온톨로지내의 계층관계를 이용한 문서검색

임수연° 송무희 이상조
경북대학교 컴퓨터공학과
nadalsy@hotmail.com °

Document Retrieval using the Ontology Hierarchy

Sooyeon Lim° Muhee Song Sangjo Lee
Dept. Computer Engineering, Kyungpook National University, Korea

요 약

온톨로지는 주어진 응용 도메인의 특성을 나타내는 관련 개념들의 집합과 정의 그리고 그들간의 관계로 이루어진다. 본 논문에서는 코퍼스에 있는 텍스트의 분석 결과를 이용한 온톨로지를 구축방안과 이를 문서의 검색에 사용함으로써 해당정보가 있는 자원을 찾는 정확도를 향상시키는 방안을 제시하고자 한다. 이를 위하여, 실험 도메인의 문서 내에 출현한 전문 용어들의 결합형태를 분석하여 계층구조를 도출해내는 알고리즘을 제안하며 구축된 온톨로지를 문서의 검색에 응용하였다. 제안된 온톨로지는 전통적인 문서검색의 인덱스 파일과 같은 역할을 하게 되며, 질의로 들어온 키워드뿐 아니라 그에 대한 온톨로지 내 하위어들에 기반하여 검색을 수행함으로써 많은 의미정보를 포함하고 있으며 검색의 정확도를 높일 수 있었다.

1. 서 론

최근 몇 년간, 온톨로지는 컴퓨터과학 분야, 특히 지식의 공유와 재사용을 목적으로 하는 인공지능 분야에서 많은 관심을 끄는 주제가 되어 왔다. 본래 철학의 주된 연구 대상으로 오랜 역사를 가지고 있는 온톨로지는 웹의 발전과 더불어 웹 기반의 지식처리, 정보통합, 데이터베이스 설계, 정보검색 그리고 정보교환 등의 여러 목적에 사용된다.

온톨로지에 대한 정의는 여러 가지가 있지만 Gruber는 온톨로지를 공유화된 개념화(shared conceptualization)에 대한 정형화되고 명시적인 명세(formal and explicit specification)라고 정의하였다[1]. 간단히 말하면 온톨로지는 어떤 특정 도메인(실세계)에서 사용되는 정보들과 그 정보들간의 관계를 정의해 놓은 것을 말하며, 이를 계층적 구조로 표현하고 이를 확장할 수 있는 추론 규칙을 포함한다. 최근 상당한 시간과 비용을 요구하는 수작업 대신 온톨로지를 (반)자동으로 구축하기 위한 방안이 계속 제시되고 있는데 이들은 시소러스와 같은 다른 자원들에 의존하는 경향이 있으며 이들은 일반적이 도메인에 대해 구축되어 왔다. 그러나 각 분야마다 단어들의 사용개념이 다르므로 실제의 응용 시스템에서는 각 도메인마다의 특정적인 지식을 포함하는 온톨로지가 필요하다.

정보검색 시스템은 문서집합에 있는 정보의 내용을 번역하고, 입력된 질의와 관련된 정보를 찾아내는 것을 목적으로 한다. 전통적인 정보검색 시스템은 문서들로부터 추출된 명사들의 리스트를 사용하게 되는데, 이 때 추출된 명사들은 다른 명사들과의 관련 의미정보를 풍부하게 가지고 있지 않으므로 특정 분야의 주제들에 관한 단어들을 계층적으로 분류해 놓은 온톨로지를 사용하게 된다[2]. 문서나 웹을 검색할 때 온톨로지를 사용하는 경우, 원하는 정보를 좀더 빨리 사용할 수 있으며 자원을 찾는 정확도를 높일 수 있다.

본 논문에서는 특정 도메인의 문서들을 수집하여 코퍼스를 만들고, 코퍼스내의 텍스트 분석 결과를 이용하여 반자동으로 온톨로지를 구축하는 방법을 제안하며 구축된 온톨로지에 정의된 개념들을 문서의 검색에 이용함으로써 검색의 정확도를 향상시킬 수 있었다.

2. 도메인 온톨러지의 구축

본 논문에서는 특정 도메인 코퍼스 내의 문서들을 학습시킨 결과를 이용하여 온톨로지를 반자동으로 구축하는 방안을 제시하고자 하며 그림 1에서와 같이 네 단계의 구축과정으로 이루어진다. 먼저 관련 도메인 내의 웹 문서들을 구조화하여 코퍼스를 형성한 뒤, 간단한 자연어 처리 과정을 거치고 개념들을 추출한다. 그리고 추출한 전문 용어들을 분석한 결과로부터 계층구조를 구한 뒤, 온톨로지에 추출한 관계들을 추가하게 된다.

2.1 온톨로지의 구조와 표현

구축할 온톨로지의 구조를 정하기 위하여 약품과 관련이 있는 웹상의 문서들을 분석하고 전문가들과 상의하여 질의에 대한 응답을 위해 필요한 개념들과 관계들을 설정하였다. 수집한 문서들은 학습(learning)을 위한 코퍼스를 형성하기 위하여 설정한 구조에 맞게 변환과정을 거친다. 그 결과 문서들은 태깅된 텍스트들로 이루어지게 되며 태

그들은 해당 약품에 대한 개념들을 각각 형성한다. 구축할 온톨로지에 존재하는 개념들과 관계들은 OWL[3]을 이용하여 표현하였다.

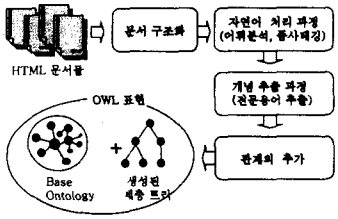

그림1. 온톨로지의 구축과정

문서내의 태깅된 텍스들은 간단한 자연어 처리 파서를 이용한 텍스트 분석 과정을 거치게 되는데, 추출한 명사들은 온톨로지의 개념을, 문서에 부착된 태그들과 동사들은 개념들간의 관계를 나타낸다.

2.2. 전문용어의 처리

실험 대상 문서인 약품 매뉴얼을 분석한 결과, 병명이나 증세 등을 나타내는 전문 용어들이 많이 출현함을 알 수 있었는데 이는 전문적인 지식을 포함하는 도메인의 특성 때문으로 추측된다. 약품 도메인에 출현하는 대부분의 전문 용어들은 복합명사의 형태로 나타나며, 크게 두 가지의 결합형태로 나타난다. 하나는 띄어쓰기가 없는 단일 어절(singleton term) 형태이고, 다른 하나는 띄어쓰기가 나타나며 앞의 어절 성분과 의미적으로 관련이 있는 두 어절 이상으로 이루어진 복합명사 형태이다.

2.2.1 단일어절의 형태

약품 도메인 내에서 전문용어를 이루고 있는 복합 명사들은 대부분 한자어로부터 파생된 경우이며, 본 논문에서는 복합명사를 구성하기 위해 결합되는 명사나 접미사를 20가지로 분류하였다. 이들은 " 염, 증, 통, 균, 성, 질환, 속, …" 이며, 의미적으로 관련이 있는 전문 용어들을 서로 연결시킨다. 실험 도메인에서 단일어절의 형태로 출현하는 전문 용어들은 " 방광염, 기관지염" 과 같이 접미사(감염증을 나타내는 " 염")의 하위 단어인 경우가 대부분이다. 따라서 이들을 "hyponymOf" 관계로 연결하고 하위관계를 자동으로 추출하기 위한 알고리즘을 제안하였다[4].

2.2.2 다중어절의 형태

실험 텍스트에 나타난 전문 용어들은 대부분 " 만성위염" 과 같이 수식어와 중심어의 관계를 가지며 중심어가 다시 단일어절로 이루어진 전문 용어인 경우가 많이 출현하였다. 본 논문에서는 이 문맥 관계들을 다섯개의 관계 패턴들로 설정하고 이에 따라 온톨로지 내에서의 의미관계를 추가하고 설정하였다.

3. 문서검색에의 응용

구축된 온톨로지는 다양한 분야에서 활용될 수 있는데, 본 논문에서는 사용자의 요구를 효과적으로 검색하는 문서 검색의 정확도을 높이기 위한 방안으로 이용하고자 한다. 문서검색에서 온톨로지를 사용할 경우, 사용자는 찾아진 검색어와 관련된 온톨로지내의 하위개념들까지 조사하게 된다. 즉 사용자는 온톨로지를 이용함으로써 자신이 입력한 단어가 검색하고자 하는 목표개념을 적절히 반영한 단어인지를 확인할 수 있으며, 필요에 따라 검색어를 수정하거나 추가함으로써 검색의 효율을 높일 수 있다.

본 논문에서는 입력으로 들어온 질의어에 대한 가중치를 부여하고 문서의 순위를 결정하기 위하여 단어들의 상대빈도수($Rf_{i,j}$)와 하위빈도수($Hf_{i,j}$)를 이용하며 다음과 같이 정의한다.

n : 문서 d_j에 나타난 단어들의 수
m : 단어 k_i에 대한 온톨로지내 하위단어들의 수
$freq_{i,j}$: 문서 d_j에 나타난 단어 k_i의 빈도수

$$Rf_{i,j} = \frac{freq_{i,j}}{\sum_{i=1}^{n} freq_{i,j}} \qquad Hf_{i,j} = \frac{\sum_{h=1}^{m} freq_{h,j}}{freq_{i,j}}$$

4. 실험 및 평가

약품 도메인 내 21,113개의 실험 문서를 대상으로 하여 추출된 명사들의 수는 총 76,782개이며 이 중 전문용어의 수는 55,870개로 70.81%를 차지하였다. 특정명사나 접미사와 결합한 단일 어절로 나타나는 전문용어를 인식한 결과 1,864개의 하위개념을 추가하고 평균 92.57%의 정확도를 보였으며, 다중 어절로 나타나는 전문용어의 경우에는 평균 66.64%의 정확도를 보였다. (그림2, 3 참조)

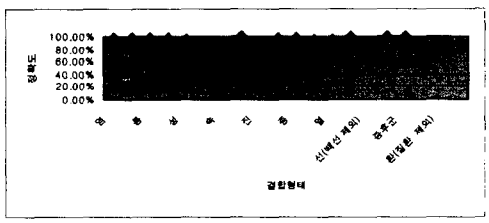

그림2. 단일어절형태 전문용어들의 정확도

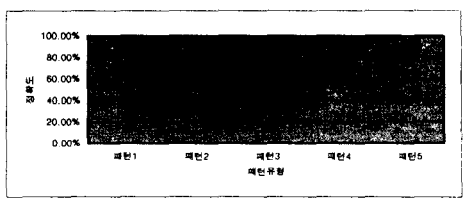

그림3. 다중어절형태 전문용어들의 정확도

본 논문에서는 전통적인 키워드기반 문서검색과 구축한 온톨로지에 기반한 문서검색의 결과를 비교하였다. 질의어" 중이염" 에 대한 예를 들어보자. 표1은 키워드기반 검색시 추출된 단어의 수와 빈도의 분포를 주며, 표2는 온톨로지 내의 하위어들까지 검색에 이용하여 추출한 단어의 수와 빈도의 분포를 보여준다. "만성 중이염", "만성 유착성 중이염" 등과 같은 36개 중이염의 하위 단어가 추가되면서 더욱 상세한 검색이 이루어짐을 알 수 있었다. 그림4는 추가된 하위어의 일부를 개념 그래프로 나타낸 것이다

표1. 키워드 기반에 의한 검색의 경우 추출된 단어수와 분포

	추출된 단어수	키워드 출현빈도	전체단어 출현빈도	키워드 점유율
doc1	81	10	142	7.04%
doc2	254	40	545	7.34%
doc3	88	18	140	12.86%
doc4	176	27	347	7.78%
doc5	102	8	171	4.68%
doc6	183	19	313	6.07%
doc7	126	18	268	6.72%
doc8	129	15	249	6.02%
doc9	171	27	302	8.94%
doc10	216	23	463	4.97%
추출된 총 단어수	1526	205	2940	

표2. 온톨로지내 하위어들을 이용한 검색의 경우 추출된 단어수와 분포

	추출된 하위단어수	하위어고려 출현빈도	전체단어 출현빈도	점유율
doc1	5	20	142	10.56%
doc2	6	49	545	23.36%
doc3	3	34	140	8.79%
doc4	2	13	347	6.14%
doc5	3	24	171	6.87%
doc6	3	30	313	8.96%
doc7	2	17	268	6.43%
doc8	0	27	249	8.94%
doc9	4	30	302	5.72%
doc10	8	68	463	9.91%

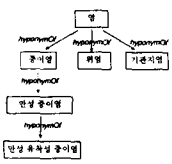

그림4. 온톨로지 내의 추가된 하위어들의 예

입력으로 들어온 질의 각각에 대한 관련문서 200개를 선택하고 이들로부터 10인의 전문가들이 30개의 정답문서집합을 만들었다. 이를 기준으로 10개 질의어에 대한 평균 정확도를 구하였으며 아래의 그림5와 같다. 그 결과, 키워드기반 검색의 평균 정확도는 43.47%이고, 온톨로지기반 검색의 평균 정확도는 57.75%였다.

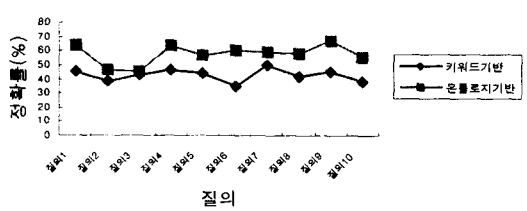

그림5. 키워드기반 검색과 온톨로지기반 검색의 정확도 비교

5. 결론

본 논문에서는 도메인 특정적인 온톨로지의 구축을 위하여 코퍼스의 분석결과를 이용한다. 온톨로지의 구축에 필요한 개념과 관계들을 추출하기 위하여 전문용어를 구성하고 있는 특정명사나 접미사들을 분류하고 이들을 이용한 전문용어의 처리방안을 제시하였다.

구축된 온톨로지는 도메인에 의존적이었으며 부착된 명사나 접미사의 형태에 따라 의미군으로 분류되는 양상을 보여주었다. 단일어절 형태의 전문용어를 인식한 결과, 2,864개의 하위개념을 추가하였으며, 평균 92.57%의 정확도을 보였고, 다중어절 형태의 전문용어들의 경우에는 66.64%의 정확도를 보였다. 구축된 온톨로지를 문서의 검색에 이용한 결과, 일반적인 키워드기반 검색보다 14.28%의 개선된 정확도을 보였다.

이렇게 특정 도메인 내의 텍스트를 분석하여 구축된 온톨로지는 자동으로 개념과 관계를 추가함으로써 좀 더 풍부한 정보를 가지게 되어 다양한 질의에 응답할 수 있으며 검색에 대한 정확도를 높일 수 있었다. 이는 온톨로지에 정의된 개념들과 규칙들이 검색을 향상시키기 위한 추론의 기반으로 이용될 수 있다는 것을 의미한다.

우리는 제안한 온톨로지의 구축방법을 일반 도메인에 확장하여 적용하는 방안에 관해 연구를 계속하고자 한다.

Reference

[1] Gruber, T. : A translation approach to portable ontologies, Knowledge Acquisition, Vol.5, No.2, pp.199-220, 1993.

[2] Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishers, Boston, 2002.

[3] Michael K. Smith, Chris Welty, Deborah L. McGuinness, "OWL Web Ontology Language Guide", World Wide Web Consortium, http://www.w3.org/TR/owl-guide, 2003.

[4] 임수연, 구상옥, 송무희, 임수연, "접미사 패턴을 이용한 온톨로지의 구축방안", 2003 가을 학술발표논문집, pp 547-549, 2003.

[5] 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, Toulouse, France, 2001.

[6] Lim, S. Y., Koo, S. O., Song, M. H., Lee, S, J., "Hub_word based on Ontology Construction for Document Retrieval", IC-AI'03, Las Vegas, USA, 2003.

[7] Michele M., Paola V. and Paolo F., "Text Mining Techniques to Automatically Enrich a Domain Ontology", Applied Intelligence 18, 322-340, 2003.