정보 검색에서 질의문 길이에 대한 가중치와 질의어 출현 빈도 가중치 적용

강승식, 전영진
국민대학교 컴퓨터학부
e-mail : {sskang, terrius7}@cs.kookmin.ac.kr

Applying the Weight for Query Length and the Frequency of Query Term to Information Retrieval

Seung-Shik Kang, Young-Jin Chun
School of Computer Science, Kookmin University

요 약

정보검색 시스템에서 긴 문장으로 질의가 들어올 경우 질의문의 길이와 시스템이 정답이라고 판단한 문서에서 질의문을 분석하여 추출한 질의어들이 출현한 빈도수를 가중치로 준다면 좀더 정확한 결과를 보일 수 있을 것이라 가정하였다. 즉 벡터 모델을 이용하여 문서와 질의와의 유사도를 계산하고 여기에 질의문의 길이에 대한 가중치와 유사도를 이용하여 얻은 결과 문서에서 질의문을 분석하여 얻은 질의 용어들의 출현 빈도에 대한 가중치를 적용하는 방법을 제안하였다.

1. 서론

현재 인터넷은 사용자들의 증가로 인하여 하루에도 셀 수 없을 정도로 많은 정보들이 쏟아지고 있다. 이런 폭발적인 정보의 양의 증가는 사용자들이 원하는 정보를 많이 얻을 수 있지만 그 정보들이 과연 사용자들이 원하는 정보인지 아닌지 판단하기가 쉽지 않다. 즉 사용자 입장에서 보면 자신이 원하는 자료는 가능한 모두 찾아내고 싶고 원하지 않는 자료는 최대한 찾아내고 싶지 않을 것이다. 그렇기 때문에 사용자들은 정확한 정보를 빠른 시간 안에 얻기 위하여 수작업 보다는 검색엔진을 이용한다. 하지만 정보의 양이 증가함에 따라 검색엔진에 의해 검색된 정보들도 실제로 필요하지 않은 정보를 보이기도 한다. 따라서 이를 정확하면서도 신속하게 효율적으로 처리할 수 있는 정보검색기술이 요구되었다. 일반 사용자들은 온라인 데이터베이스를 탐색할 때에도 10 개 이내의 질의 용어를 사용하는 경향이 있으며, 웹 검색시에는 대부분 2 개 이하의 질의어를 사용하고 있는 것으로 나타났다[1]. 일반적으로 질의문이나 문서의 내용을 분석하려면 형태소 분석과 구문 분석, 의미 분석 등이 사용되지만 현재 구문 분석과 의미 분석 기술은 실용적인 시스템에 적용하는데 많은 제약이 있다[2]. 질의문이나 문서 자료 내에서 질의어(query term) 혹은 색인어(index term)의 중요도를 반영하려면 문서 분석 기법에 의한 색인어의 가중치를 계산해야 한다[3].

본 논문에서는 벡터 스페이스 모델을 이용하여 유사도 값을 계산하고 이 값에 질의문을 분석하여 얻은 질의문의 길이에 대한 가중치와 시스템이 정답이라고 판단한 문서에서의 질의어의 출현 빈도에 대한 가중치를 적용한 방법을 제안한다.

2. 관련연구

정영미(2002)는 유사도가 탐색가중치로 최적인가라는 것에 의문을 두고 추가용어와 질의 사이의 유사도가 가지는 특성과 고정가중치를 부여한 경우를 비교해 보았다. 또한 실험집단이나 확장범위의 영향을 덜 받는 최적화된 추가용어 가중치를 찾기 위해 여러 가지 탐색가중치 공식을 이용하여 실험집단 KTSET, Medline, CACM 등을 실험하였다. 그리하여 공기기반 전역적 질의확장에서 추가용어에 부여하는 탐색가중치로 질의와의 유사도를 사용하는 것이 최선의 선택이 아님을 증명하고 0.5 내외의 고정가중치나 가중치 공식을 이용하는 것이 다소 긴 질의문에서 좋은 성능을 보이는 것을 증명하였다[4]. 김상범(2000)은 고려대학교 검색 엔진인 KUIR 을 HANTEC 2.0 의 질의 및

문서 집합과 HANTREC2000 의 질의집합에 대하여 성능을 평가하였다. KUIR 은 검색시 질의 내에서의 빈도만을 사용하여 질의어에 가중치를 할당한다. 이 질의를 사용하여 색인된 문서들과의 유사도를 비교하여 1 차 랭킹을 수행한 후, 질의와 가장 높은 유사도를 갖는 문서에 대해 자동 적합성 피드백을 1 회 수행하여 질의를 확장, 검색성능을 향상시켰다. HANTEC 2.0 에 대하여 평가한 결과는 벡터공간 모델에 비해 확률기반의 2- Poisson 모델이 월등히 더 나은 결과를 보였다고 하였다[5].

3. 질의문 길이 가중치와 질의어 빈도 가중치의 적용

본 논문에서는 벡터 스페이스 모델을 이용한 유사도 값과 이 유사도 값에서 사용하는 용어 가중치만으로는 사용자가 원하는 결과를 보여주는데 부족하다고 보고 여기에 질의문 길이에 대한 가중치와 정답이라고 판단된 문서에 질의문을 분석하여 얻은 질의어의 출현 빈도에 대한 가중치를 적용한 시스템을 구현하였다. 구현한 시스템의 구성도는 그림 1 과 같다.

3.1 가중치 기법

3.1.1 용어 가중치 기법 (TW)

용어 가중치 계산은 형태소 분석기[1]에 그 기능이 포함되어 있는데, 용어 빈도와 관련하여 자주 출현하는 용어들의 분포를 이용하고, 그 이외에 복합 명사 분해, 품사 유형 및 어절 위치 등을 고려하여 문서 내에서 용어의 중요도를 계산한다[2]. 본 논문에서는 Index Table 구성과 질의 문장의 분석에 이용하였다.

3.1.2 문장의 길이에 대한 가중치 (SLW)

문장의 길이에 대한 가중치는 일반적으로 사용자가 질의를 하게 되면 보통은 1~2 개, 많게는 3 개의 질의어를 이용하여 질의를 많이 한다[1]. 그러므로 질의를 길게 할수록 사용자가 원하는 정보를 더욱 구체적으로 알 수 있다고 보고 질의문을 분석하여 얻은 질의어들의 개수에 대해서 가중치를 부여하였다. 즉 질의문이 들어오면 질의 분석 시스템을 통하여 질의어의 개수를 구하고 구한 값을 검색시스템에서 유사도 계산시 가중치 값으로서 이용하였다.

3.1.3 질의어 출현 빈도 가중치 (QFW)

질의어 출현 빈도 가중치는 시스템이 정답이라고 판단한 문서에서 사용자가 입력한 질의문을 분석하여 얻은 질의 용어가 다수 출현할 경우 사용자가 요구하는 문서와 많은 관련이 있다고 보고 다수의 질의어가 출현하는 문서일수록 더 높은 가중치를 부여하였다. 이 가중치 역시 질의 분석 시스템을 통하여 얻은 정보로 검색 시스템에서 유사도 계산시 가중치 값으로 이용하였다.

[1] http://nlp.kookmin.ac.kr

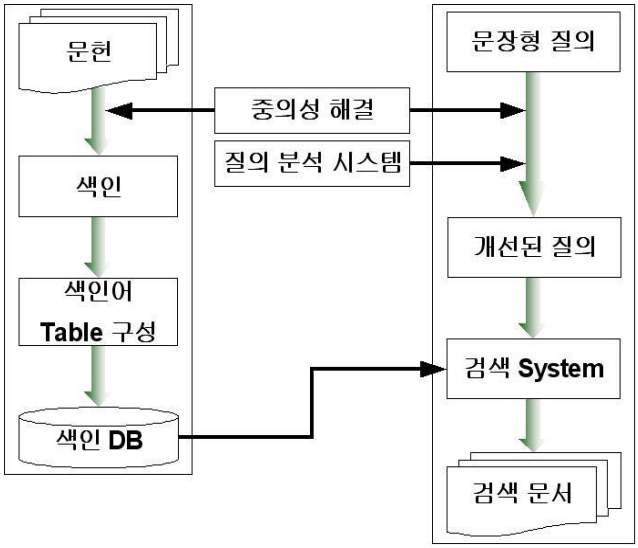


그림 1. 시스템 구성도

4. 실험 및 분석

4.1 실험 환경

본 논문에서 제안한 시스템의 성능을 테스트하기 위하여 KTSET 의 4352 개의 문서를 이용하였다. 테스트 질의는 문장형 질의(자연어 질의)에 중점을 두고 KTSET 에 포함된 1~50 번까지 총 50 개의 자연어 질의를 사용하였고 다음과 같은 4 가지의 경우로 나누어 실험을 하였다.

E1. 벡터 모델을 이용한 단순 유사도 계산
- 유사도 값을 (0.001~0.3)사이로 임계치 변경

E2. 질의문의 길이에 대한 가중치를 적용한 경우
- [(질의어 개수 / 2) * (0.1 ~ 0.9)]

E3. 질의어 출현 빈도 가중치를 적용한 경우
- [질의어 출현 빈도 * (0.1 ~ 0.9)]

E4. 질의문의 길이에 대한 가중치와 질의어 출현 빈도 가중치를 같이 적용한 경우
- [SLW (0.1 ~ 0.9) * QFW (0.1 ~ 0.9)]

2, 3, 4 번 실험 설정
- 유사도 값을 계산할 때 분자값 계산시 각 공식을 가중치로 적용
- 정답 문서 판단시 유사도 임계치 설정: 0.001

4.2 평가 방법

실험 결과에 대한 평가 방법은 정보 검색 평가 기법으로 널리 쓰이는 정확률(Precision)과 재현율(Recall)을 사용하였다. 정확률은 검색된 문서들이 얼마나 적합한가 또는 시스템이 부적합 문서를 검색해 내지 않는 능력을 뜻하고 재현율은 적합 문서가 얼마나 많이 검색되는가 또는 시스템이 적합 문서를 검색해 내는 능력을 뜻한다

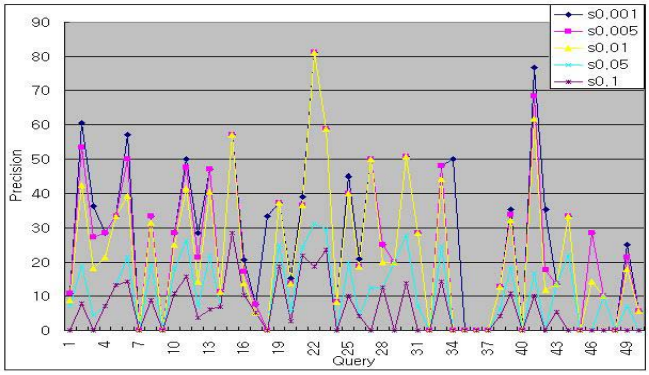
$$\text{정확률} = \frac{\text{시스템이 정답으로 판단한 문서의 수}}{\text{총 정답 문서의 수}} \times 100$$

$$\text{재현율} = \frac{\text{시스템이 정답으로 판단한 문서의 수}}{\text{시스템이 정답으로 판단한 총 문서의 수}} \times 100$$

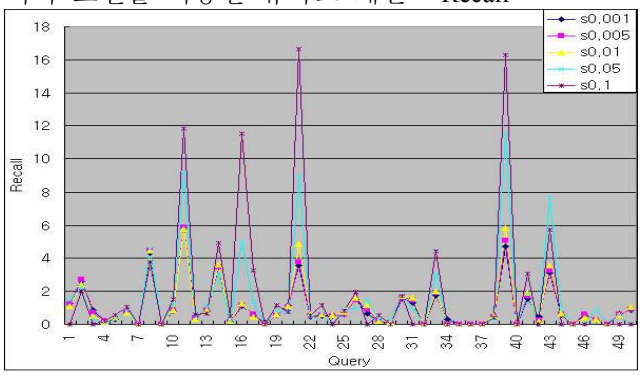
4.3 실험 및 결과

결과 그래프는 알아보기 쉽도록 간단하게 5 가지의 경우들만 표시하였다.

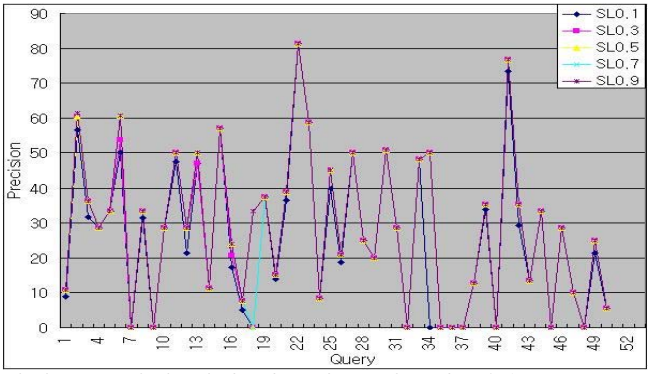
벡터 모델을 이용한 유사도 계산 – Precision



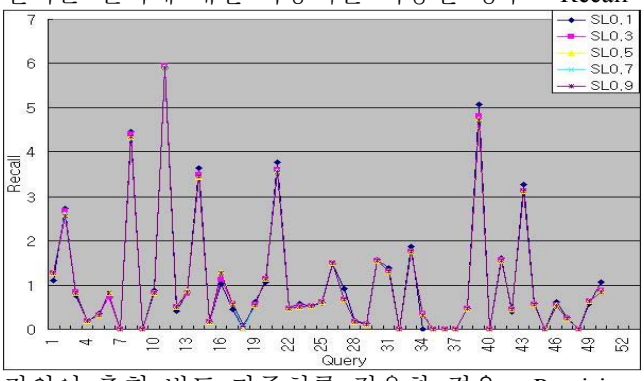
벡터 모델을 이용한 유사도 계산 – Recall



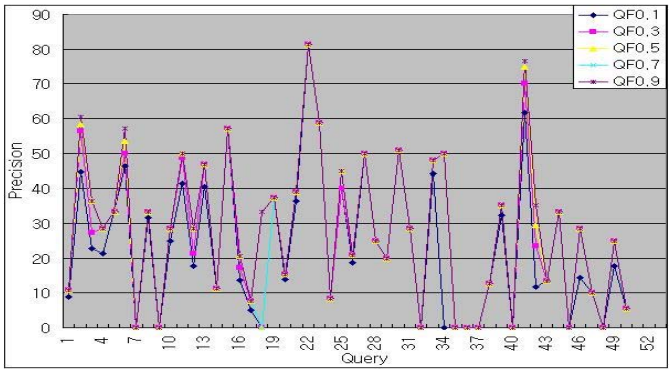
질의문 길이에 대한 가중치를 적용한 경우 – Precision



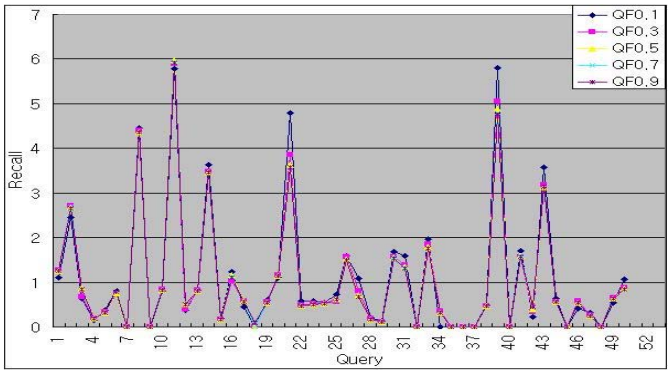
질의문 길이에 대한 가중치를 적용한 경우 – Recall



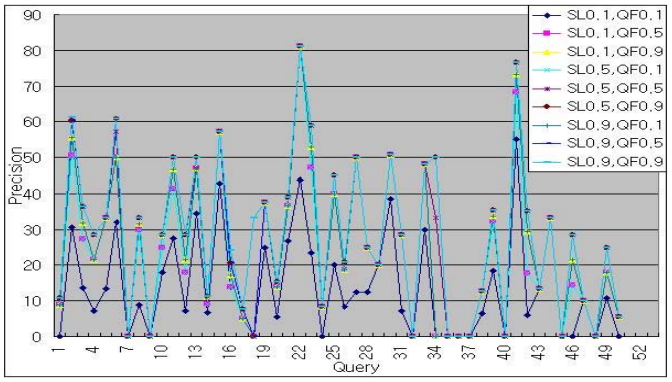
질의어 출현 빈도 가중치를 적용한 경우 – Precision



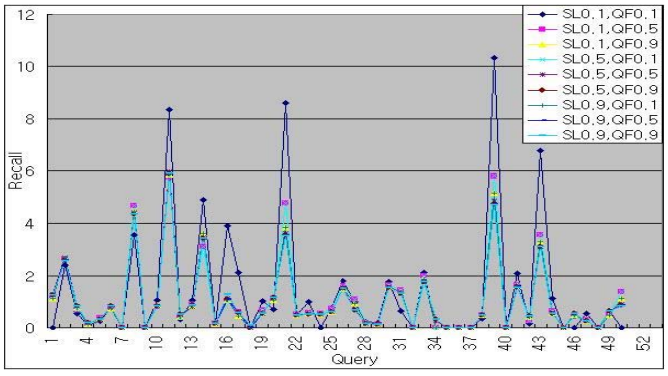
질의어 출현 빈도 가중치를 적용한 경우 - Recall



질의문 길이에 대한 가중치와 질의어 출현 빈도 가중치를 같이 적용한 경우 – Precision



질의문 길이에 대한 가중치와 질의어 출현 빈도 가중치를 같이 적용한 경우 – Recall



4.4 실험 결과 분석

E1 의 실험 결과를 살펴 보면 임계치로 0.001 이상의 유사도 값을 설정했을 경우 성능이 하락하고 0.001

이하에서는 동일한 성능을 보였다. 즉 유사도 값이 낮아질수록 오히려 성능이 높아졌고 반대로 높은 유사도 값을 임계치로 설정할수록 성능이 낮아졌다. 이 실험을 통하여 가장 성능이 좋은 유사도 0.001 을 다음 실험들에서 임계치로 쓰도록 할 것이다.

E2, E3, E4 실험에서는 유사도의 분자 값을 계산할 때 각각의 공식을 이용하여 각 특성에 0.1~0.9 사이로 값을 변화시키면서 가중치를 부여하였고 유사도 임계치는 E1 의 실험에서 보듯이 동일한 가중치 값을 적용할 경우 유사도 임계치를 0.001 로 설정할 때 가장 성능이 좋았기 때문에 0.001 로 설정하였다.

E2 의 실험 결과를 살펴 보면 SLW 의 값을 높여갈수록 정확률이 증가하면서 특이하게도 재현율도 증가하였다. 하지만 정확률이 같을 경우에는 재현율이 하락하는 것을 볼 수 있었다. 유사도로만 판단한 경우와 비교해보면 1~4%정도 정확률이 증가하였고 재현율 역시 같이 증가하였다. 유사도로만 판단하였을 때보다 SLW 값을 적용하였을 때 조금이지만 성능이 향상된 것으로 보아 질의문의 길이가 성능에 영향을 어느 정도 미치는 것을 알 수 있었다. 이 실험에서 최적의 값은 0.8 일 때 정확률이 가장 좋으면서 같은 정확률의 0.9 일 때보다 재현율이 더 높기 때문에 0.8 을 최적 값으로 판단하였다.

E3 의 실험 결과를 살펴 보면 QFW 값을 높여갈수록 정확률이 1~3%씩 증가하였고 역시 특이하게도 0.3 까지는 정확률이 증가할수록 재현율 또한 증가하는 것을 볼 수 있었다. 0.7 미만에서는 가중치를 적용하였을 때가 적용하지 않았을 때보다 낮은 성능을 보였다. 그러나 0.7 이후부터는 정확률은 동일하지만 재현율이 증가하였다. 즉, 유사도로만 판단하였을 때보다 QFW 값을 적용하였을 때 성능이 향상된 것으로 보아 질의어의 출현 빈도 또한 성능에 영향을 미친다는 것을 알 수 있었다. 최적 값으로는 2 번의 실험과 마찬가지로 0.8 을 최적 값으로 판단할 수 있다.

E4 의 실험 결과를 살펴 보면 SLW 나 QFW 의 두 값 중에 하나라도 낮은 값인 경우 상당히 낮은 성능을 보였으나 두 값을 높일수록 점차 성능이 높아짐을 보였다. 앞에서 실험한 실험들과 비교해 보면 SLW 와 QFW 를 같이 적용하였을 때 가장 높은 성능을 보이는 것을 알 수 있었다.

전체 실험결과에서 정확률과 재현율 부분이 전부 0 이 나오는 7, 9, 32, 35, 36, 37, 40, 45, 48 번 질의문의 경우 정답 문서 자체가 3, 4, 3, 5, 6, 10, 19, 5, 7 개로 적고 질의문에서 추출된 질의어의 개수도 2-6 개로 대부분 적은 경우였다. 34 번 질의문의 경우 역시 어느 정도 값이 나오긴 했지만 0 이 나오는 경우가 많았는데 이 질의문 역시 정답 문서가 6 개에 추출된 질의어가 3 개로 적은 경우였다. 즉 질의문에서 추출되는 질의어의 개수가 많고 정답 문서가 많아질수록 높은 성능을 보였다.

5. 결론 및 향후 연구

본 논문에서는 정보 검색 시스템의 더 나은 성능을 위해 벡터 모델을 이용하여 유사도를 계산하고 이 유사도 계산시에 질의문의 길이에 대한 가중치와 질의어가 출현한 빈도에 대한 가중치를 적용한 방법을 제안하였다.

위의 실험에서 볼 수 있듯이 질의어 출현 빈도 가중치를 적용하거나 질의문의 길이에 대한 가중치를 적용할 경우, 또는 두 가중치를 같이 적용할 경우 벡터 모델에 의한 단순한 유사도로만 판단할 때보다 더 높은 성능을 보였다. 이를 통해 단순히 벡터 모델만 이용하는 것보다는 질의문에 대한 분석을 통하여 좀 더 성능을 높일 수 있다는 것과 질의문이 짧을 때보다는 질의문의 길이가 길 때 더 나은 검색 결과를 얻을 수 있다는 것을 알 수 있었다.

향후 연구로는 KTSET 의 1~1000 번까지의 문서와 1001~4414 번까지 문서의 길이가 다름으로 해서 유사도 계산에 대한 득실이 있을 수 있으므로 문서의 길이에 대하여 정규화(Normalization)가 필요하다. 또한, 질의어 출현 빈도 가중치 적용 시에 그 질의어의 문서 내에서의 IDF 값을 고려하지 않았다. 일반적으로 표현되는 동일 자질인데도 단지 빈도가 높게 되어 가중치 값이 높아지는 경우는 값을 낮춰줄 필요가 있다. 예를 들면 “전자파장해에 관한 문서”라는 질의문에서 “문서”와 같은 질의어의 경우는 질의문에서는 높은 가중치를 갖는 단어지만 한편으로는 많은 문서에서 출현하는 단어이기 때문에 이에 대한 조정 값으로 IDF 값을 이용한다면 더욱 정확해진 가중치 값으로 인하여 더 좋은 성능을 나타낼 것이라 본다.

마지막으로 현재는 테스트 문서를 4 천여개의 KTSET 으로 하였으나 객관성을 위해 문서수가 12 만개인 HANTEC2.0 문서로 교체하여 테스트 할 예정이다.

참고문헌

[1] Jansen, B.J. Spink, A., & Saracevic, T. 2000. "Real life, real users, and real needs: a study and analysis of user queries on the Web". Information Processing & Management, 36(2): 207-227.

[2] Salton, G. and C. Buckley, “The Term-Weighting Approaches in Automatic Text Retrieval”, Information Processing and Management, vol. 24, no. 5, pp.513-523, 1988.

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

[4] 정영미, 이재윤, “질의확장 검색에서의 추가용어 가중치 최적화”, 제 9 회 한국정보관리학회 학술대회 논문집, 241-246, 2002.

[5] 김상범, 한경수, 이도길, 임재수, 고명숙, 임해창, “고려대학교 정보검색엔진 KUIR 의 구조 및 특징”, 제 5 회 한국 과학기술 정보인프라 워크샵 학술발표 논문집, pp.164-174, 2000.