벡터모델에서 용어 가중치 재부여를 이용한 질의 확장

Query Expansion Using Term Reweighting for Vector Model

김영천*, 이재훈*, 문유미*, 박병권**, 이성주*
Young-cheon kim, Jac-Hoon Lee,
You-Mi Moon, Byung-Gweun Park, Sung-joo Lee

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

요약

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 MMM 모델, Paice 모델, P-norm 모델이 개발되었다. 본 논문에서는 높은 검색 효과를 제공하는 벡터모델에서 용어 가중치 재부여를 이용한 정보검색 모델을 제안한다. 벡터모델에서 용어 가중치 재부여를 이용한 질의 확장 모델의 연산 특성이 MMM, Paice, P-norm 모델보다 우수함을 설명하고, 또한 성능 비교를 통하여 이를 입증한다.

Keyword : 질의확장, 벡터모델, 가중치 재부여, 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.

벡터 모델은 이진 가중치 사용이 너무 제한적이어서, 부분 정합이 가능한 틀을 제공한 것으로 인식할 수 있으며, 이는 질의나 문헌의 색인어에 비이진 가중치를 할당함으로써 가능하다[1]. 이 용어 가중치는 궁극적으로 사용자 질의와 시스템에 저장되어 있는 각 문헌과의 유사도를 계산하는데 사용되는데 검색된 문헌을 이 유사도 값의 내림차순으로 정렬함으로써 벡터 모델은 질의 용어에 부분 정합되는 문헌을 포함시킨다.

본 논문에서는 벡터모델에서 가중치를 재부여하여 질의를 확장하는 모델을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 부울 연산자를 유연하게 연산하는 기존의 방법들 MMM, Paice, P-norm 모델에 대하여 기술한다. 3장에서는 높은 검색 효과를 제공하는 벡터모델에서 가중치 재부여를 이용한 질의 확장 모델을 제안한다. 4장에서는 벡터모델에서 가중치 재부여를 이용한 질의 확장 모델과 MMM, Paice, P-norm 모델의 성능을 비교한다. 마지막으로 5장에서 결론 및 앞으로의 연구 방향을 제시한다.

II. 부울연산을 유연하게 연산하는 기존의 방법들

퍼지 집합 모델의 단일 피연산자 의존 문제를 극복하기 위해 MMM 모델, Paice 모델, P-norm 모델이 개발되었다.

퍼지 집합 모델의 부울 연산자 계산식 (a)는 두 개의 피연산자를 갖는 이항연산이고, MMM, Paice, P-norm 모델의 연산자 계산식은 2개 이상의 피연산자를 갖는 다항연산이다.

$$F(d,\ t_1\ AND\ t_2) = MIN(w_1, w_2) \quad \ldots (1)$$

$$F(d,\ t_1\ OR\ t_2) = MAX(w_1, w_2) \quad \ldots (2)$$

(a) 퍼지 집합 모델

$$F(d,\ t_1\ AND\ \ldots\ AND\ t_n) = r.MAX(w_1 \ldots w_n) + (1-r).MIN(w_1 \ldots w_n)$$
$$0 \le r \le 0.5 \quad \ldots \quad (3)$$

$$F(d,\ t_1\ OR\ \ldots\ OR\ t_n) = r.MIN(w_1 \ldots w_n) + (1-r).MAX(w_1 \ldots w_n)$$
$$0.5 \le r \le 1 \quad \ldots (4)$$

(b) MMM 모델

$$F(d, t_1\ AND \ldots AND\ t_n) = \frac{\sum_{i=1}^{n}(r^{i-1}.w_i)}{\sum_{i=1}^{n}(r^{i-1})} \quad (5)$$

($0 \le r \le 1$, w_i'는 오름차순정렬)

$$F(d, t_1\ OR \ldots OR\ t_n) = \frac{\sum_{i=1}^{n}(r^{i-1}.w_i)}{\sum_{i=1}^{n}(r^{i-1})} \quad (6)$$

($0 \le r \le 1$, w_i'는 내림차순정렬)

(c) Paice 모델

$$F(d,\ t_1\ AND \ldots AND\ t_n) = 1 - \left[\frac{(1-w_i)^p + \ldots + (1-w_n)^p}{n}\right]^{\frac{1}{p}} \quad (7)$$

$(1 \le p \le \infty)$

$$F(d,\ t_1\ OR \ldots OR\ t_n) = \left[\frac{w_i^p + \ldots + w_n^p}{n}\right]^{\frac{1}{p}} \quad (8)$$

$(1 \le p \le \infty)$

(d) P-norm 모델

확장된 부울 검색 체계를 기반으로 하는 검색 모델은 문서값을 계산하기 위하여 색인어 가중치를 사용한다. 색인어 가중치는 역문헌빈도(Inverse Document Frequency)와 색인어 출현빈도(Term Frequency)로부터 유도될 수 있다. 확장된 부울 검색 체계에서 색인어 가중치는 0부터 1 사이의 값이어야 하기 때문에 W_{ik}는 식(9)와 같이 정규화된다.

$$W_{ik} = \frac{TF_{ik}}{\max TF_i} \cdot \frac{IDF_k}{\max IDF_i} \quad (9)$$

N : 문서 집합을 구성하는 문서들의 수

IDF_k(역문헌빈도) : $\log(N/n_k)$

TF_{ik}(색인어 출현빈도): 문서 i에서 색인어 k의 출현빈도.

W_{ik} : $IDF_k \cdot F_{ik}$

III. 벡터모델에서 용어가중치 재부여

벡터 모델에서 문헌 d_j와 질의 q의 유사도 측정은 두 벡터 $\vec{d}_j$ 와 $\vec{q}$의 상관도로 구할 수 있으며, 이 상관도의 예로 두 벡터간 사이각의 코사인 값으로 정량화 할 수 있다.

$$sim(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|}$$

$$= \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \quad (10)$$

$w_{i,j}$ 와 $w_{i,q}$가 0보다 크거나 같은 값을 갖기 때문에 sim(q, d_j)값은 0과 1 사이의 값이 된다. 문헌 d_j에서의 용어 k_i의 정규화 빈도는 다음과 같다.

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \quad \ldots (11)$$

N : 시스템 내의 총 문헌 수

n_i : 색인어 k_i 가 출현한 문헌 수

$freq_{i,j}$: 문헌 d_j에서의 용어 k_i 출현 빈도수

용어 k_i의 역문헌 빈도수 idf_i는 다음과 같다.

$$idf_i = \log \frac{N}{n_i} \cdots (12)$$

최대값 max : 문헌 d_j 텍스트 내에 출현한 모든 용어 중에서 가장 빈도수가 큰 용어

가장 널리 알려진 용어-가중치 할당 기법은 다음과 같은 가중치 식(13)를 사용한다.

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i} \cdots (13)$$

질의 용어 가중치 식(14)는 다음과 같다.

$$w_{i,q} = \left(0.5 + \frac{0.5\ freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log \frac{N}{n_i} \quad (14)$$

$freq_{i,q}$: 정보 요구 q 텍스트에서의 용어 k_i 의 빈도수

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].

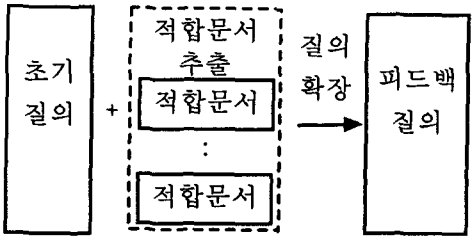

그림 3.1 가중치 재부여

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(15)와 같이 증명된다.

$$\vec{q}_{opt} = \frac{1}{|C_r|} \sum_{\forall \vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N-|C_r|} \sum_{\forall \vec{d}_j \notin C_r} \vec{d}_j \quad (15)$$

D_r :검색된 문헌 중에서 사용자에 의해 연관 문헌으로 판단된 문헌 집합

D_n : 검색된 비연관 문헌 집합

C_r :컬렉션 내 모든 문헌 중 연관 문헌 집합

$|D_r|, |D_n|, |C_r|$:각 집합 D_r, D_n, C_r의 문헌 수

α, β, γ :조절 상수

수정된 질의 $\vec{q}_m$을 계산하는 고전적인 방법은 다음 식(16)과 같다.

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j \quad (16)$$

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.

재현율= $\frac{|Ra|}{|R|}$ …(17) 정확률= $\frac{|Ra|}{|A|}$ …(18)

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

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} \cdots (19)$$

r(j) : j 번째 순위 문헌에서의 재현율

P(j) : j 번째 순위 문헌에서의 정확률

표 3.1은 MMM, Paice, P-norm, VTR(Vector Term Reweighting) 모델의 검색 효과를 보여준다. 각 질의에 대한 정확률과 재현율를 0.25, 0.5, 0.75에 고정시켜 계산된 정확률의 평균값이다.

표 3.1 각 모델의 정확률 비교

모델 \ 구분	정확률 평균값
MMM	0.327
Paice	0.318
P-norm	0.362
VTR	0.602

표 3.2는 문헌수 제한시 P-norm과 VTR의 재현율을 비교한 결과 문헌 수≤10인 경우 VTR은 P-norm 보다 39.29% 증가하였고, 문헌수 ≤20인 경우는 VTR은 P-norm 보다 46.81% 증가하였다.

표 3.2 문헌수 제한시 재현율 비교

구분 / 측정	재현율		
	P-norm	VTR	증가율
문헌수≤10	0.28	0.39	+0.11(+39.285)
문헌수≤20	0.47	0.69	+0.22(+46.81)

표 3.2는 문헌수 제한시 P-norm과 VTR의 정확률을 비교한 결과 문헌 수≤10인 경우 VTR은 P-norm 보다 50% 증가하였고, 문헌수≤20인 경우는 VTR은 P-norm 보다 48.72% 증가하였다.

표 3.3 문헌수 제한시 정확률 비교

구분 / 측정	정확률		
	P-norm	VTR	증가율
문헌수≤10	0.42	0.63	+0.21(+50.00)
문헌수≤20	0.39	0.58	+0.19(+48.72)

그림 3.1은 검색 문서수 20건으로 제한하였을 때 P-norm 초기검색과 VTR(Vector Term Reweighting) 검색 결과를 재현율과 정확률로 표현한 성능 곡선이다.

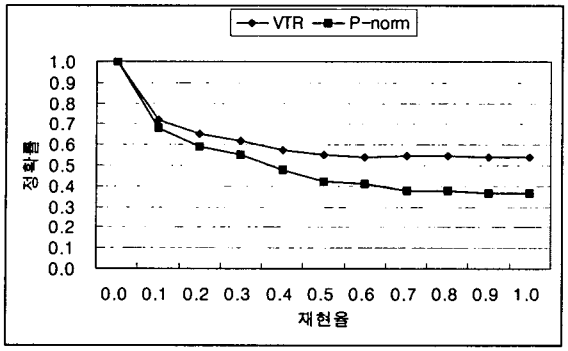

그림 3.1은 P-norm 과 VTR의 정확률, 재현율 비교

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

표 3.4 조화평균을 이용한 P-norm 모델 측정

재현율	정확률	조화평균	전체조화평균
0.1	0.68	0.174	0.430
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	0.488	
0.8	0.374	0.510	
0.9	0.366	0.520	
1.0	0.366	0.536	

표 3.5 조화평균을 이용한 VTR 모델 측정

재현율	정확률	조화평균	전체조화평균
0.1	0.72	0.176	0.51
0.2	0.65	0.306	
0.3	0.62	0.405	
0.4	0.574	0.472	
0.5	0.553	0.525	
0.6	0.54	0.568	
0.7	0.546	0.613	
0.8	0.546	0.649	
0.9	0.54	0.676	
1.0	0.54	0.702	

Ⅳ. 결론

가중치 재부여 벡터 모델의 주요 장점은 첫째, 용어-가중치 할당 기법이 검색 성능을 향상시키고, 둘째, 부분 정합 전략으로 질의 조건에 근접한 문헌 검색이 가능하며, 셋째, 코사인 순위화 수식이 문헌을 질의에 유사한 순서대로 정렬한다는 점이다.

단순성에도 불구하고 벡터 모델은 일반적인 컬렉션에 탄력적인 순위화 전략을 제공하고 있으며, 벡터 모델의 틀 내에서 질의 확장이나 연관 피드백을 사용하여 성능이 향상된 결과 집합을 제공하고 있다.

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.