유사계수에 따른 전역적 질의확장 검색 성능 비교

이재윤
연세대학교 문헌정보학과
memexlee@lis.yonsei.ac.kr

Comparing the Performance of Global Query Expansion according to Similarity Measures

Jae-Yun Lee
Dept. of Library & Information Science, Yonsei University

요 약

공기빈도를 이용한 전역적 질의확장 검색에서 공기유사도를 판정하는데 이용되는 유사계수의 특성에 따른 질의확장 성능을 비교해보았다. 먼저 각 유사계수의 통계적인 특성을 말뭉치와 검색실험 문서집단을 대상으로 살펴본 결과 코사인 계수, 자카드 계수는 고빈도어 선호경향을 보이고 상호정보량과 율의 Y는 저빈도어 선호경향을 보이는 것으로 나타났다. 질의확장 검색실험에서는 고빈도어 선호경향을 가진 유사계수에 비해서 저빈도어 선호경향을 가진 유사계수를 이용할 때 더 좋은 성능이 나타났다. 특히 율의 Y는 질의어의 DF가 1에 가깝게 매우 낮을 때 다른 유사계수와 달리 고빈도어를 선호함으로써 항상 저빈도어를 선호하는 상호정보량에 비해서 질의확장 검색에 유리함을 알 수가 있었다.

1 서론

공기기반 전역적 질의확장 검색은 이용자가 입력한 초기 질의어와 유사한 용어를 문서집단에서 통계적으로 획득하여 자동으로 질의에 추가하는 기법이다. 1969년에 Lesk[1]가 각 질의어별로 공기유사도가 높은 용어를 질의에 추가하는 방식을 제안한 이후, 여러 연구자가 이 기법을 검토하였으나 확실한 성능향상을 얻지 못했다. 그러다가 1993년 Qiu & Frei[2]가 개별 질의어 단위가 아닌 모든 질의어와의 공기유사도 평균을 이용하는 질의개념 확장 방식을 제안하면서 다시 연구가 활발해졌다.

Mandala, Tokunaga, & Tanaka[3]와 Kim & Choi[4]는 공기기반 유사도가 가장 높은 단어를 추가하는 전역적 질의확장을 수행하되 Qiu & Frei의 제안에 따라 질의 전체와의 유사도를 구하는 방식으로 성능을 향상시킬 수가 있었다. 이 두 연구는 각각 다양한 유사계수의 이용 결과를 비교 실험하였는데, 결과가 일치하지 않았다. Mandala, Tokunaga, & Tanaka[3]는 상호정보량의 성능이 자카드 계수, 다이스 계수보다 좋은 것으로 보고하였는데, Kim & Choi[4]에서는 코사인 계수, 자카드 계수의 성능이 상호정보량보다 더 좋았다고 보고한 것이다.

이 연구에서는 이러한 유사계수간 성능의 우열이 각 유사계수의 통계적 특성이 다르기 때문이라는 가정하에 질의확장 검색에 적합한 유사계수의 특성을 검증하고자 하였다.

2 유사계수의 빈도 특성

2.1 말뭉치 분석 결과

Chung & Lee[5]는 공기기반 유사계수에 대한 분석에서 공기하는 단어의 빈도 수준에 따른 유사계수의 특성을 살펴본 바가 있다. 이들은 16,555건의 신문기사에서 출현한 단어를 대상으로 빈도 수준에 따라서 문헌빈도 20인 저빈도어부터 1,820인 고빈도어에 이르기까지 10개의 기준 단어를 선정한 다음, 공기단어의 문헌빈도와 유사도 사이의 상관계수를 구해서 그림 1을 얻었다.

이 그림을 보면 상호정보량(MI)과 율(Yule)의 Y(YUL)는 모든 빈도 구간에서 공기 단어의 빈도와 유사도 값 사이의 상관이 음수이므로 저빈도어일수록 유사도가 높게 나타나는 경향, 즉 저빈도어 선호경향이 있음을 알 수가 있다. 다른 유사계수는 대부분의 구간에서 상관값이 0 이상이므로 고빈도어의 유사도를 높게 판정하는 경향, 즉 고빈도어 선호경향을 나타냈다. 그 중에서도 자카드 계수(JAC)가 고빈도어 선호경향이 가장 강하며 그 다음으로는 코사인 계수(COS), 우도비(LR)와 카이제곱 통계량(CHI)의 순서로 나타났다.

그런데 기준 단어의 빈도가 매우 낮은 지점에서는 모든 유사계수가 상관값이 음수가 되므로, 기준 단어가 저빈도어이면 모두가 저빈도어의 유사도를 높게 판정하는 저빈도어 선호경향을 가진다는 것을 알 수가 있다. 그 중에서도 상호정보량과 율의 Y는 기준 단어가 저빈도어인 경우에 더 심한 저빈도어 선호경향을 보였다.

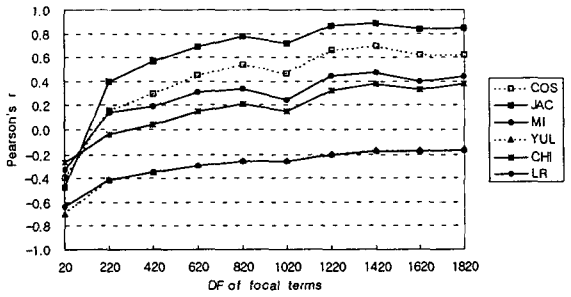


그림 1. 기준 단어의 빈도에 따른 공기단어의 빈도와 유사도 사이의 상관관계 ([5]에서 인용)

2.2 검색실험집단 질의어 분석 결과

앞에서 살펴본 유사계수의 특징이 검색실험집단에서도 나타나는가를 알아보기 위해서 CACM 실험집단(3,204건)과 Medline 실험집단(1,033건)의 질의어를 대상으로 공기기반 유사어의 빈도 수준을 분석해보았다. 이때 질의어가 아닌 용어 중에서 문서빈도(DF)가 1인 용어는 제외하였다. Qiu & Frei[2]

가 지적한 바와 같이 빈도 1인 용어는 질의확장에서 추가질의어로 사용될 경우 성능저하의 원인이 되기 때문이다.

유사계수는 강한 고빈도어 선호경향을 가진 자카드 계수와 코사인 계수, 그리고 강한 저빈도어 선호경향을 가진 상호정보량과 율의 Y를 채택하였다. 공기빈도의 2×2분할표를 이용하여 각각의 공식을 나타내면 아래와 같다.

		단어 y		
		출 현	미출현	합 계
단어 x	출 현	a	b	a+b
	미출현	c	d	c+d
	합 계	a+c	b+d	N

$$Jaccard(x,y) = \frac{a}{a+b+c}$$

$$cosine(x,y) = \frac{a}{\sqrt{(a+b)(a+c)}}$$

$$MI(x,y) = \log_2 \frac{N \times a}{(a+b)(a+c)}$$

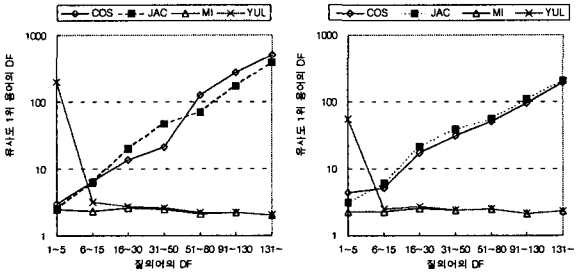
$$Yule's\ Y(x,y) = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$$

CACM의 질의어 중에서 공기 용어가 5개 이하인 질의어 3개를 제외한 289개와 Medline의 질의어 213개를 실험집단에서의 DF에 따라 각각 7개 그룹으로 나눈 다음, 각 그룹에 속한 질의어와의 공기기반 유사도가 가장 높은 용어의 평균 DF를 표 1과 그림 2에 제시하였다.

분석 결과 코사인 계수와 자카드 계수는 질의어가 고빈도어일 때 고빈도어가 유사도 상위에 속하며, 질의어가 저빈도어일 때에는 저빈도어가 높은 유사도를 가진다. 즉, 질의어와 비슷한 수준의 용어가 높은 유사도를 가지는 경향이 있다. 이와 반대로 상호정보량은 질의어의 문서빈도에 상관없이 빈도 2.5 내외의 저빈도어를 유사어로 취하는 것으로 나타난다. 빈도 1인 용어를 제외한 것을 감안하면 사실상 최저빈도어를 유사어로 선호하는 셈이다.

표 1. 질의어 DF에 따른 유사도 1위 용어의 DF 평균

빈도 구간	CACM				Medline			
	COS	JAC	MI	YUL	COS	JAC	MI	YUL
1~5	3.02	2.59	2.50	195.27	4.41	3.15	2.32	55.32
6~15	6.50	6.31	2.36	3.22	5.24	6.13	2.33	2.50
16~30	13.72	20.26	2.65	2.74	17.24	21.47	2.59	2.78
31~50	21.05	47.80	2.54	2.63	30.94	38.69	2.41	2.41
51~80	124.50	70.53	2.18	2.25	50.38	55.42	2.54	2.54
91~130	277.79	172.26	2.23	2.23	96.35	110.35	2.18	2.18
131~	491.44	391.31	2.08	2.08	194.09	206.91	2.36	2.36



(a) CACM (b) Medline

그림 2. 질의어 DF에 따른 유사도 1위 용어의 DF 평균

율의 Y는 질의어의 DF가 6이상일 경우에는 상호정보량과 거의 유사한 경향을 보이지만, 5이하인 경우에는 특이하게도 코사인이나 자카드 계수의 경우보다도 훨씬 고빈도어인 용어를 유사도 1위로 취한다. 이는 질의어의 빈도와 공기빈도가 같을 경우에 율의 Y가 상대 용어의 빈도를 무시하고 항상 최고값 1을 가지는 성질이 있기 때문이다. 상호정보량도 질의어의 빈도와 공기빈도가 같을 경우에 높은 값을 가지긴 하지만, 질의어의 빈도가 낮을수록 최고값은 높아지게 되어서 항상 저빈도어를 선호하는 경향이 유지된다. 이와 같은 이유 때문에 율의 Y는 빈도 5이하인 항목에 대해서는 사용을 삼가하도록 권하고 있기도 하다[6].

결국 검색실험집단에서도 코사인 계수와 자카드 계수의 고빈도어 선호경향과, 상호정보량과 율의 Y의 저빈도어 선호경향은 여전하였다. 다만 질의어의 빈도가 1에 가깝게 매우 낮은 경우에 율의 Y는 오히려 고빈도어 선호경향을 보인다는 점을 추가로 확인하였다.

2.3 검색실험집단 질의별 확장용어 분석 결과

실제 질의확장 검색에서는 Qiu & Frei[2]가 제안한 것처럼 개별 질의어가 아닌 질의에 포함된 전체 질의어와의 유사도 평균을 기준으로 추가 질의어를 선정한다. 따라서 실제로 질의에 추가되는 용어의 DF 수준이 유사계수마다 어떻게 다른지를 분석해보았다. 각 질의에 대해서 100개씩 확장하되 10개씩 구간을 나누어 각 구간에 포함된 확장질의어의 평균 DF를 표 2와 표 3에 제시하였다.

표 2. 확장구간별 추가 질의어의 DF 평균 - CACM

	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
COS	385.2	256.4	179.6	146.9	133.3	131.3	108.8	103.7	100.9	101.7
JAC	352.1	258.6	193.1	159.4	146.7	118.9	120.6	107.0	109.9	101.7
MI	10.1	13.5	12.0	16.4	18.8	17.0	22.1	20.9	21.3	23.7
YUL	18.7	23.3	25.9	25.3	27.6	29.6	24.5	24.7	24.3	28.4

표 3. 확장구간별 추가 질의어의 DF 평균 - Medline

	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
COS	69.2	65.5	64.7	62.0	63.8	55.8	55.5	60.6	59.0	52.5
JAC	59.5	61.5	64.1	62.0	66.2	67.3	59.8	64.7	62.3	57.5
MI	4.7	5.7	6.6	6.7	7.4	8.3	8.5	10.4	9.2	10.9
YUL	5.5	8.2	9.7	11.1	9.5	13.4	10.8	11.7	11.5	14.4

역시 코사인 계수와 자카드 계수에 의해 추가질의어로 선택된 용어가 상호정보량 및 율의 Y에 의한 경우보다 DF가 훨씬 높다는 것을 알 수 있다. 저빈도어 선호 유사계수 중에서는 율의 Y에 의해 추가된 질의어가 상호정보량에 의한 것보다 DF가 더 높았다.

3. 유사계수별 질의확장 검색 실험

CACM 실험집단과 Medline 실험집단을 대상으로 각 질의별로 공기빈도 기반 유사어를 10개부터 100개까지 10개씩 늘이면서 추가하여 확장된 질의로 검색을 수행하였다. 색인어의 가중치 $DW(t_i)$와 질의어의 가중치 $QW(t_i)$, 그리고 문서와 질의간 유사도 sim(D,Q)은 다음 공식으로 산출하였다(lncltc 방식).

$$DW(t_i) = 1 + \log tf(t_i)$$

$$QW(t_i) = (1 + \log tf(t_i)) \times \log_2 \frac{N}{DF(t_i)}$$

$$sim(D,Q) = \frac{\sum DW(t_i) QW(t_i)}{\sqrt{\sum DW(t_i)^2 \times \sum QW(t_i)^2}}$$

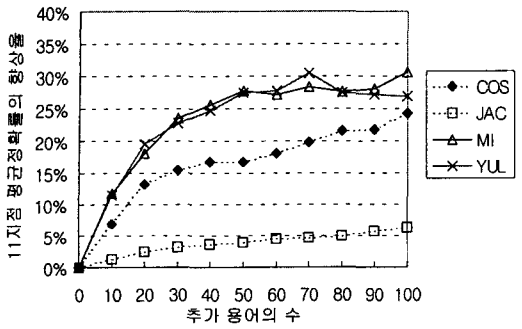


그림 3. 유사계수별 질의확장 검색성능(11-AP) - CACM

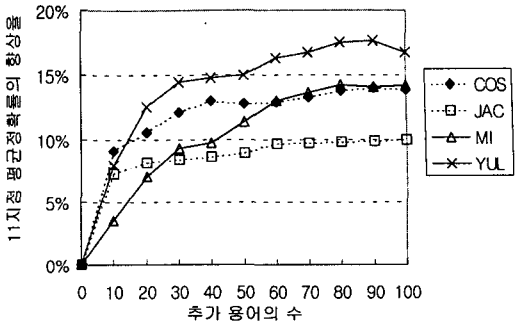


그림 4. 유사계수별 질의확장 검색성능(11-AP) - Medline

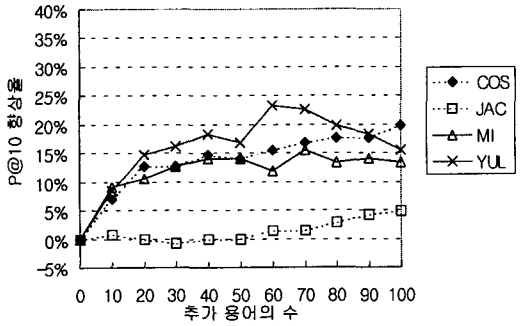


그림 5. 유사계수별 질의확장 검색성능(P@10) - CACM

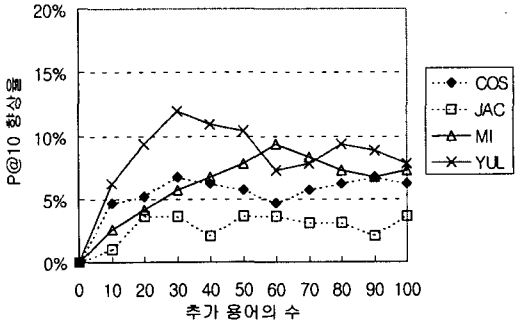


그림 6. 유사계수별 질의확장 검색성능(P@10) - Medline

추가된 질의어의 가중치는 TF가중치 부분을 초기 질의와의 유사도로 하고 IDF 가중치는 그대로 적용하였다. 단, 상호정보량은 유사도의 범위를 1에서 0 사이로 제한하기 위해서 최대값인 $\log_2 N$으로 나누고 음수인 경우는 0으로 처리하였다.

검색 결과를 11지점 평균정확률(11-AP)과 10위내 정확률(P@10) 척도로 분석하였다. 그림 3과 그림 4에 제시한 초기질의의 검색성능 대비 11지점 평균정확률의 향상율을 보면 유사계수간의 성능 우열은 대체적으로 다음과 같다.

* CACM : 율의 Y = 상호정보량 > 코사인 계수 > 자카드 계수
* Medline : 율의 Y > 코사인 계수 ≥ 상호정보량 > 자카드 계수

그림 5와 그림 6에 제시한 10위내 정확률의 향상율로 본 유사계수간의 성능 우열은 다음과 같다.

* CACM : 율의 Y > 코사인 계수 ≥ 상호정보량 > 자카드 계수
* Medline : 율의 Y > 상호정보량 ≥ 코사인 계수 > 자카드 계수

Medline에서 고빈도어 선호 유사계수의 성능이 저빈도어 선호 유사계수보다 크게 나쁘지 않은 이유는, Medline 실험집단의 질의가 주로 저빈도어로 구성되어 있기 때문이다.

저빈도어 선호 유사계수의 성능이 좋기는 하지만 상호정보량의 경우에 저빈도어 선호경향이 지나치게 강하여 오히려 성능이 향상되지 못하는 측면도 있어 보인다. 율의 Y는 저빈도어를 선호하면서도 질의어의 DF가 1이나 2처럼 매우 낮은 경우에는 오히려 고빈도어를 선호하는 것이 상호정보량보다 나은 성능을 가져오는 것으로 짐작된다.

5 결론

유사계수에 따른 공기빈도 기반 질의확장 검색실험에서 고빈도어 선호경향을 가진 유사계수에 비해서 저빈도어 선호경향을 가진 유사계수를 이용할 때 더 좋은 성능이 나타났다. 특히 율의 Y는 질의어의 DF가 1에 가깝게 매우 낮을 때 다른 유사계수와 달리 저빈도어를 선호하지 않음으로써 항상 저빈도어를 선호하는 상호정보량에 비해서 질의확장 검색에 유리하였다.

참고문헌

[1] Lesk, M. E. Word-word associations in document retrieval systems. *American Documentation*, 20: 27-38, 1969.

[2] Qiu, Y., & H. P. Frei. Concept based query expansion. *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 160-169, 1993.

[3] Mandala, R., T. Tokunaga, & H. Tanaka. Query expansion using heterogeneous thesauri. *Information Processing & Management*, 36(3): 361-378, 2000.

[4] Kim, Myoung-Cheol, & Key-Sun Choi. A comparison of collocation-based similarity measures in query expansion. *Information Processing & Management*, 35(1): 19-30, 1999.

[5] Chung, Young-Mee, & Jae-Yun Lee. A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science and Technology*, 52(4): 283-296, 2001.

[6] Delcourt, C. About the statistical analysis of co-occurrence. *Computers and the Humanities*, 26(1): 21-29, 1992.