워드넷을 이용한 검색 질의어의 모호성 해결

김형일[U] 김준태
동 국 대 학 교 컴 퓨 터 공 학 과
{salmoner, jkim}@dgu.ac.kr

Resolving Ambiquity in search query by using the WordNet

Hyoung-Il Kim[U] Jun-Tae Kim
Dept. of Computer Engineering, Dongguk University

요 약

방대한 웹에서의 자신이 원하는 정보를 정확히 얻어내기란 매우 어렵다. 현존하는 대부분의 검색엔진들은 내용기반 방식을 이용하므로, 검색 질의어의 모호성에 적절한 대응을 하지 못 하고 있다. 다시 말하면 일반 사용자들이 사용하는 질의어들은 다의어로 표현되는 것이 빈번히 나타나지만, 사용자가 나타내고 싶어하는 질의어의 정확한 의미에 대하여서는 검색엔진 자체로써는 해결할 수 없다. 특히, 빈번히 사용되지 안는 어휘의 의미를 가지고 검색엔진에 질의를 할 경우, 질의어의 형태만 같고 일반적으로 널리 사용되고 있는 어휘의 의미와 관련 있는 웹 페이지들만을 사용자에게 보여주게 된다. 이러한 점을 보완하기 위하여 본 논문에서는 사용자의 명시적 반응을 받아들이는 사용자 인터페이스와 워드넷(WordNet)을 이용하여 질의어의 모호성 해결하였다.

1. 서론

수 없이 증가되고 있는 웹 사이트에서 원하는 페이지를 검색하기란 쉬운 일이 아니다. 이러한 측면으로 하여 일반 사용자들을 위하여 많은 검색 엔진들이 개발되어 가는 추세이다[1]. 그러나 대다수의 검색 엔진들은 내용기반[1] 방법을 이용함으로써, 사용자가 만족할 만한 다수의 웹 페이지를 찾아 주지는 못하는 실정이다. 일반 사용자가 자신이 원하는 정보를 검색하기 위해 특정 질의어를 검색 엔진에 던질 경우, 대다수의 검색 엔진들은 사용자가 어떤 의미로 해당 질의어를 사용하였던지 간에 해당 질의어 형태가 포함되어 있는 웹 페이지를 보여 준다. 이러한 원인의 발생은 일반 사용자들은 어휘의 의미를 생각하고 질의어를 던지지만, 검색엔진에서는 어휘의 형태만 인식하기 때문이다.

인간이 사용하고 있는 언어는 의미의 정형화를 위하여 어휘의 형태라는 매개체를 이용하고 있다[3]. 이러한 정형화된 어휘구조를 사용하는 언어는 어느 언어이든 다의성을 갖는 어휘를 포함하게 된다[2]. 이러한 측면에서 보면, 단어가 정보 검색에서 질의어로 사용되어 질 경우 어휘의 다의성 측면에서와 같은 단점을 내포하게 된다. 예를 들어 임의의 사용자가 커피의 종류인 자바(Java)에 대하여 알고 싶을 경우, 검색엔진에 자바(Java)라고 질의어를 던지게 된다. 그러나 대다수의 검색엔진은 자바(Java)언어에 대한 웹 페이지만 사용자에게 보여주고 있다. 이러한 원인은 대다수의 검색 엔진들은 내용기반 방법을 사용하고 있으며, 현재 웹에서 인기를 끌고 있는 주제어일 경우는 웹 페이지가 당연히 많이 존재하기 때문이다. 이러한 원인으로 일반 사용자들은 자신이 원하는 문서를 찾기 위해 여러 가지 방법으로 질의어를 조합하던지, 관련 웹 페이지들의 링크를 이용하여 원하는 웹 페이지를 찾아야 한다.

이러한 검색 엔진들의 단점을 보완하기 위하여, 본 논문에서는 워드넷(WordNet)[7]을 이용하였으며, 사용자에게 어휘의 의미를 명시적으로 얻어내기 위하여 사용자 인터페이스를 설계하였다. 사용자가 인터페이스를 이용하여, 질의어의 정확한 의미를 선택하게 됨으로써, 질의어의 정확한 의미가 담긴 새로운 질의어를 재조합할 수 있다. 이러한 과정으로 만들어진 새로운 질의어를 통해 정확한 웹 페이지 검색을 할 수 있게 되었다.

2. 워드넷(WordNet)

우리가 사용하고 있는 언어는 다의어를 포함하고 있음으로 하여, 일반적으로 특정 단어나 단문의 표현 정도 가지고는 단어의 모호성을 해결할 수 없는 경우가 많다[1][5]. 간단한 예를 든다면 "나는 **배**를 보았다."라고 말할 경우 배가 과일을 의미하는지 선박을 의미하는지 청자는 알 수 없을 것이다. 그리하여 본 논문에서는 단어의 모호성을 해결하는 방안으로 워드넷(WordNet)을 이용하여 해결을 시도하였다. 또한 워드넷(WordNet)은 자연언어처리[3]에서 많은 이용이 되고 있으며, 문서 분류[4]를 위하여 워드넷(WordNet)의 어휘 관계가 이용되기도 하고 있다. 현재는 어휘의 의미와 계층을 이용한 어휘 사전의 연구가 활성화되어 가고 있는 추세이며, 워드넷(WordNet)은 현재 다국어판 구현이 되어가고 있으며, 우리 나라 또한 많은 연구가 진행되어 가고 있는 추세이다.

워드넷(WordNet)은 영어 어휘에 대한 다의성과 동의어 관계, 반의어 관계 및 어휘의 포함 관계 등을 정의해 놓은 것으로써, 동의어, 반의어, 상위어, 하위어 등에 대한 정보가 잘 표현된 사전이라 말할 수 있다[2][6]. 이러한 의미의 사용은 어휘들의 연관관계를 나타낼 수 있다. 워드넷(WordNet)에서의 단어와 의미의 관계 모델을 [표 1]과 같이 나타내기도 한다. [표 1]에서 F1과 F2는 동의(Synonym)어 관계이고 F2는 다의(Polysemy)어가 된다. M1은 단어의 의미를 나타내는 것이고 {F1,F2}는 M1의 의미에 있어서 동의어 집합(Sets of Synonyms)이 된다.

Word / Meanings	F1	F2	F3	F4	....	Fn
M1	E1,1	E1,2				
M2		E2,2				
M3			E3,3			
:					....	
Mn						Em,n

[표 1] 어휘의 개념 행렬

◈동의(Synonymy)관계와 반의(Antonymy)관계

동의 관계의 정의를 표현하여 보면, x와 y의 의미가 같다고 하면, y와 x의 의미 또한 같다고 할 수 있는 등식이 성립하여야 한다. 이러한 동의(Synonymy)관계의 정의로 동의어들을 집합으로 표현할 수 있다. 반의(Antonymy)관계는 동의관계의 중요성만큼이나 중요한 역할을 담당하고 있다. 예로써 반의(Antonymy)관계를 설명하면, 집합 A={rise, ascend}와 집합 B={fall, descend}는 각각의 집합 A와 B는 동의어 집합(Synsets)이 되고 집합 A와 집합 B는 반의(Antonymy)관계를 유지하게 된다. 또한 집합 A와 집합 B가 반의(Antonymy)관계가 되면, 집합 A의 원소들과 집합 B의 원소들 또한 반의(Antonymy)관계가 된다.

◈하위(Hyponymy)관계와 상위(Hypernymy)관계

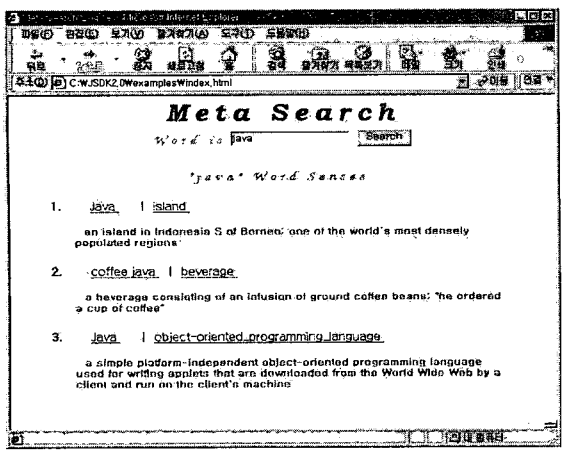
하위(Hyponymy)관계와 상위(Hypernymy)관계는 어휘에 대하여 의미 계층을(Semantic Category)을 나타낸 것이다. 예를 들어 설명하면, {maple}은 {tree}의 하위(Hyponymy)어가 되며, {tree}는 {maple}의 상위(Hypernymy)어가 된다.

◈부분(meronymy)관계와 전체(Holonymy)관계

부분(meronymy)관계와 전체(Holonymy)관계의 표현으로 단어들 사이에서의 포함 관계를 보여 줄 수 있다. 다음 예로써 부분(meronymy)관계를 설명하여 보겠다. 동의어 집합(Synset) A={ X1, X2,··,Xn}가 동의어 집합(Synset) B={ Y1, Y2,··,Yn}의 부분(meronymy)관계에 있다면 Y1은 X1을 포함하는 관계가 되고, X2는 Y2이 부분이 된다고 말할 수 있다.

3. 시스템

본 논문에서 사용자 질의어의 모호성을 해결하기 위하여 워드넷(WordNet)을 이용한 사용자 인터페이스를 설계하여 시스템을 구성하였다. [그림 1]은 사용자 인터페이스 화면으로 사용자가 "Java"라고 질의어를 던질 경우 "Java"에 대한 모든 의미를 인터페이스에서는 사용자에게 나타내 주게 된다.



[그림 1] 사용자 인터페이스

3.1 동의어, 상위어, 주석의 추출 및 질의어의 재조합

워드넷(WordNet)에서는 임의의 어휘에 대하여 동의어 집합이 나타나 있으며, 명사 사전에서는 모든 어휘들이 25개의 카테고리(Category)로 분류되어 있다[3]. 그리고 어휘들의 의미 별로 계층 분류가 되어 있음으로 하여, 임의의 어휘가 주어지게 되면 상위개념의 어휘와 하위개념의 어휘를 찾아 낼 수 있다.

워드넷(WordNet)을 구성하는 중요한 파일에는 인덱스 파일과 데이터 파일이 있다. 인덱스 파일에는 모든 어휘들과 해당 어휘의 모든 의미가 나타나 있는 데이터 파일의 위치 정보가 나타나 있다. 데이터 파일에서 위치 값은 옵셋(Offset)[7]이 사용되었다. 이 위치 값(Offset)으로 데이터 파일을 검색하여, 해당 어휘의 의미 및 동의어, 상위어, 주석 등의 정보를 추출해 낼 수 있다. 워드넷(WordNet)에서는 사용되는 어휘에 대한 설명은 주석에 나타나 있으며, 이러한 주석은 해당 어휘의 정의라 말할 수 있을 것이다. 그리고 동의어는 워드넷(WordNet)에서 특정 어휘와 같은 계층에서 같은 의미를 나타내는 어휘들의 집합이라 말할 수 있다. 그리고 상위어는 특정 어휘의 상위 계층에 있는 어휘로써, 특정 어휘의 선택 시 상위어로 특정 어휘의 의미에 대하여 대표성을 부여할 수 있다. 이러한 동의어 집합, 상위어 집합, 주석의 활용은 질의어 재조합에 있어서 중요한 역할을 할 수 있을 것이다. 그리하여 질의어 조합 과정에서 워드넷(WordNet)에서 추출할 수 있는 동의어 집합, 상위어 집합, 주석을 이용하여 본 실험에 적용하였다.

주석을 질의어와 재조합하기 위하여 질의어 확장기는 워드넷(WordNet)의 인덱스 파일을 먼저 검색하여 해당 질의어를 찾아낸 후, 해당 어휘의 옵셋(Offset)을 추출하여 낸다. 여기에서 옵셋(Offset)은 8비트로 표현[3]되어 있으며, 이러한 어휘의 옵셋(Offset)을 가지고 데이터 파일을 검색하여, 주석을 추출하게 된다. 데이터 파일에서는 주석을 표시하기 위하여 주석 앞에 " | " 기호를 사용하였다. 이러한 방식으로 추출된 주석은 해당 질의어와 재조합하기 위해 명사들만 추출하게 된다. 추출된 주석의 명사들은 해당 질의어와 재조합하여 새로운 질의어를 창출해 낸다. 다음으로는 동의어 집합을 만들기 위해 질의어 확장기는 주어진 질의어의 위치 값(Offset) 가지고 데이터 파일을 검색하여 동의어를 추출해 낸다. 데이터 파일에서는 동의어들을 나타내기 위하여 공백을 이용하였으며, 동의어들은 문자(Character)열로 나열되어 있기 때문에 한 번의 접근으로 추출할 수 있다. 이러한 과정으로 만들어진 동의어 집합은 주어진 질의어와 함께 새로운 질의어를 만들어 내게 된다. 그리고 상위어 집합을 만들기 위해서 질의어 확장기는 데이터 파일을 검색하게 되는데, 상위어의 표현을 위해 워드넷(WordNet)에서는 "@"를 사용하였다. 상위어 집합을 만들기 위해 질의어 확장기는 "@" 찾아서 상위어의 옵셋(Offset)들을 알아내고 상위어들을 추출해 낸다. 이러한 과정으로 만들어진 상위어 집합과 해당 질의어를 재조합하여 새로운 질의어를 만들어 내었다. 이와 같은 세 가지의 질의어 재조합 방법을 이용하여 본 실험에 임하였다.

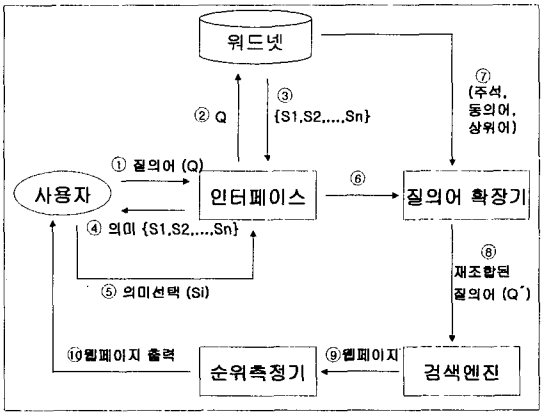
3.2 사용자 질의 시 시스템의 흐름

[그림 2]는 전체 시스템의 구성을 나타내고 있다.

사용자 인터페이스를 통하여 일반 사용자가 질의어(Q)를 던질 경우, 해당 질의어(Q)의 정확한 의미를 추출하기 위하여 사용자 인터페이스는 워드넷(WordNet)[7]을 이용하여 질의어에(Q) 해당하는 모든 의미{S1, S2, ... Sn}를 사용자 인터페이스에 전달하여 준다. 전달된 의미{S1, S2, ... Sn}는 사용자에게 보여주게 되며, 사용자가 선택한 의미(Si)는 인터페이스를 통하여 질의어 확장기에 전달된다.

질의어 확장기에서는 사용자가 선택한 의미(Si)와 워드넷

(WordNet)에서 추출된 동의어 집합, 상위어 집합, 어휘의 주석들과 재 조합하여 새로운 질의어(Q')를 만든 후, 일반 검색엔진으로 다시 질의를 하게 된다. 새로운 질의어(Q')로 검색된 웹 페이지들은 순위 결정기를 통하여 순위가 결정된 것으로 사용자에게 보여주게 된다.



[그림 2] 메타 검색엔진 시스템 구성도

4. 실험 및 결과

세 가지의 질의어 재조합 방법을 실험하기 위하여 임의의 질의어를 선택하여 실험에 임하였다. 이 실험에 이용된 9개의 의미가 다른 질의어는 구글(Google)과 알타비스타(Altavista)를 이용하여 웹 페이지를 검색하여 실험하였다. 각각의 검색엔진에서는 30개씩 페이지를 가져오게 하였다. 워드넷(WordNet)을 이용한 메타 검색엔진은 가지고온 60개의 웹 페이지를 가중치를 이용하여 30개를 선별하여 사용자에게 보여 주게 하였다. 본 시스템에서 사용한 가중치 값은 구글(Google)과 알타비스타(Altavista)에서 사용한 가중치를 이용하였다. 질의어에 대한 웹 페이지의 관련성을 측정하는 실험임으로 정량적 실험을 할 수 없었다. 또한 특정 질의어에 대하여 웹 페이지의 관련성을 나타낼 만한 기준 자료가 없는 관계로 기초적인 실험으로 임할 수 밖에 없었다. 그리하여 본 실험에서는 9명의 사용자에게 하나의 질의어를 선택하게 한 후 실험에 임하였다.

사용된 질의어		일반 검색엔진		메타 검색엔진		
질의어 형태	질의어의 의 미	Alta-vista	Google	질의어와 조합된 어휘 집합		
				동의어	상위어	주 석
J a v a	**자바커피**	**0**	**0**	**6**	**3**	**7**
J a v a	**자바 섬**	**0**	**0**	**0**	**5**	**5**
J a v a	자바언어	27	28	28	15	27
Character	**배 역**	**1**	**2**	**5**	**3**	**7**
Character	문 자	9	12	12	10	12
C u s t o m	**통 관**	**0**	**0**	**5**	**10**	**9**
C u s t o m	관 습	4	6	10	5	10
H o r s e	**마 약**	**0**	**0**	**5**	**2**	**5**
H o r s e	말	27	26	25	27	27
평균 관련문서		68	74	96	80	109
평균 정 확 도		25.2%	27.4%	35.6%	29.6%	40.4%

[표 2] 9가지의 질의어를 이용한 실험 결과

[표 2]는 9 명의 사용자가 구글(Google)과 알타비스타(Altavista), 워드넷(WordNet)을 이용한 메타 검색엔진에 질의어의 의미를 부여한 채, 질의를 할 경우에 대한 결과 표이다. 워드넷(WordNet)을 이용한 메타 검색엔진에서는 동의어, 상위어, 주석을 이용한 질의어 재조합 방법 세가지를 가지고 실험하였다. 그러한 결과 현재 많이 사용되어지고 있는 질의어들에서는 일반 검색엔진과 워드넷(WordNet)을 이용한 메타 검색엔진에서 큰 차이가 나타나지 않고 있지만, [표 2]에서 진한 색으로 표시한 어휘는 자주 사용되어지지 않는 어휘의 의미이다. 이러한 어휘의 의미로 검색을 할 경우 메타 검색엔진에서는 뛰어난 정확도를 나타낸다는 것을 알 수 있다. 또한 세 가지의 질의어 재조합 방법을 비교하여 보면, 주석을 이용하여 질의어를 재조합할 경우가 동의어나 상위어와 조합하는 경우보다 월등한 정확성을 보이는 것을 알 수 있다. 실험 결과에서 나타난 것을 보면 일반 검색엔진에서는 사용자가 일반적으로 널리 사용되지 않는 의미를 가지고 질의를 던질 경우, 적절한 대응을 할 수 없었다. 그러나 워드넷(WordNet)을 이용한 메타 검색엔진은 사용자가 질의어를 던질 경우 사용자 인터페이스를 이용하여, 해당 질의어의 모든 의미를 사용자에게 나타내 줌으로써, 사용자가 질의어의 정확한 의미를 선택 할 수 있었다. 이러한 질의어의 의미에 대한 사용자의 선택은, 해당 질의어의 정확한 의미를 파악할 수 있게 하며, 질의어 재조합 시 정확한 질의어의 의미로 새로운 질의어를 만들 수 있게 하였다.

5. 결론 및 향후 연구과제

본 논문에서는 사용자가 질의어를 던질 경우 질의어의 모호성 해결과 일반 검색 엔진들이 내용기반 검색을 기반으로 사용함으로써 원하는 웹 페이지를 얻기란 쉽지 않다는 것에 착안하여 시스템을 설계하고 실험에 임하였다. 그리하여 사용자 질의어의 모호성 해결을 위해 워드넷(WordNet)을 이용하여 사용자 인터페이스를 설계하였다. 사용자에게 인터페이스를 이용하게 함으로써 주어진 질의어에 대한 정확한 의미를 선택하게 하였다. 이러한 사용자 인터페이스의 설계로 질의어의 의미를 명시적으로 나타낼 수 있었다. 그리고 주어진 질의어와 동의어, 상위어, 주석을 이용하여 새로운 질의어를 만든 후, 검색 엔진에 던져 실험한 결과를 보면 주어진 질의어와 주석을 이용하여 질의어를 재조합한 경우가 가장 높은 정확도를 나타내었다.

본 실험에서는 정량적 실험을 할 수 없는 관계로 임의 사용자를 선택하여 실험한 결과에 의존하여야 했지만, 사용자 인터페이스의 설계로 질의어의 모호성 해결에 뛰어난 성능을 나타내었다. 향후 과제로써 문장 형태의 질의어가 들어 왔을 경우, 사용자 인터페이스에서 질의어의 의미를 사용자 반응 없이 자동으로 추출할 수 있다면, 더욱 향상된 검색엔진의 성능을 갖출 수 있을 것이다. 한국어 또한 의미와 계층을 이용한 사전이 절실히 요구되어지며, 한국어 의미 사전이 만들어진다면, 본 시스템은 한국어에서도 높은 정확도를 유지할 수 있을 것이다.

6. 참고 문헌

[1]W. Frakes, and R.Baeza-Yates, "Information Retrieval : Data Structures & Algorithm", Prentice-Hall,1992

[2]Miller, "WordNet : An On-Line Lexical Database", International Journal of Lexicography, 1990

[3]C. Fellbaum, "WordNet : An Electronic Lexical Database", MIT Press,1998

[4]S. Scott, and S. Matwin, "Text Classification Using WordNet Hypernyms", Coling-ACL '98Workshop, 1998

[5]Xiaobin Li, Stan Szpakowicz and Stan Matwin, "A WordNet-based Algorithm for Word Sense Disambiguation", IJCAI-95, 1995

[6]Eric Siegel, "Disambiguating Verbs with the WordNet Category of the Direct Object", Coling-ACL '98 workshop,1998

[7]WordNet, http://www.cogsci.princeton.edu/~wn/