U-WIN 기반의 의미적 정보검색 기술

Semantic Information Retrieval Based on User-Word Intelligent Network

임지희, 최호섭*, 옥철영
울산대학교, 한국과학기술정보연구원*

Im Ji-Hui, Choe Ho-Seop*, Ock Cheol-Young
Ulsan University, KISTI*

요약

사용자가 원하는 정보를 얼마나 정확하게 제시하느냐가 정보검색시스템 성능을 판단하는 기준이 된다. 그러나 동형이의어만을 질의어로 이용한 검색 결과는 동형이의어 각 의미에 관련된 문서가 혼재되어 있거나, 특정 의미에 관련된 문서만 집중적으로 나타나는 현상을 볼 수 있다. 그래서 본 논문에서는 한국어 사용자 어휘지능망(U-WIN)의 관계정보를 이용하여, 질의어의 모호성을 해결하는 의미적 정보검색의 기반이 되는 기술을 제안한다. 실험에서 질의어는 전문분야에 주로 사용되는 동형이의어와 보편적으로 사용하는 동형이의로 구분하고, '질의어+상위어' 형태의 확장 질의어를 설정한다. 그래서 포탈사이트의 웹 문서만을 대상으로 한 정확률은 73.5%, 통합검색의 정확률은 68.7%로 나타났다. 이것은 U-WIN 기반의 의미적 정보검색 기술이 정보검색 시스템에서 효율적임을 알 수 있다.

Abstract

The criterion which judges an information retrieval system performance is to how many accurately retrieve an information that the user wants. The search result which uses only homograph has been appears the various documents that relates to each meaning of the word or intensively appears the documents that relates to specific meaning of it. So in this paper, we suggest semantic information retrieval technique using relation within User-Word Intelligent Network(U-WIN) to solve a disambiguation of query. In our experiment, queries divide into two classes, the homograph used in terminology and the general homograph, and it sets the expansion query forms at "query + hypernym". Thus we found that only web document search's precision is average 73.5% and integrated search's precision is average 70% in two portal site. It means that U-WIN-Based semantic information retrieval technique can be used efficiently for a IR system

I. 서 론

인터넷의 등장·발달로 인해 웹 자원의 양이 방대해짐에 따라, 웹 자원으로부터 사용자가 원하는 정보를 효율적으로 검색·관리하는 기술에 대한 관심이 증가하고 있다. 자원의 방대해짐으로써 사용자에게 유용한 지식·정보를 구별해내는 비용은 증가하며, 사용자가 원하는 정보를 얼마나 정확하게 제시하느냐가 정보검색시스템의 성능을 판단하는 기준이 된다. 특히 질의어가 동형이의어일 경우, 동형이의어만을 질의어로 이용한 검색은 동형이의어의 명확한 의미 판단을 위한 근거가 부족하므로, 검색결과가 동형이의어의 각 의미에 관련된 문서가 혼재되어 있거나 특정 의미와 관련된 문서만 집중적으로 나타난다.

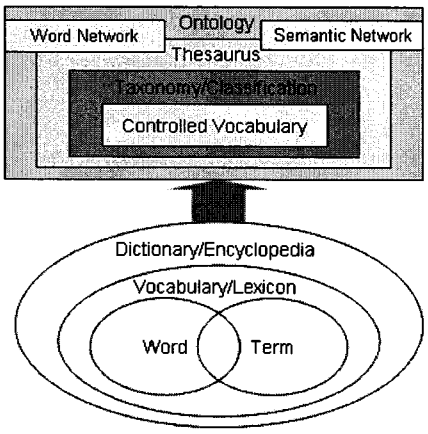
예를 들어, 동형이의어 '배'는 크게 '배(선박)', '배(열매)', '배(신체부위)'의 세 가지 의미를 가질 경우, 단순 질의어 '배'에 대한 검색 결과는 선박, 열매, 신체부위에 관련된 문서가 혼재되어 있다. 혹은 문서 집합에서 '배(열매)'와 관련된 문서가 집중적으로 많거나 사용자에 의해 선택된 빈도가 높을 경우, 검색결과로 '배(열매)'와 관련된 문서를 집중적으로 나타낸다. 이때, 검색하고자 하는 '배'의 의미가 '배(선박)'일 경우라면, 다수의 '배(열매)'와 관련된 문서 속에서 '배(선박)'과 관련된 문서를 찾아내는 데는 많은 시간이 필요하다.

위와 같이 질의어가 동형이의어일 경우 질의어의 의미를 명확하게 구분할 수 있는 특정 키워드를 추가하는 질의어 확장 방법을 사용하고 있다. 그러나 사용자가 직접 키워드를 생성·추가하는 방법은 사용자가 검색하고자 하는 정보에 대한 충분한 지식을 갖추고 있으며, 또한 키워드를 식별하는 능력이 뛰어나야 함을 전제로 한다. 그러나 대부분의 사용자는 질의어에 적합하고 효율적인 키워드를 연상하는데 많은 시간을 소비하거나 어려움을 겪고 있다.

따라서, 시소러스 기반의 정보검색 방법에 관한 연구들이 많이 선행되었다. 시소러스란 "통제된 색인 언어의 어휘집으로, 개념 간의 특정 관계를 형식적으로 조직화하여 명시한 것"으로서, 시소러스의 등가 관계(USE/UF) · 계층 관계(BT/NT) · 관련 관계(RT)와 같은 의미관계를 이용하여 질의어를 확장하는 방법을 적용했다. 그러나 시소러스가 엄밀한 의미의 계층적 상하관계가 아닌 분류적 상하관계로 이루어져 있기 때문에, 시소러스 기반의 정보검색 방법은 의미적 정보검색 방법에 한계점을 가지고 있다. 또한 특정 분야에 한정된 소규모 데이터의 형태로 시소러스가 구축되어 활용도가 낮다.

지식정보 체계화에 대한 관심이 증대하면서, 시소러스 · 어휘망 · 어휘분류 · 온톨로지 등에 대한 연구가 많이 진행되고 있다. 온톨로지는 어휘, 용어, 어휘목록, 사전, 전문분야사전 등과 같은 어휘 집합을 기반으로 하여, 시소러스 · 어휘망 · 어휘분류 등을 포함하는 개념, 관계, 속성 들이 내부적으로 형성된 상의의 지식 구조 체계라 할 수 있으며, [그림 1]과 같은 온톨로지에 대한 인식범위를 설정할 수 있다[4]. 또한 시소러스의 기본 의미관계 뿐만 아니라 기본 개념관계 · 확장 개념관계를 통해 더 많은 의미정보를 포함하고 있다.

그러므로 본 논문에서는 온톨로지적 어휘의미망인 한국어 사용자 어휘지능망의 관계정보를 이용하여 질의어의 모호성을 해결하고 의미적 정보검색 기술의 기반을 마련하고자 한다.



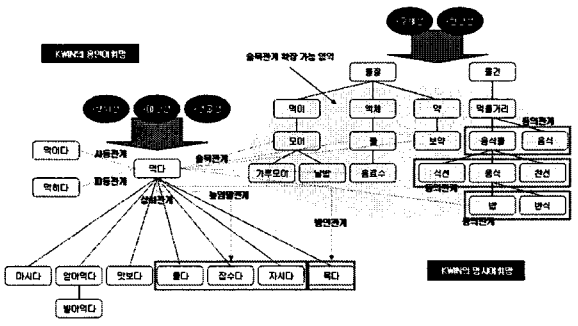
▶▶ 그림 1. 온톨로지에 대한 인식범위

Ⅱ. 지능형 한국어 어휘망

자연 언어의 어휘적 의미, 구문적 의미, 담화적 의미를 바탕으로 행위나 현상, 상태 등에 담긴 의미론적 · 개념론적 특성을 포함하고 있는 의미적 언어 자원 구축에 대한 연구는 다양하게 이루어지고 있다. 국외에서는 WordNet, EuroWordNet, Cyc, HowNet, Lexical FreeNet, EDR 등이 대표적이며, 국내에서는 카이스트의 CoreNet, ETRI의 어휘개념망, 부산대의 KorLex 등이 대표적이라 할 수 있다.

본 연구팀이 2002년부터 개발 중인 한국어 사용자 어휘지능망 (User-Word Intelligent Network, 이하 U-WIN)은 한국어의 공통적이고 개별적인 속성을 바탕으로 한국인의 보편적인 인지 체계와 개념 관계를 파악하여 이를 어휘의 의미적 · 개념적 네트워크로 형성한 온톨로지적 어휘망이라 할 수 있다.

U-WIN은 한국어정보처리를 비롯하여 정보검색, 기계번역, 시맨틱 웹 등 다양한 분야에 이용될 수 있는 대규모 어휘 지식 베이스를 목표로 하고 있다. 현재 온톨로지 기반 의미적 주석(ontology-based semantic annotation)과 유사한 단어 중의성 해소(word sense disambiguation)와 의미태깅(semantic tagging) 기술에 활용되고 있으며, 이 외에도 복합명사 자동 생성, 전문분야별 개념체계 자동 생성, 정보검색에서의 질의 확장, 어휘 학습 시스템 등 다양한 기술에서 활용되고 있다. 나아가 U-WIN 영어 버전을 구축 중에 있어 조만간 한국어를 중심으로 한 한영 대역 U-WIN이 개발될 계획이며 WordNet과의 사상 구조(mapping structure)도 기대할 수 있게 되었다.



▶▶ 그림 2. U-WIN의 구축 사례 일부

U-WIN은 현재 30만 여 어휘가 구축된 상태이다. U-WIN의 구축 대상은 한국어 어휘 전체(모든 품사 및 언어 단위)로서, 핵심적 대상은 명사, 동사, 형용사이며, 부수적 대상은 부사, 관형사, 대명사, 감탄사, 조사, 수사, 의존명사 등이며, 북한어, 방언, 옛말, 전문용어, 고유명사, 어근,

어미 등 한국어 어휘 전체를 대상으로 연구 개발 중이며, (그림 2)는 구축 사례의 일부 모습이다.

Ⅲ. 실험 및 분석

1. 실험 데이터

실험에 사용한 질의어는 두 가지로 구분하였다. 첫 번째는 전문분야 정보검색을 위한 정보검색시스템에서 사용한 80,000여 개의 한글 질의어에서, 질의어 빈도와 '국어 빈도조사'의 빈도를 반영하여 아래 [질의어 선정 기준]에 따라 선정하였으며, 그 중에서도 '국어 빈도조사'에 기반하여 동형이의어 각각의 의미 사용빈도가 균등한 어휘를 선정하였다. 두 번째는 [6]에서 WSD(Word Sense Disambiguation)의 성능을 측정하기 위해 사용된 동형이의어 리스트에서 임의적으로 선정하였다. 실험에 사용된 어휘는 <표 1>과 같다.

[질의어 선정 기준]

- 동형이의어인 어휘
- '국어 빈도조사'에서 빈도가 100이상인 어휘 리스트 중에서 어깨번호가 다른 동형이의어가 2개 이상 나타나는 어휘
- U-WIN에서 상하관계 정보가 있는 어휘
- 특정 의미가 희박하게 사용되는 어휘 제외

[표 1] 실험 어휘 리스트

실험 (1) – 전문분야에서 사용되는 동형이의어 중심으로		
기술	기술01	[명사] ① 과학 이론을 실제로 적용하여 자연의 사물을 인간 생활에 유용하도록 가공하는 수단.
	기술03	[명사] 대상이나 과정의 내용과 특징을 있는 그대로 옮겨하거나 기록하여 서술함. 또는 그런 기록.
동향	동향02	[명사] ①동쪽으로 향함. 또는 그 방향. ≒향동01(向東).
	동향03	[명사] ①사람들의 사고, 사상, 활동이나 일의 형세 따위가 움직여 가는 방향.
시장	시장03	[명사] [법률] 지방 자치 단체인 시의 책임자. 집행 기관으로서 시를 맡아서 다스린다.
	시장04	[명사] ①여러 가지 상품을 사고 파는 일정한 장소. ≒시상02(市上) · 장28(場)(2).

실험 (2) – 보편적으로 사용되는 동형이의어 중심으로		
경기	경기02	[명사] (지명)① 서울을 중심으로 한 가까운 주위의 지방.
	경기05	[명사] (경제) 메매나 거래에 나타나는 호황 · 불황 따위의 경제 활동 상태.
	경기11	[명사] 일정한 규칙 아래 기량과 기술을 겨룸. 또는 그런 일.
다리	다리01	[명사] ①동물의 몸통 아래 붙어 있는 신체의 부분.
	다리02	[명사] ①물을 건너거나 또는 한편의 높은 곳에서 다른 편의 높은 곳으로 건너 다닐 수 있도록 만든 시설물.
배	배01	[명사] ①[의학]사람이나 동물의 몸에서 위장, 창자, 콩팥 따위의 내장이 들어 있는 곳으로 가슴과 엉덩이 사이의 부위.
	배02	[명사] 사람이나 짐 따위를 싣고 물 위로 떠다니도록 나무나 쇠로 만든 물건.
	배03	[명사] 배나무의 열매.

선정된 질의어를 하나의 질의어로 판단하여, 각 질의어당 최대 50개의 웹 문서를 Google과 Naver 의 두 개의 포탈사이트에서 수집하였다. 수집된 문서들의 의미별 분포는 <표 2>에서 살펴볼 수 있다.

[표 2] 수집된 문서들의 의미별 분포

실험 (1) – 전문분야에서 사용되는 동형이의어 중심으로		
질의어	포탈 사이트	검색된 문서의 분포(%)
기술	Naver	수단(100), 서술(0)
	Google	수단(100), 서술(0)
동향	Naver	방향(0), 목표(100)
	Google	방향(0), 목표(86), 기타(14)
시장	Naver	장관(0), 장소(96), 기타(4)
	Google	장관(0), 장소(100)

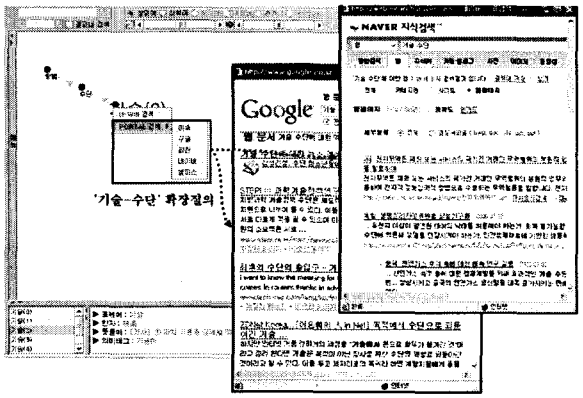
실험 (2) – 보편적으로 사용되는 동형이의어 중심으로		
질의어	포탈 사이트	검색된 문서의 분포(%)
경기	Naver	지방(40), 상태(4), 활동(56)
	Google	지방(78), 상태(0), 활동(12), 기타(10)
다리	Naver	부위(32), 시설물(66), 기타(2)
	Google	부위(38), 시설물(14), 기타(48)
배	Naver	부위(24), 구조물(10), 열매(6), 기타(60)
	Google	부위(0), 구조물(6), 열매(14), 기타(40)

질의어가 전문분야에서 주로 사용되는 동형이의어(기술, 동향, 시장)일 경우 검색 결과가 특정 의미와 연관된 문서가 집중적으로 나타남을 알 수 있다. 예를 들어, 두 포탈사이트에서 질의어 '기술'에 의해 검색된 문서 중 100%가 '과학 이론을 실제로 적용하여 자연의 사물을 인간 생활에 유용하도록 가공하는 수단'의 의미에 해당하는 것으로 나타났다. 그러나 질의어가 보편적으로 사용되는 동형이의어(경기, 다리, 배)일 경우에는 전문분야에서 사용되는 동형이의어인 경우보다 문서의 의미별 분포가 대체적으로 균등함을 알 수 있다.

2. 실험 방법 및 결과 분석

포탈 사이트는 웹문서 검색, 지식검색, 카페/블로그 검색, 뉴스 검색 등 다양한 정보를 제공하므로, 사용자는 웹문서 검색뿐만 아니라, 위와 같은 다양한 정보를 모두 활용한다. 그래서 통합검색 페이지에 나타나는 정확한 정보를 나타내는 것도 포탈사이트의 중요한 문제이다.

그러므로 실험 방법은 각 질의어의 상위어를 2차 키워드로 추가하여 확장 질의어를 생성한 후, 동일한 방법으로 웹문서를 수집할 뿐만 아니라, 통합검색 결과도 수집하였다. 이때, 2차 키워드로 상위어를 선택할 때 몇 가지 주의해야 할 점이 있다. 우선 '시장04'의 상위어 '장'과 같이 상위어의 형태가 질의어의 일부분으로 나타나거나 '경기11'의 상위어 '일'과 같이 상위어가 한 음절의 어휘인 경우에는 현재 포탈사이트의 키워드 매칭 기법으로 인해 잘못된 결과가 도출되는 경우가 많다. 그래서 이와 같은 경우에는 질의어와 조부모 관계에 있는 어휘를 2차 키워드로 활용하였다.



▶▶ 그림 3. '기술-수단' 확장 질의어를 통한 검색

[표 3] 확장 질의어의 웹 문서 검색 결과

원 질의어	확장 질의어	정확한 검색 결과(%)	
		Naver(80%)	Google(63%)
기술	기술-수단	100	100
	기술-서술	74	42
동향	동향-방향	30	0
	동향-목표	100	100
시장	시장-장관	92	50
	시장-장소	86	88

원 질의어	확장 질의어	정확한 검색 결과(%)	
		Naver(83%)	Google(65.5%)
경기	경기-지방	100	82
	경기-상태	94	53
	경기-활동	60	16
다리	다리-부위	100	98
	다리-시설물	72	94
배	배-부위	98	86
	배-구조물	54	22
	배-열매	85	94

[표 4] 확장 질의어의 통합 검색 결과

원 질의어	확장 질의어	정확한 검색 결과(%)	
		Naver(73.1%)	Daum(69%)
기술	기술-수단	96.8	96.6
	기술-서술	64.5	51.9
동향	동향-방향	20.6	0
	동향-목표	100	100
시장	시장-장관	65.6	65.4
	시장-장소	90.9	100

원 질의어	확장 질의어	정확한 검색 결과(%)	
		Naver(75%)	Daum(57.5%)
경기	경기-지방	87.2	91.7
	경기-상태	42.4	16.7
	경기-활동	45.9	3
다리	다리-부위	100	100
	다리-시설물	94.6	84.6
배	배-부위	80.5	66
	배-구조물	67.6	33.3
	배-열매	81.4	63

U-WIN의 상위어를 이용한 웹문서 검색 결과는 <표 3>과 같이, 질의어가 전문분야에서 사용되는 동형이의어인 경우에는 검색 정확률이 평균 80%(Naver), 63%(Google)이며, 질의어가 보편적으로 사용되는 동형이의어인 경우에는 평균 83%(Naver), 68%(Google)이다. 즉 '질의어+상위어' 형태의 확장 질의어에 대해 두 개의 포탈사이트(Google, Naver)를 대상으로 웹 문서를 검색한 평균 검색 정확률이 81.5%(Naver), 65.5%(Google)로 나타났다. 또한 통합검색 결과는 <표 4>과 같이, 평균 정확률이 74%(Naver), 63.3%(Daum)로 나타났다.

본 논문의 검색 정확률은 각 포탈사이트의 인덱싱 및 키워드 매칭 기법 등의 차이 때문인 것으로 판단되므로, 두 포탈사이트의 성능과는 직접적인 연관성을 가지고 있지 않으며, 본 논문에서는 상위 검색 결과에 출현하지 않았던 문서들이 "질의어+상위어"의 확장 질의어를 통한 검색 결과의 정확성, 효율성을 증명하고자 하였다.

Ⅳ. 결론 및 향후 연구 과제

본 논문에서는 질의어가 동형이의어인 경우 상위어를 추가하여 검색한 결과, 웹문서 검색은 평균 81.5%(Naver), 65.5%(Google), 통합 검색은 평균 74%(Naver), 63.3%(Daum)로 특정 질의어에 집중적으로 분포된 문서들 속에서 원하는 결과를 찾을 수 있었다. 이것은 U-WIN이 분류적 상하관계가 아닌 계층적 상하관계를 기반으로 하여 구축됨으로써, 상위어가 질의어의 모호성을 해결하는 유용한 정보로 사용됨을 알 수 있다. 향후 U-WIN의 기본의미관계, 기본개념관계, 확장개념관계 등의 여러 관계정보를 추가적으로 활용하여 사용자가 원하는 정보를 좀 더 정확하게 제공할 수 있는 정보검색 기술을 구현할 수 있을 것이며, 이를 바탕으로 의미적인 정보검색의 기반을 마련할 수 있을 것이다.

▌참 고 문 헌▐

[1] 노영희, "개념기반 검색을 위한 시소러스 관계의 효과적 활용 방안에 관한 연구", 정보관리학회지, 제17권, 제4호, pp.47~65, 2000.

[2] 박창근, 양기철, "의미정보기반 검색시스템의 설계 및 구현", 한국콘텐츠학회 종합학술대회 논문집 한국콘텐츠학회 2004 추계 종합학술대회 논문집 제2권, 제2호, pp.265~268, 2004.

[3] Snasel, V., Moravec, P, Pokorny, J., "WordNet Ontology Based Model for Web Retrieval", International Workshop on Challenges in Web Information Retrieval and Integration(WIRI) 2005, pp.220~225, 2005.

[4] 최호섭, 임지희, 배영준, 최수일, 옥철영, "온톨로지 구축 방법과 사례", 정보과학회지, 제24권, 제4호, pp.31~44, 2006.

[5] 최호섭, 임지희, 옥철영, "대규모 지능형 한국어 어휘망 구축-우리말 어휘지능망(U-WIN)을 중심으로-", 609돌 세종날 기념 한글 학회 전국 국어학 학술 대회 발표자료집, 2006.

[6] 김준수, 의미정보와 시소러스를 이용한 한국어 어휘 중의성 해소 모델, 울산대 박사학위논문, 2004.