이용자 피이드백에 의한 검색질문의 자동 수정에 관한 연구

An Experiment on Automatic Query Modification in Information Retrieval Using the Relevance Feedback

신 영 실 *

초 록

이용자와 시스템간에 상호작용이 이루어질수 있는 온라인 정보검색 시스템에서는 검색결과에 대한 이용자의 피이드백을 이용하여 검색질문을 수정함으로써 시스템의 성능을 향상시킬수 있다.

본 논물에서는 샐튼과 우가 제시한 검색질문의 자동수정 모형을 통제된 키이워어드 시스템에 적용시켜 보았다.

ABSTRACT

When an information retrieval system is implemented on-line, users can interact with the system to improve the searches. There are studies which achieved dramatic improvements in system effectiveness by using automatic relevance feedback, a technique for reformulating a patron query based on initial retrieval result.

In this thesis, an automatic query modification model was applied to a controlled keyword system.

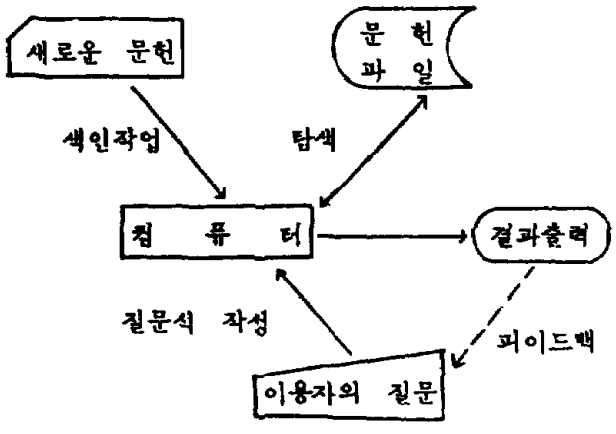
I. 서 론

정보검색 시스템이란 정보의 수집, 가공, 축적, 검색, 배포의 과정을 망라한 체계를 말하는 것으로서, 문헌의 분석과 색인(분류 포함), 검색질문의 작성, 탐색과 검색이라는 이 세가지 작업에서 지적인 노력이 요구된다. 이들은 상호 연관성이 높으므로 이 세가지 작업이 잘 통합되어야 우수한 시스템이 될수 있다는 인식이 점차 높아가고 있다. 정보검색 시스템에서의 이러한 상호 연관성을 도식으로 표시하면 그림 1 과

* 접수일자 : 1985. 5.7.

같다.[1]

그림 1 정보검색 시스템



따라서 이러한 3가지 작업이 시스템의 성능을 지배하는 주요 요인이 된다. 그러나 색인이나 분류작업은 시스템이 고안될때 이미 결정되어진 것이라고 볼 때, 이용자의 정보요구를 질문으로 작성하는 과정은 주어진 시스템의 환경하에서 시스템의 성능을 향상시킬 수 있는 보다 직접적인 요인이라 할 수 있다.

최근에는 정보의 급격한 증가추세에 따른 막대한 양의 정보와 학문의 세분화 및 상호관련화에 의하여 이용자의 정보요구 또한 보다 복합적이고 특정한 것이 되어가고 있다. 따라서 한번의 질문작성으로는 이용자의 정보요구를 정확히 표현하기가 어려우며, 또한 다양한 정보요구를 어느 한 방법에 따라 획일적으로 처리한다면 이용자의 정보요구를 충족시키기가 어려울 것이다.

또한 1970년대 초반부터 컴퓨터공학의 획기적인 발달인 컴퓨터의 시분할(time-sharing) 기술이 정보검색 시스템에 본격적으로 이용되어 종전의 오프라인(off-line)시스템이 온라인(on-line)시스템으로 바뀌면서 이용자는 시스템과 계속적인 상호작용을 할 수 있게 되었다. 따라서 시스템의 피이드백을 이용해서 각 이용자의 요구에 맞게 질문을 수정하면 검색효율을 향상시킬 수 있을 것이라고 기대된다.

오늘날 대부분의 시스템에서는 검색결과가 만족스럽지 못한 경우 시스템에서 디스플레이(display)해 준 정보를 이용해서 이용자 자신이 질문을 수정하게 되어 있으나 이는 이용자에게 지나친 부담이 되며, 또한 이용자가 해당검색 시스템에 대해서 충분히 알고 있지 못할 경우에는 검색실패를 초래하게 된다. 따라서 질문의 수정과정 중에서 이용자가 안고 있는 부담을 줄이고 시스템측에서 보다 많은 업무를 처리할 수 있는 탐색전략의 필요성이 대두되었다.

랑카스터(Lancaster)는 1968년 발표된 논문에서 적합문헌에 있는 색인어를 근거로 하여 컴퓨터에 의해 탐색전략을 자동으로 유도해 낼 수 있는 가능성에 대한 연구가 필요하다고 하였다.[2] 또한 스마트(SMART)시스템에서는[3] 로치오(Rocchio)가 제안한 적합성 피이드백(relevance feedback)과정을 적용시켜 질문을 자동수정하는 방법에 대해 연구·실험하였다. 이후 적합성 피이드백 과정을 이용한 검색질

1) Harry Wu, "On Query Formulation in Information Retrieval" (Ph.D. dissertation, Cornell University, 1981), pp. 3-4.

2) F.W. Lancaster, *Evaluation of the MEDLARS Demand Search Service*, Final Report, National Library of Medicine (Washington, D.C., Jan. 1968).

3) SMART (Salton's Magical Automatic Retrieval of Text)
시스템은 검색시스템에서 주제분석·색인작성·검색부분을 기계적 처리 즉 자동화 연구개발을 하기 위한 시스템으로서, 1962～65년에 하버드 대학에서 Gerard Salton이 개발한 후 코넬대학으로 옮겨 실험하고 있다.

문의 자동수정에 관한 연구들이 실제 시스템과[4] 실험적인 시스템을 대상으로 꾸준히 진행되고 있다.

본 논문에서는 자동검색 시스템에서 적합성 피이드백을 이용한 검색질문의 자동수정 모형을 디소오러스(thesaurus)에 의한 키이워어드(keyword)시스템에 적용시켜서 검색효율이 향상되는지를 알아 보고자 하였으며, 또한 앞으로 완전자동검색 시스템의 시대가 도래할 것에 대비하여 기존 시스템의 탐색전략개발에 대한 연구에 도움이 되고자 한다.

Ⅱ. 정보검색에서의 이용자·시스템 상호작용

1. 이용자·시스템 상호작용의 개념

지난 십여년간 정보검색 시스템의 탐색방법상에는 많은 변화가 있었다. 정보검색 시스템의 탐색과정이 근본적으로 이용자의 질의와 시스템내에서 탐색에 관여하는 모든 요소를 연결해 주는 커뮤니케이션 작업이라 할 때, 종전의 정보검색 시스템은 오프라인 시스템으로서 이용자와 검색 시스템간에 상호작용이 없어 스스로 발견하는 식의 탐색이 곤란하며, 응답시간이 상대적으로 오래 걸리고, 위탁탐색만이 가능하였다. 따라서 검색결과의 성패는 운명에 맡기게 되는(hit or miss)경우가 많았다. 그러나 오프라인 검색 시스템에서도 현황추적봉사(SDI Service)의 경우는 시스템과 이용자간에 상호작용이 이루어지고 있다. 즉, 탐색결과에 어떤 부적합한 것이 있으면 만족할 만한 검색결과가 출력될 때까지 이용자와의 접촉을 통해서 이용자의 관심 프로파일을 몇번이든 수정·보완할 수 있다.

검색 시스템의 실패원인은 시스템에 따라서 차이가 있으나, 일반적으로 색인작성, 색인언어 질문작성, 이용자와 시스템간의 상호작용을 들 수 있는데[5], 오프라인 시스템에서 그 서어비스의 주축을 이루는 소급탐색의 경우에는 이 중에서도 특히 이용자와 시스템간의 상호작용의 부족이 실패의 중대한 원인이 되기도 했다.

1970년대 초반부터 컴퓨터공학에서 획기적인 발달이라 할 수 있는 컴퓨터의 시분할 기술이 정보검색에 이용되어, 온라인 정보검색 시스템이 개발되었다. 온라인 정보검색 시스템에서는 이용자가 컴퓨터에 질의를 입력하여 탐색이 이루어지는 동안, 콘솔(console)이나 입출력 터미날을 통해서 이용자와 시스템간에 계속적인 상호작용이 유지될수 있으므로 보다 이상적이라 하겠다.

2. 검색질문의 작성

이용자는 정보가 필요하게 되면, 자연언어를 사용하여 자신의 정보요구를 질문이란 형식으로 공식화하게 되는데, 이 과정은 질문자 개개인의 특성 즉, 문헌장서의 내용에 대한 지식, 시스템의 색인·탐색과정에 대한 이해도, 탐색하려는 주제에 대한 지식, 언어표현방법 등의 영향을 받는다. 이렇게하여 이용자의 정보요구가 명확하게 파악되면 탐색자는 그 질문내용의 주제를 분석하여 중요개념을 추출하고, 그 개념들을 문헌파일을 색인할 때와 같은 형의 색인

4) C. Vernimb, "Automatic Query Adjustment in Document Retrieval," *Inf. Pro. & Man.*, 13, (1977), pp. 339-353.

5) G. Salton, *Dynamic Information and Library Processing* (N.J.: Prentice-Hall, 1975), pp. 125-128.

언어로 변환시킨다. 다음에는 탐색을 실시하기 위해서 검색 시스템에서 처리할 수 있는 질문식으로 작성하게 된다. 질문식을 표현하는 방법과 검색여부를 결정하기 위해 사용하는 질문과 문헌간의 매칭함수(matching function)즉 관련성 측정기준은 문헌을 색인하는 방식에 따라서 달라진다.

2.1 논리관계 질문식(Boolean query)

논리관계 질문식은 질문을 구성하는 탐색용어(search term)들의 관계를 논리적(AND), 논리화(OR), 논리차(NOT)등의 논리연산자로 조합하여 작성한 것으로, 현재 운영되고 있는 대부분의 검색 시스템에서는 논리관계 질문식을 이용하고 있다. 이 질문식에 의한 검색기준에 따르면 질문식에 꼭 부합되는 문헌만을 검색하게 되므로[6] 검색과정에서 장서는 검색될 문헌과 그렇지 않은 문헌의 두 부분으로 나뉘어지는데, 각 부분을 이루는 문헌들 간에는 어떤 서열이 없이 질문과의 유사정도가 모두 같은 것으로 간주된다. 따라서 이용자의 요구에 맞추어 검색되는 문헌의 양을 적당하게 조절하기가 어렵다.

2.2 벡터 질문식(vector query)

검색질문을 하나의 벡터로 나타내는 벡터 질문식은 자동색인을 하는 실험적인 시스템에서 주로 이용되고 있다. 이때 질문식은 문헌을 색인하는데 이용된 n개의 용어(term)로 구성된 어휘집(vocabulary)에서 뽑아낸 용어들로 구성되며 $Q_i = (T_1, T_2, \cdots, T_n)$으로 표시된다. 여기서 Q_i는 i번째 질문을, T_i는 i번째 탐색용어를 나타낸다.

벡터 질문식을 이용하는 시스템의 탐색전략을 보면, 문헌의 색인어와 탐색용어에는 가중치를 주고, 질문과 문헌간의 유사성(similarity)을 나타낼 수 있는 매칭함수를 이용해서 문헌을 검색하게 된다. 이때 매칭함수에 의해서 계산된 유사계수(similarity coefficient)는 해당 질문에 대한 각 문헌의 중요도를 나타내는 것으로서, 그 값이 큰 것부터 작은 것의 순으로 배열해서 일정한 순위까지를 검색하거나, 또는 유사계수에 대한 일정한 기준치(threshold)를 정해 주고 그 일정치 이상의 문헌을 검색하여 이용자에게 제시하여 준다. 따라서 벡터 질문식을 이용하는 시스템에서는 이용자의 정보요구에 부합되는 순서에 따라 문헌을 차례대로 검색해 줄 수 있다.

한편 탐색용어에 가중치를 주는 방법은 크게 두 가지로 나눌 수 있다. 하나는 장서내에서 그 용어가 갖는 특성 즉 빈도수를 이용해서 가중치를 주는 방법(statistical weighting scheme)이며, 또 하나는 문헌에서 그 용어가 지니는 중요도와는 상관없이, 이용자가 어느 탐색용어가 들어 있는 문헌을 더 중요시 하는 가에 따라서 그 질문을 구성하는 탐색용어에 상대적인 가중치를 주는 방법이다.

3. 검색질문의 수정

3.1 수정의 필요성

검색질문의 작성에서도 이미 언급했듯이, 이용자의 정보요구를 질문으로 공식화하는 과정에는 여러 요인들이 영향을 미치고 있으므로 이용자가 자신의 정보요구사항을 처음부터 정확하게 기술하여 시스템에 제출하는 경우는 매우

6) 탐색용어의 빈도와 조합방법에 따라서 검색되는 문헌의 수가 달라진다. 빈도가 높은 2개의 용어를 논리화 (OR) 로 연결하면 검색되는문헌의 수가 많아지지만, 빈도가 낮은 용어들을 논리적 (AND) 으로 연결하게 되면 극소수의 문헌만이 검색된다.

드물다고 하겠다. 콜리슨(Collison)은 이용자가 무엇을 요구하고 있는지를 정확히 파악하면 그 탐색의 반은 이미 성공한 것이라고 지적하였다.[7] 따라서 첫번째 작성된 질문, 즉 원질문(initial query)이 만족할 만한 검색 결과를 산출해 낼 것이라고 기대하는 것은 거의 불가능하다. 그러므로 이용자가 시스템과 계속적인 접촉을 유지하면서, 시스템에 제시한 관련사항들을 보고 그 결과를 이용하여 원질문을 수정해 나가는 것은 주어진 시스템의 여건하에서 검색 효율을 향상시킬 수 있는 보다 직접적인 방법이라고 할 수 있다.

3.2 수정방법

질문을 수정하는 방법은 이용자와 시스템간에 상호작용이 이루어지는 시기에 따라서 크게 다음의 두 가지 유형으로 나눌 수 있다.[8]

3.2.1 탐색전 상호작용(presearch interaction)에 의한 수정

실제로 탐색을 하기 전에 시스템이 디스플레이해 준 정보를 이용해서 질문자가 직접 질문을 수정하는 것으로서, 다음과 같은 방법들이 사용되고 있다.

첫째, 시스템에서는 질문을 구성하고 있는 탐색용어의 문헌빈도를 디스플레이해 줄 수 있으며, 질문자는 빈도가 너무 크거나 작아서 검색에 별로 유용하지 못할 것으로 여겨지는 용어는 빼거나, 다른 용어로 바꾸어 줄 수 있게 된다.

둘째, 해당되는 주제영역의 디소오러스를 디스플레이 시켜서, 원질문의 탐색용어에 대한 동의어와 관련어들을 알 수 있게 한다.

셋째로는, 검색을 하기 전에 이용자가 이미 적합문헌이라고 알고 있는 문헌을 시스템에 입력시키면 시스템에서는 그 문헌의 색인어를 디스플레이 시킬 수 있다.

3.2.2 탐색후 상호작용(postsearch interaction)에 의한 수정

원질문으로 우선 탐색을 실시하여 그 결과의 일부를 이용자에게 디스플레이하여 줌으로써 이용자는 질문을 수정하게 되는데, 시스템에서는 검색된 문헌의 표제와 초록을 디스플레이 해줄수 있다.

이러한 수정방법에서는, 일반적으로 적합문헌에 나타나는 용어를 질문에 추가시키고 비적합문헌에 있는 용어들은 빼버리는 방법으로 질문을 수정하게 된다.

이때 원질문에 의해서 검색된 문헌 중 어느 문헌이 적합문헌인지를 이용자가 판정하여 시스템에 알려주면, 시스템에서는 이 정보를 이용해서 질문을 자동으로 수정해 주는 방법이 있는데 스마트시스템에서는 이러한 수정과정을 적합성 피이드백(relevance feedback)이라고 명명하였다.

4. 적합성 피이드백

4.1 개념

피이드백이란 폐쇄 시스템에서 입력처리되어 출력된 결과가 자체 통제를 위한 자료로서 재입력되는 것을 말한다. 즉 한가지 업무를 일단 처리한 후 그 결과를 참고함으로써 시스템의 성능을 향상시킬 수 있는 장치를 뜻하는 것이다. 이러한 피이드백의 개념은 생태적 자동제어 시스

7) R.L. Collison, *Library Assistance to Readers*, 5th ed. (London: Crosby Lockwood, 1965), p. 62.

8) M.E. Lesk and G. Salton, "Interactive Search and Retrieval Methods Using Automatic Displays," in *The SMART Retrieval System*, ed. Gerard Salton (N.J.: Prentice-Hall, 1971), pp. 487-505.

템(biological and automatic control system)에서 확립되었으며, 이것이 널리 보급되기는 위너(Wiener)의 저서인 「Cybernetics」에서 였다.[9] 피이드백의 개념은 정보검색 시스템에서도 검색방침을 조정하는데 상당히 효과적으로 이용되고 있다.

적합성 피이드백이란 로치오가 제의한 상호작용 과정으로서[10], 하나의 질문을 탐색한 결과 검색된 문헌에 대하여 이용자는 각 문헌의 적합성만을 판정하고 시스템에서는 적합성 평가의 결과를 참고해서 탐색질문을 자동수정하는 방법이다. 이러한 적합성 피이드백은 이용자가 제기하는 원질문은 불완전한 것이기는 하지만 최소한 몇개의 적합문헌은 검색해 낼 수 있을 것이라는 개념에 기초한 것이다. 따라서 이것을 수정하여 최적질문(optimal query)을 만들므로써 검색효율을 높일 수 있다는 것이다. 이때 최적질문이란 적합문헌들과의 상관관계와 비적합문헌들과의 상관관계의 차이가 최대가 되어서, 문헌들을 질문과의 상관도에 따라서 순위를 매겼을때 적합문헌들이 비적합문헌들보다 상위(上位)에 오르게 하는 질문을 뜻한다. 적합문헌의 집합을 R이라하고 장서의 나머지 부분 즉 비적합문헌의 집합을 S(S=N-R) 라고 하면 최적질문이란 다음과 같은 함수 F값이 최대가 되는 질문이라고 정의할 수 있다.

$$F=\frac{1}{|R|}\sum_{d^{(i)}\in R} M(q,d^{(i)}) - \frac{1}{|S|}\sum_{d^{(i)}\in S} M(q,d^{(i)})$$

위 식에서 |R|은 적합문헌의 수, |S|는 비적합문헌의 수, $M(q,d^{(i)})$는 질문과 문헌간의 매칭함수를 나타낸다.

그러나 실제에서는 탐색이 끝나기 전에 R과 S를 알 수가 없으므로 이 공식을 이용하여 직접 최적질문을 만드는 대신에 R의 한 부분을 알아낼 수 있는 원질문을 이용해서 최적질문에 대한 근사치를 유도해 내는 것이 필요하다. 실제로 원질문인 Q_0는 적합문헌의 일부인 R_0와 비적합문헌의 일부인 S_0를 검색해 내는데 이용되며, R_0와 S_0에서 뽑아낸 색인어를 이용해서 Q_1을 만들어 탐색한 결과 R_1과 S_1을 검색하게 된다. 이렇듯 새로운 적합문헌들이 검색됨에 따라서 검색된 적합문헌의 집합은 R에, 수정된 질문은 최적질문에 수렴하게 된다.

이러한 수정과정은 반복에 의한 것으로서 Q_i, R_i, S_i를 변수로 가지는 함수로 나타낼 수 있다. 즉

$$Q_{i+1}=f(Q_i, R_i, S_i)$$

4.2 질문수정과정

적합성 피이드백을 이용하여 질문을 수정하는 과정은 일반적으로 다음과 같다.

1. 이용자가 시스템에 질문을 입력한다.

2. 시스템에서는 이용자의 요구에 적합할 것으로 여겨지는 문헌을 출력한다.

3. 그러면 이용자는 자신의 정보요구와 관련시켜서 검색결과를 살펴본 후 만일 만족할 만한 결과가 못될 경우에는, 검색된 문헌들을 적합문헌과 비적합문헌으로 나누어 그 정보를 시스템에 입력시킨다.

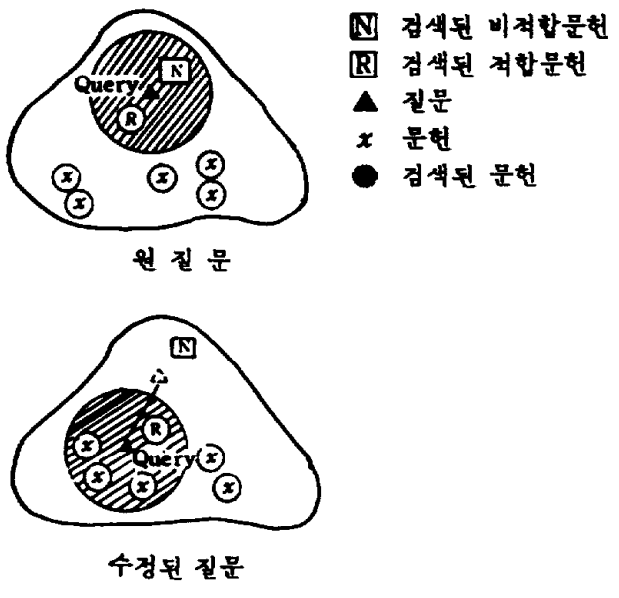
4. 시스템에서는 검색된 문헌에 대한 적합성 평가 결과와 원질문을 이용해서, 그림 2와 같이 비적합문헌과는 최대의 거리를 유지하면서 적합문헌들이 핵심적인 위치에 놓이게 되는 질

9) C.J. van Rijsbergen, *Information Retrieval*, 2nd ed. (London: Butter-worths, 1979), p. 106.

10) J.J. Rochio, Jr., "Relevance Feedback in Information Retrieval," in *The SMART Retrieval System*, ed. G. Salton (N.J.: Prentice-Hall, 1971), pp. 313-323.

문으로 만든다. 실제로는 적합문헌에 있는 색인어는 추가시키고 비적합문헌에 있는 색인어는 빼는 방법을 통해서 새로운 질문이 만들어진다.

그림 2 적합성 피이드백



즉 이때에 이용되는 기본적인 알고리듬(Algorithm)은

$$Q_{i+1} = \alpha Q_i + \beta \sum_{i=1}^{n_r} R_i - \gamma \sum_{i=1}^{n_s} S_i$$

로 표시할 수 있다. 이 식에서 파라미터 α, β, γ는 적절한 상수이며, n_r은 적합문헌의 수, n_s는 비적합문헌의 수, R_i는 적합문헌의 색인어, S_i는 비적합문헌의 색인어를 의미한다.

5. 위와 같은 과정을 거쳐서 수정된 질문을 시스템에 재입력해서 두번째 탐색을 실시한다. 그 탐색결과에 대해서 다시 적합성 평가를 한 후 만일 필요하다면 질문을 재수정하게 된다.

이러한 질문의 수정과정은 이용자가 탐색결과에 만족할 때까지 반복해서 실시할 수 있다.

적합성 피이드백을 이용한 질문수정과정은 시분할컴퓨터(time-sharing computer)와 입출력터미날을 이용하는 온라인 검색 시스템에 적합하고, 대부분의 업무는 컴퓨터에 의해 처리되기 때문에 이용자측에 별로 큰 부담이 되지 않으므로 광범위하게 연구되고 있다.

4.3 선행연구

적합성 피이드백을 이용해서 질문을 수정하기 위한 알고리듬과 기법에 대해서는 다음과 같은 연구들이 있었다.

로치오는[11] 전자계산학 분야의 「IRE Transactions on Electronic Computers」의 1958년 3월부터 9월호에 실린 405편의 초록으로 장서를 구성하고 17개의 질문을 작성하여 적합성 피이드백에 의해서 질문을 수정한 결과 검색효율이 향상되는지를 실험하였다. 이 실험에서 이용된 수정공식은 다음과 같다.

$$Q_{i+1} = Q_i + \frac{1}{n_r} \sum_{1}^{n_r} \frac{r_i}{|r_i|} - \frac{1}{n_s} \sum_{1}^{n_s} \frac{S_i}{|S_i|}$$

이러한 로치오의 연구는 적합성 피이드백에 대한 연구의 효시이다.

리들(Riddle)등은[12] 도큐멘테이션(documentation)에 관한 회의에 제출된 82개의 문헌으로 구성된 장서(ADI collection)를 대상

11) Ibid. ; *Document Retrieval System Optimization and Evaluation*, Ph.D. thesis, Report No. ISR-10 (Ma.: Harvard University, June 1965) ; J.J. Rocchio and G. Salton, "Search Optimization and Interactive Retrieval Techniques," *Proceedings of the AFIPS Fall Joint Computer Conference* (Las Vegas: Nov. 1965).

12) O.W. Riddle, T. Horwitz, and R. Dietz, *Relevance Feedback in Information Retrieval Systems*, Scientific Report No. ISR-11 to the National Science Foundation, Chapter 6 (N.Y.: Cornell University, June 1966).

으로 해서 22개의 질문을 작성하여 실험하였다. 수정공식으로는

$$Q_{i+1} = Q_i + \alpha \sum_{1}^{n_r} r_i$$

를 이용하였다. 또한 이 연구에서는 첫번째 반복검색에서 새로운 적합문헌이 검색되지 않을 경우에는 2개의 비적합문헌을 이용해서 수정하는 방법에 대해서도 실험하였다. 이 경우에는

$$Q_{i+1} = Q_i + \alpha \sum_{i}^{n_r} r_i - \sum_{1}^{2} S_i$$

를 수정공식으로 이용하였다.

크로포드(Crawford)와 멜저(Melzer)는[13] 첫번째 탐색에서 적합문헌이 검색되면, 그 후의 과정에서는 원질문을 무시하고 적합문헌만을 이용하는 방법에 대해서 연구하였다. 이때의 수정공식으로는 $Q_{i+1} = r_i$가 이용됐다. 그리고 만일 적합문헌이 없을 경우에는 일단 $Q_2 = Q_1 - S_1$을 이용하여 탐색을 하고, 그 다음 단계에서부터는 이 공식을 이용하지 않는 방법에 대해서, 기체역학분야의 200개의 문헌(Cranfield collection)을 대상으로 실험하였다.

켈리(Kelly)는[14] 적합문헌이 검색되지 않아서 새로운 색인어를 추가시킬 수 없을 경우에, 문헌장서내에서 자주 나오는 색인어를 질문에 추가시키는 전략에 대해서 연구하였다.

스타인블러(Steinbuhler)와 알레타(Aleta)는[15] 첫번째 검색결과 비적합문헌만이 검색되는 경우, 비적합문헌을 피이드백으로 이용하는 방법을 ADI 장서를 대상으로 하여 실험하였다. 수정공식으로는 로치오가 사용한 것과 동일한 수정공식을 이용하였다. 그리고 켈리가 제안한 알고리듬을 ADI 장서에서 실험하였다.

아이드(Ide)는[16] 크랜필드 장서(Cranfield collection)에 42개의 질문을 작성하고, 수정공식으로

$$Q_{i+1} = \pi Q_i + \omega Q_0 + \alpha \sum^{\min(n_s, n_r)} r_i + \mu \sum^{\min(n_s, n_s)} S_i$$

를 이용해서 다음과 같은 사항들을 실험하였다. 첫째, 크랜필드 장서와 ADI 장서의 검색결과를 비교하였고 둘째, 수정공식의 파라미터 π, ω, α를 변화시켜서 적합문헌만을 이용한 방법에 대해서 연구하였으며 셋째, 피이드백으로 이용하기 위해서 이용자에게 제시해 주는 문헌수가 결과에 미치는 영향과 넷째, 검색된 모든 문헌을 즉, 적합문헌과 비적합문헌을 함께 이용하는 방법들을 검토하였다.

보로딘(Borodin)등과[17] 베이커(Baker)

---

13) R. Crawford and H. Melzer, *The Use of Relevant Documents Instead of Queries in Relevance Feedback*, Scientific Report No. ISR-14 to the National Science Foundation, Section XIII (N. Y.: Cornell University, Oct. 1968).

14) J. Kelly, *Negative Response Relevance Feedback*, Scientific Report No. ISR-12 to the National Science Foundation, Section IX (N.Y.: Cornell University, June 1967).

15) D. Steinbuhler and C. Aleta, *Negative Relevance Feedback Process*, (student report, Computer Science 435, Cornell University, Spring 1968).

16) E.C. Ide, "Relevance Feedback in an Automatic Document Retrieval System" (Master's thesis, Cornell University, 1969).

17) A. Borodin, L. Kerr, and F. Lewis, *Query Splitting in Relevance Feedback Systems*, Scientific Report No. ISR-14, Section XII (N.Y.: Cornell University, Oct. 1968).

는[18] 해당 주제영역에 대한 이용자의 판점과 문헌색인법이 일치하지 않아서, 적합문헌들이 나뉘어지게 되는 경우에 적용시킬 수 있는 방법에 대해서 연구하였다. 이 연구에서는 이용자의 질문을 몇개의 부분으로 다시 나누는 질문분할방법(query splitting method)을 적용시켰다.

파아볼라(Paavola)는[19] 피이드백으로 이용할 각 문헌에 대해서 이전까지의 다른 이용자들이 관련이 있는 적합문헌이라고 판명한 문헌들을 추가시키는 방법을 제기하였다. 즉 이는 이용자 연구를 이용하는 것이다.

머레이(Murray)와[20] 살다나(Sardana)는[21] 피이드백 과정에 의해서 탐색용어가 계속 추가됨에 따라 탐색용어의 숫자가 지나치게 방대해지는 것을 방지하기 위한 연구를 하였다. 이러한 연구들의 결과를 보면 피이드백 질문이 방대한 경우에는 가중치가 낮은 탐색용어들을 일부 삭제하여도 검색효율상에 많은 손실을 초래하지는 않는다는 것이 밝혀졌다.

샐톤(Salton)과 우(Wu)는 정확성 가중치(precision weights)를 탐색용어에 대한 가중치로 이용하면 최적질문을 만들수 있다는 연구와[22] 적합성 피이드백을 이용해서 질문을 수정하는 로치오의 연구를 기초로 해서 새로운 질문수정 모형을 제시하였다. 그들은 424편의 크랜필드장서와 생물의학분야의 450편의 장서(MEDLARS collection)에 각각 20개와 19개의 질문을 제기하여서 검색효율이 향상되는가를 실험하였다.[23]

본 논문에서는 샐톤과 우의 모형을 자동색인 시스템이 아닌 디소오러스에 의한 키이워어드 시스템에 적용시켜 실험하였다.

Ⅲ. 질문수정 모형에 의한 검색실험

1. 실험목적

샐톤과 우의 질문수정 모형은 다음과 같은조건하에서 적용되는 것으로 본 논문의 실험은이 조건에 따라 설계되었다. 첫째, 색인어들은 장서내의 적합문헌과 비적합문헌들 안에서 독립적으로 분포되어 있다고 가정하며 둘째, 검색질문식으로는 벡터 질문식을 쓰고 셋째, 문헌은 이원

18) T.P. Baker, *Variations on the Query Splitting Technique with Relevance Feedback*, Scientific Report No. ISR-18, Section VIII (N.Y.: Cornell University, Oct. 1970).

19) L. Paavola, *The Use of Past Relevance Decisions in Relevance Feedback*, Scientific Report No. ISR-18, Section XI (N.Y.: Cornell University, Oct.1970).

20) D.M. Murray, *Document Retrieval Based on Clustered Files*, Ph.D. thesis, Scientific Report No. ISR-20 (N.Y.: Cornell University, June 1972).

21) K. Sardana, *On Controlling the Length of the Query Vector Feedback Query Vectors*, Scientific Report No. ISR-22, Section VIII (N.Y.: Cornell University, Nov. 1972).

22) D.H. Kraft and A. Bookstein, "Evaluation of Information Retrieval Systems: A Decision Theory Approach," *JASIS*, 29, (1978), pp. 31-34; S.E. Robertson and K. Sparck Jones," Relevance Weighting for Search Terms," *JASIS*, 27, (1976), pp. 129-146; C.T. Yu, W.S. Luk, and M.K. Sui, "On Models of Information Retrieval Processes," *Information Systems*, 4, 3, (1979) pp. 205-218; C.T. Yu and G. Salton, "Precision Weighting-An Effective Automatic Indexing Method," *JACM*, 23, 1, (1973), pp. 76-88.

23) Harry Wu and Gerard Salton, "The Estimation of Term Relevance Weights Using Relevance Feedback," *J. of Doc.*, 37, 4 (Dec. 1981), pp. 194-214.

색인법(binary indexing)으로 나타내주며 넷째, 문헌과 질문간의 매칭함수로는 해당 벡터간의 내적(內積)을 이용하고 다섯째, 적합성 피이드백을 이용해서 용어의 정확성 가중치(precision weights)를 추정하여 질문을 수정해서 반복검색을 실시한다.

본 실험의 목적은 이러한 탐색전략을 자동색인 시스템이 아닌 디소오러스에 의한 키이워어드 시스템에 적용시켰을 때에 검색효율이 향상되는지를 알아보고자 한다. 이때 디소오러스에 의한 키이워어드 시스템의 키이워어드들은 기능상으로 보아 문헌과 질문을 벡터(vector)로 표시하는 자동시스템의 벡터와 같다고 할 수 있다. 따라서 본 논문에서는 벡터 질문식에 의한 자동수정 모형을 이용하였다.

2. 실험설계

2.1 실험문헌집단

전자계산학의 분야중 컴퓨터 네트워크에 관한 잡지인 「Computer Networks : The International Journal of Distributed Informatique」의 1978년부터 1982년까지 발행된 것을 대상잡지로 선택하였다. 그리고는 국소지역 네트워크(local area network), 성능(performance), 협약(protocol), 모의실험(simulation)의 소주제를 중심으로 이론적인 논문기사 55편을 선정하여 표제와 초록으로 실험문헌집단을 구성하였다. 이러한 과정을 거쳐 선정된 55편의 문헌은 출판년도 순으로 일련번호를 주었으며, 같은 호에서는 잡지에 실린 순서에 따랐다.

2.2 실험순서

(1) 색인작업

「Computer Networks」에서는 각 논문기사마다 약 6～10개의 키이워어드를 주고있는데, 본 실험에서는 이 키이워어드들을 통제하여 실험문헌집단에 대한 색인어로 이용하였다.[24]

첫째, 1979년 「The Institution of Electrical Engineers Thesaurus」와 1976년 「NCC Thesaurus of Computing Terms」를 이용해서 동의어를 조절하여 주었다. 둘째, 보다 특정한 키이워어드들은 전자계산학과 석사과정 2학년생들이 동의어를 조절하였다. 셋째, 문헌집단의 규모가 작은데 비하여 키이워어드들이 너무 흩어지는 것을 방지하기 위해서 빈도는 1번이나 2번이지만 주제성이 강한 키이워어드들은 계층분류를 이용하여 동의어로 묶어 주었다. 이러한 방법으로 통제한 결과 100개의 키이워어드로 어휘집을 구성하게 되었다.

(2) 질문작성

질문자는 전자계산학과에서 컴퓨터 네트워크를 전공하는 석사과정 2학년생 2명으로 구성하였으며, 실험문헌집단을 이루는 문헌의 내용을 중심으로 해서 6개의 질문을 작성하였다. 질문은 자연언어가 아니라 통제된 키이워어드를 사용해서 작성하였으므로 질문의 주제분석 과정은 생략되었다. 이러한 원질문을 구성하는 키이워어드 즉 탐색용어에는 역문헌빈도 가중치를 부여하였다.

---

24) 「Computer Networks」 잡지에서는 년말호(No. 6)에 "Subject Index"를 만들어서 그 해에 실린 논문기사에 주었던 키이워어드들을 모두 자모순으로 배열하고 있다. 그리고는 각각의 키이워어드가 어느 논문에 있는지 그 페이지를 함께 밝혀 놓았다. 그러므로 이러한 "Subject Index"를 모아서 하나의 어휘집(vocabulary)을 만들면 각각의 문헌을 하나의 벡터로 표시할 수 있다. 동시에 질문을 구성하는 키이워어드들은 묶어서 하나의 벡터 질문식을 만들어 검색할 수 있다.

(3) 문헌검색

문헌과 질문간의 유사성을 나타낼 수 있는 매칭함수를 이용해서 유사계수를 계산하여 첫번째 검색에서는 상위(上位) 6개의 문헌을 검색하였으며, 그 후에는 첫번째 검색결과에 따라 검색기준치를 정하여 검색하였다.

(4) 질문수정과 재검색

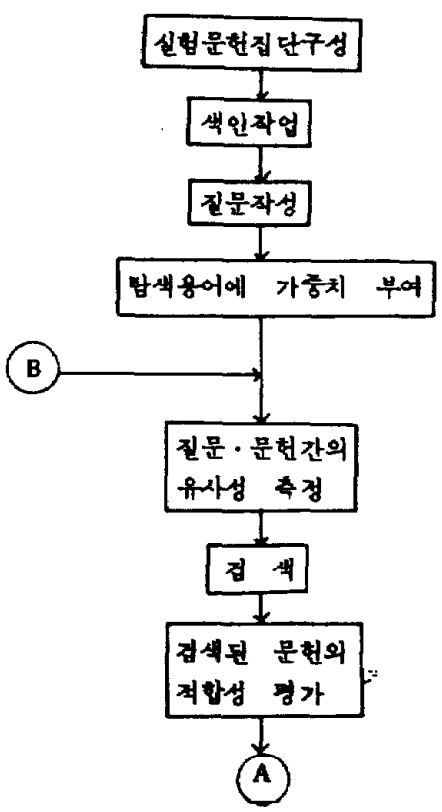
질문자들에게 검색된 문헌의 표제와 초록을 제시하여 주어서 각 문헌들이 적합문헌인지를 판정하게 한 다음, 적합문헌들의 색인어를 원질문에 추가시키고 그 용어들의 정확성 가중치(precision weights)를 추정한다. 그리고는 질문수정공식을 이용해서 질문을 수정하였다. 수정된 질문을 사용해서 55편의 실험문헌집단을 대상으로 재검색을 실시하였다.

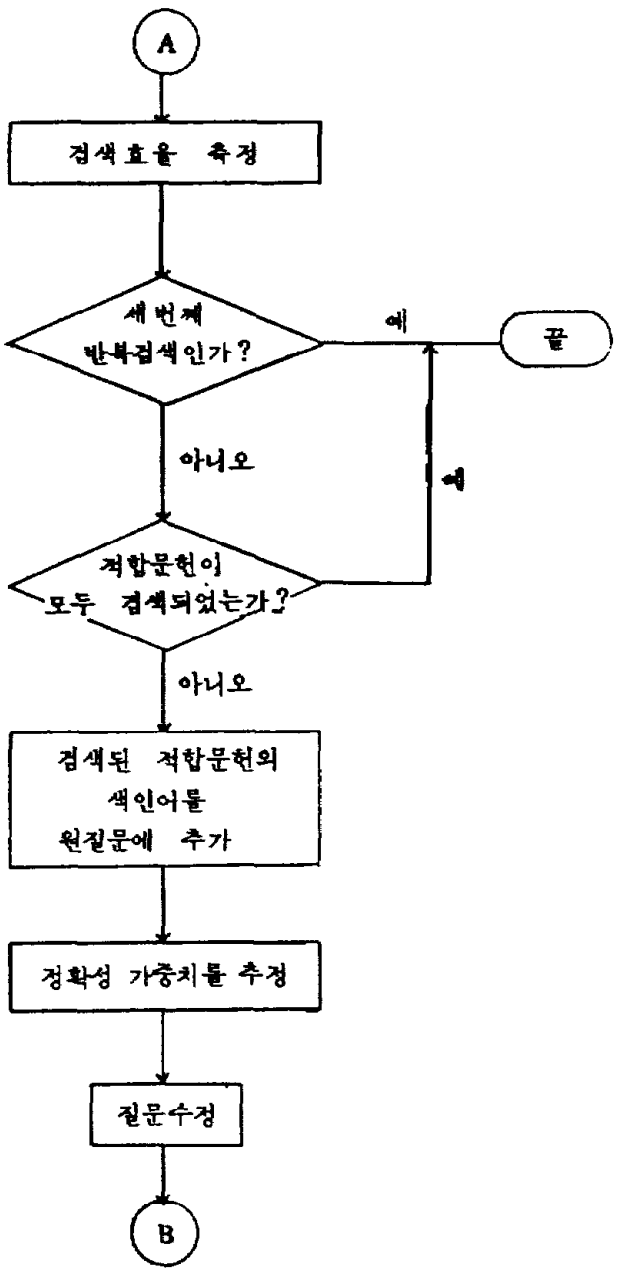
(5) 검색결과 평가

검색결과가 나오면 재현율과 정확률이라는개념을 근거로 한 평가기준을 이용하여 매번의 검색효율을 측정하였다. 실험결과의 평가 및 분석은 본 연구자가 하였다.

이상과 같은 실험순서를 흐름도로 나타내보면 그림 3과 같다.

그림 3 실험순서 흐름도





2.3 실험에서 사용한 모형

2.3.1 질문·문헌간의 유사성 측정방법

본 실험에서는 문헌의 색인어인 100개의 통제된 키이워어드로 구성된 어휘집을 이용하여 문헌과 질문을 표현하였는데 그 방법은 다음과 같다. 문헌은 하나의 백터로 표시하고 이원색인법을 사용하여 해당되는 색인어가 있을 때에는 1, 없을 때에는 0으로 표시하였다.[25]

$D_i = (d_{i1}, d_{i2}, \cdots\cdots, d_{i100})$에서 D_i는 i번째 문헌을 나타내며 $i = 1, 2, 3\cdots\cdots$, 55인 자연수이다. d_{ik}는 i번째 문헌에서 k번째 색인어의 가중치로 그값은 1 또는 0이 된다.

질문식은 문헌의 표현방법에 따라서 백터 질문식으로 작성하였으며 탐색용어에는 가중치를 주었다. $Q_j = (q_{j1}, q_{j2}, \cdots\cdots, q_{j100})$으로 질문식을 표현하게 되는데, 이 식에서 Q_j는 j번째 질문이며 q_{jk}는 j번째 질문에서 k번째 탐색용어의 가중치를 뜻하는 것으로 q_{jk}는 반드시 $q_{jk} \geq 0$를 만족시켜야 하며, $q_{jk} = 0$이라는 것은 질문 Q_j에는 k번째 용어가 들어있지 않음을 의미한다.

이러한 방법으로 표현된 문헌과 질문간의 유사성을 측정하기 위한 매칭함수로는 각 백터간의 내적(內積)을 이용하였다.

$$SIM(D_i, Q_j) = \sum_{k=1}^{100} d_{ik} \cdot q_{jk}$$

이와 같은 매칭함수에 의한 검색기준은 다음과 같다. 문헌과 질문간의 유사계수가 큰 것부터 작은 것의 순으로 문헌을 배열하여 첫번째 검색에서는 상위(上位) 6개의 문헌을 검색하도록 하여서 첫번째 피이드백으로 이용되는 문헌의 수를 통일시켰다. 그후의 반복 검색에서는 첫번째 검색에서 6번째 문헌과 7번째 문헌의 유사계수를 참작하여 각 질문의 검색기준치(*retrieval threshold*)를 정한 다음, 그 기준치 이상인 문헌을 검색하도록 하였다.

2.3.2 원질문 가중치 부여방법

원질문을 구성하는 탐색용어에 대한 가중치로는 역문헌 빈도인자(*inverse document frequency factor*)를 이용한 가중치를 주었다. 역문헌빈도 가중치란, 용어의 가중치는 그 용어가 들어있는 문헌의 수에 반비례한다는 것으로서, 다음과 같은 식을 이용해서 산출된다.

$$(IDF)_k = \log_2 \frac{N}{f_k}$$

N : 문헌의 총수, f_k : 용어 k의 문헌빈도

역문헌빈도 가중치를 계산하는데 필요한 변수는 문헌집단에 대한 통계에서 쉽게 얻을 수 있다.

본 실험에서는 원질문의 가중치로 우선 역문헌빈도 가중치를 주어 검색을 한 후, 검색된 적합문헌과 비적합문헌에서 그 용어들의 출현에 대한 정보를 이용하여 정확성 가중치를 추정해서 이를 질문수정과정에서의 가중치로 바꾸어 이용하고 있는데, 이러한 방법은 다음과 같은 역문헌빈도 가중치와 정확성 가중치의 관계에서 볼 때 그 타당성을 입증할 수 있다.[26] 첫째, 용어의 문헌빈도가 적합문헌수보다 큰 범위에서 정확성 가중치가 감소하는 비율은 역문헌빈도 가중치와 같으며 둘째, 중간빈도의 용어들

---

25) 문헌에 이원색인법을 사용한 이유는, 문헌과 질문에 모두 가중치를 부여하면 정확성 가중치를 추정하는 과정이 복잡하고 어려워지기 때문이다.

26) Harry Wu, op. cit., pp. 34-51.

은 정확성 가중치와 역문헌빈도 가중치가 거의 같고 셋째, 일반적으로 질문을 구성하는 탐색용어들에 대한 두 가중치는 별로 차이가 없다는 점을 들 수 있다.

역문헌빈도 가중치 외에도 처음부터 정확성 가중치를 추정하여 원질문에 부여하는 방법이 있을 수 있겠으나, 이 방법은 검색효율면에서는 역문헌빈도 가중치와 비슷하면서도 그 추정과정이 복잡하다.[27] 따라서 원질문에 대한 가중치로는 역문헌빈도 가중치를 이용하는 것이 바람직하다고 하겠다.

2.3.3 정확성 가중치 부여방법

정확성 가중치란 이용자가 제시하는 적합성 정보를 이용하여 가중치를 부여하는 방법으로서, 각 질문에 대한 적합문헌과 비적합문헌에서 해당용어가 어떻게 출현하는가 하는 특성에 따라서 가중치가 달라진다. 정확성 가중치는 다음과 같은 공식으로 산출된다.

$$W_i = \log \left\{ \frac{P_i}{1-P_i} \div \frac{U_i}{1-U_i} \right\} \cdots\cdots (1)^{28)}$$

위식에서 P_i 는 용어 i 가 적합문헌에 출현할 확률을, U_i 는 비적합문헌에 출현할 확률을 나타낸다.

본 실험에서는 적합성 정보를 얻는데 다음과 같은 적합성 피이드백 과정을 이용하였다. 즉 6개의 질문에 대하여 검색된 문헌의 표제와 초록을 해당 질문자에게 제시하여 주고, 질문자가 검색된 문헌중에서 적합문헌을 판정하면 그 적합문헌을 이용하여 각 용어의 정확성 가중치를 추정하였다. 정확성 가중치를 추정하는 방법은 다음과 같다. 예를들어, 검색기준치를 K라 하면 검색된 문헌들은 모두 그 유사계수가 K이상이 된다. 즉

$$\sum_{k=1}^{n} d_{ik} \cdot q_{jk} \geq K \qquad (2)$$

를 만족시킨다. 이때, 임의의 질문에 대한 검색결과에서 그 질문을 구성하는 각 탐색용어와 검색된 적합문헌들과의 관계를 생각해 보자. 그 질문을 구성하는 탐색용어들 중에서 어느 하나를 h라고 하였을때 검색된 적합문헌들은 h를 기준으로 하여 두 부류로 나눌 수 있다. 첫째는 그 문헌안에 h가 색인되어 있든, 없든 관계없이 검색된 문헌들로서, h가 문헌안에 없어도 검색될 수 있으므로 다음과 같은 조건을 만족시킨다.

$$\sum_{\substack{k=1 \\ k \neq h}}^{n} d_{ik} \cdot q_{jk} \geq K \qquad \cdots\cdots (3)$$

둘째는 h가 문헌안에 꼭 있어야만 검색되는 문헌들로 이루어진 집합으로서, h가 없을 경우의 유사계수는 K보다 작아져서

$$\sum_{\substack{k=1 \\ k \neq h}}^{n} d_{ik} \cdot q_{jk} < K \qquad \cdots\cdots (4)$$

을 만족시킨다. 그러나 h가 있으면 그 문헌의 유사계수는

---

27) Harry Wu and G. Salton, op. cit., pp. 200-202.

28) 정확성 가중치공식 W_i 는 용어의 출현 빈도를 변수로 하는 공식으로 바꾸어 이용될 수 있다. 그 공식은

$$W_i = \log \left\{ \frac{r_i}{R-r_i} \div \frac{s_i}{I-s_i} \right\}$$

로 표시할 수 있는데, 여기서 R은 주어진 질문에 대하여 장서에 들어있는 적합문헌의 수, I는 비적합문헌의 수를, r_i 는 용어 i 가 들어있는 적합문헌의 수, s_i 는 용어 i 가 들어있는 비적합문헌의 수를 나타낸다.

$$\sum_{\substack{k=1 \\ k \neq h}}^{n} d_{ik} \cdot q_{jk} + q_{jh} \geq K$$

가 된다.

이상과 같은 탐색용어와 검색된 적합문헌들의 관계를 2×2테이블(contingency table)로 나타내면 표1과 같다.

표1. 검색된 적합문헌에서 탐색용어의 출현상황

	h가 있는 문헌 $d_{ih} = 1$	h가 없는 문헌 $d_{ih} = 0$
$\sum_{\substack{k=1 \\ k \neq h}}^{n} d_{ik} \cdot q_{jk} < K$	a	0
$\sum_{\substack{k=1 \\ k \neq h}}^{n} d_{ik} \cdot q_{jk} \geq K$	b	c

표1에서 a, b, c는 각 조건을 만족시키는 문헌의 수를 나타내는 것으로서 오른쪽 윗칸은 비어있는데, 그 이유는 식(2)와 (4)를 비교해 볼때 h가 없는 문헌은 그 유사계수가 기준치보다 작아져서 검색될수 없기 때문이다.

표1과 같은 특성을 갖는 용어의 정확성 가중치를 계산하기 위해서는, 적합문헌의 완전한 집합에서 그 용어가 출현할 확률(P_h)과 비적합문헌에서 출현할 확률(U_h)을 추정하여야 한다. 일반적으로 통용되고 있는 추정방법에서는 P_h는 검색된 적합문헌의 수에 대해서 용어 h가 들어있는 검색된 적합문헌의 수의 비율이라고 정의하고 표1상의 a, b, c로 나타내면 다음과 같다.

$$P_h = \frac{a+b}{a+b+c}$$

그러나 표1에서 오른쪽 윗칸이 비어있다는 것은 이 조건을 만족시키는 문헌이 없다는 것이므로, 검색된 적합문헌의 집합 $\{a+b+c\}$는 P_h를 추정하기 위한 무작위 표본(random sample)이 못된다. 그런데 실험 시스템에 대한 전제조건에서 용어들이 적합문헌과 비적합문헌들 안에서 독립적으로 분포되어 있다고 가정하였으므로 첫번째 부류의 문헌들은 P_h를 추정하기 위한 무작위 표본이 된다. 따라서 표본$(b+c)$안에서 h의 출현을 나타내는 특성은 완전한 적합문헌 집합에서의 특성과 같다. 이것을 수학적으로 표현하면 d_{ih}와

$$\sum_{\substack{k=1 \\ k \neq h}}^{n} d_{ik} \cdot q_{jk} \geq K$$

가 서로 독립적이므로

$$P_r \{ d_{ih} = 1 \mid \sum_{\substack{k=1 \\ k \neq h}}^{n} d_{ik} \cdot q_{jk} \geq K \}$$

$$= P_r \{ d_{ih} = 1 \}$$

을 만족시킨다.

그리하여 용어 h가 적합문헌에 출현할 확률은 다음과 같이 추정할 수 있다.

$$P_h = \frac{b}{b+c} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (5)^{29)}$$

한편 비적합문헌에서 h가 출현할 확률인 U_h는 검색된 적합문헌에 대한 보집합을 이용해서 추정할 수 있다.

29) S.E. Robertson and K. Sparck Jones, "Relevance Weighting of Search Terms," *JASIS*, (May-June, 1976), pp. 129-146; D. Chow and C.T.Yu, "On the Construction of Feedback Queries," *JACM*, 29, (Jan. 1982), 127-151.

$$U_h = \frac{f_h - (a+b)}{N-(a+b+c)} \quad \cdots\cdots (6)$$

이식에서 N은 문헌의 총수, f_h는 용어 h가 색인되어 있는 문헌의 총수, $(a+b+c)$는 검색된 적합문헌의 수, $(a+b)$는 검색된 적합문헌 중에서 용어 h가 색인되어 있는 문헌의 수를 나타낸다.

본 실험에서는 식(5)와 (6)이 0으로 나뉘어지는 것을 막기 위하여 아래와 같이 약간 수정하여 이용하였다.[30]

$$P'_i = \frac{b+0.5}{b+c+1} \quad \cdots\cdots (7)$$

$$U'_i = \frac{f_i - (a+b) + 0.5}{N-(a+b+c)+1} \quad \cdots\cdots (8)$$

식(7)과 (8)을 정확성 가중치 공식인 (1)에 대입시켜서 각 용어에 대한 정확성 가중치를 추정하였다.

2.3.4 질문수정공식

본 실험에서는 원질문의 형태와 비교해 볼때 질문을 구성하는 탐색용어와 가중치를 모두 수정하여 새로운 질문을 작성하였다. 즉 수정된질문은 검색결과 적합문헌으로 판정된 문헌의 모든 색인어를 원질문의 탐색용어에 추가시켜서 탐색용어를 구성하였다. 한편, 가중치는 먼저 실시한 검색에서 이용된 질문의 가중치와 새로 추정된 정확성 가중치를 다음과 같은 비율로 더하여서 새로운 질문에 대한 가중치(피이드백 가중치)를 산출하였다.

피이드백 가중치 $= \alpha \cdot$ (먼저 질문의 가중치) $+ \beta \cdot$ (정확성 가중치)

여기서 $\alpha = 1-\beta$이며, $a+b+c \leq R$이면 $\beta = \frac{a+b+c}{R}$, 그렇지 않으면 $\beta = 1$이 된다.

이러한 수정방법을 식으로 표시하면 다음과 같다.

$$\begin{aligned} Q_{new} = (1-\beta)\ & (T_1, q_1 ; T_2, q_2 ; \cdots\cdots \\ & ; T_n, q_n ; T_{n1}, 0 ; \cdots\cdots ; T_{nm}, 0) \\ & + \beta (T_1, q'_1 ; \cdots\cdots ; T_n, q'_n ; T_{n1}, \\ & q'_{n1} ; \cdots\cdots ; T_{nm}, q'_{nm}) \end{aligned}$$

이식에서 q_i는 먼저 질문의 가중치를, q'_i는 정확성 가중치를 나타내며, $(T_1, \cdots\cdots, T_n)$은 먼저 질문의 탐색용어를, $(T_{n1}, \cdots\cdots, T_{nm})$은 검색된 적합문헌에 의해서 추가되는 용어들을 의미한다. 이때 추가되는 용어들은 먼저 질문에 들어 있지 않은 용어들이므로 먼저 질문에서 그 용어들의 가중치는 0이라고 가정한다.

2.4 실험내용

본 절에서는 실험에 이용한 6개의 질문과 각 질문에 의한 검색결과 검색된 문헌의 순위들을 표로 만들어서 나타내 주었다. 검색결과인 문헌의 순위는 적합문헌이 모두 나타나는 순위에서 적당히 끊어 주었으며 유사계수가 같은 문헌들은 그 안에서 문헌번호의 순에 따라 배열해 주었다.

그리고 적합문헌의 경우 문헌번호 뒤에 "R"을 써주어 식별할 수 있게 하였으며, 검색기준치가 위치할 곳에 "……" 표시로 해주어서 검색되는 문헌의 수를 파악하기 쉽게 하였다. 즉 각각의 검색결과에서 "……"표시위에 놓인 문헌들은 모두 검색되어서 이용자에게 제시되는 문헌들이다.

검색을 반복 시행하는 횟수는 선행연구의 결

30) S.E. Robertson and K. Sparck Jones, op. cit., pp. 143-144.

순 위	첫번째 검색		첫번째 반복검색		두번째 반복검색		세번째 반복검색	
	문헌번호	유사계수	문헌번호	유사계수	문헌번호	유사계수	문헌번호	유사계수
1	49 R	8.85	49 R	30.73	49 R	46.88	49 R	48.72
2	7	7.24	18 R	26.97	18 R	41.99	18 R	44.73
3	11	6.98	14 R	12.20	14 R	16.57	14 R	19.85
4	14 R	6.98	11	11.83	10 R	16.42	11	19.44
5	2	5.39	10 R	7.25	11	16.21	10 R	17.52
6	18 R	5.39	7	6.49	53 R	12.79	53 R	16.89
7	10 R	5.07	53 R	5.92	7	7.17	13 R	15.49
8	48	4.81	2	5.43	2	7.11	15 R	7.16
9	54	4.81	13 R	4.62	24	6.56	7	7.15
10	13 R	3.78	54	4.03	32	6.39	32	6.57
11	53 R	3.78	5	3.92	13 R	6.15	24	6.35
12	9	3.46	24	3.75	5	5.86	12	6.27
13	15 R	3.46	48	3.62	37	5.18	2	6.23
14	17	3.20	37	3.34	8	4.75	5	5.96
15	30	3.20	9	3.19	12	4.73	9	5.71
16	1	1.61	15 R	3.19	15 R	4.60	44	5.34
17	5	1.61	32	2.93	30	4.58	8	5.03

과들을[31] 참조하여 원칙적으로 세번으로 정하였으나, 질문 3, 4, 5, 6의 경우는 두번째 반복검색에서 적합문헌이 모두 검색되었으므로 그 시점에서 검색을 끝냈다.

각 질문에 대한 적합문헌의 수는, 소규모의실험문헌집단이므로 전체 55편의 문헌을 조사해서 사전에 알아낼 수 있었다. 이러한 검색과정은 수작업에 의해서 이루어졌다.

2.4.1 질문 1

질문1은 "컴퓨터 네트워크의 상호연결에서 단말상호(端末相互) 협약의 모형화와 그 모형의 유효성에 관한 문헌"이며 이 질문에 관한 적합문헌은 10, 13, 14, 15, 18, 49, 53의 7개의 문헌이다. 첫번째 검색결과에 따라 정해진 검색기준치는 5.2이며, 매번의 검색결과 실험문헌집단을 구성하는 문헌들의 순위는 위표와 같이 나타났다.

2.4.2 질문 2

질문2는 "컴퓨터 네트워크 중 국소지역 네트워크 고리구조 네트워크의 조정방식과 공보통신의 CSMA에 관한 문헌이며 적합문헌으로는 20, 34, 35, 38, 43, 45, 46, 50이 있다. 검색기준치는 5.5이며 질문을 수정하여 반복검색을 한 결과, 각각의 검색에서의 문헌순위는 다음과 같다.

31) 검색질문을 수정하여 반복검색을 실시한 연구결과들을 보면, 첫번째 수정에 의한 검색효율의 향상이 가장 크며, 그 향상의 폭은 점차 줄어들게 된다. 따라서 일반적으로 4번 이상의 반복검색은 무의미한 것으로 나타나고 있다.

순 위	첫번째 검색		첫번째 반복검색		두번째 반복검색		세번째 반복검색	
	문헌번호	유사계수	문헌번호	유사계수	문헌번호	유사계수	문헌번호	유사계수
1	20 R	8.86	20 R	14.54	43 R	16.68	45 R	22.96
2	30	8.86	43 R	11.83	45 R	16.54	43 R	19.48
3	52	8.86	52	10.10	20 R	14 63	35 R	18.28
4	1	8.01	30	9.66	35 R	13.97	20 R	14.89
5	43 R	7.85	1	7.86	52	10.28	31	12.65
6	42	5.66	33	7.10	30	9.75	52	11.28
7	45 R	5.39	32	7.04	31	8.93	30	10.00
8	35 R	4.81	42	6.47	33	7.18	46 R	9.03
9	32	4.07	45 R	6.34	1	7.12	6	8.48
10	23	3.78	35 R	6.32	42	6.67	38 R	7.41
11	34 R	3.78	46 R	5.13	46 R	6.63	1	6.92
12	36	3.20	24	4.64	6	6.31	23	6.40
13	46 R	3.20	37	4.64	54	5.36	33	6.31
14	10	2.46	54	4.64	32	5.33	34 R	5.55
15	33	2.46	23	3.87	23	5.29	17	4.96
16	38 R	2.46	36	3.85	38 R	4.71	54	4.95
17	41	2.46	38R	3.72	36	4.49	36	4.53
18	51	2.46	11	3.71	34 R	4.25	26	4.15
19			34 R	3.43	37	4.22	15	4.05
20			19	3.16	24	3.85	32	3.68
21			41	2.84	11	3.63	50 R	3.38
22			51	2.84	17	3.31		
46					50 R	1.42		
55	50 R	0.00	50 R	0.00				

2.4.3 질문 3

질문 3은 "컴퓨터 네트워크의 협약에서 단말상호 협약과 그것들의 검사에 관한 문헌" 이며, 이에 대한 적합문헌은 11, 12, 13, 14, 53, 54이다. 검색기준치는 4.5이다.

2.4.4 질문 4

질문 4는 "국소지역 네트워크에서 고리구조 네트워크에 관한 일반적인 정보와 공보통신 방식중 하이퍼채널 방식을 이용한 통신에 관한 문헌"이며, 검색기준치는 5.92이다.

질문3에 의한 검색결과에서 문헌의 순위는 다음과 같다.

순 위	첫번째 검색		첫번째 반복검색		두번째 반복검색	
	문헌번호	유사계수	문헌번호	유사계수	문헌번호	유사계수
1	10	5.07	11 R	20.12	11 R	35.52
2	49	5.07	53 R	15.85	53 R	28.02
3	11 R	4.98	13 R	14.80	13 R	25.05
4	13 R	4.98	14 R	11.60	14 R	24.69
5	53 R	4.98	18	7.79	12 R	20.36
6	18	4.81	12 R	4.99	18	16.05
7	7	3.46	32	3.84	5	5.21
8	9	3.46	9	3.72	32	5.14
9	15	3.46	15	3.72	54 R	4.52
10	24	3.39	40	3.45	40	4.43
11	32	3.39	5	3.07	15	4.30
12	37	3.39	54 R	2.91		
13	54 R	3.39	37	2.24		
14	12 R	3.20				
15	14 R	3.20				

질문 4에 의한 검색결과 문헌의 순위는 아래와 같다.

순 위	첫번째 검색		첫번째 반복검색		두번째 반복검색	
	문헌번호	유사계수	문헌번호	유사계수	문헌번호	유사계수
1	20 R	9.12	46 R	18.55	46 R	29.28
2	52	9.12	30 R	16.35	30 R	23.30
3	43 R	8.21	20 R	13.34	20 R	17.06
4	30 R	7.63	43 R	11.71	38 R	15.54
5	46 R	6.66	52	10.13	43 R	14.18
6	33	5.92	23 R	7.12	52	12.92
7	38 R	5.92	1 R	6.69	1 R	11.26
8	23 R	5.75	38 R	6.65	23 R	10.77
9	45	5.75	45	6.51	45	8.14
10	1 R	5.17	42 R	5.76	48	7.54
11	42 R	4.43	33	5.69	42 R	7.09
12	34	3.78			33	6.47
13					41	4.80

2.4.5 질문 5

질문 5는 "분산체제에서 패킷을 이용하여 데이터를 전송하는 시스템의 경로지정(經路指定) 방법과 착오교정, 또 이런 시스템의 모형화와 성능분석에 관한 문헌"이며 이에 대한 적합문헌은 23, 31, 34, 45, 46, 48, 50이다. 검색기준치는 7.5이며 각각의 검색결과에서 문헌의 순위는 다음과 같다.

순 위	첫번째 검색		첫번째 반복검색		두번째 반복검색	
	문헌번호	유사계수	문헌번호	유사계수	문헌번호	유사계수
1	48 R	11.06	46 R	20.33	46 R	25.03
2	46 R	9.86	48 R	11.21	45 R	20.89
3	30	8.86	30	11.19	23 R	12.90
4	23 R	7.53	38	10.47	50 R	12.34
5	37	7.53	50 R	9.77	30	12.11
6	50 R	7.53	23 R	9.57	38	12.10
7	17	7.40	45 R	7.82	48 R	11.11
8	45 R	6.27	34 R	5.92	34 R	8.56
9	38	5.66	26	5.78	31 R	7.94
10	27	5.33	17	5.05	26	6.85
11	47	5.33	20	4.89	16	5.99
12	26	4.40	16	4.49	17	5.90
13	20	4.20	25	3.87		
14	1	4.07	37	3.73		
15	31 R	3.46	52	3.10		
16	11	3.20	1	2.67		
17	14	3.20	54	2.64		
18	54	3.20	36	2.31		
19	5	2.20	27	2.19		
20	7	2.20	47	2.19		
21	16	2.20	44	2.09		
22	22	2.20	31 R	1.83		
23	25	2.20	43	1.68		
24	34 R	2.20				

순 위	첫번째 검색		첫번째 반복검색		두번째 반복검색	
	문헌번호	유사계수	문헌번호	유사계수	문헌번호	유사계수
1	10 R	11.50	10 R	17.79	10 R	21.14
2	49	8.53	9 R	13.69	9 R	19.18
3	8 R	8.21	8 R	11.34	8 R	14.41
4	24	6.85	15	9.65	15	12.58
5	7	6.43	49	9.02	40 R	12.18
6	9 R	6.43	28	8.11	28	10.94
7	11	5.56	7	7.64	7	9.44
8	13 R	5.56	40 R	7.00	49	8.19
9	53	5.56	24	6.24	44 R	6.11
10	18	5.07	44 R	5.38	13 R	5.63
11	28	4.75	13 R	5.24	24	5.58
12	40 R	4.75	18	4.57		
13	44 R	4.58				
14	14	3.78				

2.4.6 질문 6

질문 6은 "처리장치들의 통신을 위한 네트워크협약에서 단말상호협약과 x.25에 관한 문헌, 그리고 그것들의 유효성에 관한 문헌"이며 적합문헌은 8, 9, 10, 13, 40, 44로서 검색기준치는 5.6이다. 질문6에 의한 검색결과 문헌들의 순위는 위와 같다.

2.5 검색효율 측정

검색에 의하여 얻어진 정보가 이용자의 요구를 어느 정도 만족시켜 주고 있는가에 관점을 두고, 시스템의 재현율(recall ratio)과 정확률(precision ratio)을 근거로 한 평가기준을 사용하여 실험결과의 검색효율을 측정하였다.

2.5.1 재현율과 정확률

검색결과는 정해진 검색기준치 점에서의 재현율과 정확률로 나타낼 수 있다. 재현율이란 적합한 문헌중에서 검색된 문헌의 백분율이며, 정확률은 검색된 문헌중에서 적합문헌의 백분율을 뜻한다.

그림 4는 6개 질문의 재현율과 정확률을 각각 매번의 검색결과가 나올 때마다 측정한 것이며, 그림 5는 그림 4를 평균한 것으로서, 그 반복횟수가 다르므로 질문 1～6은 두번째 반복검색까지를 평균하였고, 세번째 반복검색까지 실시한 질문1과 2는 추가로 다시 평균을 내어 표시해 주었다.

그림 4. 6개 질문의 재현율과 정확률

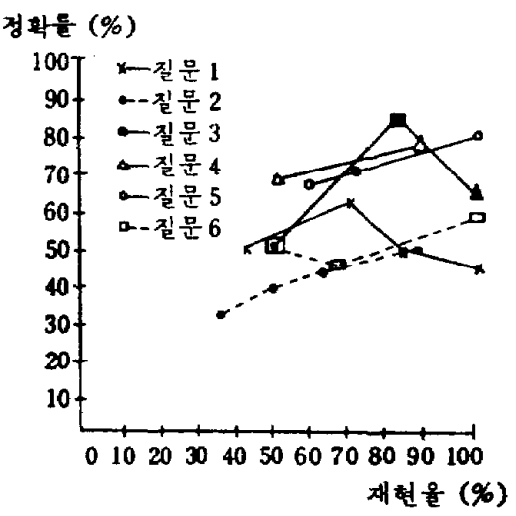
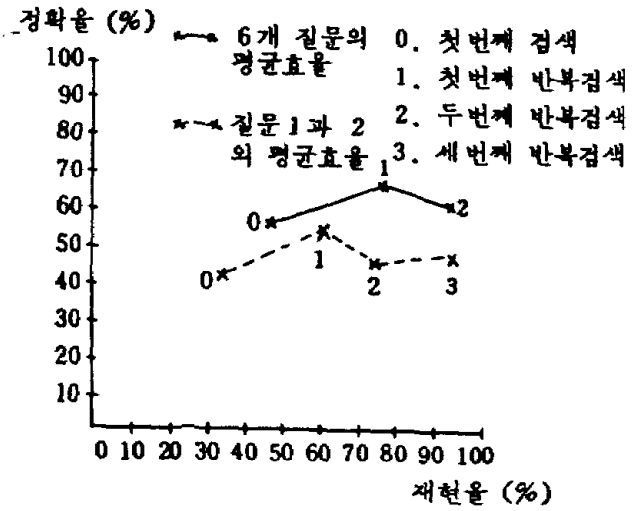


그림 5 평균 재현율과 정확률



2.5.2 표준재현율과 표준정확률[32]

'적합문헌들이 어떤 순위를 갖는가?' 하는 점에서 검색효율을 평가할 수 있다. 이때 가장 이상적인 시스템이라면, n개의 적합문헌이 있을때 그 적합문헌의 순위는 1, 2, ……, n이 될 것이다. 표준재현율(normalized recall)과표준정확률(normalized precision)은 이러한 개념에 근거한 것으로서, 이상적인 시스템의 결과와 실제의 검색결과를 동일한 그래프에 그렸을 때 그 면적의 차이를 평가기준으로 이용하는 것이다.

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

$$\text{표준정확률} = 1 - \frac{\sum_{i=1}^{n} \log r_i - \sum_{i=1}^{n} \log i}{\log\,[N!/(N-n)!\,n!]}$$

로 나타내어지는데 이 식에서 N은 총 문헌의 수, n은 적합문헌의 수, r_i는 적합문헌의 실

32) 1966년 로치오가 제안한 평가기준이다.
G. Salton, *Automatic Information Organization and Retrieval* (N.Y.: McGraw-Hill Book Co. 1968), pp. 285-289.

표 2 표준재현율과 표준정확률

		첫번째 검색	첫번째 반복검색	두번째 반복검색	세번째 반복검색
질문 1	표준재현율	0.9286	0.9554	0.9554	0.9881
	표준정확률	0.7980	0.9063	0.9156	0.9638
질문 2	표준재현율	0.6915	0.6702	0.8271	0.9282
	표준정확률	0.6762	0.6763	0.9948	0.9970
질문 3	표준재현율	0.8878	0.9762	0.9898	
	표준정확률	0.9920	0.9987	0.9994	
질문 4	표준재현율	0.9654	0.9867	0.9867	
	표준정확률	0.8941	0.9669	0.9710	
질문 5	표준재현율	0.9048	0.9315	0.9821	
	표준정확률	0.8269	0.8596	0.9542	
질문 6	표준재현율	0.9252	0.9524	0.9694	
	표준정확률	0.7998	0.8840	0.9231	
평균	표준재현율	0.8839	0.9121	0.9518	
	표준정확률	0.8312	0.8820	0.9597	

제순위를, i는 이상적인 시스템에서의 순위를 각각 나타낸다.

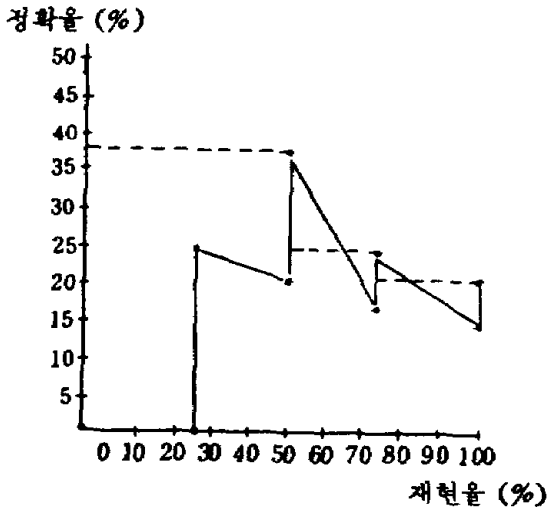
표2에서는 6개 질문의 각 검색결과를 표준정확률과 표준재현율로 측정하여 나타내 주었고, 6개 질문에 대한 평균도 산출해 주었다.

2.5.3 성능곡선(performance curve)

'일정한 재현율 수준에서 정확률이 어느 정도 향상되는가?'는 평가의 기준이 된다. 이 결과는 재현율과 정확률의 상관관계를 그래프로 표시한 성능곡선을 이용해서 나타낼수 있다. 이때 실제 관찰된 점을 이용해서 일정한 재현율 수준에 대한 정확률을 얻기 위해서는 보간법(interpolation)이 이용된다.[33]

본 논문에서는 동일한 재현율 수준에서 관찰된 정확률 중 최고점에서부터 왼쪽으로 수평선을 그어서 그보다 큰 정확률 값을 만나는 점에서 멈추는 방법인 새 클리버든 보간법(neo-Cleverdon interpolation)을 이용했다.

그림 6. 새 클리버든 보간법[34]



각 질문에 대해서 만들어진 이러한 곡선을 평균내는 데에는 산술평균(macroevaluation)을 이용하였다.[35]

그림7. 성 능 곡 선

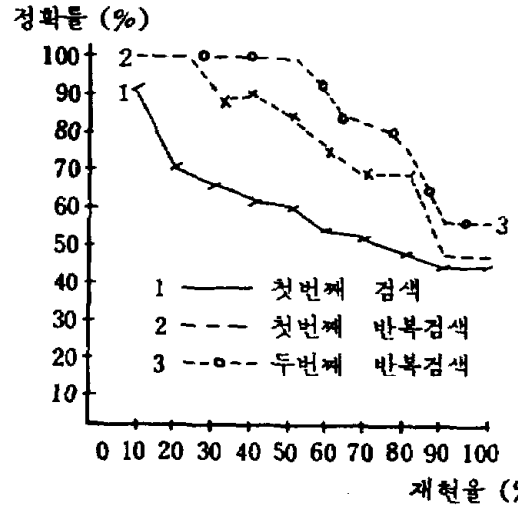


그림7에서는 각 검색결과에 대하여 질문 6개를 평균한 성능곡선을 보여주고 있으며, 표3은 그림7을 표로 나타내준 것으로서 검색효율의 상승비율을 백분율로 나타내어 함께 보여주고 있다.

2.5.4 피이드백 평가(feedback evaluation)

홀(Hall)과 와이더만(Weiderman)은 적합성 피이드백에 의해 검색효율이 향상되는 데에는 순위효과(ranking effect)와 피이드백 효과(feedback effect)라는 두 가지 요인이 있다고 하였다.[36] 즉, 질문을 수정해서 반복검

33) G. Salton, ed., *The SMART Retrieval System: Experiments in Automatic Document Processing* (N.J.: Prentice-Hall, 1971), pp. 342-344.

34) 적합문헌의 순위가 4, 6, 12, 20인 경우의 그래프이다. ●——● 점들은 실제 관찰된 점들이다.

35) G. Salton, *Automatic Information ...*, pp. 298-300.

36) H.A. Hall and N.H. Weiderman, *The Evaluation Problem in Relevance Feedback Systems*, Report No. ISR-12 to the National Science Foundation, Section XII (N.Y.; Cornell University, 1966)

표 3. 성 능 도 표

(단위 : %)

정확률 / 재현율	첫번째검색	첫 번 째 반복검색	두 번 째 반복검색	첫번째 검색에 대한 두번째 반복 검색의 상승비율
10	93	100 + 7.53	100 + 0	+ 7.53
20	69	100 + 44.93	100 + 0	+ 44.93
30	65	87 + 33.85	100 + 14.94	+ 53.85
40	62	87 + 40.32	100 + 14.94	+ 61.29
50	61	83 + 36.06	100 + 20.48	+ 63.93
60	55	72 + 30.91	81 + 12.5	+ 47.27
70	54	67 + 24.07	75 + 11.94	+ 38.89
80	50	67 + 34.	70 + 4.48	+ 40.
90	43	44 + 2.33	57 + 29.55	+ 32.56
100	43	44 + 2.33	57 + 29.55	+ 32.56
		+ 25.63	+ 13.84	+ 42.28

색한 결과를 보면 이용자가 먼저번 검색결과에서는 보지 못했던 적합문헌들이 순위가 향상되어 새로 검색되기도 하지만(피이드백 효과), 그 밖에 이미 검색되어서 이용자가 적합문헌이라고 판정한 문헌이 다시 검색되면서 순위가 향상되는(순위효과)경우가 있다는 것이다. 사실상 적합성 피이드백을 이용한 질문수정에는 몇개의 적합문헌에서 얻은 정보 즉 색인어들이 이용된다. 따라서 이러한 목적에 이용된 문헌들은 수정된 질문으로 검색한 결과에서 먼저번 보다 순위가 높아지게 된다. 그러나 이용자는 이미 검색된 문헌에 대해서는 충분히 알고 있어서, '새로운 적합문헌이 검색되는가?'에만 관심을 갖고 있을 것이므로 이 두 가지는 분리시켜야 한다고 하였다.

홀과 와이더만이 피이드백 효과에 의한 검색효율을 측정하기 위해서 사용한 방법은 다음과 같다. 처음 탐색결과 검색되어서 피이드백으로 이용된 문헌은(N개의 문헌) 그 순위를 교정시키고, 장서의 나머지 부분을 이용해서 다시 N개의 문헌을 검색한 결과를 그다음 순위인 (N+1)부터 주어서 검색효율을 평가하였다.

표 4. 피이드백 평가

(단위 : %)

정확률 / 재현율	첫번째 검색	두번째 반복검색
10	93	100 + 7.53
20	69	75 + 8.70
30	65	72 + 10.77
40	62	71 + 14.52
50	61	71 + 14.52
60	55	67 + 21.82
70	54	67 + 24.07
80	50	65 + 30.00
90	43	55 + 27.91
100	43	55 + 27.91
		+ 18.78

본 연구자는 이 방법을 응용해서 실험결과를 평가해 보았다. 홀과 와이더만은 일정한 갯수의(N개) 문헌을 검색하는 방법을 이용하였으나 본 논문에서 실시한 실험에서는 일정한 기준치 이상인 문헌을 검색하도록 하였으므로 검색기준의 형태가 다르다 하겠다. 그러나 일단 검색된 문헌들이 이용자에게 제시되어 피이드백으로 이용되는 것은 같은 원리이므로, 본 논문에서는 정해진 기준치 이상이 되어서 이미 검색이 되었던 문헌들의 순위만을 교정시키고 효율을 측정하였다. 그러므로 홀과 와이더만의 방법에서는 검색을 반복함에 따라서 고정되는 문헌의 수가 N, 2N, 3N, ……이지만, 본 실험의 결과에서는 각각의 검색결과 정해진 기준치 이상인 문헌만을 고정하였으므로 순위가 교정되는 문헌수의 증가에 일정한 규칙은 없다.

표4는 피이드백 평가방법에 의한 검색효율을 일정한 재현율 수준에 따라서 표시한 것으로서, 두번째 반복검색에 의한 검색효율의 상승비율을 함께 나타내 주었다.

2.6 실험결과 분석

본 절에서는 앞절 2.5의 평가방법에 의해서 작성된 표와 그림을 이용해서 실험결과를 분석해 보았다. 검색실험에서 사용한 검색기준치는 임의적인 것으로서, 어떤 것을 쓰느냐에 따라서 산출되는 재현율과 정확률의 값이 달라진다. 따라서 재현율과 정확률 외에, 특정한 기준치를 쓰지 않고 전체 문헌장서에 대한 검색효율을 평가하는 방법들도 함께 이용해서 실험결과를 분석해보면 다음과 같다.

표 5. 6개 질문에 대한 통계

질 문	1	2	3	4	5	6	평균
적합문헌의 수 (개)	7	8	6	8	7	6	7
첫번째 검색에서 검색된 적합문헌의 수 (개)	3	2	3	4	4	3	3.2
첫번째 검색결과의 재현율 (백분율)	43	25	50	50	57	50	46

6개의 실험질문에 대한 사항은 표5와 같다.

정해진 검색기준치에서 검색효율을 나타낸 그림4에서, 검색효율이 향상되었다는 것은 재현율이 향상됨에 따라서 정확률도 향상되어 그래프상에서는(100,100)점에 가까와지게 되는 것을 뜻한다. 그러나 재현율과 정확률은 한쪽이 상승되면 다른 한쪽은 저하되는 반비례 관계에 있다. 재현율은 향상되었으나 정확률이 떨어지는 것은 추가로 검색되는 적합문헌의 수에 비해서 비적합문헌들이 더 많이 검색되었다는 것을 의미한다. 그러한 변화가 심한 질문1의 경우를 보면 두번째 반복검색 결과 1개의 적합문헌과 3개의 비적합문헌이 추가로 검색되었고, 세번째 반복검색에서는 다시 1개의 적합문헌과 3개의 비적합문헌이 새로 검색되어서 정확률은 계속 떨어지게 된다. 첫번째 검색결과와 비교해 볼때 세번째 반복 검색에서는 4개의 적합문헌과 6개의 비적합문헌이 추가되어서 정확률은 약간 떨어진다.

이론적으로는 개개의 질문에 대해서 그 결과를 평가하는 것이 좋지만, 시스템의 전반적인 효율을 평가한다는 면에서는 모든 질문에 대해서 평균을 낸 결과를 제시하는 것이 바람직하다. 그림5는 그림4의 6개의 질문을 평균낸 것으로서 두번째 반복검색 결과까지를 평균한 결과와 세번째 반복검색까지 실시한 질문1과 2

는 따로 평균을 내어서 표시했다. 그 결과를 보면 첫번째 반복검색에 의한 효율의 상승폭이 크며 두번째 반복검색부터는 정확률이 조금씩 떨어지는 것으로 나타났다. 그러나 그 떨어지는 폭은 아주 미미하여, 전체적으로 보아 검색효율이 향상된다고 말할수 있다.

표준재현율과 표준정확률을 나타낸 표2에서는, 보통의 정확률과 재현율이 반비례하는 것과는 달리, 어느 한쪽 값이 1이면 다른 한쪽 값도 1이된다. 질문2의 첫번째 검색과 첫번째 반복검색의 효율이 다른 질문에 비해서 많이 떨어지는 이유는 1개의 적합문헌이 순위상에 나타나지 않으므로 그 문헌에는 제일 끝 순위인 55를 주었기 때문이다. 그리고 질문1과 4의 첫번째 반복검색과 두번째 반복검색에서는 표준재현율이 같은 것을 볼 수 있다. 이는 표준재현율은 모든 적합문헌들의 비중을 균일하게 두어서 순위가 54에서 50으로 변한 것과 5위에서 1위로 변하는 것이 동등하게 다루어지기 때문이다. 그러나 사실은 상위에서의 문헌들의 변화가 더 중요한 의미를 갖는다. 한편 표준정확률은 상위의 문헌들에 더 비중을 둔것으로서 이런 경우에 표준정확률의 값을 보면 순위의 변화를 보다 정확하게 알 수 있다. 표2의 6개의 질문에 대한 평균을 보면 반복검색을 함으로써 효율이 증가되고 있음을 알 수 있다.

일정한 재현율 수준에서 정확률이 어느 정도나 상승되는지를 나타낸 그림7과 표3을 보면 첫번째 반복검색의 결과 정확률의 평균 상승폭이(25.63%) 두번째 반복검색에 의해서 다시 상승되는 폭(13.84%)보다 크다. 이는 적합성 피이드백을 이용해서 질문을 수정해서 반복검색을 하면 첫번째 반복검색에 의한 향상의 폭이 제일 크며(약20%), 점점 효율의 상승폭이 줄어든다는 선행연구의 일반적인 결과와 비슷하다. 그러나, 선행연구들에서는 두번째 반복검색의 상승폭이 무척 많이 줄어서 4~6%인데 반하여 본 실험결과에서는 비교적 높은 13.8%로 나타났다. 따라서 재현율은 낮아도 정확률이 높은 결과를 원하는 경우는 1번의 반복검색이면 어느 정도 바람직하고, 재현율과 정확률이 모두 높은 결과를 얻으려면 2번의 반복검색을하는 것이 좋은 것으로 나타났다.

피이드백 평가결과를 나타낸 표4는 표3과 비교해 볼때, 두번째 반복검색에 의한 각각의 평균 상승비율이 18.78%와 42.28%로서 검색효율의 상승비율이 상당히 낮다. 그 이유는 이미 검색되었던 문헌들 즉, 상위의 문헌순위를 고정시켰으므로 새롭게 검색되는 적합문헌들의 유사계수가 앞서 검색된 적합문헌의 유사계수보다 크더라도 그보다 먼저 검색된 문헌의 상위에 오를수 없기 때문이다. 따라서 표4와 5를 비교해 볼때, 표4에서 재현율이 60% 이상인 경우에 정확률이 크게 상승되는 이유가 설명될 수 있다. 표4에서도 재현율이 높은(70%이상)경우에 정확률의 상승비율이 큰 것을 볼때(27%), 재현율과 정확률이 모두 높은 결과를 원하는 이용자의 경우에는 반복검색이 효과적임을 알 수 있다.

표4를 보면 새로운 적합문헌이 검색됨에 따라서 향상되는 효율만도 평균 19%가 된다는 것을 알 수 있다.

실험문헌집단의 규모가 달라서 직접적인 비교는 될 수 없으나, 이러한 결과는 평균상승율이 10~15%이며, 재현율이 높은 수준에서는 22~29%인 샐톤과 우의 실험결과와 비슷하다.

Ⅳ. 결 론

1960년대에는 실험적으로 이용되던 온라인 정보검색 시스템이 1970년대초부터 본격적으로 운영되면서 이용자와 시스템간에 상호작용이 가능하게 되었다. 이에 따라 이용자의 질문이 해당 시스템내에서 어느 정도의 유용성을 갖고 있는지를 반영해 주는 일련의 검색결과를 이용해서 검색질문을 수정하면 시스템의 성능을 향상시킬수 있다는 관점에서 탐색전략에 대한 연구들이 진행되었다.

본 논문에서는 이러한 탐색전략 중에서 샐톤과 우의 질문수정모형을 통제된 키이워어드 시스템에 적용시켜 보았다. 그 결과 원질문을 수정해서 실시한 첫번째 반복검색에서는 모든 재현율 수준에서 정확률이 평균 25.63%가 향상되었으며, 두번째 반복검색에서는 평균 13.84%가 다시 향상되었다. 또한 새로운 적합문헌이 검색되는 데에 따른 피이드백 평가에서는 두번의 반복검색을 실시한 결과 정확률이 평균 19% 향상되었다.

이러한 실험결과에 따라 다음과 같은 결론을 내릴수 있게 되었다.

1. 디소오러스에 의한 키이워어드 시스템에서 첫째, 문헌과 질문은 벡터의 형태로 표현하고 둘째, 검색의 기준치는 그들 벡터의 내적으로 유사계수를 계산하여 정해 주며 셋째, 탐색용어에는 적합성 피이드백 과정을 이용해서 정확성 가중치를 추정해 주는 방법으로 질문을 수정하여 반복검색을 실시하면, 수정된 질문은 이용자의 정보요구를 가장 잘 표현해 준 최적질문에 가까와지므로 검색효율 측 시스템의 성능이 향상된다.

2. 검색효율이 향상되는 폭은 첫번째 반복검색을 실시한 경우가 가장 크고 차츰 감소하므로, 질문을 수정해서 검색을 반복하는 횟수는 세번 정도가 알맞다.

3. 재현율은 낮아도 정확률이 높은 검색결과를 원하는 이용자의 경우는 한번의 반복검색이면 어느 정도 만족할 것이며, 재현율과 정확률이 모두 높은 결과를 원할 때에는 두번이상의 반복검색이 바람직하다.

參 考 文 獻

1. 랑카스터, F.W.·오웬, J.M. "컴퓨터에 의한 정보검색," 구자영 역. 문헌정보학연구, vol. 1, No.2, (1978), pp. 35-58.
2. 알탄디, S. "情報科學과 콤퓨터," 이순자 역. 문헌정보학연구, vol.1, No.3, (1978), pp.79-96; Vol.1, No.4, pp.74-81.
3. 田中哲. 〔情報檢索論〕 서울 : 아세아문화사, 1977.
4. Angione, Pauline V. "On the Equivalence of Boolean and Weighted Searching Based on the Convertibility of Query Forms," *JASIS*, (March-April 1975), pp. 112-124.
5. Attar, Rony and Fraenkel, Aviezri S. "Experiments in Local Metrical Feedback in Full-Text Retrieval Systems," *Information Processing & Management*, Vol. 17, No. 3, (1981), pp. 115-126.
6. Buell, Duncan A. "A General Model of Query Processing in Information Retrieval Systems," *Information Processing & Management*, Vol. 17, No. 5, (1981), pp. 249-262.
7. Chow, D. and Yu, C.T. "On the Construction of Feedback Queries," *J. of ACM*, Vol. 29, No. 1, (Jan. 1982), pp. 127-151.
8. Cook, Kenneth H. "A Threshold Model of

Relevance Decisions," *Information Processing & Management*, Vol. 15, (1975), pp. 125-135.

9. Dillon, Martin and Desper, James, "The Use of Automatic Relevance Feedback in Boolean Retrieval Systems," *J. of Doc*, Vol. 36, No. 3, (Sept. 1980), pp. 197-208.
10. Ide, E.C. "Relevance Feedback in an Automatic Document Retrieval System." Master's thesis, Cornell University, 1969.
11. The Institution of Electrical Engineers, *INSPEC Thesaurus 1979*. U.K., 1978.
12. Kraft, Donald H. "A Decision Theory View of the Information Retrieval Situation: An Operations Research," *JASIS*, (Sept.-Oct. 1973), pp. 368-376.
13. _____. "A Threshold Rule Applied to the Retrieval Decision Model," *JASIS*, (March 1978), pp. 77-80.
14. Kraft, Donald H. and Bookstein, A. "Evaluation of Information Retrieval Systems: A Decision Theory Approach," *JASIS*, (Jan. 1978), pp. 31-40.
15. Lancaster, F.W. *Information Retrieval Systems: Characteristics, Testing and Evaluation*. 2nd ed. New York: John Wiley & Sons, 1979.
16. Markey, K. 'Levels of Question Formulation in Negotiation of Information Need During the Online Presearch Interview: A proposed Model," *Information Processing & Management*, Vol. 17, No. 5, (1981), pp. 215-225.
17. Mazur, Zygmunt. "Properties of a Model of Information Retrieval System Based on Thesaurus with Weights," *Information Processing & Management*, Vol. 15, (1979), pp. 145-154.
18. Meadow, C.T. *The Analysis of Information Systems*. New York: John Wiley & Sons, 1967.
19. Miller, D. and Dattola, R.T. "Methods for Estimating the Number of Relevant Documents in a Collection,"*Information Processing & Management*, Vol. 18, No. 4, (1982), pp. 179-191.
20. National Computing Center. *NCC Thesaurus of Computing Terms*. 8th ed., U.K., 1976.
21. Oddy, R.N. et al., eds. *Information Retrieval Research*. London: Butterworths, 1981.
22. Robertson, S.E. "The Probability Ranking Principle in IR," *J. of Doc*, Vol. 33, No. 4, (Dec. 1977), pp. 294-304.
23. Robertson, S.E.and Sparck Jones, K. "Relevance Weighting of Search Terms," *JASIS*, (May-June 1976), pp. 129-146.
24. Robertson, S.E. and Belkin, N.J. "Ranking in Principle," *J. of Doc.*, Vol. 34, No. 2, (June 1978), pp. 93-100.
25. Salton, G. *Automatic Information Organization and Retrieval*. New York: McGraw-Hill Book Company, 1968.
26. _____. *Dynamic Information and Library Processing*. Englewood Cliffs, New Jersey: Prentice-Hall, 1975.
27. _____. "Automatic Information Retrieval," *Computer*, (Sept. 1980), pp. 41-56.
28. _____, ed. *The Smart Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.
29. Salton, G., Yang, C.S. and Yu, C.T. "A Theory of Term Importance in Automatic

Text Analysis," *JASIS*, (Jan.-Feb. 1975), pp. 33-44.

30. Salton, G., Wong, A. and Yu, C.T. "Automatic Indexing Using Term Discrimination and Term Precision Measurements," *Information Processing & Management*, Vol. 12, (1976), pp. 43-51.
31. Salton, G. and Waldstein, R.K. "Term Relevance Weights in On-Line Information Retrieval," *Information Processing & Management*, Vol. 14, (1978), pp. 29-35.
32. Salton, G., Wu, H. and Y, C.T. "The Measurement of Term Importance in Automatic Indexing," *JASIS*, (May 1981), pp. 175-186.
33. Sparck Jones, K. "Experiments in Relevance Weighting of Search Term," *Information Processing & Management*, Vol. 15, (1979), pp. 133-144.
34. Taylor, R.S. "The Process of Asking Questions," *American Documentation*, (Oct. 1962), pp. 391-396.
35. _____. "Question-Negotiation and Information Seeking in Libraries," *College & Research Libraries*, (May 1968), pp. 178-194.
36. Van Rijsbergen, C.J. "A Theoretical Basis for the Use of Co-Occurrence Data in Information Retrieval," *J. of Doc*, Vol. 33, No. 2, (June 1977), pp. 106-119.
37. _____ *Information Retrieval*. 2nd ed. London: Butterworths, 1979.
38. Van Rijsbergen, C.J., Harper, D.J. and Porter, M.F. "The Selection of Good Search Terms," *Information Processing & Management*, Vol. 17, (1981), pp. 77-91.
39. Vernimb, C. "Automatic Query Adjustment in Document Retrieval," *Information Processing & Management*, Vol. 13, (1977), pp. 339-353.
40. Wu, H.C.C. "On Query Formulation in Information Retrieval." Ph.D. dissertation, Cornell University, 1981.
41. Wu, H.C.C. and Salton, G. "The Estimation of Term Relevance Weights Using Relevance Weights Using Relevance Feedback," *J. of Doc.* Vol. 37, No. 4, (Dec. 1981), pp. 194-214.
42. Yu, C.T. and Salton, G. "Precision Weighting-An Effective Automatic Indexing Method," *J. of ACM*, Vol.23, No. 1, (Jan. 1976), pp. 76-88.
43. Yu, C.T., Luk, W.S. and Cheung, T.Y. "A Statistical Model for Relevance Feedback in Information Retrieval," *J. of ACM*. Vol.23, No. 2, (April 1976) pp. 273-286.
44. Yu, C.T. and Salton, G. "Effective Information Retrieval Using Term Accuracy," *C. of ACM*, Vol. 20, No. 3, (March 1977), pp. 135-142.
45. Yu, C.T., Lam, K. and Salton, G. "Term Weighting in Information Retrieval Using the Term Precision Model," *J. of ACM*, Vol. 29, No.1, (Jan. 1982), pp. 152-170.