Web 검색 엔진의 제목과 문서요약을 이용한 동위어와 문맥의 발견

한상용°, 이상훈
국방대학교 국방관리대학원 전산정보
ugodonly@naver.com, hoony@kndu.ac.kr

Discovery of Coordinate Terms and Context using the Title and Snippet in Web Search

Sang-Yong Han°, Sang-Hoon Lee
Computing Information, Graduate School of Defense Management, National Defense University

요 약

웹상에서의 정보량이 증가함에 따라, 사용자가 알고 싶어 하는 단어에 대해서 연관된 단어를 통해서 이해하게 된다. 동위어란 공통의 상위어를 가지는 단어이다. 이를 위한 기존의 연구로서 동위어와 상위어, 하위어 등을 찾는 연구는 많이 있었지만, 웹상의 문서를 이용하여 거대한 코퍼스를 해석해서 결과를 구하는데 많은 시간이 소요되었다.

이에 본 논문에서는 사용자의 질의어에 대해서 웹 검색엔진이 가지는 제목과 문서요약으로부터 동위어와 문맥을 빠른 시간 안에 발견하는 방법에 대해 제안한다. 어떤 단어에 대한 동위어가 병렬조사 '와'로 접속되는 것을 이용하여 웹 검색 엔진에 대한 질의어를 작성하고, 그 검색 결과로부터 동위어를 얻는다. 이와 동시에 발견된 동위어와 질의어의 배후에 있는 문맥도 얻는다. 이를 통해, 웹 검색에 있어서 질의어의 확장과 비교 대상의 발견 등 폭넓은 분야에서도 적용가능하다고 할 수 있다.

1. 서 론

우리가 어떤 단어를 알고 있다고 하는 것은 그 단어 또는 그와 관련된 다른 단어들을 알고 있다는 것이다. 단어와 단어사이의 관계는 매우 중요하기 때문에, 많은 사전에서 그러한 관계를 설명하고 있다. 또한, 어떤 단어와 관련된 단어를 발견하는 연구도 많이 행해지고 있다. 하지만 웹상의 문서를 이용하여 거대한 코퍼스를 해석해 결과를 구하는데 많은 시간이 소요되고 있다. 검색하는데 너무 많은 시간의 소모로 인하여 효율성이 적어진다.

우리는 시간을 줄이기 위하여 웹 검색 엔진이 가진 정보(제목과 문서요약)로부터 정보를 획득하는 Quick 마이닝 기법을 제안한다.

단어의 관계를 나타내는 개념에는 여러 종류가 있다. 여기서 단어는 단일어 혹은 복합어를 가리킨다. 예를 들면, '삼성'과 '삼성전자'는 모두 하나의 단어이다.

1.1 상위어, 하위어, 동의어, 유의어, 동위어

상위어와 하위어는 상위개념과 하위개념을 나타내는 단어이다. '영어'의 상위어는 '외국어'이며 '영어'는 '외국어'의 하위어이다.

동의어는 완전히 같은 의미를 가지는 단어이다. 예를 들어, '선물'과 '증정품'에서 '선물'은 동의어이며 문맥에 관계없이 그것들을 바꿔 넣어도 의미는 변하지 않는다.

유의어는 '장치'와 '설비'라는 의미가 비슷한 단어이다.

동위어는 어떤 단어에 대해서 공통의 상위어를 가지는 의미가 다른 단어이다. 예를 들어, '불어'와 '영어'는 공통의 상위어로서 '외국어'가 있기 때문에 동위어이다. 공통의 상위어를 가지는 단어에는 동위어 외에도 동의어나 유의어가 있기 때문에, 공통의 상위어를 가지는 단어로써 의미가 다른 단어만을 동위어라고 한다.

본 논문에서는 웹 검색 엔진이 가지는 정보(제목과 문서요약)만을 이용하여 동위어를 발견한다.

1.2 동위어를 발견하는 방법

동위어를 발견하기 위한 질의어가 '서울대'인 경우에 아래와 같은 절차가 적용된다.

<표 1-1> 동위어를 발견하는 방법

(1) 사용자가 질의어로 '서울대'를 입력한다.
(2) 질의어의 앞뒤에 병렬조사 '~와/과'를 덧붙여 질의어를 2개씩 작성한다. 예, '서울대와~', '~와/과 서울대'
(3) 검색 엔진을 통해서 검색한다.
(4) 검색의 결과(제목과 문서요약)를 얻는다.

이러한 절차를 거쳐 질의어와 병렬조사 '~와/과'로 접속된 단어수를 순위화 한다.

1.3 질의어와 동위어의 배경

질의어에서 발견된 동위어는 각각 서로 다른 배경이 있을 수 있다. 예를 들면, 다의어를 가진 '배'는 1)먹는

배, 2)타는 배, 3)신체의 일부인 배 등이 있다. 이러한 관계를 알기 위해서는 질의어와 동위어의 주위 단어들을 표시한다. 이렇게 단어들을 표시함으로써 어떤 단어가 왜 동위어로 판정되었는지 사용자가 이해하도록 돕는다.

동위어의 발견과 이러한 동위어를 이해하도록 제시할 수 있는 문맥은 아래의 3가지이다.

<표 1-2> 동위어의 배경 (3가지)

(1) 질의어와 발견된 동위어가 함께 사용되는 문장 (2) 발견된 동위어를 특징짓는 단어와 TF-IDF 값 (3) 질의어와 발견된 동위어와의 관계

2. 관련 연구

동위어를 얻는 시스템으로 Google Sets라고 하는 서비스가 있다. 몇 개의 질의어를 입력하면, 그것들이 속하는 동위어의 한 무리(15개 이상의 단어)를 찾아낸다.

동위어를 구하는 사전으로서는 Word-Net과 EDR 전자화 사전[2], 언어 공학 연구소의 디지털 유의어 사전(시소러스)등이 있다. 이를 이용하면 상위어나 하위어뿐만 아니라 동위어도 획득할 수 있다. 이러한 동위어의 발견에 관한 연구는 몇 가지가 존재하고 있다.

2.1 상호 정보량을 이용한 단어의 발견

상호 정보량(단어들의 출현회수)이 많은 단어끼리는 동위어일 가능성이 높다고 하는 생각은 여러 가지의 다른 연구에서도 이용되고 있다. Church는 상호 정보량을 이용해 의미적으로 관련이 있는 단어 발견하는 연구[3], Google Zoubin은 Sets와 같은 동위어의 한 무리를 찾아내는 시스템의 작성에 관해 연구[4] 했다.

2.2 유사한 단어의 클러스터링

Lin은 단어와 단어의 연관 관계를 이용하여 유사도를 계산함으로써, 유사한 말의 클러스터를 생성하는 연구[5], Shinzato는 HTML 문서로부터의 동위어를 발견하는 연구[6], Hearst는 "such as"라는 상위어와 하위어가 나타나는 몇 가지의 패턴에 들어맞는 단어를 발견[8]하는 등이 있다.

2.3 동일 / 상 · 하위 개념 추출

Sanderson는 동일한 개념 계층을 추출[9]하고, Glover는 어떤 단어에 대해서 부모의 개념을 나타내는 단어와 자기 자신을 가리키는 단어 또는 아이의 개념을 나타내는 단어를 획득하는 연구[10]를 했다.

2.4 상세어 / 추정검색건수

HTML 문서로 제목 및 본문에서 찾은 간단한 구조 정보를 이용해서 구체적인 단어를 찾는다. Oyama는 웹 검색 엔진(구글, 야후)의 결과로 얻을 수 있는 웹의 추정 검색 건수를 이용하여 웹 검색 엔진 인덱스의 Quick마이닝을 통하여 어떤 단어에 대한 상세어를 발견하는 연구[11]를 했으며, Turney와 Baroni는 유의어를 얻는데 웹 검색 엔진을 이용한 추정 검색 건수를 사용하여 공통적으로 출현하는 단어수와 상호 정보량을 계산하는 방법[12][13] 등의 연구를 하고 있다.

3. 동위어의 발견

3.1 병렬조사 '~와/과'에 주목한 동위어의 발견

본 절에서는 동위어를 발견하는 방법에 대해서 말하는데 아래와 같은 2 가지의 가정에 근거한다.

<표 3-1> 동위어의 발견을 위한 가정

(1) 병렬조사 '~와/과'는 동위어를 접속할 수 있다. (2) 단어 X와 질의어가 병렬조사 '와/과'에 의해서 접속된 "X와/과 질의어", "질의어와 X"의 패턴이 존재할 때, X와 질의어는 동위어일 가능성이 높다.

첫 번째 가정은 동위어가 '~와/과'의 앞 또는 뒤에서 나타날 것이라는 점이다. 병렬조사 '~와/과'에 의해서 접속되는 것은 단일어나 복합어나 관계없다.

두 번째 가정은 기존 연구에서 이용되고 있는 대규모 코퍼스로부터 'A나 B', 'B나 A'라고 하는 패턴이 모두 100회 이상 나오는 단어가 동위어의 정답으로서 평가에 유용하다는 것이다, 아이자와 아키코[14].

병렬조사는 '~와/과'외에도 '~(이)나' 등이 있지만 그것보다는 '~와/과'를 이용한 경우의 결과 값이 나았다.

동위어를 검색하는 단계는 아래와 같다.

<표 3-2> 동위어 검색의 단계

(1) 사용자가 하나의 질의어를 준다. (2) 웹 검색 엔진에 대한 질의어를 2개씩 작성하고 검색하여 결과를 획득한다. "~와/과 질의어", "질의어와/과~" (3) 검색 결과(제목과 문서요약)를 종합한다. (4) 동위어의 후보가 되는 단어를 순위화하여 제시한다.

예를 들면, 사용자의 질의어가 '삼성전자'일 때 질의어는 "삼성전자와~", "~와/과 삼성전자"가 된다. 인용부호로 쌍따옴표가 있는 것은 프레이즈 검색을 위한 것이다.

결과 값은 리스트 형식으로 제시되는데, 각 아이템은 제목(title), 문서요약(snippet)으로 제시된다.

질의어를 기본으로 웹을 검색해서 얻은 제목과 문서요약이 해석의 대상이 되는 텍스트이다.

해석하는 텍스트 중에서 "삼성전자와~"의 '와' 뒤의 단어와 "~와/과 삼성전자"의 '와' 앞에 나오는 단어를 얻을 수 있다. 결과 값으로 공통적인 단어가 발견되면, 동위어로 간주한다. 예를 들면, '삼성전자'의 예에서는 아래와 같은 문장이 검색 결과의 문서요약으로 발견된다.

<표 3-3> 질의어 '삼성전자'의 결과 값 (문서요약)

(S1) **삼성전자와 LG전자**의 기업이미지와 브랜드 이미지에 대해서...
(S2) 중국에 진출한 우리나라의 대표적인 전자회사인 **LG전자와 삼성전자**의 중국시장 진출 사례를 바탕으로...

이 경우, 'LG전자'라고 하는 단어가 '삼성전자'와 병렬조사 '~와/과'라는 단어로 앞뒤에 나오는 것을 알 수 있다. 그 때문에 'LG전자'라고 하는 단어를 '삼성전자'에 대한 동위어라고 생각한다. 이러한 '~와/과'에서 동시에 나오는 단어가, 둘 중에서 하나만 나오는 단어에 비해 동위어로서 적합하다고 할 수 있다.

<표 3-4> '삼성전자'의 '~와/과'의 전후 단어수

동위어의 후보	'~와/과'의 단어수	'와~'의 단어수
LG	0	2
LG전자	11	5
LG전자가	0	1
LG전자는	0	1
LG전자의	0	3
LG전자의 비교	0	1

병렬조사 '~와/과'의 앞뒤 단어수를 각각 더하고, 한쪽만 출현하는 단어는 동위어에서 제거했다.

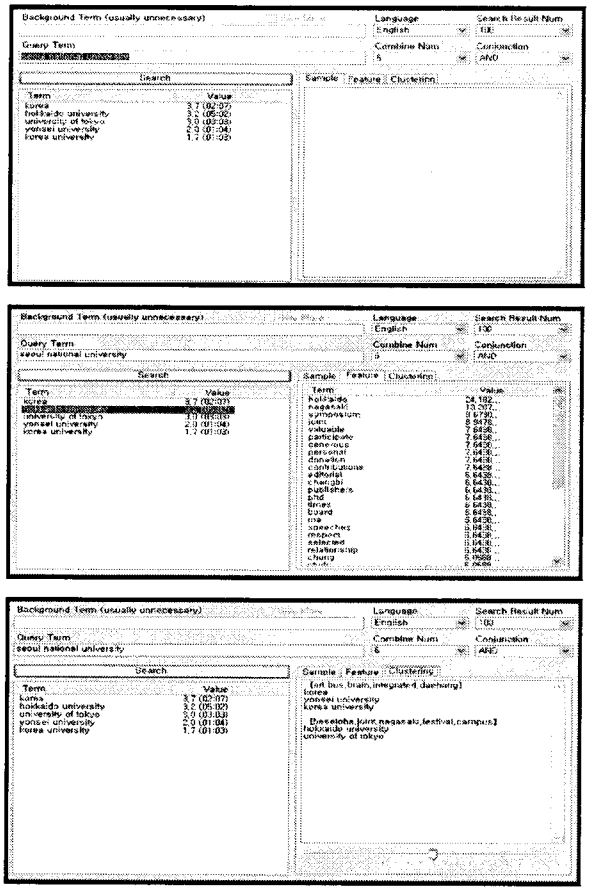
표 3-4는 '삼성전자'라고 하는 질의어가 주어졌을 때에 "삼성전자와~", "~와/과 삼성전자"라고 하는 질의어로 웹 검색하여, 각각 100건씩을 획득한 결과 값이다. 웹의 제목과 문서요약만으로 병렬조사 '~와/과'의 앞뒤에 나타한 단어와 그 단어수의 결과를 보여준다. 복합어로서 올바른 'LG전자'만이 병렬조사 '~와/과'의 양쪽 모두에서 나타났다.

이와 같이 '~와/과'의 앞과 뒤 모두에서 나타나는 결과 값을 기하평균(= (A*B)^0.5)을 이용하여 양쪽으로 균등하게 나타나는 점수가 높도록 순위화했다.

3.2 평 가

3.2.1 평가 방법

평가 방법으로는 웹 검색 엔진(구글)을 이용하여 동위어를 찾는 실험을 했다. 웹 검색 엔진의 결과를 획득하기 위해서는 프로그램을 Google API나 Yahoo! 웹 검색 서비스 등에서 이용할 수 있다. 이번에 작성한 시스템에서는 Google API를 이용했다. 또 시스템이 작성한 2가지의 질의어에 대해서 웹 검색 엔진으로부터 획득한 검색 결과의 최대수는 각각의 질의어에 대해서 100건으로 제한했다.



(그림 1) 검색 시스템 : 질의어 / 특징어 / 클러스터링

3.2.2 평가에 이용한 검색어

우리가 검색어로 이용한 것은 필수시사용어로 143개 단어이다. 현대시사용어사전을 통하여 현재 화제가 되고 있는 사건에 관한 용어이다.

다만, 복수의 단어이거나 설명하는 단어 등 한 단어로 간주하기에 부적합할 경우 적절한 단어로 조정하여 평가에 이용하는 단어를 얻었다.

<표 3-5> 평가에 이용한 검색어

(1) 단어가 함께 사용 되고 있는 경우는 최초의 단어를 선택했다. '양자회담과 다자회담'에서는 '양자회담'을 '대포동 1·2호 / 노동 1호 / 스커드 미사일'에서는 '대포동 1호'를 이용했다.
(2) 한 단어로 간주할 수 없을 경우는 주제라고 생각하는 단어를 선택했다. '급여 소득 공제의 재검토'에서는 '급여 소득 공제'를 이용했다.

3.2.3 정오의 판정

결과 값에 대하여 수동으로 정오의 판정을 실시했다.

<표 3-6> 정오 판정시 주의점

(1) 정답은 질의어와 공통의 상위어를 갖는다.
(2) 동의어나 유의어는 오답
(3) 질의어와 부분 관계(part-of)에 있는 단어는 오답
(4) '배'가 질의어일 때 공통의 상위어로서 '과일'을 가지는 '사과'와 '운송수단'을 가지는 '자동차'를 각각 정답으로 한다.
(5) 공통의 상위어는 일반적인 상위어이어야 한다.

예를 들면, '서울대'와 'Google'은 공통의 상위어로서 '조직'을 가진다고 생각할 수도 있지만, 일반적인 상위 단어보다 높기 때문에 이런 경우는 오답으로 한다.

<표 3-7> 동위어 발견의 실험 결과의 정리

구 분	143개 단어	동위어
동위어 출력어수	1086	7.59
총 정 답 수	847	5.92
총 오 답 수	239	1.67
적 합 율	78.0%	69.1%

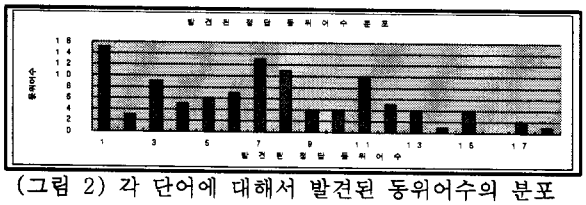
3.2.4 평가

표 3-7은 평가한 결과 값이다. 우선 143개 단어에 대해서 결과 값으로 출력된 동위어수는 1086개 단어로 질의어 당 7.59개 단어가 출력되었다. 그 중에서 정답으로 판정된 총수는 847개 단어, 질의어 한 단어에서는 평균 5.92개 단어였다. 오답으로 판정된 것은, 총회수가 239개 단어로 질의어 한 단어 당 평균 1.67어였다

얻을 수 있는 동위어의 총 회수 1086개 단어에 대해서 정답으로 판정된 단어가 847개 단어로, 전체 적합율은 78.0%였다. 또한 143개 단어 각각의 결과에 대해 적합율을 구하여 적합율의 평균을 구하면 69.1%가 되었다.

143개 단어 중에서 동위어로서 정답을 얻을 수 없었던 질의어 21개 단어는 전체의 14.7%였다. 이런 21개 단어는 웹 검색으로 동위어 '~와/과'를 사용해도 검색 결과를 얻을 수 없었던 경우이다. 정답으로 얻을 수 없는 단어의 상당수는 아직까지 별로 사용되지 않는 단어이다. 예를 들면, '특별 긴급 과세', '섬유 원료 기준', '화성 7호' 등은 웹 전체검색에서도 30건 미만으로 발견되었고, '~(이)나'를 덧붙이면 거의 웹 검색의 결과를 얻을 수 없는 것이다

'10종 경기'와 같이 단어로서는 어떤 정도의 지명도가 있어도 동위어가 원래 적은 경우도 있다. 몇 개의 단어는 분할하면 동위어를 발견할 수 있게 되는 것도 있다. 예를 들면, '세계 문화 유산 기록'을 '세계 문화 유산'이라고 하면 '동북 공정', '민속 문화', '세계 자연 유산'이라는 동위어를 획득할 수 있다.



(그림 2) 각 단어에 대해서 발견된 동위어수의 분포

그림 1은 각 단어에서 발견된 정답 동위어수의 분포이다. 143개 단어 중 정답수가 0인 것은 21개 단어로 전체의 14.7%이다. 올바른 동위어를 가장 많이 얻은 경우는 17개 단어가 있었다.

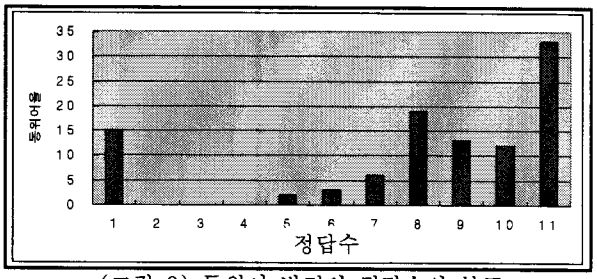
표 3-7에서는 평균 정답수가 5.92개 단어이지만 이 그래프에서 보면, 정답수가 6~7개 단어의 분포가 많은 것을 알 수 있다.

그림 2는 동위어로서 발견된 단어의 적합율 분포이다. 정답수가 0일 때에는 적합율도 0으로 되고 있기 때문에 적합율이 0인 단어의 비율이 전체의 14.7%가 되고 있다. 정답이 발견되는 질의어의 경우는 결과의 적합율은 대부분 70%이상이다.

전체에 대해서도 반수 이상의 질의어에 대해 적합율이 80%이상이다. 출력된 1086개 단어의 22.0%에 해당하는 239개 단어는 오답으로 판정되었다. 그림 2에서는 적합율이 50%이하인 것을 알 수 있다. 오답이였을 경우를 분류하면 이하와 같은 것이 있었다.

<표 3-8> 오답의 경우

(1) 복합어에서 공통부분의 생략에 의한 오류
(2) 관련 있지만 동위어가 아닌 단어
(3) 관련 없는 단어와 일반적인 단어
(4) 상위어



(그림 3) 동위어 발견의 적답수의 분포

오답의 몇 가지는 질의어와 그 동위어가 모두 복합어라서 뒷부분에 있는 공통부분이 생략된 것이었다. 예를 들면, '태풍 매미'와 '루사'라는 오답이 출력되었다. 문서요약에는 '태풍 매미와 루사'가 3회 '태풍 루사와 매미'가 33회 출현하고 있었다. 여러 분야에 있는 여러 종류의 단어 평균값이 78%의 적합율로 5.92개 단어의 동위어를 얻었다. 이로써 본 시스템의 유효성을 입증했다.

3.3 '~와/과'이외의 병렬조사 비교

병렬조사에는 '~와/과'이외에도 '~(이)나' 등이 있다. '~(이)나' 이용했을 경우에 대하여 알아본다. 표 3-9는 병렬조사'~(이)나'를 사용했을 경우를 나타낸다.

단어에 따라서는 '~와/과'보다 '~(이)나'를 사용하는 것이 좋은 경우도 있겠지만, 동위어를 구하는 데는 '~와/과'가 가장 적합하다고 할 수 있다.

<표 3-9> 병렬조사'~와/과 / ~(이)나'에 의한 결과

병렬조사	질의어 '전자상거래'에서 발견된 단어 (~와/과의 앞에 나온 회수 : 뒤에 나온 회수)
와/과	인터넷(21:10), 기술(8:9), 정보(5:6), 마케팅(8:3),세금(7:3), 비즈니스(1:5), 과세(4:1), 전기통신(3:1), 전자화폐(2:1), 텔레워크(1:1), edi(1:1), 출판업(1:1), 전자출판업(1:1), 지적재산(1:1)
(이)나	e-commerce(30:37), 인터넷(16:7), ecommerce(4:4), internet commerce(3:2), 비즈니스(1:2), IT(1:1), 정부(1:1), internet ; e-commerce(1:1), 마케팅(1:1)

4. 동위어의 문맥 발견

4.1 동위어의 문맥

동위어를 나타낼 때 문맥을 동시에 보여줌에 의해서, 사용자는 각 단어가 왜 동위어로 되었는지 이해하는데 도움을 줄 것이다. 여러 표현에 의한 문맥을 동시에 보이는 것만으로도 그 효과는 충분히 높아진다고 할 수 있다. 발견된 동위어의 문맥은 아래의 3가지로 표현된다.

<표 4-1> 동위어의 문맥

(1) (예문) 질의어와 발견된 동위어가 함께 사용된 문장
(2) (특징어) 동위어를 특징지우는 단어와 TF-IDF 값
(3) (상관관계) 질의어와 발견된 동위어와의 유명도

발견된 모두를 클러스터링 함에 의해서, 발견된 동위어끼리의 유사성도 이해할 수 있도록 했다.

4.2 동위어와 질의어가 나타난 예문

예문은 해석한 텍스트 중에서 동위어와 질의어가 병렬조사'~와/과'로 연결된 문장을 추출한 것이다. 예를 들면, 질의어가 '소니'일 때 '삼성', '엔씨소프트', '삼성전자'등이 동위어로서 나타난다.

'삼성'이 동위어가 된다는 것은 "삼성과 소니"와 "소니와 삼성"이라는 부분을 포함한 문장이 존재하기 때문이다. 예를 들면, '삼성'의 경우는 아래와 같은 예문이 제시된다.

<표 4-2> '삼성과~', '~와/과 삼성'의 예

(1) **삼성과 소니**의 합작사인 S-LCD 설립 2주년을 앞두고 이들 두 기업이 경쟁이 치열한 가전시장에서...
(2) 8세대 협력은 **삼성과 소니** 양측의 이해관계에...

이 두 문장에 의해 삼성과 소니의 합작사가 S-LCD인 점과 협력을 통한 이해관계가 이루어지고 있음을 알 수 있다. 문맥에서 사용되는 단어들을 살펴보면 이해에 도움이 된다고 할 수 있다.

4.3 동위어와 질의어 주변 문장의 특징어

특징어란 발견된 동위어와 질의어가 모두 나타나는 문장과 주변 문장의 문맥을 특징짓는 단어이다. 특징어를 구하는 방법은, 각 동위어에서 '~와/과'라는 질의어로 연결된 문장을 포함한 제목과 문서요약을 모두 모은다. 그것들로부터 모든 단어수 즉, Term Frequency (TF)를 구한다. 이 단어수가 많은 단어가 발견된 동위어와 질의어의 문맥을 특징짓는 단어(특징어)라고 생각할 수 있다.

IDF도 얻은 모든 제목과 문서요약으로 구한다. 질의어의 전후에 '~와/과'를 넣은 2가지의 질의어 각각의 검색 결과를 100건씩 구했을 경우, 합계 200건의 제목과 문서요약에서 term이 출현한 회수를 DF(term)로 하면, 그 단어의 IDF로 계산된다.

```
IDF = log( 200 / DF(term) )
```

예를 들면, 질의어가 '민주사회주의'일때 결과 값은 '공산주의', '사회민주주의', '혁명'등이다. 표 4-3은 특징어인'공산주의'와 '사회민주주의'라는 문맥의 이해를 돕는다.

<표 4-3> '민주사회주의'의 특징어

공산주의		사회민주주의	
값싼정부	19.9	선동	19.9
공산주의	14.2	세속화	13.2
매수	13.2	래스터패리언	13.2
뇌물	13.2	대중	13.2
투쟁	9.6	이데올로기	12.1
과도기	7.6	민주주의	11.3

특징어는 관련된 제목과 문서요약을 이용한 문맥으로 보다 종합적인 단어의 배경을 알 수 있을 것이다

4.4 동위어와 질의어의 상관관계

질의어와 발견된 단어의 상관관계와 유명도도 얻을 수 있다. '~와/과'의 앞뒤의 어디에서 많이 출현했는가에 의해서 나타내진다. '~와/과'의 비율로 나타내지는 값이 1에 가까울 경우에 질의어보다는 동위어가 유명하며, 0에 가까울수록 질의어가 유명하다는 것을 알 수 있다.

<표 4-4> '~와/과' 앞뒤 단어수와 웹의 총페이지수

동위어	질의어	'와'앞	'와'뒤	동위어 웹수	질의어 웹수	'와' 비율	웹 비율
백두산	한라산	42	6	4,370,000	3,700,000	0.88	0.54
덕유산	가야산	64	19	770,000	467,000	0.70	0.62
대둔산	계룡산	5	10	366,000	1,010,000	0.33	0.27
금학산	고대산	25	26	39,900	57,600	0.49	0.40
용문산	남산	2	6	275,000	4,380,000	0.25	0.05

5. 결론 및 향후연구

본 논문에서는 웹 검색 엔진의 정보(제목과 문서요약)만을 이용하여 동위어를 발견하는 방법을 제안했다.

웹 검색 엔진의 정보에는 지식으로써 이용할 수 있는 정보가 많이 존재할 뿐만 아니라, 지식을 쉽고 빠르게 얻을 수 있다.

기존의 연구는 풍부한 코퍼스를 해석해서 대량의 결과를 얻는데 시간이 오래 소모되었지만, 제안하는 방법에

서는 웹 검색 엔진의 정보만을 이용하여 사전 처리 등을 거치지 않고 모든 분야의 단어에 대한 동위어를 빠른 시간 안에 얻을 수 있다.

우선 동위어가 병렬조사 '~와/과'에 의해서 나타나는 사실을 이용하여, 질의어의 앞뒤에 '~와/과', '와/과~'를 덧붙여서 질의어를 만들었다. 검색 결과(제목과 문서요약)만으로부터 질의어와 '~와/과'로 접속되고 있는 단어를 찾아 동위어를 나타낸다.

평가를 위한 검색어를 준비해서 실험한 결과, 평균 70~80% 정도의 적합도로 질의어 당 평균 6개 단어의 동위어를 얻을 수 있었다. 사용자가 질의어와 발견된 동위어를 이해하기 위해서, 동위어와 함께 있는 문맥을 제시했다. 동위어의 문맥으로는 예문, 특징어, 상관관계의 3가지이다. 이를 통해, 사용자의 이해를 돕는다.

향후 연구방향으로는 국내뿐만 아니라 국외의 자료에 대해서도 웹 검색을 통한 질의어 확장과 동일한 종류의 다른 것과 비교 대상의 발견 등을 통하여 실생활에서 응용되도록 보다 실질적으로 연구되어야 할 것이다.

참고 문헌

[1] Miller G. A. Beckwith R. Fellbaum C.Gross D. and Miller K. J.: Introduction to WordNet: An on-line lexical database International Journal of Lexicography Vol.3 No.4 pp.235~312, 1990.

[2] 독립 행정법 인정보통신 연구 기구 : EDR 전자화 사전 2.0 판 사양 설명서 주식회사 한국 전자화 사전 연구소(2001).

[3] Church K.W. and Hanks P.: Word Association Norms Mutual Information and Lexicography Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics pp.76~83, 1998.

[4] Ghahramani Z. and Heller K.: Bayesian Sets Proceedings of the Nineteenth Annual Conference on Neural Information Processing Systems (NIPS2005) , 2005.

[5] Lin D.: Automatic Retrieval and Clustering of Similar Words Proceedings of the 36th annual meeting on Association for Computational Linguistics pp.768~774, 1998.

[6] Shinzato K. and Torisawa K.: A Simple WWW-based Method for Semantic Word Class Acquisition Proceedings of the Recent Advances in Natural Language Processing (RANLP05) pp.493~500 , 2005.

[7] Shinzato K. and Torisawa K.: Acquiring Hyponymy Relations from Web Documents Proceedings of Human Language Technology Conference/ North American chapter of the Asso ciation for Computational Linguistics annual meeting (HLT-NAACL04) pp.73~80, 200.

[8] Hearst M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora Proceedings of the Fourteenth International Conference on Computational Linguistics pp.539~545, 1992.

[9] Sanderson M. and Croft B.: Deriving concept hierarchies from text Proceedings of the 22nd ACM SIGIR Conference (SIGIR'99) pp. 206~213, 1999.

[10] Glover E. Pennock D.M. Lawrence S. and Krovetz R.: Inferring hierarchical descriptions Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM'02) pp.507~514, 2002.

[11] Oyama S. and Tanaka K.: Query Modification by Discovering Topic from Web Page Structures Proceedings of the Sixth Asia Pacific Web Conference (APWeb'04) pp. 553~564, 2004.

[12] Turney P.D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL Proceedings of the 12th European Conference on Machine Lerning (ECML 2001) pp.491~502, 2001.

[13] Baroni M. and Bisi S.: Using cooccurrence statistics and the Web to discover synonyms in a technical language Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) pp. 1725~1728, 2004.

[14] 아이자와 아키코 : 접수 관계를 이용한 유의어·예문 사전 구축법과 대규모 코퍼스에의 적용, 인공지능 학회 제 20회 전국대회 발표논문집, 2E1-5, 2006.