학문의 주제별 특성에 따른 자동색인기법의 비교연구

-약학분야와 도서관·정보학 분야를 중심으로-

A Comparative Study of Automatic Indexing Techniques in Pharmacology and Library & Information Science.

조수련* 사공철**

초 록

본 연구는 서로 다른 주제를 갖는 장서내의 통계적 용어특성에 따라 적합한 자동색인기법을 제시하는데 그 목적이 있으며 약학분야와 도서관·정보학분야를 대상으로 하여 두개의 자동색인기법을 비교, 평가하였다.

사용된 자동색인기법은 역문헌빈도가중기법과 문헌분리가(文獻分離價)가중기법이며 총장서빈도와 문헌빈도로 정의되는 용어특성과 자동색인기법으로 할당된 가중치들 간의 관계를 분석하였다.

ABSTRACT

The purpose of this study is to present a relevant automatic technigue in accordance with the statistical term characteristic in a collection comprising different subjects, by comparing and evaluating two automatic indexing techniques(Inverse Document Freguency Weighting Technigue and Term Discrimination Value Weighting Technigue) in the fields of Pharmacology and Library & Information Science.

1. 서 론

1. 연구의 필요성

정보요구가 점차 다양해지고 전문화되어 감에 따라 정보의 양적, 질적 측면뿐 아니라 요구되는 정보에의 접근(access)등과 같은 정보문제도 타학문의 적용과 함께 더욱 활발히 연구되고 있다.

북스타인(A.Bookstein)과 스완슨(D.R. Swanson)은 문헌검색시스템의 목적이 비교적 방대한 문헌장서로부터 정보요구를 만족시키기에 적합한 양의 문헌을 선정하는 것으로, 이 정보시스템에는 각 문헌을 표현하는 기능이 필요하다[1] 고 말하고 있다.

1) Abraham Bookstein and Don R.Swanson. "Probabilistic Models for Automatic Indexing," Journal of the American Society for Information Science,25 (Sept / Oct. 1974). P. 312.

* 숙명여자대학교

** 숙명여자대학교 도서관학과

색인은 문헌을 표현하는 기능이며,정보검색의 핵심과정으로 검색효율을 극대화하기 위해 여러 기법들이 제시, 실험되어 왔으며 이러한 실험을 통해서 그 타당성과 실제 적용상 효율성의 향상이 입증된 상당수의 이론들이 정보문제의 해결책으로 받아들여지고 있다.

그러나 이러한 기법들의 적용과정에서 색인환경의 차이점이 고려되지 않았으며 이제까지의 대부분의 실험이 자연과학분야(hard subject)에 국한되었다. 또한 서로 다른 주제를 대상으로 한 상이한 속성들에 대한 연구는 거의 없는 실정으로 특정장서를 대상으로 한 실험에서 검색성능의 향상을 전제로 선정된 색인기법이 통용되고 있다.

따라서 정보검색 성능을 좌우하는 색인기법들이 적절히 선택, 사용될 수 있도록 주제 간의 상이한 속성에 대한 연구가 행해져야 한다.

실제로 샐튼(G.Salton)은 서로 다른 주제를 대상으로 한 실험에서 '하드 서브젝트(hard subject)'와 '소프트 서브젝트(soft subject)'간에는 일반적 속성에 어느정도의 차이가 있다[2] 고 지적하였는데 이러한 결론이 본 연구의 필요성을 뒷받침하고 있다.

2. 연구의 목적 및 가설

본 연구의 목적은 장서내 문헌들의 주제와 그 문헌에 출현하는 용어들의 통계적 특성 간의 관계를 밝힘으로써 환경에 따라 최적인 자동색인기법을 선정하는데 기준을 제시하기 위한 것이며, 연구를 위해 다음과 같은 가설을 설정하였다.

(1) 서로 다른 주제 문헌을 표현하는 장서내의 용어특성 간에는 주제에 따라 차이가 있을 것이다.

(2) 용어의 특성이 색인기법의 성능에 영향을 미칠 것이다.

(3) 서로 다른 주제의 문헌장서에 적합한 색인기법은 다를 것이다.

3. 연구의 방법 및 범위

자동색인기법을 평가하는 방법에는 동일문헌에 대한 수작업색인과 자동색인을 비교하는 방법과 이용자에 의해 제시된 질의를 근거로 검색 실험하는 방법의 두가지가 있는데, 두번째 방법은 주관적 적합성 판정의 평가문제로 실제 적용에 어려움이 있으며 첫번째 방법이 간단히 적용될 수 있는 객관적 방법으로 널리 사용되고 있다.[3]

본 연구에서는 첫번째 방법에 따라 화학 및 화공분야의 초록지인 CA(Chemical Abstracts)와 도서관·정보학분야의 초록지인 LISA (Library & Information Science Abstracts) 가운데 무작위로 추출한 각 201개의 초록을 대상으로 하여 스파크 존스(K.Sparck Jones)의 역문헌빈도가중기법과 샐튼(G.Salton)의 문헌분리가(文獻分離價) 가중기법을 적용하여 분석, 평가하였다.

입력시 불용어(stop word)를 제거하고 용어절단(truncation)은 기존색인을 참고하였으며, 각 장서마다 두가지의 색인기법이 평가되었는데 평가방법은 기존색인과의 일치율을 근거로 하였다.

---

2) Gerald Salton and C.S.Yang. "On the Specification of Term Values in Automatic Indexing," Journal of Documentation, 29(Dec. 1973). P. 358.

3) Stephen P. Harter. "A Probabilistic Approach to Automtic Keyword Indexing : Part II," Journal of the American Society for Information Science, 26 (Sept / Oct. 1975). P. 285.

연구의 범위에는 다음의 사항들이 포함된다. 입력시 사용 패키지(package)의 제한용량과 입력시간 제한등의 문제로 초록을 201개로 한정할 수 밖에 없었으며, 총장서빈도가 1인용어가 너무 많으므로, 복합어 처리를 하지 않았다. CA의 경우 '-'로 연결된 화학명을 분리 입력하고 약어도 함께 입력하였다.

또한 기존색인과의 일치율 측정에 있어 CA에는 화학물질명만을 따로 취급하는 화학물질명색인(Chemical Substances Index)이 있으나 LISA의 색인과 비교해 볼 때 지나치게 많은 적합색인어가 생기므로 일반주제명색인 만을 대상으로 하였다.

Ⅱ. 이론적 배경

1. 자동색인의 발달과정

자동색인은 문헌을 분석하여 특성을 부여하고, 탐색과정에서 그 기술(description)을 조작(操作)하며, 이러한 목적으로 사용되는 색인어를 선정하기 위한 자동적인 기법이다.[4]

클리블랜드(D.B.Cleveland)등은 기본적으로 색인작업(indexing)을 할당색인법[5](assigned indexing)과 유도색인법[6](derived indexing)으로 나누어 설명하였으며, 자동색인은 원문내의 단어들과 그들 간의 관계가 내용개념을 표현하는데 충분하다는 가설에 근거하므로 유도색인법에 속한다[7] 고 하였다.

자동색인은 다시 통계적 기법과 비통계적 기법, 가중치기법과 비가중치기법, 소급적 탐색에 적용되는 기법과 SDI (Selective Dissemination of Information) 및 반복탐색에 적용되는 기법[8]의 세가지 측면에서의 방법론적인 분류가 가능하다.

비통계적 기법이란 어의적, 구문적 측면에서 문장을 분석하는 기법으로 촘스키(N.Chamsky)의 언어학적 모델이나 디소러스(thesaurus)의 사용이 그 대표적인 것이며, 힐만(D.J.Hillman)[9]이 구문론(syntax)을 근거로 개발한 바도 있으나 자동색인으로서의 활동단계에는 머무르지 못하고 있다.

초기 색인시스템에는 이진모드(binary mode)인 비가중치기법이 많이 활용되었는데, 색인작업이 비교적 간단하다는 이점이 있으나 모든 검색항목에 동일한 값이 주어지고 많은 수의 용어들이 질의에 할당되므로 이용자가 처리하기에 방대한 항목들이 검색되는 경향이 있어 검색항목들의 적합성 정도를 구분하기 위한 방법이 필요하게 되었다.

이에 따라 이용자 측면에서의 검색항목에 대한 중요도를 반영하기 위해 검색항목을 가정된 유용성에 따라 순위를 매기는 방법이 고려되었는데, 이것이 가중치 색인기법으로 용어의 빈도특성이나 용어의 적합성등을 사용하여 용어에 가중치를 할당, 항목을 검색하는 것이다.

이때 검색의 기준은 기준치(threshold) 나

---

4) K.Sparck Jones. "Automatic Indexing," Journal of Documentation, 30 (Dec. 1974). P. 393.

5) 할당색인법 ; 문헌내의 개념을 기술하기 위해 디스크립터(descriptor)를 색인작성자가 선정하는 방법

6) 유도색인법 ; 저자가 사용한 용어를 아무런 수정도 하지않고 그대로 사용하는 방법.

7) D.B.Cleveland and D.Cleveland. Introduction to Indexing & Abstracting. Littleton : Libraries Unlimited, 1983. P. 148.

8) S.E.Robertsom. "Specificity and Weighted Retrieval," Journal of Documentation,30 (March 1974). P. 41

9) K.Sparck Jones. Op. cit., P. 400.

스코어(score)를 사용하는 것으로, 첫번째 방법은 기준치가 T보다 높은 것을, 두번째 방법에서는 상위 N번째까지의 스코어에 속해 있는 항목을 검색하게 된다.

샐튼은 가중치용어의 사용으로 검색효율 및 이용자의 만족도를 향상시킬 수 있으며, 검색항목의 수를 통제하여 이용자의 노력을 최소화할 수 있다고 주장했으며,[10] 스파크 존스도 고정어휘(fixed vocabulary)를 선택하는 것은 장서변화에 따를 수 없으므로 대안적인 전략으로 가중치에 의해 색인어휘를 통제한다면 거의 다른 노력을 들이지 않고도 장서의 수명을 고려한 장서의 성장에 대응할 수 있을 것이라고 제안했다.[11]

실제로 많은 실험을 통해 가중치기법의 성능향상이 입증되었으며, 샐튼은 공기역학분야의 크랜필드(Cranfield)장서, 의학분야의 메들라인(Medline)장서, 세계사(world affairs)분야의 타임(Time)장서를 대상으로 실험한 결과 모든 장서에서 가중치기법이 비가중치기법에 비해 성능이 향상되었으며, 타임장서의 고(高) 재현율에서는 최고 25%까지의 정확율이 향상되었음을 보여주고 있다.[12]

통계적인 기법은 가중치색인기법에 사용되는 대표적인 기법이다. 통계적 기법에서 가장 널리 사용되는 것은 빈도특성에 대한 정보라고 할 수 있으며, 이러한 개념은 1949년 지프(G.K. Zipf)가 펴낸 저서 「Human Behavior and the Principle of Least Effort」에서 처음 찾아볼 수 있다. 그는 이 저서에서 '최소 노력의 법칙(Law of Least Effort)'을 적용시켜 언어는 최대의 정보를 전달하기 위해 최소의 단어수를 이용하기 위한 것이라고 정의했으며 원문내의 단어빈도를 그 빈도수에 따라 순위를 매기면 단어의 출현특성이 상수에 의해 표현될 수 있음을 발견하고 다음과 같은 공식을 제시했다.[13]

$$\text{Frequency} \times \text{Rank} = \text{Constant} \qquad 1)$$

그러나 이 공식은 저(低)빈도용어에는 적용되지 않은 한계가 있는데 하위 순위로 내려갈수록 상수가 점차 변수화 된다는 것이다.

따라서 지프의 제2법칙이 유도되었다.

$$I_1 / I_n = n(n+1)/2 \qquad 2)$$

위의 공식에서 I_1은 빈도가 1인 용어의 총 갯수이며, I_n은 n개의 빈도를 갖는 용어의 총 갯수이다.

부쓰(A.D.Booth)는 2) 번공식에 파라미터(parameter)를 확장시켜 지프의 제2법칙에 대한 수정안을 제시했고, 고프만(W.Goffman)은 지프의 법칙과 부쓰의 수정된 저빈도용어의 법칙을 근거로 하여 고(高)빈도용어로부터 저빈도용어의 특성으로 전환하는 임계지역(critical region)에 위치하는 용어들이 원문의 내용을 가장 잘 표현하는 단일어나 단어군(high content-bearing nature)임을 주장했다.[14]

따라서 전환의 임계점이 용어들의 빈도가 1로 접근하는 곳에 위치한다는 가정하에 2)식의 I_n에 1을 대입하고,

---

10) G.Salton,H.Wu and C.T.Yu. "The Measurement of Term Importance," Journal of the American Society for Information Science, 32 (Jan. 1981). P. 176.

11) K.Sparck Jones. Op. cit., P. 412.

12) G.Salton and C.S.Yang. Op. cit., P. 353.

13) D.B.Cleveland and D.Cleveland. Op. cit., P. 149.

14) Ibid., P. 149.

$$(I^1/1 = n (n+1) / 2$$

이로부터 n값을 산출하는 공식이 파오(M. L.Pao)에 의해 정형화 되었다.

$$n = \frac{-1+\sqrt{1+8I_1}}{2} \qquad 3)$$

1957년 룬(H.P.Luhn)은 지프의 법칙을 그래프로 나타내고, 빈도를 기준으로 하여 문헌내용의 식별자로서의 용어의 유용성을 나타내는 RP(Resolving Power)를 그래프로 그렸다.

룬은 문헌빈도의 함수로써 고빈도용어나 저빈도용어가 중간빈도용어에 비해 유용성이 떨어진다고 주장하고 상한선과 하한선을 정해 고빈도용어와 저빈도용어를 제거한 후 그 나머지 용어를 색인목적에 사용토록 제안했다.

이 RP에서의 문제점은 상한선과 하한선을 정하는 절대빈도를 정하기가 어렵고 고빈도와 저빈도용어의 제거는 반대로 재현율과 정확율의 손실을 가져올 수도 있다는 것이다.[15]

용어의 출현빈도로 총장서빈도(Total Collection Frequency) Fk와 문헌빈도(Document Frequency) Bk를 얻을 수 있으며 이 두가지가 가중치기법의 기본개념이 된다.[16]

$$Fk = \sum_{i=1}^{N} f_iK \quad (f_iK는\ i번째\ 문헌에서의\ 용어K의\ 출현빈도)$$

$$Bk = \sum_{i=1}^{N} b_iK \quad (f_iK \geq 1이면\ b_iK=1이고\ f_iK=\emptyset이면\ b_iK=\emptyset)$$

1965년 다메로우(F.J.Damerau)는 단어빈도를 사용하는 경우 색인어를 선정하는 절대적 기준을 정하기가 어렵다는 데 착안하여, 상대빈도(Relative Occurrence Frequency)를 측정하여 색인어를 선정하는 방법을 제시했으나[17] 이 방법 역시 실제 적용에 어려움이 있어 많이 사용되지 않고 있다.

1970년대에 들어서면서, 장서나 특정문헌내에서의 용어의 가치를 측정하여 단순히 색인의 목적에만 적용되던 이론에 검색의 목적을 함께 고려하게 되었다. 다시말해, 문헌을 구분하는데 유용한 용어를 색인어로 선정함으로써, 검색시 적합항목을 구분할 수 있는 새로운 이론과 기법이 개발되었는데, 그 대표적인 것이 스파크존스의 역문헌 빈도론(1972년)과 샐튼의 문헌분리가이론(1973년)으로, 본 장의 2, 3절에서 자세히 설명하기로 한다.

전술한 이론 및 기법들은 소급적 탐색에 적용되는 기법들로써 특정질의에 대한 적합하고 비적합한 문헌들에 출현하는 용어들 간의 구분이 없이, 비적합문헌에서 출현하는 용어들도 적합문헌에서 출현하는 용어들과 동일한 가중치를 얻게된다. 이러한 혼란을 배제하기 위해 이용자 지향적인 이론들이 개발되었으며, 이러한 이론들은 앞선 탐색에서의 결과를 기준으로 하여 적합성 여부를 판정, 그 측정치를 함께 사용하여 SDI 및 반복탐색에 적용이 된다. 로버트슨(S.E.Robertson)은 검색시스템에 제시되는 어떤 질문과도 관계없이 용어가 문헌 및 장서내에서 얼마나 특정한가를 나타내는 '용어 특

15) G.Salton,H.Wu and C.T. Yu. Op.cit., P. 177.

16) Ibid., P. 178.

17) F.J.Damerau. "An Experiment in Automatic Indexing," American Documentation, 16 (Oct. 1965). PP. 238 - 284.

정성(Term Specificity)'과 특정질문에 얼마나 특정한가하는 '용어 대 질의 특정성(Term-Question Specificity)'을 나누어 설명했는데,[18] 이로써 소급적 탐색에 적용되는 기법과 SDI 및 반복탐색에 적용되는 기법 간의 구분은 명확히 된다.

'용어 대 질의 특정성'을 근거로 한 이론들의 기본 개념은 최상의 색인어는 특정 질의에 대해 적합한 문헌에 출현하는 경향이 있다는 가설에서 출발한다. 1960년 마론(M.E. Maron)과 쿤(J.L.Kuhns)에 의한 확률색인기법(Probabilistic Indexing)은 검색문헌이 이용자를 충족시키는 확률에 따라 출력문헌의 순위를 매기기 위한 색인정보를 문헌검색시스템에 사용하도록 하는 것으로,[19] 장서의 전체를 아는 것은 불가능하며, 불확정한 영역에서의 디스크립터(descriptor)는 한정된 확률로만 대응하는 주제 영역에 할당되는 가능성이 있다.[20] 고 하였다. 즉, 이 이론은 색인 단계에서의 결정을 최적화하기 위한 것이다.

1975년 하터(S.P.Harter)는 동일문헌에 함께 군집하여 출현하는 용어가 색인어로 유용하다는 가설[21] 하에 포아슨(Poisson) 모델을 적용하여 용어의 출현분포에 대해 확률론적으로 모델화한 북스타인과 스완슨의 이론[22] 을 발전시켜 2-포아슨분포모델을 제시했다. 2-포아슨분포모델은 용어의 종류에는 문헌내의 빈도분포가 포아슨인 비전문어(nonspecialty word)와 전문어(specialty word)가 있는데, 각 전문어에 대해 문헌장서는 두개의 동질의 부집단으로 구성된다고 가정하고 각 부집단에서 문헌내의 빈도분포가 포아슨이므로 전체장서에는 두개의 포아슨이 있다는 것으로, 적절한 색인어(전문어)를 식별하고 문헌내의 빈도와 2-포아슨 분포의 피라미터들을 근거로 각 문헌에 전문어를 할당하는 결정기준을 제시하기 위한 것이다.[23]

1977년 마론과 쿠퍼(W.S.Cooper)는 용어가중시스템의 형태를 연구하기 위해 유용성이론(Utility Theory)을 제시했는데[24] 이 이론에서는 색인어휘가 주어져 있고, 각각의 문헌에 어떤 색인어를 할당할지의 여부를 '문헌 대 색인어'로 결정하고 이때 결정된 결과는 시스템의 수명이 지속되는 동안 주어진 단일어로 된 질의어에 대해 문헌이 검색되고 검색되지 않은 누적된 유용성에 따라 평가된다.[25]

또한, 1975년에는 바클라(J.K.Barkla)와 밀러(W.L.Miller)가 질의어는 적합문헌뿐 아니라 장서내의 모든 문헌들에 용어분포를 고려하여 가중치를 할당해야 한다는 가정하에 스

18) S.E.Robertson. Op. cit., P. 41.

19) W.S.Cooper and M.E.Maron. "Foundations of Probabilistic and Utility-Theoretic Indexing," Journal of the Association for Computing Machinery, 25 (Jan. 1987). P. 67.

20) 加藤德義. "自動索人의 動向과 逆設的 接近," 高孝坤譯, 情報管理研究, 14 (1981.12). P.184.

21) Stephen P.Harter. "A Probabilistic Approach to Automatic Keyword Indexing : Part I," Journal of the American Society for Information Science, 26 (July/Aug.1975). P.197.

22) 加藤德義. Op. cit., P. 185.

23) S.E.Robertson. "Theories and Models in Information Retrieval," Journal of Documentation, 33 (June 1977). P.140.

24) R.N.Oddy (et al.) Information Retrieval Research. London : Butterworths,1981. P. 15.

25) S.E.Robertson. "Theories and Models in Information Retrieval," Op. cit., P.138.

파크 존스의 역문헌빈도가중기법의 체계를 논리적으로 확장한 적합가중치(Relevance Weight) 이론을 제시했다.[26]

$W_i = \log N - \log n_i + \log r_i - \log R$

(N : 장서내의 문헌수, n_i ; 용어 i 의 빈도, R ; 질의에 적합한 문헌수, r_i ; 적합문헌내의 용어 i 의 빈도)

이는 TQS (Term Question Specificity) 모델이라고도 하는데, 이 공식을 적용하기 위해서는 일반 탐색상황에서 적합문헌에 속해있는 질의어의 분포가 알려져 있지 않기 때문에 확률적 분포의 견적이 제시되어야만 한다. 이는 SDI 및 반복탐색시스템에서는 선행 탐색에서 그 견적을 얻어낼 수 있으나 소급적 탐색에서는 불가능한 것이다. 이에대해 밀러는 소급적 탐색에서도 이용자나 사서에 의해 측정할 수 있다고 주장했다.

샐튼은 이용자 지향형의 SDI 및 반복탐색 시스템에 적용되는 기법들이 이론적으로는 특정하게 주어진 적절히 체계화된 조건에서 최적이지만, 적합성 관정이 특정질의에 대한 문헌을 대상으로 가능하지 않으면 계산될 수 없다는 점을 지적하고 있다.[27]

2. 스파크 존스의 역문헌빈도론

스파크 존스는 전체 문헌빈도는 낮으면서 특정문헌에서의 출현빈도는 높은 용어가 색인어로써 좋은 용어라고 가정하고, 이런 가정을 근거로 역문헌빈도(Inverse Document Frequency)가중기법을 제시했다.[28]

이 기법은 낮은 문헌빈도를 갖는 용어에 높은 가중치를 할당하는 것으로 정확율 향상에 유용하다.

N개의 문헌으로 구성된 장서 가운데 용어 i 로 색인된 문헌의 갯수 n_i 가 알려져 있고 장서로부터 무작위로 어떤 문헌을 선정할 때, 선정된 문헌이 용어 i 를 포함할 확률은 n_i/N가 되며, 용어 i 에 의해 전달되는 정보의 총량(정보이론적 의미에서 문헌에 대한 메시지(message)로서의 색인어)은 다음과 같이 정의된다.[29]

$$-\log_2 (n_i / N) = \log_2 N - \log_2 n_i \quad 1)$$

이러한 개념을 바탕으로 스파크 존스는 다음의 가중치기법을 제시했다.

$$IDF_v \approx \log_2 N - \log_2 n_i + 1 \quad 2)$$

여기서 근사치는 로그(log)를 취할때 다음으로 높은 정수를 얻기 위한 것이며, '+1' 은 어떤 용어에도 가중치 ∅가 할당되지 않도록 하기 위한 것이다.

위의 용어특정성모델을 순수한 정보이론적 모델로 변형하면

$$IDF_v = \log N - \log n_i + 1 \quad 3)$$

로 표현되며, 이때 로그의 밑수는 고려되지 않는다.

마지막으로, 3) 번 공식에 각 문헌K 마다의

---

26) K.Sparck Jones. "A Performance Yardstick for Test Collections," Journal of Documentation,31 (Dec. 1975) P.267.

27) G.Salton and J.M.Michael. Introduction to Modern Information Retrieval. N.Y.: Mcgraw-Hill, 1983. P. 206.

28) G.Salton,H.Wu and C.T.Yu. Op. cit., P. 178.

29) S.E.Robertson. "Specificity and Weighted Retrieval," Op. cit., P. 42.

용어 i의 출현빈도(f_{iK})인 로칼밸류(local value)[30] 를 곱하여 용어가중을 한다.

$$W_{iK} = f_{iK} \cdot IDF_V \qquad 4)$$

3. 샐튼의 문헌분리가(文獻分離價)이론

용어의 문헌분리가(Term Discrimination Value)가중기법은 문헌들에 용어가 할당될때 발생하는 공간분리의 변화를 측정하는 방법으로 1973년 샐튼에 의해 개발되었다.[31]

문헌분리가이론에서는 좋은 분리자(discriminator)가 색인어로 할당될 경우 문헌간의 유사성이 감소하며, 좋지 않은 분리자가 색인어로 할당되면 문헌간의 유사성이 증가한다.
즉, 좋은 분리자가 색인어로 할당됨으로써 문헌의 공간밀도를 감소시켜 적합문헌의 구분이 수월하다는 것으로 각 용어의 장서내 할당 전과 후의 공간밀도를 측정하여 그 차이에 따라 용어의 순위를 매기게 된다. 이 기법에서 사용되는 문헌 간의 유사성공식[32] 은 코싸인계수(cosine coefficient)로 t차원공간에서의 일반 벡터(vector)로 가정할때 두개의 t차원 문헌 벡터 간의 각을 측정하는 값이며 두개 벡터내의 공통항 갯수가 많아질수록 값이 증가하게 되는데 범위는 양의 벡터항목에서 $\emptyset \sim 1$ 까지이다.[33]

$$SIM(DOC_{K1}, Doc_{K2}) = \frac{\sum_{i=1}^{t} (Term_{K1i} \cdot Term_{K2i})}{\sqrt{\sum_{i=1}^{t} (Term_{K1i})^2 \cdot \sum_{i=1}^{t} (Term_{K2i})^2}}$$

(Term k_i : k번째 문헌내의 i번째 용어의 빈도)

위의 공식을 실제로 적용할때는 문헌공간의 중앙에 위치하며 전 장서를 대표하는 중심문헌(Centroid;C)을 임의로 정하고,[34]

$$C = (1/n\sum_{K=1}^{n} W_{1K}, 1/n\sum_{K=1}^{n} W_{2K}, 1/n\sum_{K=1}^{n} W_{3K} \cdots 1/n \sum_{K=1}^{n} W_{iK})^{35}$$

장서내의 각 문헌과 중심문헌 간의 유사성을 측정하여 더한 후 평균을 낸 값을 AVESIM이라 하며,

$$AVGSIM = 1/n \sum_{K=1}^{n} SIM(C, D_K)$$

장서내에서 용어 i를 모두 제거한 후 AVGSIM과 동일하게 계산된 값을 $AVGSIM_i$라 할때 용어 i의 문헌분리가($DISC_V$)는 다음과 같

30) 로칼밸류 ; $f_{ik} = i/k$ (k는 k번째 문헌내의 총단어출현빈도이며 i는 k번째 문헌내에서의 용어 i의 출현빈도이다.)

31) G.Salton,C.S.Yang C.T.Yu. "A Theory of Term Importance in Automatic Text Analysis," Journal of the American Society for Information Science, 26 (Jan./ Feb.1975). P.34.

32) 유사성공식 ; 문헌간의 유사성을 측정하는 공식에는 Jaccard계수, cosine계수, overlap계수, asymmetric계수등을 사용하는 5개의 공식이 있는데, overlap계수와 assymetric계수는 복잡하여 많이 사용되지 않는다.

33) G.Salton and J.M.Michael. Op. cit.,PP. 202-204

34) G.Salton,C.S.Yang and C.T.Yu. Op.cit., P.34.

35) $1/n \sum_{k=1}^{n} W_{ik}$; k번째 문헌에서 i번째 용어의 가중치 합의 평균

이 정의된다.[36]

$$DISC_V = AVGSIM_i - AVGSIM$$

위의 공식에서 $DISC_V$가 양수면 ($AVGSIM_i > AVGSIM$) 용어 i는 좋은 분리자이며 음수면 좋지않은 분리자로 구분되며 $DISC_V$가 Ø에 가까운 값을 갖는 경우에는 용어 i가 공간밀도에 별다른 영향을 미치지 않는다. 마지막으로 역문헌빈도가중기법에서와 같이 로칼밸류를 곱하여 용어가중을 한다.

$$W_{iK} = f_{iK} \cdot DISC_V$$

4. 선행연구

1972년 스파크 존스는 문헌기술(document description)의 망라성은 그 문헌에 포함되어 있는 용어들의 갯수이며, 용어의 특정성은 그 용어가 속해있는 문헌의 수라고 재정의하고, 어휘에 있어서의 특정성의 최적수준과 색인을 하는데 있어서의 망라성의 최적수준을 모색하기 위해 다음과 같은 실험을 하였다.[37] 대상장서는 크랜필드장서, 전기공학분야의 인스펙(INSPEC)장서, 킨(Keen)장서였으며, 각 장서로부터 문헌당 색인어와 질의당 용어의 평균치로 그 특성을 조사하고, 문헌을 기술하는 평균 용어갯수보다 특정질의어를 포함하는 문헌의 수가 더 많다는 것을 지적, 그 해결책으로 역문헌빈도가중기법을 제시하였으며, 세개의 장서에 적용, 실험한 결과 세개의 장서 모두에서 일반 비가중치기법에 비해 월등히 성능이 향상되었음을 보고했다.

1975년 샐튼은 크랜필드장서, 메들라인장서, 타임장서에 대해 문헌분리가중기법을 적용하여 실험하였다.[38] 그는 용어의 문헌빈도와 문헌분리가의 관계를 관찰하고 색인어로써 적합한 용어의 문헌빈도는 $N/100 < \chi < N/10$ (N; 총 문헌수) 구간이라는 결과를 얻었으며 문헌빈도가 낮은 하위 70%의 용어와 문헌빈도가 높은 상위 4%의 용어들에 대해 전자의 경우, 재현율을 향상시키기 위해 디소러스를 사용한 용어의 계층적 확장 전략을, 후자의 경우에는 정확율을 향상시키기 위해 용어조합 전략을 제안했다. 이러한 실험의 결과 표준 용어빈도가중치기법에 비해 높게는 메들라인장서에서 39%, 낮게는 타임장서에서 17% 성능이 향상되었으며, 문헌분리가가중기법으로 용어의 분리속성이 최적일때 고(高)재현율에서는 0.7에서 0.9로, 중간 재현율에서는 0.4에서 0.6으로, 저(低)재현율에서는 0.2에서 0.3으로 각각의 평균 재현율도 향상되었다.

한편, 역문헌빈도가중기법과 문헌분리가가중기법의 성능을 상호 비교한 실험들도 행해졌는데, 1973년 샐튼과 양(C.S.Yang)은 위와 동일한 장서를 대상으로 하여 장서의존적인 색인기법의 성능을 평가하였다. 평가방법은 가중치가 낮은 용어를 제거하는 'Cut'처리, 단어빈도가중치기법과 역문헌빈도가중기법 및 문헌분리가가중기법에 의한 가중치를 곱하는 'Mult'처리, 두가지 방법을 통합한 'Cut + Mult'처

36) G.Salton and J.M.Michael. Op. cit., PP. 66-67.

37) K.Sparck Jones. "A Statistical Interpretation of Term Specificity and its Application in Retrieval," Journal of Documentation, 28 (March 1972). P. 13.

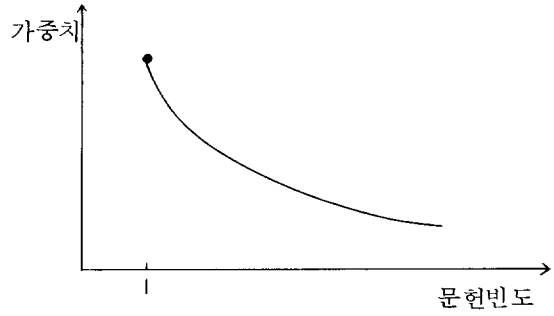
38) G.Salton,C.S.Yang and C.T.Yu. Op. cit., PP. 33-44.

리의 세가지 가중치처리방법과 단어빈도,역문헌빈도, 문헌분리가 가중치시스템을 혼합하여 사용하였다.[39] 전체적으로는 비가중치기법보다는 가중치기법이 모든 장서에서 유용하고 역문헌빈도가중기법과 문헌분리가가중기법과 모두 검색 결과 개선에 효율적이었으며 저재현율과 고정확률이 요구되는 경우에는 역문헌빈도가중기법이, 중간정도의 재현율이 요구되는 경우에는 문헌분리가가중기법이 우수한 것으로 나타났다. 장서별로는 크랜필드장서와 타임장서에서는 역문헌빈도 'Cut + Mult' 처리가 가장 우수하고, 저빈도용어가 많은 메들라인장서에서는 역문헌빈도 'Cut' 처리가 우수하게 나타났으며 가장 일반적인 방법은 문헌분리가 'Cut' 처리와 역문헌빈도 'Mult' 처리를 함께 사용하는 것이라고 제시하였다.

또한 최근에 샐튼등은 역문헌빈도가중기법 및 문헌분리가가중기법을 이용자 지향형의 용어적합성기법과 비교하여 성능 향상을 보고하였다.[40] 대상장서는 크랜필드장서와 메들라인장서였으며 각 장서에 24개의 탐색질의를 사용하여 실험한 결과 역문헌빈도가중기법, 문헌분리가가중기법, 용어적합성기법의 성능 향상이 크랜필드장서에서는 27.6 % ; 16 % ; 46.2 %로, 메들라인장서에서는 14.9 % ; 8 % ; 74.1 %로 밝혀졌다. 즉, 두개의 장서 모두에서 문헌분리가가중기법보다는 역문헌빈도가중기법의 성능이 우수했으며 용어적합성기법의 성능 향상은 괄목할 만한 것이었다.

1984년 김현희는 룬의 RP (분석력치)와 용어의 편포(skewness)를 함께 사용한 가중치기법을 제시하고, 역문헌빈도가중기법, 문헌분리가가중기법, 하터의 2-포아슨가중치기법 등과 비교하여 실험하였다.[41] 대상장서는 경구적 과식(entral hyperalimention) 주제에 대한 1333개 문헌의 초록과 표제로 구성되고 22개의 질의를 사용하였으며, 적합문헌은 정보전문가에 의해 수작업으로 검색하였다. 실험결과는 비교적 간단한 과정으로 가중치를 구할 수 있는 역문헌빈도가중기법이 가장 효율적이며, 이론적으로는 완벽하지만 상당히 복잡한 과정이 요구되는 문헌분리가가중기법의 효율이 가장 떨어진다는 것이었다.

이상에서와 같이 역문헌빈도가중기법과 문헌분리가가중기법에 대한 이론 및 선행연구를 살펴본 결과 다음과 같은 차이점이 있었다 가장 큰 차이점은 각 기법에서 제시하는 적합한 용어빈도 특성 간의 차이점으로 다음의 그래프에 잘 나타나 있다.[42]



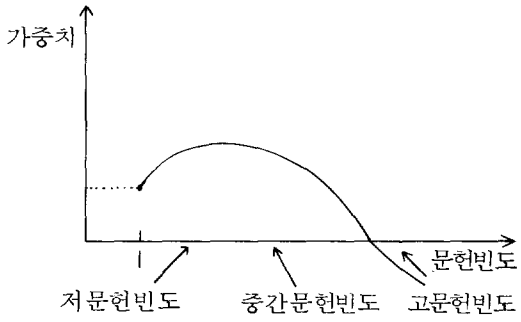
(그림-1) 역문헌빈도가중기법에서의 용어가중치 분포

39) G.Salton and C.S.Yang. Op. cit., PP. 351-372.

40) G.Salton, H.Wu and C.T.Yu. Op. cit., PP. 175-186

41) 김현희. "An Investigation of Automatic Term Weighting Techniques," 情報管會學會誌, 1(1984). PP. 43-62

42) R.N.oddy (et al). Op. cit., P. 11, P.13.



(그림 - 2) 문헌분리가가중기법에서의 용어가중치 분포

위의 그래프에서 역문헌빈도가중기법의 경우에는 저문헌빈도용어의 가중치가 높으며 음의 값은 나타나지 않는 반면, 문헌분리가가중기법에서는 중간문헌빈도용어에 높은 가중치가 할당되며 고문헌빈도용어는 가장 색인어로서의 가치가 없는 것으로 음의 값을 갖는다는 것을 알수있다. 여기서 색인기법의 환경에 적합한 용어의 특성을 규정지을 수 있다. 역문헌빈도가중기법은 낮은 문헌빈도를 갖는 특정용어가 검색의 기준으로 특정성이 요구될때 적합하며, 문헌분리가가중기법은 중간문헌빈도의 용어들이 검색의 기준이 되며 특정성보다는 망라성이 요구되는 검색시스템에 적합할 것이다.

마지막으로 문헌분리가가중기법은 역문헌빈도가중기법에 비해 계산과정이 복잡하기 때문에 실제 실험환경에서 상당한 결점으로 지적될 수도 있다.

Ⅲ. 연구방법

1. 실험환경

수작업 색인은 비용이 많이 든다는 것 외에도 일관성 문제에서의 결정적인 결점을 수반하게 되며, 정보의 출판과 그 활용을 위한 색인 및 초록작업 간의 시간지연에서도 문제가 제시된다. 따라서 일관성 및 시간지연의 해결책으로 등장한 자동색인[43]에서의 쟁점은 시스템 자체의 성능과 신속한 처리능력이며 신속성은 비용과도 직접적인 관련이 있다.

고프만과 뉴웰(V.A.Newil)은 시스템 성능평가를 위해 시스템의 성능(U:effectiveness)과 시스템의 실행비행(V:efficiency 또는 효율)의 두 변수를 사용하여 E=F(U, V)라는 공식을 제시한 바 있다.[44]

본 연구의 실험에서는 시스템을 신속하고 효율적으로 실행시키기 위해 작성된 프로그램들과 16비트 컴퓨터의 응용 패키지들을 함께 사용하였으며 구체적으로는 다음과 같다.

사용 컴퓨터는 트라이젬 88+인 16비트 AT로 두개의 드라이브(drive)와 20MB 하드디스크가 내장되어 있어 방대한 양을 신속히 처리할 수 있었다. 프로그램은 화일 입출력과 변수영역 관리가 수월한 C언어로 작성하였으며 사용 패키지에는 C프로그램을 컴파일(Compile)하기 위한 MSC series(Microsoft Corp.C-ompiler)와 화일 작성 및 편집을 위한 보석글 시스템, 결과의 소트(somt)를 위한 보석글유틸리티(utility) 및 결과 분석과정에서 통계처리에 사용된 SAS (Statistical Analysis System)등이 있다.

43) 자동색인 ; 레코드와 정보의 탐색질의에 대해 용어 및 내용식별자를 자동 할당하는 것으로,본 연구에서는 소급탐색에 적용되는 통계적 가중치 색인기법만으로 한정한다.

44) Eva L.Kiewitt.Evaluating Information Refrieval System.
Westport : Greenwood press, 1979. P. 92.

2. 입력데이타와 입력과정

입력대상 데이터는 초록지인 CA(Chemical Abstracts)와 LISA(Library & Information Science Abstracts)가운데 1987년도 3월호에서 무작위로 추출한 각 201개의 초록이며 표제를 제외한 초록만을 입력하였다.

표본 유형을 초록으로 선정한 이유는 여러실험 결과에서 밝혀진 바와 같이[45] 전문의 처리가 거의 불가능한 상황에서 표제보다는 초록을 대상으로 한 결과가 월등하기 때문이다.

CA는 화학 및 화공에 관한 주간 초록지로 색인에는 저자명색인, 일반주제명색인, 화학물질명색인, 구조식명색인, 특허색인 등이 있다.

본 실험에서는 CA의 한 분야인 약학(pharmacology) 분야만을 대상으로 한다.

또한 LISA는 도서관·정보학분야에서 가장 빈번히 사용되는 월간초록지로 인명색인과 주제명색인이 함께 수록되어 있다.

입력은 다음과 같은 몇단계에 걸쳐 행해졌다.

첫째, 불용어를 제거한다. 영어에는 전체단어의 약 40 ~ 50 %에 해당하는 고빈도용어가 있는데 이런 용어들은 문헌의 내용을 식별하는데 사용될 수 없는 불용어(stop word)들로 약 250개 정도의 용어들이 포함된다.[46] 불용어에는 전치사, 관사, 접속사, 대명사 등과 내용상 중요치 않은 명사, 형용사, 부사 및 동사와 그 변화형등이 포함되며[47] 본 논문에서 사용된 불용어 리스트(list)는 부록에 수록되어 있다.(부록Ⅰ)

둘째, 복합어를 분리한다. 복합어는 검색의 정확율 향상에 유용한 탐색전략이기도 하나 본 실험에서는 총장서빈도의 빈약으로 분리를 해도 무방한 복합어들을 분리, 입력하였으며 다음과 같은 몇몇 복합어는 의미상 그대로 입력하였다.

예) anti-inflammatory, half-life, inter-library, menu-driven, anline 등,

또한 하나의 단어로 처리하기 불가능한 복합화학명도 분리, 입력하였다.

예) L-histidinol-FUra → L / histidinol / FUra (14β-hydroxybufa-4,20,22,- trienolid 3β-yl) oxycarbany → beta / hydroxybufa / trienalid / beta / yl / oxycarbony

세째, 복수명사를 단수화한다.

네째, 동의어처리를 한다. 동의어 처리는 기존 수작업색인의 '보라'참조를 기준으로 하였으며, 화학명의 약어는 중복 입력하였다.

다섯째, 특수문자는 소리나는 대로 풀어 입력하였다.

예) α→alpha , β→beta

입력데이터에 대한 구체적 사항은 (표-1)에 나타나 있다.

	CA	LISA
주제분야	약 학	도서관·정보학
문헌수	201	201
단어의 종류	2,614	2,017
단어의 총출현 횟수	9,083	7,119
문헌당 단어의 평균갯수	45.1	30.4

(표-1) 집단간의 입력데이터 비교

45) G.Salton and M.E.Lesk. "Computer Evaluation of Indexing and Test processing," Journal of the Association for Computing Machinery, 15 (Jan.1968). P. 630.

46) G.Salton and J.M. Michael. Op. cit., P. 71

47) 司空哲. 情報檢索論. 서울 : 亞細亞文化社, 1983. P. 102.

3. 입력데이터의 처리과정

원시화일(source file)의 데이터는 출현순에 따라 임의로 주어진 문헌번호와 함께 입력했으며, 이 화일을 대상으로 작성된 프로그램들을 단계별로 실행하여 출력물을 얻는다.
각 프로그램은 부록Ⅱ에 수록되어 있으며 적용순서는 다음과 같다. 괄호안은 프로그램명이며 →다음은 출력화일명이다.

1) 원시화일을 단어들의 알파벳순으로 소트한다.→ data. ca. data. ls
2) 소트된 단어들에 대한 일반정보를 얻는다. (CMP. C) → ca. id. lisa. id

이때 총장서빈도 1인 것을 제거하며 일반정보에는 총장서빈도, 문헌빈도, 단어의 중심값[48] (centroid value)이 포함된다.

3) 201개의 각 문헌에 대해 출현단어의 정보를 얻는다.(NUMB。C) → data。ca2 data. ls2
4) *.id화일을 대상으로 하여 각 단어에 대한 역문헌빈도가 (IDFv)를 구한다. (IDF·c) → IDF.ca, IDF.Is.
5) *.* 2화일을 대상으로 문헌분리가(DISCv)를 구한다. (DISC.c) → DISC.ca,DISC.Is
6) 지금까지의 출력을 하나의 화일에 저장한다.(PFT.c) → CA all. LISA all
7) 전체적인 빈도분포를 총장서빈도와 문헌빈도 각각에 대해 구한다. (COUNT.c) →Stat Is.
8) 각 기법을 적용하여 얻은 가중치가 높은 순으로 일정비율을 선정하고 수작업색인과의 일치율을 계산한다.(EVL.c) → Evl.1,2,3,4

본 실험에서 로칼밸류는 각 문헌마다의 단어출현빈도 대신 전체장서에서의 각 단어의 문헌빈도 대 총장서빈도의 비율을 사용하였는데, 원래 기법에서 사용되는 로칼밸류는 주어진 문헌에 대한 단어빈도의 비율로써 $F_{ik} = i/k$ (k는 k번째 문헌내의 총단어출현빈도, i는 단어 i의 k문헌내 출현빈도)로 정의된다.

이를 장서에 대입시키면 총장서빈도를 문헌당 단어수로 나누고 다시 문헌빈도로 나눈 것으로 대치된다. 이때 모든 단어에서 문헌당 단어수는 상수로 CA에서는 45개 LISA에서는 30개 개로 일정하여 단어의 가중치에 변화를 주는 변수로 작용할 수 없으므로 적용하지 않는다.

따라서 각 단어의 로칼밸류는 '총장서빈도/문헌빈도'로 사용한다.

4. 출력

출력물은 CA집단과 LISA집단 각각에 대해 다음과 같은 화일로 구성된다.

1) 단어에 대한 일반정보로 구성된 화일. (CA.all, LISA.all)

단어번호	단어	총장서빈도	문헌빈도	단어의 중심값	역문헌분리가	문헌분리가	수작업색인과의 일치여부
%4d	%27s	%3d	%7.5f	%7.5f	%17.5f	%10.3f	%s

48) 단어의 중심값; 단어의 총장서빈도를 장서내 문헌수로 나누어 얻는 단어의 문헌당 평균 빈도를 말한다.
$C_V = 1/N \sum_{j=1}^{N} F_{ij}$ (F_{ij}는 j번째 문헌에서 i번째 단어의 빈도)

NO.	WORD	COL. FREQ.	DOC. FREQ.	CENT. VALUE	IDFv	DISCv	
1	aacr2*	2	1	0.00995	17.30210	0.00006220	Y
2	ability	5	4	0.02488	6.65105	0.00009390	N
3	abstract	6	6	0.02985	6.06609	0.00021340	N
4	academic	11	8	0.05473	5.65105	0.00020020	Y
5	academy	6	4	0.02985	6.65105	0.00010020	Y
6	access	22	16	0.10945	4.65105	0.00020680	N
7	accessible	2	2	0.00995	7.65105	0.00002900	N
8	achivement	7	5	0.03483	6.32912	0.00014590	Y
9	acquisition	15	15	0.07463	4.74416	0.00010370	Y
10	act	7	4	0.03483	6.65105	0.00019970	N
11	action	3	2	0.01493	7.65105	0.00005050	N
12	active	5	4	0.02488	6.65105	0.00003050	N
13	activity	16	13	0.07960	4.95061	0.00019300	N
14	adequate	3	2	0.01493	7.65105	0.00007210	N
15	administrative	3	3	0.01493	7.06609	0.00003290	Y
16	administrator	2	1	0.00995	17.30210	0.00022300	Y
17	adult	8	3	0.03980	14.13218	0.00097180	Y
18	advance	2	2	0.00995	7.65105	0.00004320	N
19	advantage	5	5	0.02488	6.32912	0.00001260	N
20	advice	3	2	0.01493	7.65105	0.00008210	N
21	advisory	3	1	0.01493	25.95315	0.00014700	N
22	aesthetic	2	1	0.00995	17.30210	0.00002720	N
23	affair	2	2	0.00995	7.65105	0.00002000	N
24	africa	2	2	0.00995	7.65105	0.00001990	Y
25	african	3	2	0.01493	7.65105	0.00008130	N
26	age	6	5	0.02985	6.32912	0.00006270	N
27	agency	3	3	0.01493	7.06609	0.00004430	N
28	agreement	2	2	0.00995	7.65105	0.00003290	N
29	agricultural	6	4	0.02985	6.65105	0.00017210	N
30	aim	7	6	0.03483	6.06609	0.00006280	N

이하생 략

(표 - 2) 출력데이터 - 1

2) 총장서빈도와 문헌빈도의 분포사항(Stat ca, Stat.ls)

빈도	총장서빈도건수	문헌빈도건수
%3d	% 4d	% 4d

```
*** FREQ. POSTING ***
          CF     DF
-------------------
      1 1342   1754
      2  489    359
      3  176    140
      4  147     86
      5  100     44
      6   59     46
      7   58     28
      8   39     27
      9   31     24
     10   24     19
     11   17     17
     12   16      9
     13    9      9
     14    8      5
     15   14      3
                  이하생략
```

(표-3) 출력데이터-2

3) 기존 수작업색인과의 일치율을 근거로 한 성능측정 결과에 대한 화일(Ev1 1,2, 3,4)로 적합색인어와 선정된 색인어의 갯수와 비율 및 정확율, 재현율, 제외율, 잡음율 등으로 구성된다.

```
*** PERFORMANCE EVALUATION OF ----v IN ---- ***

Total terms                                  : --
Total relevance index terms                  : --  --%
Total selected index terms by ----           : --  --%
Selected relevance index terms by ----       : --
Not selected relevance index terms by ----   : --
Selected non-relevance index terms by ----   : --

PRECISION ratio :  --%
RECALL ratio    :  --%
OMISSION ratio  :  --%
NOISE ratio     :  --%
```

(표-4) 출력데이터-3

Ⅳ. 실험결과 분석

본 장에서는 두집단 간의 용어특성을 비교하고 각 집단에 대해 적용된 두개의 자동색인기법의 성능을 평가함으로써 서로 다른 주제를 갖는 집단에 적합한 색인기법을 제시하고 용어특성과의 관계를 밝히기 위해 앞서 얻은 출력물을 다음과 같이 분석하였다.

1. 사용변수 및 변수 간의 상관관계 분석

분석에서 사용된 변수는 총장서빈도(CF),문헌빈도(DF), 역문헌빈도가(IDFv), 문헌분리가(DISCv)의 네가지이며 각 변수들 간의 상관계수(Correlation Coefficients : cc)는 다음과 같다.

	CF	DF	IDFv	DISCv
CF	1.00000	0.91775	-0.08809	-0.71116
EF	0.91775	1.00000	-0.30367	-0.58116
IDFv	-0.08809	-0.30367	1.00000	0.23252
DISCv	-0.71116	-0.58116	0.23252	1.00000

(표-5) CA집단의 상관계수

	CF	DF	IDFv	DISCv
CF	1.00000	0.93973	-0.13414	-0.80197
DF	0.93973	1.00000	-0.28463	-0.60330
IDFv	-0.13414	-0.28463	1.00000	0.04991
DISCv	-0.80197	-0.60330	0.04991	1.00000

(표-6) LISA집단의 상관계수

위의 표에서 CF와 DF, IDFv와 DISCv의 관계를 제외한 각 변수들 간의 관계는 감소함수관계임을 알 수 있다. 다음에서 각 변수들

간의 상관관계를 분석한다.

1) CF와 DF의 관계

두집단 모두에서 상관계수(기울기)는 CC> 0.9로 거의 정비례 관계를 형성한다. 따라서 총장서빈도가 높아질수록 문헌빈도도 높아진다.

집단별 비교에서는 LISA집단의 상관계수가 더 크다는 것을 알 수 있는데(0.93973>0.91775) 이는 CA집단에 포함된 용어들이 특정 문헌에 더 군집되어 있음을 나타내는 것으로 CA집단의 용어특성은 총장서빈도가 높고 문헌빈도가 낮은 단어들이 색인어로 유용하다는 이론에 근거한 역문헌빈도가중기법[49]에 더 적합할 것으로 보인다. 반면에 LISA집단의 상관계수가 낮다는 것은 저문헌빈도나 고문헌빈도보다는 중간문헌빈도를 갖는 단어들이 CA집단에 비해 많이 포함되어 있다는 것으로 LISA집단의 용어특성은 문헌분리가가중기법에 더 적합할 것으로 기대된다.

2) CF와 IDFv의 관계

CF와 IDFv는 특정한 관계에 있다고 할 수 없는데 이는 로칼밸류의 적용에 따른 것이며 로칼밸류를 적용하지 않은 경우에는 거의 반비례관계(-0.81902, -0.76755)를 형성한다. 즉 CF가 높아질수록 DF가 높아지고 DF가 높아진다는 것은 문헌빈도의 역함수를 취하는 역문헌빈도가중기법에서 IDFv가 낮아짐을 의미한다.

3) DF와 IDFv의 관계

CF와 IDFv에서의 관계와 동일하게 분석되며 두집단 간의 차이가 줄어들고 상관계수의 절대치(|cc|)가 높아진 것은 CF보다는 DF의 분포가 전체적으로 군집되며 저문헌빈도를 더 많이 형성하게 되기 때문이다. 또한 DF는 역문헌빈도가중기법에 직접 사용되는 변수라는 점에서 상관계수의 절대치가 높아질 수 밖에 없다.

4) CF와 DISCv의 관계

문헌분리가가중기법은 중심문헌을 구성하여 각 문헌과의 공간밀도를 측정하는 기법으로 중심문헌은 총장서빈도를 근거로 설정된다.

따라서 CF는 IDFv와 보다는 DISCv와 더 밀접한 관계를 갖는다.

5) DF와 DISCv의 관계

두집단 간의 차이가 CF와 DISCv의 상관관계에서 보다 줄어든 것은 DF와 IDFv의 관계에서와 동일하게 분석되며 DISCv에서는 IDFv에서와는 달리 CF와의 관계가 더 밀접함을 알 수 있다.

또한 두변수사이의 상관계수 절대치는 그다지 높지 않게 나타났으므로, DF가 높을수록 DISCv가 낮아진다고 할 수 없는데 문헌분리가가중기법에서는 중간문헌빈도를 갖는 용어가 가장 유용한 색인어로 높은 가중치를 갖게 되기 때문이다.

6) IDFv와 DISCv의 관계

두변수 간에는 거의 관계가 없는 것으로 나타났으므로 IDFv가 높은 단어가 반드시 높은 DISCv를 갖는다고 할 수는 없다. 따라서 두기법 간의 성능에 차이가 있을 것으로 보이며 CA집단에서의 상관계수가 더 높은 것으로 보아 LISA집단에서 두기법 간의 성능차이가 더 클 것으로 생각되는데 본 장의 4절에서 자세히 살펴보기로 한다.

---

49) G.Salton,H.Wu and C.T.Yu. Op. cit., P. 178.

2. 주제별 용어특성[50] 비교

2.1 총장서빈도(Total Collection Frequency)분포 비교

분석은 '1'의 총장서빈도를 갖는 것과 저(低)총장서빈도, 중간총장서빈도, 고(高)총장서빈도로 구간을 나누어 수행하였으며 두집단간의 크기와 빈도분포 범위가 다르므로 각 빈도별로 전체에 대한 백분율이 군집해 있는 정도에 따라 구간을 나누는 기준을 설정하였다.

두집단의 총장서빈도 분포사항은 거의 유사하게 나타났으며 표-7에 각 구간마다의 건수와 백분율이 나와있다.(상세한 데이터는 부록Ⅲ-1, 2 참조)

CF	CA	LISA	CF
1	1,342(51.3%)	1,077(53.4%)	1
2-7	1,029(39.4%)	717(35.5%)	2-6
8-22	201(7.7%)	180(9.0%)	7-19
23-133	42(1.6%)	43(2.1%)	20-344
합	2,614	2,017	

(표-7) 두집단의 총장서빈도 분포 비교

위의 표에서 구간별로 비교해 본 결과 빈도 1을 찾는 용어들은 LISA집단에서 더 높은 비율로 나타났으나 빈도 1을 포함한 총장서빈도가 낮은 용어들의 비율은 CA집단이 높았으며 중간 총장서빈도와 높은 총장서빈도를 갖는 용어들의 비율은 LISA집단에서 더 높게 나타났다.

따라서 CA집단에는 희귀어(낮은 총장서빈도를 갖는 용어)가 LISA집단에는 일반어(높은 총장서빈도를 갖는 용어)가 상대적으로 많은 것을 알 수 있다. 또한 집단 간의 비교에서는 티검증(t-test) 결과 유의한 차이가 없는 것으로 나타났는데 유의수준 $\alpha = 0.05$[51]하에서 P값은 0.8380으로 이는 (0.05<0.8380) 집단 간에 유의한 차이가 없음을 의미하는 것이다.

2.2. 문헌빈도(Document Frequency) 분포 비교

문헌빈도는 특정용어들이 전체 문헌에 어떻게 분포되어 있는가를 측정하는 값으로 색인환경에서는 총장서빈도보다 더 유용하게 사용되므로 문헌빈도의 분포는 색인기법에서 할당하는 가중치와 밀접한 관계가 있다. 문헌빈도는 총장서빈도가 1인 용어들을 제거한 후 계수되었으며 제거된 용어들은 CA집단에서 1,342개(51.3%), LISA집단에서 1,077개(53.4%)였다.

분석은 1의 문헌빈도를 갖는 것과 샐튼이 가장 적합한 것으로 제시한 $N/100 \leq X \leq N/10$ (N은 총문헌수)[52]의 문헌빈도 구간 및 그 이상의 문헌빈도의 세구간으로 나누어 수행하였다.(부록Ⅳ-1,2 참조) 다음은 각 구간마다의 건수와 백분율로 표시된 문헌빈도 분포사항이다.

---

50) 용어특성 ; 통계적 개념으로 총장서빈도와 문헌빈도로 정의되는 용어들의 빈도분포 특성을 의미한다. 또한 총장서빈도를 기준으로 하여 희귀어와 일반어를 구분하며 문헌빈도를 기준으로 특정어, 준특정어, 비특정어의 세가지로 구분하기로 한다.

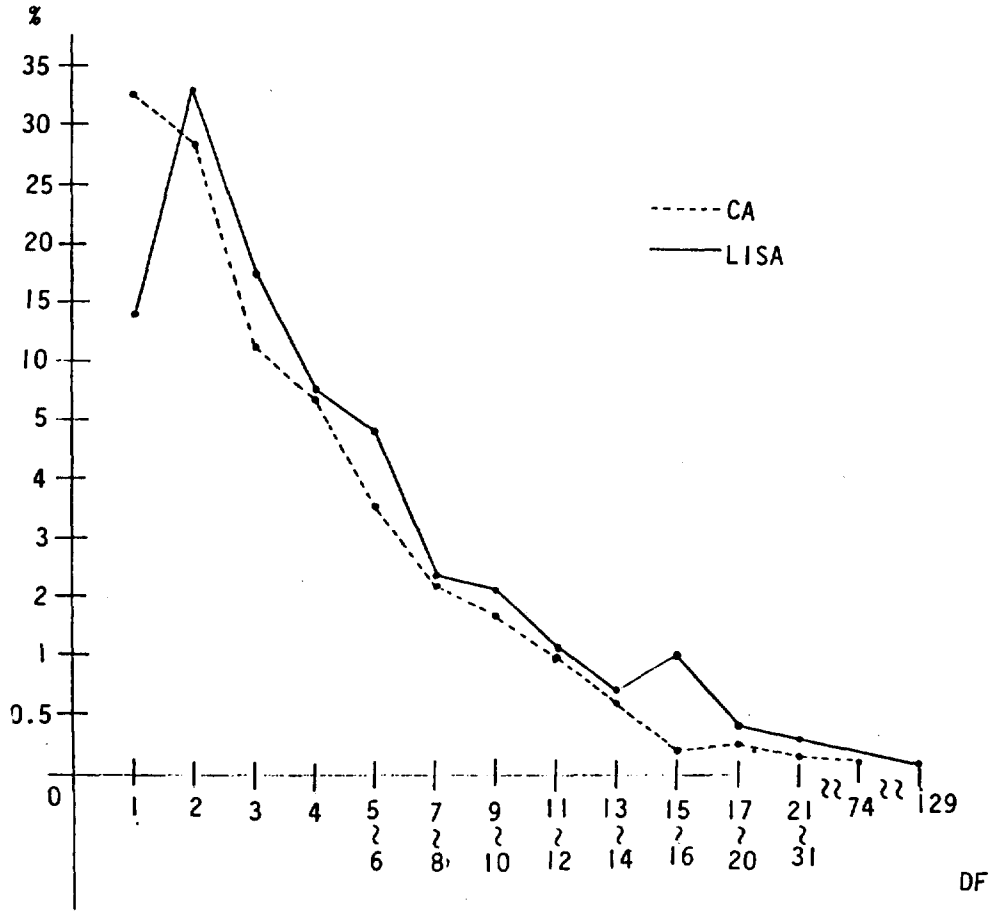
51) 유의수준 ; 일반적으로 유의수준에 사용되는 수치는 $\alpha = 0.1$ (10%), $\alpha = 0.05$ (5%), $\alpha = 0.01$ (1%)의 세가지이며 유의수준이 낮을 수록 결과가 유의하다. 또한 P값이 유의수준 보다 낮을 수록 결과의 신빙성은 높아진다.

52) G.Salton,C.S.Yang and C.T.Yu., Op. cit, P. 37.

DF	CA	LISA
1	412(32.4%)	132 (14 %)
2-20 (N/100≦χ≦N/10)	844(66.3%)	780 (83 %)
21-129 (N/10<χ)	16(1.3%)	28 (3 %)
합	1,272	940
평 균	3.9	4.8

(표-8) 두집단의 문헌빈도 분포 비교

위의 각 구간을 임의로 특정어, 준특성어, 비특정어로 나누어 본다면 CA집단에서는 특정어가, LISA집단에는 특정어 및 비특정어가 많이 포함되어 있음을 알 수 있다. 또한 문헌빈도에 따라 두집단 간에 차이가 있다는 것은 서로 다른 주제를 갖는 집단마다 적합한 색인기법의 구분이 가능하다는 것을 의미한다.



(그림-3) 두집단의 문헌빈도 분포 비교

그림에서와 같이 두집단 간에는 확실한 차이점이 있는데 문헌빈도 1을 제외한 모든 구간에서 LISA의 비율이 높게 나타나 있다. 실제로 t검증의 결과 유의수준 $\alpha = 0.05$ 하에서 P값은 0.00005로 (0.05>0.00005) 앞서 밝힌 것과 같이 매우 유의한 차이가 있었다.

총장서빈도와 문헌빈도를 대상으로 한 용어특성에 비교 결과 총장서빈도 분포로는 두집단을 구분할 만한 큰 차이점이 없었으나 문헌빈도 분포로도 두집단을 구분할 수 있었다. 즉, CA집단에는 문헌빈도 1을 갖는 특정어가 18%, LISA집단에서는 준특정어가 23% 더 많이 나타났는데 CA집단은 LISA집단에 비해 용어들이 특정문헌에 더 군집되어 있으므로 저문헌빈도를 많이 형성하기 때문이다.

따라서 용어의 통계적 특성에 근거한 로칼밸류(총장서빈도/문헌빈도)는 전반적으로 CA집단에서 더 높게 주어진다.

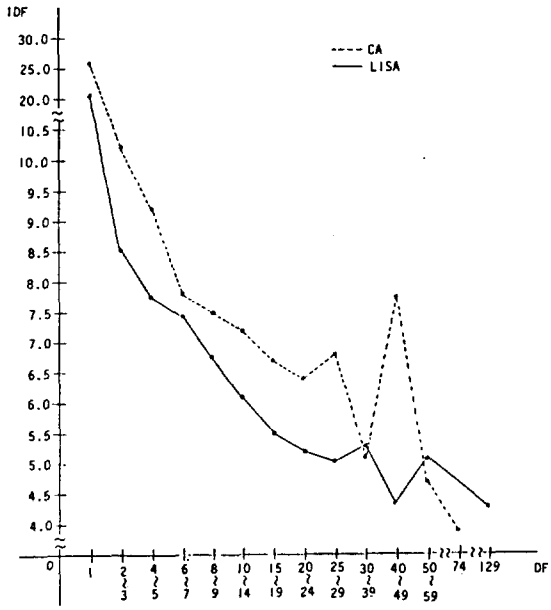
3. 문헌빈도와 색인기법 간의 관계

문헌빈도는 두기법 모두에 큰 영향을 미치며 문헌빈도 분포특성에 따라 색인기법의 성능이 달라질 것으로 여겨진다. 따라서 다음에서는 문헌빈도와 색인기법들 간의 상관관계를 살펴보기로 한다.

3.1. 문헌빈도와 역문헌빈도가중기법의 관계

로칼밸류는 적용하지 않은 역문헌빈도가는 문헌빈도의 역함수를 취하므로 거의 반비례관계를 형성하게 되며 로칼밸류를 적용한 경우에도 고문헌빈도를 갖는 몇몇 용어들을 제외하면 전체적 경향은 유사하나 상관관계는 더 낮아지게 된다.

(그림-4)는 문헌빈도에 따라 할당되는 역문헌빈도가를 나타낸 것이다.



(그림-4) 문헌빈도와 역문헌빈도가의 관계

위 그림에서와 같이 전반적으로 역문헌빈도(IDFv)는 CA집단에서 더 높게 나타나 있다.

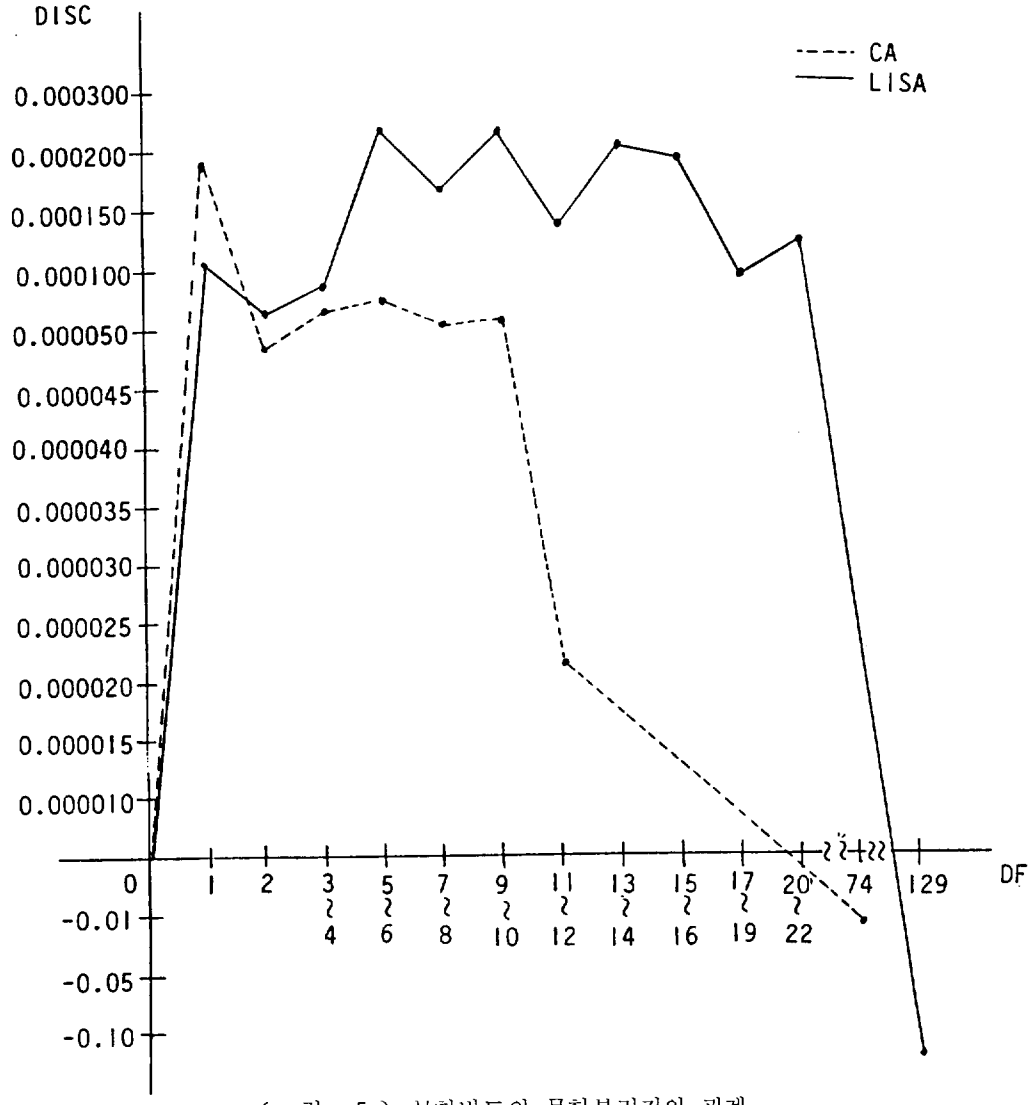
즉, 역문헌빈도가중기법에서 제시하는 적합한 용어특성은 CA집단에 더 유사하다는 것을 알 수 있다. 또한 높은 역문헌빈도가를 갖는 문헌빈도는 1～3구간으로 CA집단의 경우는 전체의 71%에 해당되며 LISA집단의 경우에는 전체의 64%에 해당하는 높은 비율로 CA집단에서 7% 더 높게 나타났다.

따라서 역문헌빈도가중기법에서는 CA집단의 용어특성이 더 적합하다고 하겠다.

3.2. 문헌빈도와 문헌분리가가중기법의 관계

문헌분리가가중기법은 중간문헌빈도를 갖는 용어에 높은 가중치를 할당하는 기법으로 샐튼은 그 적합문헌빈도를 $N/100 \leq \chi \leq N/10$ 으로 제시한 바 있다.

다음의 그래프에서 그 관계를 좀 더 구체적으로 살펴보기로 한다.(그림-5)



(그림-5) 문헌빈도와 문헌분리가의 관계

위의 그림에서와 같이 전체적으로 문헌분리가가 LISA집단에서 보다 높게 나타나 있다.

문헌빈도 1을 갖는 용어에서는 CA집단의 문헌분리가가 더 높게 나타났는데 CA와 같은 자연과학 주제에서는 저문헌빈도용어가 더 유용하기 때문이다. 또한 높은 문헌분리가를 갖는 용어들은 CA집단에서 1과 3～10구간의 문헌빈도를 갖는 용어들로 ($DISC_V > 0.000050$) 전체의 65%에 해당하며 LISA집단에서는 1～22구간의 문헌빈도를 갖는 용어들로 ($DISC_V > 0.000050$) 전체의 97%에 해당한다. 이로써 샐튼이 제시한 적합문헌빈도가 LISA 집단에서는 어느정도 적합했으나 CA집단에서는 적합하지 않음을 알 수 있다. 음의 값을 갖는 용어들은 CA집단에서 문헌빈도가 13이상인 용어들(4.8%)이었으며 LISA집단에서는 문헌빈도 23이상인 용어들(1.8%)로 나타났다. 이와 같이 문헌분리가가중기법은 중간문헌빈도의 비율이 높은 용어특성을 갖는 주제에 더 높은 가중치를 할당하게 되므로 LISA집단의 용어특성에 더 적합할 것으로 보인다.

이상에서 문헌빈도와 색인기법 간의 관계를 분석해 본 결과는 다음과 같다.

첫째, 역문헌빈도가중기법에서 적합한 문헌빈도는 1～3까지의 구간이며 이 구간에 속해 있는 용어들의 비율은 CA집단에서 7% 더 높게 나타났다.

둘째, 문헌분리가가중기법에서 적합한 문헌빈도는 중간빈도구간이며 이 구간에 속해 있는 용어들의 비율은 LISA집단에서 32%까지 높게 나타났다.

세째, 각 기법에 적합한 용어의 특성은 상이하며 약학분야(CA)의 용어특성은 역문헌빈도가중기법에 도서관·정보학분야(LISA)의 용어특성은 문헌분리가가중기법에 더 적합하게 나타났다.

4. 분야별 색인어 선정 및 색인기법 평가

4.1. 분야별 색인어 선정

색인어 선정은 역문헌빈도가와 문헌분리가에서 높은 가중치를 갖는 용어들을 선정하였는데 양집단을 비교하기 위해 적합색인어로 선정된 단어들의 전체에 대한 비율을 근거로 선정비율을 달리하였다. 즉, CA집단에서는 적합색인어의 전체에 대한 비율이 50%(635) 였으며, LISA집단에서는 35%(332)였는데 각 집단에서 약 10%씩을 더한 비율인 60%(763)와 46%(435)를 기준으로 색인어를 선정하였다.

또한 문헌빈도가 1인 용어는 색인어 선정에서 제외하였는데 너무 많은 용어들이 문헌빈도 1을 갖고 있으며 (CA집단에는 32.4%) 색인시스템의 성능 향상에 유용하지 않기 때문이다. 샐튼은 이에대해 다만 하나의 문헌에서만 K번 나타나는 희귀용어는 하나의 문헌에만 집중되어 있으므로 검색에서는 중요치 않은 용어임[53]을 밝히고 있다. 실제로 문헌빈도 1을 갖는 용어를 제외하고 선정한 색인어들에 대한 성능은 각 집단에 대해 다음과 같이 향상되었으며 각각의 선정 기준치(threshold)도 제시되어 있다.(표-9).

53) G.Salton and C.S. Yang. Op. cit., P. 355.

		CA	LISA
IDF	정확율	+ 1.8%	+ 3.4%
	재현율	+ 2.5%	− 3 %
	기준치	6.13794	7.65105
DISC	정확율	+ 0.4%	+ 0.6%
	재현율	+ 0.8%	+ 2.7%
	기준치	0.0000032	0.0000412

(표-9) 문헌빈도1을 갖는 용어 제거후의 성능향상율

전반적으로 집단으로는 LISA집단에서, 기법으로는 역문헌빈도가중에서 향상율이 높은데 이는 역문헌빈도가중기법에서는 문헌빈도가 1인 용어에 가장 높은 가중치를 할당하며 도서관·정보학분야에서는 약학분야에서 보다 저문헌빈도용어가 색인으로써 덜 중요하기 때문이다.

4.2. 색인기법 평가

색인기법 평가는 기존 수작업색인과의 비교로 행해졌으며 비교에 의한 일치여부에 따라 색인기법을 평가하기 위해 시스템의 검색효율평가에 사용되는 정확율(Precision ratio), 재현율(Recall ratio), 제외율(Omission ratio), 잡음율(Noise ratio)을 다음과 같이 변형하여 사용하였다.

$$\text{정확율} = \frac{\text{선정된 적합색인어}}{\text{선정된 색인어}} \times 100$$

$$\text{재현율} = \frac{\text{선정된 적합색인어}}{\text{전체 적합색인어}} \times 100$$

$$\text{제외율} = \frac{\text{선정되지 않은 적합색인어}}{\text{전체 적합색인어}} \times 100$$

$$\text{잡음율} = \frac{\text{선정된 부적합색인어}}{\text{선정된 색인어}} \times 100$$

두집단에서의 각 기법에 대한 성능은 다음과 같다.

CA집단에서는 역문헌빈도가중기법의 성능이 더 우수하게 나타났으나 기법 간의 큰 차이는 없었으며 LISA집단에서는 문헌분리가가중기법의 성능이 월등한데 이는 앞서 문헌빈도와의 관계로 살펴본 분석의 결과와 일치하는 것이다.

또한 전체적으로 CA집단에서의 성능이 두기법 모두에서 LISA집단보다 높은 것은 CA에 비해 LISA의 색인이 덜 상세하다는 점과 일반적으로 자동색인기법이 특정용어들이 한 문헌

```
*** PERFORMANCE EVALUATION OF IDFv IN CA ***

Total terms                                    : 1272
Total relevance index terms                    :  635   49.9%
Total selected index terms by IDFv             :  763   60.0%
Selected relevance index terms by IDFv         :  405
Not selected relevance index terms by IDFv     :  230
Selected non-relevance index terms by IDFv     :  358

PRECISION ratio :  53.1%
RECALL ratio    :  63.8%
OMISSION ratio  :  36.2%
NOISE ratio     :  46.9%
```

*** PERFORMANCE EVALUATION OF DISCv IN CA ***

Total terms	: 1272	
Total relevance index terms	: 635	49.9%
Total selected index terms by DISCv	: 763	60.0%
Selected relevance index terms by DISCv	: 399	
Not selected relevance index terms by DISCv	: 236	
Selected non-relevance index terms by DISCv	: 364	

PRECISION ratio	: 52.3%
RECALL ratio	: 62.8%
OMISSION ratio	: 37.2%
NOISE ratio	: 57.7%

*** PERFORMANCE EVALUATION OF IDFv IN LISA ***

Total terms	: 940	
Total relevance index terms	: 332	35.3%
Total selected index terms by IDFv	: 435	46.3%
Selected relevance index terms by IDFv	: 146	
Not selected relevance index terms by IDFv	: 186	
Selected non-relevance index terms by IDFv	: 289	

PRECISION ratio	: 33.6%
RECALL ratio	: 43.9%
OMISSION ratio	: 56.1%
NOISE ratio	: 66.4%

*** PERFORMANCE EVALUATION OF DISCv IN LISA ***

Total terms	: 940	
Total relevance index terms	: 332	35.3%
Total selected index terms by DISCv	: 435	46.3%
Selected relevance index terms by DISCv	: 203	
Not selected relevance index terms by DISCv	: 129	
Selected non-relevance index terms by DISCv	: 232	

PRECISION ratio	: 46.7%
RECALL ratio	: 61.1%
OMISSION ratio	: 38.9%
NOISE ratio	: 53.3%

에의 군집율이 높은 자연과학분야에 더 적합한 속성을 갖는다는 점등을 함께 고려해 볼 수 있다.

V. 결론 및 제언

선행연구에서와 같이 지금까지 제시된 자동색인기법은 매우 다양하며 계속해서 새로운 기법들이 소개되고 있다. 이러한 자동색인기법들을 적절히 사용하기 위해서는 적용 대상장서의 통계적 용어특성이 먼저 고려되어야 하며 그 특성에 적합한 자동색인기법이 선정되어야 한다.

본 연구에서는 서로 상이한 주제를 갖는 장서의 용어특성에 적합한 색인기법을 제시하기 위해 화학 및 화공분야의 초록지인 CA와 도서관·정보학분야의 초록지인 LISA를 대상으로 각 자동색인기법에 적합한 용어특성을 분석해 본 결과 다음과 같은 결론을 얻었다.

첫째, 분석을 위해 사용된 변수는 총장서빈도(CF), 문헌빈도(DF), 역문헌빈도가(IDFv) 문헌분리가(DISCv)의 4가지였으며 총장서빈도와 문헌빈도의 상관관계와 기법의 적용시 할당된 가중치들 간의 상관관계를 제외한 나머지 변수들 간의 상관관계는 감소함수 관계를 형성한다.

이러한 상관관계 분석에서 얻을 수 있는 주요 결론은 다음과 같다.

1) 용어들의 특정문헌에 대한 군집율은 CA 집단에서 높게 나타났다.

2) 역문헌빈도가중기법에서는 총장서빈도나 문헌빈도가 낮은 용어일수록 높은 가중치를 할당하는데 이러한 현상은 문헌빈도를 기준으로 할때 더 명확하다.

3) 문헌분리가가중기법은 총장서빈도나 문헌빈도나 특정한 관계를 갖지 않는데 일반적으로 중간문헌빈도를 갖는 용어에 높은 가중치를 할당하기 때문이다.

4) 기법의 적용시 할당된 가중치들 간에는 거의 관계가 없으며 CA집단에서의 상관계수가 LISA집단에 비해 높으므로 성능의 차이는 LISA집단에서 더 뚜렷이 나타나게 된다.

둘째, 서로 다른 주제를 갖는 장서들의 용어특성 간에는 차이가 있다. 총장서빈도의 분포특성에서는 별다른 차이점을 발견할 수 없었으나 문헌빈도의 분포특성에서는 저(低)문헌빈도와 중간문헌빈도에서 차이를 보였는데 CA집단에서는 저문헌빈도를 갖는 용어가 전체의 32.4%로 LISA집단에 비해 28.4% 높았으며 LISA집단에서는 중간문헌빈도를 갖는 용어가 전체의 83%로 CA집단에 비해 16.7% 높게 나타났다.

또한 CA집단은 LISA집단에 비해 한 문헌에 대한 용어들의 군집율이 높게 나타나 로칼벨류에 의한 용어가중이 전체 가중치에 큰 영향을 미친다는 것을 알 수 있다. 이로써 첫번째 가설이 검정된다.

세째, 대상장서의 용어특성에 따라 자동색인기법의 성능이 달라진다. 역문헌빈도가중기법에서 높은 가중치를 할당하는 문헌빈도는 1~3구간으로 CA집단의 경우 전체의 71%에 해당하고 LISA집단에 비해 7% 높게 나타났으며, 문헌분리가가중기법에서는 LISA집단의 경우 1~22구간의 문헌빈도로 전체의 97%에 해당하고 문헌빈도 1과 3~10구간에 높은 가중치를 할당하는 CA집단에 비해 32%높게 나타났는데 두번째 가설을 검정하는 것이다. 이렇게 주제에 따른 용어특성은 색인기법의 성능

에도 영향을 미치게 되며 각기법에서 제시되는 최적의 용어특성과 장서내의 용어특성이 일치하는 경우에 성능이 향상된다.

네째, 세번째 가설을 검정하는 결론은 서로 다른 주제를 갖는 장서에 적합한 자동색인기법에는 차이가 있는데 CA집단에서는 역문헌빈도가중기법이, LISA집단에서는 문헌분리가가중기법이 더 적합하다는 것이다. 약학분야 (CA)에는 특정어 (저문헌빈도용어)가 많이 포함되어 있으며 색인어로써 더 유용하게 사용되는 반면 도서관·정보학분야(LISA)에서는 특정어보다 준특정어 (중간문헌빈도용어)가 더 많이 포함되어 있으며 색인어로써도 더 유용하게 사용된다. 실제로 CA집단에서는 역문헌빈도가중기법에서의 성능이 정확율 및 재현율 모두에서 약 1%씩 향상되었으며, LISA집단에서는 문헌분리가가중기법에서의 성능이 정확율에서는 13% 재현율에서는 약 27% 가량 높게 나타났다.

마지막으로 좀더 다양한 주제들을 대상으로 하여 실험이 행해진다면 자연과학, 사회과학 및 인문과학에서의 일반적인 용어특성을 정형화시킬 수 있을 것으로 보이며 어떤 분야에나 적합한 색인기법을 개발하기 보다는 분야별로 적합한 색인기법을 개발함으로써 그 분야의 용어특성을 성능 향상에 최대한 활용하는 것이 바람직할 것으로 생각된다. 또한 이용자 지향적인 색인기법의 적합성 판정 문제를 체계화함으로써 실제 활용단계에서 더 나은 시스템의 성능을 기대할 수 있을 것이다.

참 고 문 헌

김 현 희 "An Investigation Automatic Technigues." 情報管理學會誌, Vol. 1, No.1(1984), PP. 43-62.

司 空 哲 情報檢索論. 서울 : 亞細亞文化社,1983.

宋 美 蓮 "自動索引方法과 自動索引 시스템 性能." 情報管理研究, Vol.17, No. 4 (1984), PP.1-15.

鄭 英 美 李泰榮. "자동색인의 통계적 기법과 한국어 문헌의 실험." 圖書館學, 9 (1982). PP.99-108.

加藤德義 "自動索引의 動向과 逆說的 接近." 高亨坤 譯, 情報管理研究, Vol.No.14, No 4 (1981), PP.182-193.

Cleveland, Donald B. and Cleveland, D. Introduction to Indexing & Abstracting. Littleton : Libraries Umlimited, 1983.

Kiewitt, Eva L. Evaluating Information Retrieval Systems. Westport : GreenWood Press, 1979.

Oddy, R.N. (et al.). Information Retrieval Research. London : Butterworths, 1981.

Salton, G and Michael, J.M. Introduction to Modern Information Retrieval. New York : McGraw-Hill, 1983.

Artandi, Susan and Wolf, Edward H. "The Effectiveness of Automatically Generated Weights and Links in Mechanical Indexing." American Documentation, Vol. 20, No. 3 (July 1969), pp. 198-202.

Barker, F.H., Veal, D.C. and Wyatt, B.K. "Towards Automatic Profile Construction." Journal of Documentation, Vol. 28, No. 1 (March 1972), pp. 44-55.

Bookstein, Abraham and Swanson, Don R. "Probabilistic Models for Automatic Indexing."

Journal of the American Society for Information Science. Vol. 25, No. 5 (Sep. / Oct. 1974), pp. 312 - 318.

Caroll, John M. and Roeloffs, Robert. "Computer Selection of Keywords Using Word-Frequency Analysis." American Documentation, Vol. 20, No. 3 (July 1969), pp. 227 - 233.

Cooper, W.S. and Maron, M.E. "Foundations of Probabilistic and Utility-Theoretic Indexing." Journal of the Association for Computing Machinery, Vol. 25, No. 1 (Jan. 1978), pp. 67 - 80.

Damerau, Fred J. "An Experiment in Automatic Indexing." American Documentation, Vol. 16, No. 4 (Oct. 1965), pp. 283 - 289.

Farradane, J. "The Evaluation of Information Retrieval System." Journal of Documentation, Vol. 30, No. 2 (Oct. 1974), pp. 195 - 209.

Harter, Stephen P. "A Probabilistic Approach to Automatic Keyword Indexing : Part 1 On the Distribution of Specialty Words in a Technical Literature." Journal of the American Society for Information Science, Vol. 26, No. 4 (July/Aug. 1975), pp. 197 - 206.

Harter, Stephen P. "A Probabilistic Approach to Automatic Keyword Indexing : Part 2. An Algorithm for Probabilistic Indexing." Journal of the American Society for Information Science, Vol. 26, No. 5 (Sep./Oct. 1975), pp. 280.- 289.

. "Statistical Approaches to Automatic Indexing." Drexel Library Quarterly, Vol. 14, No. 2 (Apr. 1978), pp. 57 - 74.

Lesk, M.E. "Word-Word Associations in Document Retrieval Systems." American Documentation, Vol. 20, No. 1 (Jan. 1969), pp. 27 - 38.

Maron, M.E. "On Indexing Retrieval and the Meaning of About." Journal of the American Society for Information Science, Vol. 28, No. 1 (Jan. 1977), pp. 38 - 43.

Miller, William L. "A Probabilistic Search Strategy for Medlars." Journal of Documentation, Vol. 27, No. 4 (Dec. 1971), pp. 254 - 266.

Minker, Jack, Peltola, Eero and Wilson, Gerald A. "Document Retrieval Experiments Using Cluster Analysis." Journal of the American Society for Information Science, Vol. 24, No. 4 (July/Aug. 1973), pp. 246 - 257.

Robertson, S.E. "On Relevance Weight Estimation and Query Expansion." Journal of Documentation, Vol. 42, No. 3 (Sep. 1986), pp. 182 - 188.

. "Specificity and Weighted Retrieval." Journal of Documentation, Vol. 30, No. 1 (March 1974), pp. 41 - 46.

. "Theories and Models in Information Retrieval." Journal of Documentation, Vol. 33, No. 2 (June 1977), pp. 126-148.

Salton, G. "Automated Language Processing." Annual Review of Information Science & Technology, Vol. 3 (1968), pp. 169 - 199.

Salton, G. "Automatic Text Analysis." Science, Vol. 168, No. 17 (Apr. 1970), pp. 335 - 343.

_____. "A Comparison between Manual and Automatic Indexing Methods." American Documentation, Vol. 20, No. 1 (Jan. 1969), pp. 61 - 71.

_____. "Mathematics and Information Retrieval." Journal of Documentation, Vol. 35, No. 1 (March 1979), pp. 1 - 29.

Salton, G. and Lesk, M.E. "Computer Evaluation of Indexing and Test Processing." Journal of the Association for Computing Machinery, Vol. 15, No. 1 (Jan. 1968), pp.621-640.

Salton, G., Wu, H. and Yu, C. T. "The Measurement of Term Importance." Journal of the American Society for Information Science, Vol. 32, No. 1 (Jan. 1981), pp. 175 - 186.

Salton, G. and Yang, C.S. "On the Specification of Term Values in Automatic Indexing." Journal of Documentation, Vol. 29, No. 4 (Dec. 1973), pp. 351 - 372.

Salton, G., Yang, C.S. and Yu, C.T. "A Theory of Term Importance in Automatic Text Analysis." Journal of the American Society for Information Science, Vol. 26, No. 1 (Jan./Feb. 1975), pp. 33 - 44.

Sparck Jones, K. "Automatic Indexing." Journal of Documentation, Vol. 30, No. 4 (Dec. 1974), pp. 393 - 432.

_____. "Does Indexing Exhaustivity Matter?" Journal of the American Society for Information Science, Vol. 24, No. 5

(Oct. 1973), pp. 313-327.

_____. "A Performance Yardstick for Test Collection." Journal of Documentation, Vol. 31, No. 4 (Dec. 1975), pp. 266-272.

_____. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." Journal of Documentation, Vol. 28, No. 1 (March 1972), pp. 11-21.

Svenonius, Elaine. "An Experiment in Index Term Frequency." Journal of the American Society for Information Science, Vol. 23, No. 2 (March/Apr. 1972), pp. 103-121.

Van Rijsbergen, C.T. and Sparck Jones, K. "A Test for the Separation of Relevant and Non-Relevant Documents in Experimental Retrieval Collection." Journal of Documentation, Vol. 29, No. 3 (Sep. 1973), pp. 252-257.

Yu, C.T., Lam, K. and Salton, G. "Term Weighting in Information Retrieval Using the Term Precision Model." Journal of the Association for Computing Machinery, Vol. 29, No. 1 (Jan. 1982), pp. 152-170.

Yu, C.T. and Salton, G. "Precision Weighting - An Effective Automatic Indexing Method." Journal of the Association for Computing Machinery, Vol. 23, No. 1 (Jan. 1976), pp. 76-88.

Yu, C.T., Salton, G. and Siu, M.K. "Effective Automatic Indexing Using Term Addition and Deletion." Journal of the Association for Computing Machinery, Vol. 25, No. 2 (Apr. 1978), pp. 210-225.

부록 I. 불용어리스트

a	alone	by	following		
able	along	came	for		
about	already	can	forthcoming		
above	also	carried	fortunately		
across	although	caused	found		
affected	always	causing	frequently		
after	among	certainly	from		
again	an	clearly	full		
against	and	come	fully		
ago	another	comparable	further		
all	any	compared	furthermore		
almost	anyone	considerable	gained		
alone	apparently	considered	generally		
along	appear	contained	get		
already	appearing	correctly	given		
also	appropriate	could	go		
although	approximately	coverd	going		
always	are	described	got		
among	around	despite	gradually		
an	as	did	great		
and	at	disccused	greater		
another	available	do	greatly		
any	aware	does	had		
anyone	away	doing	has		
apparently	based	done	have		
appear	basic	during	having		
appearing	be	each	he		
appropriate	became	easily	held		
approximately	because	either	her		
are	bccome	enable	here		
around	becoming	encountered	his		
as	been	enough	how		
at	before	especially	however		
available	begin	established	if	occurred	since
aware	begun	even	ignored	of	slightly
away	behind	ever	immediately	off	so
based	being	every	important	often	some
basic	best	everybody	in	on	sometimes
be	better	everything	included	once	soon
became	between	everytime	instead	one	specifically
because	beyond	examined	into	only	still
bccome	both	except	involved	or	substantially
becoming	brief	far	is	other	successfully
been	briefly	few	it	our	such
before	but	followed	its	out	suggested

itself
just
known
last
less
let
like
likely
made
mainly
make
making
many
markedly
may
mere
might
more
most
much
must
nearly
neither
never
newly
no
none
nor
not
nothing
now
numerous
occasionally
over
overall
overview
own
particular
particularly
per
performed
played
previously
probably
profoundly
properly
put
rather
really
recently
referred
reflected
regard
relatively
remarkably
reported
researched
resulted
same
seem
seen
several
should
shown
significantly
similarly
suitable
take
taken
tended
tested
than
that
the
their
them
then
there
thereby
therefore
these
they
this
thorough
those
though
through
thus
times
to
together
too
took
toward
ultimately
under
undertaken
unknown
unless
you
your

until
up
upon
us
used
using
usually
varied
various
very
via
was
way
we
well
were
what
when
where
whereas
whereby
whether
which
while
who
whole
whom
whose
why
will
with
within
without
worked
would
yet