Proceedings of KFIS 2002 Spring Conference, 2002. 5. 25

P-norm, RF, LCAF를 이용한 정보검색에 관한 연구

A Study on Information Retrieval Using P-norm, RF, LCAF

김영천*, 이재훈*, 박병권**, 이성주*
Young-Cheon Kim, Jae-Hoon Lee,
Byung-Gweun Park, Sung-Joo Lee

조선대학교 전자계산학과*
서강정보대학 정보통신과**
E-mail : yckim@stmail.chosun.ac.kr

요약

정보검색에서 가장 중요하면서도 어려운 문제 중의 하나는 사용자가 원하는 정보를 찾기 위한 효율적인 질의를 작성하는 일이다

순수한 부울 검색 시스템은 문서와 질의 사이의 유사도를 나타내는 문서값을 계산할 수 없기 때문에, 검색된 문서들을 질의를 만족하는 정보에 따라 정렬할 수 없다.

부울 검색 시스템의 이러한 단점을 보완하는 방법으로 P-norm 모델이 개발되었다. 본 논문에서는 높은 검색 효과를 제공하는 지역적 문맥 분석 피드백을 이용한 정보검색 모델을 제안한다. 제안한 지역적 문맥 분석 피드백모델이 적합성 피드백이나 P-norm 모델보다 우수함을 설명하고, 또한 성능 비교를 통하여 이를 입증한다.

Keyword : 질의확장, RF, LCAF, 부울 검색, 유사도

Ⅰ. 서론

정보검색에서 가장 중요하면서도 어려운 문제 중의 하나는 사용자가 원하는 정보를 찾기 위한 효율적인 질의를 작성하는 일이다. 하지만 전체 문서집합의 구성에 대해 미리 알고 있지 않는 한 이상적인 최적의 질의는 작성할 수 없다. 대신 최초에는 시험적 질의(tentative query)로 검색을 수행한 후, 이전의 검색 결과에 대한 평가에 기반하여 다음 번 검색의 질의를 확장한다.

본 논문에서는 지역적 문맥 분석 피드백(Local Context Analysis Feedback) 모델에서 가중치를 재부여하여 질의를 확장하는 모델을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 확장 불리언 모델에서 P-norm 모델에 대하여 기술한다. 3장에서는 높은 검색 효과를 제공하는 적합성 피드백(Relevance Feedback: RF)모델과 지역적 문맥 분석 피드백(Local Context Analysis Feedback: LCAF)에 대하여 기술한다. 4장에서는 P-norm, RF, LCAF 모델의 성능을 비교한다. 마지막으로 5장에서 결론 및 앞으로의 연구 방향을 제시한다.

Ⅱ. 확장 불리언 모델

불리언 검색은 단순하고 강력하지만 용어 가중치를 제공하지 않기 때문에 결과 집합의 순위화가 불가능하고, 출력 크기가 너무 크거나 작아진다. 이런 이유로 최근에는 불리언 질의에 벡터 모델의 특성을 확장한 모델들을 사용한다. P-norm 모델은 유클리디안 거리뿐만 아

니라 질의시 규정되어야 하는 새 파라미터인 $1 \le p \le \infty$인 p-거리에 대해서도 일반화한다. 일반화된 논리합과 논리곱의 질의는 식(1), 식(2)와 같다.

$$q_{or} = k_1 \vee^p k_2 \vee^p \ldots \vee^p k_m \quad (1)$$

$$q_{and} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m \quad (2)$$

각각의 질의-문헌 유사도는 식(3)과 같다.

$$sim(q_{or}, d_j) = \left(\frac{x_1^p + x_2^p + \ldots + x_m^p}{m} \right)^{\frac{1}{p}} \quad (3)$$

$$sim(q_{and}, d_j) = 1 - \left(\frac{(1-x_1)^p + (1-x_2)^p + \ldots + (1-x_m)^p}{m} \right)^{\frac{1}{p}}$$

x_i는 $[k_i, d_j]$쌍의 가중치 $w_{i,d}$를 나타내며 앞에서 정의된 p-norm은 식(4), 식(5)와 같은 몇가지 특성을 가진다.

첫째로, p=1인 경우 식(4)와 같다.

$$sim(q_{or}, d_j) = sim(q_{and}, d_j) = \frac{x_1 + \ldots + x_m}{m} \quad (4)$$

둘째로 $p=\infty$인 경우 식(5)와 같다.

$$sim(q_{or}, d_j) = \max(x_i)$$
$$sim(q_{and}, d_j) = \min(x_i) \quad (5)$$

p=1이면 논리합과 논리곱 질의 모두 벡터 기반의 유사도 공식인 벡터 내적과 같이 용어-문헌 가중치의 합이 된다. 또 $P=\infty$인 경우 질의는 불 논리의 일반화 관점에서 퍼지 논리 형식으로 계산된다. 파라미터 p를 1에서 무한대까지 변화시키면 p-norm 순위화 행위를 벡터 순위화에서 불리언 순위화까지 변화시켜 볼 수 있다.

Ⅲ. RF와 LCAF 정보검색 모델

3.1 RF 모델의 정보검색

적합한 문헌으로 판단된 문헌들의 용어-가중치 벡터는 서로 유사하다는 사실을 이용한다. 또 비적합한 문헌들은 적합한 문헌들이 갖는 용어-가중치 벡터와는 다른 벡터를 갖는다고 가정한다[2][3].

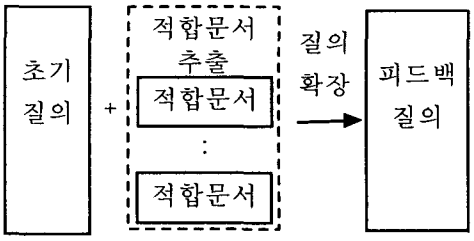

그림 3.1 적합성 피드백

비현실적이기는 하지만 주어진 질의 q에 대한 전체 연관 문헌 집합인 C_r을 이미 알고 있다고 가정하면 연관 문헌들을 비연관 문헌들로부터 구분하는 최적 질의 벡터는 식(6)과 같이 증명된다.

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\forall \vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\forall \vec{d}_j \notin C_r} \vec{d}_j \quad (6)$$

수정된 질의 $\vec{q}_m$을 계산하는 고전적인 방법은 식(7)과 같다.

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j \quad (7)$$

3.2 LCAF 모델의 정보 검색

지역적 문맥 분석 피드백 과정의 정보 검색은 그림 3.2와 같다.

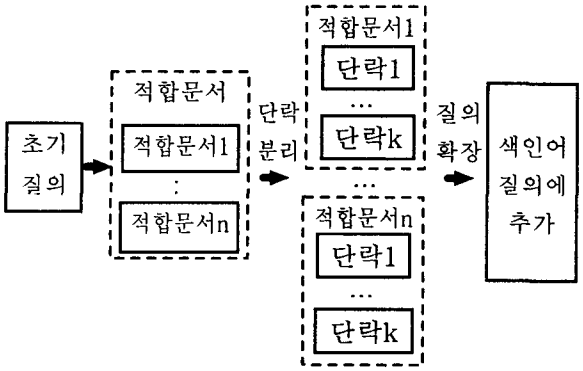

그림 3.2 지역적 문맥 분석 피드백

본 논문에서 제안한 지역적 문맥 분석 피드백 과정은 다음 3단계를 거쳐 수행된다. 첫째, 현재 질의를 사용하여 상위 n 개의 단락을 검색한다. 이 과정은 현재 질의에 의해 초기 검색된 문헌을 일정 길이의 단락으로 분할한 후 단락을 마치 문헌처럼 순위화함으로써 수행된다. 두 번째, 상위 순위 단락에 나타나는 각 개념 c에 대해 tf-idf 방법의 한 변형을 이용하여 해당 개념과 전체 질의와의 유사도 sim(q, c)를 계산한다. 세 번째, m 개의 상위 순위 개념이 원래 질의에 추가된다. 추가된 각 개념에 대해 1-0.9×i/m의 가중치를 부여한다.

이 3단계 중 두 번째 단계가 가장 복잡한데, 각 연관 개념 c와 원래 질의 q 사이의 유사도 sim(q,c)는 식(8)과 같이 계산된다.

$$sim(q,\ c) = \prod_{k_i \in q}\left(\delta + \frac{\log(f(c,k_i)\times idf_c}{\log n}\right)^{idf_i} \quad (8)$$

식(8)에서 n은 상위 순위 단락의 수이다. 함수 $f(c,k_i)$는 개념 c와 질의 용어 k_i사이의 연관도를 나타내며 식(9)와 같이 계산된다.

$$f(c,k_i) = \sum_{j=1}^{n} pf_{i,j} \times pf_{c,j} \quad (9)$$

식(9)에서 $pf_{i,j}$는 j번째 단락 내에서의 용어 k_i의 출현 빈도를 나타내며, $pf_{c,j}$는 j번째 단락 내에서의 개념 c의 출현 빈도를 나타낸다. 역문헌 빈도 인수는 식(10), (11)과 같이 계산된다.

$$idf_i = \max\left(1, \frac{\log_{10} N/np_i}{5}\right) \quad (10)$$

$$idf_c = \max\left(1, \frac{\log_{10} N/np_c}{5}\right) \quad (11)$$

식(10), (11)에서 N은 컬렉션 내의 단락 수를 나타내며, np_i는 용어 k_i를 가진 단락의 수, nc_i는 개념 c를 포함하는 단락 수를 나타낸다. 인수 δ는 sim(q, c)가 0이 되는 것을 피하기 위한 상수인데, 보통 0.1에 가까운 상수이다. 마지막으로 지수 부분의 idf_i인수는 저빈도 질의 용어들을 강조하기 위하여 도입되었다.

Ⅳ. 실험 및 결과

어떤 정보 요구 I에 대해 연관 문헌 집합을 R이라고 가정하고, |R|은 이 집합의 문헌 수를 표시한다. 어떤 검색 방법이 이 정보 요구를 처리하여 응답 문헌 집합 A를 검색하였다고 하고, |A|는 전과 마찬가지로 이 집합의 문헌 수를 표시한다. 또한 |Ra|를 R과 A의 교집합의 문헌 수라 한다[4][5].

- ▶ 재현율(Recall) : 연관 문헌 집합(집합 R) 중 검색된 문헌의 비율을 나타낸다.
- ▶ 정확률(Precision) : 검색된 문헌 집합(집합 A) 중 연관 문헌의 비율을 나타낸다.

$$\text{재현율} = \frac{|Ra|}{|R|} \quad (12)$$

$$\text{정확률} = \frac{|Ra|}{|A|} \quad (13)$$

재현율과 정확률을 결합한 단일 척도가 유용할 수도 있는데, 그러한 단일 척도 중의 하나가 재현율과 정확률의 조화 평균(F) 이다.

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} \quad (14)$$

r(j) : j 번째 순위 문헌에서의 재현율
P(j) : j 번째 순위 문헌에서의 정확률

표 4.1은 P-norm, 적합성 피드백(Relevance Feedback:RF) 모델의 검색효과를 보여준다. 적합성 피드백을 이용한 모델이 P-norm 모델을 이용한 것 보다 재현율에서는 74.59%와 정확률에서는 82.50%의 향상을 보여주고 있다.

표 4.1 P-norm과 적합성 피드백의 재현율 및 정확률 비교

	P-norm	RF	증가율
재현율	0.362	0.63	+0.27(+74.59)
정확률	0.40	0.73	+0.33(+82.50)

표 4.2는 문헌수 제한시 P-norm과 RF의 재현율을 비교한 결과 문헌수≤10인 경우 RF는 P-norm 보다 71.43% 증가하였고, 문헌수 ≤20인 경우는 RF는 P-norm 보다 65.96% 증가하였다.

표 4.2 문헌수 제한시 재현율 비교

	재현율		
	P-norm	RF	증가율
문헌수≤10	0.28	0.48	+0.20(+71.43)
문헌수≤20	0.47	0.78	+0.31(+65.96)

표 4.3는 문헌수 제한시 P-norm과 RF의 정확률을 비교한 결과이다.

표 4.3 문헌수 제한시 정확률 비교

구분 / 측정	정확률		
	P-norm	RF	증가율
문헌수≤10	0.42	0.75	+0.33(+78.57)
문헌수≤20	0.39	0.71	+0.32(+82.05)

그림 4.1은 검색 문서수 20건으로 제한하였을 때 P-norm 검색과 RF 검색 결과를 재현율과 정확률로 표현한 성능 곡선이다.

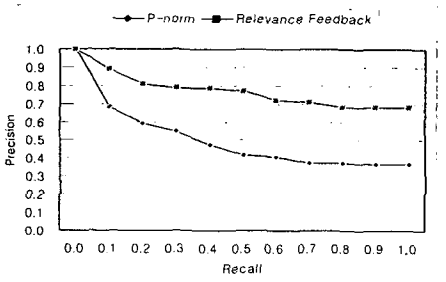

그림 4.1은 P-norm 과 RF의 정확률, 재현율 비교

재현율과 정확률을 결합한 단일 척도가 유용할 수도 있는데, 그러한 단일 척도 중의 하나가 재현율과 정확률의 조화 평균(F) 이다.

표 4.4 조화평균을 이용한 P-norm 모델 측정

재현율	정확률	조화평균	평균값
0.1	0.68	0.170	0.429
0.2	0.59	0.299	
0.3	0.55	0.389	
0.4	0.475	0.434	
0.5	0.42	0.457	
0.6	0.41	0.487	
0.7	0.374	0.487	
0.8	0.374	0.510	
0.9	0.366	0.521	
1.0	0.366	0.536	

표 4.5 조화평균을 이용한 RF 모델 측정

재현율	정확률	조화평균	평균값
0.1	0.89	0.18	0.5758
0.2	0.81	0.327	
0.3	0.79	0.435	
0.4	0.79	0.531	
0.5	0.77	0.606	
0.6	0.72	0.653	
0.7	0.71	0.704	
0.8	0.68	0.735	
0.9	0.68	0.778	
1.0	0.68	0.809	

표 4.6은 적합성 피드백(Relevance Feedback :RF) 모델과 지역적 문맥 분석 피드백(Local Context Analysis Feedback : LCAF)의 검색 효과를 보여준다. LCAF를 이용한 모델이 RF 모델을 이용한 것 보다 재현율에서는 3.173%와 정확률에서는 12.82%의 향상을 보여주고 있다.

표 4.6 RF와 LCAF의 재현율 및 정확률 비교

	RF	LCAF	증가율
재현율	0.63	0.65	+0.02(+3.174)
정확률	0.73	0.78	0.05(+12.82)

표 4.7는 문헌수 제한시 RF와 LCAF의 재현율을 비교한 결과 문헌수≤10인 경우 LCAF는 RF 보다 8.33% 증가하였고, 문헌수 ≤20인 경우는 LCAF는 RF 보다 1.282% 증가하였다.

표 4.7 문헌수 제한시 재현율 비교

	재현율		
	RF	LCAF	증가율
문헌수≤10	0.48	0.52	+0.04(+8.33)
문헌수≤20	0.78	0.79	+0.01(+1.282)

표 4.8는 문헌수 제한시 RF와 LCAF의 정확률을 비교한 결과이다.

표 4.8 문헌수 제한시 정확률 비교

	정확률		
	RF	LCAF	증가율
문헌수≤10	0.75	0.77	+0.02(+2.67)
문헌수≤20	0.71	0.74	+0.03(+4.23)

그림 4.2는 검색 문서수 20건으로 제한하였을 때 RF 검색과 LCAF 검색 결과를 재현율과 정확률로 표현한 성능 곡선이다.

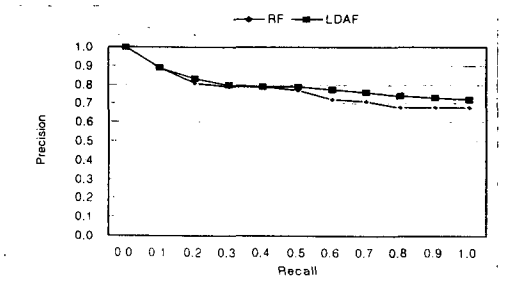

그림 4.2 RF와 LCAF의 정확률, 재현율 비교

재현율과 정확률을 결합한 단일 척도가 유용할 수도 있는데, 그러한 단일 척도 중의 하나가 재현율과 정확률의 조화 평균(F) 이다.

표 4.9 조화평균을 이용한 P-norm 모델 측정

재현율	정확률	조화평균	평균값
0.1	0.89	0.18	0.5758
0.2	0.81	0.327	
0.3	0.79	0.435	
0.4	0.79	0.531	
0.5	0.77	0.606	
0.6	0.72	0.653	
0.7	0.71	0.704	
0.8	0.68	0.735	
0.9	0.68	0.778	
1.0	0.68	0.809	

표 4.10 조화평균을 이용한 RF 모델 측정

재현율	정확률	조화평균	평균값
0.1	0.89	0.180	0.5898
0.2	0.831	0.322	
0.3	0.8	0.437	
0.4	0.79	0.531	
0.5	0.79	0.612	
0.6	0.773	0.675	
0.7	0.76	0.728	
0.8	0.74	0.769	
0.9	0.729	0.809	
1.0	0.72	0.837	

Ⅳ. 결론

지역적 문맥 분석 피드백 기술은 질의에 의해 제공된 지역 문맥을 활용한다는 점에서 매우 긍정적이다. 이러한 점에서 지역적 문맥 분석 기술은 적합성 피드백 기술보다 더 바람직하며 긍정적인 결과가 제시되었다.

앞으로 지역적 문맥 분석 기술을 웹에 활용하는 문제는 아직 깊이 연구되지 않았다. 이러한 연구의 가장 큰 어려움은 질의 시간에 문헌을 분석해야 하기 때문에 검색 엔진측에 큰 계산 부하가 걸린다. 따라서 관련된 중요 연구 주제 중 하나로 검색 엔진측의 질의 처리 속도를 향상시키는 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] E. A. Fox. Extending the Boolean and Vector space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. PhD thesis, Cornell University, Ithaca, New York, Http:// www.ncstrl.org, 1983.

[2] Donna Harman. Relevance feedback revisited. In Proc. of the 5th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1-10, Copenhagen, Denmark, 1992.

[3] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval, pages 4-11, Zurich, Switzerland, 1996.

[4] Baeza-Yates, R. and Ribeiro-Neto, Berthier. Moderm Information Retrieval, addison-wesley Pub. Co(sd), 1992.

[5] G. Salton and C. Buckley. Term-weighting approaches in automatic retrieval. Information Processing & Management, 24(5):513-523, 1988.