적합성 가중치 검색 및 P-NORM 검색에 관한 연구

-불 논리 검색의 개선을 중심으로-

A Comparative Analysis of the Relevance Weighted Boolean Model and the P-NORM Model: An Improvement on the Boolean Retrieval

이 효 숙 (Hyo Sook Lee)*

□ 목 차 □

초 록

본 연구에서는 검색실험을 통하여 질문 변환에 의한 불 논리 검색, 적합성 가중치 검색, P-NORM 검색에 대해 평가하였다. 적합성 가중치 검색은 질문 변환에 의한 불 논리 검색 및 P-NORM 검색보다 정확률과 검색순위에 있어 효과적이었다. 정보 탐색과정에서 적합성 정보의 이용수준과 용어에 대한 가중치방법은 검색성능에 영향을 주는 것으로 밝혀졌다.

ABSTRACT

To evaluate the retrieval effectiveness of the Boolean Request Conversion Model, the Relevance Weighted Boolean Model, and the P-NORM Model, the present study has been done with experimental tests. It is proven that the Relevance Weighted Boolean Model is more effective in precision and the document output ranks than the other ones. The experimental results indicate a promising application of relevance information and weighting schemes.

* 한남대학교 문헌정보학과 강사

I. 서 론

1.1 연구의 목적

본 연구에서는 불 논리 검색의 개선을 위해 개별적으로 연구된 검색방법들로서 질문 변환에 의한 불 논리 검색, 적합성 가중치 검색 및 P-NORM 검색에 대하여 연구하고자 한다. 각 검색모형에 대해 검색실험의 실시 및 평가를 통하여 가장 우수한 검색모형을 밝혀내고, 각각의 검색모형에서 검색성능에 영향을 미치고 있는 요인들을 규명한다.

1.2 가 설

본 논문에서 검색성능은 검색시스템에서의 검색효율과 검색문헌을 질문과의 유사성 순위대로 출력시킴으로써 이용자의 노력을 감소시켜주는 시스템의 능력을 나타내며, 검색성능의 척도로서는 정확률과 검색순위를 사용하였다. 검색순위는 시스템에서 예측된 적합성 정도에 따라 검색된 적합문헌이 부적합문헌보다 먼저 제공되도록 검색시에 문헌에 부여되는 순위를 의미한다.

검색실험을 통하여 검증할 연구가설은 다음과 같다.

가설 1. 적합성 가중치 검색과 P-NORM 검색은 질문 변환에 의한 불 논리 검색보다 정확률이 더 높고, 이용자 판정에 의한 검색순위와 더 높은 상관관계가 있다.

가설 2. 적합성 가중치 검색은 P-NORM 검색보다 정확률이 더 높고, 이용자 판정에 의한 검색순위와 더 높은 상관관계가 있다.

가설 3. 적합성 가중치 검색에서 완전한 적합성 정보를 이용한 검색방법은 소수의 적합성 정보를 이용한 검색방법보다 정확률이 더 높고, 이용자 판정에 의한 검색순위와 더 높은 상관관계가 있다.

가설 4. P-NORM 검색에서 탐색어에 적합성 가중치를 적용한 검색방법은 역문헌빈도 가중치를 적용한 검색방법보다 정확률이 더 높고, 이용자 판정에 의한 검색순위와 더 높은 상관관계가 있다.

1.3 연구의 방법 및 범위

실험연구에서 사용할 검색시스템, 데이터베이스 및 질문과 검색실험에 참여할 이용자는 다음과 같다.

첫째, 정보검색은 전기·전자분야를 대상으로 한다. 검색시스템은 STAIRS와 KIROS를 사용하고, 탐색될 데이터베이스는 INSPEC과 BIST이다.

둘째, 이용자의 탐색질문으로부터 작성된 66개의 탐색식(한글 33개, 영문 33개)으로 탐색을 실시한다. 적합성 가중치 검색에서는 검색결과에 대해 적합성 판정을 받은 후, 적합성 가중치를 이용하여 문헌은 이용자에게 순서적으로 제공된다.

셋째, P-NORM 검색에서는 INSPEC과 BIST로부터 다운로드한 실험집단(영문문헌 2,824건, 한글문헌 1,416건)에 대해 실험시스템에서 탐색을 실시한다. 시스템은 질문 처리, 검색, 연산, 입력 및 출력 등의 주요 기능을 갖는다. 탐색어

에 가중치, 논리연산자에 p 값을 부여하여 탐색한 후 유사도의 내림차순으로 문헌이 제공된다.

넷째, 이용자는 전기·전자 분야에서 석사학위 취득 후 해당분야에서 3년 이상 연구에 종사한 자에 대해 유층표집한 자 33명으로 이루어졌다. 그 구성은 대학 연구소의 연구원으로 박사학위 과정 중에 있는 자 17명, 정부 출연기관 및 기업체 연구소의 연구원으로 박사학위 과정에 있거나 전공분야에서 연구에만 종사해 온 연구자 8명, 그리고 정부 출연기관의 연구원 및 대학교수로서 박사학위 취득자 8명이다.

본 연구는 다음과 같은 연구범위 내에서 실시되었으며, 연구의 결론은 이 범위 내에서 타당성을 갖는다.

첫째, 탐색은 한글 및 영문 데이터베이스에 대해 free text 탐색을 실시하였다.

둘째, 문헌의 색인어에 대해서는 가중치를 부여하지 않았고, 질문의 탐색어에 대해 적합성 가중치 검색에서 적합성 가중치, P-NORM 검색에서 역문헌빈도 가중치 및 적합성 가중치를 각각 사용하였다.

셋째, 적합성 가중치 검색에서는 적합 및 부적합문헌에서 용어의 출현특성을 적용하였고, P-NORM 검색에서는 용어의 빈도 정보, 용어와 용어구 간의 관계 및 논리연산자에 파라미터 값 등을 사용하였다.

II. 이론적 배경

2.1 불 논리 검색의 개선

정보검색 분야에서는 불 논리를 적용하는 검색환경에서 검색성능을 개선하기 위해 탐색어에 가중치를 적용하거나 또는 불 논리를 확장하여 검색결과를 순위화하는 문제에 관심을 기울여 왔다.

정보검색에서 문헌의 순위화는 북스타인과 쿠퍼(A. Bookstein & W.S. Cooper, 1976)가 정보검색시스템을 구조적 모형으로서 설명한 이래 보다 구체적으로 논의되었다. 이들의 이론은 정보검색시스템의 중심기능이란 탐색시 문헌집단 내의 각 문헌들을 구분하고 문헌집단에 대해서 랜덤탐색 보다는 가능한 한 순서적으로 접근하도록 하는 것임을 지적한 것이다.

로버트슨(Robertson, 1977a)은 문헌의 순위화를 적합성의 본질적 특성에 기초를 두고 다음과 같이 두 가지 관점에서 설명하였다.

첫째, 시스템에서 검색된 문헌들을 순서적으로 제공함으로서 탐색신청자가 이에 따라 순서적으로 살펴볼 수 있도록 하여야 한다. 이것은 탐색자가 어디까지 탐색을 계속하여야 할 것인가를 결정하는 문제에 있어 시스템이 이를 지원한다는 점에 중요성을 두며, 정보검색기능의 확률적 특성과 관련한다. 둘째, 검색시스템에서 문헌을 순위화하는 이유는 적합성이란 연속성의 성질을 가지므로 시스템에서 적합문헌은 이보다 덜 적합한 문헌보다 먼저 제공되어야 하기 때문이다.

로버트슨은 문헌의 확률 순위화 원칙(the probability ranking principle)에 관한 그의 연구(1977b)에서 적합성 확률과 적합성 정도의 개념이 결합될 수 있는 최적의 순위화 원칙이 필요하고 이에 대한 해결은 문헌과 탐색신청자의 요구와의 관계를 연구하는 것으로 부터 시작될 수 있다고 보았다.

스파크 존스(K. Sparck-Jones, 1979b)는 탐색어에 가중치를 부여하는 문제와 관련하여 가장 효과적인 적합성 가중치를 규명하고 이와 관련된 일반적 가정을 밝혀내었다. 그리고 후속연구에서 탐색환경이 다를 때 적합성 가중치의 효과는 탐색에서 어떤 역할을 하며 어떤 차이를 보이는 가에 대해서 집중적으로 조사한 결과를 보고하였다. 밴 리즈버겐(C.J. van Rijsbergen, 1977)은 문헌에서 용어가 독립적으로 출현하는 것으로 가정하는 것은 수학적 근사화의 과정을 편리하게 하기 위한 것이며, 실제로 문헌에서 용어의 출현은 통계적 의존관계에 있는 것으로 보았다.

용어 간의 의존관계를 가정할 때 용어의 동시출현 확률분포에 따라 문헌의 확률분포함수 $P_t(X)$에 대한 근사화는 달라진다. 문헌 X에서 용어 x의 동시출현에 의한 확률분포함수는

$P_t(X) \doteq \prod_{i=1}^{n} P(x_{mi}|x_{mj(i)})$가 되고 x_{mi}와 $x_{mj(i)}$는 문헌 X에서 통계적 의존관계를 갖고 출현하는 용어들이다. 문헌 X가 용어벡터 형태로서 $X = (x_1, x_2, x_3, x_4)$ 일 때, 용어 간의 관계에 따른 문헌 X의 확률분포함수 $P_t(X)$는 $P_t(X) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)$가 된다. 문헌에서 용어 x_1, x_2, x_3, x_4 간의 의존관계가 다를 때 문헌 X의 확률분포함수는 다르게 표현된다.

밴 리즈버겐은 용어 간의 의존도를 트리구조로 보고, 용어 쌍에 의한 관계를 검색에서 이용하기 위해서 용어 간의 관계를 최적화하는 트리 구조의 구축과 이를 위한 계산방법을 연구하였다. 이를 위해서 최대신장트리(maximum spanning tree)를 이용할 것을 제안하였고, 최대신장트리는 용어와 용어 간의 의존도의 합인 $\sum_{i=1}^{n} I(x_i, x_{j(i)})$가 최대가 되는 경우이다. 이같은 용어 의존도 정보를 확률검색에서 문헌값을 결정하는 분리함수에 적용할 것을 제시하였다(van Rijsbergen, 1979:122-124).

불 논리 검색을 개선하기 위해 딜론과 데스퍼(M. Dillon & J. Desper, 1980)는 탐색어에 가중치를 주는 문제를 다루었다. 이들은 용어의 적합 및 부적합 정도를 구별하기 위해서 이용자의 피이드백 정보를 이용하여 용어 측정기준이 되는 가중치 방법을 유도하였다. 용어의 최종가중치(prevalence)는 양의 가중치와 음의 가중치와의 차이에 의한 값을 적용한다. 그리고 가중치가 높은 용어들이 포함된 문헌은 검색될 수 있도록 하고, 가중치가 낮은 용어에 관한 문헌은 검색에서 보다 제한되도록 하기 위해서 불 논리연산자를 이러한 목적에 맞게 사용하도록 하는 방법을 제시하였다.

라데키(T. Radecki, 1988)는 불 논리 검색에서 적합문헌이 다수의 부적합문헌과 함께 동시에 제공되는 점을 개선하기 위해서 불 대수의 논리합 정규형(disjunctive normal form)의 응용과 확률검색 이론에 기초하여 검색된 문헌에 대해 출력순위를 줄 수 있는 방법을 제안하였다. 그의 이론적 전개는 크게 두 가지 부분으로 구성된다. 첫째 단계는 검색시스템에서 탐색식을 발전된 논리합 정규형(developed DNF)으로 변환하고, 둘째 단계에서는 변환된 탐색식으로 탐색한 후 탐색어의 적합성 가중치를 계산하여 적합성 가중치합에 따라 문헌을 순위화하도록 하였다.

2.2 검색모형

2.2.1 질문 변환에 의한 불 논리 검색모형

불 논리에 의한 탐색식은 논리적으로 항등한 다른 형태의 탐색식으로 변환할 수 있다(A. Bookstein, 1978). E_i가 불 논리에 의해 논리적으로 타당한 탐색식일 때 하나의 탐색식은 E_1, E_2,.....E_n으로 이루어지고 이들 간에 논리합 관계를 갖는 형태를 이룬다. 그리고 바로 이러한 형태의 탐색식을 문헌의 순위화에서 이용한다. 이때 각 E_i에 속하는 탐색어 간의 논리관계는 AND와 NOT의 관계만을 갖는다. 탐색어 a, b, c로 구성된 탐색식이 b AND NOT (c AND a) 라고 할 때 이것은 (b AND NOT c) OR (b AND NOT a)로 변환된다.

논리합으로 변환된 질문식에 의한 검색결과는 적어도 논리합 관계에 있는 각 탐색식 가운데 최소한 한 개 이상이 '진'이 되는 문헌이 검색되게 된다. 검색시스템에서는 각 탐색식에 대해서 문헌이 논리적으로 '진'에 해당하면 1의 값을 가지며 '위'에 해당하면 0의 값을 갖는다. 결과적으로 불 논리 검색결과는 논리합으로 연결된 탐색식의 수를 가장 많이 만족시키는 문헌에서 부터 한 개만 만족시키는 문헌까지 내림차순으로 순위화 할 수 있다. 북스타인이 제시한 이 방법은 검색시스템에서 각 탐색어와 관련된 문헌집합들 간에 반순서($\leq$) 관계에 있는 점이 응용되어 검색된 문헌들을 순위화하는 것이다.

2.2.2 적합성 가중치 검색모형

적합성 가중치 함수는 정보검색시스템에서 문헌의 총수(N), 적합문헌수(R), 색인어가 출현한 문헌수(n), 색인어가 출현한 적합문헌수(r) 등으로 구성된다. 적합문헌과 부적합문헌에서 용어는 다른 용어와의 관계에서 독립적으로 출현하는 것으로 가정하며, 용어가 적합문헌에 출현한 경우와 부적합문헌에 출현한 경우를 모두 고려한 가중치 함수가 가장 효과적인 것으로 보고되었다(Robertson & Sparck-Jones, 1976: 131).

적합성 가중치 검색에서는 결정이론(D.H. Kraft, 1973)을 기초로 위험 부담을 최소화하는 검색기준을 정함으로써 검색된 문헌을 순위화하는 방식을 사용하고 있다. 그리고 적합문헌과 부적합문헌의 분리함수로 공식 (1) 이 제안되었다(Radecki, 1988b).

$$g(X_d) = \log \sum_{i=1}^{k} \left[X_{d(i)} \log \frac{\xi_i}{\eta_i} + (1 - X_{d(i)}) \log \frac{1 - \xi_i}{1 - \eta_i} \right] + \log \frac{P(W_1)}{P(W_2)} + \log \frac{\lambda_{21} - \lambda_{11}}{\lambda_{12} - \lambda_{22}} \quad \cdots\cdots\cdots\cdots\cdots (1)$$

X_d : 색인어벡터

$X_{d(i)}$: i 번째 색인어

η_i : 색인어가 부적합문헌에서 출현할 확률 $P(X_{d(i)} = 1 \mid W_2)$

ξ_i : 색인어가 적합문헌에서 출현할 확률 $P(X_{d(i)} = 1 \mid W_1)$

$X_{d(i)} = 1$ 이면 i번째 색인어가 존재하는 경우이다.

$1 - \xi_i$: 색인어가 적합문헌에서 출현하지 않을 확률

$1 - \eta_i$: 색인어가 부적합문헌에서 출현하지 않을 확률

$P(W_1)$: 적합문헌이 될 확률

$P(W_2)$: 부적합문헌이 될 확률

λ_{21} : 적합문헌을 부적합문헌으로 결정했을 때 예상되는 손실

λ_{12} : 부적합문헌을 적합문헌으로 결정했을 때 예상되는 손실

λ_{11} : 옳은 결정으로서 적합문헌을 적합문헌으로 결정했을 때의 손실 ($\lambda_{11}=0$)

λ_{22} : 옳은 결정으로서 부적합문헌을 부적합문헌으로 결정했을 때의 손실 ($\lambda_{22}=0$)

적합성 확률에 의한 출력문헌의 상대적 위치는 문헌에 질문과 일치하는 색인어가 있는 경우($X_d(i)=1$)에 색인어의 가중치합에 따라 결정된다. 색인어가 적합문헌 및 부적합 문헌에서 각각 출현할 확률의 계산은 표본문헌 집합에서 i번째 색인어의 출현특성을 이용한다. 분리함수 $g(X_d)$는 공식 (2)와 같으며, 이것을 기초로 질문에 대한 검색문헌의 적합성을 산출한다.

$$g(X_d)=\sum_{i=1}^{k} X_{d(i)}\log\frac{r_i/(R-r_i)}{(n_i-r_i)/(N-n_i-R+r_i)}+C \quad \cdots\cdots (2)$$

2.2.3 P-NORM 검색모형

문헌공간 상에서 각 지점들 간의 거리개념은 n 차원의 벡터로 정의되고, 각 지점 간의 거리는 노름(norm)으로 설명될 수 있다. 이를 공식으로 표현하면 다음과 같다.

$$\| d \|_p = \| d_1, d_2, \ldots, d_n \|_p = (d_1^p + d_2^p + \ldots + d_n^p)^{1/p}$$

질문과 문헌 간의 유사도를 측정하기 위해서는 문헌길이를 표준화한 값이 사용되어야 하므로 벡터노름은 $\| d \|_p = \left[\frac{(d_1^p + d_2^p + \ldots + d_n^p)}{n}\right]^{1/p}$ 가 된다. 불 논리 검색에서 질문에 대한 문헌의 관련성 정도를 비교하기 위해서 이와 같은 벡터노름에 의한 값을 사용한다.

어떤 문헌 D에서 용어 A_i의 가중치가 dA_i 이고, dA_i가 0에서 1 사이의 값을 가질 때 문헌 D는 $D = \{dA_1, dA_2, \ldots dA_n\}$ 이다. 그리고 질문 Q는 용어 간의 관계가 합집합 관계에 있을 때와 교집합 관계에 있을 때에 다음과 같이 표현된다.

$Q_{or(p)} = \{(A_1, a_1)\ or^p (A_2, a_2)\ or^p \ldots or^p (A_n, a_n)\}$

$Q_{and(p)} = \{(A_1, a_1) and^p (A_2, a_2) and^p \ldots and^p (A_n, a_n)\}$

여기에서 a_i는 질문 상의 용어 A_i가 갖는 중요도에 따라 가중치 $0 \le a_i \le 1$의 값을 나타내며 p 값은 1에서 ∞사이의 값을 갖는다. 용어들의 집합 $\{A_1, A_2, \ldots A_n\}$에 대하여 문헌 D와 질문 Q간의 유사도는 공식 (3), (4)와 같이 정의된다 (G. Salton, E.A. Fox, & H. Wu, 1983: 1025).

$$Sim(D, Q_{or(p)}) = \left[\frac{a_1^p d_{A1}^p + a_2^p d_{A2}^p + \ldots + a_n^p d_{An}^p}{a_1^p + a_2^p + \ldots + a_n^p}\right]^{1/p} \quad \cdots\cdots (3)$$

$$Sim(D, Q_{and(p)}) = 1 - \left[\frac{a_1^p (1-d_{A1})^p + a_2^p (1-d_{A2})^p + \ldots + a_n^p (1-d_{An})^p}{a_1^p + a_2^p + \ldots + a_n^p}\right]^{1/p} \quad \cdots\cdots (4)$$

문헌 D가 용어 A, E, F에 의해 표현되고, 용어 A, E, F에 대해서 각각 a, e, f의 가중치를 갖는 질문식이 {(A, a) or (b [(E, e) and (F, f)])}

로서 작성되었다고 하자. 그리고 b는 용어 E와 용어 F 간의 관계에서 용어구에 부여되는 가중치일 때, 용어구가 포함된 질문에 대해 계산되는 질문과 문헌의 유사도는 공식 (5)와 같다.

$$Sim(D, Q) = \left\{\frac{a^p d_A^p + b^p[1-((e^p(1-d_E)^p + f^p(1-d_F)^p) - /(e^p+f^p))^{1/p}]^p}{a^p + b^p}\right\}^{1/p}$$

..... (5)

검색에서 두 용어 간의 관계정도를 표현하기 위해 파라미터로서 p 값은 $p=1, 1 \leq p \leq \infty$, $p = \infty$ 등으로 사용한다. 질문에서 p 값은 서로 다른 값을 사용하여 용어 간의 관계를 엄격히 지키는 경우와 다소 완화된 관계를 유지하는 경우 등을 검색에서 적용할 수 있다.

III. 검색실험

본 장에서는 적합성 가중치 검색과 P-NORM 검색에서 실시한 검색과정과 그 내용을 기술한다.

3.1 적합성 가중치 검색

3.1.1 검색문헌의 적합성 판정

적합성 가중치 검색에서는 정보탐색 평가서와 함께 검색된 문헌들을 이용자에게 보내어 적합성 여부를 판정하도록 하였다. 적합성 판정 기준은 '적합', '부분적합', '부적합'으로 하였다. 적합성 가중치 검색에 의해 검색된 문헌수와 적합문헌수는 〈표-1〉, 〈표-2〉와 같다.

〈표 1〉 불 논리 검색 및 적합성 가중치 검색결과(한글)

질문번호	검색된 문헌수	적합문헌수
1	17	15
2	2	1
3	34	14
4	8	7
5	12	9
6	35	8
7	64	31
8	5	5
9	3	3
10	11	4
11	7	4
12	43	13
13	20	8
14	30	24
15	5	3
16	5	3
17	10	2
18	13	9
19	24	12
20	24	20
21	52	25
22	10	2
23	18	8
24	40	27
25	7	6
26	13	8
27	3	2
28	12	12
29	11	3
30	7	3
31	7	4
32	4	2
33	8	1
평 균	17.09	9.03

〈표 2〉 불 논리 검색 및 적합성 가중치 검색결과(영문)

질문번호	검색된 문헌수	적합문헌수
1	57	25
2	31	29
3	48	23
4	43	38
5	11	6
6	57	46
7	31	19
8	21	19
9	8	6
10	41	33
11	14	11
12	20	17
13	5	5
14	37	11
15	2	2
16	30	18
17	17	7
18	15	11
19	17	16
20	33	27
21	44	21
22	35	21
23	132	75
24	3	2
25	21	13
26	30	24
27	79	55
28	78	31
29	31	26
30	31	24
31	25	8
32	41	32
33	8	5
평 균	33.21	21.39

3.1.2 검색결과의 순위화

적합성 가중치 검색에서 탐색식은 불 논리 관계에 의한 탐색어들로 구성되고, 탐색어가 문헌에 출현한 경우와 문헌에 출현하지 않은 경우를 기초로 하여 탐색식에 일치하는 문헌이 검색된다. 탐색어를 포함하는 문헌수에 따라 동일한 탐색식에 대해 1건 이상의 문헌이 검색될 수 있으며, 검색문헌의 출력순위는 탐색어의 적합성 가중치합에 의해 결정된다.

탐색식에 포함되지 않은 탐색어에 대한 가중치는 고려되지 않으며, 탐색어의 가중치는 완전한 적합성 정보를 이용하는 경우와 소수 문헌에 의해 적합성을 예측하는 경우로 구분하였다. 전자의 경우는 검색된 문헌 모두에 대해서 이용자로부터 판정을 받은 후 이것을 기초로 적합성 가중치를 산출하였고, 탐색어의 적합성 가중치합에 따라 검색결과를 순위화하였다. 후자의 경우는 검색된 문헌 가운데 5건의 문헌에 대한 적합성 정보를 사용하여 검색결과를 순위화하였다. 적합성 가중치 산출과 검색문헌의 순위화 과정은 다음과 같다.

질문 1: '실리콘 및 화합물반도체에서 센서와 변환기 공정'

탐색식: 1. "SI" OR 실리콘 OR (화합물 ADJ 반도체)

2. 센서 OR 변환기 OR 구동$1

3. 1 AND 2

적합문헌수는 세 가지 검색모형에 의해 검색된 문헌들에 대해 이용자가 적합한 것으로 판정한 문헌수를 사용하였다. 질문 1의 탐색어 '실리콘'에 대해 적합문헌수(R), 용어가 출현한 적합문헌수(r), 용어가 출현한 문헌수(n), 데이

터베이스 내의 전체문헌수(N)는 $R_{실리콘}=15$, $r_{실리콘}=15$, $n_{실리콘}=858$, $n_{실리콘}=23,500$ 이며 이를 2×2 분할표로 작성한다.

	W_1	W_2	
X_d(실리콘)=1	r	N-r	n
X_d(실리콘)=0	R-r	N-n-R+r	N-n
	R	N-R	N

$$\Rightarrow$$

	W_1	W_2
X_d(실리콘)=1	15	843
X_d(실리콘)=0	0	22,642

'실리콘'의 적합성 가중치는

$$W_{실리콘}=\log \frac{r_{실리콘}/(R-r_{실리콘})}{(n_{실리콘}-r_{실리콘})/(N-n_{실리콘}-R+r_{실리콘})}=6.724$$

이다.

적합성 가중치 공식은 각 항에 0.5를 더해준 공식을 사용하였다. 질문 1의 탐색식 가운데 '실리콘'외의 나머지 탐색어 '센서', '변환기', '구동$1'에 대한 적합성 가중치도 계산한다.

$W_{센서}$ = 5.555

$W_{변환기}$ = 2.735

$W_{구동기}$ = 6.542

질문 1에 대해서 검색문헌 Di의 순위를 결정하는 값 R_1, R_2, R_3, R_4은 다음과 같이 된다.

$R_1 = W_{실리콘} + W_{센서} = 12.279$

$R_2 = W_{실리콘} + W_{센서} + W_{변환기} = 15.014$

$R_3 = W_{실리콘} + W_{변환기} = 9.459$

$R_4 = W_{실리콘} + W_{구동기} = 13.266$

R_1, R_2, R_3, R_4에 해당하는 문헌이 D_1, D_2, D_3, D_4 일 때 질문 1에서 검색된 문헌이 이용자에게 제공되는 순서는 D_2, D_4, D_1, D_3의 순이 된다. 순위화된 문헌들은 검색순위에 대한 평가서와 함께 이용자에게 제공되었다.

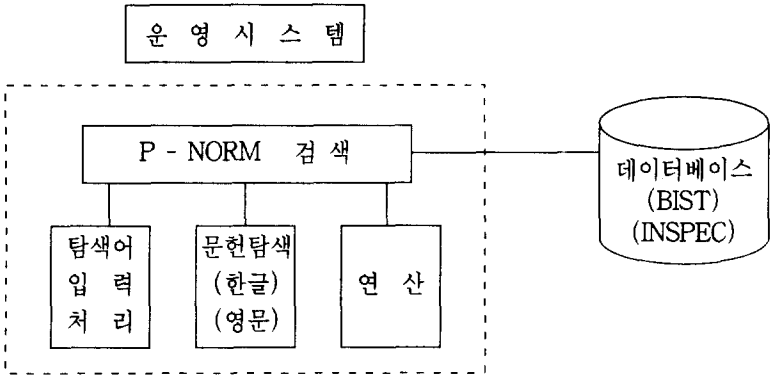
3.2 P-NORM 검색

3.2.1 데이터베이스 다운로드

데이터베이스의 다운로드를 목적으로 본 연구자가 1993년 5월 18일~1993년 5월 30일의 기간에 한글 및 영문 탐색식을 시스템의 공동 사용장소(common pool)에 저장한 후, 검색결과를 다운로드하였다. 데이터 화일은 ISO 2709 데이터교환형식에 의해 ISO 화일로 작성된 것으로 P-NORM 검색에서 사용하기 위해서 각 레코드에 지정된 태그에 따라 데이터를 변환하였다.

3.2.2 시스템 설계 및 구현

P-NORM 검색을 실시하기 위해 Clipper 언어로서 실험시스템을 구현하였으며, 시스템 구성도는 〈그림-1〉과 같다. 시스템에서는 텍스트화일, 서지정보화일, 사전화일, 키워드화일, 인덱스화일이 필요한 경우에 조회 및 결합된다. 실험시스템에서 탐색식 작성시에 탐색어와 탐색어의 가중치는 새로 추가할 수 있는 기능을 가지며, 검색시에 용어의 통제 및 확대를 위해서는 사전화일에서 해당 용어와 관련된 모든 용어를 조회한 후 검색이 실시된다. 검색방법은 한글과 영문 모두 단어 검색 및 해당문자가 있는 문자열 검색을 하여 탐색어가 출현한 모든 문헌이 검색되도록 하였다. 텍스트화일은 다운



〈그림 1〉 P-NORM 검색시스템 구성도

로드 받은 실제 데이터가 있는 화일로서 시스템에서 검색된 결과에 대해 질문과 문헌의 유사도 값의 내림차순으로 리스트가 생성된 후 원문의 출력시에 사용된다.

3.2.3 문헌검색

탐색어의 가중치로서 적합성 가중치와 역문헌빈도 가중치를 각각 적용하여 검색을 실시하였다. 탐색어의 역문헌빈도 값을 산출하기 위해 사용한 공식은 $-\log_2(n/N)$으로 n은 특정한 단어 t가 출현한 문헌수, N은 문헌집단 내 전체문헌수를 나타낸다. 중간 빈도의 탐색어에 대해서는 논리연산자의 P 값을 1.5 로 주고, 매우 낮은 출현빈도의 탐색어와 관련된 논리연산자는 P 값으로 2 를 주어 탐색식을 구성하였다. 질문에 의한 탐색식 작성과 검색과정은 다음과 같다.

① ② ③ ④ ⑤

탐색식: (Space, 3.543 or 〈1.5〉 Vector, 3.884) 3.694 and 〈2〉 (Multiple Access, 10.514) and 〈2〉 (Carrier, 7.684) and 〈2〉 (Protocol $1, 5.048 or 〈2〉 Communication, 4.620) 4.834

탐색식 구성요소:
① 탐색어
② 탐색어의 가중치
③ 불 논리 연산자
④ 불 논리 연산자에 부여된 P 값
⑤ 'Space' 와 'Vector' 의 가중치 평균값

연산과정:
1. 각 탐색어에 대한 문헌검색.
2. 탐색어의 가중치, 불 논리연산자 및 P 값 입력.
3. 불 논리 연산자 AND, OR 에 따른 유사계수 공식 적용.
4. 유사계수 공식에서 용어가 출현한 문헌일 때 문헌은 1의 값을 가지며 용어가 출현하지 않은 문헌이면 0의 값을 갖는다.
5. 탐색식의 첫단계(용어와 용어의 관계)에 대한 각 문헌의 문헌값을 산출한다.
6. 탐색식 내에 용어와 용어구, 용어구와 용어구 간의 연산과정이 있

는지 조사한다.

7. 연산과정이 계속되어야 할 탐색식에서는 전단계에서 산출된 용어 간의 가중치 평균값이 용어구의 가중치가 되며, 전단계에서 산출된 문헌값은 유사계수 공식상에서 문헌값 d_A, d_B로 되어 연산은 순환적으로 이루어진다.

기준치로 유사도가 0.2 이상인 문헌들만이 출력되도록 하였으나 검색문헌수가 많은 경우는 탐색어의 수와 내용에서 탐색질문의 주제를 최소한으로 반영할 수 있는 부분까지 검색하여 출력한다는 원칙을 세웠다.

한글질문 및 영문질문에 대한 P-NORM 검색결과와 이용자로부터 적합성 판정에 의한 적합문헌수는 〈표-3〉, 〈표-4〉와 같다. 적합문헌수는 P-NORM 검색에서 탐색어에 적합성 가중치를 사용한 경우와 역문헌빈도 가중치를 사용한 경우에 검색된 적합문헌수의 합이다.

IV. 분석 및 논의

본 장에서는 검색실험의 결과를 기술통계 및 유의도 검증에 의해 분석한다.

4.1 가설 1과 가설 2의 검증

각 검색모형에 의해 검색된 문헌수는 〈표-5〉와 같다. 불 논리 검색과 적합성 가중치 검색에서는 한글문헌 검색의 경우 질문당 평균 17.09건, 영문문헌 검색에서는 33.21건이 검색되었다.

〈표 3〉 P-NORM 검색결과(한글)

질문번호	검색된 문헌수		적합문헌수
	적합성 가중치	역문헌빈도 가중치	
1	17	17	15
2	8	8	3
3	33	34	14
4	15	15	8
5	77	77	29
6	60	60	15
7	82	82	32
8	21	21	9
9	42	42	19
10	13	13	6
11	7	7	4
12	65	75	16
13	16	20	8
14	51	51	41
15	27	27	17
16	32	25	10
17	10	10	2
18	30	30	12
19	27	27	14
20	49	49	25
21	52	52	25
22	29	29	9
23	18	18	8
24	49	49	36
25	32	36	13
26	41	41	27
27	11	11	4
28	29	29	17
29	35	35	7
30	6	8	3
31	18	18	5
32	9	13	7
33	33	8	1
합 계	1044	1037	461
평 균	31.64	31.42	13.97

〈표 4〉 P-NORM 검색결과(영문)

질문번호	검색된 문헌수		적합문헌수
	적합성 가중치	역문헌빈도 가중치	
1	28	28	25
2	31	31	29
3	38	38	23
4	43	43	38
5	11	11	6
6	71	71	51
7	43	46	22
8	77	77	35
9	47	47	18
10	76	86	58
11	84	84	38
12	40	38	26
13	32	32	25
14	75	74	27
15	11	11	8
16	30	30	18
17	25	35	8
18	21	21	12
19	36	52	31
20	33	33	27
21	44	44	21
22	57	57	29
23	158	157	83
24	46	45	27
25	46	46	29
26	97	97	78
27	109	110	67
28	101	101	35
29	31	31	26
30	40	40	27
31	42	42	10
32	80	80	37
33	29	29	13
합 계	1732	1767	1007
평 균	52.48	53.54	30.52

P-NORM 검색에서는 한글문헌 검색에서 31.64건, 영문문헌 검색에서 52.48건으로 적합성 가중치 검색과 불 논리 검색에 비해 한글은 85.13%, 영문은 58.02%의 문헌이 더 검색되었다.

검색된 적합문헌수는 〈표-6〉에서와 같이 질문 변환에 의한 불 논리 검색과 적합성 가중치 검색에서 1개 질문에 대해 검색된 적합문헌수는 한글은 9.03건, 영문은 21.39건으로 영문문헌에 대해 검색된 적합문헌수의 비율이 한글문헌보다 더 높다.

P-NORM 검색에서는 1개 질문에 대해 한글은 13.97건, 영문은 30.52건이 검색되어 불 논리 검색과 적합성 가중치 검색에 비해 검색된 적합문헌수에 있어서 한글은 1.54배, 영문은 1.42배가 더 검색되었다. 검색된 문헌수의 비율보다 적합문헌수의 비율이 낮은 것은 P-NORM 검색은 불 논리 검색이나 적합성 가중치 검색보다 검색되는 문헌수가 많아 적합문헌수가 증가하지만 또한 부적합문헌수도 역시 증가하고 있음을 나타낸다.

각 검색모형의 상대재현율은 〈표-7〉과 같다. P-NORM 검색은 불 논리 검색, 적합성 가중치 검색에 비해 상대재현율이 한글문헌 검색에서 56.3%, 영문문헌에서는 53.8% 더 높다. P-NORM 검색은 논리연산자에 P 값을 사용하므로 탐색식의 탐색어 가운데 한 개의 용어와 관련된 문헌도 검색될 수 있어서 재현율이 증가하였다.

4.1.1 정확률

각 검색모형에서 정확률은 〈표-8〉과 같다. 한글문헌 검색의 경우 58% 재현수준에서 불

〈표 5〉 각 검색모형에 의해 검색된 문헌수

구분 / 검색문헌수	한글			영문		
	논리합 형태	적합성 가중치	P-NORM	논리합 형태	적합성 가중치	P-NORM
합 계	564	564	1044	1096	1096	1732
평 균	17.09	17.09	31.64	33.21	33.21	52.48

〈표 6〉 각 검색모형에 의해 검색된 적합문헌수

구분 / 적합문헌수	한글			영문		
	논리합 형태	적합성 가중치	P-NORM	논리합 형태	적합성 가중치	P-NORM
합 계	298	298	461	706	706	1007
평 균	9.03	9.03	13.97	21.39	21.39	30.52

〈표 7〉 각 검색모형에 의한 검색결과의 상대재현율

구분 / 탐색질문수	한글			영문		
	논리합 형태	적합성 가중치	P-NORM	논리합 형태	적합성 가중치	P-NORM
평 균	0.64	0.64	1.00	0.65	0.65	1.00

논리 검색은 정확률이 0.57 이었고, 적합성 가중치 검색은 0.64, P-NORM 검색은 0.65의 결과를 보였다. 영문문헌의 검색에서는 불 논리 검색 0.72, 적합성 가중치 검색 0.77, P-NORM 검색은 0.76이었다. 한글문헌과 영문문헌의 검색에서 모두 불 논리 검색에 비해 적합성 가중치 검색, P-NORM 검색은 정확률이 더 높다. 그리고 적합성 가중치 검색과 P-NORM 검색 간에 정확률의 차이는 한글 문헌에서 0.8%의 근소한 차이로 P-NORM 검색의 정확률이 약간 높은 반면에, 영문문헌의 검색에서는 적합성 가중치 검색이 정확률에서 1.7% 더 높은 결과를 보였다.

〈표 8〉 각 검색모형에 의한 검색결과의 정확률

질문번호	한 글				영 문			
	재 현 율	논리합 형태	적합성 가중치	P-NORM	재 현 율	논리합 형태	적합성 가중치	P-NORM
1	0.9	0.88	0.94	0.93	0.9	0.93	0.92	0.92
2	0.3	0.50	0.50	0.50	0.6	0.90	0.95	0.90
3	0.9	0.41	0.42	0.42	0.8	0.58	0.66	0.68
4	0.8	0.88	0.88	0.88	0.9	0.81	0.85	0.85
5	0.3	0.75	0.75	0.75	1.0	0.55	0.55	0.55
6	0.4	0.23	0.35	0.35	0.8	0.84	0.86	0.86
7	0.7	0.45	0.49	0.49	0.5	0.65	1.00	1.00
8	0.5	1.00	1.00	1.00	0.5	0.91	0.90	0.90
9	0.1	1.00	1.00	1.00	0.3	0.75	0.75	0.75
10	0.6	0.36	0.36	0.36	0.5	0.80	0.97	0.97
11	1.0	0.57	0.57	0.57	0.3	1.00	1.00	1.00
12	0.5	0.33	0.77	0.77	0.5	0.80	0.87	0.87
13	0.8	0.40	0.56	0.50	0.2	1.00	1.00	1.00
14	0.4	0.80	0.90	0.90	0.2	0.32	0.38	0.38
15	0.1	0.50	1.00	1.00	0.2	1.00	1.00	1.00
16	0.3	0.60	1.00	1.00	0.8	0.60	0.83	0.83
17	1.0	0.20	0.20	0.20	0.8	0.41	0.41	0.41
18	0.5	0.69	0.70	0.70	0.6	0.70	0.73	0.73
19	0.5	0.50	0.54	0.54	0.5	0.94	1.00	1.00
20	0.7	0.84	0.82	0.82	0.7	0.78	0.93	0.83
21	0.9	0.48	0.49	0.49	1.0	0.48	0.48	0.48
22	0.2	0.20	0.20	0.20	0.7	0.60	0.60	0.60
23	1.0	0.44	0.44	0.44	0.7	0.56	0.56	0.56
24	0.6	0.67	0.68	0.68	0.1	0.67	1.00	0.75
25	0.4	0.86	0.86	0.86	0.4	0.62	0.67	0.67
26	0.2	0.63	0.64	0.64	0.3	0.80	0.80	0.64
27	0.5	0.67	1.00	1.00	0.6	0.69	0.73	0.73
28	0.7	1.00	1.00	1.00	0.6	0.41	0.53	0.53
29	0.2	0.29	0.33	0.33	1.0	0.84	0.84	0.84
30	1.0	0.43	0.50	0.50	0.8	0.77	0.77	0.77
31	0.8	0.57	0.57	0.57	0.6	0.33	0.40	0.40
32	0.2	0.50	0.67	1.00	0.7	0.78	0.79	0.79
33	1.0	0.13	0.13	0.13	0.2	0.78	0.79	0.79
평 균	0.58	0.57	0.64	0.65	0.58	0.72	0.77	0.76

4.1.2 검색순위

불 논리 검색, 적합성 가중치 검색 및 P-NORM 검색에 의한 검색순위와 이용자 판정에 의한 검색순위와의 관계검증을 위해서 스피어맨(Spearman)의 순위상관계수 ρ (rho) 값을 구

〈표 9〉 각 검색모형에 의한 검색결과의 질문별 상관계수 값

질 문	한 글				영 문			
	문헌그룹수	논리합 형태	적합성 가중치	P-NORM	문헌그룹수	논리합 형태	적합성 가중치	P-NORM
1	4	0.25	0.40	0.40	4	0.89	0.80	0.80
2	2	0.00	-1.00	-1.00	4	0.25	0.40	0.40
3	2	0.00	1.00	1.00	8	0.36	0.53	0.35
4	2	0.00	1.00	-1.00	6	-0.18	0.02	0.14
5	2	0.00	-1.00	-1.00	2	1.00	1.00	1.00
6	5	-0.28	-0.20	-0.20	7	0.39	0.35	0.35
7	3	-0.86	-0.50	-0.50	4	-0.26	0.40	0.40
8	4	-0.89	-0.80	-0.80	6	0.41	0.60	0.06
9	2	-1.00	-1.00	-1.00	2	0.00	-1.00	-1.00
10	2	1.00	1.00	1.00	2	0.00	1.00	1.00
11	2	0.00	-1.00	-1.00	5	-0.58	-0.20	0.10
12	5	0.57	0.90	0.70	3	0.87	1.00	1.00
13	5	0.00	0.50	0.80	2	1.00	1.00	1.00
14	2	0.00	1.00	1.00	3	-0.87	-1.00	-1.00
15	4	-0.44	0.60	0.60	2	-1.00	-1.00	-1.00
16	2	0.00	1.00	1.00	4	0.26	0.80	0.80
17	3	0.86	0.50	0.50	2	0.00	-1.00	-1.00
18	2	0.00	1.00	1.00	3	-0.87	-0.50	-0.50
19	2	0.00	1.00	1.00	4	0.26	0.80	0.80
20	5	0.00	0.60	0.60	3	0.87	1.00	1.00
21	2	0.00	1.00	1.00	3	0.00	-0.50	-0.50
22	2	0.00	-1.00	-1.00	8	-0.35	-0.67	-0.78
23	2	0.00	1.00	1.00	3	-0.87	-1.00	-1.00
24	4	-0.25	0.20	0.20	2	0.00	1.00	1.00
25	2	-1.00	-1.00	-1.00	3	-0.87	-0.50	-0.50
26	3	0.00	-0.50	-0.50	2	0.00	1.00	1.00
27	2	0.00	1.00	1.00	3	0.87	1.00	1.00
28	3	0.00	-0.50	-0.50	6	0.41	0.94	1.00
29	5	0.22	0.10	0.10	2	-1.00	-1.00	-1.00
30	3	0.86	1.00	1.00	2	-1.00	-1.00	-1.00
31	2	1.00	1.00	1.00	3	0.00	0.50	0.50
32	3	0.86	1.00	1.00	3	0.87	1.00	1.00
33	2	-1.00	-1.00	-1.00	5	0.58	0.50	0.30

하여 이들 간의 양의 관계, 또는 음의 관계를 조사하였다. 스피어맨의 순위상관계수를 사용한 것은 분석하고자 하는 측정치가 선형적이지 않고 서열등급에 의한 특성을 갖기 때문이다. 한글문헌 및 영문문헌 검색에서 각 검색모형의 질문별 순위상관계수 값은 〈표-9〉와 같다.

ρ값이 0.5 이상인 것으로 각 검색모형별로 높은 상관관계를 갖는 질문의 비율을 조사한 결과 한글문헌 검색에서 불 논리 검색 18.18%, 적합성 가중치 검색 54.54%, P-NORM 검색은 51.51%로서 적합성 가중치 검색에 의한 검색순위가 이용자 판정에 의한 검색순위와 가장 상관관계가 높다. 영문문헌의 검색에서는 ρ값이 0.5 이상의 값을 갖는 질문의 비율은 불 논리 검색 24.24%, 적합성 가중치 검색 51.51%, P-NORM 검색은 42.42%로서 영문문헌 검색에서도 적합성 가중치에 의한 검색순위가 이용자 판정에 의한 검색순위와 가장 상관관계가 높은 것으로 해석된다.

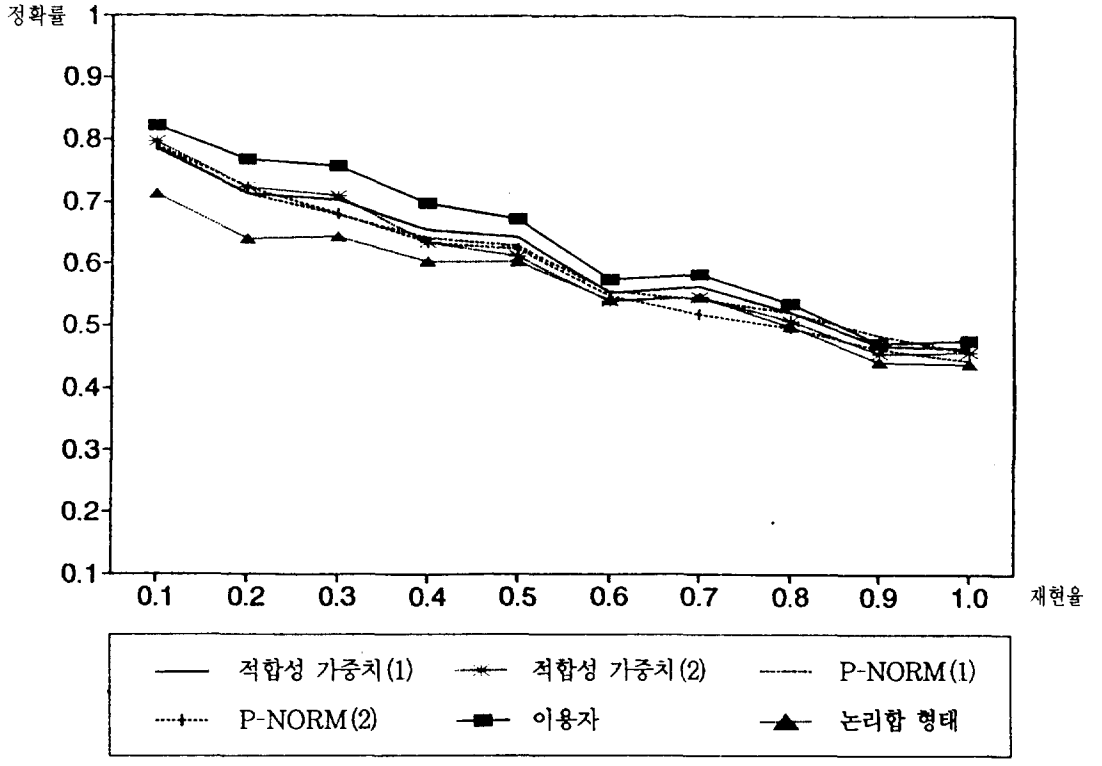
4.1.3 검색성능

각 검색방법에 의한 평균정확률은 〈표-10〉, 〈표-11〉과 같으며, 성능곡선은 〈그림-2〉, 〈그림-3〉과 같다.

평균정확률은 한글문헌의 검색에서 불 논리 검색에 비해 적합성 가중치 검색은 7.9%, P-NORM 검색은 0.5%가 높다. 영문문헌의 검색에서는 불 논리 검색에 비해 적합성 가중치 검색은 정확률이 3.7% 높고, P-NORM 검색에서는 0.6%가 낮았다. 결과적으로 한글문헌 및 영문문헌의 검색에서 적합성 가중치 검색은 불 논리 검색에 비해 높은 정확률을 원하는 검색에서 비교적 효과적인 것으로 분석된다. 반면에 P-NORM 검색은 〈표-7〉에서 살펴본 바와 같이 높은 재현율을 원하는 검색에서는 효과적이나 정확률은 크게 개선하지 않는 것으로 해석된다.

〈표 10〉 각 재현수준에서 평균정확률(한글)

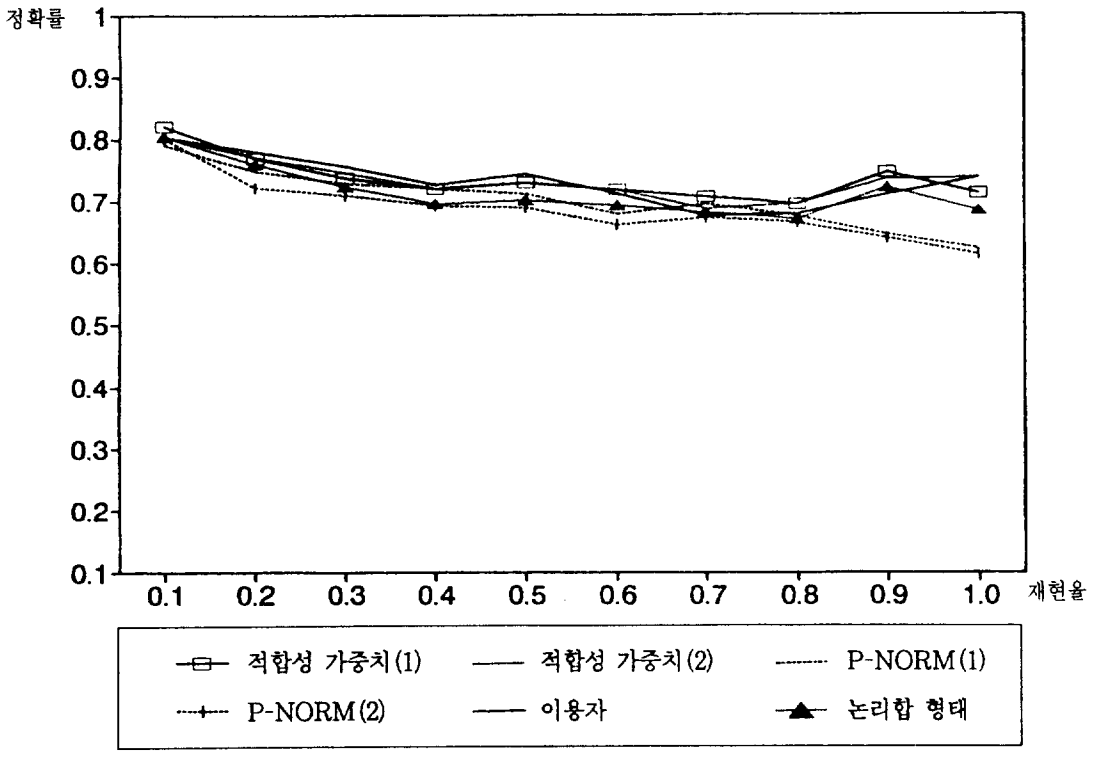
	논리합 형태	적합성 가중치 (1)	적합성 가중치 (2)	P-NORM (1)	P-NORM (2)	이 용 자
0.1	0.71	0.78	0.79	0.78	0.79	0.82
0.2	0.63	0.71	0.72	0.71	0.72	0.76
0.3	0.64	0.70	0.70	0.67	0.68	0.75
0.4	0.60	0.65	0.63	0.64	0.63	0.69
0.5	0.60	0.64	0.61	0.62	0.62	0.67
0.6	0.54	0.55	0.53	0.55	0.54	0.57
0.7	0.54	0.56	0.54	0.54	0.51	0.58
0.8	0.49	0.52	0.50	0.52	0.49	0.53
0.9	0.44	0.46	0.45	0.48	0.46	0.47
1.0	0.43	0.46	0.45	0.45	0.44	0.47



〈그림 2〉 한글문헌 검색결과의 성능곡선

〈표 11〉 각 재현수준에서 평균정확률(영문)

	논리합 형태	적합성 가중치 (1)	적합성 가중치 (2)	P-NORM (1)	P-NORM (2)	이 용 자
0.1	0.80	0.82	0.80	0.79	0.80	0.80
0.2	0.76	0.77	0.77	0.75	0.72	0.78
0.3	0.72	0.74	0.75	0.73	0.71	0.76
0.4	0.70	0.72	0.72	0.72	0.69	0.73
0.5	0.70	0.73	0.73	0.71	0.69	0.74
0.6	0.69	0.72	0.72	0.68	0.66	0.71
0.7	0.68	0.71	0.69	0.69	0.67	0.68
0.8	0.67	0.69	0.69	0.68	0.66	0.68
0.9	0.72	0.75	0.74	0.65	0.64	0.71
1.0	0.68	0.71	0.74	0.62	0.61	0.74



〈그림 3〉 영문문헌 검색결과의 성능곡선

4.1.4 유의도

사전검증은 SAS(Statistical Analysis Package)를 사용하여 일원분산분석을 하였다. F 검증에서 한글문헌 검색은 확률값 0.0009, 영문문헌 검색은 확률값 0.0002로서 유의수준 0.05 에서 차이가 있는 것으로 판정되었다. 개별 검색모형 간의 정확률이 유의할 차이가 있는지 비교하기 위해 던컨(Duncan)의 다중비교검증을 하였다.

각 검색모형은 정확률과 검색순위에서 모두 차이가 있으나, 개별 검색모형 간의 직접비교에서 정확률에 차이가 없으므로 귀무가설 1은 기각될 수 없다. 적합성 가중치 검색과 P-NORM 검색은 영문문헌의 검색에서 정확률, 순위상관계수에서 모두 차이가 있으므로 유의수준 0.05에서 귀무가설 2는 기각되었다.

4.2 가설 3과 가설 4의 검증

적합성 정보 및 탐색어의 가중치가 각 검색방법의 검색성능에 어떤 영향을 주는지에 대해 분석한 내용은 다음과 같다.

4.2.1 정확률

적합성 가중치 검색에서 완전한 적합성 정보를 이용한 경우와 소수의 적합성 정보를 이용한 경우에 한글문헌 검색결과는 〈표-12〉와 같다. 58% 재현수준에서 정확률은 각각 0.64, 0.63으로서 완전한 적합성 정보를 이용한 경우가 1.6% 더 높다. 영문문헌의 검색에서는 〈표-13〉

과 같이 정확률이 각각 0.77, 0.74로 완전한 적합성 정보를 이용한 경우가 4.1% 더 높다.

P-NORM 검색에서 탐색어에 적합성 가중치를 사용한 경우와 역문헌빈도 가중치를 사용한

〈표 12〉 적합성 가중치 검색 및 P-NORM 검색결과의 정확률(한글)

질문번호	재 현 율	적합성 가중치 (1)	적합성 가중치 (2)	P-NORM (1)	P-NORM (2)
1	0.9	0.94	0.88	0.93	0.88
2	0.3	0.50	1.00	0.50	0.50
3	0.9	0.42	0.41	0.42	0.41
4	0.8	0.88	0.88	0.88	0.88
5	0.3	0.75	0.75	0.75	0.75
6	0.4	0.35	0.35	0.35	0.35
7	0.7	0.49	0.41	0.49	0.41
8	0.5	1.00	1.00	1.00	1.00
9	0.1	1.00	1.00	1.00	1.00
10	0.6	0.36	0.36	0.36	0.36
11	1.0	0.57	0.57	0.57	0.57
12	0.5	0.77	0.23	0.77	0.77
13	0.8	0.56	0.50	0.50	0.39
14	0.4	0.90	0.90	0.90	0.90
15	0.1	1.00	1.00	1.00	1.00
16	0.3	1.00	1.00	1.00	1.00
17	1.0	0.20	0.20	0.20	0.20
18	0.5	0.70	0.70	0.70	0.70
19	0.5	0.54	0.45	0.54	0.54
20	0.7	0.82	0.84	0.82	0.83
21	0.9	0.49	0.48	0.49	0.49
22	0.2	0.20	0.20	0.20	0.20
23	1.0	0.44	0.44	0.44	0.44
24	0.6	0.68	0.68	0.68	0.75
25	0.4	0.86	0.86	0.86	0.86
26	0.2	0.64	0.64	0.64	0.62
27	0.5	1.00	1.00	1.00	1.00
28	0.7	1.00	1.00	1.00	1.00
29	0.2	0.33	0.25	0.33	0.33
30	1.0	0.50	0.50	0.50	0.38
31	0.8	0.57	0.57	0.57	0.57
32	0.2	0.67	0.67	1.00	0.67
33	1.0	0.13	0.13	0.13	0.13
평 균	0.58	0.64	0.63	0.65	0.63

경우에 정확률은 〈표-12〉, 〈표-13〉과 같다. 적합성 가중치를 사용한 경우가 역문헌빈도 가중치를 사용한 경우보다 한글문헌 검색에서 3.2%, 영문문헌 검색에서 2.7% 더 높다.

〈표 13〉 적합성 가중치 검색 및 P-NORM 검색결과의 정확률(영문)

질문번호	재 현 율	적합성 가중치 (1)	적합성 가중치 (2)	P-NORM (1)	P-NORM (2)
1	0.9	0.92	0.92	0.92	0.92
2	0.6	0.95	0.86	0.90	0.86
3	0.8	0.66	0.50	0.68	0.68
4	0.9	0.85	0.85	0.85	0.92
5	1.0	0.55	0.55	0.55	0.55
6	0.8	0.86	0.86	0.86	0.80
7	0.5	1.00	1.00	1.00	1.00
8	0.5	0.90	0.86	0.90	0.86
9	0.3	0.75	0.75	0.75	0.75
10	0.5	0.97	0.97	0.97	0.97
11	0.3	1.00	1.00	1.00	1.00
12	0.5	0.87	0.87	0.87	0.87
13	0.2	1.00	1.00	1.00	1.00
14	0.2	0.38	0.38	0.38	0.38
15	0.2	1.00	1.00	1.00	1.00
16	0.8	0.83	0.83	0.83	0.57
17	0.8	0.41	0.41	0.41	0.41
18	0.6	0.73	0.69	0.73	0.69
19	0.5	1.00	0.94	1.00	0.94
20	0.7	0.93	0.72	0.83	0.72
21	1.0	0.48	0.48	0.48	0.48
22	0.7	0.60	0.60	0.60	0.60
23	0.7	0.56	0.56	0.56	0.61
24	0.1	1.00	0.60	0.75	0.75
25	0.4	0.67	0.67	0.67	0.62
26	0.3	0.80	0.80	0.64	0.80
27	0.6	0.73	0.73	0.73	0.73
28	0.6	0.53	0.48	0.53	0.35
29	1.0	0.84	0.84	0.84	0.84
30	0.8	0.77	0.77	0.77	0.77
31	0.6	0.40	0.43	0.40	0.40
32	0.7	0.79	0.78	0.79	0.78
33	0.2	0.75	0.75	0.75	0.75
평 균	0.58	0.77	0.74	0.76	0.74

4.2.2 검색순위

적합성 가중치 검색에서 적합성 정보 이용에 따른 검색결과를 검증한 결과, 상관계수 값은 〈표-14〉, 〈표-15〉와 같다. 한글문헌 검색에서 비교적 높은 ρ값(ρ>0.5)을 갖는 질문의 비율은 완전한 적합성 정보를 이용한 검색에서는 54.54%, 소수의 적합성 정보를 이용한 검색에서는 42.42%이고, 영문문헌의 검색에서는 각각 51.51%, 45.45%이다. 따라서 완전한 적합성 정

〈표 14〉 적합성 가중치 검색 및 P-NORM 검색결과의 질문별 상관계수 값(한글)

질문번호	문헌그룹수	적합성 가중치 (1)	적합성 가중치 (2)	P-NORM (1)	P-NORM (2)
1	4	0.40	-0.20	0.40	-0.40
2	2	-1.00	1.00	-1.00	-1.00
3	2	1.00	-1.00	1.00	-1.00
4	2	1.00	-1.00	-1.00	-1.00
5	2	-1.00	-1.00	-1.00	1.00
6	5	-0.20	-0.20	-0.20	-0.10
7	3	-0.50	-1.00	-0.50	-1.00
8	4	-0.80	-0.80	-0.80	-0.80
9	2	-1.00	-1.00	-1.00	-1.00
10	2	1.00	1.00	1.00	1.00
11	2	-1.00	-1.00	-1.00	1.00
12	5	0.90	0.80	0.70	0.70
13	5	0.50	0.50	0.80	-0.10
14	2	1.00	1.00	1.00	1.00
15	4	0.60	0.60	0.60	0.60
16	2	1.00	1.00	1.00	-1.00
17	3	0.50	0.50	0.50	0.50
18	2	1.00	1.00	1.00	1.00
19	2	1.00	-1.00	1.00	1.00
20	5	0.60	-0.30	0.60	-0.50
21	2	1.00	1.00	1.00	1.00
22	2	-1.00	-1.00	-1.00	-1.00
23	2	1.00	-1.00	1.00	-1.00
24	4	0.20	-0.40	0.20	-0.80
25	2	-1.00	-1.00	-1.00	-1.00
26	3	-0.50	-0.50	-0.50	0.50
27	2	1.00	1.00	1.00	1.00
28	3	-0.50	-0.50	-0.50	-1.00
29	5	0.10	0.10	0.10	0.10
30	3	1.00	1.00	1.00	-0.50
31	2	1.00	1.00	1.00	1.00
32	3	1.00	1.00	1.00	-0.50
33	2	-1.00	-1.00	-1.00	-1.00

〈표 15〉 적합성 가중치 검색 및 P-NORM 검색결과의 질문별 상관계수 값(영문)

질문번호	문헌그룹수	적합성 가중치 (1)	적합성 가중치 (2)	P-NORM (1)	P-NORM (2)
1	4	0.80	0.80	0.80	0.80
2	4	0.40	-0.80	0.40	-0.80
3	8	0.53	0.06	0.35	0.35
4	6	0.03	0.03	0.14	-0.08
5	2	1.00	1.00	1.00	1.00
6	7	0.36	0.36	0.35	0.25
7	4	0.40	0.20	0.40	0.40
8	6	0.60	-0.08	0.05	-0.08
9	2	-1.00	-1.00	-1.00	-1.00
10	2	1.00	1.00	1.00	1.00
11	5	-0.20	0.60	0.10	0.10
12	3	1.00	1.00	1.00	1.00
13	2	1.00	1.00	1.00	1.00
14	3	-1.00	-1.00	-1.00	-1.00
15	2	-1.00	-1.00	-1.00	-1.00
16	4	0.80	0.80	0.80	-0.20
17	2	-1.00	-1.00	-1.00	-1.00
18	3	-0.50	-1.00	-0.50	-1.00
19	4	0.80	0.60	0.80	0.00
20	3	1.00	0.50	1.00	0.50
21	3	-0.50	-0.50	-0.50	-0.50
22	8	-0.66	-0.66	-0.79	-0.87
23	3	-1.00	-1.00	-1.00	-1.00
24	2	1.00	1.00	1.00	-1.00
25	3	-0.50	-1.00	-0.50	-1.00
26	2	1.00	-1.00	1.00	-1.00
27	3	1.00	1.00	1.00	1.00
28	6	0.94	0.77	1.00	-0.54
29	2	-1.00	-1.00	-1.00	-1.00
30	2	-1.00	-1.00	-1.00	-1.00
31	3	0.50	0.50	0.50	0.50
32	3	1.00	0.50	1.00	0.50
33	5	0.50	0.50	0.30	0.80

보를 이용한 경우가 소수의 적합성 정보를 이용한 경우보다 이용자 판정에 의한 검색순위와 상관관계가 더 높다. P-NORM 검색에서는 적합성 가중치를 적용한 경우에 비교적 높은 ρ값(ρ)0.5)을 갖는 질문의 비율이 한글문헌 검색에서 51.51%, 영문문헌 검색에서 42.42%이다. 반면에 역문헌빈도 가중치를 적용한 경우는 한글문헌 검색에서 39.39%, 영문문헌 검색에서 30.30%인 것으로 나타나 적합성 가중치를 적용한 검색이 상관관계가 더 높은 결과를 보였다.

4.2.3 유의도

적합성 가중치 검색에서 적합성 정보 이용수준에 의한 검색방법 간의 차이에 대해 paired t 검증을 실시한 결과, 유의수준 0.05에서 한글문헌의 검색은 확률값이 0.4053 이므로 차이가 없다. 영문문헌의 검색에서는 확률값이 0.0125로서 차이가 있는 것으로 나타났다. P-NORM 검색에서 적합성 가중치를 적용한 경우와 역문헌빈도 가중치를 적용한 경우에 한글문헌의 검색에서는 확률값이 0.1163 이므로 두 검색방법 간에 차이가 없다. 반면에, 영문문헌의 검색에서는 확률값이 0.0006 으로서 유의수준 0.05에서 차이가 있다. 결론적으로 적합성 가중치 검색은 영문문헌의 검색에 있어서 완전한 적합성 정보를 이용한 경우에 소수의 적합성 정보를 이용한 경우보다 정확률이 더 높으며, P-NORM 검색은 영문문헌의 검색에 있어서 탐색어에 적합성 가중치를 적용한 경우가 역문헌빈도 가중치를 적용한 경우보다 정확률이 더 높은 것으로 요약된다.

V. 결 론

불 논리 검색의 개선을 위한 검색모형들에 대한 실험을 통해 검증된 내용은 다음과 같다.

첫째, 질문변환에 의한 불 논리 검색에 대해 적합성 가중치 검색과 P-NORM 검색의 정확률은 유의할 차이가 없다. 반면에 검색순위에 의한 순위상관관계와 이용자 판정에 의한 검색결과와의 비교에서 적합성 가중치 검색은 질문변환에 의한 불 논리 검색과 P-NORM 검색보다 더 효과적이다.

둘째, 영문문헌의 검색에서 적합성 가중치 검색이 P-NORM 검색보다 정확률이 더 높고, 검색순위는 이용자 판정에 의한 검색순위와 더 높은 상관관계가 있다.

셋째, 적합성 가중치 검색은 영문문헌의 검색에서 완전한 적합성 정보를 이용한 경우가 소수의 적합성 정보를 이용한 경우보다 정확률이 더 높고 이용자 판정에 의한 검색순위와 더 높은 상관관계가 있다.

넷째, P-NORM 검색은 영문문헌의 검색에서 탐색어에 적합성 가중치를 적용한 경우가 역문헌빈도 가중치를 적용한 경우보다 더 높은 정확률을 나타내며, 이용자 판정에 의한 검색순위와 더 높은 상관관계가 있다.

한글문헌과 영문문헌의 검색에서 차이를 보이는 이유는 언어적 특성에 의한 것이기보다는 데이터베이스에서 적합문헌의 분포특성에 의한 영향인 것으로 판단된다. 본 실험을 수행하는 과정에서 탐색질문이 특정적이고 최근 연구와 관련된 용어에 관한 탐색일 때, 한글문헌 탐색에서는 탐색결과를 보고(browse) 탐색식을 수정(modify)하여 탐색하게 되는 경우가 많아 탐

색에 사용된 명령싸이클수가 증가하였다. 그럼에도 불구하고 적합성 판정에 의한 결과는 영문문헌 검색에서는 적합문헌의 비율이 부분적합문헌보다 9.5% 높은 반면에, 한글문헌 검색에서는 적합문헌보다 부분적합문헌의 비율이 11% 높다(〈표-6〉, 〈표-7〉 참조). 그리고 탐색된 질문수 전체에 대한 정확률이 한글문헌의 검색보다 영문문헌의 검색에서 16.6% 높은 결과를 보인 것 등은 이같은 점을 잘 나타낸 것이다(〈표-10〉, 〈표 -11〉 참조).

본 연구와 관련하여 다음과 같은 영역에서 후속의 연구가 필요하다.

첫째, 데이터베이스 및 탐색 질문이 전문성 수준에 있어서 보다 더 동등한 조건에서 탐색을 실시한다. 둘째, 탐색어와 색인어에 모두 가중치를 준 경우에 정확률과 검색 순위에서 어떤 차이를 보이는지 분석한다. 세째, 탐색을 점진적 과정으로 보고 질문식 수정을 통하여 적합성 가중치 검색과 P-NORM 검색에서의 변화를 평가한다.

참고문헌

노정순 (1991). 우리나라 온라인 탐색자의 탐색형태와 탐색전략 개발에 관한 연구. 『한남대학교 인문과학 논문집』, 21, 127-161.

정영미 (1993). 『정보 검색론』. 서울: 구미무역.

Bookstein, A. & W.S. Cooper (1976). General Mathematical Model for Information Retrieval *System. Library Quarterly,* 46, 153-167.

Bookstein, A. (1978). On the Perils of Merging Boolean and Weighted Retrieval Systems. *JASIS,* 29, 156-158.

Bookstein, A. (1981). A Comparison of Two Weighting Schemes for Boolean Retrieval. *JASIS,* 32, 275-279.

Buell, D.A. (1981). A General Model of Query Processing in Information Retrieval. *IPM,* 17, 249-260.

Croft, W.B. (1986). Boolean Queries and Term Dependencies. *JASIS,* 37, 71-77.

Cooper, W.S. (1988). Getting Beyond Boole. *IPM,* 24, 249-255.

Croft, W.B. & D.J. Harper (1979). Using Probabilistic Models of Document Retrieval Without Relevance Information. *JD,* 35, 285-295.

Dillon, M., J. Ulmschneider, & J. Desper (1983). A Prevalence Formula for Automatic Relevance Feedback in Boolean Systems. *IPM,* 19, 27-36.

Fox, E.A. & M.B. Koll (1988). Practical Enhanced Boolean Retrieval: Experiences with the SMART and the SIRE Systems. *IPM,* 24, 257-267.

Harper, D.J. & C.J. van Rijsbergen (1978). An Evaluation of Feedback in Document Retrieval Using Co-occurrence Data. *JD,* 34, 189-216.

Harter, S.P.(1986). *Online Information Retrieval: Concepts, Principles, and Techniques.* London: Academic Press.

Kraft, D.H. & A. Bookstein (1978). Evaluation of Information Retrieval Systems: A

Decision Theory Approach. *JASIS*, 29, 31-40.

Lee, J.H., M.H. Kim, & Y.J. Lee (1993). Information Retrieval Based on Conceptual Distance in Is-a Hierarchies. *JD*, 49, 188-207.

McGill, M.J. (1976). Knowledge and Information Spaces: Implications for Retrieval Systems. *JASIS*, 27, 205-210.

Miller, W.L. (1971). Probabilistic Search Strategy for MEDLARS. JD, 27, 254-266.

Noreault, T., M. McGill, & M. Koll (1981). A Performance Evaluation of Similarity Measures, Document Term Weighting Schemes and Representation in a Boolean Environment. In *Information Retrieval Research*, ed. by R.N. Oddy et al. London: Butterworths.

Rada, R. & E. Bicknell (1989). Ranking Documents with a Thesaurus. *JASIS*, 40, 304-310.

Radecki, T. (1982). A Probabilistic Approach to Information Retrieval in Systems with Boolean Search Request Formulations. *JASIS*, 33, 365-370.

Radecki, T. (1988). Probabilistic Methods for Ranking Output Documents in Conventional Boolean Retrieval Systems. *IPM*, 24, 281-303.

Ro, J.S. (1985). An Evaluation of the Applicability of Ranking Algorithms Improving the Effectiveness of Full Text Retrieval. Unpublished Ph.D. dissertation, Indiana University.

Robertson, S.E. & K. Sparck Jones (1976). Relevance Weighting of Search Terms. *JASIS*, 27, 129-146.

Robertson, S.E. (1977a). Theories and Models in Information Retrieval. *JD*, 33, 126-146.

Robertson, S.E. (1977b). The Probabilistic Ranking Principle in IR. *JD*, 33, 294-304.

Robertson S.E., C.J. van Rijsbergen, & M.F. Porter (1981). Probabilistic Models of Indexing and Searching. In *Information Retrieval Research*, ed. by R.N. Oddy et al. London: Butterworths.

Robertson, S.E. (1990). On Term Selection For Query Expansion. *JD*, 46, 359-364.

Salton, G., A. Wong, & C.T. Yu (1976). Automatic Indexing Using Term Discrimination and the Term Precision Measurements. *IPM*, 12, 43-51.

Salton, G., E.A. Fox, & H. Wu. (1983). Extended Boolean Information Retrieval. *CACM*, 26, 1022-1036.

Salton, G., C. Buckley, & E.A. Fox (1983). Automatic Query Formulations in Information Retrieval. *JASIS*, 34, 262-280.

Salton, G. (1988). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Reading, Massachusetts: Addison-Wesley.

Salton, G. & C. Buckley (1988). Term-Weighting Approaches in Automatic

Text Retrieval. *IPM*, 24, 513-523.

Salton, G. & C. Buckley (1990). Improving Retrieval Performance by Relevance Feedback. *JASIS*, 41, 288-297.

Salton, G. (1992). The State of Retrieval System Evaluation. *IPM*, 28, 441-449.

Sparck-Jones, K. (1979a). Search Term Relevance Weighting Given Little Relevance Information. *JD*, 35, 30-48.

Sparck-Jones, K. (1979b). Experiments in Relevance Weighting of Search Terms. *IPM*, 15, 133-144.

van Rijsbergen, C.J. (1977). A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval. *JD*, 33, 106-119.

van Rijsbergen, C.J. (1979). *Information Retrieval*, 2nd ed. London: Butterworths.

van Rijsbergen, C.J., D.J. Harper, and M.F. Porter (1981). The Selection of Good Search Terms. *IPM*, 17, 77-91.