순위화시스템의 효과측정척도에 관한 연구

A Study on the measurement of the system effectiveness with ranked results

노정순(Jung-Soon Ro)*

목 차

초 록

본 연구는 IR시스템 평가에 가장 많이 사용되는 R과 P가 순위화된 검색결과를 제공하는 시스템의 효과를 측정하는데 적절한가를 논의하고, R과 P 대신 순위화된 검색결과를 평가하는데 사용되고 있는 평가척도들을 고찰하고, 새로운 평가척도를 제안하였다. 또한 이들 평가척도가 어떤 환경에서 타당한지를 이론적으로 규명하고 사례를 들어 검증하였다.

11-포인트 평균정확률(평균)이 11-포인트 평균정확률(최고)보다 판별력이 있는 것으로 나타났고, 보다 간편한 여러 측정척도가 11-포인트 평균정확률을 대신할 수 있을 정도로 충분히 유사도가 높은 것으로 검증되었다.

ABSTRACT

This study discussed why Precision(& Recall) is not a good effectiveness measurement of IR systems providing ranked results, reviewed other effectiveness measurements appropriate for ranked results, and proposed new measurements based on the average rank of relevant documents retrieved.

The 18 case-sets of ranked results were used for evaluating the 10 effectiveness measurements including proposed measurements. Simple measurements were significantly similar with the 11-Point Precision requiring complicated calculation.

키워드 : 순위화, 시스템효과측정척도, 11-포인트 P, First-n P, 정확률

* 한남대학교 문헌정보학과 교수

■ 논문 접수일 : 2000년 11월 8일

1 연구목적 및 방법

최근 대부분의 웹(web) 탐색엔진과 같이 검색결과를 순위화시켜 제공하는 정보검색(IR)시스템이 다양해지면서 새로운 시스템평가 척도가 요구되고 있다. 기존의 대부분의 IR시스템 평가에서는 재현율(R)과 정확률(P)이나 재현율과 정확률을 단일수치로 표현하는 단일가 E 등이 시스템의 성능을 평가하는데 사용되었다. 그러나 재현율과 정확률은 순위화 성능을 평가하는데 판별력이 부족하기 때문에 TREC(Text REtrieval Conference)를 비롯한 여러 순위화 평가연구에서는 재현율-정확률곡선(R-P곡선)이나 이 곡선을 단일수치로 표현하는 11-포인트 평균정확률이 순위화된 시스템의 평가척도로 사용되고 있다. 그러나 R-P곡선과 11-포인트 평균정확률은 재현율 0.0, 0.1에서부터 0.9, 1.0 시점에서 각각의 정확률을 구하여야 하기 때문에 간단하지가 않다. 뿐만 아니라 R-P곡선은 하나의 평면 위에 여러 탐색결과를 모양이 다른 여러 선으로 표시하여 비교하기 때문에 비교 가능한 시스템(혹은 탐색)의 수에 제한을 받는다. 더욱이 대규모 데이터베이스(DB)를 대상으로 하는 실용 IR시스템에서 이용자의 정보요구에 대한 탐색효과를 측정하는 평가에서는 검색되지 않은 적합문헌을 알 수가 없기 때문에 재현율을 기반으로 하는 IR시스템 효과측정 척도를 사용하는데 어려움이 많다.

본 연구에서는 최근 여러 연구에서 사용되고 있는 순위화된 검색결과의 평가척도들을 고찰하고 새로운 평가척도를 제안하며, 이들 평가척도가 어떤 환경에서 타당한지를 이론적으로 규명하고 가상적 사례를 들어 검증하는데 그 목적이 있다. 다음과 같은 연구질문이 규명되었다.

연구질문

1. R과 P는 순위화된 결과를 제공하는 시스템의 탐색성능을 평가하기에 적당한 측정척도인가?
2. 11-포인트 평균정확률을 계산할 때 평균정확률과 최고정확률 중 어느 것이 더 판별력이 좋은가?
3. 순위화된 검색결과를 평가하는데 11-포인트 평균정확률보다 간편한 평가척도로는 무엇이 가능한가?
4. 11-포인트 평균정확률을 대신할 가장 좋은 평가척도는 무엇인가?

본 연구에서는 선행연구에서 사용된 평가척도들과 함께 이론적으로 가능한 새로운 척도들이 논의되었으며(2장), 연구질문 1~3은 가상적으로 작성된 18개 탐색결과를 예로 들어 규명하였고(3장), 연구질문 4는 18개 탐색결과의 효과를 측정한 다양한 평가척도 간의 상관관계를 규명함으로써 검증하였다(4장).

2 순위화시스템의 평가척도

2.1 P-R과 r^2/n

순위화된 검색결과를 제공하는 시스템의 성능은 사용된 언어통제, 파싱 룰 등 각종의 색인 언어처리알고리즘, 연산자처리알고리즘, 순위화알고리즘 등에 복합적으로 영향을 받고, 순위화 성능은 적합문헌을 부적합문헌보다 얼마나 더 먼저 출력시키느냐의 문제이다.

순위화된 검색결과를 제공하는 대부분의 web 탐색엔진의 성능을 비교한 국내외 많은 연구에

서 R(상대재현율)과 P가 평가척도로 사용된 것을 볼 수 있다. Timaiuolo & Packer(1996), Gauch & Wang(1996), 이명희(1998), 오삼균과 박희진(2000)의 연구에서는 P가 사용되었다. Clarke & Willett(1997), 이명희(1997), 이은주와 정영미(1997), 정영미와 김성은(1997)의 연구에서는 P와 함께 R(상대재현율)도 사용되었다.

실용 시스템에서 R과 P를 계산할 때 정보요구에 적합한 DB내의 모든 적합문헌을 알지 못하기 때문에 R을 계산하기가 불가능하다. 두 개 이상의 시스템을 비교할 때는 상대재현율을 사용할 수 있지만, Frants 등(1993)은 r^2/n(n : 검색된 문헌 수, r : 검색된 적합문헌 수)를 R과 P 대신 사용할 수 있는 타당성을 다음과 같이 논술하였다.

i번째 문헌이 적합문헌이면 $K_i=1$, 적합문헌이 아니면 $K_i=0$, i번째 문헌이 검색되면 $V_i=1$, 검색되지 않으면 $V_i=0$라 하면, 검색된 문헌 set과 적합문헌 set은 다음과 같은 두 개의 백터로 표현될 수 있다.

$K = (K_1, K_2, \cdots\cdots, K_n)$

$V = (V_1, V_2 \cdots\cdots, V_n)$

두 백터 사이의 거리(근접도)를 코사인 공식으로 계산하면,

$$\cos\phi = \frac{\sum^{n} (K_i, V_i)}{\sqrt{\sum^{n} (K_i)^2} \cdot \sqrt{\sum^{n} (V_i)^2}}$$

이 공식에서 $\sum^{n}(K_i)^2$는 적합문헌 총수(C)이고, $\sum^{n} (V_i)^2$는 검색된 문헌 총수(n), $\sum^{n}(K_i \cdot V_i)$는 검색된 적합문헌총수(r)와 같으므로, $\cos\phi$는 $\sqrt{R \cdot P}$와 동일하게 된다.

$$\cos\phi = \frac{r}{\sqrt{c} \cdot \sqrt{n}} = \sqrt{\frac{r^2}{c \cdot n}} = \sqrt{\frac{r}{c} \cdot \frac{r}{n}} = \sqrt{R \cdot P}$$

$R_i = \frac{ri}{\quad}$, $P_i = \frac{ri}{\quad}$이므로 $R_i \cdot P_i = \frac{ri^2}{\quad}$

두 개의 탐색결과 set를 $\sqrt{R_1 \cdot P_1}$과 $\sqrt{R_2 \cdot P_2}$라는 성능평가 척도를 사용한다면

$$\sqrt{R_1 \cdot P_1} : \sqrt{R_2 \cdot P_2} = \frac{r_1^2}{c \cdot n_1} : \frac{r_2^2}{c \cdot n_2}$$

각각에 c를 곱하면

$$\sqrt{R_1 \cdot P_1} : \sqrt{R_2 \cdot P_2} = \frac{r_1^2}{n_1} : \frac{r_2^2}{n_2}$$

결국 각각의 결과 set은 r^2/n으로 표시 가능하다.

2.2 R-P곡선과 11-포인트 평균정확률

11-포인트 평균P값은 R-P곡선의 성능을 단일가로 표현하는 것이다. 11-포인트 평균P는 TREC에서 시스템의 성능을 평가하는 표준척도로 사용된 이후 순위화된 검색결과를 제공하는 여러 실험시스템에서 평가척도로 사용되고 있을 뿐만 아니라 web 탐색 평가에서도 사용되고 있다(Chu & Rosenthal, 1996). 11-포인트 P나 R-P 곡선을 그리기 위해서는 순위화된 문헌 리스트에서 검색된 문헌을 1에서 2, 3 …으로 하나씩 늘려가면서 각각 R과 P를 구한 후 R=0.0, 0.1, …, 0.9, 1.0 일 때의 P를 계산하여 그래프(R-P 곡선)을 그리거나 평균(11-포인트)을 낸다.

〈표 1〉은 상위 10개의 문헌 중 1, 2, 4, 6등

〈표 1〉 S4의 문헌순위에 의한 R과 P

문헌순위	적합문헌	R	P	P(최고)	P(평균)
1	R	1/4 = 0 25	1/1=1.0	1.0	1.0
2	R	1/2 = 0.5	2/2 = 1.0	1.0	0.835
3		1/2 = 0.5	2/3 = 0.67		
4	R	3/4 = 0.75	3/4 = 0.75	1 75	0.675
5		3/4 = 0.75	3/5 = 0.6		
6	R	4/4 = 1.0	4/6 = 0.67		
7		4/4 = 1.0	4/7 = 0 57		
8		4/4 = 1.0	4/8 = 0 5	0.67	0.516
9		4/4 = 1.0	4/9= 0.44		
10		4/4 = 1 0	4/10 = 0.4		
평 균					0.7565

〈표 2〉 S4의 보간법을 사용한 R과 P

R	최고값(11Pm)	평균(11Pa)
0.0	1	1
0 1	1	1
0 2	1	1
0.3	1	0.97
0.4	1	0.90
0.5	1	0.84
0.6	0.90	0.77
0.7	0.80	0.71
0.8	0.73	0.64
0.9	0.70	0.58
1.0	0.67	0.52
11-포인트 평균 P	0.89	0 81

문헌이 적합문헌일 경우 문헌순위별로 검색된 문헌을 1개에서 10개로 늘려갈 때 각각의 R과 P를, 〈표 2〉는 R=0.0, 0.1, …, 0.9, 1.0 일 때 P값으로 〈표 1〉의 값을 환원한 것이다. 〈표 1〉에서 보는 바와 같이 문헌순위에 의한 R과 P를 산출하기 위해 재현율값을 고정시키고 정확률을 구하면 재현율 0.5와 0.75에서는 각각 2개의 정확률을, 재현율 1.0에서는 5개의 정확률을 갖게 된다. 1개의 재현율이 여러 개의 정확률을 가질 때는 최고값, 최소값, 중간 위치의 값, 평균

값, 최고값과 최소값의 평균값 중의 하나의 정확률을 택할 수 있다. 〈표 1〉에서 P(최고)는 한 개의 R이 여러 개의 P를 가질 때 그중 최고값의 P를 선택한 것이고, P(평균)은 여러 개 P의 평균값을 P값으로 선택한 것이다.

〈표 2〉에서 11Pm은 〈표 1〉의 P(최고)값을 가지고, 11Pa는 〈표 1〉의 P(평균)값을 가지고 각각 R값을 11개 값(0.0, 0.1, …, 1.0)으로 고정시켜 선형보간법에 의해 산출한 P값과 평균 P(11-포인트 P)를 나타낸다.

복잡한 11-포인트 P 대신 상대적으로 덜 복잡한 평균정확률(Pa)를 척도로 사용하는 것도 가능하다. Pa는 11-포인트 R에서 P를 계산하기 이전에 얻은 R값의 평균 P이다. 〈표 1〉의 예에서 Pa는 4개의 R(R=0.25, 0.5, 0.75, 1.0) 지점에서 얻은 4개의 P(평균)의 평균값(0.7565=(1.0+0.835+0.675+0.516)/4)이다. 즉 11Pa는 R=0.5에서 P(=0.81)값을 의미하나, Pa는 R=0.625(=(0.25+0.5+0.75+1.0)/ 4)에서 P(=0.7565)값을 의미한다.

2.3 First-n P

확률이나 백터모델을 사용하는 시스템에서는 DB내의 모든 문헌을 질문과의 유사도 혹은 적합도 순위로 순위화하기 때문에 검색되는 문헌수가 제한된 일정수에 미치지 못하는 경우가 드물겠지만, 불리언모델을 사용하여 순위화된 결과를 제공하는 시스템에서는 불연산조건에 만족시키는 문헌에 한하여 순위화를 하기 때문에 검색된 문헌수가 제한된 문헌수 보다 작게 검색되는 경우가 있다. 검색된 문헌수가 서로 다른 탐색을 평가할 때는 문헌의 순위 뿐만 아니라 검색된 문헌수도 반영되어야 한다.

Leighton & Srivastava(1999)의 First-n Precision은 web 탐색엔진의 성능측정에 사용하기 위해 정확률에 문헌의 순위를 응용시킨 척도이다. n은 web 탐색으로 검색된 순위화된 싸이트 중 몇 등까지의 싸이트를 링크하여 적합성을 판정하였는지를 의미한다. Leighton & Srivastava는 검색된 싸이트를 20개까지 링크하여 적합성을 판정하였다. 1~3등 문헌이 적합문헌이면 20점, 4~10등 문헌이 적합문헌이면 17점, 11~20등 문헌이 적합문헌이면 10점을 배정하여, 적합문헌의 점수합계를 구하였다. 이를 분자값으로 하여 20개 문헌이 모두 적합문헌일 경우 합계점수(279점)로 나눈 값을 first-20 P로 사용하였다. 검색된 싸이트 수가 20개가 안될 때는 20개 적합문헌 총합계(279점)에서 링크당 10점을 마이너스하여 분모값에 변화를 줌으로써 검색된 문헌건수에 영향을 받는 정도를 반영시켰다.

$$\text{First-20 P}=\frac{(\text{1-3등 링크 20})+(\text{4-10등 17})+(\text{11-20등 10})}{279-[(20-\min(\text{검색된 링크수},20))\ 10]}$$

Leighton & Srivastava의 연구에서는 1~3등, 4~10등, 11~20등으로 문헌을 그룹화함으로써 같은 그룹내의 문헌순위에는 영향을 받지 않게 하였다. 그러나 적합문헌이 4등에 위치하는 것과 10등에 위치하는 것은 차이가 있으므로 보다 정확한 판별력을 위해서는 등수간의 차이를 더욱 민감하게 하거나 모든 순위에 서로 다른 점수를 부여할 수 있을 것이다.

2.4 적합문헌의 평균등수에 기반한 척도(P/e와 r/e)

상위 n개의 문헌으로 출력을 제한하여 탐색결과를 평가할 때 n개의 문헌중 적합문헌(r)의

평균등수 e는 11-포인트 P 대신 사용될 수 있는 비교적 간단한 측정척도이다. 원하는 적합문헌수를 얻기 위해 훑어보아야 할 부적합문헌수로 성능을 평가하는 ESL(Expected Search Length : Cooper 1968)과 상대적 개념으로 Losee(1994)는 검색된 적합문헌의 평균등수 e를 기반으로한 ASL(Average Search Length)을 제안하였다. ASL은 원하는 R수준의 적합문헌을 얻기 위해 검색된 문헌수를 제한했을 때 검색된 적합문헌의 평균등수 e를 성능측정척도로 사용한 것이다. 그러나 ASL에서 DB내 전체 적합문헌수를 알아야 원하는 R 수준의 적합문헌을 알 수 있기 때문에 상위 n개의 문헌으로 출력을 제한할 경우엔 적당치 않다. 상대재현율과 비슷하게 상대(Relative) ASL(RASL)로 응용할 수도 있다. 예를 들어 4개의 시스템 탐색결과를 20개로 제한하여 평가한다면 80개의 검색된 문헌 중 고유한(unique) 적합문헌수를 전체 DB내 적합문헌수로 간주하여 RASL을 구할 수 있겠다.

본 연구에서는 적합문헌의 평균순위 e를 기반으로한 척도 P/e와 r/e의 가능성이 검토되었다. 검색된 문헌을 10개로 제한한 〈표 3〉에서 e는 검색된 문헌 중 적합문헌의 평균등수를 나타낸다. S5~S7 혹은 S8~S11 내에서와 같이 적합문헌수가 동수일 경우에 평균등수는 판별력을 갖는다. 이 경우에 ESL 보다 더 좋은 판별력을 갖는 것으로 보인다(ESL로 평가하면 S9 = S10). 그러나 검색된 적합문헌수가 다를 경우엔 판별능력을 상실한다. 평균등수로 볼 때 S5와 S12가 가장 좋은 탐색결과이다. 10개의 문헌이 모두 적합문헌인 S1의 평균등수는 5.5등이다. 그러나 10개중 첫 문헌과 끝 문헌만이 적합문헌일 경우에도 적합문헌의 평균순위는 5.5등이 된다.

〈표 3〉에서 볼 때 e값 자체는 탐색성능을 나타내지 못하지만 그 안에는 어떤 질서가 있음을 볼 수 있다. S2와 S6과 S9 혹은 S5와 S12의 예에서 e값이 동일한 경우 탐색결과가 나쁠수록 P가 낮음을 볼 수 있다. 또한 S5~S7 혹은 S8~S11에서와 같이 P값이 동일한 경우 탐색결과가 나쁠수록 e값이 커지는 것을 볼 수 있다. 그러므로 적합문헌의 평균순위를 이용한 탐색평가척도는 e/P로 수정할 수 있다. e/P값은 오름차순으로 문헌을 정렬하기 때문에 P나 11Pa 등 다른 척도와 마찬가지로 좋은 결과일수록 높은 값을 갖도록 내림차순으로 정렬하기 위해서 e/P값은 1/e/P 즉 P/e로 바꿀 수 있다.

검색된 문헌수를 상위 n개로 제한할 경우 e를 이용한 다른 척도로는 r/e를 생각해 볼 수 있다. 10개 모두 적합문헌일 경우(S1)와 첫문헌과 끝문헌만이 적합문헌일 경우 e는 모두 5.5이지만, 전자(S1)의 경우 e는 평균 5.5등 내에 10개의 적합문헌이 있지만, 후자의 경우엔 5.5등 내에 2개의 적합문헌만이 존재한다. 그러므로 1개의 등수간격 내에 존재하는 적합문헌수 즉 r/e로 평가척도를 삼을 수 있다.

적합문헌의 평균순위 e를 기반으로 하는 두 평가척도 P/e와 r/e는 비교하는 결과set의 검색된 문헌수가 동일할 경우에는 같은 기능을 수행한다. 즉, P=r/n 이므로, P/e : r/e는 r/n/e : r/e = r/ne : r/e가 되므로써 n이 동일할 경우 P/e와 r/e는 같은 결과를 가져온다(〈표 3〉 참조).

그러나 〈표 4〉에서와 같이 검색된 문헌수가 다른 상황에서 n은 동일하지 않으므로 P/e와 r/e는 다른 결과를 가져온다. r/e가 P/e보다 좋은 평가척도로 보이나 부적합문헌은 무시되었기 때문에 S3, S14, S15, S16은 같은 r/e값

을 갖게 되었다. 이 문제는 r/e에 부적합문헌수($\bar{r}$)를 고려하고, 부적합문헌에는 적합문헌보다 낮은 가중치를 줌으로써 $r-\bar{r}$이 음수값이 되지 않게 하므로써 해결할 수 있다($\frac{r-(a)\bar{r}}{e}$).

2.5 적합성가중치 기반의 척도

단순 2진척도(적합 혹은 부적합)가 아닌 적합성정도를 점수로 판정한 평가에서는 적합성 가중치를 응용한 변형된 P를 사용하기도 한다. Ding & Marchionini(1996), Chignell, Gwizdka, & Bodner(1999), 우유미,정영미(1998)는 web 탐색에서 검색된 싸이트의 적합도를 0~3점(0: 부적합, 3: 가장 적합)으로 판정하여, P(Best)와 P(Full)이라는 정확률로 web 탐색엔진의 성능을 비교평가 하였다. P(Best)는 가장 적합한(3점) 싸이트만을 적합 싸이트로 간주하여 산출한 정확률이다. P(Full)은 검색된 싸이트의 적합성 점수 합계를 검색된 싸이트가 모두 가장 적합한(3점) 싸이트 일 때 적합성합계로 나누었다.

$$P(Best) = \frac{\text{가장적합(3점)싸이트수}}{\text{검색된 싸이트수}}$$

$$P(Full) = \frac{\text{검색된 싸이트의 적합성 합계}}{\text{검색된 싸이트 수}\times 3}$$

$$\text{순위정확률} = \frac{\Sigma(V_i\times R_u\times 2)+\Sigma(V_i\times R_u\times 1)}{3\times 5\times 2+3\times 5\times 1}$$

P(Full)은 가중치정확률이란 명칭으로 우유미와 정영미(1998) 연구에서 사용되었으며, 가중치정확률과 함께 순위정확률도 사용하였다. 가중치정확률 P(Full)은 P에 적합성가중치를, 순위정확률은 First-n P에 적합성가중치를 적용한 척도이다(V_i : 각 문서의 적합성가중치(3, 2, 1, 0), R_u: 1~5위 문서중 적합문서수, R_u: 6~10위 문서중 적합문서수, 순위가중치: 2(1~5위)~1(6~10위)). Web 탐색엔진으로 검색된 상위 10개의 문서를 0(부적합)~3(아주 적합)점으로 적합성가중치를 부여하고, 1~5위의 문서에는 2점, 6~10위의 문서에는 1점의 순위가중치를 부여하였다. 그러나 이 측정방법은 First-n P에서 처럼 검색된 문헌이 상위 n개가 되지 못할 경우 검색된 문헌수는 반영하지 못하고 있다. Web을 대상으로한 선행연구에서처럼 연구된 Web엔진이 모두 상위 n개문헌를 제공할 때는 좋은 척도이나, 검색된 문헌수가 서로 다른 탐색결과를 평가할 때는 검색된 문헌수를 반영할 수 없다.

3 평가척도의 성능

본 연구에서는 가중치 적합도가 아닌 단순 2진 적합도로 문헌의 적합성을 판정할 때 사용될 수 있는 측정척도 P, r^2/n, 11Pa, 11Pm, Pa, First-n P, P/e, r/e, (r-$\bar{r}$)/e를 검색된 문헌수가 일정할 때와 다를 때로 나누어 비교 분석하였다.

〈표 3〉은 web 탐색에서와 같이 탐색결과를 10개까지만 출력한다고 가정하여 임의로 작성된 13개의 검색결과 set과 연구된 성능측정척도로 탐색결과를 평가한 것이다. 〈표 4〉는 비교평가하는 시스템에서 검색된 문헌수가 다를 경우로 가정하여 작성된 7개의 결과 set과 그 평가를 보여주고 있다. 〈표 3〉에서는 13종류의 검색된 결과 set을 검색된 10개의 문헌 중 적합문헌 개수가 보다 많고 적합문헌이 보다 상위에 출현하는 순으로 set번호를 부여하였다.

〈표 3〉 상위 10개의 문헌으로 탐색결과를 제한한 경우

Set	문헌순위										척 도									
	1	2	3	4	5	6	7	8	9	10	P	r^2/n	11Pa	11Pm	Pa	first-n P	e	P/e	r/e	$(r-\bar{r})/e$
S1	R	R	R	R	R	R	R	R	R	R	1.00	10 00	1.00	1.00	1.00	1.00	5.50	.182	1.82	1.82
S2	R	R	R	R	R	X	X	X	X	X	.50	2.50	.96	1.00	.94	.64	3.00	.167	1.67	1 50
S3	R	R	R	R	X	X	X	X	X	X	.40	1.60	.95	1.00	.91	.52	2.50	.160	1.60	1.36
S4	R	R	X	R	X	R	X	X	X	X	.40	1.60	.81	.89	.76	.48	3.25	.123	1.23	1 05
S5	R	R	R	X	X	X	X	X	X	X	.30	.90	.91	1.00	.85	41	2.00	.150	1.50	1.15
S6	X	R	R	R	X	X	X	X	X	X	.30	.90	.45	.50	41	38	3.00	100	1.00	.77
S7	X	R	X	R	X	R	X	X	X	X	.30	.90	.34	.40	.32	.33	4.00	.075	.75	.58
S8	X	R	R	X	X	X	X	X	X	X	.20	40	.33	.41	.29	26	2.50	.080	.80	.48
S9	X	R	X	R	X	X	X	X	X	X	.20	.40	.28	.36	.24	26	3.00	.067	.67	.40
S10	X	X	R	R	X	X	X	X	X	X	.20	.40	.23	29	.21	.23	3.50	.057	.57	.34
S11	X	R	X	X	X	X	X	X	X	R	.20	.40	.14	.28	.14	.22	6.00	.033	.33	.20
S12	X	R	X	X	X	X	X	X	X	X	.10	.10	.11	25	.11	.15	2.00	.050	50	.05
S13	X	X	X	X	X	X	X	X	X	R	.10	10	.05	.05	.05	07	1000	.010	.10	.01

(R : 적합문헌, X : 부적합문헌)

〈표 4〉 검색된 문헌수가 일정치 않을 경우

Set	문헌순위										척 도									
	1	2	3	4	5	6	7	8	9	10	P	r^2/n	11Pa	11Pm	Pa	first-n P	e	P/e	r/e	$(r-\bar{r})/e$
S1	R	R	R	R	R	R	R	R	R	R	1.00	10.00	1.000	1.00	1.000	1.00	5.50	.180	.182	1.80
S14	R	R	R	R							1.00	4.00	1.000	1.00	1.000	92	2 50	400	1 60	1.60
S15	R	R	R	R	X						.80	3.20	978	1 00	975	82	2.50	.320	1.60	1.56
S16	R	R	R	R	X	X					.67	2.70	.959	1.00	.955	.73	2 50	.268	1.60	1.52
S17	R	R	R	X							.75	2.25	.960	1.00	.958	.72	2.00	.375	1.50	1.45
S18	X	R	R	R	X	R					.67	2.70	.520	.52	.500	.63	3.75	.179	1.07	1.01
S3	R	R	R	R	X	X	X	X	X	X	.40	1.60	.950	1.00	.910	.52	2.50	.160	1.60	1.36

3.1 순위화된 검색결과를 평가하는데 R과 P는 적당한가?

P(=r/n)와 r^2/n은 공식이 의미하는 바와 같이 이론적으로 상위 n개의 문헌 중 적합문헌의 수만을 반영할 뿐, 그 적합문헌의 순위는 반영하지 못한다. 특히 〈표 3〉에서 보는 바와 같이 검색된 문헌수 n이 동일할 때는 P(=r/n)와 r^2/n에서 n은 상수이기 때문에 동일한 결과를 가져오며, n개 문헌 중 적합문헌수(r) 밖에 반영하지 못함으로써 좋은 순위화 측정척도가 되지 못함을 볼 수 있다.

S5와 S6, S7 모두에서 검색된 10개의 문헌 중 3개의 적합문헌을 얻었다면 세 결과 모두 정확률은 30%이다. 그러나 S5는 처음 세 문헌이 모두 적합문헌이고, S6은 2등, 3등, 4등 문헌이, S7은 2등, 4등, 6등 문헌이 적합문헌이기 때문에 S5의 순위화 결과는 S6보다, S6의 순위화 결과는 S7보다 우수하다. 그러나 P와 r^2/n은 이 세 탐색결과의 순위화 효과를 판별하는 능력을 갖지 못하고 있다.

〈표 4〉에서와 같이 검색된 문헌수가 일정치 않을 경우 r^2/n과 P는 서로 다른 결과를 가져오며, 11Pm보다 나은 비교적 좋은 판별력을 지니고 있음을 볼 수 있다. 비교평가하는 두 시스템 중 한 시스템은 S1을, 다른 시스템은 S14를 탐색결과로 출력한다면 이용자는 S1을 S14보다 더 좋게 평가하지 않을까? 그렇다면 r^2/n이 P보다 나은 판별력을 보인다고 할 수 있다. S16과 S17 혹은 S17과 S18을 비교한다면 어느 쪽이 더 좋은 검색결과라고 말하기는 어렵지만, P와 r^2/n은 다른 척도에 비해 여전히 가장 비슷한 척도임을 볼 수 있다.

3.2 11-포인트 평균정확률(11Pa와 11Pm)

〈표 3〉과 〈표 4〉에서 11Pa와 11Pm은 한 개의 R이 여러 개의 P를 가질 때 그 중 평균값의 P와 최고값의 P를 각각 선택하여 산출한 11-포인트 평균정확률을 나타낸다. 11-포인트 평균정확률을 계산할 때 평균정확률(11Pa)과 최고정확률(11Pm) 중 어느 것이 더 판별력이 좋은가? 보다 간편한 Pa와는 어떠한가?

〈표 3〉과 〈표 4〉에서 보는 바와 같이 11Pa가 11Pm보다 더 판별력이 있는 것으로 보인다. 〈표 3〉에서 최고의 P값으로 R-P곡선에서 11-포인트 평균정확률(11Pm)을 산출했을 때는 S1과 S2, S3 세 결과 set이 모두 같은 값을 갖게 됨으로써 판별력이 부족함을 볼 수 있다. 같은 탐색질문을 가지고 3개의 시스템에서 탐색을 수행하여 검색된 문헌수를 10개로 제한했다면 10개가 모두 적합문헌인 결과를 제공하는 시스템(S1)이 10개 중 처음 5개만(S2) 혹은 처음 4개만(S3) 적합문헌을 제공하는 시스템보다 더 좋은 시스템으로 평가되어야 할 것이다. 〈표 4〉는 검색된 문헌수가 다를 경우 11Pm은 P보다 더 판별력이 없음을 보여준다.

11Pa나 11Pm에 비해 상대적으로 복잡한 계산을 덜 필요로 하는 Pa는 〈표 3〉과 〈표 4〉에서 모두 11Pa와 가장 유사한 척도이며, 11Pm보다 나은 판별력을 갖는 것을 볼 수 있다.

〈표 5-a〉, 〈표 5-b〉는 각각 검색된 문헌수가 일정할 때(〈표 3〉)와 검색된 문헌수가 다를 때(〈표4〉), 11Pa와 11Pm, Pa를 포함한 9개 측정척도 간의 상관관계를 피어슨의 상관계수로 표현한 것이다. 계산이 복잡한 11-포인트 평균정확률(11Pa 혹은 11Pm) 대신 Pa를 사용할 수 있음을 보여주며(α=0.01), 11Pa는 11Pm보다는 Pa와 더 유사함을 보여준다.

3.3 단순척도

11-포인트 평균정확률은 P와 R에 비해 순위화된 검색결과를 평가하는데 상대적으로 분별력이 좋으나, 복잡한 계산을 필요로 하는 만큼 간단한 계산을 요구하는 평가척도에 비해 그 만큼 더 좋은 분별력을 보여준다고 말하기는 어렵다. 〈표 4〉의 S3과 같이 10개의 문헌을 검색하여 1등부터 4등까지만 적합문헌인 경우 11-포인트 평균정확률 11Pa는 95%이지만, 4개의 문

헌을 검색하여 4개 모두 적합문헌인 경우(S14)의 11Pa는 100%이다. 10개의 문헌을 검색하여 10개 모두가 적합문헌일 경우에도(S1) 11Pa는 100%이다. 이렇게 비교하는 탐색결과 set의 문헌수가 동일하지 않을 때 First-n P는 11Pa 보다 더욱 분별력이 있는 성능측정 척도인 것으로 보인다.

본 연구에서 First-10 P는 1등~2등 적합문헌에는 10점, 3~5등 적합문헌에는 8점, 6~10등 적합문헌에는 5점을 부여하여 P값을 다음과 같이 변화시켰다.

$$\text{First-10 P} = \frac{(1\sim2\text{등 }10) + (4\sim10\text{등 }8) + (6\sim10\text{등 }5\text{점})}{69-[(10-\min(\text{검색된 문헌수}, 10))\ 5]}$$

〈표 5-a〉에서 보는 바와 같이 First-10 P는 검색된 문헌수가 같을 경우에는 11Pa나 Pa와 높은 상관관계를 보이고 있다. 이는 P가 11Pa나 Pa와 갖는 상관계수보다 더 높은 수치이다. 그러나 〈표 5-b〉에서와 같이 검색된 문헌수가 다를 경우 First-10 P는 11Pa나 Pa와는 상관관계를 보이지 않고 P나 $(r-\bar{r})/e$와 유사한 상관관계를 보이고 있다.

적합문헌의 평균등수 e를 기반으로 한 척도 P/e와 r/e는 검색된 문헌수가 일정할 때는 동일하다. 〈표 3〉에서 S12를 S11보다 우수하게 잘못 평가하고 있지만 비교적 좋은 분별력을 보이고, 11Pa나 Pa와 가장 높은 상관관계를 보이고 있다(〈표 5-a〉 참조).

그러나 검색된 문헌수가 다를 때는 r/e가 P/e보다 좋은 척도임을 볼 수 있다. 〈표 4〉에서 P/e는 S14를 S1보다 우수하게, S17를 S16보다 우수하게 잘못 평가하고 있다. 〈표 5-a〉에서 P/e는 다른 어떤 척도와도 유의한 상관관계를 갖지 못함을 보여주고 있으나, r/e는 11Pa, 11Pm, Pa와 높은 상관관계를 보여주고 있다.

〈표 3〉과 〈표 4〉에서 $(r-\bar{r})/e$는 부적합문헌에 가중치 $\alpha=0.1$을 부여한 것으로$\left(=\frac{r-0.1\bar{r}}{e}\right)$, 검색된 문헌수가 동일 할 때나 동일하지 않을 때 모두 매우 판별력이 좋은 순위화 측정척도임을 볼 수 있다. 〈표 3〉에서 전체 13개 탐색결과의 $r-\bar{r}$이 음수값이 되지 않도록 부적합문헌에 0.1의 가중치만을 부여한 것이다(S12와 S13에서 $r=1$, $\bar{r}=9$이므로).

4.11-포인트 평균정확률을 대신할 간편한 평가척도

〈표 5-a〉는 상위 n개로 순위화된 문헌을 제한하여 시스템을 평가할 때의 사례 〈표 3〉에서 측정척도간의 상관관계를, 〈표 5-b〉는 불리언시스템에 기반을 둔 환경에서 검색된 문헌을 순위화할 때와 같이 검색된 문헌수가 일정치 않을 때의 사례 〈표 4〉에서 측정척도간의 상관관계를 나타낸다.

〈표 5-a〉의 경우 e를 제외하고 모든 측정척도가 매우 유의한 상관관계를 보이고 있다. r^2/n만 0.05 수준에서, 기타 모든 측정척도는 0.01 수준에서 유의한 상관관계를 보이고 있다. 11Pa와 가장 높은 상관관계를 보인 것은 Pa이고, 그 다음 11Pm, r/e, $(r-\bar{r})/e$순으로 높은 상관관계를 보이고 있다.

〈표 5-b〉에서 보는 바와 같이, 검색된 문헌수가 일정치 않은 사례의 경우 e와 P/e는 다른 측정척도와 상관관계가 높지 못함을 볼 수 있다. 〈표 5-a〉에서와 마찬가지로 Pa는 11Pa와 상관관계가 가장 높았으며, 단순측정척도 r/e와 $(r-\bar{r})/e$는 11Pa와 상관관계가 가장 높았다. 그러나 〈표 5-a〉와는 달리 P와 r^2/n은

〈표 5-a〉 〈표 3〉의 데이터를 사용한 측정척도 간의 상관계수

	e	first-n P	p	11Pa	11Pm	Pa	P/e	r^2/n	r/e	$(r-\bar{r})/e$
e										
first-n P	-1.58									
p	-.006	980**								
11Pa	-.354	.863**	.766**							
11Pm	-.423	.838**	.730**	.993**						
Pa	-.326	880**	.789**	.999**	.990**					
P/e	-.442	.895**	.797**	.981**	.976**	.981**				
r^2/n	128	.901**	.964**	.606*	.564*	.635*	.650*			
r/e	-.442	.895**	.797**	.981**	.976**	.981**	1.000**	.650*		
$(r-\bar{r})/e$	-.279	.947**	.876**	974**	.954**	.979**	.980**	737**	.980**	

* 유의수준 0.05
**유의수준 0 01

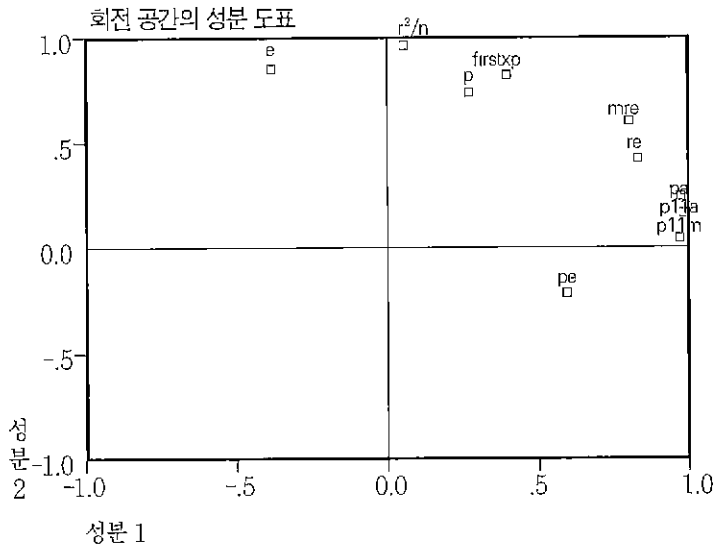
〈표 5-b〉 〈표 4〉의 데이터를 사용한 측정척도 간의 상관계수

	e	first-n P	p	11Pa	11Pm	Pa	P/e	r^2/n	r/e	$(r-\bar{r})/e$
e										
first-n P	.472									
p	.413	.963**								
11Pa	-.192	.453	.284							
11Pm	-.260	.355	181	.993**						
Pa	-.175	.508	.347	.997**	985**					
P/e	- 600	.370	.470	.421	.400	.461**				
r^2/n	.875**	.799*	.705	.252	.167	.283	-225			
r/e	158	.557	.346	.931**	.908**	.927**	.130	.533		
$(r-\bar{r})/e$	.236	.786*	.620	881**	.830**	.902**	.298	656	.946**	

First-n P와는 높은 상관관계를 보이나 11Pa-유사측정척도(11Pm, Pa, r/e 등 포함)와는 유사도가 낮은 것으로 나타났다. 〈표4〉에서 보는 바와 같이 이 차이는 사례 S3에서 기인된 것으로 보인다. S3은 P와 First-n P에서는 최하의 탐색결과이나, 11-포인트 P에서는 S18보다 좋은 탐색으로 산정되었기 때문이다.

〈그림 1〉은 〈표 4〉의 데이터를 2차원으로 요인분석한 것이다. 측정척도가 First-n P 유사그룹(P와 First-n P)과 11Pa 유사그룹으로 분류됨을 보여준다. $(r-\bar{r})/e$는 두 그룹과 모두 유사관계를 갖고 있음을 보여준다.

〈표 5-c〉는 검색된 문헌수가 다를 때 사례를 달리할 경우 성능측정척도간의 유사도가 어떤 영향을 받는지를 보기 위해 〈표 4〉의 7건의 사례에 〈표 3〉의 11건의 사례를 통합하여 18건의 사례를 대상으로 척도간의 상관관계를 산출한 것이다. 불리언탐색에서 순위화결과를 10개로



〈그림 1〉 〈표 4〉의 데이터 요인분석

제한검색할 때 10개 이상 검색된 사례는 13건(S1-S13), 10개 이하 검색된 사례는 5건(S14-S18)인 경우로 생각할 수 있다. 〈표5-c〉에서 보는 바와 같이 척도간의 상관계수는 10건 검색한 사례가 13건으로 10개 이하 검색한 사례 5건보다 많기 때문에, 〈표5-c〉에서의 상관관계는 〈표5-a〉의 상관관계와 더 비슷한 경향을 보인다.

전체적으로 이 세 상관관계표에서 일관성있

〈표 5-c〉 〈표 3〉과 〈표 4〉의 데이터를 통합하여 사용한 측정척도 간의 상관계수

	e	first-n P	p	11Pa	11Pm	Pa	P/e	r^2/n	r/e	(r - r̄)/e
e										
first-n P	-.310									
p	-.219	.981**								
11Pa	-.456	.881**	798**							
11Pm	-.509*	.848**	.755**	.994**						
Pa	-.437	.897**	820**	.998**	.990**					
P/e	-457	.860**	872**	812**	.789**	.830**				
r^2/n	035	.838**	.838**	.626**	.586*	.646**	.502*			
r/e	-.520*	.893**	.801**	.982**	976**	.979**	777**	.672**		

게 11Pa는 11Pm보다는 Pa와 더 유사한 측정척도이며, 단순측정척도 중 r/e 혹은 $(r-\bar{r})/e$가 11Pa나 Pa, 11Pm과 가장 유사하며 유의한(α=0.01) 상관관계를 갖는 것을 볼 수 있다.

18개의 결과 set을 서열화시켜 스피어만의 상관계수로 e를 제외한 9개의 측정척도 간의 관계를 분석한 결과에서도 모든 측정척도가 0.01 수준에서 높은 상관관계를 보였다. 11Pa는 Pa, $(r-\bar{r})/e$, First-n P, r/e, 11Pm, P, P/e, r^2/n 순으로 높은 상관관계를 보였다.

5 결 론

순위화된 검색결과를 제공하는 시스템에서 좋은 시스템은 보다 많은 적합문헌을 부적합문헌보다 먼저 상위에 출력하는 시스템이다. 순위화된 웹탐색시스템의 출현으로 순위화시스템은 우리와 친숙하게 되었다. 전통적으로 순위화시스템을 평가하는 실험연구에서 11-포인트 P가 대표적인 평가척도로 사용되어 왔으나 그 복잡성 때문에 웹탐색엔진과 같은 순위화시스템을 평가하는 연구에서도 비순위화시스템의 대표적인 평가척도 R과 P가 사용되는 것을 볼 수 있다.

본 연구에서는 복잡한 11-포인트 P를 대신할 수 있는 간편한 평가척도의 가능성을 연구하였다. 단순 평가척도의 가능성은 평가환경에 따라 차이가 있었다.

검색된 문헌수를 동일하게 제한했을 때 P와 r^2/n는 적합문헌의 위치까지도 반영하는 11-포인트 P나 First-n P, P/e, r/e 척도보다 분별력은 낮지만 유의한 상관관계를 지닌 것으로 보였다.

그러나 순위화된 완전일치 불리언탐색과 부분일치 유사도탐색시스템을 비교하는 연구에서와 같이 검색된 문헌수가 동일하지 않을 경우, P와 r^2/n은 First-n P와 높은 상관관계를 갖고, Pa와 r/e는 11-포인트 P와 높은 상관관계를 갖는 것으로 보였다. 제안된 $(r-\bar{r})/e$는 First-n P그룹과 11-포인트 P그룹 모두와 높은 상관관계를 보였다.

검색된 문헌수가 동일하든 동일하지 않든 어떤 척도가 다른 척도보다 더 분별력이 있는지 단정하기란 쉬운 일이 아니다. 예를 들어 상위 10개 문헌중 4등과 5등 문헌이 적합문헌일 경우와 2등과 7등 문헌이 적합문헌일 경우 어느 것이 더 좋은 탐색결과인지는 판단하기 매우 어렵기 때문이다. 〈표 4〉에서 S3과 S18 중 어느 것이 더 좋은 검색결과인지 먼저 판정이 되지 않고는 S3을 더 상위로 출력하는 척도가 좋은지 S18을 더 상위로 출력하는 척도가 더 좋은지 분별하는 것은 불가능하다. 이에 대한 보다 깊은 연구가 필요할 것이다.

본 연구의 결론은 논리적으로 측정단위를 설명하는데 필요한 사례를 인위적으로 가정하여 검증한 것이기 때문에 실제 대규모의 DB에서 검색된 데이터에서도 타당한지 연구가 필요하다.

참고문헌

고미영, 정영미. 1999. P-norm 검색의 문헌순위 화기법에 관한 실험적 연구. 『정보관리학회지』, 16(1): 7-30.

이명희. 1997. 네트웍 데이터베이스에서의 주제별 디렉토리와 키워드 검색엔진의 검색효율에 관한 탐색적 연구. 『한국문헌정보학회지』, 31(2): 177-197.

이명희. 1998. 교육학 분야 주제전문가와 탐색전문가의 인터넷 검색엔진을 사용한 정보탐색 형태 비교 연구. 『한국문헌정보학회지』, 32(3): 5-22.

이은주, 정영미. 1997. WWW 탐색도구의 검색성능에 관한 실험적 연구. 『제4회 한국정보관리학회 학술대회 논문집』, 59-62.

오삼균, 박희진. 2000. 국내 인터넷 탐색엔진에 대한 이용자 중심의 평가에 관한 연구: 한글 알타비스타와 네이버를 중심으로. 『한국문헌정보학회지』, 34(2): 117-133.

우유미, 정영미. 1998. 웹 검색엔진의 피드백 기능 평가. 『제5회 한국정보관리학회 학술대회논문집』, 69-72.

정영미, 김성은. 1997. WWW 탐색도구의 색인 및 탐색기능 평가에 관한 연구.『한국문헌정보학회지』, 31(1): 153-174.

Chignell, Mark H., Jacek Gwizdka, & Richard C. Bodner. 1999. "Discriminating Meta-Search: a Framework for Evaluation." *Information Processing & Management*, 35: 337-362.

Chu, H. & M. Rosenthal. 1996. "Search Engines for the World Wide Web:A Comparative Study and Evaluation Methodology." In: *ASIS 1996 Annual Conference Proceedings*, Baltimore, MD, October 19-24, 1996, 127-135. 〈http://www.asis.org/annual-96/Electronic Proceedings/chu.html〉

Clarke, Charles L. A. et. al. 2000. "Relevance Ranking for One to Three Term Queries." *Information Processing and Management*, 36: 291-311.

Clarke, W. & P. Willett. 1997. "Estimating the Recall Performance of Web Search Engines." *Aslib proceedings*, 49(7): 184-189.

Cooper, W. S 1968. "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on Weak Ordering Action of Retrieval Systems." *JASIS*, 19: 30-41.

Ding, W & G. Marchionini. 1996. "A Comparative Study of Web Search Service Performance." In: *ASIS 1996 Annual conference Proceedings*, Baltimore, MD, October 19-24, 1996, 136-142.

Frants, V., J. Shapiro & V. G. Voiskunskii. 1993 "Multiversion Information Retrieval Systems and Feedback with Mechanism of Selection." *JASIS*, 44(1): 19-27.

Gauch, S. & G. Wang. 1996. "Information Fusion with Profusion." *Webnet 96 Conference*, San Francisco, CA, Octover 15-19, 1996. http://curry.edschool.virginia.edu/aace/conf/webnet/html/155.htm〉

Leighton, H. V. & J. Srivastava. 1999. "First 20 Precision Among World Wide Web Search Services(Search Engines)." *JASIS*, 50(10): 870-881.

Losee, R. M. 1994. "Upper Bounds for Retrieval Performance and Their Use Measuring Performance and Generating Optimal Boolean Queries: Can It Get Any Better Than This?" *Information Processing & Management*, 30(2): 193-203.

Lowley, S. 2000. "The Evaluation of WWW Search Engines." *J. of Documentation*, 56(2): 190-211.

Timaiuolo, N. G. & J. G. Packer. 1996. "An Analysis of Internet Search Engines: Assessment of Over 200 Search Queries." *Computers in Libraries*, 16(6). 〈http://neal.ctstateu.edu:2001/htdocs/websearch.html〉