검색 성능 향상을 위한 약품 온톨로지 기반 연관 피드백

Relevance Feedback based on Medicine Ontology for Retrieval Performance Improvement

임 수 연(Soo-Yeon Lim)*

초 록

기계가 정보의 의미를 이해하고 처리할 수 있도록 기존의 웹을 확장하는 것을 목적으로 하는 시멘틱 웹은 온톨로지를 이용하여 지식을 공유하게 된다. 본 논문에서는 정교한 질의의 처리를 위하여 온톨로지 내에 존재하는 의미 관계들을 질의의 확장을 위한 연관피드백 정보로 이용하는 방안을 제안한다. 실험은 도메인 온톨로지인 Medicine 온톨로지를 대상으로 하였으며, 출현 용어들의 빈도정보만을 이용한 키워드 기반 문서검색과 제안한 온톨로지기반 문서검색의 성능을 비교하였다. 이 때, 두 시스템의 정확률과 재현율을 성능 평가의 기준으로 삼았다. 그 결과, 검색 엔진은 온톨로지에 정의된 개념들과 규칙들을 활용하면서 검색의 정확률을 향상시키는데 도움이 되었고 검색 성능을 향상시키기 위한 추론의 기반으로도 사용될 수 있었다.

ABSTRACT

For the purpose of extending the Web that is able to understand and process information by machine, Semantic Web shared knowledge in the ontology form. For exquisite query processing, this paper proposes a method to use semantic relations in the ontology as relevance feedback information to query expansion. We made experiment on pharmacy domain. And in order to verify the effectiveness of the semantic relation in the ontology, we compared a keyword based document retrieval system that gives weights by using the frequency information compared with an ontology based document retrieval system that uses relevant information existed in the ontology to a relevant feedback. From the evaluation of the retrieval performance, we knew that search engine used the concepts and relations in ontology for improving precision effectively. Also it used them for the basis of the inference for improvement the retrieval performance.

키워드: 온톨로지, 의미관계, 문서검색, 질의 확장
ontology, semantic relation, document retrieval, query expansion

---

* 경북대학교 컴퓨터공학과 연구원(nadalsy@hotmail.com)
■ 논문접수일자 : 2005년 4월 19일
■ 게재확정일자 : 2005년 6월 20일

1. 서 론

시멘틱 웹(semantic web)은 기계가 정보의 의미를 이해하고 처리할 수 있는 거대한 정보의 공간으로 정의할 수 있다. 즉, 사람의 머리에 있는 거대한 세상에 대한 정보들을 컴퓨터 언어를 이용하여 표현하고 이것을 컴퓨터가 사용할 수 있게 만드는 것이며, 특별히 분산 환경을 갖춘 웹에 구현하자는 시도이다.

기계가 정보의 의미를 이해해야 하는 이유는 정보 검색과 같은 사람의 요구를 더 잘 이해하고 적절하게 반응하기 위한 것이다. 사람과 기계 사이에 커뮤니케이션이 가능하기 위해서는 사람이 이해하는 수준으로 기계도 세상을 이해할 수 있어야 한다. 사람들이 세상을 이해하는 방식을 개념화(conceptualization)라고 말한다.

기계가 정보의 의미를 이해하고 처리할 수 있도록 기존의 웹을 확장하는 것을 목적으로 하는 시멘틱 웹은 온톨로지(ontology)를 이용하여 지식을 공유하게 된다. 온톨로지는 넓은 의미에서는 데이터베이스라고 할 수 있지만, 데이터보다 복잡한 형태의 지식과 관련되어 있다는 의미에서 지식 베이스(knowledge base)라고 부르기도 한다. 그러나 이는 지식의 내용과 절차적인 추론 과정을 포함하는 포괄적인 의미의 지식 보다는 용어 사이의 개념적 관계에 국한되어 있기 때문에 지식 베이스와 구별되는 또 다른 형태의 데이터베이스라고 할 수 있다(김홍기 2000).

온톨로지는 주어진 응용도메인 의 특성을 나타내는 관련 개념들의 집합과 정의, 관계들로 이루어진다. 따라서 개념적이고 술어적인 혼란을 감소시킬 수 있다.

웹상의 문서가 증가함에 따라 사용자는 검색엔진(search engine)이 보여주는 수많은 검색결과들 중에서 자신이 원하는 정보를 찾기 위해 시간을 소비한다. 이런 경우 검색엔진이 온톨로지에 정의된 개념들과 규칙들을 활용하게 되면 중요한 정보가 있는 자원을 빠르게 찾아서 사용할 수 있고 자원을 찾는 정확도를 향상시킬 수 있다는 장점이 있다.

본 논문에서는 구축한 약품 온톨로지 내의 의미관계들을 연관 피드백 정보로 이용하여 검색의 성능을 향상시키는 것을 목적으로 한다.

2. 관련연구

2. 1 온톨로지

가장 본질적인 정의를 들어 온톨로지를 설명하면 "온톨로지는 공유하기(shared)위한 개념들의 개념화를 형식적(formal)이고 명백하게(explicit) 설명해놓은 명세서(specification)"라고 할 수 있다. 개념화는 어떤 현상에 대해서 관련이 있는 개념들을 식별하여 그 현상을 추상화한 모델로 설명하는 것이며, 명백하게 표현한다는 것은 개념들의 사용유형과 사용된 유형의 제약조건(constraints)을 명백하게 정의한다는 것을 의미한다. 형식적이라는 것은 기계가 읽을 수 있어야 한다는 것을 말하며, 공유는 온톨로지가 표현하는 개념이 개별적이 아닌 해당 그룹 구성원간의 합의된 지식에 바탕을 두고 있다는 것을 의미한다(박정오 2000).

온톨로지를 개발하는 목적은 정보를 상호교환하고 필요한 정보를 추출하기 위하여 정보의

구조를 사람들과 소프트웨어 에이전트들 간에 서로 공동 합의된 형태로 구성하고 이를 효율적으로 재사용(reusability)함으로써 다양한 애플리케이션으로 적용하도록 하는 것이다.

Guarino는 개념들의 분류에 따른 여러 가지 온톨로지의 종류를 제안하였다. 어떤 문제나 도메인에 독립적인 개념을 나타내는 Top-Level 온톨로지, 특정 도메인이나 일반적인 도메인의 개념을 나타내는 Domain 온토롤지, 주석 분석과 같은 특정 작업이나 문제해결과 같은 보편적인 작업을 위한 Task 온톨로지, 그리고 가장 일반적이고 공통적인 개념들, 즉 세상의 모든 사물들에 대한 일반적인 지식이나 개념들을 포함하는 온톨로지인 Generic(or Application) 온톨로지이다(Guarino 1998). 본 논문에서는 약품이나 병명, 증세 등과 관련된 약품 도메인 온톨로지인 Medicine 온톨로지를 대상으로 한다.

2. 1. 1 온톨로지의 표현언어

시멘틱 웹에서는 서로 다른 사전 체계를 가진 문서들 간의 변환이 쉽게 이루어지도록 개념간의 계층관계나 개념 정의간의 정합성 등을 자동으로 계산할 수 있는 온톨로지 기술언어를 제공하려고 하고 있다. 시멘틱 웹 환경에서 지식을 표현하기 위한 언어는 W3C(World Wide Web Consortium)에서 개발된 다양한 언어가 존재한다.

인간이 웹 문서를 읽어보는 것만을 목적으로 하는 HTML과 비교하여 XML은 문서의 구조와 태그를 자유로이 규정함으로써 표현의 유연성이나 확장성을 가진다. 그러나 XML은 임의의 문서구조를 지정할 수 있는 반면에 문서구조가 갖는 의미에 대해서는 정의하지 않으므로 표현한 문서의 의미를 해석하는 역할은 하지 않는다.

RDF는 XML의 한계를 극복하기 위해 W3C의 주도로 제정된 가장 기본적인 시멘틱 웹 언어로서 의미의 손상없이 응용 프로그램간의 정보가 교환되도록 해당 정보를 호환성있게 표현하는 프레임 워크를 제시하며 메타 데이터를 인코딩, 교환, 재사용할 수 있는 기반을 제공한다.

DAML+OIL은 DAML(DARPA Agent Markup Language) 프로그램의 DAML- ONT와 유럽에서 개발된 OIL(Ontology Inference Layer)의 결합을 통하여 만들어진 웹 온톨로지 언어이다. 이의 목적은 관심영역(domain)의 구조를 서술하기 위함인데 이러한 구조는 객체지향적인 방법으로 클래스(class)와 속성(property)을 써서 표현된다.

시멘틱 웹을 위한 온톨로지 언어의 가장 최근의 동향이며 DAML+OIL 언어가 계승 발전된 형태인 OWL은 정보 기술을 필요로 하는 커뮤니티 사이에서, 보다 고도화된 데이터의 통합과 상호 운용성을 제공하며 웹에 근거한 구조적인 온톨로지를 정의한다. OWL은 DAML+OIL의 네임스페이스와 속성 클래스 이름 등을 변경하고 RDF 및 RDF Schema의 변화를 수용하였다

2004년 2월에 W3C에서는 웹상의 데이터의 공유와 재사용을 제공하는 시멘틱 웹의 표준우로 개정된 자원 기술 프레임워크(RDF)와 웹 온톨로지 언어(OWL)를 인정하였다고 발표하였다. XML, RDF 그리고 OWL에 의하여 웹은 문서와 데이터가 공유 가능한 글로벌한 정보 공유 기반으로 되며, 보다 용이하게 그리고 신뢰성이 높은 정보의 검색과 재사용이 가능해진다.

2. 1. 2 온톨로지의 구축사례

이미 구축되어있거나 현재 갱신되고 있는 온톨로지의 예는 다음과 같다.

Mikrokosmos는 대규모의 실용적인 기계번역을 지향하고 있는 미국방성의 지원 아래 미국 뉴멕시코 주립대학에서 개발된 지식 기반의 기계번역 시스템으로 5,000여개의 개념과 7,000여 단어의 스페인어 사전을 구축하여 회사 간 인수, 합병에 관한 스페인어 기사에 대한 고품질의 의미 분석이 가능하다(Mahesh, K. 1996).

HowNet은 중영 기계번역 시스템의 개발을 위해 만들어진 중국어 온톨로지이다. 총 53,000개의 중국어 사전과 57,000개의 영어사전을 구축하고 있으나 중국어에 의존적인 것이 문제점이다(Dong, Z. 1999).

Cyc는 인공지능을 응용하여 인간과 같은 추론을 수행할 수 있게 하려는 목적으로 MCC(The Microelectronics and Computer Technology Corporation)에서 10년 전부터 구축한 일상의 상식(commonsense)들을 데이터베이스로 만든 것이다. 일반적인 개념 3,000개로 구성된 상위 Cyc 온톨로지 아래에 많은 양의 사실들이 연결되어 있는 구조이며, 약 10만개의 개념과 100만개의 사실, 규칙 등을 가지고 있다(Lenat, D. B. 1995).

WordNet은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스턴대학 인지과학연구실이 구축해온 단어간의 관계를 표현하는 영어어휘 데이터베이스로 자연언어처리와 정보검색의 여러 분야에서 널리 이용되고 있다. 단어 중심으로 표현되어 있는 사전과 달리, WordNet은 단어형이 아닌 단어의 의미를 구성요소로 하여 네트워크 형태로 구성되어 있다는 것이 특징이며, 현재 대략 14만 단어를 포함하고 있으며 다국어판의 구현도 시도되고 있다(Miller, G. A. 1990).

국내 온톨로지의 대표적인 구축사례로는 한국 전자 통신 연구소에서 만든 한국어 명사 어휘로 표현되는 개념들 간의 다양한 관계를 연결시켜 놓은 어휘 데이터베이스인 ETRI 명사 개념망이 있다. 현재 일반명사 약 5만 단어와 경제 명사 약 1만 5천 단어로 구성되어 있으며, 이를 확장하는 작업과 함께 동사 개념망을 구축중이다.

코어넷은 KAIST에서 일본국립국어연구소의 어휘 분류표에 근거하여 어휘 의미 속성 체계를 개념 체계로 설정하고, 단어들의 의미와 개념들을 연결한 것으로, 총 2,938개의 계층적 개념과 총 92,448개의 어휘의미가 구축외어 있다(최기선 2001).

이외에도 국내에서는 포항공대의 LIP 온톨로지(강신재 2002), 울산대학교의 UOU온톨로지(옥철영 2004), 한국어 명사 워드넷(문유진 1996)등의 다각적인 지식 베이스 구축 방법이 개발되고 있다.

2. 2 연관피드백

대부분의 정보검색 시스템의 경우, 처리속도와 반응시간을 줄이기 위하여 역화일(inverted file)구조를 이용하고 있다. 역화일 구조는 갱신비용이 많이 들며 수집한 데이터에만 의존하여 검색하기 때문에 대부분의 사용자들은 만족스러운 결과를 얻을 때까지 질의를 수정해가면서 탐색을 반복해서 해나가게 된다.

검색 결과를 개선하는 정보 접근 과정의 중

요한 부분은 질의 재형식화로 이를 위한 효과적인 방법으로는 연관 피드백이 알려져 있다. 이는 검색된 문서들이 사용자에게 적합한가에 대한 판단을 기반으로 하여 질의를 재구성함으로써 검색 성능을 향상시키는 것이 목적이다(Ricardo, B. Y. 1999).

원래 연관 피드백은 사용자가 질의와 관련된 소량의 문서들을 선택하면 시스템은 이를 수정한다. 사용자가 연관성 평가 집합을 선택하고 재검색 명령을 내리면, 시스템은 자동으로 질의에 대한 가중치를 재산정해서 검색을 다시 수행하거나 원래의 질의를 확장해간다.

질의를 수정하는 방법으로는 선택된 문서들로부터 단어나 구를 추출하여 질의를 확장하는 것이 일반적이다. 첫 번째 검색에서 초기 키워드로부터 검색된 문서와 검색되지 않은 문서들에 의하여 키워드들의 가중치를 재산정하고, 검색한 문서 중에서 적합하다고 생각되는 문서들에서 추출된 키워드를 사용자가 검색할 키워드에 추가해 나간다. 즉, 수정된 질의를 이용해서 새로운 문서가 검색되는 반복적인 상호작용을 연관 피드백이라고 한다. 따라서 원래의 문서들도 새로운 결과물로 나타날 수 있지만, 다른 순위로 나타나게 되는 것이다.

기존의 연관 피드백을 이용하는 시스템들은 여러 가지 방법들을 사용하고 있다. 하나는 사용자가 처음에 검색한 문서들 중에서 적합하다고 판단된 문서들로부터 추출한 키워드를 질의에 추가하여 사용하는 방법이다. 다른 하나는 사용자 키워드를 수정하는 과정에서 불리언 연산을 수행하는 방법이다. 연관 피드백을 이용하는 시스템들은 사용자가 정보를 검색하려고 질의를 수행했을 때 즉각 결과를 보여주지 못하는 단점이 있어 응답시간이 길어지는 단점이 있다. 대부분의 사용자는 연관 피드백에 익숙하지 않기 때문에 여러 번 연관성 평가를 수행하는 것은 어려운 일이다.

본 논문에서는 이러한 문제를 해결하기 위한 노력으로 온톨로지 내의 의미관계 정보들을 이용한 연관 피드백 방법을 제안하고자 한다. 제안한 시스템은 사용자 질의와 관련된 단어들을 온톨로지로부터 추출하여 이를 추가한 검색을 실시함으로써 응답시간이 짧아지고, 재현율은 유지하면서도 정확률은 향상됨을 알 수 있었다.

3. 문서검색 시스템

문서나 웹을 검색할 때 온톨로지를 사용하는 경우, 원하는 정보를 빨리 찾을 수 있으며 자원을 찾는 정확도를 높일 수 있는 장점이 있다. 검색엔진은 온톨로지에 정의된 개념들과 규칙들을 활용하면서 이를 검색의 성능을 향상시키기 위한 추론의 기반으로 이용한다.

특정 분야의 주요 문서 집합을 선정한 후, 이들 문서들의 내용을 분석하여 개념들을 추출하고 이들을 링크로 연결한 것이 온톨로지이며 개념 추출의 목적은 문서들을 가장 잘 대표할 수 있는 명사들을 추출하는 것이다. 특히 가중치가 부여된 온톨로지를 이용한 검색 시스템이나 질의응답 시스템의 경우에는 가중치에 따라 선별된 소수의 정보들만을 보여줌으로써 사용자의 판단에 도움을 줄 수 있다.

본 논문에서는 전문용어의 처리를 통하여 구축한 Medicine 온톨로지를 실험대상으로 하여 입력된 질의어에 해당하는 개념 뿐 아니라 그의

온톨로지내 하위 개념들까지 탐색하는 것을 목적으로 하는 검색을 진행하고자 한다.

온톨로지 내의 가중치 부여는 주어진 문서들 내에서 특정 단어가 얼마나 자주 사용되는가를 나타내는 출현 빈도(term frequency)로부터 유도될 수 있다. 빈도수가 지나치게 높거나 지나치게 낮은 단어들은 개념 추출 대상에서 제외한다. 즉, 문헌에서 빈번하게 나타나는 단어들의 기록인 불용어 리스트를 사용하여 고빈도어를 먼저 제거한 다음 나머지 단어들을 빈도수 순으로 배열함으로써 질의어에 대한 문서들의 연관 순위를 결정시킨다.

벡터 모델에서 계산되는 유사도의 질의 문서 순위 결정의 정확성은 문서와 질의를 표현하기 위해 사용되는 색인 방법뿐만 아니라 색인어에 가중치를 부여하는 기법에 의해 크게 영향을 받는다.

일반적으로 벡터 모델에서 가중치 부여를 위하여 단어들의 빈도수를 이용할 때, 대부분의 사용자들은 문서집합의 구성이나 검색환경에 대한 자세한 지식이 없으므로 자신의 검색 목적에 잘 맞는 질의를 작성하는 것이 매우 어렵다. 따라서 효과적인 검색을 위하여 질의를 재작성하게 되는데 여기에 사용자들은 많은 시간을 소비하게 된다.

3. 1 온톨로지를 이용한 질의 확장

질의 재작성 과정은 원래의 질의에 새로운 질의어를 추가하는 작업과 확장된 새 질의에 대한 가중치를 재계산하는 작업으로 이루어진다. 이 때 처음의 질의는 적합한 문헌을 검색하기 위한 초기 시도로 볼 수 있으며, 초기 검색 이후 검색된 문헌이 연관성 평가를 받고 좀 더 많은 연관 문헌을 검색하기 위하여 개선된 질의가 작성된다.

전통적인 $tf \cdot idf$ 방법을 개선하기 위한 연관 피드백은 작은 실험 문서집합을 대상으로 할 경우 정확률이 많이 개선되는 것으로 알려져 있다.

본 논문에서는 사용자 연관 피드백 과정에 온톨로지 내의 계층관계를 이용한다. 입력으로 들어온 질의어와 관련된 온톨로지 내의 하위 정보로 출현하는 용어들을 이용하여 질의를 확장하고, 재작성된 질의에 대한 가중치를 다시 계산한다. 이 때 온톨로지내의 노드를 탐색할 하위어 검색 레벨은 2로 정하였다.

예를 들어 온톨로지 내에서 노드 '중이염'에 대한 하위노드로 '삼출성 중이염'과 '급성 삼출성 중이염'이 존재한다고 가정하자. 입력으로 질의어 {중이염}이 들어온 경우, 온톨로지를 탐색한 후 질의어 집합은 {중이염, 삼출성중이염, 급성 삼출성중이염}으로 확장되고 이들의 가중치를 기반으로 유사도를 다시 계산하게 되는 것이다.

다음의 〈그림 1〉은 질의의 확장에 이용할 OWL로 표현된 Medicine 온톨로지의 일부분을 보여주고 있다.

3. 2 문서의 순위화

검색 시스템은 검색된 문서들 각각에 대하여 순위 결정을 적용한다. 문서 순위 결정 방법은 문서와 질의 사이의 관련 정도를 나타내는 유사도를 계산하고, 계산된 유사도에 따라 문서들에 순위를 부여한다. 높은 순위를 갖는 문서일수록

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
   <!ENTITY owl  "http://www.w3.org/2002/07/owl#" >
   <!ENTITY xsd  "http://www.w3.org/2001/XMLSchema#" >]>
<rdf:RDF
 xmlns:owl = "http://www.w3.org/2002/07/owl#"
 xmlns:rdf = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:rdfs= "http://www.w3.org/2000/01/rdf-schema#"
 xmlns:xsd = "http://www.w3.org/2001/XMLSchema#">
 <owl:Ontology rdf:about="">
   <rdfs:comment>Medicine OWL ontology</rdfs:comment>
   <rdfs:label>Medicine Ontology</rdfs:label>
 <owl:Class rdf:ID="Medicine">
   <rdfs:subClassOf rdf:resource="eating" />
  <rdfs:subClassOf>
   <owl:Restriction>
    <owl:onProperty rdf:resource="#productedBy" />
       <owl:allValuesFrom rdf:resource="#Company" />
       <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger"> 1
       </owl:cardinality>
   </owl:Restriction>
  </rdfs:subClassOf>
          :
   <rdfs:label xml:lang="en">Medicine</rdfs:label>
  <rdfs:label xml:lang="kr">약품</rdfs:label>
 </owl:Class>
          :
<owl:Class rdf:ID="급성중이염">
  <rdfs:subClassOf rdf:resource="#중이염" />
 </owl:Class>
 <owl:Class rdf:ID="급만성기관지염">
  <rdfs:subClassOf rdf:resource="#기관지염" />
 </owl:Class>
 <owl:Class rdf:ID="중이염">
  <rdfs:subClassOf rdf:resource="#염" />
 </owl:Class>
```

〈그림 1〉 OWL로 표현된 Medicine 온톨로지

질의에 대한 만족도가 크며, 사용자는 높은 순위를 갖는 문서를 우선적으로 검토함으로써 필요한 정보를 얻는 데 소모되는 시간을 최소화할 수 있다(Ricardo, B. Y. 1999).

벡터 모델에서 용어 문헌 쌍 (k_i, d_j) 의 가중치 $w_{i,j}$ 는 양의 비이진 값이며, 질의 색인어도 가중치를 가진다. $[k_i, q]$ 의 가중치를 $w_{i,q} \geq 0$ 이라 하면 질의 벡터 $\vec{q}$ 는 $\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$ 로 정의되며, 여기서 t 는 시스템 내의 전체 색인어 수이다. 문헌 d_j 는 $\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$ 로 표현된다.

따라서 벡터 모델에서 t 차원 벡터로 표시된

문헌 d_j와 사용자 질의 q의 유사도 측정은 두 벡터 $\overrightarrow{d_j}$와 $\overrightarrow{q}$의 상관도로 구할 수 있으며, 이는 두 벡터간 사이각의 코사인 값으로 정량화 할 수 있다. $w_{i,j}$와 $w_{i,q}$가 0보다 크거나 같은 값을 갖기 때문에 $sim(d_j, q)$의 값은 0과 1사이의 값이 된다.

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

시스템 내의 총 문헌 수를 N이라 하고 색인어 k_i가 출현한 문헌 수를 n_i라 하자. 문헌 d_j에서의 용어 k_i의 출현 빈도수를 $freq_{i,j}$라 할 때, 문헌 d_j에서의 용어 k_i의 정규화 빈도는 다음과 같다.

$$tf_{i,j} = \frac{freq_{i,j}}{\max_i freq_{i,j}}$$

여기서 최대값 max는 문헌 d_j 텍스 내에 출현한 모든 용어 중에서 가장 빈도수가 큰 용어이며, 용어 k_i의 역문헌 빈도수 idf_i는 다음과 같다.

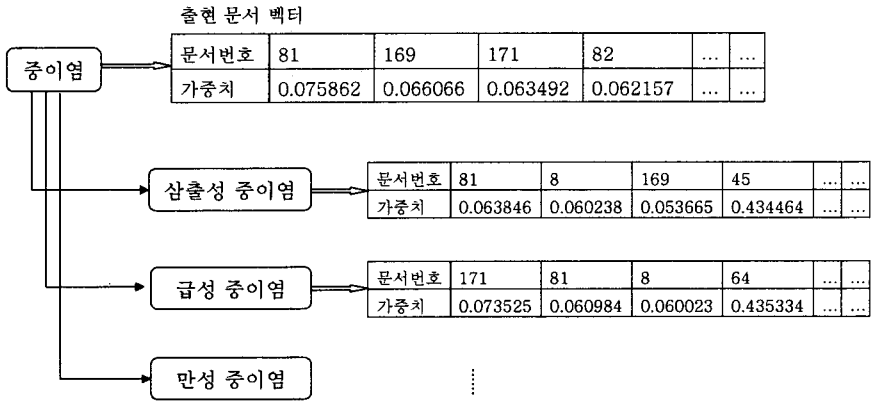
$$idf_i = \log \frac{N}{n_i}$$

가중치의 부여는 가장 널리 알려진 $tf \cdot idf$ 기법(용어-가중치 할당 전략)을 이용하여 계산한다.

$$w_{i,j} = tf_{i,j} \times idf_i$$

계산된 가중치는 출현 문서 벡터에 문서번호와 함께 정렬된 순으로 저장됨으로써 검색의 속도를 향상시키고 더 정확한 검색을 가능하게 해준다.

〈그림 2〉는 출현 문서 벡터의 구조를 보여주고 있다.



〈그림 2〉 출현 문서 벡터의 구조

제안한 시스템의 검색 성능은 사용자의 질의에 대한 재현율과 정확률을 구함으로써 평가할 수 있다.

〈그림 3〉은 지금까지 설명한 문서 검색을 위한 전체 시스템의 구성도를 보여준다. 제안된 시스템은 크게 전처리 모듈과 검색 모듈로 구성된다.

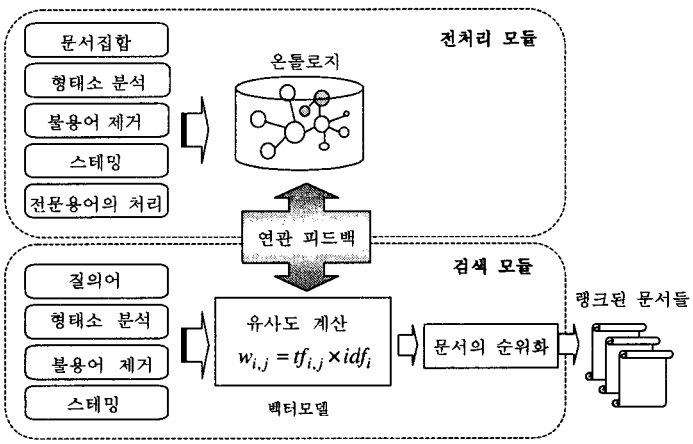
먼저 전처리 모듈에서 대상문서들에 대한 색인어 집합을 구성하기 위하여 형태소 분석 과정을 거친다. 그 결과 중에서 명사만을 추출하여 색인어 집합으로 사용하게 되는데 명사는 정보검색이나 분류에서 문서를 대표할 수 있는 통계적 정보를 얻는데 주로 사용된다. 이 시스템의 경우에는 온톨로지가 검색을 위한 색인어 집합의 역할을 하게 된다. 문서에서 추출되어진 단어들은 불용어를 포함하고 있다.

불용어란 색인할 때 무시될 수 있는 단어를 의미하며 대부분이 숫자, 기호나 기호를 포함하고 있는 무의미한 단어들로 이루어져 있으며, 띄어쓰기가 제대로 되어있지 않은 단어 등을 말하며 이들은 불용어 리스트로 만들어둔다.

실험 대상인 특정 도메인 내에서의 문서검색을 위한 질의어에는 전문 용어의 출현이 많고 의미가 가지지 않으며 자주 쓰이는 조사 등은 불용어 리스트에 의해 대부분 처리가 이루어진다. 이렇게 제거되어지고 남은 단어들로 온톨로지의 노드들을 설정한다. 본 논문에서는 181개의 용어로 이루어진 불용어 리스트를 이용하였으며 다음의 〈그림 4〉는 사용한 불용어 리스트의 일부분을 보여준다.

가	과의	그런줄
거나	그	그럴수록
건	그간	그로
결	그것	그무렵
것	그곳	그외
게	그당시	그이
고	그대신	그전
곳	그동안	그전날
곳곳	그들	그쪽
과	그때	까지 ...

〈그림 4〉 불용어 리스트



〈그림 3〉 문서검색 시스템의 전체 구성도

제안한 온톨로지는 또한 약품 검색을 위한 간단한 질의 응답 시스템에도 이용할 수 있다. 입력으로 들어온 병명이나 증세에 대하여, 기존의 질의응답 시스템은 해당되는 모든 약품명을 일정한 기준없이 나열하는 수준에 그치고 있다.

그러나 온톨로지를 이용한 질의 응답 시스템의 경우에는 처방빈도나 해당 약품의 효능을 고려한 가중치에 따라 선별된 소수의 약품명들만을 보여줌으로써 사용자의 판단에 도움을 줄 수가 있다.

간단한 질의 응답 시스템의 경우, 먼저 질의 유형을 정의하고 정의된 질의 유형에 기반하여 정답문장에 있는 적절한 개체명들을 탐색해 나간다. 질의로부터 키워드와 함께 나타나는 개체명을 찾은 뒤, 정답으로서 가장 좋은 개체명을 돌려준다. 질문과 답사이에는 형태소분석(lexical analysis), 구문분석(syntax analysis), 의미분석(semantic analysis) 등의 질의처리(query processing) 과정이 필요하다.

기존의 일반적인 질의응답 시스템에서는 〈효능, 효과〉 태그에 "중이염"이라는 키워드가 들어있는 약품들의 목록들을 추출하여 이들을 대개 가나다순으로 나열하여 보여주지만, 온톨로지의 계층관계를 이용하는 제안한 시스템의 경우에는 사용자가 입력한 질의어에 대하여 자동적으로 분류된 정보들을 제공받을 수 있으며 더욱 상세한 정보를 얻을 수 있다.

〈그림 5〉는 온톨로지기반 검색을 수행했을 때 하위 노드를 탐색함으로써 검색의 범위가 확장되어진다는 것을 보여주는 그림이다. 질의어 "중이염"은 병명을 나타내는 특정명사인 "염"과 결합된 전문용어로서 "염"과 하위어 관계로 연결되어 있다. 질의어 "중이염"에 대하여 일반적인 키워드기반 검색의 경우에는 ①의 네 개 노드들에 대한 탐색만 이루어지지만, 온토로지를 이용한 검색의 경우에는 질의에 대한 하위개념들인 삼출성중이염, 급성중이염, 만성중이염과 연결된 ②의 일곱 개 노드들에 대한 검색이 이루어진다. 이는 온톨로지의 계층관계를 질의의 확장에 이용했을 때 검색의 효율을 높일 수 있음을 의미한다.

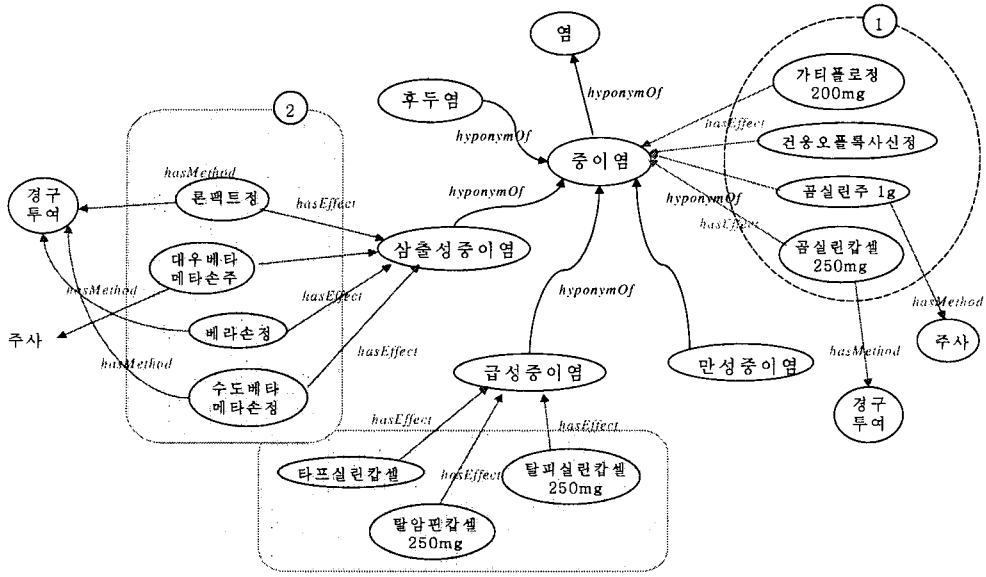
4. 실험 및 평가

구축한 온톨로지의 효용을 보이기 위하여 전통적인 $tf \cdot idf$ 방법을 이용하여 가중치를 부여하는 키워드 기반 문서 검색과 온톨로지 내의 하위 정보를 연관 피드백에 이용하고 가중치를 계산하는 온톨로지 기반 문서 검색의 결과를 비교 분석하였다.

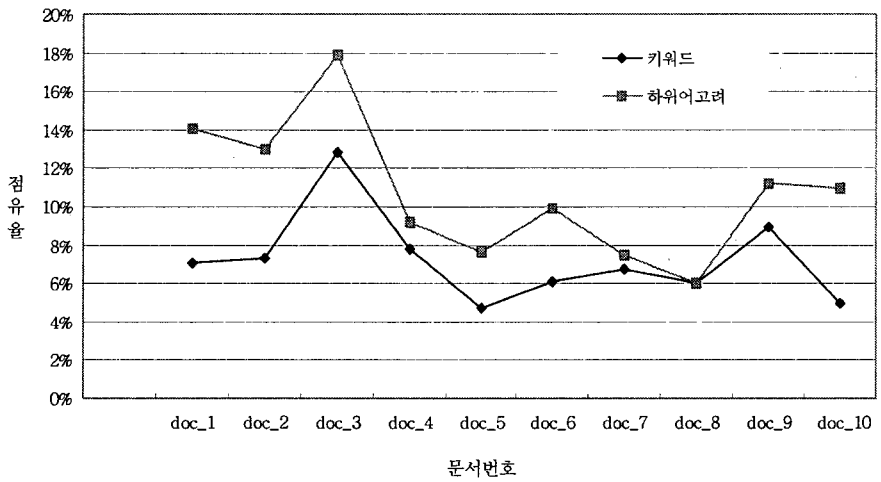
다음의 〈그림 6〉은 질의어가 "중이염"인 경우, 10개 문서들에 대한 점유율의 분포를 비교한 그래프이다. 질의어"중이염"의 경우 만성중이염", "만성 유착성 중이염"등과 같은 36개의 중이염의 하위 단어가 추가됨을 알 수 있었다. 이는 하위 단어로 검색레벨이 확장된다는 것이 검색의 정확률에 영향을 미칠 수 있음을 의미한다.

온톨로지를 이용한 벡터기반 검색모듈에서는 색인 파일을 구성하기 위하여 온톨로지 내의 노드들을 이용하고 출현 문서 벡터에 저장되어 있는 가중치를 유사도 계산에 이용함으로써 문서들의 순위화가 이루어진다.

검색 성능을 평가하기 위해서는 실험참조 컬렉션과 평가척도를 사용한다. 실험 참조 컬렉션



〈그림 5〉 온톨로지기반 검색시의 하위노드확장



〈그림 6〉 검색레벨 확장에 따른 점유율의 분포

은 문헌 집합, 정보 요구 예제, 각 정보요구에 대한 연관 문헌 집합으로 구성된다. 여기서 연관 문헌 집합은 전문가에 의해 제공되고, 평가 척도는 각 정보 요구에 의해 검색된 문헌 집합과 전문가에 의해 제공된 연관 문헌 집합 사이의 유사도로 계산된다.

본 논문에서는 참조 컬렉션을 구성하기 위해 대한의사협회 홈페이지(http://www.kma.org)에서 제공하는 건강/질병 정보 문서 750개를 수집하였으며 다음과 같은 10개의 질의로 구성된 정보 요구를 구성하였다.

질의 1 : {중이염}
질의 2 : {삼출성 중이염, 치료, 약}
질의 3 : {만성 중이염, 진단, 치료}
질의 4 : {고열, 병, 귀}
질의 5 : {중이염, 증상, 종류}
질의 6 : {고혈압, 당뇨병, 인슐린}
질의 7 : {피부, 발진, 감염 }
질의 8 : {소아, 발열, 감염경로}
질의 9 : {갱년기, 위장, 장애, 소화}
질의 10 : {기침, 두통, 콧물}

검색 성능의 비교는 추출한 750개의 문서를 대상으로 하였으며 전문가 5인의 자문을 구하였다. 10개의 질의 각각에 대하여 문서들을 분류하고 순위를 매긴 결과, 평균 30개의 문서들이 질의와 관련된 문서들로 판단됨에 따라 상위 30개의 문서들을 정답 문서 집합으로 정하였다. 따라서 본 논문에서는 상위 30위까지만 정확률을 계산하며, 적합문서 31위부터는 검색되지 않았다고 가정하고 정확률은 0으로 계산하였다. 이를 기준으로 각각의 질의에 대한 재현율과 정확률을 구하고, 전체 질의에 대한 평균 재현율과 정확률을 구하였다.

탐색 작업의 속도를 향상시키기 위해 텍스트에 대한 색인을 만들게 되며, 이들은 어휘와 출현빈도의 두 요소로 구성된다. 어휘는 텍스트에 나타나는 모든 단어들의 집합이며, 각 단어에 대한 출현 문서 벡터를 가지게 된다. 출현 문서 벡터는 빈도를 고려한 가중치와 함께 출현 문서의 위치를 저장하고 있다. 아래의 〈표 1〉은 질의1 : {중이염}에 대해 부여된 가중치와 이를 이용하여 계산된 상위 10개 문서와의 유사도를 보여주고 있다.

〈표 1〉 질의어 "중이염"에 대한 문서별 가중치

순위	문서번호	가중치	유사도
1	doc_81	0.086142	0.098016
2	doc_169	0.072683	0.097477
3	doc_82	0.061670	0.082352
4	doc_171	0.061206	0.078861
5	doc_48	0.050151	0.073900
6	doc_64	0.048935	0.072776
7	doc_381	0.046505	0.072516
8	doc_198	0.056599	0.072050
9	doc_430	0.040887	0.072012
10	doc_194	0.035664	0.071538

다음의 〈표 2〉과 〈표 3〉은 입력으로 들어온 10개 질의에 대한 재현율과 정확률의 분포를 각각 보여주고 있다.

비교한 두 가지 검색에 대하여 평균 재현율과 평균 정확률을 각각 구하고 이를 그래프로 나타낸 것이 〈그림 7〉이다.

이로부터 우리는 온톨로지 내에 존재하는 질의어와 관련있는 하위 노드 정보들을 질의의 확장에 이용하고 가중치를 부여하는 방법에 의한 온톨로지 기반 문서 검색이 전통적인 $tf \cdot idf$ 방법을 이용한 검색보다 재현율은 1.87%, 정확률은 7.9 % 향상됨을 알 수 있었다.

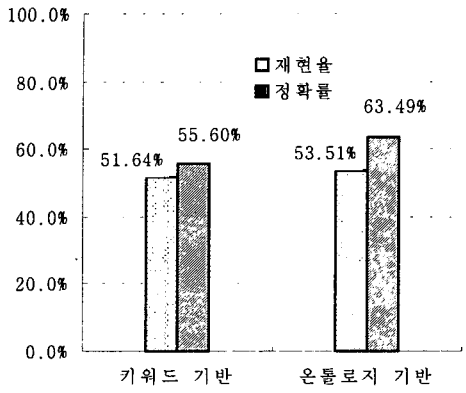
결과적으로 문서의 검색에 온톨로지 내의 의미 관계 정보들을 질의의 확장을 위한 연관 피드백 정보로 이용하면 재현율의 향상에는 별 영향을 주지 않지만 정확률을 향상시키는데 효용이 있어 사용자의 요구에 좀더 정확하게 응답할 수 있음을 알 수 있었다.

〈표 2〉 10개 질의에 대한 재현율의 비교

질의	키워드 기반	온톨로지 기반
질의1	45.52%	46.07%
질의2	52.62%	53.73%
질의3	50.33%	53.16%
질의4	53.58%	56.44%
질의5	55.61%	55.91%
질의6	51.01%	52.71%
질의7	44.59%	50.23%
질의8	52.89%	54.00%
질의9	60.55%	61.47%
질의10	49.68%	51.36%

〈표 3〉 10개 질의에 대한 정확률의 비교

질의	키워드 기반	온톨로지 기반
질의1	47.84%	55.16%
질의2	51.33%	59.15%
질의3	72.61%	86.31%
질의4	52.36%	68.61%
질의5	53.03%	59.56%
질의6	67.05%	72.06%
질의7	47.90%	62.00%
질의8	53.44%	56.71%
질의9	51.73%	54.78%
질의10	58.67%	60.58%



〈그림 7〉 평균 재현율과 평균 정확률의 비교

5. 결 론

시멘틱 웹의 출현과 더불어 온톨로지의 중요성이 인식되면서 온톨로지에 관한 연구는 인공지능 분야의 시작과 함께 지식 표현 분야의 핵심으로 활발히 연구가 이루어져 온 분야이다. 다수의 사용자가 광범위하고 손쉽게 때와 장소를 가리지 않고 공유할 수 있는 웹이라는 지식의 체계는 인류 문명이 추구해 온 기본적인 목표하고 할 수 있다. 시멘틱 웹 온톨로지는 다수에 의한 소규모 온톨로지의 개발을 추구하고 있다.

본 논문에서는 더욱 정교한 질의의 처리를 위하여 온톨로지 내에 존재하는 의미 관계 정보들을 연관피드백을 위한 질의의 확장에 이용하였다. 온톨로지내의 여러 의미관계들이 문서의 검색에 효용이 있음을 보이기 위하여 용어들의 출현빈도 정보만을 이용하여 가중치를 부여한 키워드기반 문서검색과 온톨로지 내의 관련 정보들을 연관 피드백에 이용한 온톨로지기반 문서검색을 비교하였다. 검색의 성능을 평가한 결과, 재현율의 변화는 거의 없었으나 정확률이 7.9% 향상되는 것을 알 수 있었다. 이는 온톨로지 내의 의미관계를 연관 피드백 정보로 이용하면 검색의 정확률을 향상시킬 수 있음을 의미한다.

주어진 도메인의 용어들의 정의와 그들 간의 관계를 나타내고 있는 온톨로지는 풍부한 시소러스로 볼 수도 있다. 질의의 확장을 위한 피드백 정보뿐 아니라 온톨로지 내에 존재하는 각종 메타 데이터들도 검색의 효율을 높이기 위한 방안으로 사용될 수 있을 것이며 이에 대한 연구가 병행되어야 한다.

실험대상인 약품 온톨로지는 특정 도메인에 맞추어 관련문서 내에 출현하는 용어들의 형태를 분석한 결과를 이용하는 텍스트 마이닝 기술을 이용하여 구축된 도메인 온톨로지이다.

의미적으로 풍부한 온톨로지의 구축은 검색의 효율을 증대시키기 위한 중요한 문제 중의 하나이다. 따라서 현재 약품 온톨로지에 설정되어 있는 33개의 의미관계들 외에 필요한 의미

관계들을 추가하며 개념들을 확장해나가는 것이 필요하며, 특정 도메인이 아닌 다양한 도메인들에 대한 온톨로지에 관한 연구가 계속 되어야 할 것이다.

참 고 문 헌

강승식. 1998. 형태소 해석기 HAM. [cited 2004. 12.26.] 〈http:// nlp.kookmin. ac.kr〉

강신재. 2002. 실용적인 온톨로지의 반자동 구축 및 어휘 의미 중의성 해소를 위한 응용. 포항공과대학교 대학원 컴퓨터공학부 박사학위논문.

김홍기, 김학래, 이강찬, 정지훈, 이재호 외. 2002. 웰드와이드 웹에서 시멘틱 웹으로. 『마이크로소프트웨어』, 4월호: 242-301.

문유진. 1996. 한국어 명사를 위한 WordNet의 설계와 구현. 『정보과학회 논문지』, 2(4): 437-445.

박정오, 황도삼. 2000. 전문용어 추출시스템. 『정보과학회 봄 학술발표 논문집』, 27(1): 381-383.

신효식, 김재호, 이해윤, 최기선. 2002. 텍스트로부터 용어 정의문 자동 추출방법. 『제14회 한글 및 한국어 정보처리 학술발표 논문집』, 292-299.

오종훈, 이경순, 최기선. 2002. 분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출. 『정보과학회 논문지』, 29(4): 258-269.

옥철영. 2004. 한국어 정보처리와 온톨로지. 『2004 한국어 정보처리 연구회 동계 튜토리얼 자료집』, 81-123.

이재호. 2003. 시멘틱 웹의 온톨로지 언어. 『정보과학회지』, 21(3): 18-27.

정도헌. 2003. 시멘틱 웹을 위한 온톨로지 언어와 구현사례 연구. 『정보관리연구』, 34(3): 87-109.

최기선. 2001. KAIST 대용량 코퍼스. [cited 2004.12.27.] 〈http:// kibs. kaist.ac.kr〉

Guarino,N. 1998. "Formal Ontology and Information Systems", *Proceedings of FOIS'98*, 3-15.

Kang,S.J. and Lee, J.H. 2001. "Semi- Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries, and Large Corpora", *ACL 2001 Workshop on Human Language Technology and Knowledge Management*, 45-52.

Klavans,J. and Muresan,S. 2000. "DEFINDER : Rule-based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text", *Proceedings of AMIA Symposium*, 201-202.

Lee,J.H. 1995. "Combining Multiple Evidence from Different Properties of Weighting Schemes", *ACM SIGIR Conference on Research and Development in Infor-*

mation Retrieval, 180-188.

Lenat,D.B. 1995. "Cyc: A Large-Scale Investment in Knowledge Infrastructure", *Communications of the ACM*, 38(11): 33-38.

Lim,S.Y. Koo,S.O. Song,M.H. and Lee,S.J. 2003. "Hub-word based on Ontology Construction for Document Retrieval", *Proceedings of IC-AI'03*, 549-552.

Maedche,A. 2002. *Ontology Learning for the Semantic Web*. Kluwer Academic Publishers.

Missikoff,M. Velardi,P. and Fabriani,P. 2003. "Text Mining Techniques to Automatically Enrich a Domain Ontology", *Applied Intelligence*, 18: 322-340.

Ricardo,B.Y. and Berthier,R.N. 1999. *Modern Information Retrieval*, Addison Wesley Professional.

Salton,G. and McGill,M.J. 1983. *Introduction to Modern Information Retrieval*, McGraw-Hill.

Smith,M.K., Welty,C. and McGuinness,D.L. 2003. OWL Web Ontology Language Guide, World Wide Web Consortium. 〈http://www.w3.org/TR/owl-guide〉

Volz,R. Studer,R. Maedche,A. and Lauser,B. 2003. "Pruning-based Identification of Domain Ontologies", *Journal of Universal Computer Science*, 9(6):520-529.

Vossen,P. "Extending, trimming and fusing WordNet for technical documents", *Proceedings of NAACL-2001 Workshop on WordNet and Other Lexical Resources : Applications, Extensions and Customizations*, 208-215.[cited 2004. 12.27.] 〈http://www. w3.org〉