의미관계 정보를 이용한 약품 온톨로지의 구축과 활용

(Medicine Ontology Building based on Semantic Relation and Its Application)

임 수 연 † 박 성 배 †† 이 상 조 †††
(Soo-Yeon Lim) (Seong-Bae Park) (Sang-Jo Lee)

요 약 온톨로지는 주어진 응용 도메인의 특성을 나타내는 관련 개념들의 집합과 정의, 그리고 그들 간의 관계로 이루어진다. 본 논문에서는 온톨로지를 구축하고 갱신할 때의 시간과 비용을 줄이기 위하여 텍스트의 분석결과를 이용한 도메인 온톨로지의 반자동 구축방안을 제안한다. 이를 위하여 관련 문서들 내에 출현한 전문용어들의 처리방안을 제시하고, 추출한 개념들과 그들간의 관계를 온톨로지의 구축에 활용한다. 실험 도메인은 약품분야로 정하였으며, 구축한 온톨로지는 문서의 검색에 활용하였다. 온톨로지 내의 계층관계들이 문서검색에 효용이 있음을 보이기 위하여 일반적인 키워드기반 문서검색과 온톨로지 내의 관련 정보들을 연관피드백에 이용한 온톨로지기반 문서검색을 비교한 결과, 후자의 경우 정확률이 4.97%, 재현율이 0.78% 향상됨을 알 수 있었다.

키워드 : 온톨로지, 전문용어, 문서검색, 계층관계, 연관피드백

Abstract An ontology consists of a set and definition of concepts that represents the characteristics of a given domain and relationship between the elements. To reduce time-consuming and cost in building ontology, this paper proposes a semiautomatic method to build a domain ontology using the results of text analysis. To do this, we propose a terminology processing method and use the extracted concepts and semantic relations between them to build ontology. An experiment domain is selected by the pharmacy field and the built ontology is applied to document retrieval. In order to represent usefulness for retrieving a document using the hierarchical relations in ontology, we compared a typical keyword based retrieval method with an ontology based retrieval method, which uses related information in an ontology for a related feedback. As a result, the latter shows the improvement of precision and recall by 4.97% and 0.78% respectively.

Key words : ontology, terminology, concept, relation, document retrieval

1. 서 론

온톨로지는 특정주제에 대한 간단한 규칙들, 의미적 연관관계와 단어들을 포함하는 지식용어들(knowledge terms)의 집합으로 정의된다. 즉, 온톨로지는 도메인에 관련된 개념(concept)과 개념을 표현하는 특징(feature), 개념간의 관계(relation) 그리고 특징들이 갖는 제약조건(constraint)으로 구성된 노드들로 표현할 수 있다. 도메인 온톨로지의 유용성은 IT 사회에 널리 알려져 있으며 가장 중요한 문제는 개념들을 식별하고 정의하여 온톨로지를 구축하는 것이다.

크고 복잡한 응용 도메인의 경우 온톨로지의 구축작업은 시간이 너무 오래 걸리고 비용이 많이 들며, 같은 개념에 대해서도 사람마다 다른 관점을 가지므로 논쟁의 여지가 많다. 이들은 대부분 수작업으로 구축되어 왔지만 이 방법은 상당한 시간과 비용이 들므로 최근에는 온톨로지를 반자동으로 구축하기 위한 방안이 활발히 연구되고 있다.

온톨로지의 학습(learning)은 전혀 구조화되지 않았거나 반구조화, 혹은 완전히 구조화된 여러 가지 데이타 유형들을 대상으로 온톨로지의 구축작업을 반자동으로 이루어지게 하며, 온톨로지를 구축하고 갱신할 때의 시

† 비 회 원 : 경북대학교 컴퓨터공학과 연구원
nadalsy@hotmail.com
†† 비 회 원 : 경북대학교 컴퓨터공학과 교수
seongbae@knu.ac.kr
††† 종신회원 : 경북대학교 컴퓨터공학과 교수
sjlee@knu.ac.kr
논문접수 : 2004년 10월 20일
심사완료 : 2005년 3월 12일

간과 비용을 줄여준다. 이 때, 해당 도메인의 개념들과 그들 간의 의미관계를 추출하는 텍스트 마이닝(text mining)기술이 큰 의미를 지니게 된다[1].

온톨로지를 (반)자동으로 구축하는 방안들은 기존의 시소러스나 사전 등과 같은 기존의 자원을 이용하는 방법[2]과 기존의 자원을 이용하지 않고 텍스트의 분석결과로 얻어지는 단어들의 분포를 이용하여 베이스 온톨로지를 구축하고 확장하는 방법[3] 등이 있다.

전자의 경우에는 개념이 부착된 대용량의 사전을 미리 확보함으로써 추가의 사전 작업 없이 바로 활용할 수 있는 지식베이스를 구축할 수 있으므로 활용 가능한 언어 자원만 있으면 제한된 시간과 인력으로 온톨로지를 개발할 수 있는 방법이다. 그러나 사람들이 수동으로 구축한 자원들에 의존하는 경향이 있고 인간의 직관에 의한 계층적인 구조를 벗어나지 못한다. 후자의 경우에는 개념과 관계를 추출하기 위해 텍스트 분석 결과를 이용하므로 노드들의 확장이 용이한 것이 장점이나 온톨로지의 구축을 위해 학습한 문서들에 의존하는 경향이 있다. 따라서 정제된 온톨로지의 구축을 위해서는 학습을 위한 문서 집합을 선정하는 일이 매우 중요하다.

본 논문에서는 해당 도메인의 개념들과 그들 간의 의미관계를 추출하는 텍스트 마이닝(text mining)기술을 이용하여 온톨로지를 구축하고, 외부에 존재하는 대용량의 사전을 이용하여 확장하고자 한다. 이를 위하여 한국어 문서 내에 복합명사의 형태로 출현하는 전문용어들의 패턴들을 분류하고 이들의 구조를 분석한다. 그 결과로부터 도출해낸 의미군과 계층구조를 온톨로지 내의 의미관계로 부여함으로써 도메인 온톨로지의 구축작업이 이루어진다[4].

어떤 주제에 관한 단어들을 계층적으로 분류해 놓은 온톨로지는 다양한 분야에서 활용될 수 있는데, 본 논문에서는 문서 검색의 성능을 향상시키기 위한 방안으로 제안한다. 검색엔진은 온톨로지에 정의된 개념들과 규칙들을 검색의 성능을 활용시키기 위한 추론(inference)의 기반으로도 이용할 수 있다. 이 때, 약품 분야와 관련된 문서 집합에 있는 텍스트들을 실험대상으로 삼았으며 구축한 온톨로지는 약품 온톨로지라고 부르기로 하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 관련 연구에 관해 설명하고, 3장에서는 전문용어를 추출하고 이를 이용하여 도메인 온톨로지를 구축하는 과정에 대해 설명한다. 4장에서는 구축한 온톨로지를 문서검색에 이용하는 방안에 대해 논의하고, 5장에서는 제안한 온톨로지를 검색에 활용하는 실험과 함께 성능을 평가한다. 마지막으로 6장에서 앞으로의 연구 방향과 함께 결론을 맺는다.

2. 관련연구

이미 구축되어있거나 현재 갱신되고 있는 국내외 연구를 살펴보면 다음과 같다.

먼저 국외의 사례를 살펴보면, 미국 뉴멕시코 주립대학에서 개발된 지식기반 기계번역 시스템인 Mikrokosmos[5]는 5,000여개의 개념과 7,000여 단어의 스페인어 사전을 구축하고 있으며, HowNet[6]은 중영 기계번역 시스템의 개발을 위해 만들어진 중국어 온톨로지로 총 53,000개의 중국어 사전과 57,000개의 영어사전을 구축하고 있다. 이는 다른 지식 베이스에 비해 자세한 분류정보를 가졌으나 아직 상용 시스템이 만들어지지 않았고 중국어에 의존적이므로 호환성에 문제가 있다. 또한 인공지능을 응용하여 인간과 같은 추론을 수행할 수 있게 하려는 목적으로 MCC(The Microelectronics and Computer Technology Corporation)에서 10년 전부터 구축한 용어와 일반상식을 데이타베이스로 만든 Cyc[7]가 있다. WordNet[8]은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스턴 대학 인지과학연구실이 구축해온 단어간의 관계를 표현하는 영어어휘 데이타 베이스이다. 1990년 첫판이 완성된 이래, 현재 WordNet 2.0판이 브라우저를 포함해 소스까지 웹사이트(http://www.cogsci.princeton.edu/~wn)에 공개되어 있으며 대략 14만 단어를 포함하고 있다. 단어 중심으로 표현되어 있는 사전과 달리, WordNet은 단어형이 아닌 단어의 의미를 구성요소로 하여 네트워크 형태로 구성되어 있다는 것이 특징이다. WordNet은 현재 자연언어처리와 정보검색의 여러 분야에서 널리 이용되고 있으며 다국어판의 구현도 시도되고 있다. 그 중의 하나인 EWN(EuroWordNet)은 유럽 8개 국어를 대상으로, 각 나라의 언어에 대해 개념 분류(ontology)를 통해 어휘 데이타베이스를 구축하고, 미국 프린스턴 대학의 워드넷을 기반으로 하여 각각의 어휘들의 공통개념을 추출한 중간 개념 지표(Inter-lingual Index)를 이용해 각 언어들을 연결시킨 다국어 어휘 데이타베이스이다.

국내 온톨로지의 대표적인 구축사례로는 한국 전자통신 연구소에서 만든 ETRI 명사 개념망을 들 수 있다. 이는 한국어 명사 어휘로 표현되는 개념을 정확하게 파악하기 위하여 개념들 간의 다양한 관계를 연결시켜 놓은 어휘 데이타베이스이다. 개념망의 노드는 국어사전에 등재된 단어 중에 개념어를 선별하고, 사전 뜻풀이를 기준으로 상하관계, 동의관계, 유의관계, 부분-전체관계 등 언어학적인 의미관계(semantic relation)를 이용하여 구축되어 있는 언어자원이다. 현재 일반명사 약 5만 단어

와 경제 명사 약 1만 5천 단어로 구성되어 있으며, 이를 확장하는 작업과 함께 동사 개념망을 구축 중에 있다. 코어넷은 KAIST에서 일본국립국어연구소의 어휘 분류표에 근거하여 어휘 의미 속성 체계를 개념 체계로 설정하고, 단어들의 의미와 개념들을 연결한 것이다. 이는 한국어, 영어, 중국어를 일대일 대응한 다국어 목록으로 총 62,281개의 어휘를 구축하였다. 이외에도 국내에서는 포항공대의 LIP(Language Independent and Practical) 온톨로지[9], 울산대학교의 UOU온톨로지[10], 한국어 명사 워드넷[11] 등의 다각적인 지식 베이스 구축 방법이 개발되고 있다.

지금까지 소개한 대부분의 구축작업은 막대한 비용과 인원을 기반으로 장기간에 걸쳐 수동으로 진행되거나 사전이나 시소러스 같은 기존의 자원을 변환함으로써 이루어져 왔다. 일반적인 도메인의 경우는 이와 같은 작업이 효율적이라고 할 수 있으나, 의학, 공학 등과 같은 특정 도메인의 경우는 많은 전문적인 지식을 내포하고 있어 기존 자원들에 내포된 지식만으로는 부족한 실정이다.

따라서 본 논문에서는 도메인을 약품분야로 정한 뒤, 텍스트 내에 출현하는 전문용어들의 출현패턴을 분석한 결과를 약품 온톨로지의 구축에 이용하고자 한다. 이를 위하여 전문용어들을 구성하고 있는 접미사나 특정 명사들의 의미정보에 기반한 규칙들을 설정하고, 전문용어들과 그들 간의 관계를 추출함으로써 온톨로지 내의 노드들을 확장해나간다.

3. 도메인 온톨로지의 구축

온톨로지의 개발은 도메인의 특성과 온톨로지의 사용용도를 고려하여 온톨로지의 구성영역을 정하고, 온톨로지가 응답해야 할 질의들에 대하여 온톨로지 내에 포함되어야 할 세부 사항들을 먼저 정한다. 이를 위하여 해당 도메인 전문가들과의 협의에 의하여 개념들과 그들의 관계들의 구조를 정한 뒤 이들을 기반으로 구축하게 되며, 실제의 응용 시스템에서는 도메인마다의 특정적인 지식을 포함하는 온톨로지가 필요하다.

본 논문에서는 실험 도메인을 약학 분야로 정하고 학습을 위한 문서들은 약학 도메인 내의 문서들로 한정하였다. 즉, 구체적인 온톨로지의 구축은 약품과 관련된 도메인 내에서 행해지며, 병명이나 증상에 따른 약품명이나 관련 문서들을 검색하기 위한 약품 온톨로지를 구축하는 것을 목적으로 한다.

3.1 구축과정

제안한 온톨로지의 구축과정은 그림 1과 같은 네 개의 단계로 이루어진다. 첫 번째 단계에서는 코퍼스를 형성하기 위해 관련 도메인 내의 웹 문서들을 수집하고,

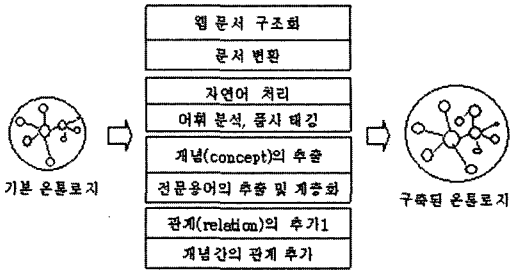

그림 1 기본 온톨로지의 구성

수집한 문서들은 문서변환 과정을 거쳐 구조화된다. 두 번째 단계에서는 간단한 자연어 처리과정을 거친 뒤, 개념이 될 수 있는 명사나 추출된 개념들 사이의 관계를 나타내는 동사를 추출한다. 세 번째 단계에서는 추출된 개념들 중 전문용어들을 추출하고 이들의 구조를 분석한 결과로부터 계층구조를 도출해낸다. 마지막으로, 추출한 관계들을 이미 존재하고 있는 온톨로지에 개념들과 함께 추가시킨다.

약품과 관련이 있는 데이타베이스(http://www.druginfo.co.kr)내의 문서들을 분석한 결과를 이용하여 구축할 약품 온톨로지의 개념들과 이들을 연결시킬 관계들을 설정한다. 수집된 문서들은 반구조화된(태깅된) 문서들로 설정한 구조에 맞도록 변환 과정을 거친 뒤 부착된 태그에 따라 개념들을 형성하게 된다. 구축된 온톨로지에 존재하는 개념들과 그들의 관계는 OWL을 이용하여 표현한다.

3.1.1 기본 온톨로지

온톨로지를 구축할 때 설계자는 상위레벨에 있는 소수의 노드들을 결정하고 이를 기반으로 온톨로지를 구축한 뒤 확장하게 되는데, 이를 기본 온톨로지(Base Ontology)라고 한다. 본 논문에서는 그림 2와 같은 48개의 어휘들로 구성된 기본 온톨로지를 구성하였다.

이를 위하여 병명, 증세, 약품 개념을 최상위 노드로 설정하고 그에 대한 45개의 하위노드를 설정하였다. 하위 노드들은 약학 도메인에서 병명이나 증세를 구성하고 있는 특정 명사나 접미사들의 분류에 따른 20개의 노드들과 설정한 구조에 필요한 15개의 노드, 그리고 출현빈도가 높은 일반 명사를 나타내는 10개의 노드들이다.

3.1.2 개념의 추출

문서 내의 텍스트들은 형태소 분석 과정과 태깅 과정을 거친 후, 텍스트 내의 불용어들을 제거하고 스테밍한 뒤, 각 문장에 대한 문서 내의 모든 명사와 동사들을 추출한다. 이를 위하여 한국어의 형태론적 특성을 고려하여 181개의 어휘들로 구성된 불용어 리스트를 작성하였으며, 스테밍할 때 일부 접미사들을 제외시켰다. 이유는 이들 접미사들이 전문용어의 추출에 유용하게 쓰일 수

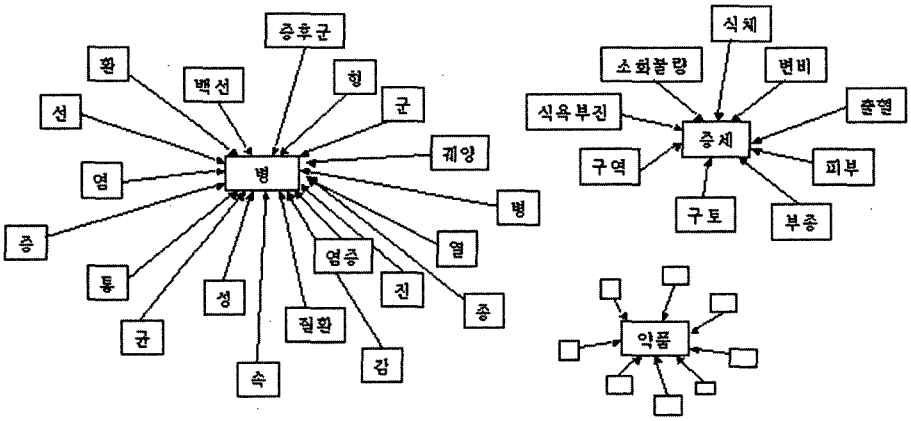

그림 2 기본 온톨로지의 구성

있기 때문이다.

많은 어휘들로 구성된 네트워크의 일종인 온톨로지에서는 텍스트로부터 추출한 명사들은 온톨로지의 개념을 나타내고, 문서에 붙은 태그들과 동사들은 개념들 간의 관계를 나타내며 개념들을 연결짓는 링크로서의 역할을 한다.

우리가 관심을 갖는 약품과 관련된 문서들에서는 병명, 증세, 성분 등을 나타내는 고유명사나 복합명사들이 등장하였다. 주요 개념들을 나타내는 고유명사들은 일반명사와 같은 방법으로 처리하였으며, 해당 도메인에서 복합명사의 형태로 출현하는 전문용어들을 추출하고 계층화하여 이를 온톨로지에 추가하였다.

3.1.3 관계의 추가

본 논문에서는 추출된 개념들에게 관계를 부여하기 위하여 두 가지 방법을 이용한다. 하나는 텍스트 앞에 부착된 태그 값을 이용하는 방법이고, 다른 하나는 텍스트 내에 존재하는 동사들을 추출하여 이용하는 방법이다. 그림 3은 부착된 태그 값에 따른 15개의 의미관계 유형들을 보여주며 그림 4는 이 관계들 중의 일부를 보여준다.

의미 관계들	
productedBy (제조회사)	hasMethod (복용방법)
hasInsCode (보험코드)	hasContra (금기사항)
hasComCode (성분코드)	hasSideEffect (부작용)
hasClsCode (분류코드)	byMean (음용법)
hasColor (색깔)	byAmount (용량)
hasKind (모양)	byUnit (복용단위)
hasForm (형태)	byAge (연령)
hasEffect (효능, 효과)	

그림 3 태그 값에 따른 의미 관계 유형

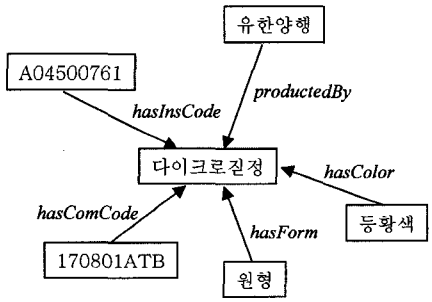

그림 4 의미 관계 유형의 예

또한, 개념들 간의 관계를 정의하기 위하여 주변에 나타난 명사들을 연결짓는 동사들을 추출하였다. 그 결과, 빈도가 200이상인 35개 동사들의 출현빈도는 11,250으로 전체 동사들의 출현빈도인 23,453의 47.97%를 차지하고 있었다 우리는 이들 동사들을 의미패턴으로 분류하여 그림 5와 같은 18개의 의미 관계를 설정하였다.

추출한 명사와 동사들 간의 연관관계는 공기(co-occurrence) 정보를 이용하여 파악한다. 만약 명사와 동사들 간에 연관관계가 형성된다면 그들 사이에 링크를 형성하고, 그렇지 않다면 다른 명사와 동사들 간을 비교한다.

3.1.4 전문용어의 추출

전문용어(terminology)란 주어진 도메인 안에서 의미를 가지고 있는 단어들의 집합으로 도메인 내에서 사용되는 개념을 표현시켜 줌으로서 주제를 특성화 시켜주는 어휘적 단위를 말한다. 이러한 전문용어는 하나의 도메인을 이해하는데 필요한 요소이기 때문에 특정 도메인에 대한 기계번역이나 정보검색을 보다 효율적이고 정확히 수행하기 위해서 전문용어에 대한 언어자원은 중요하다. 본 논문에서는 전문용어들을 자동으로 추출하

동사의 추출에 의한 의미 관계들	
관계	해당하는 동사들
appear	일어나다, 나타나다, 발현하다, 생기다, 발생하다
beWorse	증상이 악화되다
inject	주사하다, 근육주사하다
noTake	투여를 중지하다, 투여를 피하다, 복용을 금하다
reduce	감량하다, 감소하다
return	재발하다
rise	상승하다, 증가하다, 증량하다
take	복용하다, 투여하다, 경구투여하다
use	사용하다, 이용하다
cause	유발하다, 일으키다, 기인하다
accompany	수반하다
prevent	예방하다, 막다
control	조절하다
infect	감염되다
cure	치료하다, 처치하다
improve	개선되다
maintain	유지하다
relax	완화되다

그림 5 빈도가 높은 동사들에 따른 의미 관계 분류

기 위하여 그들의 출현형태를 분석하였다. 전문용어의 형태결합 방식은 매우 다양하다. 해당 도메인에 출현하는 대부분의 전문용어들은 복합명사의 형태로 출현하였으며, 크게 두 가지의 결합형태로 나눌 수 있다. 하나는 단일어절(singleton term) 즉, 띄어쓰기가 없는 한 어절로 나타나는 단순한 결합형태이고, 다른 하나는 다중어절(multi-word term)의 형태로 띄어쓰기가 나타나며 앞의 어절성분과 의미적으로 관련이 있는 두 어절이상으로 이루어진 복합명사이다.

단일어절형태 전문용어들을 구성하고 있는 명사나 접미사들은 20가지로 분류하였다. 이들은 "염, 증, 통, 균, 성, 질환, 속, 염증, 진, 감, 종, 병, 열, 궤양, 선, 백선, 증후군, 형, 환, 군"이며, 특정명사의 하위단어들인 경우가 대부분으로 "hyponym Of" 관계로 연결한다.

다중어절형태 전문용어들은 "급성 기관지염"과 같이 대부분 수식어와 중심어의 관계를 가지며, 중심어가 다

패턴	패턴의 형태
패턴1	N1(~성, ~형)+N2
패턴2	N1(~에 의한, ~(으)로 인한, ~(으)로 인해 유발된)+N2 N1(~에 따른)+N2 N1(~시(의), ~상태에서, ~후(의))+N2
패턴3	N1+ 의 +N2, N1+N2
패턴4	N1+ 및 +N2, N1+ +N2, N1+ 또는 +N2
패턴5	N1 (suffix_ 1)+ ,및 +N2(suffix_ 2)+N3 (if suffix_ 1=suffix_ 2)

그림 6 다중어절형태 전문용어를 처리하기 위한 관계 패턴

시 단일어절로 이루어진 전문용어로 이루어진 경우가 많았다. 우리는 그림 6과 같은 다섯 개의 관계패턴들을 설정하여 이에 따라 온톨로지 내의 의미관계를 설정하였다[4].

3.2 온톨로지의 확장

구축된 온톨로지는 다른 온톨로지나 시소러스, 사전 등과 같은 다른 자원들에 의존하여 확장할 수가 있다. 기존의 자원을 이용하여 이미 정의된 개념들과 규칙을 활용함으로써 온톨로지의 확장에 드는 시간과 노력을 줄일 수 있다.

본 논문에서는 약품 온톨로지의 확장을 위하여 의약용어 약어사전(http://www.nurscape. net/nurscape/dic/frames.html)과 두산 세계대백과 엔싸이버(http://www.encyber.com)를 이용하였다. 이 때, 텍스트의 범위가 너무 광범위하므로 대상 개념의 조회 결과 중에서 전문용어들과 그들의 하위개념만을 추출하였다. 추출된 개념들은 온톨로지의 상위개념 아래에 추가시켜 줌으로써 확장이 이루어지게 되며, import, extract, append의 세 단계로 이루어진다. Import는 외부의 자원을 가져와서 이용하는 것을 의미하며, 본 논문에서는 관련이 있는 문서들을 수집하기 위하여 두 개의 외부 자원들을 이용한다. 배경 지식의 형태로 주어진 사전 등의 import된 자원들로부터 관련된 어휘항목인 전문용어들을 추출(extract)한 뒤, 추출된 개념들을 온토로지 내의 상·하위 관계들을 고려한 적당한 위치에 링크로 연결(append)하게 된다. 두개의 기존 자원들을 이용한 온톨로지의 확장 과정을 개괄적으로 나타낸 것이 다음의 그림 7이다.

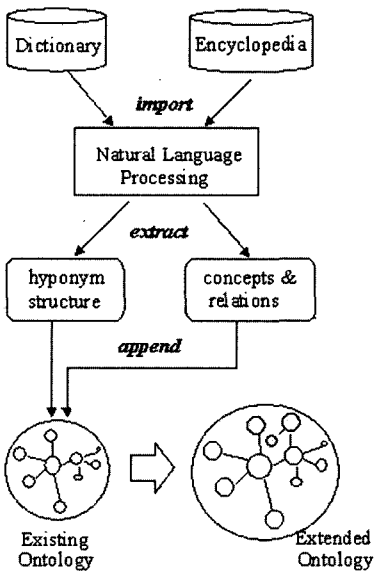

그림 7 외부자원을 이용한 온톨로지의 확장

4. 문서 검색에의 이용

구축된 온톨로지는 다양한 분야에서 활용될 수 있는데, 본 논문에서는 문서 검색의 효과를 향상시키기 위한 방안으로 이를 제안하고자 한다. 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 제안한 전문용어의 추출방법을 이용하여 구축한 약품 온톨로지를 대상으로 입력된 질의어에 해당하는 개념뿐 아니라 그의 온톨로지 내 하위 개념들까지 탐색하는 것을 목적으로 하는 검색을 진행한다.

온톨로지 내의 가중치 부여는 벡터모델을 기반으로 주어진 문서들 내에서 특정 단어가 얼마나 자주 사용되는가를 나타내는 출현빈도로부터 유도될 수 있다. 일반적으로 벡터모델에서 가중치 부여를 위하여 단어들의 빈도수를 고려한 $tf \bullet idf$ 방법을 사용할 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 된다.

정보 접근 과정의 중요한 부분인 질의 재형식화를 위한 효과적인 방법으로는 연관 피드백(relevance feedback)이 널리 알려져 있다. 전통적인 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 온톨로지 내의 노드를 탐색할 하위어 검색 레벨은 2로 정하였다. 예를 들어 온톨로지 내에서 노드 '중이염'에 대한 하위노드로 '삼출성중이염'과 '급성삼출성중이염'이 존재한다고 가정하자. 입력으로 질의어 {중이염}이 들어온 경우, 온톨로지를 탐색한 후 질의어 집합은 {중이염, 삼출성중이염, 급성삼출성중이염}으로 확장되고 이들의 가중치를 기반으로 유사도를 다시 계산하게 되는 것이다.

벡터모델에서 t차원 벡터로 표현된 문헌 d_j와 사용자 질의 q의 유사도 측정은 두 벡터 $\vec{d_j}$와 $\vec{q}$의 상관도로 구할 수 있으며, 이는 두 벡터간 사이각의 코사인 값으로 정량화 할 수 있다.

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

이 때, 가중치의 부여는 가장 널리 알려진 $tf \bullet idf$ 기법(용어-가중치 할당 전략)을 이용하여 계산한다. 계산된 가중치는 출현 문서 벡터에 문서번호와 함께 정렬된 순으로 저장됨으로써 검색의 속도를 향상시키고 더 정확한 검색을 가능하게 해준다. 그림 8은 출현 문서 벡터의 구조를 그림으로 나타낸 것이다.

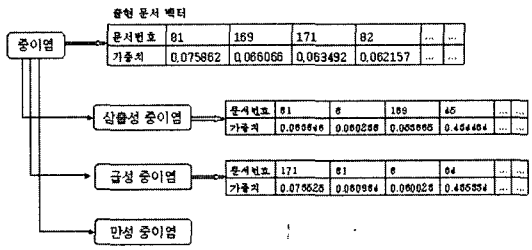

그림 8 출현 문서 벡터의 구조

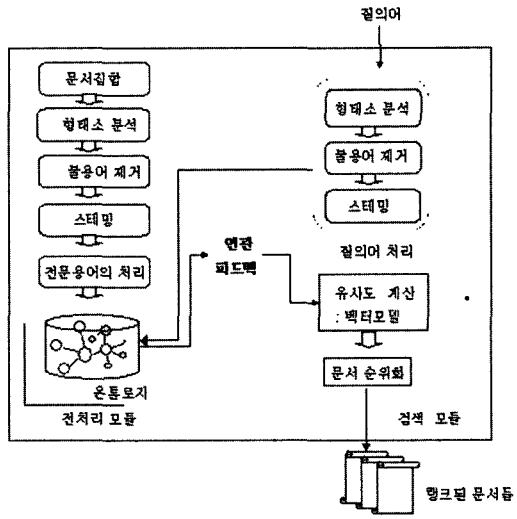

그림 9 문서검색 시스템의 전체 구성도

그림 9는 제안한 문서 검색 시스템의 구성도를 보여주고 있으며, 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.

검색의 성능을 비교하기 위하여 430개의 문서들을 대상으로 전문가 5인의 자문을 구하여 10개의 질의에 대한 상위 30개의 정답 문서 집합을 정하였다. 이를 기준으로 각각의 질의에 대한 재현율과 정확률을 구하고 이들의 평균을 구하였다.

5. 실험 및 평가

형성한 코퍼스 내의 텍스트들에 대한 형태소를 분석한 결과, 추출된 명사들 중 특정 어구나 패턴이 이용된 용어들을 전문 용어로 추출한다. 띄어쓰기 오류나 타이핑의 오류로 인하여 형태소 분석이 실패한 경우에는 해석의 오류를 수정하였다.

5.1 전문용어 인식

약품 도메인 내의 텍스트들을 실험 대상으로 하여 본 논문에서 제안한 전문 용어의 추출 방법을 적용하였다. 실험 대상이 된 실험 문서의 수는 21,113개이다. 구문분석을 실시한 결과 추출된 전체 명사수는 총 78,902개이다. 추출된 전체 명사수 78,902개에 대하여 전문용어들의 수는 55,870개로 전체 명사수의 약 70.8%를 차지한다. 이는 특정 도메인 내에서 전문용어의 비중이 아주 높음을 의미한다. 그림 10은 전문용어들의 출현형태에 따른 분포도를 보여준다.

본 논문에서 제안한 단일어절형태의 전문용어로부터 계층관계를 추출하는 알고리즘을 적용한 결과 출현한 전문 용어들의 인식과 함께 2,864개의 하위 개념이 추가되었으며 온톨로지 내의 노드들의 평균 레벨은 1.8로 나타났다. 추출한 전문 용어에 대한 평가는 세 명의 전문가에 의해서 수작업으로 조사하고 결과는 추출 정확도로 평가하였다. 추출 정확도는 추출된 전문 용어들 중 올바른 관계로 연결된 전문 용어들의 비율을 나타낸다.

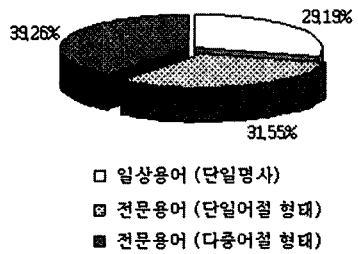

그림 10 용어들의 분포도

그림 11은 단일어절형태 전문용어들의 추출 정확도를 그래프로 나타낸 것이다. 그 결과, 단일어절형태 전문용어들의 평균 정확도는 92.57%로 제안된 알고리즘이 비교적 좋은 성능을 나타냄을 알 수 있었다.

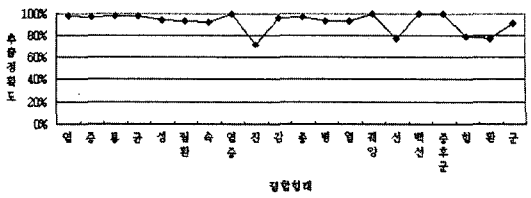

그림 11 단일어절형태 전문용어들의 추출 정확도

예를 들어 명사 "진"의 경우, "촉진, 증진, ..."과 같은 개념들이 "습진, 농가진, ..." 개념들과 같은 군으로 형성된다. 이와 같이 잘못된 군이 많이 형성되는 경우에는 오류로 인식하였으므로 71.87%라는 낮은 정확도가 발생하였다.

그림 12는 다중어절형태 전문용어들의 추출 정확도를 그래프로 나타낸 것이다. 그 결과, 다중어절형태 전문용어들의 평균 정확도는 79.96%였으며, 574개의 개념이 추가되었다.

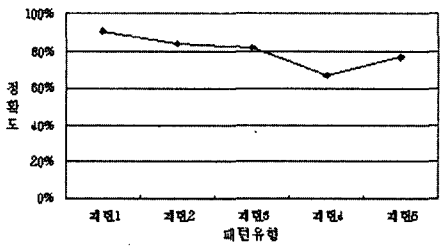

그림 12 다중어절형태 전문용어들의 추출 정확도

실험결과, 제안된 전문용어 처리방법에 의해 인식하지 못하는 오류형태는 표 1과 같다. 단일 어절 형태의 경우에는 결합하는 명사나 접미사가 동일한 전문용어와 단일명사를 구분하지 못하였다. 이는 명사사전의 불충분으로 인한 문제이므로 특정 접미사나 명사로 종결되는 단일명사사전을 보강하고 이를 우선적으로 검색함으로써 간단히 해결될 수 있다. 하지만 다중어절인 경우에는 "다음의 ... 아래에 의한 ..."과 같은 특정 어휘에 대한 별도의 정교한 처리가 필요함을 알게 되었다.

표 1을 분석한 결과, 낮은 정확도가 발생하는 원인이 다양하지 못한 패턴의 종류에 있다는 것을 알았다. 따라

표 1 인식하지 못한 오류의 형태와 예

오류의 형태	오류의 예
단일어절	변형, 합병, 옷감, 전환, 고통, 증진, 촉진, 감염, 배열, 지속, 신속, …
다중어절	균형 유지, 각종 원인에 의한, 아래의 질환, 급만성 질환, …
그 외	띄어쓰기, 맞춤법, 외래어 표기상의 오류 등

서 이를 보완하기 위해서는 정교한 텍스트의 분석을 통해 패턴들을 확장해 나가는 것이 필요하며 이는 향후과제로 남겨두기로 한다. 또 다른 오류의 형태는 띄어쓰기나, 맞춤법, 외래어 표기상의 오류 등에서 찾아볼 수 있었다. 그리고 전문가들의 습관적 오류도 한 몫을 하는 것으로 생각된다.

5.2 검색 성능 평가

구축한 온톨로지의 효용을 보이기 위하여 전통적인 *tf•idf* 방법을 이용하여 가중치를 부여하는 키워드기반 문서검색과 온톨로지 내의 하위 정보를 연관 피드백에 이용하고 가중치를 재계산하는 온톨로지기반 문서검색의 결과를 비교, 분석하였다.

질의어 "중이염"에 대한 예를 들어보자. 표 2는 키워드기반 문서검색시 추출된 단어의 수와 빈도의 분포를 보여주며, 표 3은 온톨로지 내의 하위어들까지 검색에 이용하여 추출한 단어의 수와 빈도의 분포를 보여준다. 두개의 표를 분석한 결과, 질의어 "중이염"에 대하여 "만성 중이염", "만성 유착성 중이염" 등과 같은 36개의 중이염의 하위 단어가 추가됨을 알 수 있으며, 하위어로

표 2 키워드기반 검색시 추출된 단어수와 분포

문서번호	추출된 단어수	키워드 출현빈도	전체단어 출현빈도	점유율
doc_1	81	10	142	7.04%
doc_2	254	40	545	7.34%
doc_3	88	18	140	12.86%
doc_4	176	27	347	7.78%
doc_5	102	8	171	4.68%
doc_6	183	19	313	6.07%
doc_7	126	18	268	6.72%
doc_8	129	15	249	6.02%
doc_9	171	27	302	8.94%
doc_10	216	23	463	4.97%
추출된 총 단어수	1,526	205	2940	7.24%

표 3 온톨로지기반 검색시 추출된 단어수와 분포

문서번호	추출된 하위단어수	하위어고려 출현빈도	전체단어 출현빈도	점유율
doc_1	5	20	142	14.08
doc_2	6	71	545	13.03
doc_3	3	25	140	17.86
doc_4	2	32	347	9.22
doc_5	3	13	171	7.60
doc_6	3	31	313	9.90
doc_7	2	20	268	7.46
doc_8	0	15	249	6.02
doc_9	4	34	302	11.26
doc_10	8	51	463	11.02
추출된 총 하위단어수	36	312	2,940	10.75%

검색레벨이 확장된다는 것이 검색의 정확률에 영향을 미칠 수 있음을 짐작할 수 있다.

제안한 방법이 문서검색에 효용이 있음을 보이기 위하여 두 가지 방법에 의한 문서 검색을 비교하였다. 하나는 전통적인 *tf•idf* 방법을 이용하여 가중치를 부여한 키워드기반 검색의 경우이고, 다른 하나는 온톨로지 내의 계층 정보들을 연관 피드백에 이용한 온톨로지기반 검색이다. 정보 검색 시스템의 검색 성능을 평가하기 위해서는 실험 참조 컬렉션과 평가 척도를 사용한다. 실험 참조 컬렉션은 문헌 집합, 정보 요구 예제, 각 정보 요구에 대한 연관 문헌 집합으로 구성된다. 본 논문에서는 참조 컬렉션을 구성하기 위해 대한의사협회 홈페이지(http://www.kma.org)에서 제공하는 건강/질병 정보 문서 430개를 수집하였으며 다음과 같은 10개의 질의로 구성된 정보 요구를 구성하였다.

실험은 추출한 430개의 문서를 대상으로 하였다. 10개 질의들에 대한 재현율과 정확률을 구하는 것을 목표로 하였으며, 입력된 각 질의에 대한 정답 집합으로는 전문가들이 정한 문서의 순위를 기준으로 정하였다.

[질의]

질의 1 : {중이염, 증상, 종류}

질의 2 : {중이염, 치료}

질의 3 : {중이염, 치료, 약}

질의 4 : {중이염, 특징}

질의 5 : {중이염}

질의 6 : {만성 중이염, 진단, 치료}

질의 7 : {급성 중이염}

질의 8 : {중이염, 감염경로}

질의 9 : {고열, 병}

질의 10 : {귀, 병}

탐색 작업의 속도를 향상시키기 위해 텍스트에 대한 색인을 만들게 되며, 이 들은 어휘와 출현 빈도의 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.

다음의 표 4는 질의 5:{중이염}에 대해 부여된·가중치와 가중치를 이용하여 계산된 상위 10개 문서와의 코사인 유사도를 보여준다.

그림 13과 그림 14는 입력된 각 질의들에 대한 두 가지 문서검색의 정확률과 재현율의 분포를 비교해서 보여주고 있다. 비교하는 두 가지 검색에 대하여 평균 재현율과 평균 정확률을 각각 구한 결과가 표 5에 나타나 있다.

이로부터 우리는 온톨로지 내의 하위정보를 질의의 확장에 이용하고 가중치를 부여하는 방법에 의한 문서

표 4 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

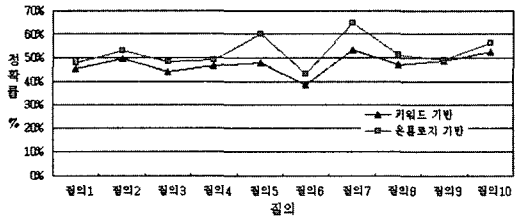

그림 13 정확률의 비교

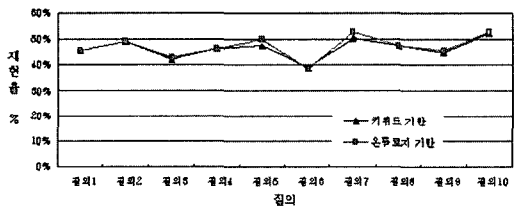

그림 14 재현율의 비교

의 검색이 전통적인 방법을 이용한 검색보다 재현율은 0.78%, 정확률은 4.97% 높게 나타났다. 이는 문서의 검색에 온톨로지 내의 계층관계를 연관 피드백 정보로 이용하면 재현율의 향상에는 별 영향을 주지 않지만 정확률을 향상시키는데 효용이 있음을 뜻한다.

표 5 평균 재현율과 평균 정확률

	키워드기반 검색	온톨로지기반 검색
재현율	46.34%	47.12%
정확률	47.28%	52.25%

6. 결 론

본 논문에서는 텍스트의 분석결과를 이용한 도메인 온톨로지의 반자동 구축 방안을 제시하고 구축한 온톨로지를 문서의 검색에 활용하였다. 실험 도메인은 약학 분야로 정하고, 웹으로부터 약품과 관련된 문서들을 수집하여 코퍼스를 형성하였다. 형성된 코퍼스 내에 있는 텍스트들의 구조를 분석하여 온톨로지의 구조를 정하고 개념과 관계를 추출하기 위한 관계 설정 유형을 설정하였다. 특히 관련 문서 내에 출현하는 용어들의 형태를 분석한 결과, 약품 온톨로지의 구축에 필요한 개념과 관계들을 추출하기 위하여 특정 명사나 접미사와 결합한 전문용어의 처리방안을 제안하였다. 제안한 방법은 텍스트 마이닝 기술을 이용한 반자동화된 방법으로서 수동으로 구축할 때의 인간의 노력과 시간을 줄일 수 있다.

특히, 온톨로지 내의 계층관계들이 문서의 검색에 효용이 있음을 보이기 위하여 출현 빈도 정보만을 이용하여 가중치를 부여한 키워드기반 문서검색과 온톨로지 내의 관련 정보들을 연관 피드백에 이용한 온톨로지기반 문서검색을 비교하였다. 이 때 검색의 성능을 평가한 결과, 재현율은 비슷하게 유지하면서 정확률이 4.97% 향상되는 것을 알 수 있었다. 이는 온톨로지 내의 계층관계를 연관 피드백 정보로 이용하면 검색의 정확률을 향상시킬 수 있다는 것을 의미한다.

더욱 정교한 질의의 처리를 위해서는 온톨로지를 수정 및 확장해 나가는 것이 필요하다. 또한 현재 약품 온톨로지에서 정의하고 있는 33개의 의미관계가 부족할 수도 있다. 이런 경우에는 해당 도메인에 필요한 기타 의미 관계를 다시 정의하고 약품 온톨로지를 확장해 나가야 한다. 또한 약품 온톨로지는 특정 도메인에 맞추어 구축되었기 때문에 이를 범용적인 목적으로 사용하는 방법에 대한 연구가 필요하다. 즉, 다양한 도메인들에 대한 온톨로지의 구축이 필요하며, 제안된 온톨로지의 구축방법을 일반 도메인에 적용하도록 확장하는 방안에 관해 계속 연구해 나가야 할 것이다.

참 고 문 헌

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

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

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

[4] 임수연, 송무희, 이상조, "전문용어의 처리에 의한 도메인 온톨로지의 구축", *정보과학회 논문지(B)*, 제 31 권 3호, pp. 353-360, 2004.

[5] Mahesh, K., "Ontology Development for Machine Translation: Ideology and Methodology," *Technical*

Report MCCS 96-292, Computer Research Laboratory, New Mexico State University, 1996.
[6] Dong, Z. and Dong, Q., *HowNet*.http://www.keenage.com/zhwang/e_zhi wang.html, 1999.
[7] Lenat, D. B., "Cyc: A Large-Scale Investment in Knowledge Infrastructure," *Communications of the ACM*, Vol. 38, No. 11, pp. 33-38, 1995
[8] Miller, G. A., Chodorow, M., Landes, S., Leacock, C. and Thomas, R. G., "WordNet:An On-line Lexical Database," *International Journal of Lexicography*. Vol. 3, No. 4, pp. 235-244, 1990.
[9] 강신재, "실용적인 온톨로지의 반자동 구축 및 어휘 의미 중의성 해소를 위한 응용", 포항공대 박사학위논문, 2002.
[10] Ock, C. Y. and Choe, H. S., "The Fundamental Construction Princi- ples of UOU Ontology," *1st International Workshop of The Korean Thesaurus Association*, 2003.
[11] 문유진, "한국어 명사를 위한 WordNet의 설계와 구현", *정보과학회 논문지*, 제2권, 4호, pp. 437-445, 1996.
[12] 박정오, 황도삼, "전문용어 추출시스템", *정보과학회 봄 학술발표 논문집*, 제27권, 1호, pp. 381-383, 2000.
[13] 이경희, 이주호, 최명석, 김길창, "한국어 문서에서 개체명 인식에 관한 연구", *제12회 한글 및 한국어 정보처리 학술대회 학술발표논문집*, pp. 292-299, 2000.
[14] Klavans, J. and Muresan, S., "DEFINDER:Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text," In *Proceedings of AMIA Symposium*, pp. 201-202, 2000.
[15] 신효식, 김재호, 이해윤, 최기선, "텍스트로부터 용어 정의문 자동 추출방법", *제14회 한글 및 한국어 정보처리 학술발표 논문집*, pp. 292-299, 2002.
[16] 오종훈, 이경순, 최기선, "분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출", *정보과학회 논문지*, 제29권, 4호, pp. 258-269, 2002.
[17] 황이규, 윤보현, "HMM에 기반한 한국어 개체명 인식", *정보처리학회 논문지(B)*, 제10권, 2호, pp. 229-236, 2003.

임 수 연

1988년 2월 경북대학교 전자공학과(공학사). 1991년 2월 경북대학교 컴퓨터공학과(공학석사). 2004년 8월 경북대학교 컴퓨터공학과(공학박사). 2004년 9월~현재 경북대학교 컴퓨터공학과 연구원. 관심분야는 정보검색, 온톨로지, 기계학습, 자연어처리

박 성 배

1994년 2월 한국과학기술원 전산학과(학사). 1996년 2월 서울대학교 컴퓨터공학과(석사). 2002년 8월 서울대학교 전기컴퓨터공학부(박사). 2004년 3월~현재 경북대학교 컴퓨터공학과 교수. 관심분야는 기계학습, 자연언어처리, 정보검색, 바이오 인포매틱스

이 상 조

1974년 2월 경북대학교 수학교육과(이학사). 1976년 2월 한국과학기술원(이학석사). 1994년 2월 서울대학교 컴퓨터공학과(공학박사). 1976년~현재 경북대학교 컴퓨터공학과 교수. 관심분야는 언어처리, 지식처리, 정보검색, 기계학습, 시멘틱웹