동사사전를 이용한 의미 기반 정보 검색 시스템의 설계

이용훈*, 이상범*
*단국대학교 전자계산학과
e-mail:karismanj@dankook.ac.kr

Design of An Information Retrieval System using Verb Dictionary

Yong-Hun Lee*, Sang-Bum Lee*
*Dept of Computer Science, Dankook University

요 약

본 논문에서는 문장에서 동사를 파악하여 명사간의 의미를 부여하는 자동학습 온톨로지 기반 정보 검색 시스템을 제안한다. 정보의 양이 무한히 증가하고 있으며 웹의 발전에 따라 적합한 정보를 찾아내야 하는 효율적인 정보 검색 시스템의 필요성이 증대되고 있다. 단순히 키워드의 가중치에 따른 검색의 순위화는 사용자의 의미를 이해하지 못한 검색 결과로서 사용자로 하여금 결과를 다시 한번 직접 검색해야 하는 불편함을 제공하고 있다. 이러한 문제점을 해결하기 위해서 문장에서 동사를 파악하여 명사 간의 의미를 부여하고 문서 내에서 단어 간의 의미를 파악하여 검색의 질을 개선하는 방법을 논의한다. 또한, 문서에서 단어의 관계를 스스로 학습 가능하여 구축되는 자동학습 온톨로지 기반의 정보 검색 시스템을 제안한다.

1. 서론

정보검색이란 데이터 웨어하우스 또는 데이터 집합소에서 사용자의 질의로부터 원하는 정보를 찾아오는 것을 말한다. 오늘날의 데이터 집합소는 거대한 정보의 바다라고 일컫는 인터넷이 그 역할을 하고 있다. 인터넷은 일반 데이터 집합소와 다르게 구조화 되어 있지 않은 정보를 가지고 있다. 즉, 모든 정보들이 정해진 규칙없이 산재된 상태로 존재하고 있는 것이다. 인터넷은 웹2.0 시대를 맞이하면서 정보를 제공하는 업체들만이 아닌 많은 일반 사용자의 정보가 공유되는 세상이 되었다. 일반 사용자와 정보를 제공하는 업체들 간의 정해진 구조 없이 정보를 마구 생산해 내고 있는 것이다. 이러한 정보의 증가에 따른 정보검색의 필요성은 나날이 증가하고 있다. 인터넷으로 부터 정보를 검색 가능하게 하는 사이트들이 등장하였으며 이러한 회사들은 점점 규모가 커져서 Google이나 네이버와 같은 IT 업계의 선두 주자가 되었다. 정보를 가진 자가 세상을 지배하는 시대가 도래한 것이다. 정보의 양이 무한대에 가까워진 요즘 같은 시대에 검색이 되지 않은 정보는 더 이상 정보로써의 가치를 잃어 버리게 되는 것이다. 하지만 지금의 정보 검색의 모습은 웹에서 존재하는 정보의 단위인 문서에서 단순한 키워드 매칭을 이용하여 정보를 찾아내는 것 이상의 역할을 할 수 없는 것이 사실이다. 사용자의 질의어로부터 해당하는 단어를 포함하는 문서를 순위화 하여 가져오는 것이다. 사용자의 의도나 의미를 파악하지 못한 단순한 키워드 매칭은 사용자로 하여금 검색된 결과에서 직접 사용자의 노력으로 검색과정을 거쳐 자신의 원하는 정보를 찾아 내야 하는 것이다. 이러한 문제점은 정보의 양이 점점 증가하면서 질적인 정보가 아닌 양적인 정보의 증가로 인하여 더욱 필요성을 느끼게 된다. 사용자의 의미를 파악하여 검색을 한다는 것은 쉬운 것이 아니다. 사용자의 검색어에서 어떤 의미를 찾아 낸다는 것은 단순한 몇 가지의 단어를 가지고 검색해야 하는 현 정보검색 시스템의 기능으로는 힘든 것이 사실이다.

본 논문에서는 이러한 문제점을 개선하기 위한 방법으로 문장으로부터 명사 간의 의미를 찾아 내는 방법을 제안한다. 기본 문장인 주어, 목적어, 동사 순의 문장에서 동사를 파악하여 두 명사(주어, 목적어) 간의 관계를 저장하여 색인어 간의 관계를 동사로 표현하는 정보검색 시스템을 제안한다.

본 과제는 한국소프트웨어진흥원의 SW공학 요소기술 개발과 전문인력 양성사업의 결과물임을 밝힙니다.

2. 관련 연구

2.1 형태소 분석 및 색인어 추출

형태소 분석은 문장을 구성하는 단어들의 품사를 파악하여 문장의 구조를 분석해내는 방법이다. 한국어 형태소 분석은 많은 문법적 변화를 가지고 있으며 단어가 축약, 탈락, 불규칙 활용 등으로 인하여 여러 가지 형태로 변경되어 문장에서 표현된다[1]. 이러한 변화가 많은 한국어 문장에서 색인어를 추출하는 통상적인 과정은 불용어 제거, 접미사 절단, 동일 어근 검출과정을 통하여 색인어를 추출하게 된다[2]. 문서에서 중요한 단어 추출에 사용하는 품사로써 명사를 추출하여 검색시스템의 색인어로 사용하고 있다. 명사는 문장의 중심적인 역할을 하며 문장 및 문서를 표현하는데 중요한 의미를 지니고 있다. 색인어로 추출된 단어들을 이용하여 가중치를 계산하는 단어 빈도수(TF, Term Frequecey) , 역문헌의 수(IDF, Inverted Document Frequecy)등의 계산을 통하여 검색어에 순위화된 결과 정보를 사용자에게 제공한다[3]. 산재된 정보를 가지고 있는 웹에서는 정확한 형태소 분석이란 쉬운 방법이 아니다. 웹에서 제공되는 정보의 단위인 문서는 정확한 문법을 지키며 쓰여지지 않는 경우가 많기 때문에 이를 분석하는 과정에서 여러 가지 잘못된 품사가 결정되며 이는 가중치 계산에 오차를 발생하게 된다.

문장에서 가장 중요한 역할을 하는 것은 명사인 것이 사실이나 명사가 문장의 전체의 의미를 대변한다고는 할 수 없다. 주어와 목적어의 관계가 동사에 의해 결정되어 지므로 동사 파악은 문장의 의미를 파악하는데 중요한 역할을 한다. 과거 검색 시스템이 단순한 키워드 검색을 통한 방법을 사용할 때는 명사를 색인어로 추출하는 방법이 가장 적합하나 의미까지 검색 가능한 시스템에서는 명사만을 색인어로 추출하는 것은 문장과 문서의 의미를 파악하는데 한계가 있다.

2.2 연관 관계 추출 : 시소러스와 온톨로지

시소러스는 특정 주제 영역에서 사용되는 용어와 이들 용어간의 의미 관계를 체계적으로 제시한 색인어휘집으로서 색인과 검색과정에서 디스크립터와 검색어를 선정하기 위한 도구로 사용된다[4].

시소러스 용어의 의미 체계에서 가장 중요한 관계는 광의어(Broad Term : BT), 협의어(Narrow Term: NT), 관련어(Related Term : RT)로 표현되는 개념 간의 관계이다. BT와 NT로 표현되는 계층관계는 집합과 요소와의 관계를 나타내는 속-종(generic) 관계, 신체 조직이나 지리상의 위치 또는 학문 분야 등을 나타내는 부분-전체(part-whole) 관계, 카테고리와 그에 포함된 예시를 나타내는 사례(instance) 관계로 엄격하게 제한되는 것이 일반적이다. 그러나 RT 로 표현되는 연관관계는 물체와 특성, 인과관계, 조작과 행위자, 개념과 측정 단위 등 계층적이지도 않고 등가 관계도 아니면서 상당한 관련성이 있는 관계를 모두 포함한다. 따라서 연관관계는 관계의 규정범위가 지나치게 단순하고 포괄적이다.

온톨로지는 개념들의 집합들을 나타내는 것으로 개념 간에 존재하는 관계를 명백하게 기술하여 개념과 그러한 관계성립의 공유를 목적으로 사용된다[5]. 시소러스가 용어관계에 대한 정의라면 온톨로지는 용어간의 의미관계라고 정의할 수 있다. 또한, 사람의 마음속에 존재하는 내재적 생각이나 외재적 세계의 현상과 대상에 대하여 공유하는 개념을 컴퓨터가 이해할 수 있는 형식으로 명확하고 명시적으로 정의하고 규정하는 것을 말한다[4]. 온톨로지가 시소러스와 구별되는 또 다른 특성은 일반화 혹은 상호운영성의 규칙을 적용함으로써 구조화된 지식으로부터 새로운 지식을 추론할 수 있다는 점이다. 이러한 점을 이용하여 시맨틱 웹을 실현하기 위하여 온톨로지에 대한 연구가 활발히 진행되고 있는 것이다.

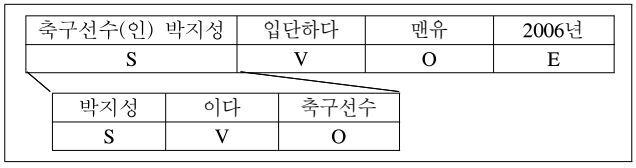
3. 본론

3.1 형태소 분석을 통한 문장 구조 파악

단어 간의 관계를 동사를 통하여 정의하는 본 논문의 큰 주제를 가능하게 하기 위해서는 정확한 형태소 분석을 통한 명사 추출과 동사 추출이다. [표 1]은 예제에 해당하는 문장을 분석한 결과이다.

예제) “2006년, 축구선수 박지성은 맨유에 입단했다.”

[표 1] 문장을 분석한 결과



[표 1]은 형태소 분석을 통하여 “박지성은 축구선수이다”, “박지성은 맨유에 입단하다” 라는 두가지 문장을 얻을 수 있게 된다.

형태소 분석 과정 중 동사로 파악된 “입단했다”라는 동사는 동사의 원형인 “입단하다”라는 형태의 원형으로 복원하여 저장된다. 또한 “축구선수 박지성”이라는 파싱된 토큰에서 “박지성은 축구선수 이다”라는 관계를 파악하기 위해서는 명사와 명사가 연속적으로 나오는 관계에서 “이다” 동사 관계로 파악하면 가장 적절할 것이다. 이와 같이 여러 상황 속의 문자에서 명사 간의 관계를 추출해야 하는 것이다.

[표 2]는 또 다른 문장의 형태소 분석의 결과 이다.

예) “박지성은 2002년 월드컵에서 엄청난 활약으로 해외 진출에 성공했다.”

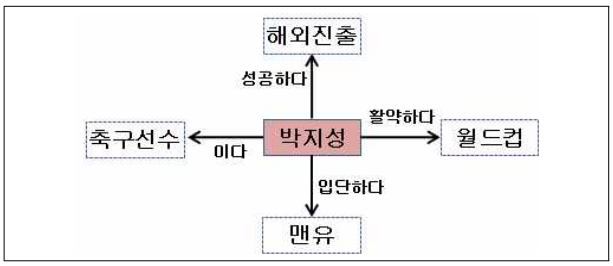
[표 2] 문장을 분석한 결과

박지성	활약하다	월드컵	2002년 월드컵에서
S	V	O	E

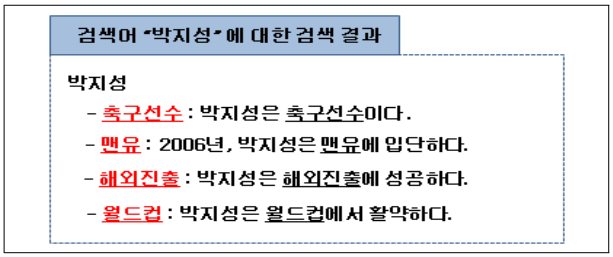
박지성	성공하다	해외진출
S	V	O

[표 2]의 결과도 문장에서 주어와 목적어를 동사관계에 따라 분석한 것이다.

이렇게 분석된 문장을 통하여 사용자가 “박지성”이라는 키워드를 질의로 입력하게 되면 [그림 1]과 같은 연관 관계도가 그려 지며, [그림 2]와 같은 검색 결과를 얻을 수 있을 것이다.



[그림 1] 키워드에 해당하는 연관관계도

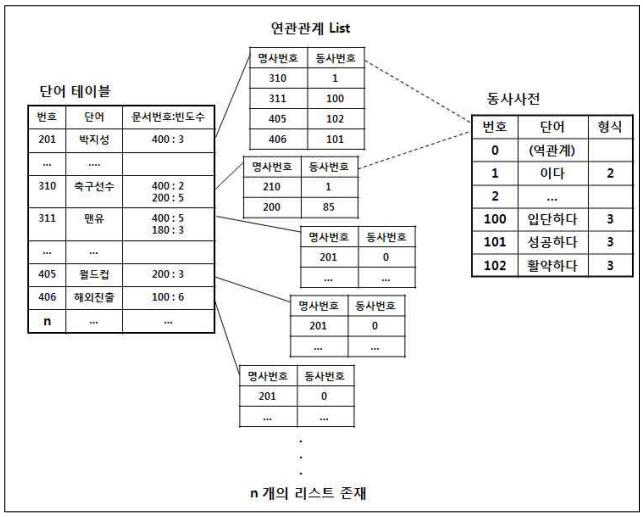


[그림 2] 질의에 대한 검색 결과

문장의 동사를 파악하게 되면 명사 간의 의미 관계를 파악할 수 있게 되며 이를 통하여 사용자는 검색 결과에서 자신이 원하는 질의를 확장하면서 검색을 계속 할 수 있게 된다. [그림 2]에서 “박지성”이라는 질의로 “축구선수”, “맨유”, “해외진출”, “월드컵” 이라는 단순한 연관어가 아닌 의미 관계의 단어를 추가적으로 알 수 있게 된다. 따라서 사용자는 추가적인 질의 확장을 할 수 있게 되며 확장된 단어들을 이용하여 다시 검색을 하면 자신이 찾고자하는 정보를 보다 의미적으로 빠르게 접근할 수 있을 것이다.

3.2 역화일 구조

본 논문에서 제시하는 동사를 파악하여 명사간의 관계를 추출하는 과정에서 명사와 동사간의 관계는 시스템 내부적으로 저장되어 있어야 한다. 따라서 [그림 3]과 같은 역화일 구조를 구성하는 것이 효율적이다.



[그림 3] 역화일 구조

[그림 3]의 역화일 구조는 단어테이블, 연관관계 리스트테이블 그리고 동사사전 테이블로 3가지 나눌 수 있다.

첫 번째로 단어테이블은 단순한 해쉬 구조를 가진 단어와 단어의 정보쌍을 가진 구조이다. 저장된 단어는 시스템에서 색인어로 선정된 것이며, 색인어들은 각각의 고유한 번호를 가지게 된다. 각 색인어의 정보에 해당하는 문서 번호와 빈도수를 색인어와 쌍으로 저장한 구조이다.

두 번째로 연관관계 리스트 테이블은 각 색인어마다 하나씩 테이블이 생성되며, 색인어 수 N개 만

큼 생성되는 것이다. 이 테이블은 색인어와 연관이 있는 명사 번호와 동사 번호를 가지고 있다. 명사 번호는 단어 테이블의 색인어 번호 이며, 동사 번호는 동사사전 테이블의 번호가 되는 것이다. 하나의 색인어는 연관관계리스트 테이블에서 관계있는 다른 단어를 알 수 있으며 그 관계를 동사사전 테이블와 찾아와 의미 관계를 짓게 되는 것이다.

세 번째로 동사 사전 테이블은 문장 분석 과정 중 동사로 판단된 단어가 들어가는 테이블이다. 연관관계리스트 테이블에서 동사사전 테이블에 해당하는 번호의 동사 정보를 가지고 있어서 색인어 간의 동사 관계를 파악할 수 있게 된다. 형식이라는 필드는 동사가 어떤 문형의 형식으로 쓰이는지를 표시한 필드이며 동사 번호가 0 인 것은 색인어가 역으로 관계를 가진다는 것이다. 즉, “축구선수” 라는 색인어는 “박지성”이란 관계가 “박지성”이라는 색인어에서 “축구선수”를 찾을 수 있다는 반대되는 관계에서 정보를 찾을 수 있다는 표시이다.

3.3 검색과정 및 결과 도출 과정

사용자가 원하는 검색 질의를 색인구조에서 찾게 되며 본 논문에서 제안된 역화일 구조에서는 단어 테이블에서 단어를 찾게 된다. 단어 테이블에서 발견된 단어는 그 단어와 연결되어 있는 연관관계 리스트 테이블에서 명사번호와 동사번호를 찾아와 새로운 문장을 다시 만들어내야 한다. 동사사전에 형식이라는 필드에서 2형식과 3형식이 있는데. 2형식일 경우 “○○는 ○○이다.”형식으로 결과를 만들어 낼 수 있으며 3형식일 경우 “○○는 ○○을 (동사)하다”라는 형식으로 문장화 하여 사용자에 질의 해당하는 연관관계를 이룬 명사와 동사가 갖춰진 문장으로 제공 받게 된다.

4. 결론

본 논문에서는 문장의 구조 중 동사를 파악하여 명사 간의 의미 관계를 추론한 검색시스템을 제안하였다. 문장에서 동사와 명사 관계를 추출하여 색인어로 추출된 명사와 연관을 가지는 동사와 다른 명사를 연관관계화 하여 사용자의 질의로부터 확장된 문장형식의 결과를 받아 보게 된다. 단순히 명사만을 파악하여 문서를 이해하던 기존 검색 엔진 방식이 아닌 문장을 통한 문서의 의미를 찾는 과정이기에 의미론적인 검색이 가능해진 검색 시스템이라 할 수 있다. 하지만 이러한 엔진을 구성하는 것은 더 많은 외부 변수에 대한 계산이 필요하게 될 것이다. 같은 단어인데 다른 뜻을 가지는 단어들에 대한 링크를 설정하기 위해서는 여러 가지 상황을 고려하는 계산이 필요할 것이며, 중요도가 낮은 문장 또는 잘못 쓰여진 문장에서도 의미를 추출하여 연관관계를 맺게 되면 분명 의미적 오류를 발생할 수 있을 것이다. 이러한 부분은 차후에 많은 문장을 테스트 하면서 조율되어져야 할 문제일 것이다.

시멘틱 웹은 인터넷의 정보 간의 의미 관계를 가지게 된다. 시멘틱 웹을 바라보고 있는 이 시점에 검색엔진의 의미적 추론에 따른 검색은 현 정보 검색 시스템이 나가야할 방향이며 앞으로 활발한 연구가 진행 되어야 할 것이다.

참고문헌

[1] 심광섭, 양재형, “인접 조건 검사에 의한 초고속 한국어 형태소 분석”, 정보과학회논문지, 제31권, 제1호, pp. 89-99, 1월, 2004.

[2] 강승식, “한글 문서의 색인어와 색인 기법”, 정보과학회지, 제22권, 제4호, pp. 72-77, 4월, 2004.

[3] 송원문, 김영진, 김은주, 김영원, “단어 빈도수와 공기 정보를 이용한 효율적인 핵심어 추출 기법 개발”, 한국지능시스템학회 2008년도 추계학술대회 학술발표논문집, 제18권, 제2호, pp. 193-196, 10월, 2008.

[4] 고영만, “시소러스 기반 온톨로지에 관한 연구”, 정보관리학회지, 제5집, 2006.

[5] 양정진, “시맨틱 웹에서의 온톨로지 공학”, 정보확학회지, 제21권, 제3호, pp. 28-35, 3월, 2003.