위키피디아를 이용한 지식베이스 개념 확장 방법

황명권, 최동진, 김판구
조선대학교 컴퓨터공학과
mghwang@chosun.ac.kr, Dongjin.Choi84@gmail.com, pkkim@chosun.ac.kr

Knowledge Base Population Method using Wikipedia

Myunggwon Hwang, Dongjin Choi, Pankoo Kim
Dept. of Computer Engineering, Chosun University

요 약

다양한 분야에 소속된 사람들이 사용하고 있는 개념들을 기존의 워드넷과 같은 지식베이스가 모두 포함하지 못한다는 한계점이 지적되었다. 본 연구에서는 이를 해결하기 위해 위키피디아 문서집합의 분석을 통하여 해결하고자 한다. 위키피디아는 현재 320만개 이상의 유/무형의 개체에 대한 상세한 설명을 포함하고 있으며, 현재도 해당 분야의 전문가들에 의해 지속적으로 제목(주제) 생성 및 내용 작성이 수행되고 있다. 이에, 위키피디아 문서는 지식베이스의 개념 확장을 위해 아주 유용한 자원이 될 수 있으며, 본 논문에서는 이러한 위키피디아 문서 제목의 개념화를 통해 기존의 지식베이스와 연결하는 의미적인 방법을 기술한다. 이를 이용한 간단한 실험을 통하여 본 연구가 우월한 가능성이 있음을 파악하였다.

1. 서 론

무수한 정보들이 웹에 게시되면서 검색의 중요성이 높아지고 있다. 또한 시맨틱 웹(Semantic Web) 기술의 발전은 사람의 요구에 의미적으로 더욱 가깝게 검색 결과를 보여주며, 질의확장[1], 연관검색[2] 등의 추가 기능을 제공할 수 있게 되었다. 이러한 연구 및 서비스들의 대부분은 온톨로지 형식의 지식베이스를 기반으로 하며, 가장 대표적으로는 프린스턴 대학 인지과학 연구실의 워드넷[1]이 있다[3]. 워드넷(WordNet)은 인간의 지식체계와 유사하게 만든 개념 네트워크(concept network)이지만 실세계의 모든 개념을 모두 커버하지는 못한다는 한계점이 지적되고 있다[4, 7]. 이에 다양한 연구들에서 그 한계점을 극복하기 위한 연구들이 수행되었다. Velardi는 각 동일한 도메인의 문서집합에서 도메인 용어를 추출하여 워드넷의 개념과 연결하는 방법을 제시하였다[7]. 또한 Hwang은 대용량의 문서를 분석하여 고유명사를 파악하고 고유명사와 의미적으로 관련된 어휘를 파악하기 위한 연구가 수행되었다[4]. 하지만, 이들이 이용하는 문서집합의 대부분은 웹에 존재하는 문서 또는 논문들로, 추출된 어휘들과 워드넷의 연결에 대한 정확도가 낮다. 이에 본 연구에서는 위키피디아 문서를 이용하여 지식베이스의 개념들을 추가하는 방법을 제안한다. 위키피디아[2]는 현재 320만개 이상의 문서를 포함하고 있다. 그리고 각 문서는 하나의 제목(주제)을 갖고 그것과 아주 관련 깊은 내용으로 기술되고 있다. 또한 시대의 흐름과 함께 생성되는 개념들로 앞으로도 꾸준히 증가할 것이다. 이에 지식베이스 확장을 위해 다른 자원보다 근거가 정확하고 일반적인 개념만을 다루는 기존의 지식베이스보다 더욱 다양한 분야에 적용할 수 있는 장점을 갖는다.

본 연구에서는 위키피디아 문서의 제목을 하나의 개념으로 간주하고 워드넷의 계층구조와 연결하고자 한다. 위키피디아의 제목은 일반 개념, 학문 분야 제목, 알고리즘 제목, 건물 이름, 회사 이름, 제품 이름, 노래 제목 등 하나의 유/무형의 개체를 의미하기 때문에, 이러한 제목들의 개념화는 이미 형성된 지식베이스의 하위개념 또는 인스턴스로 연결될 수 있는 유용한 자원이다.

본 논문에서는 위키피디아 문서의 제목을 추출하여 기존의 워드넷과 의미적으로 연결하는 방법을 2장에서 다룬다. 그리고 3장에서 간단한 실험을 통하여 본 연구의 가능성에 대해 살펴본다. 마지막으로 4장에서 본 연구에 대한 요약과 현재 진행 상태 및 향후 방향에 대해 기술하며 마무리한다.

2. 지식베이스 개념 확장 방법

위키피디아의 제목을 하나의 개념으로 간주하고 기존에 형성된 지식베이스에 추가하려는 본 연구의 전체 과정은 다음과 같이 5가지로 구성된다.

(a) 위키피디아 문서 집합: 지식베이스 확장을 위해 처리해야할 모든 정보를 포함하고 있다.

(b) e-WordNet: 기존의 연구를 통해 워드넷의 개념 관계쌍을 확장한 것으로 본 연구에서 의미 처리를 위한 지식베이스로 사용된다.

(c) 문맥 정보: 최근 수행된 연구를 통해 위키피디아 문서의 핵심이 되는 정보와 제목을 매칭해놓은 지식베이스로 본 연구에서 상위어 결정을 위한 근거데이터로 사용된다.

(d) 후보 단어 리스트 생성 모듈: 위키피디아 문서 제목의 상위어 후보를 찾는 부분이며, 구문

1) WordNet : http://wordnet.princeton.edu/
2) Wikipedia : http://en.wikipedia.org/wiki/Main_Page

분석기, 카테고리 정보 분석기, 명사구에서 주요부(head) 추출기 및 상위 단어 후보 리스트 생성기로 구성되어 있다.

(e) 지식베이스 확장 모듈: 후보 단어 리스트 생성 모듈에서 전달해오는 후보 단어들에서 워드넷과 매칭하여 개념들을 파악하고, 이들 중에서 최종적으로 하나의 개념을 위키피디아 문서 제목의 상위어로 결정한다.

본 장에서는 논문의 이해를 위해 기존에 수행된 (b)와 (c)에 대해서 간략히 설명하며, 본 연구의 핵심인 (d)와 (e)에 대해서 예제와 함께 기술한다.

2-1. 확장된 워드넷

워드넷은 사람의 지식체계와 유사하게 형성한 지식베이스로 미국 프린스턴대학의 인지과학연구소에서 개발하였다. 의미적 문서처리를 위한 분야에서 워드넷의 역할은 아주 중요하지만, 실세계의 모든 개념 관계쌍(concept-pairs)을 포함하지 못한다는 한계점이 지적되었다 [4, 5, 6, 7]. 이에 워드넷의 의미 관계 망(semantic relation network)를 확장하기 위한 연구가 수행되었으며 [6, 7], [6]에서 확장된 워드넷(enriched WordNet)은 그 규모에 따라 라이트(light)와 헤비(heavy) 버전으로 나누었다. 라이트 버전은 워드넷에 정의된 개념이 갖는 정의구문(glossary)을 분석하여 총 114,400개의 개념 관계쌍을 확장함으로써 총 318,160개(WordNet 2.1 + 확장된 관계쌍)의 개념 관계쌍을 포함하고 있다. 또한 [5]에서는 [6]의 결과에서 중요하지 않은 개념관계쌍이 일부 포함된 것을 지적하고 이를 필터링(filtering)함으로써 워드넷을 포함하여 총 306,022개의 개념 관계쌍을 형성하였다. 이렇게 확장된 각 지식베이스를 Senseval-3를 이용한 WSD(Word Sense Disambiguation) 평가에서 최종적으로 [5]에 의한 라이트 e-WordNet이 가장 높은 정확도를 보였다. 이에 본 연구에서는 후보 상위단어 매칭, 각 단어의 개념파악, 그리고 상위어 결정을 위한 관계성 추출을 위해 [5]의 e-WordNet을 활용한다.

2-2. 위키피디아 문서의 문맥 정보

위키피디아의 특징은 하나의 주제에 대해 정의 및 관련 깊은 내용을 상세하게 기술하고 있는 것이다. 또한 위키피디아는 역사적 사건, 자연, 사물, 장소, 인물, 학문 등 한정되지 않은 주제들을 다루고 있으며, 해당 주제에 대한 전문가들에 의해 지속적으로 작성되고 있다. 이에, 위키피디아는 이미 기존의 개념 관계망만을 다루는 지식베이스와는 다른 새로운 개념의 지식베이스로써 활용도가 높다. 이러한 위키피디아를 지식베이스로써 활용할 수 있도록 DBPedia[3]에서 각 문서의 제목, 초록(short abstract과 extended abstract), 그림, 카테고리, 인포박스(infobox) 등으로 나누어 제공하고 있다. 특히, 위키피디아의 초록 정보는 해당주제에 대한 핵심적인 내용만을 포함하고 있어 의미적인 관계성이 아주 높다. 이러한 근거를 바탕으로 [8]에서는 제목과 초록(extended abstract)을 이용하여 위키피디아 제목-문맥정보 관계쌍을 의미적으로 추출한 연구를 진행하였다. [표 1]은 문서 'Academic major[4]'와 'Amshuverma[5]'의 초록을 분석하여 추출한 문맥정보를 보이고 있다.

[표 1] 'Academic major'의 문맥정보

제목	문맥정보#워드넷센스	문맥가중치
Academic major[4]	student#1	0.654
	university#3	0.344
	major#4	0.312
	study#6	0.259
	education#1	0.222
	college#1	0.181
	curriculum#1	0.147
	…	…
Amshuverma[5]	dynasty#1	0.487
	emperor#1	0.435
	kingdom#3	0.412
	territory#1	0.294
	battle#1	0.190
	…	…

[8]에 의해 추출된 문맥정보는 가중치에 따른 상위 30%의 적합성(relevance) 평가에서 약 82%에 도달하였으며, 이는 기존의 방법보다 11% 이상 향상된 결과였다. 또한 상위 30%에 해당하는 문맥정보는 각 제목에 대해 평균 약 11개의 단어로 구성되어 이를 확장한 연구에 중요한 역할을 수행할 수 있다. 본 연구에서는 이렇게 형성된 위키피디아의 문맥정보를 제목의 상위어 결정에서 관계성 파악을 위한 핵심정보로 사용한다.

2-3. 상위어 후보 리스트 생성

위키피디아 제목을 지식베이스로 확장하기 위해서는 먼저 상위어를 파악해야 한다. 또한 이를 가장 간단하고 정확하게 표현하고 있는 것은 초록에서 첫 번째 문장과 해당 문서의 카테고리 정보라 할 수 있다. 적합한 상위어 후보들을 추출하기 위해, 문장의 구문 분석(syntactic sentence pattern analysis)을 이용한다. 첫 문장의 패턴은 일반적으로 다음과 같은 두 가지 규칙이 대부분이다.

패턴 1: (a)문서 제목 + (b)be동사 + (c)관사 + (d)명사구

(예1) (a)AMSD Ariadna[6] (b)is (c)the (d)first Russian web browser …

(예2) (a)AMSDOS[7] (b)is (c)a (d)disk operating system …

(예3) (a)Amshuverma[5] (b)was (c)the

3) DBpedia: http://dbpedia.org/About

4) Academic major: http://en.wikipedia.org/wiki/Academic_major

5) Amshuverma: http://en.wikipedia.org/wiki/Amshuverma

6) AMSD Ariadna: http://en.wikipedia.org/wiki/AMSD_Ariadna

7) AMSDOS: http://en.wikipedia.org/wiki/AMSDOS

(d)Licchavi king of Nepal ...

패턴 2: (a)문서 제목 + (b)be동사 + (c)종류유형 + (d)명사구

(예4) (a)Ahmad Motevaselian[8] (b)is (c)one of the (d)four Iranian diplomats and commander ...

(예5) (a)Alan Dutton[9] (b)is (c)a member of the (d)Canadian Anti-racism Education and Research Society ...

위의 패턴과 일치하는 문장이 출현하지 않을 경우는 위키피디아의 카테고리 정보를 이용한다. 카테고리의 경우 역시 DBPedia에서 NT 파일로 제공을 하고 있어, 간단한 분석을 통해 정보 추출이 가능하다. [표 2]는 위에 기술한 [표 1]과 패턴 예제의 일부에 대한 카테고리 용어들을 보이고 있다.

[표 2] 제목 및 카테고리 용어

제목	카테고리 용어
Academic major[4]	School terminology
AMSD Ariadna[6]	Windows web browsers, Internet history
AMSDOS[7]	Amstrad CPC, Disk operating systems, Operating system stubs

위와 같이 문장 또는 카테고리에서 추출된 용어들은 하나 이상의 단어로 형성된 명사구 형식을 갖는다. 이러한 각 용어들을 워드넷과 매칭하여 존재하는지 여부를 확인하기 위해, 다음의 과정을 거친다. 이해를 위해 패턴 1에서 이용된 (예1)과 (예3)을 이용하여 설명한다.

(a) 명사구(수식어부+주요부)에서 형성 가능한 주요부를 모두 추출한다.
(예1) {first Russian web browser, Russian web browser, web browser, browser}
(예3) {Licchavi king of Nepal, Licchavi king, king}

(b) 각 단어를 워드넷과 매칭하여 존재하지 않는 것을 제거한다.
(예1) {web browser, browser}
(예3) {king}

(c) 각 집합에서 하나의 원소가 다른 원소를 포함하면 작은 부분의 원소를 제거한다.
(예1) "web browser"⊃"browser" → {web browser}
(예3) {king}

상위어 후보 리스트는 경우에 따라서 2개 이상이 형성될 수 있다. 본 예에서는 문장에서 추출된 경우만을 고려하였지만, 실제로 문장에서 추출되지 않고 [표 2]와 같은 카테고리 용어들을 이용하는 경우도 존재하기 때문이다. 그렇다 할지라도 다음 장에서 하나의 상위 개념을 선정하는 방법은 동일하기 때문에, 본 논문에서는 그 경우는 특별히 다루지 않겠다. 본 장에서 형성된 상위어 후보 리스트를 상위 개념 선정 모듈(2-4)로 전달하여 최종적으로 하나의 개념을 선정할 수 있도록 한다.

2-4. 상위 개념 선정

앞의 과정에서 상위어 후보 리스트를 추출하였다. 위키피디아의 제목을 하나의 개념(또는 인스턴스)으로 간주하고 워드넷의 하위개념으로 연결하기 위해서는, 후보 단어가 표현하는 여러 의미(개념, 센스)들 중에서 하나를 선택해야 한다. 예를 들어, 2-3에서 넘어오는 (예3)의 후보 "king"은 10개의 의미를 표현하고 있다. 본 장에서는 2-2에서 기술한 위키피디아의 문맥 정보를 이용하여 후보 단어의 개념들과 관계성을 파악하여 최대값을 갖는 것으로 그 상위 개념을 결정한다. 다음은 이 과정을 기술하고 있다.

(a) 워드넷과 매칭하여 후보 단어들의 워드넷 센스 리스트(sense list, $SL=\{s_i,\ 1\le i\le n\}$, s는 SL의 원소, n은 SL의 크기를 의미)를 형성한다.
$SL_{web\_browser} = \{web\ browser\#1\}$
$SL_{king} = \{king\#1,\ king\#2,\ \dots,\ king\#10\}$

(b) 후보 단어의 개수가 1이면서 SL에 포함된 센스의 개수가 1이면 그 센스를 상위개념(SC, super concept)으로 결정한다. 그렇지 않으면 (c) 단계를 수행한다.
$|SL_{web\_browser}|=1$, SC(AMSD Ariadna) = web browser#1 (a program used to view HTML documents)

(c) SL에 포함된 각 센스와 해당 문서의 문맥정보(context information, $CI=\{c_k,\ 1\le k\le m\}$, c는 CI의 원소, m은 CI의 크기를 의미) 사이의 관계성을 계산하여 최대값을 갖는 센스를 상위개념(SC)로 결정한다.
$SL_{king} = \{king\#1,\ king\#2,\ \dots,\ king\#10\}$
$CI_{Amshuverma} = \{dynasty\#1,\ emperor\#1,\ kingdom\#3,\ territory\#1,\ battle\#1\}$

(c-1) SL의 센스와 CI의 정보 사이의 관계를 e-WordNet을 통해 파악

(c-2) 수식 (1)을 이용하여 상위개념 가중치(W_{SC}) 측정

$$W_{SC}(s_i)=\sum_{k=1}^{m}\frac{1}{shortest\,distance(s_i,c_k)} \quad (1)$$

수식 (1)에 의한 king#1과 king#2의 상위어 가중치는 다음과 같다.
$W_{SC}(king\#1)=1/3(emperor\#1)+1/3(dynasty\#1)+1/2(kingdom\#3)+1/5(territory\#1)\div 1.36$
$W_{SC}(king\#2)=1/5(battle\#1)=0.2$

(c-3) 각 가중치의 결과를 비교하여 최대값을

8) Ahmad Motevaselian: http://en.wikipedia.org/wiki/Ahmad_Motevaselian
9) Alan Dutton: http://en.wikipedia.org/wiki/Alan_Dutton

갖는 것을 상위개념으로 결정한다.
SC(Amshuverma) = king#1 (a male sovereign; ruler of a kingdom)

위의 과정을 통해 위키피디아의 제목과 워드넷의 개념을 연결할 수 있다. 또한 위키피디아에서는 동일한 제목을 갖는 다른 내용이 다수 기술되어 있는데, 본 논문에서 제안하는 방법은 그러한 문제에 대해서도 어려움 없이 구분이 가능하다.

3. 실험 및 평가

현재는 본 연구에 대한 가능성을 보이기 위해 간단한 실험만을 수행하였으며, 그 결과는 본 연구에 참여한 연구자들에 의해 분석되었다. 위키피디아의 문서에서 임의로 50개의 초록(extended abstract)을 추출하고, 2장에서 기술한 과정을 따라 분석하였다. 평가는 정확률(precision rate)과 재현율(recall rate)을 이용하여 계산하였으며, 결과는 [표 3]과 같다.

[표 3] 평가결과

재현율	정확률
50/50	46/50
100%	92%

위키피디아 문서에서 작성된 첫 번째 문장이 대부분 2-3에서 기술한 패턴을 따르며, 그렇지 않은 문장에서는 카테고리 정보를 이용하기 때문에 모든 문서제목의 분류가 가능하였다. 하지만 정확률에서는 일부 잘못 판단한 경우가 발생하였는데, 이는 제목의 일반성 때문으로 분석되었다. 예를 들어, [표 1]에서 예를 보인 'Academic major' 문서에서 추출된 상위어 후보는 'term'이다. 그 후보는 크게 '용어'와 '기간'이라는 의미로 분류되며, 그 문서에서 의도하는 것은 '학교 용어'를 포함하는 '용어'로써 'term'을 의미한다. 하지만 문맥정보에 추출된 것들은 대부분 학교, 특히 대학교와 관련되어 있어 이를 '기간'의 하위로 연결하는 문제점이 발생하였다. 하지만, 다른 연구들과 비교해 볼 때 이는 월등히 높은 수치이며, 방법 또한 간단하면서 의미적으로 접근할 수 있다는 장점이 있다.

4. 결론 및 향후 연구

지식베이스는 다양한 연구에 기반데이터로 활용된다. 하지만 세상의 모든 개념들을 포함하지 못하는 한계점이 지적되어, 본 연구에서는 이를 위키피디아 문서에서 추출하는 방법으로 접근하였다. 위키피디아는 현재 320만 개 이상의 문서제목(워드넷은 8만 여개의 명사개념)을 포함하고 지식베이스 확장을 위해 다른 자원보다 근거가 정확하다는 장점을 가졌으며, 시대의 흐름과 함께 생성되는 개념들로 앞으로도 꾸준히 증가할 것이다. 이에, 위키피디아의 문서 제목의 개념화는 지식베이스의 확장을 위해 아주 유용한 자원이라 할 수 있다.

특히 본 연구를 위해 기존에 수행된 두 가지 연구(확장된 워드넷과 위키피디아의 문맥정보)를 이용함으로써 본 연구의 성능이 더욱 우수해졌다. 현재는 본 연구를 평가하기 위해 단순히 50개의 문서집합만을 이용했지만, 지속적으로 위키피디아 제목의 개념화가 진행되고 있으며, 이에 대한 결과의 확인 작업을 거쳐 워드넷의 일부로 확장할 것이다. 향후 본 연구를 통해 확장된 지식베이스는 다양한 연구에 활용될 수 있는 귀한 자료가 될 것으로 기대된다.

감사의 글

"본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음" (NIPA-2010-(C1090-1011-0009))

참고문헌

[1] Liu, S., Liu, F., Yu, C., & Meng, W. "An effective approach to document retrieval via utilizing WordNet and recognizing phrases", In Proceeding of SIGIR 2004, pp. 266-272, 2004.

[2] Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E., & Milios, E., "Information Retrieval by Semantic Similarity", International Journal on Semantic Web and Information Systems, 2(3), pp. 55-73, 2006.

[3] C. Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press.

[4] Hwang, M.G. and Kim, P.K., A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary. International Journal on Semantic Web & Information Systems. 5(1), pp.48-64, January-March, 2009.

[5] 황명권, 김판구, "워드넷의 의미 관계망 증축 방법", 한국정보기술학회논문지, 7(5), pp. 209-215, 2009년 10월

[6] Hwang, M.G., Choi, C., and Kim, P.K. Automatic Enrichment of Semantic Relation Network and its Application to Word Sense Disambiguation. IEEE Transaction on Knowledge and Data Engineering (will be published)

[7] P. Velardi, A. Cucchiarelli, and M. Petit, "A Taxonomy Learning Method and Its Application to Characterize a Scientific Web Community," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 2, pp. 180-191, Feb., 2007.

[8] Dongjin Choi, Myunggwon Hwang, and Pankoo Kim, "Semantic Context Extraction from Wikipedia Document", The 2010 International Conference on Semantic Web and Web Services (will be published)