퍼지 및 지능시스템학회 논문지 2001, Vol. 11, No. 9, pp. 811-816

정보 검색에서 용어 가중치 재부여를 이용한 성능 증진에 관한 연구

A Study on Improving the Effectiveness Using Term Reweighting for Information Retreival

김영천*, 이재훈*, 문유미*, 박병권**, 이성주*

Young-cheon kim, Jac-Hoon Lee,
You-Mi Moon, Byung-Gweun Park, and Sung-joo Lee

*조선대학교 전자계산학과

**서강정보대학 정보통신과

요 약

정보 검색 시스템의 중요한 목적중의 하나는 단순히 사용자 질의를 만족하는 문서들의 집합을 검색하는 것이 아니라, 질의를 만족하는 정도에 따라 검색된 문서들에 순위를 부여함으로써 사용자들이 필요한 정보를 얻는데 소모되는 시간을 최소화시키는 것이다.

순수한 부울 검색 시스템은 검색 전략이 이진값에 근거하여 순위 구분 없이 연관/비연관 중의 하나로 결정된다. 딸서 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.

Key Word : 질의 확장, 벡터모델, 가중치 재부여, 부울 검색, 유사도

1. 서 론

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. 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 확장모델을 제안한다. 4장에서는 벡터모델에서 가중치 재부여를 이용한 질의 확장 모델과 MMM, Paice, P-norm 모델의 성능을 비교한다. 마지막으로 5장에서 결론 및 앞으로의 연구 방향을 제시한다.

접수 일자 : 2001년 9월 15일
완료 일자 : 2001년 12월 1일

2. 부울연산을 유연하게 연산하는 기존의 방법들

순수한 부울 검색 모델은 문서와 질의 사이의 유사도를 나타내는 문서값을 계산할 수 없기 때문에, 검색된 문서들을 질의를 만족하는 정도에 따라 정렬할 수 없다는 단점을 지니고 있다[1, 2].

순수한 부울 검색 시스템의 단점을 보완하기 위하여 퍼지 집합 모델(Fuzzy Set Model)이 개발되었다.

퍼지 집합 모델은 색인어가 문서 내에서 갖는 중요성을 반영하는 색인어 가중치를 이용하여 문서값을 계산함으로써 부울 검색 시스템의 문제점을 극복하였다. 그러나 퍼지 집합 모델은 많은 경우에 부정확한 문서값을 생성하기 때문에 정보 검색 모델로서 부적합하다고 비판되어 왔다. 이것은 AND와 OR 연산을 위하여 사용하는 MIN과 MAX 연산자가 단일 피연산자 의존 문제(Single

Operand Dependency Problem)를 발생시키기 때문이다.

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

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

(a) 퍼지 집합 모델

퍼지 집합 모델의 단일 피연산자 의존 문제를 극복하기 위해 MMM 모델, Paice 모델, P-norm 모델이 개발되었다. 이들 모델들은 AND와 OR 연산을 위하여 MIN과 MAX 대신에 부울 연산자를 유연하게 연산하는 새로운 연산자를 사용함으로써 퍼지 집합 모델, MMM 모델, Paice 모델, P-norm 모델을 기반으로 하는 정보 검색 시스템은 <T, Q, D, F>로 정의되는 확장된 부울 검색 체계(Extended Boolean Retrieval Framework) 내에서 설명될 수 있다[3].

① T는 질의와 문서를 표현하기 위해 사용되는 색 인어들의 집합이다.
② Q는 시스템이 인식할 수 있는 질의들의 집합이 다. Q에 속하는 각각의 질의 q는 색인어들과 부 울 연산자 AND, OR, NOT으로 구성된 부울 수식이다.
③ D는 문서들의 집합이다. D에 속하는 각각의 문 서 d는 wi가 색인어 ti의 가중치일 때, {(t1, w1), ..., (tn, wn)}와 같이 표현된다. 색인어 가중치 w1는 0부터 1 사이의 값을 갖는다.
④ F는 문서값을 계산하는 순위 결정 함수(Ranking Function)로서 다음과 같이 정의된다.

$$F : D \times Q \rightarrow [0, 1]$$

검색함수 F는 각 쌍의 (d, q)에 0부터 1사이의 값을 지정한다. 이 값은 문서 d와 질의 q사이의 유사도를 의미하며, 질의 q에 대한 문서 d의 문서값이다.

검색 함수 F(d, q)는 다음과 같은 2단계 과정을 거쳐서 계산된다.

(i) 질의에 나타난 각각의 색인어 ti에 대하여, F(d, ti)는 문서 d에서 색인어 ti의 가중치 wi로 정의된 다.
(ii) 부울 연산자 AND와 OR는 (a), (b), (c), (d)에서 주어진 식들을 이용하여 계산되고, NOT은 F(d, NOT ti)=1-wi로 계산된다.

두 개 이상의 부울 연산자를 포함하는 부울 질의는 가장 안쪽에 위치하는 절부터 순환적으로 계산된다.

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. 이것은 퍼지 집합 모델의 MIN과 MAX 연산자가 결합법칙을 만족하는데 비하여 MMM, Paice, P-norm 모델의 연산자는 결합법칙을 만족하지 못하기 때문이다. 결합법칙을 만족하지 못할 경우, 임의의 문서에 대하여 두 개의 동일한 질의((t1 AND t2)AND t3와 t1 AND (t2 AND t3))의 문서값이 서로 다르다. MMM, Paice, P-norm 모델은 이러한 문제점을 다항연산을 가능하게 함으로써 극복하였다.

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

$$0 \le r \le 0.5 \quad (3)$$

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

$$0.5 \le r \le 1 \quad (4)$$

(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≤r≤1, wi'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≤r≤1, wi'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 Wik는 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} : $IDF_k \cdot F_{ik}$

3. 백터모델에서 용어가중치 재부여

벡터 모델은 이진 가중치 사용이 너무 제한적이어서, 부분 정합이 가능한 틀을 제공한 것으로 인식할 수 있으며, 이는 질의나 문헌의 색인어에 비이진 가중치를 할당함으로써 가능하다. 이 용어 가중치는 궁극적으로 사용자 질의와 시스템에 저장되어 있는 각 문헌과의 유사도

를 계산하는데 사용되는데 검색된 문헌을 이 유사도 값의 내림차순으로 정렬함으로써 벡터 모델은 질의 용어에 부분 정합되는 문헌을 포함시킨다. 결과적으로 순위화된 문헌 집합이 불리안 모델에서 검색된 문헌 집합보다 사용자 정보 요구에 더 잘 맞는다고 볼 수 있다.

벡터 모델에서 용어 문헌 쌍(ki, dj)의 가중치 wi,j는 양의 비이진 값이며, 질의 색인어도 가중치를 가진다. [ki, q]의 가중치를 wi,q≥0이라 하면 질의 벡터 $\vec{q}$는 $\vec{q}=(w_{1,q}, w_{2,q}, \cdots, w_{t,q})$로 정의되며, 여기서 t는 시스템 내의 전체 색인어 수이다. 문헌 dj 벡터는 $\vec{d}_j=(w_{1,j}, w_{2,j}, \cdots, w_{t,j})$로 표현된다.

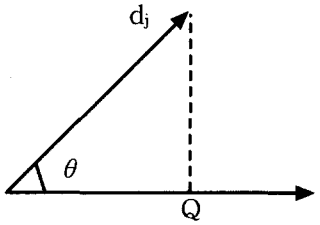


그림 3.1 코사인 θ 값이 sim(dj, q)로 적용
Fig. 3.1 The cosine of θ is adopted as sim(dj, q)

따라서 문헌 d_j와 사용자 질의 q는 그림 3.1와 같이 t 차원 벡터로 표시된다. 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 정량화할 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, dj)값은 0과 1 사이의 값이 된다. 문헌 dj에서의 용어 ki의 정규화 빈도는 식(11)을 이용하여 계산한다.

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

N : 시스템 내의 총 문헌 수
n_i : 색인어 ki 가 출현한 문헌 수
$freq_{i,j}$: 문헌 dj에서의 용어 ki 출현 빈도수

용어 k_i의 역문헌 빈도수 idf_i는 식(12)을 이용하여 계산한다.

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

최대값 max : 문헌 dj 텍스트 내에 출현한 모든 용어 중에서 가장 빈도수가 큰 용어 가장 널리 알려진 용어-가중치 할당 기법은 식(13)을 이용하여 계산한다.

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

질의 용어 가중치는 식(14)를 이용하여 계산한다.

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

$freq_{i,q}$: 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[5, 6, 7].

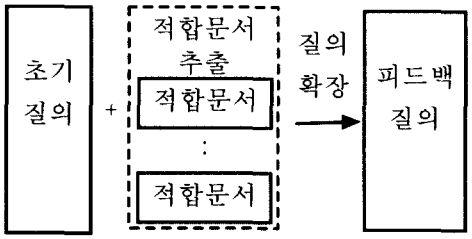


그림 3.2 용어 가중치 재부여
Fig. 3.2 Term Reweighting

비현실적이기는 하지만 주어진 질의 q에 대한 전체 연관 문헌 집합인 Cr을 이미 알고 있다고 가정하면 연관 문헌들을 비연관 문헌들로부터 구분하는 최적 질의 벡터는 식(15)와 같이 증명된다.

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

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

식(15)의 문제점은 집합 Cr을 구성하는 연관 문헌들을 미리 알 수가 없다는 것이며 사실 이 문헌들을 지금 찾고 있는 것이다. 이 문제를 해결하는 자연스러운 방법은 초기 질의를 작성한 다음 점차 초기 질의 벡터를 변화시키는 것이다. 이러한 점진적 변화는 사용자에 의해 연관 문헌으로 판정된 문헌들에게로 계산 방향을 제한함으로써 가능하며, 수정된 질의 $\vec{q}_m$을 계산하는 고전적인 방법은 식(16)과 같다.

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

4. 실험 및 결과

어떤 정보 요구 I에 대해 연관 문헌 집합을 R이라고 가정하고, |R|은 이 집합의 문헌 수를 표시한다. 어떤 검

색 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[9, 8]. 그림 4.1은 집합들을 그림으로 표시하고 있다. 재현율과 정확률은 식(17), (18)과 같이 정의된다.

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

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

$$\text{재현율} = \frac{|Ra|}{|R|} \tag{17}$$

$$\text{정확률} = \frac{|Ra|}{|A|} \tag{18}$$

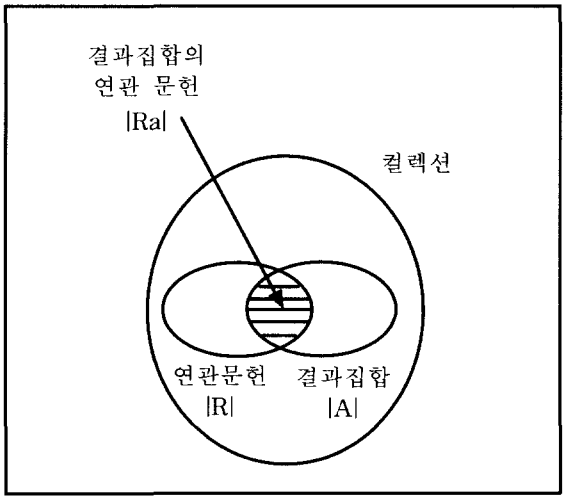


그림 4.1 정보 요구에 대한 정확률과 재현율
Fig. 4.1 Information request for Precision and recall

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

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

식(19)에서 r(j)는 j 번째 순위 문헌에서의 재현율을 의미하며, P(j)는 역시 j 번째 순위 문헌에서의 정확률이다. 또한 F(j)는 r(j)의 조화 평균이며, F 함수는 [0,1] 사이의 값을 가진다. 이 값은 연관 문헌이 하나도 검색되지 않았을 경우 0이 되며, 검색된 문헌이 모두 연관 문헌일 경우 1이 된다. 또한 재현율과 정확률이 모두 높아야 F 값이 커지는 성질이 있다. 따라서 F값을 최대화한다는 것은 재현율과 정확률 사이의 가장 적절한 타협점을 구하는 작업이다.

P-norm 검색과 VTR 검색 결과를 비교 분석하는 데는 P-norm 검색 실험 결과와 VTR 검색 실험 결과의 효율차이를 산출하여 비교하는 방법이 주로 이용된다. 여기에서 효율차이는 두 실험 결과의 검색 효율에 대한 산술적인 차이를 의미하므로 VTR 검색 실험이 P-norm 검색 실험에 비해 상대적으로 얼마만큼 검색 효율을 향상시켰는지를 나타내지는 못하는 문제점을 갖고 있다. 이와 같은 문제점을 해결하기 위해서는 효율차이 대신에 증진율을 산출해야 하는데 구체적으로 증진율은 P-norm 검색 실험 결과와 VTR 검색 실험 결과의 효율차이를 P-norm 검색 실험 결과의 효율로 나누어 백분율로 환산한 값이다.

표 4.1은 MMM, Paice, P-norm, VTR(Vector Term Reweighting) 모델의 검색효과를 보여준다. 각 질의에 대한 정확률과 재현율를 0.25, 0.5, 0.75에 고정시켜 계산된 정확률의 평균값이다.

표 4.1 검색 모델의 정확률 비교
Table 4.1 Precision comparison of retrieval model

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

표 4.2는 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 4.2 P-norm과 VTR의 재현율 비교(문헌수 제한)
Table 4.2 Recall comparison of P-norm and VTR(document number limit)

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 4.2는 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 4.3 P-norm과 VTR의 정확률 비교(문헌수 제한)
Table 4.3 Precision comparison of P-norm and VTR(document number limit)

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 4.2은 검색 문서수 20건으로 제한하였을 때 P-norm 초기검색과 VTR(Vector Term Reweighting) 검색 결과를 재현율과 정확률로 표현한 성능 곡선이다.

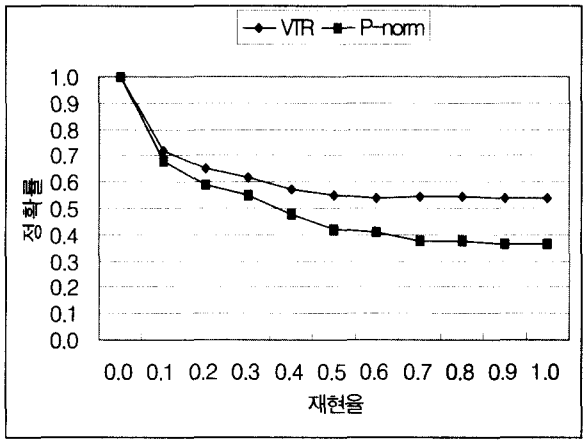


그림 4.2 P-norm 과 VTR의 정확률, 재현율 비교
Fig 4.2 Precision, Recall comparison of P-norm and VTR

표 4.4 조화평균을 이용한 P-norm 모델 측정
Table 4.4 Measure of P-norm model using harmonic mean

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 4.5 조화평균을 이용한 VTR 모델 측정
Table 4.5 Measure of VTR model using harmonic mean

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	

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

조화평균을 이용하여 P-norm 검색 모델을 이용하여 평가했을 때 0.430을 나타내고, VTR 검색 모델을 이용하여 평가했을 때는 0.51을 나타내고 있다. 그러므로 VTR 정보검색 모델이 P-norm 정보검색 모델 보다 0.08 정도 향상됨을 보여주고 있다.

5. 결 론

정보 검색 시스템의 중요한 목적중의 하나는 단순히 사용자 질의를 만족하는 문서들의 집합을 검색하는 것이 아니라, 질의를 만족하는 정도에 따라 검색된 문서들에 순위를 부여함으로써 사용자들이 필요한 정보를 얻는데 소모되는 시간을 최소화시키는 것이다.

가중치 재부여 벡터 모델의 주요 장점은 첫째, 용어가중치 할당 기법이 검색 성능을 향상시키고, 둘째, 부분정합 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] 신은자 · 정영미, "피드백 정보를 이용한 불논리 검색 시스템의 성능 증진에 관한 실험적 연구", *정보처리학회지*, 제15권 제1호, pp. 129-147, 1998.

[2] 이준호 · 이기호 · 조영화, "정보검색에서 부울연산자를 연산하는 식의 수학적 특성", *정보처리학회지*, 제12권 제1호, pp. 87-96, 1995.

[3] 이준호, "부울 연산자에 대한 효율적이며 효과적인 연산 방법", *한국정보과학회논문지*, 제21권 제3호, pp. 440-445, 1994.

[4] 이재윤 · 최보영 · 정영미, "문헌 자동분류에서 용어가중치 기법에 대한 연구", *한국정보관리학회 학술대회 논문집*, 제7회, pp. 41-44, 2000.

[5] 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.

[6] 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.

[7] 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.

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

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

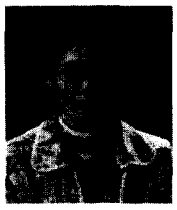
저 자 소 개

김영천(Young-Chon Kim)
1992년 : 광주대 전자계산학과(공학사)
1996년 : 조선대 컴퓨터공학과(공학석사)
1998년 ~ 현재 : 조선대 전자계산학과 박사과정

관심분야 : 객체지향시스템, 소프트웨어공학, 유전자알고리즘, 정보검색

이재훈(Jae-Hoon Lee)
2000년 : 조선대 전자계산학과 졸업 (이학사)
2000년 ~ 현재 : 조선대 전자계산학과 석사과정

관심분야 : 객체지향시스템, 소프트웨어공학, 정보처리

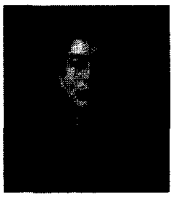


문유미(You-Mi Moon)
1983년 : 조선대학교 전자계산학과 (이학사)
1987년 : 조선대학교 전자계산학과 (이학석사)
1998 ~ 현재 : 조선대학교 전자계산학과 박사과정

관심분야 : 소프트웨어공학, 객체지향시스템, 러프 집합, 퍼지집합, 전자상거래

박병권(Byung-Gweun Park)
1988년 : 조선대학교 전자계산학과 (이학사)
1990년 : 조선대학교 전자계산학과 (이학석사)
2000년 : 조선대학교 전자계산학과 (이학박사)
1991년 ~ 1994년 : 광주은행 전산업무부
1995년 ~ 현재 : 서강정보대학 정보통신과 조교수

관심분야 : 소프트웨어공학, 객체지향시스템, 퍼지집합, 러프집합



이성주(Sung-Joo Lee)
1970년 : 한남대학교 물리학과(이학사)
1992년 : 광운대학교 전자계산학과 (이학석사)
1998년 : 대구가톨릭대학교 전자계산학과 (이학박사)
1988년 ~ 1990년 : 조선대학교 전자계산소 소장
1995년 ~ 1997년 : 조선대학교 정보과학대학장
1981년 ~ 현재 : 조선대학교 컴퓨터공학부 교수

관심분야 : 소프트웨어 공학, 프로그래밍 언어, 객체지향 시스템, 러프 집합