모든 품사 색인을 이용한 정보 검색 시스템

전영진° 강승식
국민대학교 컴퓨터학부
{terrius7°, sskang}@cs.kookmin.ac.kr

Information Retrieval System by Indexing All the Parts of Speech

Youngjin Chun° Seungshik Kang
School of Computer Science, Koomin University

요 약

사용자가 원하는 정보를 찾아주는 정보 검색 시스템은 최근까지 많은 연구가 이루어지고 있다. 일반적인 검색 엔진에서는 질의문이나 문서들을 색인할 때 명사를 중심으로 색인을 한다. 본 논문에서는 사용자가 한단어로 질의하지 않고 서술문 형식으로 질의를 할 경우 사용자가 원하는 정확한 정보를 검색해 내기 위하여 품사에 관계없이 사용자가 선택한 질의어들과 문서에서 나타나는 모든 용어들을 중요하다고 보고 명사만 이용하는 경우와 모든 품사를 이용하는 경우로 나누어 실험해 보았다.

1. 서 론

오늘날은 컴퓨터, 인터넷 등의 발달로 인하여 유용한 정보들이 빠른 속도로 증가하고 있다. 정보의 대상도 수치 정보, 사실 정보, 문서 정보, 서지 정보, 그림 정보, 음성 정보와 복합 정보 등 다양한 형태를 지니고 있다. 최근에는 전자 매체의 발달로 인해 검색의 대상이 본문 검색(text retrieval), 화상(image), 음성(sound), 화학식의 구조 등으로 확대되고 있다[1]. 이런 정보들이 증가함으로써 많은 유용한 정보들을 얻을 수 있지만 원하는 정보나 인터넷 사이트 등을 수작업으로 찾아내기란 거의 불가능해졌다. 그렇기 때문에 사용자들은 대부분 정보 검색 시스템이라는 도구를 이용하게 되었고 간편하고 빠른 속도로 정보를 찾는 것이 가능해졌다.

정보 검색 시스템이란 정보 수요자가 필요하다고 예측되는 정보나 데이터를 미리 수집, 가공, 처리하여 찾기 쉬운 형태로 축적해 놓은 데이터베이스로부터 정보 요구자가 선택한 질의를 이용하여 적절한 정보 검색 처리 과정을 거친 후 순위가 매겨진 문서를 신속하게 찾아내어 정보 요구자에게 제공하는 일종의 인공지능형 software이다.

정보 검색 처리과정은 미리 선택된 검색 모델로 사용자 질의와 문서간의 유사도 측정 공식에 따라 사용자의 질의를 받아 문서와의 유사도를 계산하는 과정으로 정보 검색 시스템의 핵심은 유사도 측정이다. 이는 데이터 검색과 뚜렷이 구분되는 정보 검색의 특징이기도 하다. 정보 검색 처리과정에서 질의문이나 문서 자료 내에서 질의어(query term) 혹은 색인어(index term)의 중요도를 반영하려면 문서 분석 기법에 의한 색인어를 추출하여 이 색인어의 가중치를 계산해야 한다[2].

본 논문에서는 사용자의 질의문과 문서와의 유사도를 계산하기 위해 벡터 공간 모델을 이용하였고 질의문과 문서를 색인할 때 추출하는 색인어의 대상을 명사에 한정하지 않고 모든 품사의 용어들로 확대하는 방법을 제안하였다. 그리고 이 방법의 성능을 비교하기 위하여 기존의 논문에서 행했던 명사만을 이용한 방법과 비교하여 보았다.

본 논문의 구성은 다음과 같다. 1장에서 연구의 목적 및 배경에 대하여 기술하고, 2장에서는 벡터 모델에 대한 소개 및 추가 가중치 기법과 본 논문에서 제안한 방법에 관하여 소개 한다. 3장에서는 제안한 방법으로 실험을 행하고 이에 대한 결과 분석을 통하여 기존에 행했던 방법과 비교 분석을 한다. 4장에서는 제안한 방법에 대한 결론을 맺고 향후 연구할 방향에 대하여 기술한다.

2. 벡터 공간 모델과 가중치 기법, 색인 기법

2.1 벡터 공간 모델

벡터 공간 모델은 코넬대학의 Gerald Salton 교수가 만든 것으로 상당히 간단하면서도 어느 정도 질의와 문서와의 관련성을 통계적 및 의미적으로 접근하고 있다. 이 모델은 벡터 공간 모델 상에서 각 문서들과 질의어들은 n차원 공간 속의 벡터들로 취급되며, 이때 각 차원들은 색인 용어들로 표현된다. 이 모델의 장점은 용어 가중치 기법이 검색 성능을 향상 시킬 수 있고 질의에 근접한 부분 정합 문서 검색과 질의와의 유사도에 따라 문서의 순위화가 가능하다. 또한 질의 확장이나 연관 피드백을 사용하여 성능을 향상 시킬 수 있고 무엇보다 속도가 빠르고 단순하기 때문에 가장 대중적인 검색 모델이라 할 수 있다. 반면에 색인어간의 상호 독립성을 가정함으로써 실제적으로 용어 종속이 있을 경우 성능에 악영향을 미칠 수 있다는 단점이 있다[3].

2.2 추가 가중치 기법

이 추가 가중치 기법은 기존의 논문에서 제안했던 기법으로서 3가지 기법을 제안하고 각 기법의 성능에 대하여 실험하였는데 본 논문에서는 같은 상황에서 성능을 비교하기 위해 같은 방법으로 가중치를 사용하였다[4].

2.2.1 용어 가중치 기법(TW)

1) 본 연구는 첨단정보기술연구센터를 통하여 과학재단의 지원을 받았음.

이 기법은 용어빈도를 이용하고, 복합 명사분해, 품사 유형 및 어절 위치 등을 고려한 가중치 기법이다.

2.2.2 문장의 길이에 대한 가중치(SLW)

질의문을 분석하여 얻은 질의어 개수에 대한 가중치

2.2.3 질의어 출현 빈도 가중치(QFW)

질의문을 분석하여 얻은 질의어들이 시스템이 검색한 문서에서 다수 출현하는 문서에 대한 가중치

2.3 질의문과 문서에서의 모든 품사를 이용한 색인기법

일반적인 검색시스템에서는 주로 명사만을 이용한다. 사용자의 질의문을 분석할 경우에도 명사만을 이용하기 때문에 나머지 품사에서 사용자가 보내주는 많은 정보를 놓칠 수도 있다.

본 논문에서는 사용자가 선택한 질의문과 수집되어 있는 문서에서 명사만이 정보를 갖고 있다고 보지 않고 다른 품사의 용어들도 많은 정보를 갖고 있을 수 있다고 가정하였다. 즉 사용자가 보여준 단서에서 최대한의 많은 정보를 이용하자는 것이고 또한 문서에서도 마찬가지로 이용할 수 있는 모든 정보를 활용하자는 것이다. 따라서 질의문과 문서에 대한 역파일을 만들기 위하여 색인어 추출 시 얻을 수 있는 모든 품사, 즉 명사, 동사, 형용사, 관형사, 부사 그리고 감탄사에 대하여 추출한다. 이 추출한 용어들에 대하여 가중치를 부여하고 유사도 계산에 이용하였다.

3. 실험 및 결과 분석

3.1 실험 환경

본 논문에서는 KTSET의 4352개의 문서를 이용하여 실험하였다. 테스트 질의는 KTSET에 포함된 자연어 질의 50개를 사용하였고 기존 논문의 결과와 비교하기 위하여 기존의 논문에서 사용한 다음과 같은 4가지 경우로 나누어 실험을 진행하였다.

1) 색인어 가중치(TW) 적용
   - 유사도 값을 (0.001~0.3)사이로 임계치 변경
2) 질의문 길이 가중치(SLW)를 적용
   - [(질의어 개수 / 2) * (0.1 ~ 0.9)]
3) 질의어 출현빈도 가중치(QFW)를 적용
   - [질의어 출현 빈도 * (0.1 ~ 0.9)]
4) 질의문 길이 가중치(SLW)와 질의어 출현 빈도 가중치(QFW)를 같이 적용
   - [SLW (0.1 ~ 0.9) * QFW (0.1 ~ 0.9)]

실험 2)~4)의 유사도 임계치 설정: 0.001

3.2 실험 평가 방법

실험 결과의 평가 방법은 정보 검색 평가에서 널리 쓰이는 정확률과 재현율을 사용하였다.

기존의 명사만을 이용한 방법과 모든 품사를 이용한 방법에 대하여 3.1에서 언급한 4가지 경우에 대하여 실험하고 이에 대한 정확률과 재현율의 평균값을 구한 다음 그래프로 표현하였다.

$$\text{정확률} = \frac{\text{시스템이 판단한 문서중 실제 정답문서수}}{\text{시스템이 판단한 모든 문서수}} \times 100$$

$$\text{재현율} = \frac{\text{시스템이 판단한 문서중 실제 정답문서수}}{\text{실제 정답 문서수}} \times 100$$

3.3 실험 및 결과

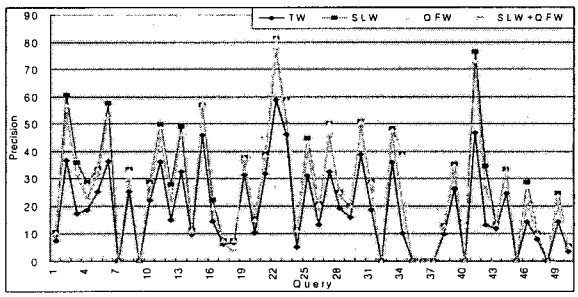

그림 1. 기존의 명사만 이용할 경우 - Precision

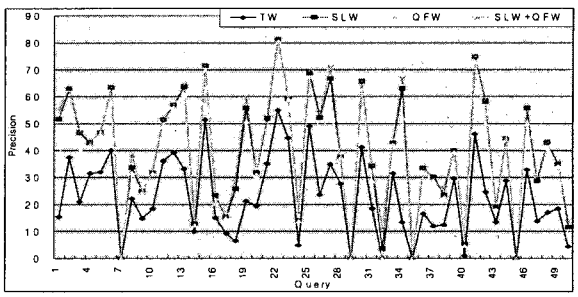

그림 2. 모든 품사를 이용할 경우 - Precision

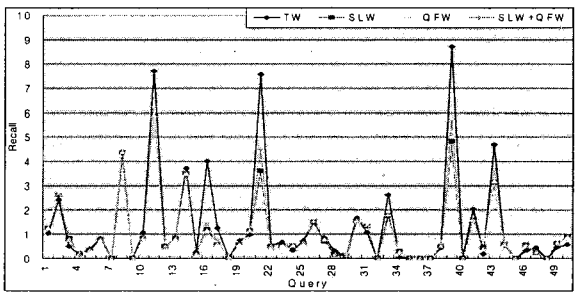

그림 3. 기존의 명사만 이용할 경우 - Recall

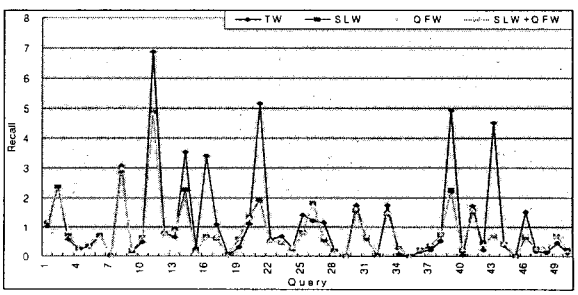

그림 4. 모든 품사를 이용할 경우 - Recall

3.4 실험 결과 분석

그림 1과 그림 2는 정확률을 나타낸 것이다. TW를 적용한 경우, 명사만을 이용한 방법이 모든 품사를 이용하는 방법보다 높은 수치를 보인 질의는 9개, 동일한 경우 3개, 두 방법 모두 정답을 찾지 못한 질의는 4개로 명사만을 이용한 방법이 모든 품사를 이용한 방법보다 높은 경우는 18%였고 모든 품사를 이용한 방법이 높은 경우는 68%로 나타났다. 또, 명사만을 이용한 방법이 검색 실패한 질의를 모든 품사를 이용한 방법이 검색해 내는 경우는 6번 이었고 반대의 경우는 1번이었다.

SLW를 가중치로 적용한 경우에는 명사만을 이용한 방법이 모든 품사를 이용한 방법보다 높은 수치를 보인 질의는 2개, 동일한 경우가 2개, 두 방법 모두 정답을 찾지 못한 질의는 3개로 SLW를 적용했을 때 명사만을 이용한 방법이 높은 경우는 4%였고 모든 품사를 이용한 방법이 높은 경우는 86%였다. 또, 명사만을 이용한 방법이 정답을 검색하는데 실패했으나 모든 품사를 이용한 방법이 검색한 경우의 질의개수는 6개였고 반대는 1개였다.

QFW를 가중치로 적용한 경우에는 명사만을 이용한 방법이 모든 품사를 이용한 방법보다 높은 수치를 보인 질의는 3개, 동일한 경우가 3개, 두 방법 모두 정답을 찾지 못한 질의는 4개로 SLW를 적용했을 때 명사만을 이용한 방법이 높은 경우는 6%였고 모든 품사를 이용한 방법이 높은 경우는 80%였다. 또, 명사만을 이용한 방법이 정답을 검색하는데 실패했으나 모든 품사를 이용한 방법이 검색한 경우는 5개였고 반대는 1개였다.

SLW+QFW를 가중치로 적용한 경우에는 명사만을 이용한 방법이 모든 품사를 이용한 방법보다 높은 수치를 보인 질의는 1개, 동일한 경우가 0개, 두 방법 모두 정답을 찾지 못한 질의는 3개로 SLW를 적용했을 때 명사만을 이용한 방법이 높은 경우는 2%였고 모든 품사를 이용한 방법이 높은 경우는 92%로 최대치를 보였다. 또, 명사만을 이용한 방법이 검색 실패한 질의의 정답을 모든 품사를 이용하는 방법이 검색해 내는 경우는 6개였고 반대는 1개였다.

실제적인 수치를 보면 명사만을 이용한 방법이 모든 품사를 이용한 방법보다 높은 경우는 10point 이내로 성능 차이가 크지 않은 반면, 반대의 경우는 2배에서 성능 차이가 클 경우 최고 8배까지의 차이를 보였다.

그림 3과 그림 4는 재현율을 나타난 것이다. 재현율의 경우는 그래프에서 알 수 있듯이 두 방법 모두 비슷한 수치를 보였고 명사만을 이용한 방법이 높은 수치를 보인 경우는 TW에서 27개, SLW에서 21개, QFW에서 19개, SLW+QFW에서 25개로 모든 품사를 이용한 방법이 높은 경우와 비슷하였고 또, 성능의 차이가 나더라도 대부분 소수점 이하의 수치로 나기 때문에 정확률에 비해서 비교하기가 어려운 측면이 있다. 이렇게 낮은 재현율이 나타나는 이유는 낮은 유사도 임계치를 사용함으로써 정답 문서의 개수에 비해서 시스템이 찾아내는 문서의 수가 10배 이상으로 많기 때문이다.

전체적인 성능을 보면 명사만을 이용한 방법보다 모든 품사를 이용한 방법이 평균적으로 2~4배 정도의 높은 정확률을 보였고 재현율은 비슷한 경향을 보였다. 즉 정확도는 증가하면서도 재현율은 떨어지지 않는 것으로 보아 정확률에 대한 성능이 많이 향상된 것으로 볼 수 있다. 위의 결과에서 보듯이 명사만을 이용한 것보다 여러 품사들을 같이 이용할 경우에 성능이 향상되는 것을 알 수 있었다.

4. 결론 및 향후 연구

본 논문에서는 사용자 질의문과 수집된 문서들을 분석할 때 명사만이 정보를 갖고 있는 것이 아니라 다른 품사들도 많은 정보를 갖고 있어서 보다 사용자가 원하는 정보를 찾아내는데 기여할 수 있다고 가정하고 질의문과 문서를 색인할 때 이를 이용하였다. 그리고 이에 대해서 기존의 명사만을 사용한 방법과 비교하여 보았다. 그 결과 다른 품사들도 동시에 이용하는 것이 성능에 영향을 미친다는 것을 알 수 있었다. 즉 명사만을 이용하는 것보다 명사만이 아니고 동사, 형용사, 관형사, 부사, 감탄사와 같은 여러 가지 품사들을 같이 이용하는 것이 약 2배에서 최고 4배 정도의 더 정확한 검색 결과를 보임을 증명하였다. 이것으로 보아 명사들이 가장 많은 정보를 가지고 있지만 사용자의 질문이 단일어 질문이 아닌 서술문으로 주어질 경우 다른 품사들 또한 많은 정보를 가질 수 있으므로 시스템에 이를 이용한다면 검색에 있어서 더 좋은 성능을 보일 것이라 기대한다.

향후 연구로는 이처럼 주어지는 모든 단어들을 이용할 경우 DB색인 파일인 역파일이 커지는 문제가 발생하는데 이것을 해결하기 위하여 어떠한 품사가 얼마만큼 많은 영향을 미치는지 알아내어 정보를 많이 가진 품사와 정보를 가장 적게 가진 품사를 구분하는 연구가 필요하다. 또한 이것을 이용하여 사용하거나 사용하지 않을 품사를 구분하는 것만이 아니라 각 품사의 가중치 값을 차별화하는데 이용한다면 더욱 더 성능을 높일 수 있을 것이라 본다. 마지막으로 실험에 대한 다른 방법으로 문서를 분석할 때만 명사를 이용하고 사용자의 질문을 분석할 때는 모든 품사를 이용하는 경우와 이 반대의 경우도 같이 실험해 본다면 더 명확한 가중치 부여가 가능할 것으로 본다.

참고문헌

[1] 강승식, "한국어 형태소 분석과 정보 검색", p441, 2002.

[2] 강승식, 이하규, 손소현, 홍기채, 문병주, "조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법", 제28회 한국정보과학회 학술발표 논문집(II), 28권 2호, pp.196-198, 2001.

[3] G.Salton and C.Buckely, "Improving Retrieval Performance by Relevance Feedback, "Journal of the American Society for Information Science, 41(4), 1990.

[4] 전영진, 강승식, "정보 검색에서 질의문 길이에 대한 가중치와 질의어 출현 빈도 가중치 적용", 제 23회 한국정보처리 학회, 2005.

[5] 정영미, 이재윤, "질의확장 검색에서의 추가용어 가중치 최적화", 제 9회 한국정보관리학회 학술대회 논문집, p241-246, 2002.