색인언어의 효율성 측정에 관한 연구 (2)

최 수 연

4) 입력자료 (input sources)

색인작성시 입력자료로는 표제 (title), 초록 (abstract), 전체문헌 (full text) 등이 있다. 사라세빅 (Saracevic)씨는 비교시스팀연구소 (Comparative Systems Laboratory ; CSL)에서 이에 관해 실험한 것을 다음과 같이 평가했다.[24]

CSL에서는 표제, 초록, 전체문헌 사이의 문헌당 색인어 수효의 비율을 1 : 4 : 6으로 주어 실험한 결과, 일반적으로 색인의 길이가점차 늘어나면 답변의 전체 수도 점차 늘어나는 경향이 있었다. 즉, 색인어의 수가 늘어나면 검색되는 문헌의 수 또한 점차 증가된다는 것을 의미한다. 이 실험에서 측정방법으로 사용한 재현성은 표제→초록→전체문헌 순으로 효율이 증가했다. 동시에 같은 파일에서의 특수성 (파일내의 비관련 문헌 총수에 대한 검색안된 비관련문헌 총수의 비율)은 점차 줄어드는데, 특히 광범위한 탐색을 할 경우에는 현저히 감소된다. 그래서 개개 파일에 대한 효율성은 거의 비슷하나 전체문헌으로 색인해 준 파일에서는 매우 낮게 된다. 따라서 대체로 초록을 입력자료로 하는 색인파일이 가장 효과적이라는 평가를 할 수 있다. 하지만 물론 어떤 경우든 입력자료의 다양성에 따라 다양한 결과가 나온다는 것을 알 수 있다.

5) 색인작성자

시스팀의 성능에 인간적인 요인이 미치는 영향이 막대하다는 것은 많은 실험에서 밝혀지고 있다. 사라세빅 (Saracevic)씨의 평가에 의하면 검색효율이 떨어진 원인의 82%가 색인작성자의 실수임을 주장했다.[27]

색인작성자가 잘못하는 경우는 두가지를 들 수 있다. 하나는 실수로 중요용어나 적절한 용어를 누락시킨 경우이고 하나는 색인작성자의 의사결정에 관련된 사항으로 문헌주제에 대하여 부적당한 용어를 사용한 경우이다. 용어의 누락은 일반적으로 재현율을 감소시킨다. 부적당한 용어의 사용은 그 부적당한 용어로 검색했을 때, 관련문헌이 적어 정확율을 감소시키고 또한 탐색자는 적절한 용어로 검색한다고 해도 관련문헌들이 부적당한 용어밑에 들어가 있어 재현율이 떨어지게 된다. 이는 색인작성자의 특정주제 분야에 대한 배경도 관련이 된다.

3.2.2 검색작업시

1) 탐색전략

색인작업의 망라성과 색인언어의 특정성은 재현성과 정확성에 영향을 준다. 그러나, 탐색자가 탐색전략을 적용시킴으로써 재현성과 정확성을 변화시킬 수 있다. 즉, 주제를 일반적으로 잡거나 자세히 할 수 있고 동의어를 대치시키기도 하여 탐색방법을 다양하게 함으로써 (시스팀의 상황에 따라 다량 또는 소량의 문헌을 검색하기 위해서) 여러 수준으로 탐색을 수행할 수 있다. 어떠한 탐색에서도 필요한 수준의 성능을 달성하기 위해서 정확성을 희생시켜 높은 재현성을 바랄 수도 있고, 반대로 재현성을 줄여서 높은 정확성을 얻을 수 있거나 혹은 양자를 절충할 수도 있다. 이처럼 색인작업에서와 같이 망라성과 특정성이라는 용어를 탐색작업에도 적용시킬 수 있다.[26] 예를 들어, "shielded arc welding of chromium-nickel steels"라는 주제를 다루는 문헌을 찾는다고 가정하자. 이 질문에는 2가지

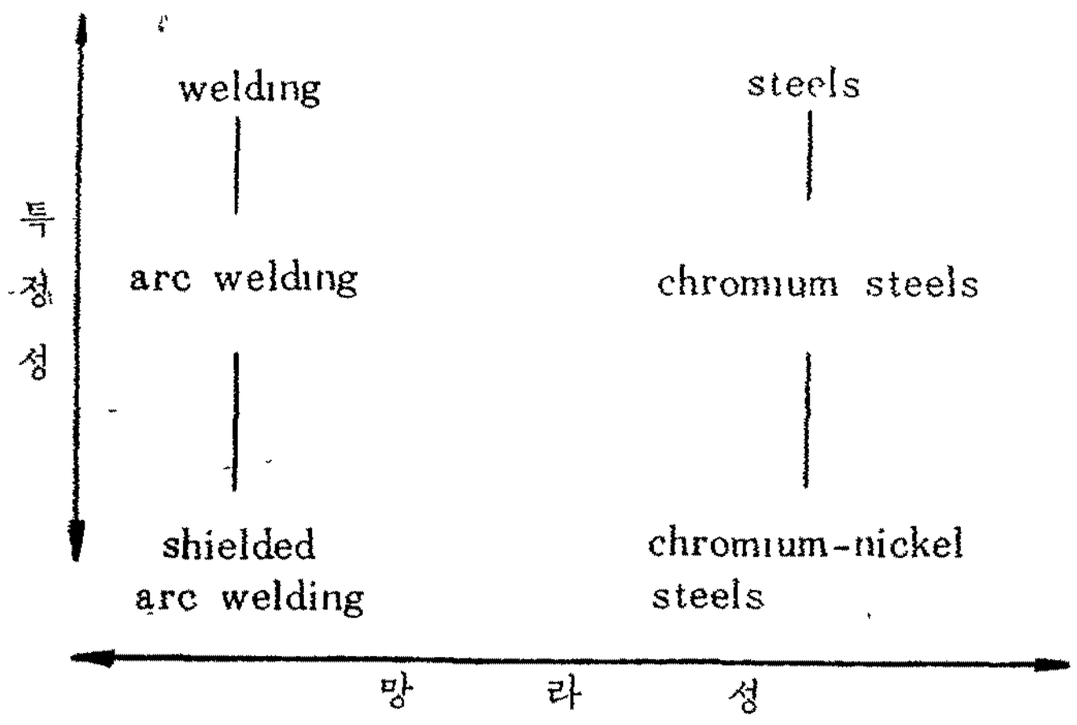
뚜렷한 범주를 포함하고 있음을 알 수 있다. 즉, 제작에 관한 사항(shielded arc welding)과 재료에 관한 사항(chromium—nickel steels)이다. 이 두가지 사항 모두를 기술해 준다면 질문에 대해서는 완전히 망라적인 표현을 해 주는 것이다. 게다가 각 사항들이 질문자가 요구한 가장 정확한 수준(세분해서)으로 표현해 주면 이 질문에 대한 기술은 완전히 망라적인 동시에 완전히 특정하다. 이렇게 동시에 망라적이고 특정한 수준으로 탐색할 경우 100%의 정확률을 얻을 수 있을지 모르나 반면에 질문과 약간 관련된 문헌이 누락되어 재현율이 떨어진다. 재현율을 개선시키기 위해서는 특정성을 좁히거나 혹은 망라성을 줄이거나 둘 다를 줄일 수 있다.

표 4에서 보듯이, "arc welding", "chromium steels" 아래에 있는 모든 문헌을 검색하는 것이다. 좀더 특정성을 줄이기 위해서는 shielded arc welding→arc welding→welding 같은 식으로 각 사항의 특정성을 줄인다.

반대로, 망라성을 줄이는 경우는 1개의 범주를 제외시키는 것이다. 즉, "shielded arc welding" 하나로 탐색하는 것인데 이 경우도 특정성은 완전하게 유지하고 있다.

다음으로 망라성과 특정성을 동시에 감소시키는 것이다(예: arc welding과 chromium-nickel에 관련된 것만 검색한다). 이 경우는 광범위하게 관련문헌이 검색되어 재현율이 개선되는 동시에 비관련문헌 또한 많이 검색되어 정확률이 떨어진다. 따라서 탐색작업에서 특정성과 망라성을 조절함으로써 재현성과 정확성에 대해 우리가 원하는 수준으로 탐색을 통계할 수 있다.

표 4. 탐색전략 수준



2) 탐색자

수동식 시스팀에서 탐색전략을 짜는 책임을 맡은 중개자들은 이용자의 정보요구를 파악하는데 명석해야 한다. 때로 질문분석을 잘못하여 탐색결과 재현율이 0가 되는 수가 있다.[27] 이것 역시 탐색자가 주제분야에 대해 얼마나 아는가와 결부된다.

또 탐색자가 탐색에 있어서 적절한 접근을 태만히 하여 질문에서 분명히 언급된 사항을 탐색시 누락할 경우[28] 탐색자가 동의어를 검색어에 포함하지 않는 경우[29]에는 재현실패의 원인이 된다.

3) 탐색출력(search output)형태

이용자에게 제공하는 탐색출력에는 원문헌을 제공하기 전에는 표제와 초록을 제공하는 경우를 들 수 있다. 보통 표제만으로도 어느 정도 수준의 적합성을 판정하기는 가능하지만 만일 표제가 주제를 표현하는데 너무 추상적이거나 애매모호할 경우, 잠재적인 정보를 놓치는 수가 있어 재현율이 떨어진다. 본드(Bond)[30]씨도 표제만으로는 약간의 실패가 있으리라고 느끼기 때문에 초록에 대한 이용자의 신뢰가 높은 반응이 나왔다고 했다.

3.2.3 기타원인

1) 이용자와 시스팀 상호작용

이용자와 검색시스팀간의 상호작용에 있어서의 결함은 어떤 시스팀에서도 특히 큰 시스팀에서는 재현성과 정확성 실패의 중요요인이 된다.[31]

검색시스팀의 이용자가 탐색질문의 답에 완전히 만족한 적은 거의 없다[32]고 한다. 이용자는 시스팀의 조직을 완전히 이해하지 못해서 적절한 질문을 하지 못하는 수가 있다. 즉, 표현한 질문은 일반적인데 실제 이용자가 원하는 정보요구는 특정한 경우 정확률이 떨어지고 표현한 질문은 특정하나 실제 정보요구는 일반적인 경우, 필요없는 문헌이 검색되어 재현율이 저하된다. 또한, 표현한 질문이 요구의 일부만을 나타내고 있을 경우 재현성, 정확성 양쪽의 실패를 초래할 수 있다.

2) 기계처리 시스팀에서

정보를 컴퓨터에 축적하여 놓고 검색을 하는 시스팀에서도 곳곳에 검색효율을 지배하는 요인들이 있다.

자연언어를 컴퓨터에 입력시키기 위해 기계어로 코드화할 때 인간적인 실수, 키펀쳐(key puncher)의 실수, 기계처리의 제한성으로 인한 잘못된 탐색논리는 검색효율을 떨어뜨리는 요인이 된다.[33]

4. 주제명표목과 표제에서 뽑은 키워드 간의 효율성에 대한 평가시험

4.1 실험에 관하여

4.1.1 목 적

이 실험의 목적은 (1) 전산화를 지향하는 정보검색 시스팀을 설계하는데 있어서 사용가능한 색인언어간의 효율성을 비교해 보고 또한 색인언어의 효율성을 높일 수 있는 방법을 고찰하고자 한다. (2) 특히, 인문과학과 자연과학 분야에서 각각 한 주제를 선택하여 각 주제에 주어진 두 가지 색인언어의 효율을 고찰함으로써 그런 특수주제 분야의 시스팀설계에 도움이 되고자 한다.

4.1.2 효율측정방법

이 실험에서의 효율측정방법은 3장에서 언급한 바 있는 정확률과 재현율을 사용한다. 이미 상술하였듯이,

$$\text{정확률} = \frac{\text{정보요구와 관련된 검색문헌수}}{\text{검색된 문헌의 총수}} \times 100$$

$$\text{재현율} = \frac{\text{정보요구와 관련된 검색문헌수}}{\text{파일내의 정보요구와 관련된 문헌의 총수}} \times 100$$

으로 측정하는데 여기서 문제가 되는 것은 재현율을 구할 때 파일내에 있는 관련문헌을 계산하는 일이다. 200~500개 정도의 문헌으로 파일이 구성되어 있는 시스팀의 경우는 파일을 일일이 훑어보는 것이 가능하겠지만, 규모가 조금 큰 시스팀에서는 그런 방법이 이론적으로는 가능하나 현실적으로는 불가능하다.

본 실험에서는 다음과 같은 방식으로 파일내의 관련문헌을 계산해 냈다. 즉 각각의 문헌은 주제명표목과 표제에서 뽑은 키워드를 색인언어로 하는 두가지 색인파일내에 들어가게 된다. 검색시에는 동일한 질문을 가지고 여러가지 탐색전략에 의해서 두 파일내의 관련문헌을 검색한다. 따라서 개개 질문과 관련있는 파일속의 문헌의 수는 두 색인시스팀에서 여러가지 탐색전략에 의해서 관련문헌을 검색하며, 따라서 개개 질문과 관련있는 파일속의 문헌의 수는 두 색인시스팀에서 여러가지 탐색전략에 의해서 관련문헌을 검색한다. 이에 따라 개개질문과 관련있는 파일속의 문헌의 수는 두 색인 시스팀에서 여러가지 탐색전략에 의해 검색된 관련있는 문헌의 합으로 계산된다.

4.1.3 실험 구성요소

1) 주제분야 : 인문과학에서 교육심리학, 자연과학에서는 생화학을 선정했다. 교육심리학은 Psychological Abstracts(Washington : American Psychological Association)에서 교육심리학 부분에 수록된 자료를 대상으로 했다. 이용부분은 Vol. 57, No. 1-6, 1977, Vol. 55, No. 4-6, 1976이다.

생화학은 Biological Abstracts(Philadelphia : Biosciences Information Service of Biological Abstracts)에서 생화학부분에 수록된 자료를 대상으로 삼았다. 이용부분은 Vol. 62, No. 1-12, 1976, Vol. 64, No. 1-11, 1977을 이용하였다.

2) 파일 크기 : 교육심리학 분야의 문헌 1,000개, 생화학 분야의 문헌 1,000개로서 도합 2,000개의 문헌[34]으로 구성된다.

3) 입력자료 : 색인작성에 필요한 입력자료는 문헌의 표제를 기본으로 심되 주제명표목을 주는 시스팀에서 표제만으로 문헌내용을 알 수 없을 경우 초록을 참조한다.

4) 색인언어 : 각 문헌은 다음 2가지의 색인언어로 색인된다.

A. 주제명표목(fixed language : 고정언어)

B. 표제에서 뽑은 키워드(free language : 자유언어)

주제명표목은 주로 1975년에 간행된 LC주제명표목 제8판속에 수록된 표목을 사용하였고 필요에 따라 약간 확장시켜 사용하였다. LC주제명표목의 선택이유는 첫째, 교육심리학과 생화학에 주어지는 주제명표목의 특정성 수준을 같게 하기 위해서이고 둘째, 색인대상이 된 초록지 속의 주제명과 LC주제명표목을 비교해본 결과 특정성 수준이 어느 정도 일치되었기 때문이다.

5) 실험자 : 실험에서 필요한 인적요소는 색인자와 탐색자인데 본 연구에서는 색인자와 탐색자가 다르므로 인해 야기되는 시스팀에 대한 이해의 차이를 줄이기 위해 색인작업과 탐색작업을 동일한 사람이 수행했다. 실험자는 각 주제분야 전공자 4학년생 1명과 도서관학 전공대학원생인 본 연구자로 모두 3명으로 구성되었다. 색인작업에 들어가기 앞서 약 1주간의 연습기간을 두었다.

6) 색인파일 구성 : 주제명표목으로 색인된 색인파일은 주제명의 자모순 배열을 하여 각 주제명아래 문헌번호가 들어가게 했고 키워드를 색인언어로 하는 색인파일도 각 키워드아래 문헌번호를 주고 키워드들을 자모순으로 정리하였다. 그리고 문헌의 서지의 주부(main part)는 후에 적합성을 판정할 때 사용하기 위해 1~1,000번의 문헌번호순으로 정리하여 두었다.

7) 질문작성 : 질문은 실험자에 의해서 두가지 원칙에 따라 작성했다. 즉,

A. 전체 문헌파일에서 대충 뽑은 문헌내용에 기초한 질문

B. 문헌파일이 망라하고 있는 주제분야에 대한 일반적 지식에 의거하여 질문들이 작성되었는데 각 주제분야당 53개의 질문작성으로 두 주제분야에서 도합 106개의 질문으로 구성되었다.

8) 질문분석 : 질문은 그 내용을 분석하여 가장 적합한 주제명표목을 주었고 질문의 내용이 복합적일 때는 주제명표목 2개를 주었다. 질문에서 뽑은 키워드는 보통 2~3개 정도로 추출되었는데, 명사의 경우 복수형을 단수형화하는 정도의 통제를 가했다.

9) 탐색전략 : 일차로 질문분석에 따라 주제명표목과 통제하지 않은 키워드의 조합[35]으로 탐색하고, 다음 질문속의 키워드를 탐색한다.

10) 적합성 판정 : 검색된 문헌은 관련문헌과 비관련문헌으로 대별했는데 질문에 대해 극소의 관련문헌은 비관련문헌으로 간주한다.

11) 평가 및 분석 : 실험내용의 평가 및 분석은 본 연구자에 의해 행해진다.

4.2 실험내용

4.2.1 색인작업

시스팀 설계를 위해서 색인작성의 대상이 된 문헌들은 Psychological Abstracts의 교육심리학 분야에 포함된 잡지기사, 단행본, 논문 등으로서 1,000개의 문헌을 추출했다. 따라서 문헌파일을 구성한 문헌들은 1~1,000까지의 고유번호를 갖게 된다. 마찬가지로 Biological Abstracts의 생화학부분에 실린 문헌 1,000개를 추출하여 1~1,000까지의 고유번호를 주었다. 두 abstracts journal의 선정이유는 초록대상이 되는 문헌들의 범위가 광범위하며 각각의 연구분야에서 이용도가 높기 때문이다.

이 두 주제분야는 각각 두가지 색인언어 즉, 주제명표목과 문헌의 표제에서 추출한 키워드들로 색인해 주어 표 5와 같이 4가지 색인파일을 구성한다.

따라서 본 실험은 표준전거목록을 이용한 건조합, 고정언어(precoordinated, fixed language)와 검색시 키워드들을 자유로이 조합하는 후조

표 5. 색인파일 구성

주 제	색인파일	색 인 언 어	입력자료
교육심리학 (E)	E 1	주제명표목 (건조합고정언어)	표제, 초록
	E 2	표제에서 추출한 키워드 (후조합자유언어)	표제
생 화 학 (B)	B 1	주제명표목 (전조합고정언어)	표제, 초록
	B 2	추출한 키워드 (후조합자유언어)	표제

표 6. 색인파일의 망라성

색인언어	색인파일	문헌당 색인어수
주제명표목	E 1	1.3 개
	B 1	1.1 개
키이워어드	E 2	4.5 개
	B 2	5 개

합, 자유언어 (postcoordinated, free language)의 비교라고도 볼 수 있다.

색인작업은 수작업으로 행해졌다. 색인자는 각 주제를 전공한 주제전문가 A와 B 그리고 본 연구자 C로 구성되었는데 A와 C가 한 주제분야를, B와 C가 다른 한 주제분야의 색인작성업무를 행했다.

색인작업의 망라성 수준은 색인어의 수로 결정될 수 있겠는데 각 색인파일의 망라성 수준은 표 6과 같다.

파일 E1 및 파일 B1과 관련된 주제명시스팀의 경우, 표제와 초록을 참고로 하여 준색인언어의 특정성 수준은 키워드시스팀과 비교할 때 일반적인 편이다,

파일 E2와 파일 B2는 문헌의 표제로부터 추출한 키워드들로 구성되었는데 키워드의 선정 및 색인어 배정은 다음과 같다. 즉,

A. 문헌의 표제속의 불용어휘 (stop word) 즉 관사, 전치사, 접속사 등은 제외하고 나머지 명사, 형용사, 동사, 부사… 등을 채택한다.

B. 채택한 키워드 중에서 복수로 된 용어는 복수와 단수의 개념의 차이가 뚜렷하지 않는 한 단수형을 채택한다.

C. 약자(略字)를 사용해도 검색에 지장이 없는 것은 약자를 사용했다. 그 예로, "adenosine triphosphate"는 "ATP"로 "messenger RNA"도 "m-RNA"식으로 색인해 주었다.

D. 대부분이 단일용어를 주었으나 복합어로 의미가 확정된 것은 복합어로 주었다(예 : high school)

4.2.2 각 파일의 조직

파일 E1과 파일 B1은 수작업으로 주제명을 자모순으로 배열하고 그 주제명 아래에 문헌번호들이 들어갔다.

파일 E2와 B2의 경우 키워드들을 모아 자모순으로 배열하고 그 키워드 아래 문헌번호를 넣었다. 이 작업은 문헌당 4~5개의 키워드가 배정되어 4×2,000(문헌수)=8,000 8,000개의 키워드들을 자모순으로 정리하기란 어려운 일이어서 컴퓨터로 처리하였다. 이는 단순히 색인자가 선정한 키워드들을 컴퓨터에 의해 배열 정리한 것에 불과하다.

4.2.3 질문작성

앞에서도 고찰한 바와 같이 검색효율은 시스팀과 이용자 사이의 상호관계에 의해서도 영향을 받게 된다. 탐색자가 축적된 컬렉션과 이용자의 요구사이의 관계를 완전히 파악하고 있다면 예의적으로 훌륭한 결과를 유도해 낼 수 있으나 탐색자가 탐색하는 동안에 이용자의 마음이 변하는 수도 있어 처음 요구내용과 차이가 나는 경우도 있다. 이러한 가변적 요소가 검색효율에 영향을 미치게 된다.

물론 본 실험에서처럼 실험자가 질문작성하는 경우도 인간적인 실수나 시스팀과의 상호작용에서 파생되는 실수를 완전히 배제할 수 없겠지만 어느 정도까지는 가변적 요소를 극소화 내지 고정시킬 수 있다는 전제하에 실험자들이 이용자의 역할을 대신한다.

질문은 전체 테스트 컬렉션에서 대충 뽑은 문헌내용에 기초한 질문과 테스트컬렉션이 망라하고 있는 주제분야에 대한 일반지식에 의거한 질문이 각각 반정도씩으로 구성됐다.

질문의 수는 파일 E1과 E2에 동일한 질문 각각 53개씩 106개, 파일 B1과 B2에도 동일한 질문 각각 53개씩 106개를 주어서 전체 4개의 파일을 대상으로 약 212번의 기본적 탐색을 시도했다.

4.2.4 질문분석과 탐색전략

주제명시스팀에 주어진 질문들은 질문의 내용을 가장 적절하게 표현해 주는 주제명표목을 선정하였고 질문내에 복합적인 주제가 있는 경우, 복합적인 주제를 표현하는 주제명표목을 부차적

으로 주었다. 그 예로,

"creative thinking of student"라는 질문은 "creativity"와 "thought and thinking"이라는 표목을 갖게 된다.

키워드시스팀의 경우는 색인작업시 키워드 추출방법과 같은 요령으로 분석 추출하였다. 질문당 추출한 키워드 수는 파일 E2에 대한 질문이 약 2.5개, 파일 B2에 대한 질문에서는 약 2.2개로 나타났다. 다시 말해 이 두 주제분야에 주어지는 질문들은 평균 2내지 3개의 키워드들을 포함하고 있다는 것이다.

탐색전략은 우선적으로 질문분석에서 주어진 주제명표목과, 키워드의 조합으로 탐색한다. 조합수준은 키워드 2개를 조합하여 탐색을 시도하였다. 그런 다음, 2차로 키워드시스팀의 경우 검색효율을 향상시키기 위해 질문의 키워드에 대

표 7. 키워드 시스팀에서 질문확장 내용

교육심리학			생화학		
질문번호	질문속의 키워드	키워드 확장	질문번호	질문속의 키워드	키워드 확장
2	training counselor	training, counselor education	10	preparing t-RNA	preparation, preparing t-RNA
13	environment	environment, behavior, attitude	16	protein denaturation	protein denaturation inactivation
14	violence treatment	agressiveness violence treatment	23	DNA structure	DNA, structure mapping
15	childhood development	childhood, development kindergarten preschool	25	DNA separation	DNA, separation isolation
21	teacher training	teacher, training educator instructor	26	protein determining	protein, determining determination identification
29	teacher job-satisfaction	teacher job-satisfaction work-satisfaction	28	porphyrines synthesis	porphyrines, synthesis biosynthesis
32	job-satisfaction teacher retarded	job-satisfaction teacher retarded handicapped	31	vitamine C analysis	vitamine C, analysis ascorbic acid water-soluble acid
37	training creative thinking	training creative creativity	45	amino acid analysis	amino acid analysis, identification
			48	isolation RNA	isolation, RNA separation
			50	protein conformational structure	protein conformational structure, conformation

해 동의어 및 관련어를 확장시켜 주고, 단어형태를 조절하였다. 이 작업은 생화학계통의 용어집과 교육심리학의 디소오러스를 참조하거나 주제분야의 경험적 배경을 통해서 질문에 사용된 키워드들을 확장시켰다.

그 예로, 생화학에서 "DNA structure"는 "DNA mapping"으로도 찾을 수 있었다. 이 경우 "structure"와 "mapping"은 관련어이다. 교육심리학의 경우는 그 빈도가 더욱 심했다. 몇가지 예를 들면, "attitude"와 "behavior," 국민학교를 뜻하는 "primary school", "elementary school" 등이 혼용되며 또 "지체부자유자(physically handicapped)"나 "정신적 지진아(mentally retarded)"는 때로 "handicapped"라는 용어나 "disadvantaged"라는 용어로 대용되기도 한다. 또한 단어형태를 조절하여 "structural" 이라는 키워드일 경우 "structure"로도 탐색하였다.

용어를 확장해서 검색한 결과 관련문헌이 추가된 키워드의 자세한 내용은 표 7과 같다.

4.2.5 결과 및 분석

색인언어간의 효율성의 비교는 주제분야별로 비교하기로 한다.

1) 생화학 분야에서의 색인언어의 효율성

파일 B1과 B2의 검색결과 나타날 주제명시스팀과 키워드시스팀간의 효율은 표 8과 같다.

아래 표에서 보면 주제명시스팀에서 재현율은 높고 정확률이 낮은 반면에 키워드시스팀에서는 정확률, 재현율 모두 어느 정도 선까지는 유지되는 것으로 나타냈다. 그 이유는 파일이 다루고 있는 주제의 성격이 뚜렷하기 때문으로 생각할 수 있다. 파일을 구성하고 있는 문헌들의 표제 속에는 생화학의 물질명들이 많이 들어있기 때문이다. 따라서, 주제명시스팀의 경우 한 주제명표목(그것이 비교적 일반적인 화학물질명이라면)아래 그에 관련된 복합물이나 치환물이 들어가게 되어 높은 재현율을 얻을 수 있었으나, 또한 비관련문헌 역시 많이 검색되어 정확률이 떨어졌다.

반면에 키워드시스팀의 경우 질문속의 키워드를 2개 수준에서 조합하였는데, 표현의 애매성이 적어서 검색된 문헌은 대다수가 질문과 부합되는 문헌들로 77%의 정확률을 보였다. 그러나 재현율의 경우는 50%의 보통 수준은 되었지만 주제명시스팀의 79%보다는 낮았다. 어휘의 통제를 하지 않았기 때문에 vitamine을 ascorbic acid 또는 water-soluble acid라고 하거나 화학물질의 분리를 뜻할 때 다른 물질에서의 분리를 하는 경우 isolation, separation이라는 용어를 자주 사용하고 같은 물질내에서의 분리는 fractionation이라는 용어를 사용하여 관련문헌이 누락되어 재현율이 떨어졌다. 그래서 키워드시스팀의 경우 개현율을 높이기 위해서 탐색질문에 상기에서 살펴본 바의 용어들의 동의어, 관련어, 상위개념, 단어형태 조절을 통한 탐색질문을 확장하여 10개 질문에 대해 탐색한 결과 1차에서 검색되지 않은 관련문헌 15개를 더 얻을 수 있어 약 8%의 재현율이 증가하였다.

과학분야의 이용자들의 검색문헌에 대한 만족성향을 보면 일부분이 관련된 다량의 문헌보다는 소량일지라도 질문에 가까운 답을 원하고 있다[36]고 한다. 이는 이용자가 높은 정확률을 요청한다는 것을 의미한다. 또한, 장래 거계처리시스팀의 도입을 염두에 두고 있을 경우 색인작성자가 어떤 경험없이도 표제에서 키워드 추출로 쉽게 시스팀을 운영할 수 있고 더 나아가서 컴퓨터내에서 자동으로 불용어휘(stop word)만 제거하고 나머지는 키위드로 처리할 수 있는 방법으로까지 발전시킬 수 있다는 이점을 감안할 때 효율 10%, 79%의 주제명 시스팀보다는 76%, 50%의 효율을 보여준 비통제 자유언어가 과학분야 문헌들의 색인에 꽤 타당성이 있다. 그러나 좀더 관련문헌을 검색하기 위해서는 색인과정이나 탐색과정에서 어휘통제를 해주는 것이 더욱 적합하다고 보여진다.

2) 교육심리학 분야에서 색인언어의 비교

표 8. 생화학분야의 색인언어 효율성

파일	색인언어	검색문헌수			검색효율	
		관련문헌	비관련문헌	Total	정확률	재현율
B1	주제명표목	166	1,454	1,620	10%	79%
B2	키워드	104	32	136	77%	50%

표 9. 교육심리학분야의 색인언어의 효율성

시스템	색인언어	검색문헌수			검색효율	
		관련문헌	비관련문헌	Total	정확률	재현율
E 1	주제명시스템	174	1,060	1,134	15%	82%
E 2	키워드	82	133	215	38%	39%

교육심리학에서의 검색 결과는 표 9와 같다. 표에서 보듯이 주제명시스템의 경우 정확률 15%, 재현율 82%로 질문에 대해 꽤 많은 관련문헌이 검색되었지만 또한 꽤 많은 비관련문헌도 함께 검색되었다. 그 이유는 탐색전략을 짤 때 질문의 내용을 표현해 주는 검색으로 일반적인 용어를 선택했기 때문이다. 이는 탐색전략의 특정성과 관련이 있는데 검색어를 특정한 용어로 주었을 경우(모든 질문에 특정성을 줄 수는 없었고 약 10개 정도에 좀더 좁은 개념의 용어를 배정했다) 정확률이 1% 개선되었고 재현율은 9%가량 떨어졌다.

따라서 전 질문을 통해 특정성 수준을 조절하여 원하는 수준의 정확률과 재현율을 얻을 수 있으리라는 것이 예상되었다. 상기 표 9의 재현율 82%는 이 주제분야의 이용 성격상 연구시 많은 자료를 참고로 한다는 특성을 생각할 때 꽤 타당한 효율을 주었다고 볼 수 있다.

반면에 키워드시스템의 경우는 심각한 문제점이 들어났다. 이 시스템은 복수를 단수로 조절한 이와에는 통제를 하지 않은 어휘를 사용하였기 때문에 언어의 애매성, 추상적인 표현, 동의어, 관련어, 구문적 요인 등이 노출되어 조합의 실패를 초래하여 정확률, 재현율 모두 낮았다. 이 분야에서 주제명시스템으로 관련문헌이 검색된 질문이 키워드시스템으로 검색에 완전히 실패한 경우가 셋이나 되었다. 그 예로, "correlations between environment and a view of value"(이 질문은 학생의 가치관과 환경과의 관계를 뜻한다)라는 질문에서 키워드로 "environment", "value", "view"를 추출 조합한 경우 파일내에 꼭 "view of value"로 표현된 문헌이 있기란 쉽지 않다. 실제, 파일내의 이 질문과 관련된 문헌의 표제를 보면 "impact of the high school environment", "student perception of the school environment as related to moral development" 등으로 표현되어 있어 표제에서 뽑은 키워드와의 매치는 힘들다.

동의어 문제 또한 심각하다. 그 예로 교사를 나타내는 용어로 "teacher", "instructor", "educator" 또는 "counselor"로 혼합 사용하고 있어 언어의 통제가 절실히 요구되고 있다. 또한 언어의 구문적 관계의 실못에 의한 실패도 있었다. 예를 들면, "teachers attitude to student"를 원했는데 "students attitude to teacher"가 검색된 예이다. 이와 같이 같은 내용을 표현해도 표현하는 사람의 주관에 따라 나양하게 표현할 수 있는 언어적 특성때문에 질문과 관련있는 문헌들이 많이 누락되었고(재현율 39%), 검색된 문헌도 관련있는 문헌이 많지 않았다(정확률 38%).

이런 경우를 위해 약 10개의 질문을 가지고 동의어, 관련어 및 단어형태를 조절하여 검색을 시도했다. 통제를 위한 도구로 Information Retrieval Thesaurus of Education Terms, The Press of Case Western Reserve Univ., Cleveland, 1968을 참고하거나 주제분야의 경험에 의지하여 질문을 확장시켜 주었는데 탐색결과, 관련문헌 약 17개를 더 얻을 수 있었으나, 워낙 주제의 애매성이 심해 이 17개로 전체효율이 향상되었다고 단정할 수는 없겠다. 다만 다양한 질문의 어휘통제에 의해 잠재적인 관련문헌을 검색해 낼 수 있다는 것을 보여 주었다고 하겠다. 여기서는 비관련문헌도 함께 상당수 검색되었다.

이렇게 인문과학 분야에서의 색인언어로 키워드시스템을 채택할 경우 잠재적인 관련문헌을 잃지 않기 위해서는 반드시 인어의 예증관계에 의한 통제를 필요로 하게 된다. 그에 필요한 도구로 강력한 예증관계를 보여주는 디소오러스의 구성이 요청된다. 그에 따라 주제전문가의 언어적 배경, 디소오러스 구성에 드는 시간과 경비 등도 고려하지 않을 수 없게 된다.

인문과학 분야의 연구자들은 좀더 많은 관련자료들을 참고로 하는 경향이 있어 재현율의 측면이 강조된다 하겠다. 이렇게 볼 때에 한 주제

명 아래 포괄적으로 관련문헌이 들어가는 주제명시스팀이 복잡한 언어의 통제를 필요로 하는 키워드시스팀보다 효율적이라고 보겠다. 이것은 실험결과에서 주제명 시스팀의 재현율이 82%인데 비해 키워드시스팀은 39%로 나타난 것으로 증명된다.

5. 일반적 결론

색인언어 효율성의 크기는 효율성을 지배하는 요인들과 밀접하게 관련되어 있다는 것을 알 수 있다. 인간의 의사결정에 의해 정해지는것 즉, 망라성의 정도, 색인언어의 특정성, 탐색전략 등 인간이 시스팀의 성능을 좌우하는 정도는 실로 크다 하겠다. 어떠한 실험에서건 실험을 주도하는 인간이 실험내의 크 많은 가변적 요소들을 파악하기는 어렵기 때문에 실험결과를 절대적이라고 할 수는 없다.

그러나, 실험내의 여러 불변적 요소와 가변적 요소를 포함하는 특정상황에서 진행된 실험결과를 통하여 시스팀 설계에 고려해야 할 사항을 제시할 수 있다. 본 실험결과에 의한 일반적 결론은 다음과 같다.

우선 자연과학 분야의 정보검색시스팀설계시, 전산화를 계획하는 경우에 특히 주제가 뚜렷한 자료들을 색인해 주기 위해서는 실험결과에서 보는 바와 같이 정확률 77%, 재현율 50%의 효율을 보인 비통제 키워드 시스팀이 정확률이 낮은 주제명 시스팀보다 더 유리하리라 본다. 이러한 자유언어의 사용은 색인자가 색인 작업시 어떤 경험이나 두뇌적인 사고과정을 필요로 하지 않고도 쉽게 시스팀을 운영할 수 있다는 이점이 있으며 자료의 자동 색인처리도 가능하다. 또한 검색시에도 이용자가 그 자신의 용어로 검색질문을 한정할 수도 있어서 이용자와 컬렉션사이의 해석자가 필요없게 되고 형식적인 디소오러스를 계속 편찬해야 하는 어려움을 피할 수 있다. 그러나, 어떤 특정 시스팀에서 좀더 많은 관련문헌을 검색하기 위해서는(높은 재현성을 바란다면) 동의어, 관련어, 좀더 일반적인 용어의 채택, 단어형태의 조절 등의 통제를 색인과정이나 탐색전략을 짤 때 채택하는 것이 좋다고 본다. 이러한 통제는 과학분야에서 사용되는 기성 디소오러스를 사용하여 해결할 수 있다고 하겠다.

따라서 과학분야의 문헌을 다루고 있는 특수도서관이나 정보관리센티에서는 당 시스팀을 이용하게 될 이용자의 특성, 주제분야의 특성 등을 고려하여 키워드의 통제여부, 통제수준을 결정하여 시스팀의 최적화를 기하도록 해야 할 것으로 생각한다.

다음으로 인문과학분야를 전산화할 경우 비통제 키워드 시스팀은 인이의 추상성, 애매성, 함축성때문에 효율을 기대하기 어려우리라 본다. 만일 이 분야에서 키워드 시스팀을 사용할 경우 색인언어에 대한 강력한 통제가 요구된다. 이경우, 기성 디소오러스를 사용한다면 그 디소오러스가 어느 정도 강력한 예증관계를 보여 주는가가 문제가 되고(본 실험에서 사용한 디소오러스는 그다지 강력한 예증관계를 보여주지 못했음), 디소오러스를 구성하는 경우에는 그에 필요한 인적요소, 시간, 경비 등을 고려해야 할 것이다.

인문과학분야의 연구 성격이 많은 자료를 참고로 한다는 점을 감안한다면, 상술한 바의 문제점이 노출된 키워드 시스팀보다는 재현율이 82%로 높은 효율을 보인 주제명 시스팀이 더 효율적이다.

또한 주제명 시스팀의 경우 색인언어의 특정성 수준을 조절하여 시스팀에 적합한 재현율과 정확률을 얻을 수 있어 원하는 성능수준을 달성할 수 있으리라 본다.

参考文獻 및 註

24) Tefko Saracevic, "Selected Results From an Inquiry into Testing of Information Retrieval Systems," Journal of the American Society for Information Science, Vol. 22, No. 2 (1971), p. 133.

25) Ibid., p. 135.

26) F. W. Lancaster, Information Retrieval Systems : Characteristics, Testing and Evaluation (New York : John Wiley, 1968), p. 71.

27) Gerald, Salton, "A Comparative Between Manual and Automatic Indexing Methods," Journal

of the American Society for Information Science, Vol. 20, No. 4 (1969), p. 66.
28) Julie A. Virgo, "An Evaluation of Index Medicus and MEDLARS in the Field of Oplthalmdogy," Journal of the American Society for Information Science, Vol. 21, No. 4 (1970), p. 262.
29) Saracevic Tefko, "Selected Results from an Inquiry into Testing of Information Retrieval Systems," Journal of the American Society for Information Science, Vol. 22, No. 2 (1971), p. 135.
30) Bond and Others, "A Computerized Current Awareness Service Using Chemical-Biological Activities (CBAC)," Journal of Documentation, Vol. 9, No. 3 (1969), p. 161.
31) 吳哲 哲, 情報檢索論, (서울 : 韓細亞文化社, 1977), 173面.
32) Betrand C. Landry and James E. Rush, "Opimon Paper : Toward a Theory of Indexing Ⅱ," Journal of the American Society for Information Science, Vol. 21, No. 3 (1970), p. 365.
33) G. Olive and Others, "Studies to Compare Retrieval Using Titles with that Using Index Terms : SDI from Nuclear Science Abstracts," Journal of Documentation, Vol. 29, No. 2 (1973), p. 186.
34) 여기서 문헌이란 사일내에 수록된 표제 하나하나를 뜻한다.
35) 조합방법은 수학의 집합이론 가운데서 적집합 방식으로 소개했다. 즉 키워드 A와 B, 조합시 A and B(집합기호는 A ∩ B)로 조합한다. 이는 A와 B를 동시에 갖는 문헌을 탐색하는 방법이다.
36) 崔成鎭, 情報學原論, (서울 : 亞細亞文化社, 1976), 220面.

情報管理研究 累加索引

Vol. 10(1977)~Vol. 11(1978)

編輯者註 : 지금까지 本索引은 1978年 以前分에 대한 累加索引을 2回에 걸쳐서 不定期的으로 收錄하였으나, 1979年부터는 每年 마지막 號의 卷末에 當該年度의 索引을 收錄하고 每 5年마다 累加索引을 收錄할 豫定이다.

凡 例

1. 收錄範圍

1977年 1月부터 1978年 12月까지 情報管理研究誌에 揭載되었던 모든 記事를 收錄하였다.

2. 排 列

모든 記事는 情報管理一般, 情報源, 情報蒐集·處理 및 提供과 情報檢索, 情報시스템·流通 및 시어미스, 기타 등 6主題로 大別하여 各 項目內에서 標題名의 가나다順으로 排列하였으며, 外國語로 된 것은 우리말 다음에 「ABC順」으로 排列하였다.

3. 記述方法

이 索引에서 記事의 記述方法은 ① 論題, ② 筆名, ③ 譯者(但 있을 경우), ④ 卷, 號數, 收錄「페이지」, ⑤ 刊行年의 順으로 하였다.

<例>特許制度의 國際化 動向. 李鉉澈. **10**, 4 (103-108) 1977.

情報管理一般

計劃像算制度의 導入에 대하여.
李帖範. **11**, 3 (62-73) 1978
目標管理制度(MBO)의 理論的 考察.