색인언어의 효율성 측정에 관한 연구 (1)

최 수 연

1. 서 론

1.1 연구의 필요성

근래에 이르러서 정보혁명, 정보의 폭발, 정보의 홍수 등의 표현을 비롯하여 정보관리 또는 정보처리라는 용어를 많이 사용하고 있다. 이는 전 세계에서 생산되는 정보의 양이 감당할 수 없을 정도로 늘어나고 있음과 또한 우리가 정보의 중요성을 인식하고 있음을 암시하는 것이다. 현대 정보사회에서는 정보의 양이 폭발적으로 증가되고 있는 동시에 정보의 내용이 더욱 세분화되고 있으며 그 정보가 여러 분야에 복합 이용됨으로써 필요한 정보의 검색이 어렵게 되고 있다.

이렇게 여러 분야에 관계된 문헌의 양이 급증하고 이용자들의 요구가 다양화됨에 따라 사서 및 정보전문가들은 그 많은 문헌들을 다양한 요구에 맞추어 수집하고 분석 및 조직을 통해 무질서한 정보들을 어떤 입장에 따라 질서를 부여하여 축적해 놓게 된다. 그런 다음, 이용자의 요구가 있을 때 이용자를 대신하여(또는 이용자 자신이)그 축적매체에서 요구하는 정보에 적합한 것을 찾아내게 되는데 그 일련의 과정을 정보의 축적과 검색(information storage and retrieval)이라고 하며 흔히 정보검색(information retrieval ; IR)이라고 쓰여진다.

정보검색이라는 용어는 1950년 무어즈(C. W. Mooers)씨가 처음으로 사용하였으나 일반화된 것은 1954년 클리버든(C. W. Cleverdon) 씨와 돈(R. G. Thorne)[1]씨가 그들의 보고서에서 이 말을 사용한 후 부터라 하겠다.

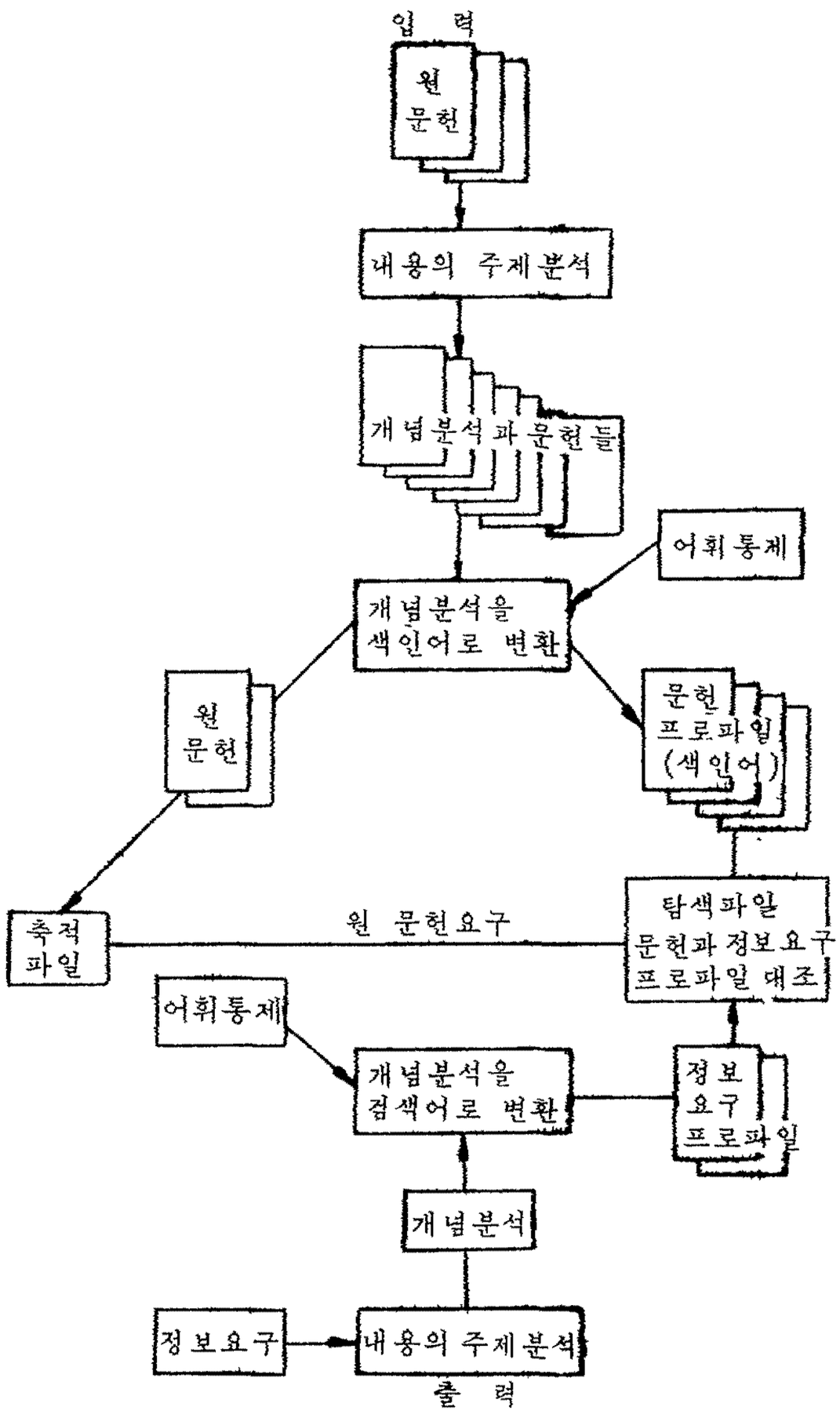
정보검색 활동은 크게 축적과정과 검색과정으로 구분된다. 랑카스터(Lancaster)씨가 제시한 플로우 차트(flow Chart)를 통해 정보검색 내용을 살펴보면 표 1[2]과 같다.

즉, 축적과정에서 새로운 정보를 입수하였을 때 시스팀내에 축적할 가치가 있는 것만을 선택하여 그 내용을 일정한 방법에 따라 분석·가공해야 한다. 그러기 위해서는 각 정보의 내용을 분석해서 그 구성요소중 중요한 개념들을 추출한 다음, 키워드(keywords)나 주제명표목(subject headings) 또는 분류기호와 같은 색인어(index terms)[3]로 변환 표시한다. 색인어로 변환 표시한 정보의 내용을 일정한 배열규칙에 따라 정리하여 문헌 프로파일(document profiles)을 구성하게 되고 원정보는 서가나 서고와 같은 축적매체에 축적하여 축적파일(storage file)을 구성한다.

이 후에 검색과정에서는 이용자가 요구한 정보의 내용을 분석하여 원 정보의 문헌 프로파일에 주어진 색인어와 동일한 형태의 검색어(retrieval terms)로 변환시킨 다음, 이 정보요구 프로파일(request profiles)과 문헌 프로파일을 대조하여 정보요구와 관련된 문헌목록, 또는 초록을 제공하게 되며 이러한 목록들을 검토한 이용자가 자신이 필요한 원 문헌(original document)을 요구하면 축적파일에서 탐색하여 제공하는 것이다.

이러한 일련의 과정을 통해서 볼 때, 정보검색은 정보검색 시스팀을 통하여 행해지며, 정보검색 시스팀을 구성하는 요소중 색인어는 시스

표 1. 정보 검색 내용



팀의 매체로서 시스팀의 기본단위가 된다. 따라서, 사용하는 색인어군(群), 즉 색인언어의 질은 정보검색 시스팀의 성능(performance)을 지배하는 요소중 가장 중요한 단일요소가 된다. 다시 말해서 색인언어의 효율성은 그 다소(多少)에 따라서 정보검색 시스팀의 성능이 결정된다. 이에 따라 다양한 색인언어들에 대한 연구가 요청되고 있고 색인어의 효율성을 증대시킴으로써 시스팀의 성능을 높여 효과적인 정보를 제공할 수 있다.

1.2 연구의 목적 및 방법

이용자의 특정한 정보요구(information request)를 충족시키는 문헌들을 검색하기 위해 사용되는 도구를 총괄해서 색인이라 하는데 이 색인을 구성하는 색인언어는 검색의 대상이 되는 문헌이 주제를 표현하기 위하여 선정한 기호군으로서 여기에는 용어 뿐만 아니라 숫자, 부호 등 모든 기호가 포함된다.[4] 앞에서 언급한 바와 같이, 시스팀의 매체이며 단일요소인 색인언어는 정보검색활동 과정에서 그 효율성에 따라 시스팀의 성능을 좌우한다. 따라서 도서관 및 정보센터를 비롯한 정보검색 시스팀에서는 어떠한 색인시스팀을 채택하느냐에 따라 당 정보검색 시스팀의 성능이 크게 좌우된다. 현재 우리나라의 특수 정보관리센터에서 다루고 있는 단행본, 잡지, 논문, 기타 문헌들을 보면 많은 부분이 영어로 된 자료로 구성되어 있다. 또한 그러한 기관을 이용하는 이용자들은 어느 정도 수준의 영어 교육을 받아 영어를 해득할 수 있는 언어적 배경을 가진 사람들로써 국제경제연구원, KIST 등 전산화된 시스팀에서도 영어로 색인 표목을 표기해 주고 있다.

따라서 본 연구의 목적은 색인시스팀의 효율성과 관계된 사항들을 고찰함과 동시에 실제로 상이한 두 주제분야(교육심리학과 생화학)의 문헌들을 실험용 데이터로 삼아 영어로 표기한 색인언어(주제명표목과 문헌의 표제에서 추출한 키워드)의 효율성을 측정 비교하여 색인어의 검색효율을 향상시킬 수 있는 방법을 도모하여 특수주제분야의 정보관리업무에 적합한 정보검색 시스팀을 설계하는데 도움을 주고자 한다.

정보검색 시스팀의 성능은 검색에 의하여 얻어진 정보가 이용자의 요구에 일치하는 율을 말하는 검색효율, 시스팀 설계 및 운영에 요하는 경비 즉 경제성 그리고 이용자가 시스팀에 대하여 질문한 시점에서 답을 얻기까지의 시간, 즉 신속성 등을 고려하여 평가할 수 있으나 본 연구는 시스팀 성능 평가방법중 가장 핵심이 되는 검색효율 측정에 국한시키기로 한다.

2. 선행연구 및 색인언어의 종류

2.1 선행연구

1953년 토브(Mortimer Taube)씨가 설립한 정

보학회(Documentation Inc)에서는 조합색인법(Coordinate indexing)에 의한 유니텀 시스팀(Uniterm System)과 미군사기술정보국(Armed Services Technical Information Agency: ASTIA)이 개발한 주제명표목에 의한 재래식 목록의 성능을 비교 연구[5]한 것이 있으나, 실제 종합적인 색인언어에 대한 평가시험은 1957년부터 클리버든씨의 지휘하에 시행된 애슬립 크랜필드 프로젝트(Aslib Cranfield Project)로서 그후에 진척상황이 여러 차례 보고[6]되거나 평가[7] 되었다. 애슬립 크랜필드 프로젝트는 크랜필드 프로젝트 I과 크랜필드 프로젝트 II로 나뉘는데 제 1단계 기간중인 1961년 클리버든(Cleverdon)씨는 국립과학재단(National Science Foundation)의 요청을 받아 4가지 색인언어(UDC, 파셋분류, 자모순 주제목록, 유니텀 시스팀)의 성능을 비교했다. 이 실험은 1,200건의 탐색질문에 대한 18,000종의 금속공학 관계문헌을 기초로 했는 데 이 실험을 계기로 재현율(recall)과 정확율(precision ratio)의 개념이 보급되었고, 실패원인을 상세히 규명한 것은 높이 평가할 만하다. 크랜필드 프로젝트 II에서는 색인언어의 디바이스(devices)들을 각각 독자적으로 또는 상호결합하여 시행했다.[8]

이렇게 1960년대 초부터 시작한 평가시험들은 규모가 크고 작은 시험들로 그 수가 막대하다. 과거의 시험들을 보면 MEDLARS같은 대규모 현장 시스팀을 대상으로 하는 시험과 규모가 작은 연구시험들로 나누어지며 대상문헌들의 주제분야는 항공학, 의학, 핵물리학, 화학, 생물학 등 순수과학 및 응용과학이 대부분이었다. 다만 ISILT 컬렉션은 비교적 "soft subject"인 정보과학 분야를 다루었으며[9] 인문과학이나 사회과학적인 주제를 대상으로 하거나 또는 주제가 상이한 문헌들을 대상으로 한 비교연구는 없었다. 연구대상이 된 색인언어는 비통제언어(uncontrolled language 또는 free language)와 통제언어(controlled language)로 그 유형을 대별할 수 있는 데 대부분이 다양한 통제언어들이었고 비통제언어를 시험한 연구는 그다지 많은 편이 아니다.

2.2 색인언어 종류

색인언어란 색인 및 검색의 대상이 되는 문헌들의 주제를 표현해 주기 위해 선정한 용어나 숫자, 기호 등의 총체를 말한다.

색인언어는 과거로부터 사용되고 있는 저자명, 주제명, 계층분류 및 파셋분류와 근래에 이르러서는 조합이론을 도입한 후조합색인법에서 사용하는 키워드를 들 수 있다. 본고에서는 연구와 관련된 주제명시스팀과 키워드시스팀에 관해서만 간단히 설명하기로 한다.

2.2.1 주제명 시스팀

주제명 시스팀의 대표적 예로는 자모순 주제목록을 들 수 있다. 주제명 시스팀은 보통 주제명표목이라고 불리는 용어를 사용하여 문헌의 주제를 색인해 주는 시스팀이다.

주제명표목은 사전에 준비된 주제명표목표에서 선정되거나 색인자에 의하여 문헌의 주제분석단계에서 작성된다. 그것은 어느 경우나 문헌내용의 주요 주제만을 표시하며 보통 2, 3개의 표목이 선정 또는 작성된다. 이러한 주제명표목은 이미 조합된 상태의 것을 사용하기 때문에 전조합색인언어(precoordinated index language)[10] 라고도 부른다. 주제명표목들 사이의 개념상 관계는 보라(see) 또는 또보라(see also) 등의 참조표시가 사용된다.

2.2.2 키워드 시스팀

키워드 시스팀은 색인대상이 되는 문헌속에 직접 또는 간접으로 포함되어 있는 이른 바 키워드들을 뽑아 모은 것으로 키워드로 사용되는 단어의 품사는 대부분이 명사, 형용사, 동사, 부사, 숫자 등이 쓰이나 전치사, 접속사, 분사 및 기타의 기능어들은 보통 사용하지 않는다. 키워드 시스팀에서는 문헌속의 자연언어에서 그대로 추출하여 전혀 통제하지 않은(uncontrolled) 키워드와 동사를 명사화한다든지, 동의어나 관련어, 용어의 상하위 개념을 통제하여(controlled) 사용하는 키워드로 나눌 수 있다. 그러한 통제

된 키워드들로 이루어진 통제어휘목록을 디소오러스(thesaurus)라고 부른다.

디소오러스를 다시 요약하면, 조합색인법[11]을 사용하는 정보검색 시스팀에서 색인어를 조절하기 위해 사용하는 어휘집으로서 각 키워드간의 개념상의 관계를 표시해 준다. 즉 상위개념어, 하위개념어, 관련어 등의 관계표시 및 선택되지 않은 단어와 선택된 단어사이의 연결을 위해 흔히 BT(broader term), NT(narrower term), RT(related term)들의 관계표시와 Use, UF (Used For) 등의 참조표시가 사용된다. 색인해 준 키워드들은 검색시 조합하여 검색하기 때문에 후조합 색인언어(postcoordinated index language)라고도 부른다.

3. 검색효율 측정방법과 효율지배요인

3.1 효율 측정방법

이상적인 검색 시스팀은 이용자가 원하는 모든 문헌과 또한 꼭 필요로 하는 문헌만을 검색할 수 있어야겠지만 실제로 100%의 정확성을 기대하기란 어렵다. 정확성이란 개개인 이용자의 상대적인 만족도에 의해서 측정되기 때문이다.

효율측정 방법은 언제나 논쟁의 대상이 되어 왔는데 정확율(precision)과 재현율(recall ratio)은 가장 널리 사용되어 온 측정수단으로서 1956년에 이미 페리(Perry)씨와 켄트(Kent)씨에 의하여 제안[12]된 것인데, 1962년 클리버든씨가 크랜필드 실험결과에 의해 이 명칭으로 부른 이래로 널리 쓰이기 시작했고, 보통 크랜필드 측정법(cranfield measures)이라고 알려져 있다.

재현율과 정확율은 다음과 같은 검색결과에 따라 정의된다. 즉, 표 2와 같다.

표 2. 정보요구와 문헌과의 관계

	검색된 문헌	검색되지 않은 축적파일내의 문헌
정보 요구에 적합한 문헌	A	B
정보요구에 부적합한 문헌	C	D

표 2에서 정확율은 $\frac{A}{A+C}\times 100$으로 표시하고 재현율은 $\frac{A}{A+B}\times 100$으로 표시된다[13].

결국 정확율은 검색된 정보속에 불필요한 것이 얼마나 많이 들어 있는지를 표시하는 것이고 재현율은 검색된 정보속에 필요한 것이 얼마나 많이 들어 있는지를 표시한 것이다. 여기서 「B」는 축적파일 속에 있으나 불완전한 검색으로 누락된 정보이고, 「C」는 부정확한 검색으로 검색되었으나 정보요구와는 관련없는 불필요한 정보이다. 검색시스팀에서 원 축적파일을 샅샅히 훑어 관련문헌을 모두 찾아내어 재현율 100%를 달성한다거나 검색된 문헌이 모두 정보요구를 만족시켜 100%의 정확율을 달성하기란 어렵다.

보통 더 많은 문헌을 검색해 내면 그만큼 더 많은 관련문헌을 얻을 수 있으나, 또한 비관련 문헌도 많이 따라 나오는 문제가 있고 반대로 적은 수의 문헌이 검색되면 그만큼 관련문헌이 적게 얻어지나 관련없는 문헌은 거의 검색되지 않는 이점이 있다.

이러한 정확율과 재현율 사이의 대체적인 반비례법칙(inverse law)은 이른바 크랜필드 프로젝트에서 클리버든씨에 의해 고찰되었다[14]. 그러나 이러한 반비례 관계가 반드시 존재하는 것은 아니며, 실제 시스팀에 적절한 변화를 주어 재현율과 정확율 모두를 높이는 것이 가능하다.

3.2 효율 지배요인

검색효율을 지배하는 요인들은 시스팀을 구성하는 여러 변수들(variables)들에 의해 영향을 받는다. 그러한 요인들은 색인작성과정, 검색작업과정 그리고 기타 요인으로 나누어 살펴보면 다음과 같다.

3.2.1 색인작업시

1) 색인작업의 망라성

망라성이란 내용파악의 척도로서 특정문헌 속에서 취급하고 있는 주제 개념들을 모두 색인작성시 색인언어로 변환시키는 것을 말하는데[15], 일반적으로 문헌당 배정된 색인어의 수[16]로서 나타낸다. 가령 4개의 주제개념(A, 4B, C, D)을

다루고 있는 문헌이 있다고 가정하자. 만일 색인작성 과정에서 주제 분석시 이 4개의 주제를 모두 파악해서 적당한 색인어로 변환해 주었다면 이 문헌은 완벽하게 망라적으로 색인했다고 할 수 있다. 따라서 이 4개중 어떤 한가지 주제로 질문한다 해도 검색해 낼 수 있다는 것은 자명하다. 그러므로, 망라성의 수준이 높으면 약간이라도 관련성있는 문헌은 모두 검색되어 재현율이 올라가는 반면 정확율은 낮아진다.

정확율이 낮아지는 이유는 두가지로 볼 수 있다. 첫째 검색문헌 가운데에는 정보요구에 극히 적은 부분만 관련된 문헌이 포함될 우려가 있고, 둘째로 후조합색인의 경우, 조합시 착오를 일으켜 전혀 관련없는 문헌이 검색될 수 있다. 그 예로 앞에서 보기를 든 문헌이 A와 B, C와 D가 서로 관련지어져 있다고 가정하면 A와 C, A와 D, B와 C, B와 D 식으로 되는 질문에 대해서 이 문헌이 검색되어 실제 정보요구와는 어긋나게 된다.

이런 관점에서 높은 망라성은 높은 재현율과 낮은 정확율을 초래하고 반대로 낮은 망라성은 낮은 재현율대신 높은 정확율을 초래한다. 따라서 망라성 수준의 결정에 따라 검색효율이 결정된다.

2) 색인언어의 특정성

시스템의 재현성이 색인언어의 고유한 특성에서 보다는 색인작업의 망라성과 관련되어 결정되는 반면에 시스템의 정확성은 전적으로 주제를 색인언어가 얼마나 명확히 표현했느냐에 따라 결정된다[17].

예를 들어 A라는 시스템은 2,000개의 통제어휘를 갖고 있고, 시스템 B는 1,000개, 시스템 C는 500개를 갖고 있다고 가정하자. 이 세 시스팀이 모두 "pulsejet engines"란 주제를 색인코자 한다면 어휘수의 풍부성에 따라 시스템 A는 "pulsejet engines"로, 시스템 B는 좀 더 일반적인 주제인 "jet engines"로, 시스템 C는 그저 "engines"로 특정지을 수 있다.

시스템 A인 경우, "pulsejet engines"에 관해 탐색을 하면 대부분 적합한 문헌을 검색해 내어 높은 정확율을 기대할 수 있으나 잠재적으로 그에 관련된 문헌이나 "pulsejet engines"의 다른 동의어로 색인해 준 문헌이 누락되어 재현율이 낮아진다. 같은 질문을 시스템 B에 주었을 때 재현율이 약간 개선될 수 있으나 여기서도 좀더 일반적인 색인어인 "engines"로 색인된 문헌이나 "jet engines"와 동의어로 색인된 문헌이 누락될 수 있다. 이렇게 부차적으로 관련있는 문헌은 "engines"라는 일반적인 용어아래서 찾을 수가 있다. 이렇게 시스템 A에서 B, C로 가는 동안 문헌을 표현하는 특정성이 줄어들수록 검색되는 문헌의 수는 증가하게 된다. 다시 말해, 색인언어의 특정성이 클수록 탐색시 높은 정확율을 허락하나 재현율은 떨어진다.

3) 통제여부

문헌의 표제나 본문의 자연어에서 표목을 선정하는 시스템에서는 주제의 표현이 애매하고 관련관계가 모호하거나 표목이 불필요하게 길어지게 되는 등 여러가지 문제점이 야기되고 있어서 표목을 적절히 조절하여 사용해야 할 필요성이 생기고 있다. 즉, 색인언어의 재현율과 정확율을 향상시키기 위한 부가적인 방법(devices)을 도입시키는 것이다.

색인언어로서 가장 기본적인 것을 문헌에서 추출한 일군의 키워드로 보고 여기에 재현성과 정확성을 개선시키기 위한 수단을 덧붙일 수 있다. 표 3[18]은 가장 일반적으로 사용되는 재현율과 정확율을 높여주는 수단이다.

A. 재현율 향상수단(recall devices)

재현능력을 높이기 위해서 동의어를 통제할 수 있다. 예를 들어, "콤퓨터", "컴퓨터", "전자 계산기"와 "전산기"라는 용어들이 어떤 특정 색인시스팀에서 동의어로 결정되었다고 생각해 보자. 이는 네 용어로 쓰인 문헌간의 개별성을 인정하지 않기로 결정했다는 것을 의미한다. 이 경우, 우리는 네가지 문헌 클래스를 표기해 주고 서로 참조가 되는 표시를 해준다. 즉, "전자 계산기는 콤퓨터를 보시오"와 같은 「보라」참조를 "컴퓨터," "전산기"에도 해 줌으로써 색인어인 "콤퓨터"라는 용어의 범위를 넓혀주고 여러 클래스를 하나로 합쳐 색인어휘의 수를 줄이는 것이다.

표 3. 재현율, 정확율 향상 수단

재현율향상수단 (recall devices)	정확율향상수단 (precision devices)
문헌클래스의 규모를 넓히고 용어정의를 넓혀 어휘수를 줄인다.	문헌클래스의 규모를 좁히고 용어의 정의를 한정시켜서 어휘수를 늘린다.
동의어통제 (synonym control)	
유사동의어 조절 (quasi-synonym control)	조합 (coordination)
계층적결합 (hierachical linkage)	가중치 (weighting)
클럼핑과 클러스터링 (clumping & clustering)	링크 (links)
단어형태조절 (word form control)	로울 (roles)
단순한 키워드들 (통제하지 않음)	

다음으로 어떤 주제분야에서 일반적으로 동의어로 간주되지는 않지만 동의어처럼 또는 실제로는 동의어로 쓰인 것들이 있다. 예를 들어, 항공학 분야에서 "supersonic speed," "supersonic flow", 그리고 "supersonic flight"는 보통 마하 1 이상 속도의 움직임을 나타내는데 쓰인다. 여기서 "speed," "flight," "flow"는 유사동의어라 할 수 있다. 유사동의어들을 하나로 모아 주는 것도 동의어 조절과 같은 효과를 얻을 수 있다.

또한, 재현율을 높여주기 위해서 단어형태를 조절할 수 있다. 이것은 같은 어원을 가진 용어(예 : coat, coated, coating)를 모아주는 것이다. 다시 말해, "coats"에 관한 모든 포괄적인 문헌들을 근원어인 "coat"로 표시해 주는 것이다.

매우 효능이 있는 재현율 향상수단중의 하나로 계층적 연결이 있다. 이것은 분류법의 체계에서도 볼 수 있듯이 개념적인 상하관계를 정하는 것이다. 이 계층적 연결은 반대로 일반적 용어에서 특정용어로 옮겨가면서 정확율을 높여주어 정확율 향상수단이 될 수도 있다[19]

재현율 개선을 위한 방법 중 또 다른 하나는 색인어 사이의 통계적인 관련도를 근거로 해서 색인어들을 짝지어 주는 것이다. 즉, 클럼핑 또는 클러스터링이라고 부른다. 예를 들어 문헌에 대한 색인 프로파일속에서 A라는 용어가 자주 나타날 때 용어 B, H, I가 함께 나타나고, L이라는 용어도 B, H, I와 상호적으로 강하게 나타난다면, 용어 A와 L사이의 다소의 관계를 예측할 수 있고 두 용어는 탐색시에 대치될 수 있을 것이다.

지금까지 살펴본 재현율 향상수단은 대개 한두가지를 적용할 수 있다. 색인언어 구성부분에서 클래스를 넓혀주거나 탐색과정에서 그렇게 할 수가 있다.

B. 정확율 향상수단 (precision devices)

정확율 향상수단 중에서 가장 효능이 있는 것은 조합이랄 수 있다. 조합은 문헌류를 좁혀서 색인어의 어휘 수를 증가시킨다. B와 꼭 함께 나타나는 A를 얻기 위해서 AB라는 클래스를 요구하는 것이다. 이런 클래스의 조합은 정확성을 개선시키기도 하지만 어느 정도 재현성에도 영향을 미친다.

둘째로, 용어에 가중치를 주는것 (term weighting)이 있다. 만일 A, B, C로 색인된 문헌이 있다고 하자. A는 가장 중요한 개념을 나타내는 용어이고 B는 부차적으로 중요한 개념을 나타내고, C는 그저 주변적인 것을 나타낸다고 할 때 우리는 용어의 중요도에 비례해서 A에 3

이라는 가중치를, B에는 2를, C에는 1을 줄 수 있다. 그런 다음 검색할 때는 A가 3을 가진 ABC로 된 문헌을 요구하거나 ABC의 가중치가 합해서 4이상인 문헌을 요구할 수가 있다. 이렇게 재래식 색인작성에서는 어떤 한 색인어가 특정문헌에 적합한지 않은지의 결정 뿐이지만 색인어에 가중치를 줌으로써 좀더 유동성있게 색인하여 주고 또한 검색함으로써 정확율을 높여 줄 수 있다.

대부분의 문헌들은 복합주제를 다루고 있다고 볼 수 있다. 한 문헌속에서도 여러 주제를 나타내는 용어가 있어서 후조합색인방법을 채택한 시스팀에서는 그러한 용어들이 잘못 조합되어 정확율이 떨어지게 된다. 이런 점을 개선하기 위한 정확율 향상수단으로 링크와 로울이 있다.

링크란 문헌의 구분상 관계있는 것끼리 연결지우는 것이다[20]. 예를 들어, "황산제조와 촉매정화" 대신 "황산정화와 촉매제조"라는 본래 내용과는 전혀 다른 문헌이 검색되는 결점을 해결하기 위해서 연관된 그룹마다 연결부호를 부기하여 각 그룹을 구별한다.

즉, 전자의 문헌번호가 100이면 "황산"과 "제조"는 100A, "촉매", "정화"는 100B라 준다. 만일 "염산제조와 촉매정화"라는 문헌이 또 있다고 가정하자. 이 문헌번호가 200이라면 "염산" "제조"는 200A, "촉매", "정화"는 200B라고 준다. 이렇게 해서 A, B가 링크되어 검색시 번호를 조합할 때 문헌번호가 100과 200이었더라도 실제 100A, 100B, 그리고 200A, 200B로 내용이 구분되어 검색의 착오를 막을 수 있다. 링크는 A, B와 같은 문자로 쓰는 예도 있고 1, 2 또는 01, 02, 03과 같이 숫자로 표시하기도 한다[21].

링크를 사용해도 해결되지 않는 문제가 있다. 즉, 후조합 검색시스팀에서 "toxins produced by fish"라는 주제에 관한 탐색을 생각해 보자. 만일 우리가 "toxins"라는 부류와 "fish"라는 부류를 논리적으로 조합시켜 문헌을 탐색한다면 실제 "toxins affecting fish"라는 문헌이 검색될 우려가 있다.

이렇게 검색되어 나온 문헌은 검색시 조합된 용어들 아래에 색인이 되어 있지만 우리가 원하는 문헌은 아니다. 두 문헌 모두 "toxins"와 "fish"라는 색인어는 동일하지만 각 문헌에서의 기능(역할 : role)은 다른 것이다.

이와 같이 언어의 구문적, 어순적인 문제에서 야기되는 혼잡때문에도 검색에 한계가 생긴다. 이러한 것을 해결하기 위한 것이 로울 또는 로울 인디케이터(role indicator)라고도 불리는 것으로 하나의 링크내에서 개개의 개념의 기능과 어순(語順)을 표시하는 것이다[22]. 상술한 예들에 다음과 같이 로울을 준다.

toxins produced by fish	toxins affecting fish
Toxins (1)	Toxins (7)
Fish (5)	Fish (9)

여기서 (1)은 생산물(product)을 나타내는 로울인디케이터이고 (5)는 생산자(producer), (7)은 영향을 주는 것(agent), (9)는 영향을 받은 것(patient)을 나타내는 로울인디케이터이다. 로울의 사용은 정확율을 높이기 위한 것이나, 잘못 사용하면 재현율을 저하시킬 우려가 있다고 몽타그(Montague)씨[23]는 지적하고 있다.

參考文獻 및 註

1) C. W. Cleverdon and R. G. Thorne, "A Brief Experiment with the Uniterm System of Coordinate Indexing for the Cataloging of Structural Data," R. A. E. Library Memo 7 (1954).
2) F. W. Lancaster, Information Retrieval Systems : Characteristics, Testing and Evaluation, CN. Y.: John Wiley a Sons, Inc., (1968), p. 4.
3) 정보검색시스팀에서 정보의 내용을 표현해 주는 용어를 말하는데 색인과정에서는 색인어, 검색과정에서는 검색어로 쓰이기도 하나 내용은 동일하다. 또한 색인어의 총체를 색인언어(Index Language)라고 부른다. 색인언어에는 용어뿐 아니라 숫자, 기호도 포함된다.
4) 司空 哲, 情報檢索論(서울 : 亞細亞文化社, 1977), 121面.
5) C. W. Cleverdon, "Progress in Documentation; Evaluation Test of Information Retrieval Systems," Journal of Documentation, Vol. 25, No. 1 (1970), p. 55.
6) C. W. Cleverdon and J. Mills, "The Testing of Index Language Devices," Aslib Proceedings, Vol. 15, No. 4 (1963), pp. 106–130.

7) 1단계 Aslib—Cranfield Project는 O'Connor에 의해 Journal of Documentation, Vol. 17 (Dec. 1961), p. 261과 Phyllis Richmond에 의해 American Documentation, Vol. 14 (Oct., 1963), pp. 307—311에서 평가되었다.
8) C. W. Cleverdon and others, Factors Determining the Performance of Indexing System, (2 Vols: Cranfield, 1966) Qusted in Cyril Cleverdon, "Progress in Documentation: Evaluation Tests of Information Retrieval Systems," Journal of Documentation, Vol. 26, No. 1 (1970), p. 58.
9) E. M. Keen and J. A. Digger, Report of an Information Science Index Language Test, (2 Vols; Wales, Aberystwyth: College of Librarianship, 1972).
10) A. C. Foskett, The Subject Approach to Information, London: Bingley, 1971, pp. 49—50.
11) Mortimer Taubè and H. Wooster, Information Storage and Retrieval : Theory, Systems and Devices (New York: Columbia Univ. Press, 1958), p. 8.
12) J. W. Perry, et al, Machine Literature Searching (New York: Interscience, 1956), p. 43.
13) C. W. Cleverdon, "The Cranfield Tests on Index Language, Devices," Aslib Proceedings, Vol. 19, No. 6 (1967), p. 610.
14) J. Farradane, "The Evaluation of Information Retrieval Systems," Journal of Documentation, Vol. 30, No. 2 (1974), p. 200.
15) F. W. Lancaster, Information Retrieval Systems: Characteristics, Testing and Evaluation (New York: John Wiley, 1968), p. 66.
16) K. Spark Jones, "Does Indexing Exhaustivity Matter?" Journal of the American Society for Information Science, Vol. 24, No. 5 (1973), p. 313.
17) F. W. Lancaster, Op. Cit., p. 68.
18) Ibid., p. 85.
19) B. C. Vickery, "Structure and Function in Retrieval Languages," Journal of Documentation, Vol. 27, No. 2 (1971), p. 70.
20) Mortimer Taube, "Notes on the Use of Roles and Links in Coordinate Indexing," American Documentation, Vol. 12, No. 2 (1961), p. 98.
21) 司空 哲, "整合索引法에 있어서 Uniterm System에 관한 硏究" 圖協月報, 7卷, 6號 (1966), 247面.
22) 上揭書.
23) B. A. Montague, "Testing, Comparison, and Evaluation of Recall Relevance and Cost of Coordinate Indexing with Links and Roles," American Documentation, Vol. 16, No. 3 (1965), p. 367.