A Study of the Education of Information Specialists

Sung Jin Choi*

Abstract

The purpose of this study is to evaluate the information science education provided by the undergraduate courses of the departments of library science of the Korean universities by looking at major topics included in the syllabi distributed to students in the past three years. It is important to determine the evaluation of the professional education for information specialists by the graduates of the departments of library science who have acquired a critical appreciation of their professional studies and speak from experience about the relavance of the programme to their work and careers, and by the managers of information service units where the graduates would eventually make their careers. Specifically, the study addresses the following four questions. (a) To what excent do the information science curricula contribute to advancement of theory and practice of the information profession? (b) To what extent do the information science curricula contribute to students in acquiring the knowledge and skills required of the information specialist? (c) To what extent are the employers' concerns reflected in the information science curricula? (d) What reforms are needed to bring the current information science

* Professor, Sung Kyun Kwan University.

curricula closer to the present and future needs of the information profession?

To answer these questions, the study is conducted in two main parts: an in-depth subject analysis of the articles of three important journals in the field of information science published during the past ten years and of the syllabi used for information science subjects taught in the departments of library science during the past three years and an extensive survey of the graduates of departments of library science and their principal employers. The major findings are as follows. The average number of 4.1 subjects of information science is offered in departments of library science, and the most common subjects offered are introduction to information science, information storage and retrieval, and library automation. Approximately two thirds of the total output of research and development in the field of information science are taught at one or more departments of library science in Korea. Majority of the graduates of the departments of library science comment that their professional education did not offer to them systematic orientation to the specifics of the first job. The employers of the graduates believe that departments of library science should provide sufficient practicums to enable students to understand and apply the theory.

검색효율 측정척도에 관한 연구

윤 구 호*

<목 차>

1. 서 론

제2차 세계대전 이후 과학기술의 고도의 발전으로 인한 정보량의 기하급수적 증가와 수많은 정보이용자들의 다양화된 정보요구는 기존의 도서관 및 정보시스템에 많은 혁신을 초래하였다. 소위 정보화시대에 대처할 수 있는 새로운 정보검색기법의 개발을 촉진하였다.

1950년대초 토오브(Taube), 무어스(Mooers) 등에 의한 새로운 색인작성법의 시도와 때마침 실용화되기 시작한 컴퓨터의 활용은 정보검색시스템의 고도의 기계화 내지는 자동화를 전제로 한 보다 효율적인 정보시스템 구축을 촉진함으로써 이를 위한 시스템 평가는 필연적으로 수행되어야 했다. 따라서 보다 효율적인 평가기법의 개발을 위한 연구와 실험이 꾸준히 시도되어 왔다.

그러나 정보검색시스템 평가에 있어 가장 중요한 문제는 적용된 평가기준과 평가방법의 객관적 타당성 및 수량적 측정의 가능성이라고 본다. 오늘날 가장 널리 적용되고 있는 평가기준은 검색효율, 검색시간 및 검색비용이다.

* 계명대학교 도서관학과 부교수

이 중에서 검색효율을 제외한 두 기준은 비교적 용이하게 측정할 수 있기 때문에 별로 문제가 되지 않는다. 그러나 정보의 '적합성'이라는 개념이 개입되는 검색효율은 이용자의 정보요구에 대한 적합성 판정이 최종적으로는 이용자에 의해 결정되므로 다분히 주관적일 수 밖에 없을 뿐만 아니라 또한 여러가지 변수에 의해 영향을 받기 때문에 객관적 타당성이나 수량적 측정 기법이 인정받기 어려워 뚜렷한 의견의 일치를 보지 못하고 논란의 대상이 되어 왔다.

검색효율(retrieval effectiveness)은 정보검색시스템 평가의 제일기준으로서 이는 이용자가 요구하는 수준의 정보서비스를 제공하는 시스템의 능력을 측정하는 척도다. 즉 정보서비스의 질(quality)에 대한 평가척도인 것이다. 이용자의 궁극적인 목적이 필요한 적합문헌을 검색해 내는 것이므로 여러가지 평가기준 가운데서 검색효율이 가장 중요한 기준으로 간주되고 있으며 이에 관한 많은 연구와 실험이 수행되어 왔다.

본고는 검색효율 측정에 있어서 가장 중요한 개념이 되는 정보의 '적합성'(relevance)에 관해 고찰해 보고, 지금까지 연구개발된 다양한 검색효율 측정척도 중에서 괄목할만한 척도들의 측정기법을 상세하게 검토 분석하고자 한다. 본고가 앞으로 반드시 수행되어야 할 우리나라 도서관 및 정보시스템의 평가실험에 다소나마 도움이 되기를 바란다.

2. 정보의 적합성

정보의 적합성(relevance 또는 pertinence)이란 이용자의 특수한 정보요구를 만족시켜 줄 수 있는 정보의 특성으로서 검색효율 측정의 바로미터가 되는 것이다. 즉 검색된 문헌의 적합성 판정결과에 따라 시스템의 검색효율이 측정되는 것이다.

정보검색시스템의 검색효율이 이용자의 특정 정보요구를 만족시켜 줄 수 있는 적합정보(실제로는 적합정보가 들어 있는 적합문헌)의 검색능력과 적합하지 않은 부적합정보(실제로는 부적합문헌)의 검색배제능력으로 측정되

기 때문에 이용자의 정보요구에 대해 시스템이 탐색할 문헌의 적합성 여부의 판단 즉, '적합성 결정'(relevance decision)과 또한 검색된 문헌이 이용자의 요구를 어느 정도까지 충족시켜 줄 수 있느냐 하는 '적합성의 정도'(degree of relevance)는 검색효율 측정의 열쇠가 된다.

이와 같은 적합성의 개념정의, 적합성의 결정 및 적합성의 정도 등의 문제는 필자의 선행연구[1]에서 이미 연구 분석되었으므로 본고에서는 생략한다.

정보의 적합성 판정은 시스템 위주의 판정과 이용자 위주의 판정으로 구분될 수 있다. 전자는 시스템의 정보전문가나 혹은 주제전문가에 의해 결정되므로 보다 객관적인 판정이라고 할 수 있으며, 후자는 이용자에 의해 결정되므로 특정시점에서의 이용자의 지식상태 또는 지식상태의 변화 등이 개입될 수 있어 가변적이고 일시적인 것으로서 보다 주관적인 판정이라고 할 수 있다. 시스템에 의한 객관적 적합성은 탐색전이나 탐색도중의 이용자의 지식상태는 고려하지 않은 것으로서 결국 시스템에 제시된 이용자의 질문과 검색된 문헌간의 관련도를 나타내는 것이며, 이용자에 의한 주관적 적합성은 질문으로 표현되기 이전 단계의 잠재적인 정보요구 또는 탐색도중 재형성된(변화된) 정보 요구와 검색된 문헌과의 관련도를 나타는 것이다.[2]

한편, 정보시스템 설계자들이 적합성 판단능력을 향상시키기 위한 컴퓨터의 활용방법을 연구검토하였지만 적합성 판정은 시스템이나 이용자를 막론하고 다분히 주관적일뿐 어떤 수량적이거나 객관적인 기준이 없기 때문에 별로 성과를 거두지 못하고 있는 실정이다. 스타일스[3]의 적합성 산정기법이나 가필드[4]의 인용문헌색인법에 의한 적합정보의 제시방법 등 상당수의 적합성 측정기법이 제시되었으나 대부분의 검색시스템이 이러한 기법들을 적용하지 않고 오히려 정보요구자에게 적합성의 결정을 맡기고 있으며 가끔 시스템의 성능측정을 위한 수단으로서 정보요구자의 적합성 판정을 통보해

1) 윤구호, 情報檢索效率에 관한 硏究, 圖書館學, 8집(1981). p.73~101.
2) 정영미, 정보검색론, 서울, 정음사, 1988. p.296.
3) Stiles, H.E., The association factor in information retrieval. *J. of the Association for Computing Machinery*, 8(1961) : p.271~279.
4) Garfield, E., Citation indexing; a natural science literature retrieval system for the social sciences. *Am. Behavioral Scientist*, 5(1964) : p.58~61.

줄 것을 요구하고 있다.[5] 이는 적합성 판정이 최종적으로는 이용자의 주관적 판단에 의해 결정될 수 밖에 없기 때문이라고 본다.

3. 크랜필드 척도(Cranfield measures)

검색효율은 이용자의 정보요구를 만족시켜 줄 수 있는 문헌을 검색해 내는 시스템의 능력으로 이는 결국 검색된 적합문헌과 부적합문헌의 비율로서 측정되는 것이다. 따라서 검색효율의 측정은 시스템이 탐색할 문헌파일(데이타베이스)을 적합문헌과 부적합문헌으로 구분하여 나타난 검색결과가 수량적으로 표시되어야 하기 때문에 도표 1과 같은 분할표(contingency table ; 일명 2×2테이블(two by two table)이라고도 함)가 일반적으로 널리 사용되고 있다.

〈도표 1〉 문헌파일의 분할표

	적 합 문 헌(r)	부적합문헌($\bar{r}$)	전 체 문 헌(N)
검색된문헌 (R)	적중(a)	잡음(b)	$a+b$
검색되지 않은 문헌 ($\bar{R}$)	누락(c)	배제(d)	$c+d$
전 체 문 헌(N)	$a+c$	$b+d$	$a+b+c+d$

이와 같은 분할표로부터 검색효율의 측정척도가 다양하게 제시되었으나 가장 일반적으로 적용되고 있는 척도는 '크랜필드 척도'라고 알려진 재현율(recall ratio)과 정확률(precison ratio)이다. 이들의 산출공식은 다음과 같다.

$$\text{재 현 율} = \frac{\text{검색된 적합문헌수}}{\text{전체의 적합문헌수}} = \frac{a}{a+c}$$

$$\text{정 확 률} = \frac{\text{검색된 적합문헌수}}{\text{검색된 전체문헌수}} = \frac{a}{a+b}$$

5) Meadow, C.T., The Analysis of Information Systems, 2nd ed. Los Angeles, Melville, 1973. p.169.

$$\text{적 합 률}^{*}=\frac{\text{전체의 적합문헌수}}{\text{전체의 문헌수}}=\frac{a+c}{a+b+c+d}$$

$$\text{부적합률}=\frac{\text{검색된 부적합문헌수}}{\text{전체의 부적합문헌수}}=\frac{b}{b+d}=1-\text{제거율}$$

$$\text{잡 음 률}=\frac{\text{검색된 부적합문헌수}}{\text{검색된 전체문헌수}}=\frac{b}{a+b}=1-\text{정확률}$$

$$\text{누 락 률}=\frac{\text{검색되지 않은 적합문헌수}}{\text{전체의 적합문헌수}}=\frac{c}{a+c}=1-\text{부적합률}$$

$$\text{배 제 율}=\frac{\text{검색되지 않은 부적합문헌수}}{\text{전체의 부적합문헌수}}=\frac{d}{b+d}=1-\text{재현율}$$

재현율과 정확률척도는 1955년 켄트[6] 등에 의해 제안된 것이나 클레버든[7]의 크랜필드 평가실험 이후 일반적으로 '크랜필드 척도' (Cranfield measures)라고 알려져 널리 사용되고 있다.

재현율은 시스템이 적합문헌을 검색하는 능력을 나타내며 정확률은 시스템이 부적합문헌을 검색하지 않는 능력을 나타내는 것으로, 결국 전자는 검색의 완전성을, 후자는 검색의 정확성을 측정하는 것이라고 말할 수 있다.

재현율과 정확률에 의한 효율측정은 시스템의 문헌파일의 크기와 그 안에 있는 적합문헌에 의해 영향을 받는다고 볼 수 있다. 따라서 이 비율을 나타내는 적합률은 재현율과 정확률에 영향을 미칠 뿐만 아니라 또한 부적합률에도 영향을 미칠 것이다. 이는 페어손[8]에 의해 최초로 암시되었다.

이용자의 정보요구에 대해 시스템이 탐색할 문헌파일(N)이 일정한데 적합률(G)이 상승하면(또는 G가 일정한데 N이 감소하면) 검색된 적합문헌수는 평균적으로 증가될 것이다. 따라서 적합률의 상승은 재현율(R), 정확률(P), 부적합률(F)에 다음 그림과 같이 영향을 미칠 것이다.

$$R=\frac{a}{a+c}=\frac{\uparrow}{\uparrow},\quad P=\frac{a}{a+b}=\frac{\uparrow}{\rightarrow},\quad F=\frac{b}{b+d}=\frac{\rightarrow}{\rightarrow}$$

---

* 'Generality'로서 '보편율'이라고 표현할 수 있으나 의미를 명확하게 하기 위해 '적합률'이라고 하였음.

6) Kent, A. et al. Machine literature searching. *Am. Doc.*, 6(1955) : p. 93~101.

7) Cleverdon, C.W., Progress in documentation: evaluation tests of information retrieval systems. *J. of Doc.*, 26(1970) : p. 55~67.

8) Fairthorne, R.A., Basic parameters of retrieval tests. *Proceedings of the Am. Doc.*

즉, 분모와 분자가 같은 방향으로 움직이는 재현율과 부적합률은 적합률에 변화해도 논리적으로 대략 일정한 값을 유지한다고 보겠으나 정확률은 적합률의 변화에 따라 직접적으로 값이 달라질 것이다.

이와 같은 관점에서 재현율과 정확률 척도가 바람직하지 못하다는 반론이 상당히 제기되었으며 상대적으로 시스템이 탐색할 문헌파일과 적합률의 변화에 별로 영향을 받지 않는 재현율과 부적합률이 새로운 척도로서 제안되었다.[9]

시스템의 문헌파일내 적합문헌의 밀도를 나타내는 적합률이 효율측정에 개입됨으로써 재현율, 정확률, 적합률, 부적합률 사이에는 '베이스의 정리'(Bayes Theorem)와 같은 함수 관계가 성립한다.[10] 따라서 이들은 상호 독립적인 것이 아니고 어느 것이든 세가지의 값을 알면 나머지 하나의 값은 자동적으로 결정된다.

$$P=\frac{R\times G}{R\times G+F(1-G)}$$

재현율과 정확률에 의한 검색효율의 평가는 이용자 지향적이라고 볼 수 있다. 왜냐하면 이 척도는 시스템의 적합문헌의 검색능력만을 나타내는 것으로 이용자는 일반적으로 적합문헌의 검색을 극대화하는데 관심이 높기 때문이다. 그러나 부적합률은 부적합문헌의 배제능력(비검색능력)의 측정으로 문헌파일 안에 있는 부적합문헌수에 의해 영향을 받는다. 따라서 재현율과 부적합률은 적합문헌의 검색능력과 부적합문헌의 검색배제능력을 나타내기 때문에 시스템 지향적이라고 볼 수 있다.

두가지 척도가 모두 특수한 환경이나 목적을 위해서 장단점을 가지고 있기 때문에 어느 척도가 보다 합리적이라고 단정할 수는 없다. 그러나 시스템의 궁극적 목표가 이용자를 위해서 존재한다고 본다면 이용자 지향적인

Inst., 1(1964) : p. 343~345.

9) van Rijsbergen, C.J., Retrieval effectiveness. In: Progress in Communication Science, ed. by M.J. Voigt and G.J. Hanneman. Norwood(N.J.), Ablex Pub. Co., 1979. p. 96~97.

10) ibid., p. 95.

검색효율의 향상이 보다 바람직하다는 관점에서 아마도 재현율과 정확률 척도가 훨씬 보편적으로 활용되고 있다고 볼 수 있을 것이다.

그러나 샐튼[11]은 재현율-정확률, 또는 재현율-부적합률의 어느 척도도 모든 상황에 다 만족스러운 것은 아니라고 지적하고 검색효율 측정척도의 요건을 다음과 같이 제시하였다.

1) 검색비용처럼 독립적인 기준으로서 단지 검색효율만을 나타낼 수 있어야 한다.
2) 특정탐색에서 검색된 문헌의 수와 같은 특정 검색기준치와는 무관해야 한다.
3) 재현율-정확률처럼 두개의 값 대신에 하나의 값 즉, 단일가로 표현될 수 있어야 한다.

한 시스템의 검색효율은 충분한 수의 정보요구에 대한 각각의 재현율과 정확률의 평균치를 산출한 평균재현율과 평균정확률로 측정되어야 한다. 평균치를 구하는 방법에는 질문지향적 방법과 문헌지향적 방법이 있다. n개의 질문이 있을때 두 방법에 의한 평균재현율과 평균정확률의 산출공식은 다음과 같다.[12]

질문지향적 방법의

$$\text{평균재현율}=\frac{1}{n}\sum_{i=1}^{n}\text{재현율 } i=\frac{1}{n}\sum_{i=1}^{n}\frac{\text{검색된 적합문헌수 } i}{\text{전체의 적합문헌수 } i}$$

$$\text{평균정확률}=\frac{1}{n}\sum_{i=1}^{n}\text{정확률 } i=\frac{1}{n}\sum_{i=1}^{n}\frac{\text{검색된 적합문헌수 } i}{\text{검색된 전체문헌수 } i}$$

문헌지향적 방법의

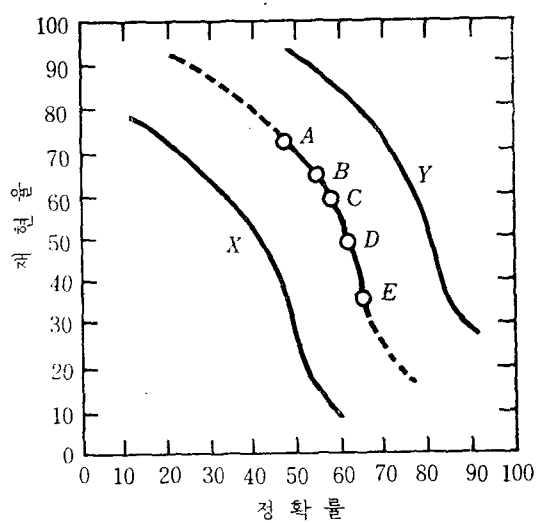
$$\text{평균재현율}=\frac{\sum_{i=1}^{n}\text{검색된 적합문헌수 } i}{\sum_{i=1}^{n}\text{전체의 적합문헌수 } i}$$

11) Salton, G. & McGill, M.J., Introduction to Modern Information Retrieval. N.Y., McGraw-Hill, 1983. p.177.
12) ibid., p.181.

$$\text{평균정확률} = \frac{\sum_{i=1}^{n} \text{검색된 적합문헌수 } i}{\sum_{i=1}^{n} \text{검색된 전체문헌수 } i}$$

크랜필드 평가실험은 재현율과 정확률간의 반비례관계를 최초로 논증하였으며 그 후 많은 시험과 연구에서 재확인 되었다.[13] 재현율과 정확률간의 상반관계를 그래프로 표시한 것을 성능곡선이라고 하는데 이러한 성능곡선은 도표 2와 같이 여러 쌍의 재현율과 정확률 값으로 작성된다.

〈도표 2〉 재현율과 정확률의 성능곡선



재현율과 정확률간의 성능곡선은 주제분야의 언어적 특성이나 이용자의 적합성 판정기준에 따라 차이가 있게 된다. 주제나 개념을 비교적 명확하게 표현할 수 있는 언어의 전문성이 높은 자연과학분야는 도표 2의 Y곡선과 같이 오른쪽으로 기울게 되며, 추상적이고 포괄적인 단어가 많은 인문·사회과학분야는 X곡선처럼 왼쪽으로 기울게 된다. 또한 SDI(selective dissemination of information)서비스의 경우에는 이용자 판정기준이 RS(retrospec-

13) Cleverdon, C.W. op. cit.

tive search)서비스의 경우보다 낮게되어 SDI검색의 성능곡선은 오른쪽으로, RS검색은 왼쪽으로 기울게 될 것이다.

색인언어의 특정성과 색인작성의 망라성도 성능곡선의 위치를 결정하는 특성으로서, 특정성이 높은 색인언어일수록 높은 정확률을 가져오며 망라성이 높을수록 재현율은 높아진다.

오늘날과 같은 정보의 홍수시대에는 이용자의 입장에서 볼 때 적합문헌을 구별하기 위해 시간과 노력이 더 요구되는 재현율보다는 덜 요구되는 정확률이 높은 검색을 선호할 수 있을 것이다. 따라서 이용자의 한정된 요구에 대한 시스템의 성능측정을 위해 다음과 같은 상대재현율을 적용할 수도 있을 것이다.

$$\text{상대재현율} = \frac{\text{검색된 적합문헌수}}{\text{이용자가 원하는 적합문헌수}}$$

상대정확률도 생각할 수 있으나 그 산출공식은 결국 상대재현율과 동일하게 된다.

4. 복합척도(Composite measures)

재현율과 정확률, 또는 재현율과 부적합률과 같이 한 쌍의 값 즉, 두가지 값으로 검색효율을 측정하는 기법은 불편하며 만족스럽지 못하다는 관점에서 연구된 방안이 복합척도와 단일가척도이다. 복합척도는 여러개의 척도를 결합시켜 단일가를 산출하는 기법이고 단일가척도는 별도의 측정기법을 사용하여 단일가를 산출한다.

여러가지 복합척도가 제시되었으나 이들이 정당화될 수 있는 합리성을 찾기란 어렵다. 재현율(R)과 정확률(P), 그리고 부적합률(F)을 결합시킨 몇가지 복합척도(CM)는 다음과 같다.[14]

14) van Rijsbergen, C.J., Information Retrieval, 2nd ed. London, Butterworths, 1979. p.154.

$$CM_1=P+R \qquad CM_2=P+R-1$$

$$CM_3=\frac{R-F}{R+F-2RF} \qquad CM_4=1-\frac{1}{2\left(\frac{1}{P}\right)+2\left(\frac{1}{R}\right)-3}$$

5. 단일가 척도(Single-number measures)

재현율과 정확률, 또는 재현율과 부적합률 등은 다음과 같은 문제점을 지니고 있다.[15] 첫째는 검색기준치의 변화에 따라 여러개의 다른 값들을 갖게 되고, 둘째는 한 쌍의 값 즉, 두개의 값이 함께 사용되어야 의미가 있으며, 세째는 시스템이 갖고 있는 적합문헌의 수 즉, 적합률로 표현되는 적합문헌의 밀도에 의해 영향을 받는다. 이러한 문제를 해결하기 위한 방법은 하나의 값으로 검색효율을 측정하는 것으로 이러한 척도를 단일가 척도라고 부른다.

단일가 척도로 잘 알려진 것으로는 스웨츠(Swets)의 모델, 쿠퍼(Cooper)의 모델, SMART 모델, 섀논(Shannon)의 정보이론에 기초한 몇가지 모델이 있다.

5.1. 스웨츠 모델

1963년 스웨츠[16][17]는 종래의 검색효율 측정척도에 대해 이의를 제기하고 잘 알려진 통계적 결정이론(statistical decision theory)에 기초한 척도를 제시하였다. 일반적으로 'E척도'(E measure)라고 불리우는 이 척도는 탐색질문의 일반성이나 검색된 문헌수에 영향을 받지 않는 단위척도로 평가되고 있다.

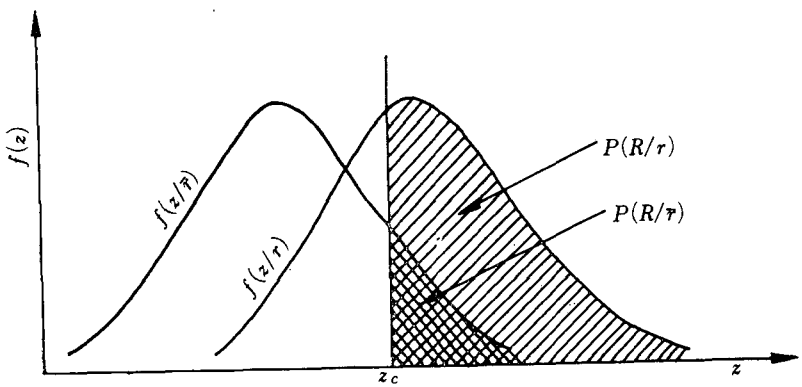
통계적 결정이론이란 표본집단을 두가지 확률분포가운데 어느 하나에 할당하는 문제와 관련된 것으로, 정보검색 문제에 대한 이 이론의 적합성은 마론과 쿤스[18]에 의해 최초로 관찰되었다. 스웨츠도 정보검색 과정을 결정

15) 정영미, op. cit., p.314.
16) Swets, J.A., Information retrieval system. *Science*, 141(1963) : p.245~250.
17) Swets, J.A., Effectiveness of information retrival system. *Am. Doc.*, 20(1969) : p.72~89.
18) Maron, M.E. and Kuhns, J.L., On relevance, probabilistic indexing and information

이론적으로 해석하였는데 그는 정보검색이란 특정질문과 관련하여 축적된 표본문헌을 적합성 판정에 따라 적합문헌을 검색하거나 부적합문헌을 배제하는 두가지 범주가운데 하나에 할당하는 것이라고 보았다.

스웨츠의 모델에서는 일반적으로 어떤 탐색질문이 시스템에 입력되면 시스템은 질문에 대한 적합성의 정도를 나타내는 색인가 z를 각 문헌에 부여하고, 다시 각 문헌이 특정한 z값을 가질 확률을 산출하여 확률분포함수 $f(z)$를 작성한다. 이때 각 문헌은 적합성 판정에 따라 적합문헌과 부적합문헌으로 구분되므로 실제로는 적합문헌의 확률분포함수와 부적합문헌의 확률분포함수가 생기게 된다. 도표 3은 이 두개의 함수와 색인가 z와의 관계를 보여주는 것으로서 x축은 z값에 의해 색인된 적합성의 정도를 나타내고 y축은 각 z값에 부여될 확률을 나타낸다.

〈도표 3〉 적합문헌과 부적합문헌의 확률분포함수



$f(z/r)$: 적합문헌의 확률분포함수

$f(z/\bar{r})$: 부적합문헌의 확률분포함수

$p(R/r)$: 적합문헌이 검색될 조건확률

$p(R/\bar{r})$: 부적합문헌이 검색될 조건확률

일반적으로 검색시스템은 변수 z의 절단가(cutoff value) 즉, 검색기준치

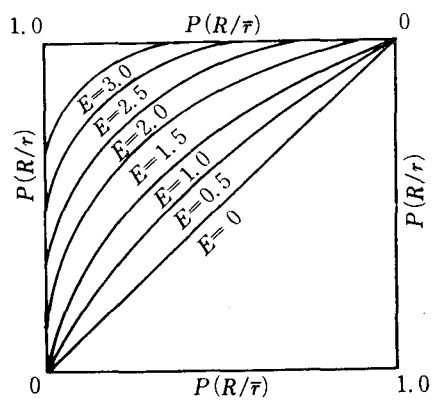
retrieval. *J. of the Association for Computing Machinery*, 7(1960) : p. 216~244.

z_c를 설정하여 z_c 값보다 큰 z값을 갖는 문헌은 모두 검색하고 z_c 값보다 작은 z값을 갖는 문헌은 모두 배제한다. 즉 도표 3에서 검색기준치 z_c가 넘는 문헌이 검색되는 것이며, 두개의 분포함수에서 빗금친 부분이 검색되는 문헌들의 비율을 나타내는 것이다.

스웨츠의 모델에서는 적합문헌이 검색될 조건확률인 $P(R/r)$와 부적합문헌이 검색될 조건확률인 $P(R/\bar{r})$이 효율측정의 기본 변수가 되는데 전자는 재현율을, 그리고 후자는 부적합률을 나타낸다. 이들은 검색기준치 z_c의 변화에 따라 달라진다. z_c를 낮추면 즉, 왼편으로 옮기면 $P(R/r)$이 증가되어 보다 많은 적합문헌이 검색되지만 동시에 $P(R/\bar{r})$도 증가되어 부적합문헌도 더 많이 검색되는 결과를 가져온다.

검색기준치 z_c 값의 변화에 따른 $P(R/r)$과 $P(R/\bar{r})$의 비율을 산출하여 그래프로 그리게 되면 도표 4와 같이 두 집단의 복합적 특성을 나타내는 곡선이 형성되는데, 결국 이 곡선은 재현율과 부적합률을 좌표로 하는 성능곡선으로 통계학에서는 '*OC*곡선' (Operating Characteristic Curve ; 검사특성치 곡선)이라고 한다.[19]

〈도표 4〉 *OC*곡선



19) Swets는 나중에 이 OC곡선을 'ROC곡선' (Relative-Operating Characteristic Curve)이라고 하였다. (Ref. : Swets, J.A., Effectiveness of Information Retrieval Methods. AM. Doc., 20(1969) : p.72~89).

도표 4에서 OC곡선이 대각선과 일치하는 경우는 두 집단($P(R/r)$과 $P(R/\bar{r})$의 분포비율이 일정하다는 것을 의미하며, 적합문헌과 부적합문헌의 두 문헌집단을 분리하는 능력이 큰 시스템일수록 OC곡선은 왼편 상단에 가까워진다. 이 OC곡선의 위치가 검색효율 측정척도로 사용되며 위치를 나타내는 한개의 값을 선택한 것이 E값이 되는 것이다.[20]

E값은 수학적으로는 두 문헌집단의 확률분포함수 $f(z/r)$와 $f(z/\bar{r})$의 평균간의 거리를 두 분포함수의 표준편차의 평균으로 나누어 준 값으로 공식은 다음과 같다.[21]

$$E = \frac{\mu_2 - \mu_1}{\frac{1}{2}(\sigma_1 + \sigma_2)}$$

$\mu_1 : f(z/r)$의 평균

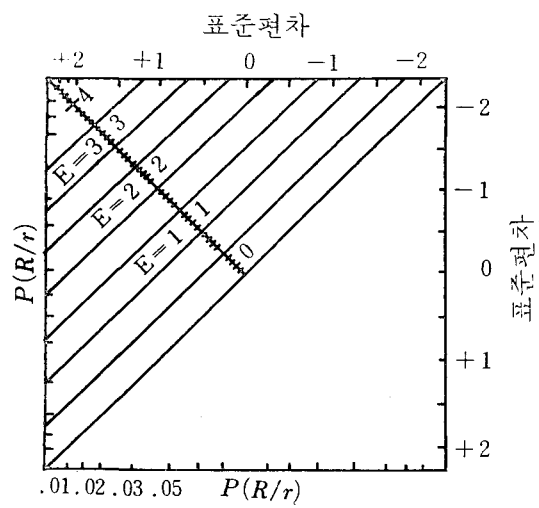
$\mu_2 : f(z/\bar{r})$의 평균

$\sigma_1 : f(z/r)$의 표준편차

$\sigma_2 : f(z/\bar{r})$의 표준편차

E값은 결국 적합문헌 집단과 부적합문헌 집단을 분리시키는 시스템의 능력을 나타내는 것으로 E값이 클수록 검색효율은 높아진다고 할 수 있다.

〈도표 5〉 이중확률지상의 OC곡선



20) Swets, J.A., Effectiveness... op. cit., p.75.

21) van Rijsbergen, C.T., Information Retrieval, op. cit., p.157.

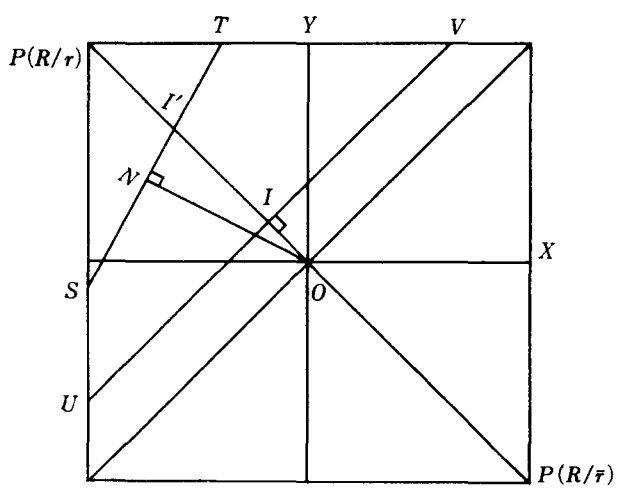
이 OC곡선을 도표 5와 같이 이중확률지에 그리면 직선을 이루게 되는데 각 직선이 갖는 E값은 x축과 y축의 표준편차의 차이가 된다.

스웨츠는 이중확률지상의 음대각선 OR에 눈금을 매겨놓고 E값을 구할 수 있도록 하였으며 실제로 E값은 대략 5.0에 가까운 최대값을 가질 수 있다고 하였다. 그러나 대부분의 경우 두 분포함수는 분산도가 다르므로 OC곡선의 기울기는 대각선과 일치하지 않게 된다.

도표 6에서 OC곡선 UV는 기울기가 1인 경우로서 이 때의 E값은 OC곡선과 대각선의 거리 OI(실제로는 $\sqrt{2}OI$)가 되는데 이는 두 분포함수의 분산도가 같을 때 갖게 된다. 그러나 OC곡선 ST는 기울기가 1보다 큰 경우로서 이처럼 기울기가 1이 아닌 경우에는 효율척도로서 E값만을 사용하는 것은 문제가 있으므로 기울기를 나타내는 S값을 두 분포함수의 표준편차의 비율로 구하여서 함께 사용하는 것이 바람직할 것이다.

스웨츠는 두 분포함수가 정규분포한다는 가정에 대한 반론에 대응하는 방법으로서 비정규분포함수의 경우를 위한 'A척도'를 추가하였다. A척도란 도표 5에서 각 OC곡선의 아래부분의 면적의 비율로서 측정되는데, 이때

〈도표 6〉 두개의 OC곡선



A값은 적합문헌 집단과 부적합문헌 집단으로부터 임의로 추출된 한쌍의 문헌들 즉, 하나의 적합문헌과 하나의 부적합문헌으로부터 적합문헌을 선택하고자 할 때 시스템이 올바로 선택할 수 있는 비율과 같다고 하는 개념적 목적(conceptual purposes)을 위해 유용한 속성이라고 하였다.[22] 따라서 스웨츠는 S값에는 상관없이 E값만으로도 효율측정의 척도로서 사용할 수 있다고 하였다. 이를 입증하기 위해 스웨츠는 약 50가지의 상이한 검색방법을 세 종류의 실험적 검색시스템에 적용하였는데 그 결과로 얻어진 대부분의 OC곡선은 직선에 거의 일치함을 보여 주었다.[23]

브룩스[24]는 기울기가 1이 아닌 OC곡선을 위해 스웨츠의 모델을 수정하였는데 이것은 스웨츠의 공식에서 두 분포함수의 평균을 두 함수의 표준편차의 평균으로 나눈 대신 제곱근으로 나눈 것으로 그 공식은 다음과 같다. 따라서

$$S=\frac{\mu_2-\mu_1}{(\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}}$$

도표 6에서 기울기가 1이 아닌 OC곡선 ST의 E값은 OI' 대신 ON의 거리가 되는 것이다.

그러나 로버트슨[25]은 브룩스의 S척도는 사실상 스웨츠의 A척도와 같은 것으로 A의 표준편차라고 하였다. 한편 북스타인[26]은 스웨츠 모델을 일반적인 검색과정과 비교하여 분석하였다. 그는 시스템이 결정한 적합성과 이용자가 판정한 적합성을 구별하는 것이 필요하다고 보고 기준치 z_c로서 시스템이 채택하는 검색영역과 이용자에게 적합한 문헌이 실제로 포함되어 있는 영역을 비교분석하였다. 적합문헌의 표준편차가 부적합문헌의 그것보다

22) Swets, J.A., Effectiveness... op. cit., p.76~77.
23) ibid., p.72~89.
24) Brookes, B.C., The measure of information retrieval effectiveness porposed by Swets. *J. of Doc.*, 24(1968) : p.41~54.
25) Robertson, S.E., The parametric description of retrival tests, part 2, Overall measures. *J. of Doc.*, 25(1969) : p.94.
26) Bookstein, A., When the most 'pertinent' document should not be retrievedan analysis of the Swets model. *Inf. Processing & Management*, 13(1977) : p.377~383.

클 경우에는 가장 적합한 문헌뿐만이 아니고 적합성이 매우 낮은 문헌을 검색해야만 할 수도 있고, 부적합문헌의 표준편차가 클 경우에는 부적합문헌이 포함된 영역뿐만이 아니라 적합성이 매우 높은 문헌이 포함된 영역도 배제될 수 있다고 지적하였다. 따라서 두 함수의 표준편차가 같은 경우에만 적합문헌을 가장 높은 비율로 검색할 수 있다고 하였다. 그는 또한 두 분포함수가 정규분포한다는 가정에 반론을 제기하고, 실제로 두 분포함수의 표준편차는 다른 경우가 많으므로 스웨츠 모델을 사용할 경우에는 두 분포함수를 정규분포함수가 아닌 포아송분포함수로 바꿔야한다고 결론지었다.

그러나 스웨츠 모델은 다양한 값의 성능곡선 대신에 OC곡선을 직선화하여 E값을 생산할 수 있는 기법을 논증하였으며, 또한 잘 알려진 통계적 결정이론에 입각한 확률이론으로부터 유도한 척도를 제안함으로써 근본적으로 정보검색의 확률적 특성을 강조하였다고 본다.

5.2. 쿠퍼 모델

1968년 쿠퍼[27]는 "검색시스템의 주요기능은 적합문헌을 검색하기 위한 탐색에서 검색되어지는 부적합문헌을 훑어보아야 하는 이용자의 노력을 가능한 한 최대로 절약시켜 주는 것"이라고 보고 이 노력의 '절약'(saving)을 측정하기 위한 단일가 척도로서 '예상탐색길이 감소인자'(expected search length reduction factor)를 제시하였다.

일반적으로 '예상탐색길이척도'라고 불리어지는 이 척도는 이용자가 원하는 수의 적합문헌이 발견될 때까지 탐색해야만 하는 부적합문헌의 수에 기초한 것으로, 질문과 문헌간의 유사도 즉, 적합성의 수준에 따라 출력문헌의 순위가 주어지는 시스템의 검색효율 측정에 적합한 척도라고 볼 수 있다. 쿠퍼가 제시한 이 척도의 장점은 다음과 같다.[28]

1) 검색효율 측정을 위한 단일가 척도다.

27) Cooper, W.S., Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *Am. Doc.*, 19(1968) : p. 30~41.

28) ibid.

2) 약순(弱順 : weak ordering으로서 뒤에서 설명함)에 대한 수학적 개념을 적용하여 검색상태를 등급화하는 척도다.

3) 임의탐색(random searching)에 관련된 검색성능을 평가하는 척도다.

4) 이용자가 원하는 적합문헌의 수를 고려한 척도다.

쿠퍼는 효율측정기법에서 가장 중요한 요소는 이용자가 실제로 요구하는 적합문헌의 양이라고 보고 요구된 수 만큼의 적합문헌이 적합성의 수준에 따라 단계적으로 검색되는 기법을 고려해야 한다고 하였다.[29] 예컨대 후조합색인에서 N개의 용어(색인어)가 일치되어 검색되는 문헌은 $N-1$개의 용어가 일치되어 검색되는 문헌보다 적합성 수준이 높은 단계로서 먼저 검색되어야 할 것이다. 이와 같은 검색결과는 문헌들의 적합성 수준에 따른 순위를 형성한다고 볼 수 있다. 도표 7은 이용자가 원하는 8개의 적합문헌을 검색하기 위해 적합성 수준에 따라 단계적으로 검색한 20개의 문헌인데 결국 네번째의 수준에서 이용자의 요구는 만족되는 것이다.

〈도표 7〉 적합성 수준별로 출력된 20개의 검색문헌

수준 1	□□▨	
수준 2	▨□▨▨▨	▨ 적합문헌
수준 3	□▨▨□□	□ 부적합문헌
수준 4	□□□□□▨□	

쿠퍼 모델에서는 탐색길이(search length)와 예상탐색길이(expected search length)의 개념과 그들의 측정기법이 중요하다. 탐색길이란 특정질문에 대해 출력되는 각 문헌에 일련번호와 같은 단순(單順 : simple ordering)의 고유한 순위가 주어지는 시스템에서 이용자가 원하는 수 만큼의 적합문헌을 검색하기 위해 훑어보아야할 부적합문헌의 수를 의미한다. 도표 8과 같이 문헌이 단순으로 순위가 주어져 출력되고 각 문헌의 적합성이 적합(○표)과 부적합(×표)으로 판정되었을 때, 만일 이용자가 두 개의 적합문헌만을 원한다면 탐색길이는 2가 되고 6개의 적합문헌을 원한다면 탐색길이는 3, 8개의 적합문헌을 원한다면 탐색길이는 7이 되는 것이다.

29) ibid.

〈도표 8〉 단순으로 고유한 순위가 주어진 20개의 검색문헌

순 위	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
적 합 성	×	○	×	○	○	○	○	×	○	×	×	×	○	×	○	×	×	×	×	×

따라서 여러 검색시스템의 효율을 비교하고자 할 때에는 검색하고자 하는 적합문헌의 수를 동일하게 했을 때의 탐색길이가 짧은 시스템일수록 성능이 더 좋다고 할 수 있다.

그러나 질문과 문헌간의 유사도(적합성수준)를 측정하는 매칭함수(matching function)에 의해 생성되는 순위는 불행하게도 단순인 경우는 드물고 오히려 약순인 경우가 일반적으로 더 많다. '약순'이란 도표 7과 같이 적합성 수준이 같은 문헌들은 동일한 순위를 갖게 되는데 이처럼 각 수준별로 동일한 순위를 갖는 순서를 의미한다.

예상탐색길이란 출력되는 각 문헌이 고유한 순위를 갖지 않고 동일한 순위를 갖는 문헌들이 있는 경우 즉, 출력문헌이 약순으로 순위가 주어지는 경우 확률적으로 산출해 낸 평균탐색길이를 의미한다. 이러한 경우에는 동일한 순위를 갖는 문헌들이 도표 7과 같이 한 수준의 문헌집단을 형성하게 되며 한 수준의 문헌집단 내에서의 문헌배열순서는 랜덤하다고 볼 수 있다. 따라서 이용자가 원하는 수의 마지막 적합문헌이 속해있는 수준에서의 문헌배열에 따라 탐색길이가 달라지게 된다.

〈도표 9〉 약순으로 순위가 주어진 20개의 검색문헌

순 위	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	4	4	4	4
적 합 성	×	×	○	○	×	○	○	○	×	○	○	×	×	×	×	×	×	×	○	×

예컨대 특정질문에 대한 검색결과가 도표 9와 같이 약순에 의해 4개의 문헌집단으로 나뉘어지고 각 문헌의 적합성이 판정되었을 때, 만일 이용자가 원하는 적합문헌의 수가 6개라면 이용자의 정보요구는 세번째 수준(순위 3)의 문헌집단에 이르러 만족된다. 왜냐하면 첫번째와 두번째 수준에서 이미 5개의 적합문헌이 검색되었으므로 세번째 수준에서 나머지 1개의 적

합문헌이 검색되어야 한다. 이때 두번째 수준까지의 탐색길이는 첫번째와 두번째 수준의 문헌집단의 문헌배열에 상관없이 3이 되지만 세번째 수준의 집단에서는 문헌의 배열순서에 따라 탐색길이가 달라진다. 따라서 가능한 탐색길이는 6개의 적합문헌 앞에 몇개의 부적합문헌이 오느냐에 따라 3,4,5 또는 6 중의 하나가 될 것이다. 그런데 세번째 수준의 5개 문헌 가운데 2개의 적합문헌이 배열되는 방법은 $\binom{5}{2}=\frac{5!}{2!(5-2)!}=10$이 되므로 10가지의 다른 배열이 가능하다. 이 중에서 4개의 배열은 탐색길이가 0, 3개의 배열은 1, 2개의 배열은 2, 한개의 배열은 3이 되어 결국 전체의 탐색길이는 각각 3,4,5,6이 된다. 그런데 이들의 발생확률은 첫번째 4개의 배열은 4/10, 두번째 3개는 3/10, 세번째 2개는 2/10, 그리고 마지막 1개는 1/10이 된다. 따라서 예상탐색길이는 다음과 같이 4가 된다.[30]

$$(4/10\cdot 3)+(3/10\cdot 4)+(2/10\cdot 5)+(1/10\cdot 6)=4$$

보다 간편한 산출방식은 세번째 수준에서 가능한 10가지 배열에 따른 평균탐색길이를 구하여 이 값을 이전 수준까지의 탐색길이에 더하는 것이다. 즉, 세번째 수준의 평균탐색길이는 $(4/10\cdot 0)+(3/10\cdot 1)+(2/10\cdot 2)+(1/10\cdot 3)=1$이 되어 결국 예상탐색길이는 3+1로 4가 된다.[31]

이와 같이 발생확률에 의해 산출한 마지막 수준에서의 여러개의 가능한 랜덤탐색의 평균은 결국 이 수준에서 적합문헌이 고르게 분포되어 있는 경우의 탐색과 같게 되기 때문에 예상탐색길이를 측정하는 하나의 공식이 유도될 수 있다.

$$ESL(q)=j+\frac{i\cdot s}{r+1}$$

$ESL(q)$: 특정질문(q)에 대한 예상탐색길이

j : 마지막 수준 이전의 모든 수준에서의 부적합문헌의 수 즉, 탐색길이

r : 마지막 수준에서의 적합문헌의 수

i : 마지막 수준에서의 부적합문헌의 수

30) ibid.

31) 정영미, op. cit., p. 320.

s : 이용자의 요구를 만족시켜 주기 위하여 마지막 수준에서 검색해야 할 적합문헌의 수

이 예상탐색길이(ESL)는 문헌집단과 질문집단이 고정되는 경우에만 충족된다. 이러한 경우 전체척도는 예상탐색길이의 평균을 구하는 것으로 그 공식은 다음과 같다.

$$\overline{ESL}=\frac{1}{|Q|}\sum_{q\in Q} ESL(q)$$

$\overline{ESL}$: 평균예상탐색길이
Q : 질문집단

그러나 문헌집단과 질문집단은 가변적이기 때문에 ESL을 어떤 방법으로든지 표준화시켜 줄 필요가 있다. 쿠퍼는 전체문헌집단이 한 수준에서 랜덤하게 검색될 때의 예상탐색길이인 예상랜덤탐색길이(ERSL: expected random search length)를 사용하여 표준화시키고 이 척도를 '예상탐색길이 감소인자'라고 부르고 다음과 같은 공식으로 나타내었다.

$$ESL(q)\text{감소인자}=\frac{ERSL(q)-ESL(q)}{ERSL(q)}$$

이 공식에서 질문 q에 대한 예상랜덤탐색길이 $ERSL(q)$의 산출공식은 다음과 같다.

$$ERSL(q)=\frac{S\cdot I}{R+1}$$

R : 문헌집단내 적합문헌 총수
I : 문헌집단내 부적합문헌 총수
S : 이용자가 원하는 적합문헌의 수

그런데 앞의 $ESL(q)$감소인자의 산출공식에서 $ESL(q)$, $ERSL(q)$, $ERL(q)$감소인자는 각각 한개의 질문에 대한 값이므로 한 검색시스템의 성능을 나타내기 위해서는 N개의 질문에 대한 평균값을 구해야 한다. 따라서 시스템의 전체척도는 평균예상탐색길이($\overline{ESL}$)와 평균예상랜덤탐색길이($\overline{ERSL}$)로부터 다음과 같은 공식으로 산출한다.

$$\overline{ESL}\text{ 감소인자}=\frac{\overline{ERSL}-\overline{ESL}}{\overline{ERSL}}=1-\frac{\sum_{n=1}^{N}\left(j_n+\frac{i_n\cdot S_n}{r_n+1}\right)}{\sum_{n=1}^{N}\frac{S_n\cdot I_n}{R_n+1}}$$

쿠퍼의 모델은 앞에서 말한 바와 같이 질문과 문헌간의 적합성 수준 즉, 유사도의 순위대로 검색문헌을 출력함으로써 이용자가 부적합문헌을 훑어보아야 하는 노력을 감소시켜 주는 시스템의 능력을 측정하는 것이다. 즉, 예상탐색길이 감소인자는 결국 부적합문헌의 검색을 줄이는 시스템의 능력을 나타내므로 정확률을 대신하는 좋은 대체척도는 될 수 있으나 시스템 내의 전체 적합문헌에 대한 검색비율을 나타내는 재현율은 완전히 무시되므로 쿠퍼의 효율척도는 논란의 소지가 있다고 본다.

5.3. SMART 척도

1966년 로치오[32]는 재현율과 정확률에 기초한 두개의 검색효율 측정척도를 제시하였다. 하나는 표준재현율(normalized recall)이며 다른 하나는 표준정확률(normalized precision)로서, 이 두 척도는 쿠퍼의 예상탐색길이척도처럼 검색된 문헌에 순위가 주어지는 시스템의 성능평가에 적합하다.

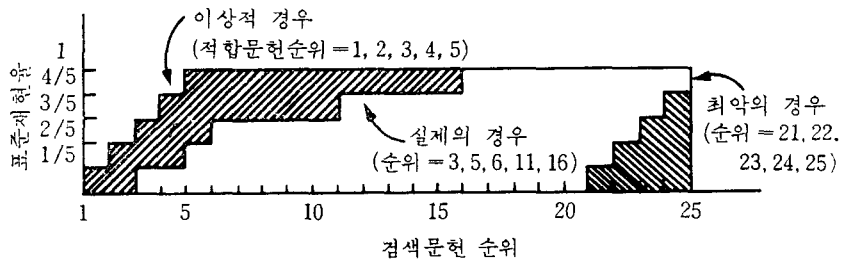
재현율과 정확률은 검색되지 않은 문헌들로부터 검색된 문헌들을 구별하기 위해 선정된 기준점(cutoff point)에 의해 영향을 받는다. 그러므로 사용된 기준치에 따라 여러개의 상이한 재현율—정확률 값이 가능하다. 그러나 검색문헌에 순위가 주어지는 시스템에서는 문헌이 질문과의 유사도 계수 순으로 출력될 때 적합문헌의 순위에 기초한 재현율과 정확률을 한정함으로써 특정한 기준치를 선택할 필요성이 배제될 수 있다. 따라서 표준재현율과 표준정확률은 특정한 검색기준치와는 무관하다.[33]

표준재현율은 실제시스템과 이상적 시스템의 재현율의 차이를 측정하는 것이다. 측정기법은 도표 10과 같이 두 시스템의 재현율을 나타내는 두 계

32) Rocchio, J.J., Document retrieval systems optimization and evaluation. Doctoral thesis. Harvard Univ., 1966.

33) Salton, G., Automatic information Organization and Retrieval. N.Y., McGraw-Hill, 1968, p. 284.

〈도표 10〉 표준재현율의 측정



단함수의 면적의 차이를 측정하는 것이다. 도표의 x축은 검색문헌의 순위를, y축은 재현율을 나타낸다.

n개의 적합문헌을 갖는 실제시스템의 재현율은 첫번째 적합문헌이 검색될 때까지는 0의 값을 가지며 첫번째 적합문헌이 검색될 때 $1/n$로 상승한다. 재현율 $1/n$은 다음 적합문헌이 검색될 때까지 같은 값을 유지하며 그 다음 적합문헌이 검색되면 $2/n$의 값을 갖게 되는 방식으로 마지막 적합문헌이 검색되어 n/n 즉, 1이 될 때까지 계속된다. 이상적 시스템이란 적합문헌이 모두 먼저 검색되는 시스템으로 그 순위는 1, 2, 3, 4, 5가 된다.

이상적 시스템과 실제시스템의 재현율 곡선을 나타내는 두 계단함수의 면적의 차이(빗금친 부분)가 재현율의 측정척도가 되는 것으로 산출공식은 다음과 같다.[34]

$$A_b - A_a = \frac{\sum_{i=1}^{n} ri - \sum_{i=1}^{n} i}{n}$$

A_b : 이상적 시스템의 면적
A_a : 실제 시스템의 면적
n : 적합문헌의 수
ri : 적합문헌의 검색순위
i : 이상적 시스템에서 n개의 적합문헌의 순위

34) ibid., p. 285.

따라서 도표 10의 두 시스템의 면적의 차이는 $\frac{41-15}{5}=\frac{26}{5}=5.2$가 된다.

검색가능한 적합문헌수가 N개인 경우 위의 공식에 의한 값은 최상의 검색일 경우 0이 되며, 최악의 검색일 경우 $N-n$이 된다. 도표 10에서 최악의 경우는 적합문헌의 검색순위가 21($N-n+1$), 22($N-n+2$), 23, 24, 25가 된다.

$$\text{최상의 검색}=\frac{15-15}{5}=0$$

$$\text{최악의 검색}=\frac{115-15}{5}=20 \quad (N-n=25-5)$$

표준재현율은 앞의 공식을 $N-n$으로 나누어 표준화시킨 다음 1에서 빼줌으로써 최상의 경우에는 1의 값, 최악의 경우에는 0의 값이 되도록 한 것이다.

$$\text{표준재현율}=1-\frac{\sum_{i=1}^{n} ri-\sum_{i=1}^{n} i}{n(N-n)}$$

따라서 도표 10의 표준재현율은 0.74(74%)가 된다.

$$1-\frac{41-15}{5(25-5)}=0.74$$

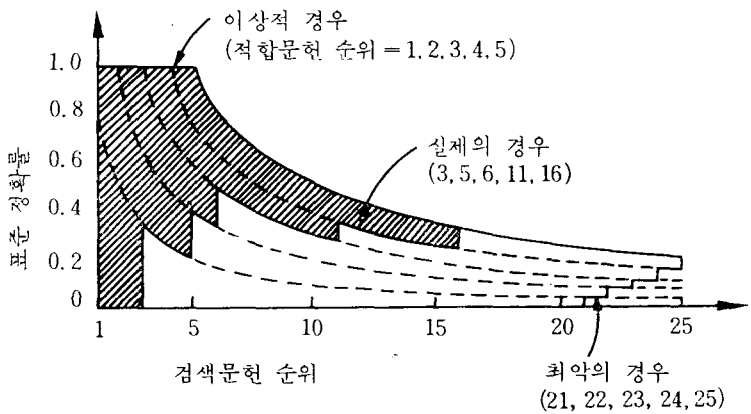
표준정확률도 실제시스템과 이상적 시스템의 정확률의 차이를 측정하는 것으로 이의 산출방법은 표준재현율의 것과 유사하다. 도표 11과 같이 두 시스템의 곡선에 의한 면적의 차이를 측정하는 것으로 각 산출공식은 다음과 같다.[35]

$$A_b-A_a=\sum_{i=1}^{n} \log ri-\sum_{i=1}^{n} \log i$$

35) ibid., pp. 287~289.

$$\text{표준정확률}=1-\frac{\sum_{i=1}^{n}\log ri-\sum_{i=1}^{n}\log i}{\log\left(\frac{N!}{(N-n)!n!}\right)}$$

〈도표 11〉 표준정확률의 측정



표준정확률도 표준재현율과 마찬가지로 최상의 경우에는 1의 값, 최악의 경우에는 0의 값이 되도록 한 것이다.

표준재현율과 표준정확률의 두 값을 사용하지 않고 단일가로 효율측정을 원하는 경우에는 두 값의 복합척도인 표준전체척도(normalized overall measure)를 사용하면 된다. 이것은 가중치를 주어 두 표준값을 더해준 것인데 가중치 5를 준 이유는 두 요소의 값의 크기를 비슷하게 하기 위한 것이다.[36]

표준전체척도=1−5(표준재현율)+표준정확률

5.4. 정보이론적 척도

섀논(Shannon)의 정보이론은 검색시스템의 성능평가에도 응용되고 있는

36) ibid., p. 303.

데[37] 코켈[38], 구아조[39], 미탐[40] 등은 정보이론의 엔트로피(entropy)개념을 이용한 검색효율의 척도를 제시하였다.

코켈은 불완전한 정보검색시스템을 잡음있는 채널을 통한 문헌전송시스템으로 해석하고 전송효율인자(transmission efficiency factor)라고 불리우는 함수 Ht를 검색효율 측정척도로 제시하였는데 그 산출 공식은 다음과 같다.[41]

$$Ht=H(x)+H(y)-H(x,y)\cdots\cdots\text{공식 (1)}$$

Ht : 전송효율인자

$H(x)$: 시스템이 전송한 정보량

$H(y)$: 시스템이 입수한 정보량

$H(x,y)$: 잡음에 의해 발생한 불확실성

여기서 $H(x), H(y), H(x,y)$는 섀논의 엔트로피 공식에 의해 산출된다.

$$H=-\sum_{i=1}^{n} pi \log_2 pi\cdots\cdots\text{공식 (2)}$$

〈도표 12〉 시스템의 조건

a : 검색된 적합문헌
b : 검색된 부적합문헌
c : 검색되지 않은 적합문헌
d : 검색되지 않은 부적합문헌
0100100100←전송
0101000100←입수
dadbcddadd

(a)

입수 \ 전송	1 적합문헌 $(a+c)$ 3	0 부적합문헌 $(b+d)$ 7
1 검색문헌 $(a+b)$ 3	a_2	b_1
0 비검색문헌 $(c+d)$ 7	c_1	d_6

(b)

37) 정영미, Shannon 정보이론의 도서관학·정보학적 해석에 관한 연구, 연세논총, 20(1983), pp. 83~102.

38) Cawkell, A.E., A measure of 'efficiency factor'-Communication theory applied to document selection systems. *Inf. Processing & Management*, 11(1975) : pp. 243~248.

39) Guazzo, M., Retrieval performance and information theory. *Inf. Processing & Management*, 13(1977) : pp. 155~165.

40) Meetham, A.R., Communication theory and the evaluation of information retrieval system. *Inf. Storage & Retrieval*, 5(1969) : pp. 129~134.

41) Shannon, C.E., A mathematical theory of communication. *Bell Syst. Tech. J.*, 27 (1948) : pp. 379~656 (ref. to p. 409).

예컨대 도표 12와 같이 10개의 문헌이 있는 경우 각 문헌은 재현율, 정확률 산출에서와 같이 4개의 범주로 나누어지며, 확률 pi는 각 범주에 속할 확률로서 다음과 같다.[42]

Prs : 검색된 적합문헌일 확률
$P\bar{r}s$: 검색된 부적합문헌일 확률
$Pr\bar{s}$: 검색되지 않은 적합문헌일 확률
$P\bar{r}\bar{s}$: 검색되지 않은 부적합문헌일 확률

$H(x)$는 이용자 판정에 의해 측정되는 정보량이므로 문헌은 적합문헌과 부적합문헌의 두 범주로 구분되며 pi는 적합문헌일 확률 $Pr(=Prs+Pr\bar{s})$과 부적합문헌일 확률 $P\bar{r}(=P\bar{r}s+P\bar{r}\bar{s})$의 두 가지를 갖는다. 반면 $H(y)$는 시스템판정에 의해 측정되는 정보량이므로 문헌은 검색된 문헌과 검색되지 않은 문헌의 두 범주로 구분되며 pi는 검색문헌일 확률 $Ps(=Prs+P\bar{r}s)$와 검색되지 않은 문헌일 확률 $P\bar{s}(=Pr\bar{s}+P\bar{r}\bar{s})$의 두 가지를 갖게 된다. 따라서 공식 (1)의 Ht는 다음과 같이 풀어 쓸 수 있다.

$$Ht=(Pr\log Pr+P\bar{r}\log P\bar{r})+(Ps\log Ps+P\bar{s}\log P\bar{s})-(Prs\log Prs+Pr\bar{s}\log Pr\bar{s}+P\bar{r}s\log P\bar{r}s+P\bar{r}\bar{s}\log P\bar{r}\bar{s})\cdots\cdots\text{공식 (3)}$$

그러므로 도표 12의 경우 Ht의 값은 공식 (3)을 공식 (2)에 대입시켜 다음과 같이 산출된다.

$$\begin{aligned}Ht&=-\{(0.3\log_2 0.3+0.7\log_2 0.7)+(0.3\log_2 0.3+0.7\log_2 0.7)\\&\quad-(0.2\log_2 0.2+0.1\log_2 0.1+0.1\log_2 0.1+0.6\log_2 0.6)\}\\&=0.1916\end{aligned}$$

코켈의 효율척도 Ht를 특정시스템의 재현율—정확률 척도와 비교해 보자. 도표 13과 같은 특정시스템의 3가지 다른 검색결과로부터 재현율—정확률 및 Ht의 값을 산출해 보면 세번째 경우가 잡음이 전혀 없는 완전한

42) 정영미, op. cit., p. 325.

〈도표 13〉 세 경우의 검색결과

경 우	a	b	c	d	재현율	정확률	$H(x)$	Ht
1	10	90	0	0	100%	10%	0.469	0
2	5	5	5	85	50	50	0.469	0.0904
3	10	0	0	90	100	100	0.469	0.469

(시스템의 소장문헌 : 10개의 적합문헌과 90개의 부적합문헌)

검색에 해당되며 이 때의 전송효율인자 Ht의 값은 시스템이 전송한 정보량 $H(x)$ 즉, 이용자 판정에 의해 측정되는 정보량의 값과 같아진다.

먼저 $H(x)$의 값은 1, 2, 3의 경우 다음과 같이 모두 같다.

$$H(x)=-(0.1\log_2 0.1+0.9\log_2 0.9)=0.469$$

Ht의 값은 3의 경우 다음과 같이 0.469가 된다.

$$\begin{aligned}Ht=&-\{(0.1\log_2 0.1+0.9\log_2 0.9)+(0.1\log_2 0.1+0.9\log_2 0.9)\\&-(0.1\log_2 0.1+0+0+0.9\log_2 0.9)\}\\=&0.469\end{aligned}$$

미탐과 구아조의 척도도 개념적으로는 코켈과 유사하며 조건엔트로피를 사용하였다는 점에서 차이가 있을 뿐이다.

6. 결 론

검색효율은 정보검색시스템 평가의 제일기준으로서 많은 연구와 실험의 대상이 되어 왔다. 크랜필드 평가실험 이후 검색효율의 측정을 위해 제시된 다양한 척도들을 검토 분석한 본 연구의 요점은 다음과 같다.

1) 검색효율 측정에서 가장 중요한 요소가 되는 정보의 적합성 판단은 시스템위주의 판정과 이용자위주의 판정으로 구분되어 평가되어야 할 것이다.

2) 재현율과 정확률은 이용자지향적인 평가척도이며, 재현율과 부적합률은 시스템지향적인 평가척도라고 본다.

3) 재현율과 정확률, 또는 재현율과 부적합률을 결합시킨 복합척도들은 대부분이 논리적 합리성을 찾기가 어렵다.

4) 스웨츠의 E척도는 다양한 값의 성능곡선대신에 OC곡선을 직선화하여 E값을 생산할 수 있는 기법을 논증하였으며, 또한 잘 알려진 통계적 결정이론에 입각한 확률이론으로부터 유도한 척도를 제시함으로써 근본적으로 정보검색의 확률적 특성을 강조하였다고 본다.

5) 쿠퍼의 예상탐색길이는 이용자가 원하는 적합문헌의 수를 고려한 척도로서 적합성의 수준에 따라 검색문헌의 순위가 주어지는 시스템의 효율측정에 적합하다고 본다.

6) 로치오의 표준재현율과 표준정확률은 특정한 검색기준치와는 무관한 척도이며 쿠퍼의 척도처럼 검색문헌에 순위가 주어지는 시스템의 효율측정에 적합하다고 본다.

7) 코켈의 척도는 섀논의 엔트로피 개념이 검색효율의 측정기법으로 응용될 수 있음을 논증하였다.

(제출일자 '89.6.12)