P-norm 모델의 구현과 검색효율 향상

*최진석, *박민식, *김진혁, *김영관, **권혁철
*부산대학교 전자계산학과, **부산대학교 정보·컴퓨터공학부
jschoe@solge.cs.pusan.ac.kr, sadwind@solge.cs.pusan.ac.kr
variant@borame.cs.pusan.ac.kr, leon@borame.cs.pusan.ac.kr
hckwon@hyowon.pusan.ac.kr

Implementation of P-norm Model and Improvement of Retrieval Efficiency

Jin-Seok Choe, ,Min-Sik Park, Jin-Hyuk Kim,
Young-Kwon Kim, Hyuk-Chul Kwon
Department of Computer Science, Pusan National University
Division of Information and Computer Engineering, Pusan National University

요 약

본 논문에서는 불리언 모델의 단점을 극복하기 위해 만들어진 확장 불리언 모델 중 가장 검색효율이 높다고 알려진 P-norm 모델을 구현하고 검색효율을 최대화하는 여러 가지 방법에 대해 연구하고 실험하였다. P-norm 모델에서 검색효율에 영향을 미치는 요소는 색인어 가중치 부여방법, P값 부여방법, 클로즈 가중치 부여방법이다. 각 요소별로 검색효율을 향상시킬 수 있는 여러 가지 방법을 제시하고 실험을 통하여 최적화된 방법을 찾아내었다. 최적화 후 P-norm 모델에서 정확도는 최적화 전보다 10.8% 증가하였다.

1. 서론

본 논문의 목적은 정보검색 시스템에서 P-norm 모델을 구현할 때 검색효율에 영향을 주는 파라미터를 정의하고, 검색효율을 최대화하는 파라미터 결정방법을 찾아내는 것이다.

순수한 불리언 모델(Boolean Model)은 구현이 간편하고 검색속도가 빠르다는 장점이 있으나, 문서에 가중치를 주지 않아 검색된 문서를 순위화할 수 없다는 단점이 있다[5]. 이러한 불리언 모델의 단점을 극복하기 위하여 제안된 모델을 확장불리언 모델(Extended Model)이라고 한다. 확장 불리언 모델로서 Fuzzy Set 모델, Waller-Kraft 모델, Paice 모델, P-norm 모델 등이 제안되었는데 이들 모델 중 검색효율이 가장 높은 모델은 P-norm 모델이다[1].

본 논문은 부산대학교 정보검색 시스템 "미리내"에서 P-norm 모델을 구현하던 중 고려되었던 파라미터들을 제시하고 이 파라미터들을 결정하는 여러 가지 방법을 소개하며 실험을 통하여 최적화된 방법을 찾아낸다. P-norm 모델에서 검색효율에 영향을 미치는 파라미터들은 다음과 같다.

첫째, 색인어 가중치 부여방법(Term Weighting Methods)이다. 본 논문에서 사용하는 색인어 가중치 부여방법은 tf·idf-cosine weights[2] 방법과 Fox weights[3] 방법이다. 이 두 가지 방법은 단어빈도(TF)와 역문헌빈도(IDF)의 개념을 사용하고, 색인어가 문서에 1개도 나타나지 않는 것보다 1개라도 나타나는 색인어의 가중치를 더 높여 준다. 두 가지 방법의 차이점은 정규화 방법에 의한 차이이다. tf·idf-cosine weights 방법은 코사인 정규화 방법을 사용하고 Fox weights 방법은 콜렉션에 등록된 문서 총 수의 로그값으로 나누어준다.

위의 두 가지 방법은 문서에서 단어의 상대적인 중요도를 나타내기 위해 Max TF(문서의 색인어들 중 빈도가 가장 높은 색인어의 빈도)를 사용한다. 본 논문에서는 추가적으로 Max TF 와 Sum TF(문서에 나타나는 색인어 빈도의 합)를 사용하여 실험하

였다.

둘째, P값 부여방법이다. AND 연산자와 OR 연산자에 똑같은 P값을 부여하는 것이 아니라, AND 연산자에는 큰 P값을 OR 연산자에는 작은 P값을 부여하는 것이 효과적이다[5].

셋째, 클로즈의 가중치 부여방법이다. P-norm 거리(P-norm distance)를 이용한 클로즈 가중치 부여방법만을 사용하면, AND 클로즈와 OR 클로즈의 차이가 거의 없어진다. 그래서 AND 클로즈의 가중치를 높이기 위해 sum-weights 방법과 sum-weights-modified 방법[4]을 사용한다. 본 논문에서는 이 두 가지 방법을 비교분석한다.

본문에서는 검색효율에 영향을 미치는 위의 3가지 파라미터를 차례로 적용하고 실험함으로써 P-norm 모델을 최적화시킨다. 2절에서 P-norm 모델에 대해 개괄적으로 설명하고 3절에서는 검색성능을 결정하는 세 가지 요소를 최적화하기 위한 연구와 실험을 하고 실험 결과를 평가한다. 그리고 4절에서 결론을 맺는다.

실험은 KT Set 2.0에서 이루어지며 50개의 불리언 질의어를 사용한다. 정확도의 계산은 Salton의 '11 points recall-precision' 계산법[6]을 사용한다. 앞으로 실험에서 사용하는 정확도라는 용어는 50개 질의어의 평균 11지점 정확도의 평균을 의미한다.

2. P-norm 모델

P-norm 모델은 불리언 모델의 단점을 극복하기 위하여 만들어진 확장 불리언 모델의 하나이다. P-norm 모델은 질의어에 의해 문서공간(Document Space)이 만들어지고 각각의 문서는 질의어에 의해 만들어진 문서공간 안에서 한 점으로 표현된다.

p-norm distance는 다음과 같이 정의된다.

임의의 벡터 $\vec{X}=(x_1, x_2, \cdots, x_n)$의

p-norm distance : $\|\vec{X}\|_p=\sqrt[p]{|x_1|^p+\cdots+|x_n|^p}$

p-norm distance를 정보검색에 적용한 것이 P-norm 모델이다. p-norm distance는 AND, OR 연산자를 계산할 때 사용된다.

P-norm 모델에서 가장 핵심적인 사항은 AND, OR 연산이다. 질의어에 포함된 색인어의 가중치를 1로 본다면 AND 연산과 OR 연산은 다음과 같이 이루어진다.

AND 연산

$$Q=[t_1 AND^p t_2 AND^p \cdots AND^p t_n]$$

$$D=(d_1, d_2, \cdots, d_n)$$

d_i = weight of t_i in D

$$Sim(D, Q)$$

$$=1-\frac{L_p dist((1,1,\cdots,1),(d_1,d_2,\cdots,d_n))}{L_p dist((0,0,\cdots,0),(1,1,\cdots,1))}$$

$$=1-\sqrt[p]{\frac{(1-d_1)^p+(1-d_2)^p+\cdots+(1-d_n)^p}{n}}$$

OR 연산

$$Q=[t_1 OR^p t_2 OR^p \cdots OR^p t_n]$$

$$D=(d_1, d_2, \cdots, d_n)$$

d_i = weight of t_i in D

$$Sim(D, Q)$$

$$=\frac{L_p dist((0,0,\cdots,0),(d_1,d_2,\cdots,d_n))}{L_p dist((0,0,\cdots,0),(1,1,\cdots,1))}$$

$$=\sqrt[p]{\frac{d_1^p+d_2^p+\cdots+d_n^p}{n}}$$

질의 색인어에 가중치를 준다면 위의 OR 연산은 다음과 같이 바뀐다.

OR 연산(질의어 가중치를 사용할 때)

$$Q=(q_1, q_2, \cdots, q_n)$$

$$D=(d_1, d_2, \cdots, d_n)$$

$$Sim(D, Q)$$

$$=\frac{L_p dist((0,0,\cdots,0),(q_1d_1,q_2d_2,\cdots,q_nd_n))}{L_p dist((0,0,\cdots,0),(q_1,q_2,\cdots,q_n))}$$

$$=\sqrt[p]{\frac{q_1^p d_1^p+q_2^p d_2^p+\cdots+q_n^p d_n^p}{q_1^p+q_2^p+\cdots+q_n^p}}$$

AND 연산도 OR 연산과 비슷한 방법으로 계산하며, AND 연산과 OR 연산이 섞여 있는 수식은 위의 AND, OR 연산을 재귀적으로 수행한다.

기본적으로 질의 색인어 가중치는 IDF를 사용하고 문서 색인어 가중치는 Fox weights를 사용하며 P값은 AND, OR 모두 1.5를 사용한다.

3. P-norm 모델의 최적화

3.1 색인어 가중치 부여방법

벡터공간 모델이나 P-norm 모델을 사용하는 정보검색 시스템에서 색인어의 가중치를 설정하는 방법은 매우 중요하다. 벡터공간 모델에서는 색인어 가중치를 주로 TF · IDF로 결정한다[6]. P-norm 모델에서도 TF · IDF의 개념을 사용한다.

색인어 가중치는 두 종류가 있다. 질의어 색인어 가중치(Query Term Weight)와 문서 색인어 가중치(Document Term Weight)가 있다. 질의어의 색인어는 TF의 의미가 없기 때문에 다음의 IDF(Inverse Document Frequency)만을 사용한다[4].

$$\log(\frac{N}{n})$$

N : 검색대상이 되는 문서집합의 문서 수,
n : 전체 문서집합에서 현재 색인어를 포함하는 문서 수

따라서 문서 색인어 가중치가 중요해진다. 본 논문에서는 문서 색인어 가중치 부여방법으로 tf · idf-cosine weights 방법[2]과 Fox weights 방법[3]을 사용하고, 또 이 두 가지 방법에서 Max TF(문서의 색인어들 중 빈도가 가장 높은 색인어의 빈도)와 Sum TF(문서에 나타나는 색인어 빈도의 합)를 사용하여 실험하고 결과를 비교 분석한다.

tf · idf-cosine weights 방법과 Fox weights 방법은 다음과 같다.

어떤 문서에 대한 가중치를 $\overline{w}=(w_1, w_2, \ldots, w_m)$라 하면,

- tf · idf-cosine 방법

$$w_i=(r+(1-r)\frac{tf_i}{\max tf})\log(\frac{N}{n_i}) \quad \text{if } tf_i>0$$

$$w_i=0 \quad \text{if } tf_i=0$$

여기서 계산된 w_i는 $\sqrt{\sum_{i=0}^{m} w_i^2}$ 으로 정규화한다.

- Fox Weights 방법

$$w_i=(r+(1-r)\frac{tf_i}{\max tf})\frac{\log(\frac{N}{n_i})}{\log(N)} \quad \text{if } tf_i>0$$

$$w_i=0 \quad \text{if } tf_i=0$$

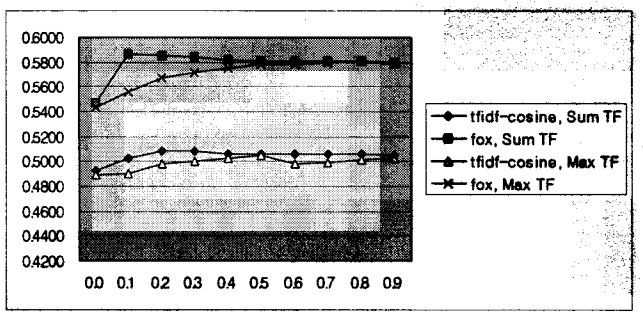
tfi : i번째 색인어가 현재 문서에 나타나는 빈도
maxtf : 현재 문서에서 빈도가 가장 높은 색인어의 빈도
N : 검색대상이 되는 전체문서 집합의 문서 수
n_i : 전체 문서집합에서 현재 색인어를 포함하는 문서 수

위 수식에서 r의 의미는 문서에 한 번 나타나는 색인어와 한 번도 나타나지 않는 색인어는 차이를 둔다는 뜻이다[4]. r값이 커짐에 따라 그 차이도 더욱 커진다. $(r+(1-r)\frac{tf_i}{\max tf})$는 TF의 개념이며, $\log(\frac{N}{n_i})$는 IDF이다. 따라서 P-norm 모델도 색인어 가중치TF · IDF의 개념[6]을 사용한다.

위 수식에서 두 문서 색인어 가중치 부여방법의 유일한 차이점은 정규화 방법이다. tf · idf-cosine weights 방법은 코사인 정규화 방법을 사용하고 있고, Fox weights 방법은 단순히 log(N)으로 나누어 준다. 코사인 정규화 방법은 길이가 긴 문서의 가중치를 낮추어 준다는 것이 장점이고, 한 문서 안에서 자체적으로 정규화함으로써 다른 문서와의 상대적인 중요도를 측정하기가 어렵다는 것이 단점이다. log(N)으로 나누어서 정규화하는 것은 다른 문서와의 상대적인 중요도를 잘 나타낸다[4].

위와 같이 정규화를 하는 이유는 색인어의 가중치가 1을 넘지 못하도록 하기 위해서이다. 불리언 모델에서 가중치 1은 완전한 일치를 의미한다. 본 논문에서는 P-norm 모델도 불리언 모델의 확장으로 간주하여 가중치는 1을 넘지 못하도록 하였다.

r값의 변화에 따른 정확도의 변화를 tf · idf-cosine weights 방법과 Fox weights 방법으로 나누어 실험하였다. 실험 결과는 다음 [그림 1]과 같다.



[그림 1] 색인어 가중치 부여방법에 따른 검색효율의 변화 : x축 - r, y축 - 정확도

[그림 1]에서 위에 있는 두 개의 그래프가 Fox weights를 적용한 정확도 그래프이고, 아래의 그래프 2개가 tf·idf-cosine weights 방법을 사용한 정확도 그래프이다. Fox weights 방법의 정확도가 tf·idf-cosine weights 방법의 정확도보다 평균 14.5% 높다.

Fox weights가 tf·idf-cosine weights보다 더 좋은 성능을 나타내는 이유는 log(N)으로 정규화하기 때문이다. tf·idf-cosine weights 방법에서 정규화는 지역적이기 때문에 다른 문서와의 중요도 차이를 줄인다. 어떤 문서 셋에서 log(N)은 절대적인 값이다. log(N)으로 정규화하면 문서간의 중요도 차이를 줄이는 일 없이 정규화한다.

또 위 [그림 1]에서 보듯이 Max TF를 사용할 때보다 Sum TF를 사용할 때 정확도가 평균 1.3% 높다. TF를 Max TF나 Sum TF로 나누어주는 것은 한 문서에서 차지하는 단어의 중요도를 나타내기 위함이다. 따라서 Max TF나 Sum TF는 문서를 대표하는 역할을 한다. 이렇게 대표성 면에서 보면, Sum TF가 Max TF보다 우수하다. Max TF는 Max TF가 아닌 다른 단어의 빈도는 모두 무시하지만 Sum TF는 최대 빈도의 단어뿐만 아니라, 문서에 나타나는 전체단어를 고려하므로 대표성이 크다.

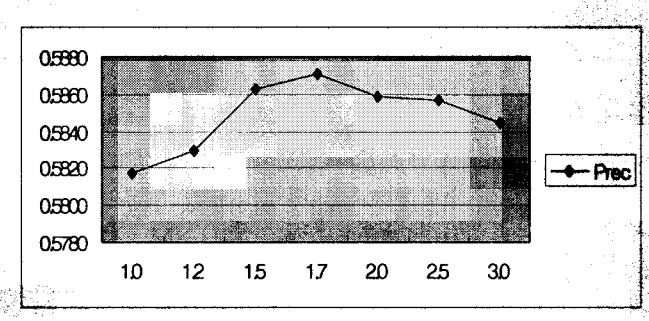
r값은 적용하지 않을 때보다 적용했을 때 정확도가 높다. 이 실험에서 가장 높은 정확도는 Sum TF를 이용하는 Fox weights 방법을 사용하고, r값이 0.1일 때 0.5863이다.

3.2 P값의 부여방법

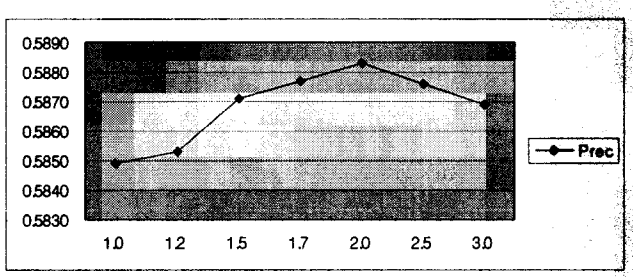
P값은 P-norm 모델의 검색성능을 결정하는 중요한 요소이다. P값은 보통 1.0에서 ∞까지 변화시키는데 1.0에 가까울수록 AND 연산과 OR 연산의 차이가 없어지면서 벡터모델에 가까워지고, ∞에 가까울수록 AND 연산에서는 최소값, OR 연산에서는 최대값을 선택하게 됨으로써 불리언 모델 중의 하나인 Fuzzy Set Model에 가까워진다. 또 AND P값은 비교적 높게 OR P값은 비교적 낮게 설정하는 것이 검색효율을 높인다는 연구결과도 나와 있다[5].

앞의 실험에서는 AND P값과 OR P값을 모두 1.5로 설정하고 실험하였다. 이번 실험에서는 최적화된 P값을 찾아내기 위하여 먼저 AND P값을 1.5에 고정하고, 가장 정확도가 높은 OR P값을 찾아낸 후, 이 OR P값을 고정하고 가장 정확도가 높은 AND P값을 찾아낸다.

실험결과는 다음과 같다.



[그림 2] OR P값의 변화에 따른 정확도의 변화 : x축 - OR P값, y축 - 정확도



[그림 3] AND P값의 변화에 따른 정확도의 변화 : x축 - AND P값, y축 : 정확도

[그림 2]은 AND P값을 1.5에 고정시키고, OR P값을 변화시키면서 정확도의 변화를 관찰한 그래프이다. OR P값이 1.7일 때 정확도가 0.5871로서 가장 높다. [그림 3]는 OR P값을 1.7에 고정시키고 AND P값을 변화시키면서 정확도의 변화를 관찰한 그래프이다. AND P값이 2.0일 때 정확도가 0.5883으로서 가장 높다. 기존 연구와 마찬가지로 본 실험에서도 OR P값보다 AND P값이 클 때 정확도가 더 높았다.

3.3 클로즈의 가중치 부여방법

불리언 질의어에서 AND 클로즈는 피연산자의 개념이 모여 더 강한 개념이 되고, OR 클로즈는 피연산자 각각의 개념보다 더 약한 개념이 된다. 앞의 실험에서 AND 클로즈와 OR 클로즈의 특성을 반영하는 계산법을 적용하기는 했으나, 그것으로는 부족하다. 여기서는 OR 클로즈는 이전과 같이 계산하고 AND 클로즈의 특성을 반영하여 그 가중치를 올려주는 방법에 대해 실험한다. Cornell 대학의 Smith는 AND 클로즈의 가중치를 높여 주는 방법으로 sum-weights 방법과 sum- weights-modified 방법을 제안하였다[4]. 이 두 가지 방법은 다음과 같다.

$Q=(q_1, q_2, \ldots, q_n)$

$D=(d_1, d_2, \ldots, d_n)$

- sum-weights 방법

$Sim(D,Q)=q_1^p d_1^p + q_2^p d_2^p + \cdots + q_n^p d_n^p$

이와 같이 단순히 각 피연산자의 가중치를 더해 준다.

- sum-weights-modified 방법

$Sim(D,Q)=(q_1^p d_1^p + q_2^p d_2^p + \cdots + q_n^p d_n^p) \times k$

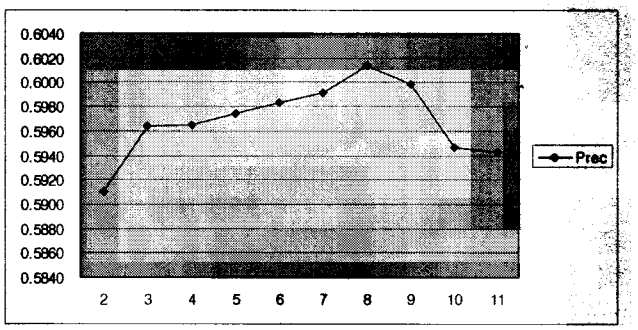
sum-weights 방법으로 계산된 결과를 k배 한다.

- 이 두 가지 수식으로 나오는 결과는 1.0이 넘을 수도 있는데, 이 때는 모두 1.0으로 처리한다. 가중치가 1.0이라는 의미는 완벽히 일치한다하는 말이기 때문이다.

이렇게 하면, 가중치가 낮은 클로즈는 위 수식대로 가중치가 증가하지만, 가중치가 높은 클로즈는 위의 수식대로 가중치가 증가하지 않는다.

앞선 실험에서 Sum TF를 이용하는 Fox weights 방법을 사용하고, r=0.1일 때, 그리고 AND P=2.0, OR P=1.7일 때 가장 좋은 검색결과를 얻었다. 이 파리미터 값을 그대로 이용하면서 위 두 가지 방법을 실험하였다.

sum-weights 방법을 적용하면, 정확도는 0.5891이 되고, sum-weights-modified 방법을 적용하면 다음 [그림 4]와 같다.



[그림 4] sum-weights-modified 방법에서 k값의 변화에 따른 정확도의 변화 : x축 - k값, y축 - 정확도

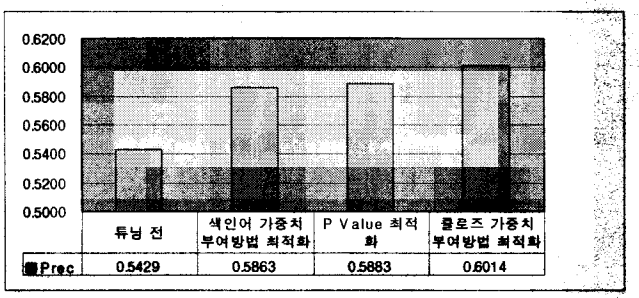
[그림 4]에서 보듯이 k가 8이 될 때까지 정확도는 계속 증가한다. AND 클로즈의 가중치를 높일수록 검색효율이 증가한다는 뜻이다. k가 8을 넘어서면 정확도가 떨어지기 시작하는데, 이유는 가중치를 증가시켰을 때 1.0에 가까워지거나 1.0을 넘어서는 경우가 많기 때문이다. 이 실험에서 가장 높은 정확도는 k값이 8인 sum-weights-modifed 방법에서 0.6014이다. 이는 튜닝 전 Max TF를 이용하는 Fox weights 방법을 이용할 때의 정확도 0.5429에 비해 10.8% 증가한 것이다.

3.4 실험 결과

지금까지 P-norm 모델을 구현하면서 고려해야할 사항들을 살펴보고, 검색효율을 향상시키기 위한 다양한 방법을 실험하였다. 실험을 통하여 [표 1]과 같은 최적화방법을 찾아내었으며, 이 방법으로 [그림 5]과 같이 검색효율을 향상시켰다.

P-norm 모델 구현 시 고려해야 할 사항	실험 결과
색인어 가중치 부여방법	Sum TF를 이용하는 Fox weights 방법을 적용하고, r=1일 때 검색효율 최대.
P값 부여방법	AND P=2.0, OR P=1.7일 때 검색효율 최대.
클로즈 가중치 부여방법	k = 8인 sum-weights-modified 방법을 적용할 때 검색효율 최대.

[표 1] 실험 결과(구현방법 최적화)



[그림 5] 실험 결과(검색 정확도 향상)

[그림 5]에서 보듯이 검색효율에 영향을 미치는 여러 요소들 중 가장 중요한 요소는 색인어 가중치 부여방법이다. 그 다음이 클로즈 가중치 부여방법으로 나타났고, P값의 최적화는 검색효율에 크게 영향을 미치지 못하였다.

그리고, 최적화하지 않았을 때에 비해 최적화를 수행함으로써 정확도가 10.8% 증가하여 최종 정확도는 0.6014가 되었다.

4. 결론

본 논문에서는 정보검색 시스템에서 P-norm 모델을 구현하고, 검색성능을 최적화하였다.

P-norm 모델에서 검색효율에 영향을 미치는 3가지 요소를 구분하고, 각 요소별로 최적화 방법을 제시하고, 실험을 통하여 실제 시스템에서 가장 효율적인 수치를 구하였다. 결과적으로 최적화 전보다 10.8%의 정확도 향상을 이루었으며, 검색효율에 가장 크게 영향을 미치는 요소도 찾아내었다.

P-norm 모델에서 향후 연구방향은 다음과 같다.

- P-norm 모델은 검색효율은 좋지만, 많은 계산량으로 인해 검색속도가 떨어지는 것으로 알려져 있다. 검색속도 향상방안을 마련하여야 한다.
- 본 논문의 실험결과에서 보다시피 색인어 가중치 부여방법이 검색효율을 결정하는 가장 중요한 요소이다. 새로운 색인어 가중치 부여방법을 개발하고 실험하여야 한다.
- 검색 효율을 향상시키는 또 다른 방법으로 적합성 피드백이 있다. P-norm 모델에서 적합성 피드백을 최적화하여야 한다.

참고 문헌

[1] Joon Ho Lee, "Properties of Extended Boolean Models in Information Retrieval", SIGIR '94 Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 182-190, 1994

[2] Gerard Salton, Ellen Voorhees. "Automatic assignment of soft boolean operators", In Proceedings of the Eighth Annual SIGIR Conference, pp. 54-59, 1985

[3] Edward Alan Fox, "Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types", Dissertation, Cornell University, 1983

[4] Maria Elena Smith, "Aspects of the P-norm Model of Information Retrieval : Syntactic Query Generation, Efficiency, and Theoretical Properties", Dissertation, Cornell University, 1990

[5] Gerard Salton, Edward A. Fox and Harry Wu, "Extended Boolean Information Retrieval", ncstrl.cornell/TR82-511, 1982

[6] Gerard Salton, Michael J. McGill, "Introduction to Modern Information Retrieval", McGrawHill, 1983

[7] Robert R. Korfhage, "Information Storage and Retrieval", Wiley Computer Publishing, 1997

[8] Ian H. Witten, Alistair Moffat, Timothy C. Bell, "Managing Gigabytes", Van Nostrand Reinhold, 1994

[9] 고미영, "P-norm 검색의 문헌 순위화 기법에 관한 실험적 연구", 박사학위논문 연세대학교 문헌정보학과, 1999

[10] 이효숙, "적합성 가중치 검색 및 P-norm 검색에 관한 연구", 정보관리학회지 11권 1호(통권 20호), p31-56, 1994

[11] 정영미, "정보검색론", 구미무역(주) 출판부, 1993