허브 단어에 기반한 온톨러지의 반자동 구축

임수연[O] 구상옥 송무희 이상조
경북대학교 컴퓨터공학과
nadalsy@hotmail.com [O], tomato@sejong.knu.ac.kr, mhsong@knu.ac.kr, sjlee@knu.ac.kr

Semi-automatic Ontology construction based on Hub word

Sooyeon Lim[0] Sangok Koo Muhee Song Sangjo Lee
Dept. Computer Engineering, Kyungpook National University, Korea

요 약

본 논문은 문서검색을 위한 온톨러지(Ontology)의 반자동 구축방안을 제시한다. 이를 위하여 우리는 다른 단어들과 특히 많은 관련이 있는 단어를 허브 단어(hub word)라고 정의하며 경제분야에 특정적인 온톨러지의 구축을 위하여 TREC 문서집합의 Wall Street Journal 문서들을 분석하였다. 문서집합 내의 모든 단어들의 tf, idf 값를 이용하여 허브 단어를 결정짓고 이렇게 선택된 허브 단어들을 중심으로 온톨러지를 구축하였다. 우리는 허브 단어와 다른 단어들간의 관계를 문서로부터 자동으로 추출하고 그 정보를 이용하여 온톨러지를 확장해나간다. 제안된 온톨러지는 전통적인 문서 검색의 인덱스 파일과 같은 역할을 하게 되며, 간단한 역파일(inverted file) 구조보다 더 많은 의미정보(semantic information)를 제공할 수 있다.

1. 서 론

정보검색 시스템은 문서집합에 있는 정보의 내용을 번역하고, 입력된 질의와 관련된 정보를 찾아내는 것을 목적으로 한다. 대부분의 전통적인 정보검색 시스템은 문서들로부터 추출된 명사들의 리스트를 사용하게 되는데, 이때 추출된 명사들은 다른 명사들과의 관련 의미정보를 가지고 있지 않다. 의미정보를 나타내기 위해, 많은 분야에서 어떤 주제에 관한 단어들을 계층적으로 분류해놓은 온톨러지(Ontology)를 사용하게 된다[1]. 온톨러지란 어떤 특정 도메인(실세계)에서 사용되는 정보들과 그 정보들간의 관계를 정의해 놓은 것을 말하며, 기존의 온톨러지들로는 MikroKosmos[9], HowNet[3], SENSUS[5], CYC[7], 그리고 WordNet[10] 등이 있다. 최근에는 정보검색, 정보교환 등에서 발생하는 많은 전통적인 문제를 극복하기 위한 많은 연구가 진행 중이며, 특히 상당한 시간과 비용이 드는 수작업 대신 온톨러지를 (반)자동으로 구축하기 위한 방안이 계속 제시되고 있다[4].

본 논문에서는 경제분야와 관련된 문서검색을 위한 온톨러지를 구축하는 방법을 제시하고자 한다. 우리는 온톨러지를 단어들로 연결된 일종의 네트워크로 생각하며 다음의 과정에 따라 구축한다. 첫째, 문서집합에서 높은 출현빈도를 가진 단어들을 발견할 수 있었는데 이들 단어들이 문서 내에서 다른 많은 단어들과 유기적으로 연결되어 있다고 가정한다. 둘째, 우리는 이들 단어들을 허브 단어(hub word)라고 규정하고 이들 단어들을 이용하여 기초적인 네트워크를 구축한다. 마지막으로, 선택된 허브 단어들과 관련이 있는 단어들을 네트워크에 추가함으로써 온톨러지를 확장해나가게 된다. 이를 위해서 우리는 TREC 문서집합에 있는 Wall Street Journal 문서들을 분석하였다.

본 논문의 구성은 다음과 같다. 2장에서는 허브 단어들을 추출하는 방법과 그를 중심으로 반자동으로 온톨러지를 구축하는 방법에 대해 논하고, 3장에서는 구축된 온톨러지를 문서검색에 응용하는 과정을 간단히 보여주며, 4장에서는 구축된 온톨러지와 그에 대한 실험결과를, 그리고 마지막 장에서는 본 논문에 대한 결론을 맺는다.

2. 제안된 온톨러지의 구축

이 장에서는 본 논문에서 제안하는 온톨러지의 구축과정에 대해 논하고자 한다. 2.1절에서는 허브 단어에 대한 정의를 살펴보고, 2.2절에서는 허브단어들을 중심으로 온톨러지를 구축해가는 과정에 대해 기술한다.

2.1 허브 단어의 추출

우리는 온톨러지를 많은 어휘들로 이루어진 네트워크로 여기기로 한다. 대부분의 네트워크에서는 아주 많은 링크들로 연결된 소수의 노드들이 존재하며, 이들은 전체 네트워크에서 중요한 역할을 하게 된다. 본 논문에서는 웹상이나 문서집합에서 많은 다른 단어들과 관련되어 있는 이들 단어들을 허브 단어라 부르기로 하며, 허브 단어들을 중심으로 문서 검색을 위한 기초적인 온톨러지를 구축한다. 이 때 허브 단어를 결정하기 위하여 우리는 문서에 나타난 단어들의 *tf* (term frequency:단어 빈도수) 값과 *idf* (inverted document frequency:역문헌 빈도수) 값을 계

산한 뒤, 높은 tf 값을 가지는 단어들을 허브 단어들로 선택한다. 그 결과, 우리는 각 전문분야에 따라 허브 단어들이 다른 것을 알 수 있었다. 예를 들면, 단어 'company'는 경제분야의 문서 내에서 'share', 'trade', 'stock', 'loan' 등의 단어들과 많은 관련이 있지만, 의학분야의 문서 내에서는 이들 단어들과 별로 연관되어 있지 않다. 즉, 허브 단어들은 특정 도메인에 의존적이며, 본 논문은 Wall Street Journal 문서들을 이용하여 경제분야의 허브 단어들을 추출해내고 이를 중심으로 온톨러지를 구성해나가고자 한다.

2.2 온톨러지의 구축 과정

이 절에서 우리는 추출된 허브 단어들을 중심으로 온톨러지를 어떻게 구축하는가에 관해 이야기 하고자 한다. 만약 우리가 시소러스나 사전 등과 같은 기존의 자원을 기반으로 온톨러지를 구축해 간다면 인간의 직관에 의존하는 계층구조에서 벗어나지 못할 것이다. 따라서 우리는 온톨러지를 구축할 때 기존의 자원을 이용하지 않고 문서를 분석하면서 온톨러지를 구축하고 확장해나간다.

표1은 2.1절에서 선택된 몇 개의 허브 단어들과 주변에 나타나는 명사들의 목록을 나타내며 이를 중심으로 우리는 네트워크를 구축한다. 이 네트워크에서 각 단어들은 하나의 노드로 표현되며 주변에 나타나는 명사들과 관련이 있어 링크로 연결된다.

표1. 허브단어와 주변에 나타나는 명사들의 목록

허브단어	주변에 나타나는 명사들
company	million, mr, year, share, stock, company, quarter, earning, market, sale, business, insurance, oil, spokesman, bank, billion, years, debt, companies, products, corp, officials, group, time, plan, analysts, plans,, staff.....
trade	market, stock, mr, board, futures, deficit, prices, japan, group, year, agreement, stocks, investors, trade, exchange, billion, trading, dollar, industry, chicago, shares, world, program, london, bank, ...
stock	exchange, market, york, trading, company, shares, price, prices, million, stock, cents, mr, investors, year, tokyo, futures, fund, dividend, issues, volume, cash, times, earnings, options, record, bond.....
staff	mr, house, members, company, john, chief, member, year, president, committee, employees, million, reporter, people, time, cuts, commission, wall, plans, reductions, support, work, office, sec, bank, headquar.....

그림1은 구축된 온톨러지의 일부분을 보여주는데 선택된 허브 단어들은 Company, Stock, Trade, staff…이다. 이들은 서로 관련되어 있으며, 많은 다른 단어들과도 관련을 가지고 있다. 그러나, 그림1은 의미관계가 첨가되어 있지 않은 단순한 구조로 완전한 온톨러지로 볼 수 없으며, 우리는 수동으로 의미관계를 넣어주고자 한다.

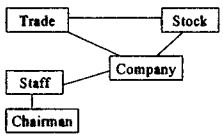

그림1. 허브단어에 기반한 온톨러지 구조의 예

관계를 정의하기 위해 우리는 먼저 문서 내에 있는 모든 동사들을 추출한다. 명사와 마찬가지로 출현 빈도가 높은 동사들을 선택하고 우리는 50개의 관계를 설정하였다. 그림2는 관계가 추가된 온톨러지의 예를 보여준다.

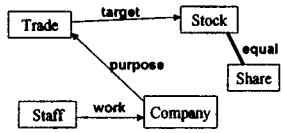

그림2. 관계가 추가된 온톨러지 구조의 예

기본적인 온톨러지(Ontology Base)를 구축한 뒤에는 다음의 단계에 따라 관계를 추가하게 되는데, 먼저 허브 단어 주변에 있는 명사들을 추출한 다음 그 명사들 사이의 관계를 relation extraction rule에 의하여 설정한다. 예를 들면, 동사 'belong', 'include'의 동사가 어떤 두 명사 사이에 존재한다면 우리는 온톨러지에 관계 'belongTo'를 추가한다. 대부분의 경우, 명사들은 온톨러지의 개념을, 동사들은 개념들 사이의 관계를 나타낸다. 그림3은 허브단어 'company'와 연결된 개념들과 그들의 관계를 추가한 결과를 대략 나타낸 것이다.

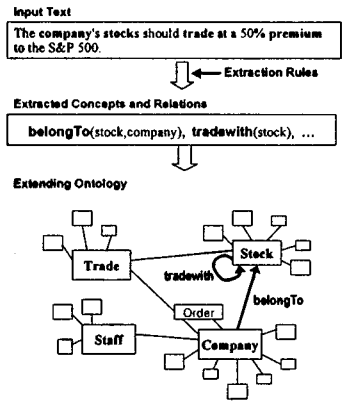

그림3. 허브단어 주변의 개념과 관계의 설정

3. 온톨러지의 문서검색에의 응용

이 장에서는 앞에서 구축한 온톨러지가 정보검색에 어떻게 적용되는가를 보여주고자 한다. 우리는 먼저 입력된 질의를 메타 데이타의 기술과 교환을 위한 프레임 워크인 RDF(Resource Description Framework) 포맷[6]으로 바꾼 뒤, 온톨러지를 참조하면서 질의에 대한 처리를 해나가게 된다. 그림4는 질의 처리 과정의 간단한 예이며, 그림5와 6은 정보 검색에 사용되는 텍스트의 RDF 표현 예와 간단한 추론의 결과를 보여준다.

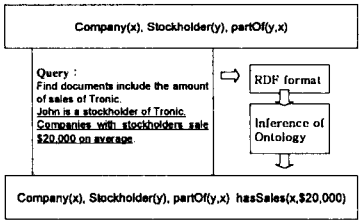

그림4. 질의처리과정의 간단한 예

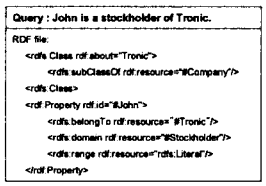

그림5. 텍스트의 RDF표현 예

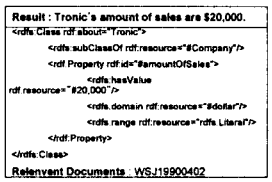

그림6. 간단한 추론의 결과

4. 실험

본 논문에서는 구축될 온톨러지의 근간을 이루는 허브 단어를 결정짓기 위하여 TREC 문서집합에 있는 627,649개의 Wall Street Journal 문서들에 대한 실험을 행하였다. 먼저, 우리는 문서들로부터 명사들을 추출한 뒤, 불용어를 제거하고 각 단어에 대한 원형 복원 과정을 거친 뒤, 각 단어들의 tf, idf 값을 계산하였는데, 우리는 이를 단어의 가중치라고 부르기로 한다. 그 결과, 195개 문서 내에서 높은 가중치를 갖는 단어들과 1,967개 문서 내에서 높은 가중치를 갖는 단어들이 거의 일치함을 알 수 있었으며 표2은 그 결과를 보여준다.

표2. 195개 문서와 1,967개의 문서에 나타나는 단어들의 tf, idf 값의 비교

문서수	195		1,967	
단 어	tf	idf	tf	idf
company	294	0.698288580060364	2951	0.70025741874245O1
trade	125	1.3811792604531203	1485	1.299114307398816
stock	230	1.2299482907291965	2004	1.3121863889661687
bank	178	1.444358162074657	1752	1.4297344898158801
fund	123	1.6354133988373611	1194	1.858730095334238?
share	249	1.0103196815224313	2433	1.060871960685262O
business	88	1.466337068793127	971	1.4067449715911815
credit	33	2.4397862145075306	460	2.177183826452773
thrift	53	2.788092908775746I	240	3.2068032436330315
court	86	1.940795048388543	640	2.2756238992660256
Lawyer	51	2.970414465569701	265	2.8974819960786693
drug	15	2.3285605793973065	304	2.7766455229353957
environment	21	3.886705197443856	245	3.2068032436339315
color	11	3.6635616461296463	16	4.154842673822666S
foot	9	3.181240089335692	47	3.951243718581427S
master	4	4.174387269895637	47	4.2566253681326O9
concept	4	3.886705197443856	39	4.292993012303184S

본 연구에서는 높은 가중치를 갖는 단어들을 허브 단어로 결정하고 이를 중심으로 기본적인 온톨러지를 구축한 뒤 문서의 분석을 통하여 , 단어들간의 관계를 추출하고 그 정보를 온톨러지에 추가하였다. 이 때, 허브 단어와 주변의 다른 단어들이 관련이 있다고 가정하고 관계를 설정해 나가며 온톨러지를 확장해 나간다.

5. 결론

본 논문에서는 문서 분석을 통하여 특정 도메인의 문서들이 소수의 허브 단어들과 많이 관련됨을 알았다. 따라서 허브 단어들을 추출하는 과정과 그를 중심으로 온톨러지를 구축하는 방법을 제안하였으며 온톨러지의 구축과 문서 검색 과정에 대한 간단한 예를 보였다. 구축된 온톨러지는 전통적인 문서 검색의 인덱스 파일과 같은 역할을 하게 되며, 검색에 있어서 역파일보다 더 많은 의미정보를 제공할 수 있다. 앞으로 우리는 정교한 질의의 처리와 문서 검색을 위해 온톨러지를 수정 및 확장해 나갈 것 이며 또한 경제 분야 외의 다른 분야에 대한 온톨러지도 구축해보고자 한다.

Reference

[1] Baeza-Yates, R., Robeiro-Neto, B. Modern Information Retrieval, 1999.

[2] Connolly, D., Harmelen, F., Horrocks, I., Deborah L. McGuinness , Lynn Andrea Stein DAML+OIL, Reference Description http://www.w3.org/TR/daml+oil-reference, March 2001.

[3] Dong, Z. and Dong, Q. HowNet. http://www.keenage.com/zhiwang/e_zhiwang.html, 1999.

[4] 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.

[5] Knight, K. and Luk, S. K. Building a Large Knowledge Base for Machine Translation. Proceedings of the American Association of Artificial Intelligence Conference AAAI-94. Seatle, WA. 1994.

[6] Lassila, O., Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specification, World Wide Web Consortium, 1999. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222.

[7] Lenat, D.B. Cyc: A Large-Scale Investment in Knowledge Infrastructure. Communication of the ACM, 1995.

[8] Maedche, A. Ontology learning for the Semantic Web. Kluwer,2002.

[9] Mahesh, K., Ontology Development for Machine Translation:Ideology and Methodology, Technical Report MCCS 96-292, Computer Research Laboratory, New Mexico State University, Las Cruces, NM, 1996.

[10] Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. WordNet:An On-line Lexical Database, International Journal of Lexicography., 1990.