자연어 질의 정보 검색 시스템의 비주제어 탐색 방법을 통한 성능 개선

o
이승률, 강현규, 박세영, 이상조*

한국 전자 통신 연구소, 언어 정보 연구실
*경북 대학교 컴퓨터 공학과

Improving the performance of natural language information retrieval system by using non-keyword search methods.

o
Seung-Ryul Lee, Hyun-Kyu Kang, Se-Young Park, Sang-Jo Lee*

요 약

본 논문에서는 한글 문서 검색 시스템에서 자연어 질의어로 검색할경우, 질의어를 주제어와 참조어로 나누어 재구성하여 검색하는 방법을 제시 하였다. 먼저 주제어로 전문 검색을 하여 후보 카드들을 추출한 후 비주제어로 다시 본문 탐색을 하여 추출된 카드의 가중치를 재조정 함으로써 카드추출의 정확성을 높였다. 이 논문에 제시된 방법의 실험은 한국전자통신연구소 언어정보연구실에서 개발한 멀티미디어 전자 백과 사전의 자연어 검색 모듈에서 행하여 졌다. 이 방법으로 별다른 검색속도의 저하나, 저장공간의 추가가 없이 기존의 검색 방법에서보다 약 58%정도의 검색의 정확성이 올라갔다. 본논문에서 제시한 검색의 방법은 여러가지 응용의 자연어 인터페이스에서 데이타를 검색하는 정보검색의 분야에 적용되어 정확성을 높일 수 있을 것이다.

1. 서 론

최근들어 무수히 쏟아지고 있는 텍스트데이타를 가공하여 사용자의 요구에 해당하는 정보만을 추출하여 사용자에게 제공하여 주는 정보 검색 시스템이 각광 받고 있다[1].

본 논문에서는 사용자가 자연어로 질의를 하였을 경우에 이 자연어 질의에서 카드 단위로 된 전자사전 검색에 대한 효과적인 기법을 제시한다.

정보 검색 시스템에서의 주제어란 다량의 텍스트 데이타 베이스에서 각 카드들을 다른 카드와 구분하여 대표할 수 있는 단어들을 말하며, 키워드, 또는 검색어라고 불리기도 한다.

자연어 질의어로서 정보검색을 행하는 방법은 질의어에서 키워드를 추출하여 이 키워드가 각 카드와 유사도를 구하여 카드를 추출해 내는 것과 구문분석을 행하여 질의어보다 작은 단위의 심층구조로 표현하여 그것을 가지고 검색을 하는 방법이 있다. 후자가 보다 정확한 검색을 할 수 있지만 실지로 구문분석이 어려워 그 복잡도에 비해 좋은 효과를 기대하기 힘들다.

그래서 일반적으로 자연어 검색시 구문 분석을 하지 않고 주제어만 추출하여 검색을 행한다[4].

일반적으로 사용자가 자연어 질의를 하는 형태는 무형적이다. 사용자의 임의로 질의 형태가 달라지기 때문이다. 이러한 경우 키워드에 등록되지 않은 단어가 질의어에 포함될 수 있다.

키워드에 등록되지 않은 단어는 원칙적으로는 검색에 참여를 하지 않는다. 그러나 이 단어가 키워드에 등록되어 있는 단어와 이형동의어일 경우 이 단어를 전거어라고 하여 키워드에 등록되어진 단어로 바꾼 후 검색에 참여하게 된다.

제2장에서는 비주제어 탐색방법의 필요성으로써 주제어와 비주제어 그리고 자연어 질의 형태를 분석하고 방법을 제시한다. 제3장에서는 실제로 구현에 대하여 설명하고, 4장에서는 실험 및 평가를 한다. 마지막 5장에서 결론을 맺는다.

2. 비주제어 탐색의 필요성

2.1 주제어 (키워드)

주제어는 정보 검색 시스템에서 문헌을 유사한 것들끼리 묶을 수 있는 능력을 갖춘 단어로써 검색시 이것을 가지고 검색을 행한다. 이것을 키워드라고도 한다.

Salton은 특정한 단어가 한 문헌 집단 속의 상호 관련없는 문헌들을 분리 시키는 능력치가 큰 것이 좋은 키워드가 되고 나쁜 키워드일수록 관련없는 문헌들을 묶어 준다고 하였다[3].

실지로 문헌 분리도가 높은 단어들은 명사와 고유명사 등에 밀집되어 있고 문헌 분리도가 낮은 단어들은 수사, 대명사, 전치사, 관형사, 형용사, 부사등에 몰려 있다는 점과 한국어에서는 체언에 비해 용언이 변형이 많아, 용언의 형태소 분석은 어려워 모호성을 많이 발생한다 . 이러한 모호성은 키워드 성능에 치명적 악영향을 미칠 수 있다.

그래서 보통의 한국어 정보 검색 시스템에서는 주제어를 명사와 명사구만을 가지고 사용한다.

2.2 비주제어가 가진 정보

위에서 정의된 주제어가 아닌 단어들인 형용사, 부사, 수사 같은 것들을 비주제어라고 하자. 이러한 것들은 그 의미가 명사와 같이 고정되어 있지 않고 명사에 비해 모호하다. 그래서 키워드로 적당하지 않아 키워드에서 제외되었다.

그러나 이러한 단어들은 홀로 사용되었을 경우에는 문헌분리도가 낮아 키워드로 부적당하지만 키워드(명사)와 함께 사용되어 있을 때에는 그 키워드를 보조하여 의미가 명확하게 한다. 결국, 보다 나은 정보 검색을 위해서라면 이러한 정보를 버릴수는 없을 것이다. 실지로 구문분석을 하여 정보검색을 하는 기법은 이러한 기능어의 정보를 이용한다고 볼 수 있다. 그러나 이 방법은 구문분석 자체의 어려움으로 인하여 그 복잡성에 비추어 효과가 떨어진다.

그러면 실지 비주제어가 자연어 질의문에서 어느정도 나타나는지 보자. 대학생 15명에게 실험용 자연어 질의어를 각 10개씩 150개를 받았다. 이가운데 주제어와 불용어만으로 구성되어 있는 것이 22개이고 나머지 128개는 비주제어를 포함하고 있었다. 이 비주제어 중에는 그 문장의 형태에 따라 실지 의미에 영향을 미치는 정도가 다르지만 어떤 것들은 아주 큰 영향을 미치기도 하였다. 아래 표에서 볼 수 있듯이 150개의 질의어중에서 순수하게 탐색어로만 이루어 진 것은 15%에 불과하다. 키워드만을 가지고 검색 할 경우 나머지 85%의 자연어 질의의는 각자 어느 정도씩 정보손실을 한다고 볼 수 있다.

전체 샘플 질의어 문장 - 150개
주제어와 불용어만 있는 것 - 22개 (A type)
참조어를 가지고 있는 것 - 128개(B type)

<그림 1> 비주제어의 존재여부에 대한 통계

2.3 자연어 질의어 형태 분류

- 보편적 객체에 관한 것
(명사에 대한 관형어의 제약이 없는 것)
파충류의 생활에 대하여
8월의 꽃은?

- 구체적 객체에 대한 것
(명사에 대한 관형어의 제약을 가진 것)
세계에서 가장 높은 빌딩은?
가장 빠른 비행기에 대해
긴 잎을 가진 식물을 말해 주시오

- 서술 형태의 제약
지구가 도는 이유는?
다람쥐와 비슷한 동물은?

- 사용자가 해답 형태 요구
(질의 응답시스템)
원자의 종류는 몇가지 인가?
영어로 "바둑"이란?
북극성을 기준으로 북극성과 카시오페이아의 거리비는?

- 예/아니오 질문
미국에 북한대사관이 있는가?
인류의 조상은 원숭이인가?

위 분류중 사용자가 요구하는 형태의 질문이나 Yes/No 질문은 정보 검색시스템이라기 보다는 질의응답시스템에 가까우므로 조사대상에서 제외되었다.

보편적 객체에 관한 것은 오직 명사와 불용어만 질의어에 사용 되어진 것으로 전거어 처리만 해주고 탐색을 하면 된다. 그러나 구체적 객체에 대한 질문과 서술 형태의 제약에 관한 질문은 비주제어인 형용사나 부사와 같은 것들이 들어가 있다. 이러한 형태의 질문들은 키워드만으로 검색하였을 때 추출 된 카드가 많을 경우에는 비주제어의 역할이 아주 중요하다고 할 수 있다.

2.4 전문 검색과 본문 검색

전문 검색은 미리 구축하여 놓은 하부 화일을 이용하여 검색의 단위(카드)를 검색하므로 속도의 이익이 있으나 미리 하부화일을 만들어 놓아야 하여 부가적인 저장 공간이 필요하다. 반면 본문 검색은 카드속의 하나의 문장단위까지 검색을 할 수 있어 보다 정교하게 검색을 하지만 인덱스된 하부화일이 존재하지 않으므로 속도가 느리다는 약점이 있다.

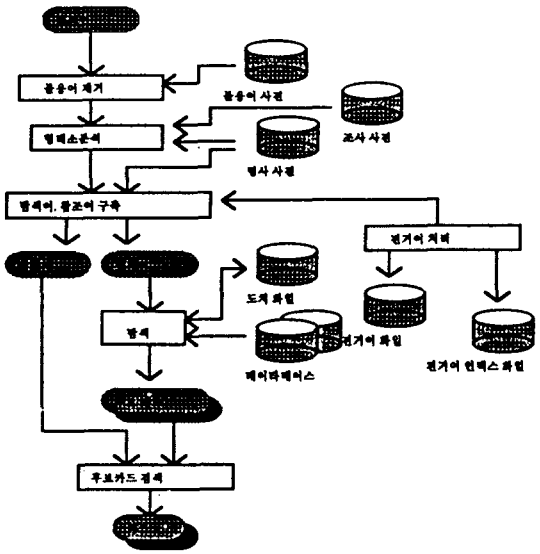
여기서는 자연어 질의어를 주제어와 비주제어로 나눈후 주제어로 전문 검색을 행한후 추출된 후보 카드들에 대하여 다시 본문 검색을 하여 카드 가중치를 재조정 하여 카드추출의 정확성을 높인다. 전문 검색을 하여 추출된 카드는 전체 카드의 극히 일부분의 카드이므로 본문 검색을 하는데 거의 실시간 검색을 할 수 있다.

3. 구 현

한국전자통신연구소 언어정보연구실에서 개발한 멀티미디어 전자사전검색시스템인 "옥서"에서 구현되었다. 옥서는 "(주)계몽사 학생 대백과 사전"[2]을 근간으로 하여 10M byte정도의 데이타, 32,000여개의 카드(항목)에 키워드가

10만여개, 명사사전, 조사사전, 전거어 사전을 갖추고 있는 사전검색 시스템으로 IBM-PC Windows 3.1 환경에서 구축 되었고 수행된다.

옥서의 자연어질의 정보 검색 모듈의 구조는 아래 그림2와 같다.



<그림 2> 자연어질의 처리용 모듈

3.1 형태소 분석

주어진 자연어 질어의로부터 최장 일치 방법으로 명사와 기본 명사구를 추출한다.

[명사구 기본형]
(1) 명사 + 속격조사 + 명사 (예:인간의 신체)
(2) 명사 + 복합조사 + 명사 (예:자유에의 갈망)
(3) 명사 + ∅ + 명사 (예: 송아지 눈)

기본 명사구는 위의 기본형의 반복으로 이루어져 있다. 추출된 명사나, 기본 명사구에서 모든 조사를 제외한다. 추출되지 않은 어절을 참조어 리스트에 넣는다.

3.2 주제어, 참조어 구축

형태소 분석결과 명사나 기본명사구를 제외한 비 주제어를 참조어로 정의하여 참조어 리스트에 수록하고, 명사나 기본명사구를 검색어 리스트에 수록한다. 그리고 기본명사구를 분리하여 만들어 지는 부분 명사, 기본명사구를 추출하여 검색어 리스트에 수록 한다.

[정의] 참조어: Wref

Wref = Wtot - Wkwd - Wstp
Wtot : 전체 단어 집합
Wkwd : 명사, 명사구의 주제어 집합
Wstp : 지시 대명사, 질의에 쓰이는 불용어 집합

3.3 전거어 처리

하나의 뜻을 나타내는 다른 형태의 키워드들 즉, 이형동의어를 처리하는 것을 말한다. "배", "선박"과 같은 관계가 그것이다. 이것은 사용자가 하나의 미리정의된 형태로 통일 시켜주는 것이 필요하다. 이것은 전거어 사전을 검색하여 표준형태를 찾아서 검색어와 바꾸어 줌으로써 해결이 가능하다.

3.4 검색

검색은 도치화일 검색기법을 사용하여 행한다. 위에서 나타난 검색어 리스트에 있는 검색어가 나타나는 카드들을 도치화일을 검색하여 찾아내어 가중치를 계산한다. 찾아낸 후보 카드들을 가중치순으로 특정 갯수만큼 추출한다.

[도치화일 형식]
키워드^Card_Number^Weight^Card_Number^Weight^...

[Card Weight계산]

$$CWi = \sum CW(i, j) * K^{dn}$$

CWi = i 카드의 가중치
CW(i, j) = i 카드의 j 검색어에 대한 weight
K = 상수
dn = 카드 추출 횟수

3.5 후보카드 조사

키워드만으로 검색된 카드들을 다시 본문검색을 행하여 추출된 카드의 가중치를 재조정 한다 본문검색은 추출된 카드의 내용을 검사하여 탐색어들이 인접하여 출현하는 정도를 찾아낸다. 인접한 정도를 구하는 것은 다음과 같은 것이 있을 수 있다.

(1) 문장 단위
- 인접한 단어사이의 범위를 한문장으로 정의한다.
- 인접한 단어사이의 간격을 사이에 오는 문장의 최대수로 정의 한다.

(2) 어절 단위
-인접한 단어사이의 간격을 사이에 오는 어절의 최대수로 지정한다.

(3) 구 단위
- 인접한 단어사이의 범위를 구문분석하여 묶여지는 구로 결정한다.

구단위로 인접한 단어범위를 결정하는 것이 보다 정확할 것이나 이것은 문장에서 구문 분석을 통하여 구단위 추출을 행하여야 하므로 실지로 정확한 추출이 힘들어 현재로서는 구현이 힘들다.

그래서 여기서는 문장단위의 범위결정과 어절단위의 범위결정을 혼합하여 가중치를 다르게 부여한다. 즉, 키워드로 추출된 카드속에 참조어들의 존재가능성을 보고 존재하면 키워드와의 거리의 반비례하는 가중치(Cdis)를 구하여 이러한 것들의 합을 참조어와 카드와의 유사도라 한다.

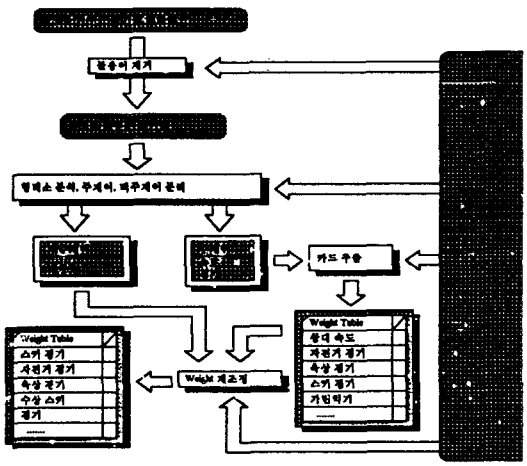
Cdis는 참조어와 주제어 두 인접단어에 따른 거리에 따라 주어지는 패러미터로 많은 실험을 통하여 결정되어야 한다. 실지 거리가 어느 정도 이상되었을 경우 이 값을 0으로 해주어 두 단어가 전혀 관계가 없음을 나타낼 수 있다.

새롭게 갱신된 카드의 가중치는 주제어와의 유사치와 참조어와의 유사치와의 곱으로 결정한다.

3.6 검색예

자연어 질의어 : "속도를 겨루는 경기에 대하여 ?"

"속도를 겨루는 경기에 대하여"라고 사용자가 물었다. 이 질의 문장은 불용어인 "대하여?"가 제거되고 주제어와 참조어로 나뉘어 진다. 주제어인 "속도", "경기"를 가지고 추출한 카드의



< 그림 3 > 자연어 질의 검색의 예

타이틀이 가중치별로 정렬 되어 나온다. 그림3에서 보다시피 "상대속도"와 "가변익기"와 같은 카드들이 "속도"에 많은 가중치를 가져 나올 수 있다. 이 카드들을 참조어 "겨루는"을 가지고 본문 탐색을 하면, 본문 중에, "속도", "경기"와 "겨루는"이 한 문장속에 나타나는 것들은 가중치가 올라 가서 보다 만족할 만한 결과를 보여 줄 수 있다.

4. 실험 및 평가

전체 표본 질의어 문장 150개 중에서 위에서 본 것과같은 잘못된 질문형태와 본문중에 내용이 없는 것을 제외한 24개의 질문을 선택하여 일반적인 방법으로 추출한 카드들과 여기에 다시 본문검색을 하였을 때의 카드들을 가중치별로 정렬하여 올바른 카드가 순위가 높아 졌으면 +1, 그대로이면 0, 낮아졌으면 -1을 각기 매겨 주었다. 전체적으로 나아진 개선된 카드가 14개, 그리고 변화 없는 카드가 7개, 그리고 마이너스 값이 나오는 카드가 3개가 생겼다. 즉, 58%의 카드의 정보추출이 보다 정확하여 졌다고 할 수 있다. 그리고 12%의 값이 마이너스가 나오는 데 이것은 주제어가 다수 있을 경우 비주제어가 잘못된 주제어에 붙기 때문에 발생한다.

5. 결론

이 논문에서는 질의어를 명사, 명사구를 키워드로 하여 전문검색하고 질의어에서 키워드로 선택되지 못한 나머지에 대하여 키워드와 관련지어 본문검색을 하였다. 이 방법의 장점은 다음과 같다.

첫째, 명사에 대하여만 키워드추출을 행하기 때문에 형태소 분석이 간단하며, 추출된 키워드가 명확하며 키워드의 수를 줄여 효과적이다.

둘째, 추출된 카드에 대하여 본문 검색을 하여 가중치를 조정함으로써 보다 정확한 카드추출에 도움을 준다.

보다 많은 실험을 하여야 하겠지만 현재의 실험에서는 약 58%의 성능 향상을 보였다.

그러나 구문분석을 하지 않아 구현이 쉽기는 하지만 보다 정확한 의존관계에 기초하지 않아 참조어와 키워드간의 연관관계가 잘 못 생성될 수 있다. 여기서는 이것을 단어 거리에 따라 가중치를 조정하여 줌으로써 해결하였지만 정확한 가중치를 만들기 위해서는 보다 많은 실험과 연구가 뒤따라야 할 것이다.

참고 문헌

[1] William B.Frakes, Ricardo Baeza-Yates, Information Retrieval, Prentice Hall, pp. 15-18, 1992.

[2] 계몽사 학생 백과 사전 1-6, 계몽사, 1992.

[3] 정영미, 정보 검색론, 정음사, pp.296-312, 1986.

[4] 이창열, 강현규, 박세영, "자동 키워드 제작기 시스템 설계", 제 5회 한글및 한국어 정보 처리 학술 대회, 한국 인지 과학회, 한국 정보 과학회, pp.71-73, 1993.

[5] 강현규, 장호욱, 이승률, 박세영, "옥서에서의 표제어와 자연어 검색의 설계 및 구현", 한국 정보 과학회 추계 학술발표 논문집(A), pp.633-636, 1994.