관정보를 사용자에게 제공함으로써 향후 전자도서관 활성화에 기여할 수 있을 것이다.

참 고 문 헌

강래구. 2007.『시맨틱 웹 환경에서 온톨로지기반의 지능형 상품 검색 시스템 설계 및 구현』. 석사학위논문. 조선대학교.

김영민, 이상준. 2003. 시맨틱을 이용한 연구 논문 검색 시스템.『한국인터넷정보학회』, 4(3): 15-22.

김태환, 전호철, 최중민. 2008. 시맨틱 웹 서비스 기반 커뮤니티 정보 검색 시스템.『한국컴퓨터종합학술대회 논문집』, 35(1): 299-304.

박종욱. 2008.『온톨로지기반 검색을 이용한 지능형 통계 검색 모델에 관한 연구』. 석사학위논문. 공주대학교.

박지형, 박상언, 이명진, 홍준석, 김우주. 2007. 다중 온톨로지를 이용한 시맨틱 웹 포털에서의 의미형 검색.『한국지능정보시스템학회』, 11: 463-467.

박진석, 양기철, 오정진. 2004. 시맨틱 웹 기반 박물관 유물 검색을 위한 온톨로지 설계 및 구현.『한국콘텐츠학회』, 2(2): 269-274

심재문. 2008.『온톨로지 추론엔진 성능 평가 및 지능형 엔진 선택 기법에 대한 연구』. 석사학위논문. 경희대학교.

이재원, 박성찬, 이상근, 박재휘, 김한준, 이상구. 2010. 개념 망을 통한 전자 카탈로그의 시맨틱검색 및 추천.『한국전자거래학회』, 15(3): 131-145.

정은경. 2003.『시맨틱 웹 환경에서의 온톨로지기반 정보검색 시스템』. 석사학위논문. 제주대학교.

하상범, 한은영, 최호준. 2005. OWL 기반의 SPARQL을 이용한 시맨틱검색.『한국정보과학회』, 32(2): 706-708.

한동일, 권혁진, 정학진. 2007. 시맨틱검색시스템의 구현과 평가에 관한 연구.『한국IT

서비스학회』, 7(3): 253-269.

Carroll, J. J., L. Dickinson, D. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinosn. 2004. "Jena: Implementing the Semantic Web Recommendations." *Proceedings of the 13th International World Wide Web Conference*, New York. 74-83.

Chen, H. and V. Dhar. 1991. "Cognitive Process as a Basis for Intelligent Retrieval Systems Design." *Information Processing & Management*, 27(5): 405-432.

Christophides, V., et al. 2003. "The ICS-FORTH SWIM: A Powerful Semantic Web Integration Middleware." *In Proceedings of the First International Workshop on Semantic Web and Databases(SWDB)*, Co-located with VLDB 2003.

Haarslev, V., R. Moller, and M. Wessel. 2004. "Querying the Semantic Web with Racer + nRQL." *Proceedings of the Ki-04 Workshop on Applications of Description Logics.*

Karvounarakis, G., A. Magganaraki, S. Alexaki, V. Christophides, D. Plexousakis, Michel Scholl, and Karsten Tolle. 2003. "Querying the Semantic Web with RQL." *Computer Networks*, 42(5): 617-640.

Kevin, W., C. Sayers, and H. Kuno. 2003. "Efficient RDF Storage and Retrieval in Jena 2." *Proceedings of First International Workshop on Semantic Web and Databases*, 131-151.

Lee, M. C., H. K. Jan, Y. S. Paik, S. E. Jinf, and S. Lee. 2006. "A Ubiquitous Device Collaboration Infrastructure: Celadon." 3rd Workshop on Software Technologies for Future Embedded & Ubiquitous Systems.

Ma, L., G. Xie, T. Yang, and L. Zhang. 2006. "IODT: IBM Integrated Ontology Development Toolkit." [cited 2009.12.16]. Available at: 〈http://www.alphaworks.ibm.com/tech/semanticsik〉, 2004.

Maedche A, B. Motik, L. Stojanovic, R. Studer, and R. Volz. 2003. "Ontologies for Enterprise Knowledge Management." *IEEE Intelligent Systems*, 18(2): 26- 33.

Noh, Younghee. 2001. "A Study on the Estimation of Performance of the Concept-based Information Retrieval Model for Searching the Web." *Journal of Information Science*, 28(5): 407-415.

Noh, Younghee. 2011. "A Study on Constructing the Ontology of LIS Journal." *Journal of the Korean Society for Information Management*, 28(2): 177-193.

Pan, J. 2006. "HAWK: OWL Repository and Toolkit 1.5 Releases." [cited 2009.10.6]. Available at: 〈http://swat.cse.lehigh.edu/downloads/hawk.html〉.

Parsia, B., B. Sirin, M. Grov, and R. Alford. 2003. "Mindswap Project: Pellet." Available at:

〈http://www.mindswap.org/2003/pellet/〉.

Prud'Hommeaux, E. and A. Seaborne. 2006. SPARQL Query Language for RDF. W3C Working Draft 4.

Salton, G., Edward A. Fox, and Harry Wu. 1983. "Extended Boolean Information Retrieval." *Communications of ACM*, 26(12): 1022-1036.

Seaborne, Andy. 2003. A Query Language for RDF. W3C Member Submission 9-January-2004. [cited 2009.10.15]. Available at: 〈http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/〉.

Sirin, E., B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz. 2007. "Pellet: A Practical OWL-DL Reasoner." *Journal of web Semantics*, 5(2): 1-26.

Wessel, M. and R. Moler. 2005. "A High Performance Semantic Web Query Answering Engine." International Workshop on Description Logics(DL2005), Edinburgh, Scotland, UK.

Zhou, J., L. Ma, Q. Liu, L. Zhang, Y. Yu, and Y. Pan. 2006. "Minerva: A Scalable OWL Ontology Storage and Inference System." The Semantic Web-ASWC 2006, Volum LNCS 4185, 429-443.