한국어 정보검색 시스템의 성능 향상을 위한 용언 색인

박진희○, 박대원, 박민식, 남현숙, 김광영, 권혁철
부산대학교 전자계산학과 인공지능연구실
jinhee@solge.cs.pusan.ac.kr

Predicates Indexing for efficiency improvement in Korean Information Retrieval System

Jin-Hee Park○, Dae-Won Park, Min-Sik Park, Hyeon-Sook Nam,
Kwang-Young Kim, Hyuk-Chul Kwon
Dept. of Computer Science, Pusan National University

요 약

지금까지 대부분의 정보검색 시스템은 명사만을 색인어로 추출하여 사용하였다. 명사는 문서를 대표할 수 있는 어휘 요소이다. 그러나 명사 색인어만 가지고는 문서의 주제를 정확하게 나타낼 수 없다. 본 논문은 명사 색인어와 함께 용언도 색인어로 추출하여 사용하는 한국어 정보 검색시스템을 제시한다. 또한, 용언 색인어와 명사 색인어의 상대적 가중치를 검색에 이용하여 사용자의 질의에 적합한 문서를 검색할 수 있도록 한다. 이러한 과정에서 발견된 문제점은 향후 연구 과제로 계속 향상시켜나갈 것이다.

1. 서론

오늘날 인터넷의 발달로 웹 문서는 기하급수적으로 증가하고 있다. 수많은 문서들 속에서 원하는 정보를 효과적으로 검색할 수 있는 정보검색시스템의 개발이 매우 중요하다.

정보검색 시스템의 성능 향상은 사용자가 원하는 문서를 더욱 높은 순위에서 제시하는 것이다. 이를 위해서는 무엇보다 색인어를 정확하게 추출해야 한다.

명사는 문서를 가장 잘 나타낼 수 있는 어휘 요소이지만 명사 색인어만을 이용하는 검색 시스템은 사용자가 원하는 정보를 정확히 찾기가 어렵고, 검색 결과 또한 상당한 양으로 사용자가 원하는 정보인지를 쉽게 판단하기가 어렵다.

용언은 명사를 서술하여 명사의 의미를 구체화하고 명확히 하는 역할을 하므로 이를 색인어로 이용함으로써 검색의 정확도를 높일 수 있다.

본 논문에서는 명사 뿐만 아니라 용언도 색인어로 추출하여 사용하는 한국어 정보 검색시스템을 제시하고, 용언 색인어와 명사 색인어의 상대적 가중치를 검색에 이용하는 방법에 대해 제안한다.

2. 색인

색인 시스템은 저장 공간과 검색 효율을 높이기 위해서 문서를 대표할 수 있는 색인어를 추출해야 한다. 명사가 문서를 대표할 수 있는 가장 적합한 어휘 요소이다. 그러나 문서에서 색인어를 명사로만 한정하는 것은 부적절하다.

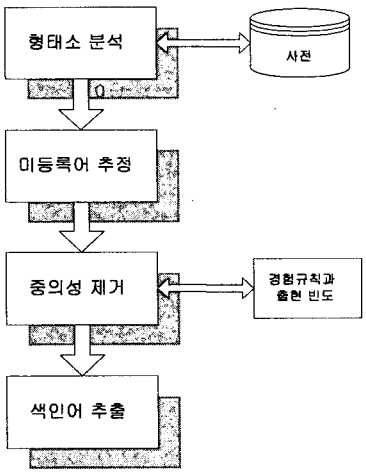
검색 시스템의 사용자들은 더 정확한 검색 결과를 얻기 위해서 단순히 핵심단어가 아니라 얻고자 하는 내용을 잘 설명하기 위해 여러 가지 단어로 구성된 질의문을 사용한다. 이 때 이 질의문은 명사를 나열한 것일 수도 있겠지만, 용언을 이용하여 명사를 수식한 것일 수도 있다. 이때 명사만을 색인어로 이용한 검색 시스템보다 용언까지 색인어로 이용한 검색 시스템이 사용자가 원하는 더 정확한 정보를 제공한다.

2.1 색인 과정

색인어 추출은 세 단계의 과정을 거쳐야 한다. 일단 주어진 문장이나 어절의 형태소 분석을 하는데, 속도 개선을 위하여 상위빈도의 말뭉치를 미리 분석해 놓은 상위 빈도 어절사전에서 먼저 찾고 결과가 없는 어절에 대해서만 기존의 방식으로 형태소 분석을 한다.

둘째, 사전에 등록되지 않은 미등록어를 추정하고, 셋째, 중의성을 제거한다. 중의성 제거에서는 앞 뒤 어절간의 경험 규

칙 적용과 말뭉치 자료 분석에 의해 얻어진 단어의 출현 빈도를 이용한다. 그리고 마지막으로 색인어 생성 규칙에 따라 색인어를 추출한다.



[그림 1] 색인 과정

2.2 명사 색인

문서에서 보통 색인어를 명사로 국한하는데 이는 명사가 문서의 중요한 의미를 나타낼 수 있는 어휘 요소이기 때문이다.

명사를 추출할 때는 단일 명사 뿐만 아니라 직접적으로 복합명사의 범주에 정확하게 넣을 수는 없지만 복합명사로 받아 들일 수 있는 명사 유형을 판별하여 복합명사 형태로 색인하고, 띄어쓰기 없이 표기된 복합명사를 단일명사로 분리하여 색인한다. 단, 영어 단어는 복합명사로 색인을 하지는 않는다.

원 문	공유자원의 정확한 운용이나 프로세스 사이의 커뮤니케이션을 위하여 중요한 ...
색인어	공유,공유자원, 자원, 정확, 운용, 프로세스, 사이, 커뮤티케이션, 중요

2.3 용언 색인

동일한 명사를 가지고 있는 문서라도 전혀 다른 의미를 가질 수 있기 때문에 명사만으로는 문서의 의미를 정확히 나타내기 어렵다.

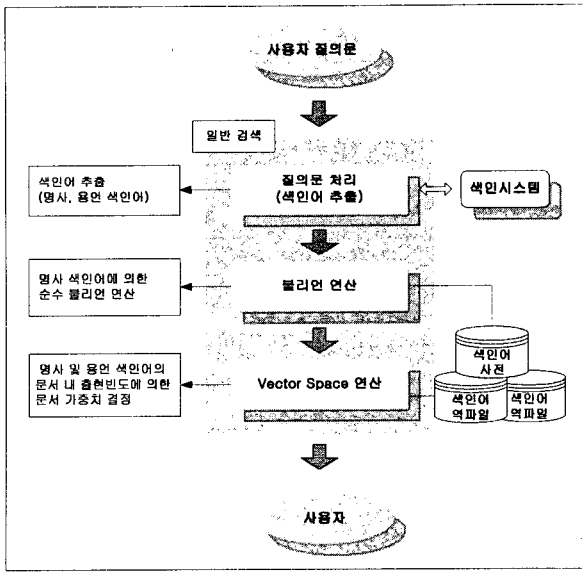
한국어에서 용언은 명사를 서술하여 명사의 의미를 구체화하고 명확히 하는 역할을 하기 때문에 단독으로 검색에 이용되지는 않고 명사와 함께 색인어로 사용된다.

따라서 용언을 색인어로 추출하되 명사 색인어보다 낮은 가중치를 둔다.

용언 색인어를 이용한 어절 거리 연산을 위해 명사 색인어와 함께 포스팅 파일을 구성하여 검색에 이용한다.

원문	색인어
세상에서 가장 빠른 사나이	세상, 빠르다, 사나이
김치 담그는 법	김치, 담그다, 법
밤으로의 긴 여행	밤, 길다, 여행

3. 검색



[그림 2] 검색 과정

검색시스템은 사용자 질의문을 처리하여 사용자 질의에 적합한 문서를 검색하여 제시한다. 색인어 빈도에 의한 검색은 사용자 질의문 처리, 불리언 연산, 문서 순위 결정 과정을 거쳐 결과를 얻는다.

먼저, 사용자 질의문을 분석하여 색인어를 추출한다. 색인 시스템을 이용하여 사용자의 자연언어 질의문에서 검색을 위한 색인어를 추출한다.

두 번째, 색인어 사전의 색인어 ID로 색인어를 포함하는 문서 리스트를 얻는데, 이를 질의어 처리 규칙에 따라 AND 또는

OR연산을 수행한다. 이때 용언 색인어 단독으로는 검색에 이용되지는 않으므로 용언 색인어는 제외시켰다.

세 번째로 문서 내 색인어 등장 빈도 tf(term frequency)와 색인어를 포함하는 문서 빈도 df(document frequency)를 이용하여 불리언 연산에서 얻은 문서리스트와 사용자 질의문과의 유사도를 벡터 스페이스 모델에 의해 계산하여 문서의 순위를 결정한다.

4. 실험

본 논문에서는 사용자 질의문과 문서와의 유사도를 계산할 때 명사 색인어로 유사도를 계산하고, 여기에 용언 색인어에 의한 가중치를 더해서 유사도를 계산한다. 이때 용언 색인어의 가중치는 명사 색인어로 계산한 가중치보다 낮게 두었다.

$$Weight_{I_{Pred}} = \alpha \times \mathrm{Sim}\,(Q_{I_{pred}}, D_d)$$

I_{Pred} : 용언 색인어

α : 용언 색인어 가중치 상수 $(0 < \alpha < 1)$

본 논문의 실험은 신문기사 데이터 40만 건을 대상으로 하였고, 총 데이터의 크기는 630,495KB이다. 그리고, 검색 결과의 정확도는 20개 질의문의 검색 결과 상위 10개 문서에 대한 적합성 판단으로 비교하였다. 실험에 사용된 질의문은 "맑은 물 공급 대책", "임금 상승에 따른 채산성 악화", "멸종위기에 처한 동식물" 등 이었다.

먼저 색인어 개수는 용언 색인어를 포함했을 때, 평균 1.17배 증가했다.

검색시스템의 검색 정확도 실험을 위해 검색 Mode를 두어 검색 결과를 비교 분석하였다. Mode 1은 기존의 검색 시스템에서 사용하던 방법으로서 명사 색인어만을 이용하여 사용자 질의문의 색인어로 OR 연산을 하고 Vector Space Model의 유사도 함수로 문서순위를 결정하고, Mode 2는 명사 색인어와 용언 색인어를 모두 이용하며 OR 연산과 벡터 스페이스 모델의 유사도 계산 함수로 순위를 결정한 것이다.

적합 문서의 비율은 Mode1에서 Mode2로 18.2%의 향상을 보였다.

5. 결론 및 향후 연구

이 논문에서는 명사 뿐만 아니라 용언도 색인어로 사용하는 한국어 정보검색 시스템을 제시하였다. 이 때, 용언 색인어의 가중치를 명사 색인어의 가중치와 동일하게 두지 않고 상대적으로 낮게 설정하여 검색의 효율을 높였다.

이 연구에서는 문서 내에서 색인어의 위치는 고려하지 않았다. 문서 내에서 색인어 위치는 출현 빈도와 함께 검색의 정확도를 높이는 중요한 요소로 작용할 수 있다. 그러므로, 명사 색인어와 용언 색인어의 문서 내 위치 정보를 포스팅 파일(Posting File)로 구성하여 색인어의 위치 정보를 검색에 이용할 수 있는 시스템을 연구하여야 한다.

그리고, 향후 보다 나은 시스템 개발을 위해 명사, 용언 색인어와 함께, 부사, 관형사를 색인어로 추출하여 검색에 이용하는 검색시스템을 고려해 볼 수 있다. 이 때 명사, 용언, 부사, 관형사 등 여러 색인어를 검색에 적용하기 위한 효과적인 색인어 역파일 구성에 대한 연구가 필요하다.

또한, 숫자와 영어를 포함하는 명사구를 색인하는 시스템과 명사, 용언 색인어와 함께 부사, 관형사를 색인어로 추출하여 검색에 이용하는 시스템도 고려해 볼 수 있다.

참고문헌

[1] Robert R. Korfhage, "Information Storage and Retrieval", Willey Publishing, 1997

[2] Dornna Harman, "An Experiment Study of Factors Important in Document Ranking", Information Retrieval, 186-193, 1986, 8.

[3] Gerard Salton, Michael J. McGill, "Introduction to Modern Information Retrieval", McGrawHill, 1983.

[4] 정영미, 정보검색론, 구미무역 출판부, 1993.

[5]박대원, "용언 색인을 적용한 한국어 정보 검색시스템의 검색효율 향상", 이학석사 학위논문, 부산대학교 전자계산학과, 2000

[6] 최종희, 최동시, 박세영, 오희국, " 다중단어를 사용한 정보검색 시스템에서의 재현정확도 향상방법", 정보과학회 가을 학술발표논문집, 1998