질의확장 검색에서의 추가용어 가중치 최적화

Optimizing the Weight of Added Terms in Query Expansion

정영미, 이재윤, 연세대학교 문헌정보학과

Young-Mee Chung, Jae-Yun Lee, Yonsei University

전역적 질의확장 검색에서 단어간 공기기반 유사도를 사용할 경우에는 질의에 추가되는 용어에 부여하는 탐색가중치로 질의와의 유사도를 사용하는 것이 일반적이다. 그러나 과연 유사도가 탐색가중치로 최적인가는 의문의 여지가 있다. 추가용어와 질의 사이의 유사도가 가지는 특성을 살펴보고 고정가중치를 부여한 경우와 비교해보았다. 또한 실험집단이나 확장범위의 영향을 덜 받는 최적화된 추가용어 가중치를 찾기 위해 여러 가지 탐색가중치 공식을 실험하였다.

1. 서론

인터넷에 익숙한 최종 이용자에 의한 정보검색 작업이 보편화되면서 비전문가인 일반 이용자를 지원할 수 있는 지능적이고 효과적인 정보검색 기법에 대한 필요성이 더욱 커지고 있다.

전문가가 길고 상세한 탐색문을 작성하는 데 반해서 일반 이용자들은 온라인 데이터베이스를 탐색할 때에도 10개 이내의 질의어를 사용하는 경향이 있으며, 심지어 웹 검색엔진에서는 대부분의 이용자가 2개 이하의 질의어를 사용하고 있는 것으로 분석되어 있다(Jansen, Spink, & Saracevic 2000). 또한 일반 이용자들은 이와 같이 적은 수의 질의어를 사용하면서도 검색 결과를 보고 질의수정을 거의 하지 않는 것으로 나타났다(Fenichel 1981).

정보요구가 구체적으로 표현되지 않은 짧은 탐색문을 사용하는 검색 결과는 상대적으로 나쁠 수밖에 없다. 이런 문제를 해결하기 위해서 이용자의 초기 질의와 관련된 용어를 새로운 질의에 추가하는 자동 질의확장(automatic query expansion)에 대한 연구가 1970년대에 시작되었으며, 최근 들어 앞에서 언급한 필요성으로 인해 더욱 활발하게 연구되고 있다.

새로 작성되는 질의에 추가될 용어를 어떻게 획득하는가에 따라 질의확장 기법은 두 가지로 구분된다. 즉, 초기 질의에서 검색된 문헌들을 이용하는 지역적(local) 또는 질의 기반(query specific) 질의확장과 전체 문헌집단을 이용하는 전역적(global) 또는 말뭉치 기반(corpus specific) 질의확장이 있다.

전역적 질의확장은 다시 단어간의 공기빈도 기반 방식과 문맥벡터 기반 방식으로 나눌 수가 있다. 공기빈도 기반 전역적 질의확장에서는 초기 질의에 포함된 질의어들과의 유사도 평균이 높은 용어를 추가하게 된다. 이는 개별 질의어와의 유사도보다는 질의 전체와의 유사도를 반영하기 위한 것이다.

질의와의 유사도에 따라서 선정된 추가용어는 확장하기 전 원래 질의에 포함된 질의어와 비교해서 어느 정도로 신뢰할 수 있는가를 반영하는 탐색가중치를 가지게 된다.

기존 연구에서는 주로 원 질의와의 유사도를 추가 질의어의 가중치로 사용하였다(Mandala et al. 1998). 추가되는 질의어에 별도의 가중치를 주지 않고 초기 질의어와 추가 질의어를 대등하게 취급할 경

우에는 오히려 성능이 저하되는 것으로 보고되었다 (Kim & Choi 1999).

본 연구에서는 공기빈도 기반 전역적 질의확장에서 추가용어를 초기 질의어와 대등하게 취급하는 것이 과연 나쁜 결과를 가져오는지를 알아보고, 유사도를 탐색가중치로 사용하는 것이 최선인가를 검증함은 물론, 두 방법보다 더 나은 대안을 모색해보았다.

2. 실험 환경 및 실험집단의 특성

실험에서 사용한 검색시스템은 백터공간모형을 사용하였으며, 용어가중치는 질의어에는 로그TF 가중치를 적용하고 문헌의 용어에는 로그TF · IDF 가중치를 적용하였다. 용어간의 유사도 계산은 상호정보량(MI) 공식을 적용하였다. 상호정보량의 범위를 0에서 1 사이로 정규화하기 위해서 최대정보량으로 나눈 값을 사용하였고 0 이하의 음수값은 부정적 관계로 보아 0으로 처리하였다.

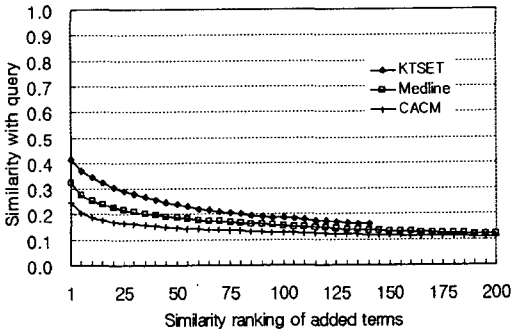
이 연구에서 사용한 대상 실험집단의 특성은 표 1과 같다.

<표 1> 실험집단의 특성

	KTSET	Medline	CACM
언어	한국어	영어	영어
분야	전산학, 정보학	의학	전산학
성격	초록	초록	초록
문헌 수	1000	1033	3204
질의문 수			
(전체)	30	30	45
(질의어 5개 이하)	29	8	16
(질의어 6개 이상)	1	22	29
질의어 수			
(질의문 당 평균)	3.0	8.5	8.6
(최대)	6	19	22
(최소)	2	2	2

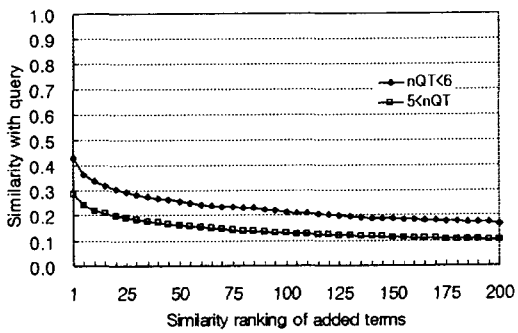
실제로 질의에 추가되는 용어가 가지는 유사도가 어떤 수준의 값을 가지는지를 알아보기 위해서 각 실험집단별로 질의에 용어를 200개까지 추가할 때의 유사도 수준을 그림 1에 나타내었다. KTSET은 일부 질의에서 추가용어가 200개에 못미치는 경우가 있어서 140개까지만 확장하였다.

KTSET의 30개 질의에 대해서 최근접 용어의 유사도는 평균 0.41이었으며, Medline의 30개 질의에서는 평균 0.32, CACM의 45개 질의에서는 평균 0.25였다. 이와 같이 유사도 수준이 차이가 나는 것은 실험집단의 규모와 질의당 질의어 수, 질의어의 문헌빈도 수준과 상관이 있는 것으로 생각된다. 질의당 질의어 수가 매우 적은 KTSET에서 최근접 용어의 유사도 평균이 가장 높게 나타났다. Medline의 경우에는 질의어의 문헌빈도 수준이 CACM보다 상대적으로 낮기 때문에 상호정보량의 수준이 CACM에서보다 높게 나타났다고 볼 수 있다.

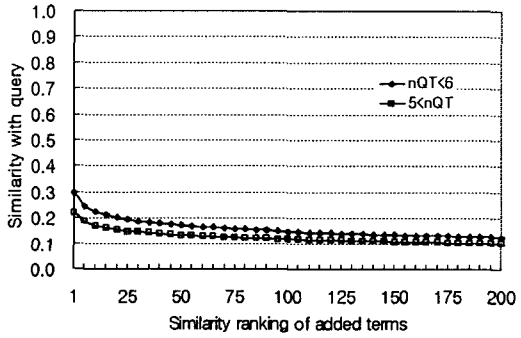


<그림 1> 추가용어의 유사도 변화 - 실험집단별

질의당 질의어의 수가 유사도 수준에 큰 영향을 끼친다는 것을 각 실험집단 내에서 질의어의 수에 따라서 질의를 구분해서 유사도를 확인해봄으로써 알 수가 있다. Medline과 CACM에서 질의어 5개 이하인 질의와 나머지 질의를 구분해보면 그림 2, 그림 3과 같이 질의어 5개 이하인 질의에서 추가용어의 유사도가 높게 나타난다.



<그림 2> 추가용어의 유사도 변화 - Medline



<그림 3> 추가용어의 유사도 변화 - CACM

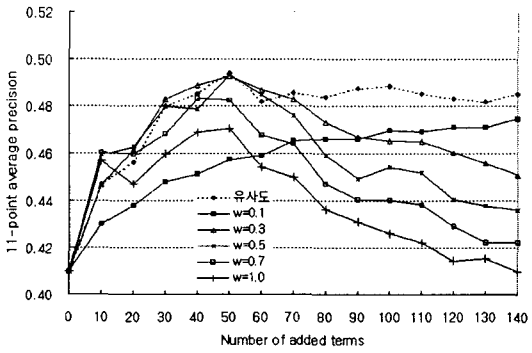
3. 고정 가중치를 적용한 질의확장 검색 실험

질의확장 검색에서 원 질의어의 가중치를 1이라고 하였을 때, 기존 연구에서는 추가되는 용어의 가중치는 고정값 1로 하거나 유사도로 설정하였다. 여기서는 고정 가중치를 0.1에서 1까지 0.1씩 증가시키면서 설정한 실험 결과를 유사도를 가중치로 사용한 경우와 비교해보았다. 그림에서는 공간 문제로 0.1, 0.3, 0.5, 0.7, 1.0으로 가중치를 설정한 경우만 나타내었다.

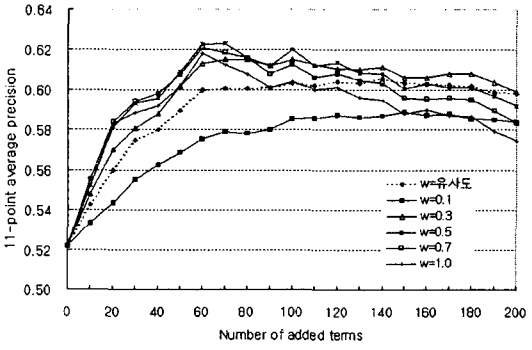
그림 4, 5, 6에 나타난 실험 결과를 보면 추가용어의 고정 가중치가 0.1에 가깝게 낮을 경우에는 확장 초기에 낮은 성능을 보이다가 점차 유사도에 가깝게 향상된다. 반대로 고정 가중치가 1.0에 가까우면 확장 초기에 매우 높은 성능을 보이다가 일정 시점을 넘기면 급격하게 성능이 저하된다.

그러나 가중치를 1.0으로 하였을 때, 즉 초기 질의어와 추가용어를 대등하게 간주하였을 때에도 일정 수준까지는 성능이 향상됨은 물론이고 유사도를 사용한 경우보다 높은 성능을 보이기도 하는 것으로 나타났다. 전체적으로는 고정 가중치를 0.5 전후의 값으로 설정하는 것이 바람직한 것으로 보인다.

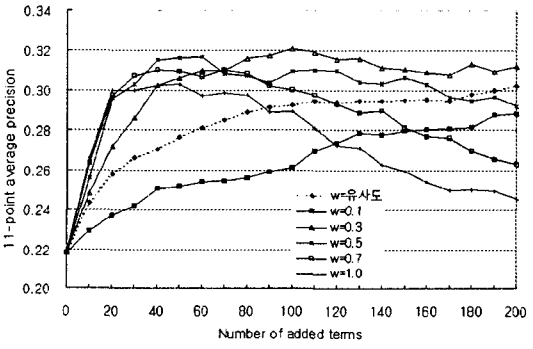
실험집단 별로는 KTSET에서는 50개까지 확장하였을 때 고정 가중치를 적용한 성능과 유사도를 가중치로 사용한 경우가 비슷하며, 나머지 두 실험집단에서는 0.1처럼 지나치게 낮은 값만 아니면 고정 가중치를 적용한 성능이 뚜렷하게 우세하게 나타났다.



<그림 4> 탐색가중치 고정 실험 - KTSET

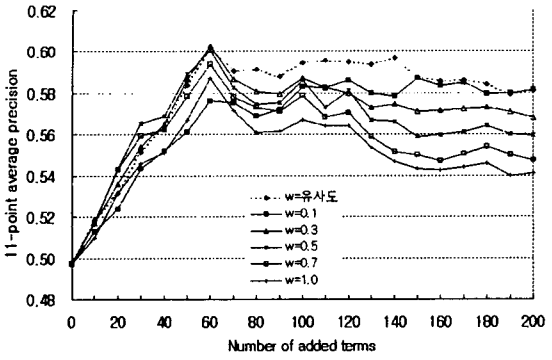


<그림 5> 탐색가중치 고정 실험 - Medline

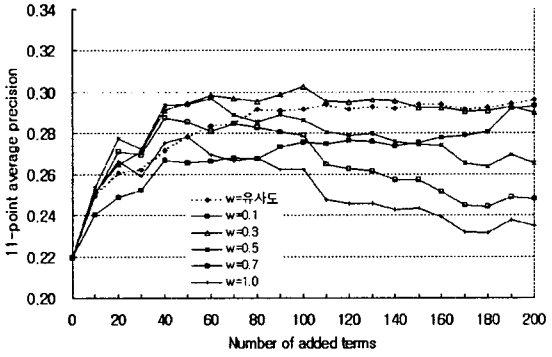


<그림 6> 탐색가중치 고정 실험 - CACM

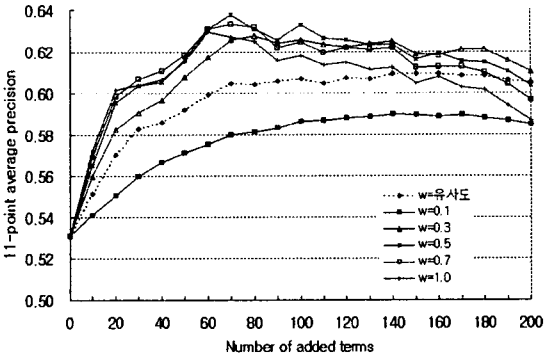
KTSET에서 고정 가중치를 적용한 성능이 유사도를 가중치로 사용한 경우에 비해서 우세하지 못한 것은 질의어의 수 때문인 것으로 짐작된다. 이를 확인하기 위해서 Medline과 CACM에서 질의어가 5개 이하인 질의를 별도로 구분하여 성능을 평가해본 결과 그림 7, 8과 같이 KTSET과 마찬가지로 두 방법 사이의 성능 차이가 별로 없는 것으로 나타났다. 반



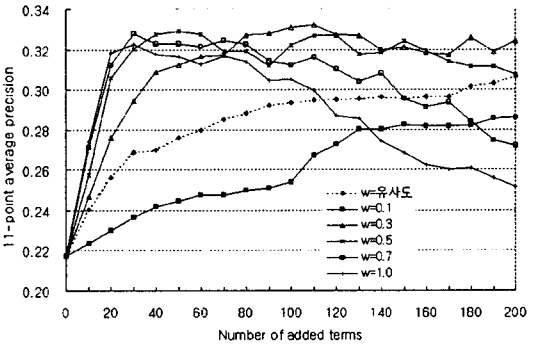
<그림 7> 탐색가중치 고정 실험 - Medline
(질의어 5개 이하 8건 평균)



<그림 8> 탐색가중치 고정 실험 - CACM
(질의어 5개 이하 16건 평균)



<그림 9> 탐색가중치 고정 실험 - Medline
(질의어 6개 이상 22건 평균)



<그림 10> 탐색가중치 고정 실험 - CACM
(질의어 6개 이상 29건 평균)

대로 질의어 6개 이상인 질의들의 성능은 그림 9, 10에서처럼 유사도를 사용한 것에 비해서 고정 가중치를 적용한 경우가 월등하게 나타났다.

질의어 수가 많을 수록 유사도를 가중치로 쓰는 것이 불리한 이유는 질의어가 많을 경우에는 추가되는 용어를 더 신뢰할 수 있음에도 불구하고 2장에서 살펴본 것처럼 오히려 더 낮은 평균 유사도를 가지게 되기 때문이다. 고정 가중치 중에서도 매우 낮은 0.1의 경우에는 성능이 낮은 것을 확인할 수가 있다.

4. 가중치 공식을 적용한 질의확장 검색 실험

적정 수준의 고정 가중치를 사용한 질의확장은 유사도를 가중치로 사용한 경우에 비해서 검색성능 및 효율 면에서 오히려 우월하다는 것을 확인할 수 있었다. 그러나 고정 가중치가 높을 수록 성능의 고점은 확장을 적게 한 부분에서 나타나며 그 지점을 전후로 하여 성능이 급하게 저하된다는 단점이 있다. 즉, 최고 성능을 보이는 지점을 미리 알 수가 없으므로 질의확장 범위를 임의로 정했을 때 최고에 가까운 성능을 얻을 가능성이 그다지 높지 않다는 것이다.

이런 문제를 해결하기 위해서는 질의에 추가되는 용어 중에서 상위 순위의 용어에는 높은 값을, 하위 순위의 용어에는 낮은 값을 가중치로 부여하는 것이 바람직하다. 사실 이것은 유사도를 가중치로 사용하였을 때 지켜지는 원칙이기도 하다. 그러나 앞에서 살펴본 것처럼 유사도는 실험집단의 특성이나 실험질의의 특성에 좌우되어 바람직한 탐색가중치에 비해서 지나치게 낮은 값을 가지는 경향이 있다. 따라

서 추가용어 t가 질의와의 유사도 순위 r에 따라서 감소되는 가중치 w를 가지도록 인위적인 공식을 아래와 같이 몇 가지 고안하였다.

[탐색가중치 공식 a] $w_a(t_r) = \frac{Sim_{qt}(t_r)}{Sim_{qt}(t_1)}$

[탐색가중치 공식 b] $w_b(t_r) = \frac{1}{r^{0.25}}$

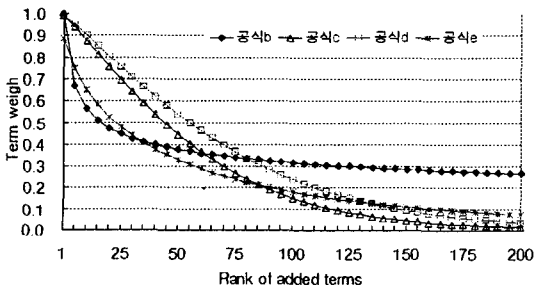
[탐색가중치 공식 c] $w_c(t_r) = \frac{2}{1+e^{0.025r}}$

[탐색가중치 공식 d] $w_d(t_r) = \frac{2}{1+e^{0.02r}}$

[탐색가중치 공식 e] $w_e(t_r) = \frac{2}{1+10^{0.1\times\sqrt{r}}}$

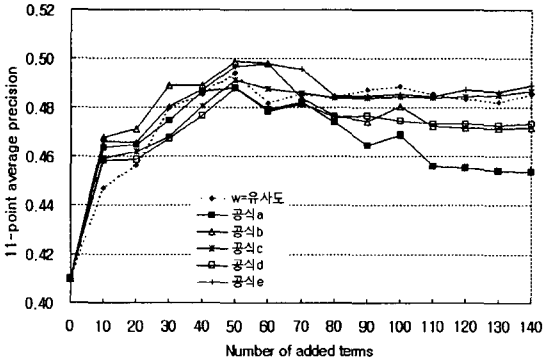
공식 a는 최상위 유사어의 가중치를 1로 하고 나머지 용어의 가중치는 r번째 용어가 질의와 가지는 유사도($Sim_{qt}(t_r)$)와 최상위 유사어의 질의와의 유사도($Sim_{qt}(t_1)$) 사이의 비율만큼 반영하는 것이다. 나머지 공식들은 최근접 용어의 가중치를 1로 하고 나머지 용어의 가중치는 순위에 따라서 조금씩 감소하게 되어있다. 단, 이때 적용하는 순위는 유사도가 같은 경우에도 동순위로 취급하지 않고 질의에 추가되는 순서대로 임의로 순위를 구분하였다.

유사도 순위에 대해서 공식을 적용하여 얻어지는 탐색가중치의 감소 추세는 그림 11과 같다. 공식b와 e는 공식c와 d에 비해서 상위 유사어의 가중치를 낮게 설정하는 효과가 있으며, 공식b는 다른 공식에 비해서 하위 유사어에도 상대적으로 높은 가중치를 부여하는 특징이 있다.

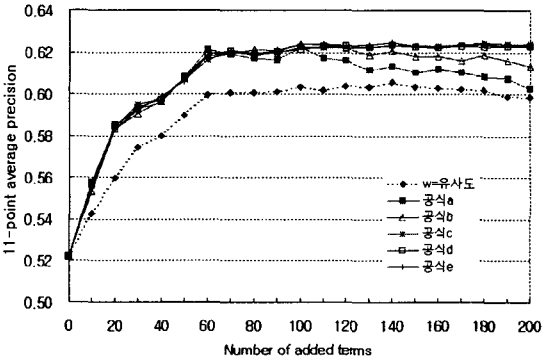


<그림 11> 공식에 따른 추가용어 가중치의 변화

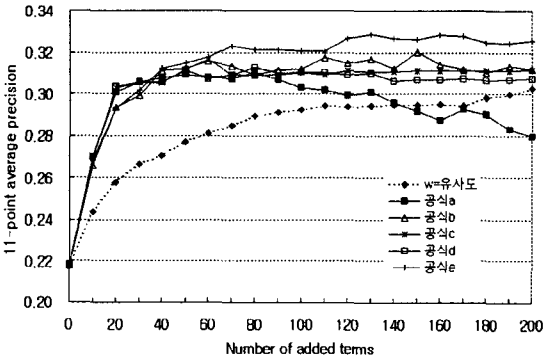
탐색가중치 공식을 적용한 질의확장 검색 실험결과는 그림 12, 13, 14와 같다. 공식 간의 차이는 크지 않으나 공식e의 경우에 상대적으로 실험집단이나 확장범위에 영향을 덜 받아 안정적인 것으로 나타났다.



<그림 12> 탐색가중치 공식의 성능 - KTSET

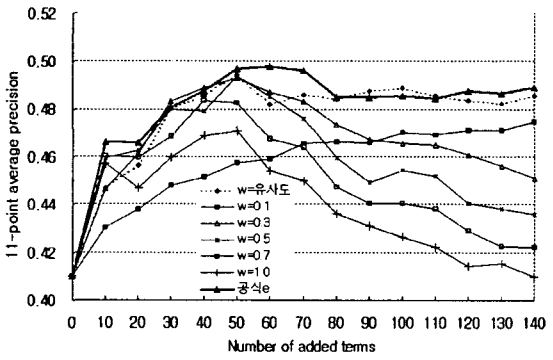


<그림 13> 탐색가중치 공식의 성능 - Medline

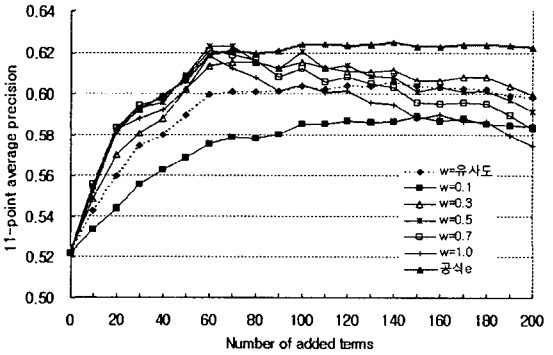


<그림 14> 탐색가중치 공식의 성능 - CACM

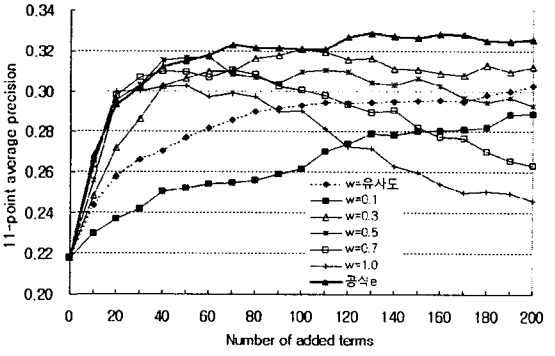
공식e를 이용한 질의확장 검색 성능을 고정가중치와 비교해보면 그림 15, 16, 17과 같이 실험집단이나 확장범위를 불문하고 거의 최고 성능을 얻게 되는 것으로 나타났다.



<그림 15> 고정가중치와 가중치 공식의 성능 비교 - KTSET



<그림 16> 고정가중치와 가중치 공식의 성능 비교 - Medline



<그림 17> 고정가중치와 가중치 공식의 성능 비교 - CACM

4. 결론

공기기반 전역적 질의확장에서 추가용어에 부여하는 탐색가중치로 질의와의 유사도를 사용하는 것이 최선의 선택이 아님을 알 수가 있었다. 0.5 내외의 고정가중치나 가중치 공식을 이용하는 것이 특히 다소 긴 질의문에서 좋은 성능을 보이는 것으로 나타났다.

지나치게 많은 용어를 질의에 추가하게 되면 시스템의 효율이 저하된다는 점을 감안하면, 추가용어의 범위를 20~30개 정도로 한정할 경우에는 초기 질의어와 추가용어를 구분하지 않고 탐색가중치로 1.0을 사용해도 좋은 성능을 얻을 수 있을 것이다.

참고문헌

Fenichel, C.H. 1981. "Online searching: measures that discriminate among users with different types of experiences." *Journal of the American Society for Information Science*, 32(1): 23-32.

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.

Kim, M.C., & Choi, K.S. 1999. "A comparison of collocation-based similarity measures in query expansion." *Information Processing & Management*, 35(1): 19-30.

Mandala, R., Tokunaga, T., Tanaka, H. 1998. "Query expansion using heterogeneous thesauri." *Information Processing & Management*, 36(3): 361-378.