

# 자동색인의 통계적기법과 한국어 문헌의 실험

鄭 瑛 美\*  
李 泰 榮\*\*

## <차 례>

- |                 |                   |
|-----------------|-------------------|
| 1. 서 론          | 4. 한국어문헌의 색인어선정실험 |
| 2. 자동색인기법       | 4.1 단어빈도론의 실험     |
| 3. 통계적 색인어 선정기준 | 4.2 문헌빈도론의 실험     |
| 3.1 단어빈도론       | 4.3 하터의 확률이론의 실험  |
| 3.2 문헌빈도론       | 5. 결 론            |
| 3.3 확률이론        |                   |

## 1. 서 론

색인작업은 문헌의 내용을 분석하여 그 중심주제를 표현하기에 가장 적합한 용어나 기호 즉, 키워드, 디스크립터, 분류기호, 주제명표목 등의 색인어를 선정하여 문헌에 부여하는 일로서 이 결과 작성된 색인은 정보원과 정보입수자 사이에 위치하여 특정한 주제의 문헌들을 선별하여 주고 선별된 자료의 소재를 지시하여 주는 기능을 수행한다. 따라서 효과적인 정보검색은 적절한 색인의 사용을 전제로 하고 있으며 색인의 성능이 결국은 정보검색 시스템 전체의 성능을 좌우하게 된다. 왜냐하면 정보검색시스템의 궁극적인 목적은 이용자가 원하는 내용의 정보나 정보자료를 제공하는 것이기 때문이다.

전통적으로 색인작업은 훈련된 사서나 주제전문가에 의해 수행되어 왔으

\* 연세대학교 도서관학과 부교수

\*\* 중의여전 도서관과 강사

나 1950년대에 이르러 색인해야 할 문헌의 급속한 증가와 색인경험과 주제 지식을 제대로 갖추고 있는 색인자의 절대적인 부족현상이 인식되면서부터 컴퓨터를 사용한 자동색인기법이 출현하게 되었다.

자동색인의 효시를 이룬 연구가 1957년 룬(Luhn)에<sup>1)</sup> 의하여 발표되었는데 룬의 사상은 이후 개발된 다양한 자동색인기법의 기초를 이루었다. 룬은 기본적으로 기록물에 의한 사상의 커뮤니케이션에 있어서 저자는 자신의 사상을 잘 전달할 수 있도록 단어를 결합하고 메시지의 특정성수준을 선택하게 되는데 이때 가장 적합한 수준을 선택할 확률에 근거하여 커뮤니케이션의 성패가 결정된다고 보았다. 즉 저자는 특정한 단어들을 결합하여 사용함으로써 독자에게 자신의 경험이나 사고과정을 전달하고자 하는데 이때 저자는 독자와 커뮤니케이션이 성취될 수 있는 수준까지 자신의 사상을 점점 작은 사상으로 쪼개가게 된다는 것이다. 이 수준은 두 사람이 갖는 공통되는 경험의 유사성 정도에 따라 정해진다. 즉 공통되는 경험이 적을수록 저자는 더 많은 단어를 사용하여 사상을 표현하고자 한다. 따라서 저자의 사상은 독자와의 공통경험의 정도에 따라 한개의 문장을 통해 전달될 수도 있고, 또는 여러 문장으로 구성된 하나의 문단을 통해서, 아니면 여러 문단으로 구성된 하나의 문헌을 통해서 전달될 수도 있는 것이다.

이러한 커뮤니케이션이론에 근거한 룬의 가설은 문헌에 나타나는 단어들을 문헌의 내용분석을 위해 사용할 수 있으며 단어의 출현빈도가 단어의 중요성 내지는 의미성을 결정하는 기준이 된다는 것으로 룬 이후 개발된 대부분의 자동색인기법에서는 실제로 이 가설에 근거하여 색인어를 선정하고 있음을 볼 수 있다.

자동색인방법은 색인어를 선정하는 기준에 따라 통계적인 기법과 비통계적인 기법으로 나누어지는데 본고에서는 통계적인 기법에 관해 상세히 고찰하고 그 중에서 특히 주목할만한 몇가지 기준을 한국어문헌을 대상으로 실험하여 그 응용성을 평가해 보고자 한다.

1) Luhn, H.P. "A Statistical Approach to Mechanized Encoding and Searching of Library Information," IBM J. of Research and Development, 1/4, 309-17 (1957).

## 2. 자동색인기법

자동색인의 기본원리는 문헌을 구성하는 단어들을 일정한 기준에 의해 주제어와 비주제어, 의미어와 무의미어, 또는 전문어와 비전문어로 구분하고 주제어, 의미어, 또는 전문어로 평가된 단어들로부터 색인어를 선정하는 것이다.

자동색인의 가장 초보적인 단계는 KWIC 색인으로 이 색인방식은 문헌의 표제를 입력하여 표제를 구성하는 각 단어를 분리한 다음 불용어리스트에 나와 있는 전치사, 조사, 관사 등의 비주제어를 제외한 나머지 단어를 색인어로 선정하는 방법이다. 그러나 KWIC 색인은 문헌에 표제를 부여하는 저자에 의해 이미 색인어가 선택되는 것이므로 순열색인, 인용색인 등과 함께 자동색인이라기 보다는 컴퓨터인쇄색인 또는 기계색인이라고 부르고 있다. 불용어리스트를 사용하는 다른 예로는 온라인정보검색시스템인 다이알로그(DIALOG)의 자동색인방식을 들 수 있으며 여기에서는 an, and, by, for, from, of, the, to, with의 9개 불용어 이외에는 표제와 초록에 출현하는 모든 단어를 색인어로 채택하고 있다.

색인어 선정기준에 의해 자동색인방법은 통계적기법과 비통계적기법으로 대분할 수 있다. 통계적기법은 단어의 출현빈도를 기준으로 하여 색인어를 선정하는 것으로 한 단어가 각 문헌에 출현한 빈도(단어빈도)나 문헌집단내 총빈도(전체문헌단어빈도), 또는 단어가 출현한 문헌의 빈도(문헌빈도), 그리고 상대빈도나 또는 문헌집단내에서 한 단어가 몇개의 문헌에 몇번씩 출현했는가를 설명하는 빈도분포모형을 기준으로 사용한다. 이때 일반적으로 고빈도의 공통어나 저빈도의 희귀단어는 색인어대상에서 제외하고 나머지 단어들은 출현빈도통계나 빈도분포모형에 의해 단어의 의미성을 측정하고 이 측정치가 일정한 한계치범주 이내에 속하는 단어들을 색인어로 선택하게 된다.

전형적인 영어문헌에서는 대략 본문의 50%정도는 접속사, 대명사, 전치사, 조사, 수량형용사 등의 주제어로서는 별 의미가 없는 단어들로 구성된

다.<sup>2)</sup> 가장 보편적인 색인방법에서는 이러한 무의미어들로 불용어리스트를 작성하여 무의미어들은 일단 색인어대상에서 제외하고 나머지 단어들은 같은 어간을 갖는 단어들을 하나의 단어로 취급하는 등의 통제를 가한 다음 각 단어의 출현빈도를 계산하여 빈도순으로 배열하고 일정한 빈도 이상 나타난 단어나 또는 본문의 일정한 비율이상을 점유하는 단어들을 색인어로 선택하도록 하고 있다.

비통계적인 기법으로는 “결론”, “결과”, “요약”, “입증하다” 등과 같이 문헌의 주제를 축약적으로 표현해주는 특정한 의미의 단어를 찾아 이러한 단어가 출현한 문장속에 함께 출현한 단어들을 색인어로 선택하는 단서어기법과 구두점이나 전치사, 접속사 등을 단서로 하여 문장을 문법적으로 분석하여 전치사구나 명사구 등 한 단위로 처리되는 단어군을 찾아낸 다음 빈번히 나타나는 단일어나 복합어를 색인어로 선택하는 부분적인 문장분석에서부터 단어구의 유형과 구조를 자동적으로 식별하는 완전한 문법적 문장분석법에 이르기까지 다양한 구문론적인 기법이 있다.<sup>3)</sup>

또한 문헌속에서 서론, 결론, 요약 등의 제목을 갖는 특정한 부분에 나타나는 단어들만을 색인어로 선택하는 방법과 각 문단의 첫 문장과 마지막 문장에 나타나는 단어들을 색인어로 선택하는 방법과 같은 문헌내 소재에 의한 색인어 선정기준이 사용되고 있다.<sup>4)</sup> 비통계적인 기준은 통계적인 기준대신 사용하거나 통계적인 기준을 보완할 목적으로 사용하고 있다.

보코브스키(Borkowski)와 마틴(Martin)은<sup>5)</sup> 자동초록에 관한 연구에서 자동초록기법(정확히 말하면 초록문발췌기법)을 아래와 같이 다섯 부류로 분류하고 있는데 이 분류는 자동색인에도 그대로 적용가능하다. 즉 (1) 문헌내 출현빈도가 일정한 범주에 속하는 단어나 단어군을 포함하는 문장을 발

2) Lancaster, F.W. *Information Retrieval Systems: Characteristics, Testing, and Evaluation*. New York: Wiley, 1968, p.100.

3) Sparck Jones, K. "Automatic Indexing," *J. of Documentation*, 30/4, 393-432(1974).

4) Baxendale, P.B. "Machine-made Index for Technical Literature—An Experiment," *IBM J. of Research and Development*, 2/4, 354-61 (1958).

5) Borkowski, C. and J.S. Martin. "Structure, Effectiveness and Benefits of Lextractor, an Operational Computer Program for Automatic Extraction of Case Summaries and Dispositions from Court Decisions," *JASIS*, 26/2, 94-102 (1975).

체하는 통계적 기법, (2) “이 논문은”, “결론”, “요약”, “결과” 등과 같은 특별한 의미의 단어나 단어군을 포함하는 문장을 발췌하는 단서어기법, (3) 문헌의 표제에 나타난 단어들을 포함하는 문장을 발췌하는 표제어기법, (4) 초록, 결론, 요약 부분과 같은 문헌내 특정한 위치에 나타나는 문장을 발췌하는 문헌내소재기법, (5) 문헌에 나타나는 단어나 단어군을 먼저 어의적이거나 문법적인 클래스에 배정한 다음 이러한 클래스들이 일정한 패턴에 따라 동시에 출현하는 문장을 발췌하는 어의적/구문론적기법으로 분류하고 있다.

앞에서 대략적으로 고찰한 자동색인기법도 단어의 출현빈도에 기초한 통계적기법, 단서가 되는 단어를 이용하는 단서어기법, 특정한 위치에 나타나는 단어를 선택하는 문헌내소재기법, 문장의 문법적인 분석을 통한 구문론적기법, 그리고 KWIC색인과 같이 불용어리스트를 사용하는 불용어기법으로 세분할 수 있다. 이러한 기법들은 단독적으로 사용되거나 또는 함께 보완적으로 사용되고 있다.

### 3. 통계적 색인어 선정기준

통계적인 색인어 선정기준은 단어의 의미성을 어떻게 측정할 것인가 하는 문제에 해답을 제시해주며 구체적으로 단어의 출현빈도 계산과 한계치 결정 방법을 정해주고 있다. 실제로 측정되는 단어의 출현빈도는 앞서서도 이미 언급한 바와 같이 각 문헌에 특정한 단어가 출현한 횟수를 계산하는 단어빈도, 특정한 단어가 한 문헌집단내에 출현한 횟수를 나타내는 전체문헌단어빈도, 특정한 단어가 출현한 문헌의 수를 나타내는 문헌빈도, 단어빈도를 문헌빈도로 나누어주므로써 전체문헌에 나타나는 횟수가 적은 희귀단어가 특정한 문헌에 여러번 나타났을 때 이 단어를 의미어로 결정하는 역문헌빈도로 구분된다. 역문헌빈도는 단어가 예상되는 빈도와 다른 출현빈도를 가질 때 의미어로 판정하는 상대적인 빈도측정기준이다. 다메로(Damerau)는<sup>6)</sup> 자

6) Damerau, F.J. “An Experiment in Automatic Indexing,” American Documentation, 6/4, 283-289 (1965).

동색인 실험에서 통계학적으로 볼 때 예상할 수 없을 정도로 높은 빈도를 갖는 단어를 색인으로 선택하였는데 이때 측정된 단어빈도는 대표적인 상대빈도로서 각 단어의 한 문헌에서의 상대빈도를  $f$  라고 하고 대규모 샘플내에서의 상대빈도를  $r$  이라고 할 때  $f-r$ ,  $f/f+r$ ,  $f/r$ 의 세 공식과 포아슨분포 공식을 사용하였다.

이외에 샤논(Shannon)의 정보이론을 응용하여 시그널—잡음 비율로 단어의 의미성을 측정하여 색인어를 선정한 기법과 단어빈도분포의 평방편차를 이용하여 색인어를 선정하는 기법 등이 제시되었다.<sup>7)</sup>

### 3.1 단어빈도론

문헌에 출현하는 단어들의 빈도에 관한 통계적인 법칙이 1949년 지프(Zipf)에 의해 발표되었다.<sup>8)</sup> 잘 알려져 있는 이 지프의 법칙은 어느 한 문헌에 나타나는 단어들을 출현빈도순으로 배열하여 순위를 매기면 출현빈도와 순위를 곱한 값이 일정하다는 것으로 지프의 제 1 법칙이라고도 한다. 이 법칙은 고빈도의 단어에는 잘 적용이 되나 저빈도의 단어에는 적용되지 않아서 지프는 저빈도 단어에 맞는 제 2 법칙을 제안하였고 후에 이 제 2 법칙은 부스(Booth)에 의해 수정되었다.<sup>9)</sup>

수정된 지프의 제 2 법칙은  $I_n/I_1=2/n(n+1)$ 의 공식으로 표현되며  $I_1$ 은 한 번 출현한 단어의 수이고  $I_n$ 은  $n$ 번 출현한 단어의 수를 나타낸다. 지프의 제 2 법칙은 한 문헌에 한번만 출현한 단어의 수와  $n$ 번 출현한 단어의 수의 비율은 문헌의 길이에 관계없이 일정하다는 것을 보여주고 있다.

룬은 1957년의 논문에서 하나의 개념이나 복합개념이 문헌속에 나타나는 빈도가 클수록 저자는 그 개념에 보다 큰 의미를 부여하고 있다고 전제하고 하나의 문단에 두번이상 나타나는 개념은 중요한 개념이며 또한 바로 전후에 오는 문단에 나타난 동일한 개념은 한번만 나타나더라도 중요한 개념이

7) Salton, G. Dynamic Information and Library Processing. Englewood Cliff: Prentice-Hall, 1975, p.81-82.

8) Zipf, G.K. Human Behavior and the Principle of Least Effort. Boston: Addison-Wesley, 1949.

9) Booth, A.D. "A Law of Occurrences for Words of Low Frequency," Information and Control, 10/4, 386-393 (1967).

라고 보았다. 문은 1958년 발표된 자동초록에 관한 논문에서<sup>10)</sup> 한 논문에 나타나는 단어의 출현빈도에 의해 단어의 의미성을 측정하는 방법을 제시하고 의미어를 포함하는 문장은 그 의미어의 문장내 위치에 따라 주제문장으로서의 중요성을 결정할 수 있다고 기술하였다. 여기에서 문은 가장 빈번히 나타나는 단어는 너무 일반적인 단어이므로 주제어로서의 가치가 없으며 너무 빈도가 낮은 단어 또한 주제어로서 별 의미가 없다고 보고 이러한 무의미어를 제외한 나머지 단어가 의미어로서 주제색인어의 대상이 된다고 보았다.

문은 지프의 발상에 단어의 의미성을 결합하여 단어의 출현빈도를 색인어 선정기준으로 제시한 최초의 사람으로 고빈도단어와 저빈도단어를 무의미어로 간주하고 단어의 빈도분포에서 최고와 최저의 두개의 한계빈도를 설정하여 한계빈도범위안에 속하는 중간빈도의 단어들을 색인어로 선정하도록 하였다. 그러나 문은 두 한계빈도의 구체적인 산출방법을 제시하지는 않았다.

파오(Pao)는<sup>11)</sup> 단어빈도리스트로부터 고빈도와 저빈도 단어들을 제외한 중간빈도어로부터 색인어를 선택하는 기준을 실험하였다. 파오가 사용한 기준은 고프만(Goffman)의 발상으로서 문헌의 주제를 가장 잘 나타내는 단어들은 지프의 제 1 법칙이 적용되는 고빈도단어로부터 지프의 제 2 법칙이 적용되는 저빈도단어들로 전환되는 영역에 있다고 보고 이 영역에 속하는 단어들을 색인어로 선택하는 것이다. 이 전환점을 산출하는 공식은 부스가 수정한 지프의 제 2 법칙으로 부터 유도한 것으로 다음과 같다.

$$n = (-1 + \sqrt{1 + 8I_1}) / 2$$

$n$ 은 전환점에 있는 단어의 출현빈도를 나타내며  $I_1$ 은 색인대상 문헌에 한번씩 출현한 단어들의 총수이다. 즉 고빈도단어들은 거의 모두가 각각 다른 빈도를 갖는다는 특징에 착안하여 수정된 지프공식의  $I_n$ 을 1로 대치하여 전환점을 구하는 것이다.

파오는 고프만의 전환공식을 사용하여 전환점을 구한 다음 이 지점을 중

10) Luhn, H.P. "The Automatic Creation of Literature Abstracts," IBM J. of Research and Development, 2/2, 159-165 (1958).

11) Pao, M.L. "Automatic Text Analysis Based on Transition Phenomena of Word Occurrences," JASIS, 29/3, 121-124 (1978).

십으로 하여 이 점의 출현빈도보다 상위빈도를 갖는 단어수 만큼의 하위빈도 단어들을 선택하여 이 두 단어그룹에 속하는 단어들이 문헌의 주제를 대표하는 색인어가 될 수 있다고 보고 이 방법을 실제로 실험해 보았다. 실험 대상 문헌은 도서관학분야의 문헌으로 단어의 유형은 559개, 한번 출현한 단어의 수  $I_n$ 은 256이었으며 전환공식에 의해 설정된 전환영역에 속하는 모두 32개의 단어가운데 10개가 주제어, 22개가 비주제어였는데 이 주제어들은 실제로 수작업에 의해 선정된 색인어와 크게 유사하였다.

문의 단어빈도론에 근거한 대부분의 통계적 색인기법에서는 임의로 한계빈도를 정하고 있음에 비해 파오의 실험에서 사용한 색인어선정기준은 문의 단어빈도론에서 설정한 최고 한계빈도와 최저 한계빈도의 타당성을 잘 뒷받침해주고 있다.

### 3.2 문헌빈도론

샬톤(Salton)과 그의 동료들이<sup>12)</sup> 제시한 단어의 문헌분리가(discrimination value)에 의한 색인어선정기법은 좋은 색인어는 문헌집단내의 문헌들을 가능한 한 서로 분리시키며 나쁜 색인어는 문헌들을 오히려 함께 무리짓는다는 색인이론으로부터 발전되었다. 즉 좋은 색인어일수록 문헌들을 분리시켜 문헌집단의 밀집도를 낮추므로써 한 주제를 다루고 있는 문헌들을 이웃 문헌들로부터 쉽게 구별하도록 한다는 것이다.

단어  $k$ 의 문헌분리가의 산출은 먼저 문헌집단내의 각 문헌과 중심문헌(centroid)과의 유사성을 유사계수 산출공식에 의해 측정하고 그 합을 문헌집단의 밀집도라고 하여  $Q$ 로 나타내는데  $Q$ 가 클수록 문헌들이 밀집해 있음을 의미한다. 단어  $k$ 가 문헌집단내의 모든 문헌으로부터 제거되었을 때의 밀집도를  $Q_k$ 라고 한다면  $Q_k$ 에서  $Q$ 를 뺀 값이 단어  $k$ 의 문헌분리가가 된다. 이때 좋은 색인어는 문헌분리가( $Q_k - Q$ )가 양수가 되며 나쁜 색인어는 음수가 되는데 문헌분리가가 높은 단어일수록 색인어로서 적합하다는 것이다. 이 문헌분리가에 의한 색인어 선정기법은 좋은 검색결과를 가져올 수 있다는 것

12) Salton, G. et al. "A Theory of Term Importance in Automatic Text Analysis," JASIS, 26/1, 33-44 (1975).



이 실험을 통하여 입증되었다.<sup>13)</sup>

샐튼 등은 또한 문헌분리가에 의한 단어의 순위와 문헌빈도와의 관계를 실험을 통해 규명하였는데 실험결과 아주 낮은 빈도의 단어들과 고빈도의 단어들은 순위가 낮았고 반면에 중간층에 속하는 단어들 즉, 전체문헌수가  $n$  일 때  $n/100$ 과  $n/10$  사이에 오는 단어들이 문헌분리가가 높은 것으로 나타났다.<sup>14)</sup> 즉 중간층의 문헌빈도를 갖는 단어들이 색인어로서 적합하다는 것을 입증하였다.

샐튼 등은 이 실험결과에 기초하여 문헌빈도에 의한 색인어선정기준을 다음과 같이 제시하였다. 즉 첫째, 문헌빈도가  $n/100$ 에서  $n/10$  사이의 단어들은 그대로 색인어로 채택하고 둘째, 빈도가  $n/10$  이상인 고빈도단어들은 너무 일반적인 단어이므로 보다 낮은 빈도의 단어로 변환시키며 셋째, 빈도가  $n/100$  이하인 저빈도단어들은 너무 희귀하거나 특정한 단어이므로 보다 높은 빈도의 단어로 변환시키도록 하는데 구체적으로는 색인어구를 형성함으로써 빈도를 낮추는 방안과 디소오러스나 동의어사전을 사용하여 같은 개념의 다른 단어들을 한 색인어클래스로 모아주어 빈도를 높이는 방안을 제시하고 있다.

### 3.3 확률이론

색인작성의 확률이론은 1960년 마론(Maron)과 쿤스(Kuhns)가<sup>15)</sup> 문헌의 검색확률에 근거한 확률색인의 개념을 제창한 이래 계속 발전되어 왔으며 특히 IBM 연구소의 아브라함 (C.T. Abraham)이 제시한 포아슨분포론, 하터(Harter) 등이 제시한 2-포아슨분포론에 기초한 색인어 선정기법이 주목할 만하다.

#### 3.3.1 포아슨분포론

총출현빈도가  $R$ 인 하나의 단어가 한 문헌집단을 이루는  $A$ 개의 문헌들 간에 랜덤하게 분포되어 있는 현상을 포아슨분포함수에 의해 설명한 것으로 특

13) Salton, G. and C.S. Yang. "On the Specification of Term Values in Automatic Indexing," J. of Documentation, 29, 351-372 (1973).

14) Salton et al. op. cit., p.35.

15) Maron, M.E. and J.L. Kuhns. "On Relevance, Probabilistic Indexing and Information Retrieval," J. of the ACM, 7/3 (1960).

정한 단어가 한 문헌에  $k$ 번 출현할 확률을 다음의 포아슨분포공식으로 나타내고 있다.

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

위 식에서 파라미터  $\lambda$ 의 값은 특정한 단어가 한 문헌에 출현한 평균빈도로서  $R/A$ 에 해당된다. 예를 들어 100개의 문헌에 정보라는 단어가 200번 출현했다면 평균출현빈도  $\lambda$ 는 2이며 이 단어가 한번 출현한 문헌의 수는  $200P(1)$ , 두번 출현한 문헌의 수는  $200P(2)$ 와 같이 확률적으로 산출될 수 있다는 것이다.

북스타인(Bookstein)과 스완슨(Swanson)은<sup>16)</sup> 비주제어는 전체문헌속에 랜덤하게 분포되지만 주제어는 몇몇 문헌속에 집중적으로 출현한다는 가설 아래 단어들이 포아슨분포현상으로부터 벗어나서 적은 수의 문헌들에 집중되어 있는 정도를 측정하여 색인어 선정기준으로 사용하였다. 집중화정도(clusteredness)를 나타내는 변수  $x$ 는 단순히 한 단어의 문헌집단내 총출현빈도에서 문헌빈도를 뺀 것으로 예를 들어  $a$ 라는 단어가 문헌집단내에 총 75번 출현하고 이 단어가 출현한 문헌수가 65개라면 이때의 집중도  $x$ 의 값은 10이 된다. 클러스터확률  $g(x)$ 는 랜덤분포에서 집중도가  $x$ 이상일 확률이며  $R$ 의 총출현빈도를 갖는 단어가  $A$ 문헌속에 출현할 때 정확히  $l$ 개의 출현빈도를 갖는 문헌의 수가  $S_l$ 이 될 확률을 나타내는 아래와 같은 점유분포함수로부터 확률이 계산되는데 이  $g(x)$ 값에 따라 단어의 순위를 매기게 된다. 이때  $g(x)$ 가 적을수록 집중성이 큰 단어로서 색인어로 적합하다는 것이다.

$$P(S_0, S_1, \dots, S_R) = \frac{R!A!}{A^R(S_0!S_1!S_2!\dots S_R!)[2^{S_2}(3!)^{S_3}(4!)^{S_4}\dots(R!)^{S_R}]}$$

북스타인 등이 실제로 심리학분야의 초록 650개에 출현한 단어들을 같은 주제분야의 출판된 색인지에 근거하여 색인어와 비색인어로 구분한 다음 각 단어에 대한 클러스터확률  $g(x)$ 를 계산한 결과 색인어들은 비색인어보다 훨씬 고도의 집중화현상을 나타내고 있음을 입증하고 만일  $g(x)$ 의 한계치를  $10^{-2}$

16) Bookstein, A. and D.R. Swanson. "Probabilistic Models for Automatic Indexing," JASIS, 25/5, 312-318 (1974).

로 하여  $g(x) < 10^{-2}$ 가 되는 단어들을 색인어로 선택한다면 선택된 단어들 가운데 69%가 실제 색인어이고 이 단어들은 전체 색인어의 82%에 해당된다고 실험결과를 분석하였다.

하터는<sup>17)</sup> 포아슨분포함수는 색인어로 부적합한 비주제어의 분포양상에 보다 잘 적용된다고 지적하고 북스타인 등과 함께 주제어의 분포모형으로 다 음에 상술할 2-포아슨분포함수를 채택하였다.

### 3.3.2 2-포아슨분포론

북스타인과 스완슨은<sup>18)</sup> 한 문헌이 어떤 특정한 개념을 얼마나 깊이있게 다루고 있는가 하는 것은 이 개념을 나타내는 단어의 출현빈도에 의해 결정할 수 있다고 보고 특정한 개념의 취급정도에 따라 한 문헌집단을 여러개의 동일한 수준의 클래스로 나눈 복수포아슨분포모형을 제시하였다.

복수포아슨모형에서 문헌집단은 특정한 개념과의 관련도에 따라 가장 적합한 문헌클래스, 두번째로 적합한 문헌클래스, 세번째로 적합한 문헌클래스 등과 같이 여러 계층으로 구분되어진다는 것인데 실제로 클래스를 세개 이상으로 나누게 되면 베개의 파라미터를 갖게 되어 계산상의 어려움이 따르기 때문에 경제성과 간단성을 고려하여 문헌집단을 적합한 클래스와 부적합한 클래스의 두 클래스로 나눈 2-포아슨 모형에 채택한 것이다.

특정한 주제어가 한 문헌에  $k$ 번 나타난 확률은 다음과 같이  $\pi, \lambda_1, \lambda_2$ 의 세 파라미터를 갖는 2-포아슨 분포모형으로 설명된다.

$$P(k) = \pi \frac{\lambda_1^k e^{-\lambda_1}}{k!} + (1 - \pi) \frac{\lambda_2^k e^{-\lambda_2}}{k!}$$

위에서  $\lambda_1$ 과  $\lambda_2$ 는 각각 문헌클래스 1과 문헌클래스 2에서의 단어의 평균출현 빈도이며  $\pi$ 는 문헌클래스 1에 속하는 문헌의 비율을 나타낸다. 여기에서 클래스 1에 속하는 문헌들은 특정한 개념을 중심주제로 다루고 있는 반면 이 개념을 주변적으로 다루고 있는 문헌들은 클래스 2에 속하게 된다.

비주제어의 분포모형인 포아슨분포는 2-포아슨분포의 특수한 경우로서 두

17) Harter, S.P. "A Probabilistic Approach to Automatic Keyword Indexing," JASIS, 26/4, 197-206 (1975).

18) Bookstein and Swanson, op. cit., P.316-318.

클래스중의 하나가 비어있을 경우에 해당되며 특히  $\pi=1$  일 때 2-포아슨 분포모형은 단일포아슨공식과 같아진다.

2-포아슨 분포모형에서  $\pi, \lambda_1, \lambda_2$  값의 추정은 샘플문헌의 실제 빈도데이터로부터 첫 세 모우먼트값을 계산한 다음 2-포아슨 모우먼트 산출공식에 대입하므로써 가능하다.

하터의 실험에서 실제로 관찰된 빈도분포가 이 모형에 의한 이론치를 따르지 않는 경우가 적지 않았는데 이유로는 이 2-포아슨 분포모형은 문헌집단을 단지 두개의 클래스로만 나누고 각 클래스에 속하는 문헌들에 대해  $\lambda_1$ 과  $\lambda_2$ 가 각각 일정한 값을 갖기 때문인 것으로 보인다.

주어진 공식에 의해<sup>19)</sup>  $\pi, \lambda_1, \lambda_2$ 의 값이 정해지면 하터가 제시한 주제어와 비주제어의 식별공식에 의해 색인어를 선정할 수 있다. 각 단어에 대해 문헌집단이 두개의 클래스로 나누어진다고 가정한 2-포아슨모형에서 각 클래스에 속하는 문헌이 특정한 정보요구에 적합한 확률을  $u_1, u_2$ 라고 한다면  $u_1 \gg u_2$ 가 되어야 한다. 따라서 가장 좋은 색인어는  $u_1$ 이  $u_2$ 보다 훨씬 큰 단어가 될 것이다. 하터는 어떤 단어가 색인어로 적합한가는 이 단어가 문헌들을 얼마나 명확히 두 클래스로 구분짓는가에 달려있다고 보고 두 클래스사이에 중복도(degree of overlap)가 클수록, 즉  $\lambda_1$ 이  $\lambda_2$ 에 가까울수록 두 클래스는 잘 구분되지 않게 되고 따라서  $u_1$ 은  $u_2$ 에 가깝게 되므로 이러한 단어는 색인어로 적합하지 않으며  $\lambda_1 \gg \lambda_2$ 이면  $u_1 \gg u_2$ 가 되어 이러한 단어는 색인어로 적합하다고 보았다. 위의 가설 아래 색인어의 적합성 판정을 위하여 다음공식과 같이 Z치로 중복도를 측정하였다.

$$Z = \frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}}$$

위 공식에서  $\lambda_1$ 이  $\lambda_2$ 에 가까울수록 Z치가 적어지므로 Z치가 큰 단어가 좋은 색인어가 될 수 있는 것이다.

심리학분야의 문헌초록을 대상으로 한 실험에서 하터는 183개의 주제어와 175개의 비주제어의 Z치를 구한 결과 주제어들은 비주제어에 비해 큰 값의

19) Harter, op. cit., p.202.

Z치를 갖는 비율이 훨씬 높음을 발견하므로써 Z치를 색인어 식별기준으로 사용할 수 있다는 것을 입증하였다. 이 실험에서  $Z \geq 1.5$  범위에서는 모든 단어가 주제어였고  $\leq Z 0.5$  범위에서는 3/4이 비주제어였다.

#### 4. 한국어문헌의 색인어선정실험

본장에서는 앞에서 고찰한 통계적 색인어선정기준인 단어빈도론, 문헌빈도론, 하터의 확률이론을 한국어문헌을 대상으로 실험하여 단어의 출현빈도에 관한 통계적인 현상과 여기에 기초한 색인어 선정기법이 한국어문헌에도 적용가능한가를 평가해 보았다.

##### 4.1 단어빈도론의 실험

한 문헌에 출현한 단어들을 출현빈도순으로 순위를 매겼을 때 최고빈도, 최저빈도의 단어들을 제외한 중간빈도의 단어들이 색인어로 적합하다는 이론을 농학분야 한국어문헌에 적용하여 보고 고프만의 전환점공식을 이용한 파오의 색인어 선정기법을 실험해 보았다.

실험대상 문헌은 2개로 하여 파오의 실험결과와 비교하였는데 하나는 한국축산학회지(1978)에 실린 박희규 저 “돈 냉동 정액 연구”이고 다른 하나는 한국농공학회지(1980)에 실린 김진하 등저 “동진강 제수문 시공”이라는 논문이었다.

첫번째 실험에서는 다음과 같이 단어통제를 가한 다음 단어빈도를 계산하였다.

- (1) 단어는 명사, 조사, 동사 등 9품사로 구분하였다.
- (2) 동음이의어는 한 단어로 취급하였다.
- (3) 단어구분은 관사, 조사 외에는 띄어쓰기가 이루어진 것을 한 단어로 인정하였다.
- (4) 논문의 본문 즉, 서론, 본론, 결론부분에 나타난 단어만을 대상으로 하고 초록, 표제, 저자명, 참고문헌, 주기, 서지사항, 특수기호, 도표, 삽도, 수학기식, 감사문 등은 제외하였다.
- (5) 동사와 형용사는 어미변화를 원형으로 통제하지 않고 어미가 다른 단

어는 각각 별개의 단어로 취급하였다.

위와 같은 원칙하에 단어의 출현빈도를 계산한 결과 문헌 1에 나타난 단어 유형은 187개, 출현빈도가 한번인 단어의 수는 112개였으며 문헌 2는 단어 유형 431개, 출현빈도 한번인 단어의 수는 332개였다.  $I_n$ 이 1이 되는 전환점에 오는 단어의 빈도  $n$ 을 계산한 결과 문헌 1의 경우가 29, 문헌 2가 50으로 29는 문헌 1의 최고출현빈도인 30(“에”)에 너무 가깝고 50은 문헌 2의 최고출현 빈도인 32(“이”)를 초과하는 값으로 파오의 방식이 두 문헌에 모두 적합하지 않음을 볼 수 있다.

두번째 실험에서는 영어가 단일어를 원칙으로 함을 감안하여 한국어도 단일어로 분리하여 통계를 낸 결과 문헌 1에서는 단어유형 181개,  $I_1$ 은 97개, 전환점의 빈도는 27이었고 문헌 2에서는 단어유형 449개,  $I_1$ 은 325개, 전환점의 빈도는 50으로 나타났다.

세번째 실험에서는 다시 동사와 형용사의 어미변화를 통제하여 같은 어간을 갖는 단어들은 한 단어로 취급하였다. 영어의 동사가 과거, 과거분사, 현재와 같이 구분되는 점을 감안하여 한국어의 동사형을 능동형 현재와 과거(～하다, ～하였다), 수동형 현재와 과거(～되다, ～되었다)와 같이 귀속 통제하였다. 이 결과 문헌 1에서는 단어유형이 165개,  $I_1$ 은 79개, 전환점의 빈도는 24, 문헌 2에서는 단어유형이 401개,  $I_1$ 은 277개, 전환점의 빈도는 46으로 각각 나타났다. 여기에서 동사를 다시 현재원형으로 통제한다고 해도 전환점은 크게 달라지지 않을 것으로 보여 더 이상 통제를 가하지 않았다.

표-1과 표-2는 세번째 실험결과 문헌 1과 문헌 2에 나타난 단어의 출현빈도 순위로서 대략 중간빈도를 갖는 단어들 가운데 주제어(고딕체로 표시)가 많이 들어있음을 알 수 있다.

문헌 1에서 출현빈도 1의 단어들과 문헌 2에서 출현빈도 3이하의 단어들은 거의가 비주제어로 판단되어 표에 수록하지 않았다. 문헌 1(표-1)에서는 상위빈도의 비주제어와 주제어의 빈도구분이 뚜렷하지 않으나 대부분의 저빈도 단어는 비주제어임이 확실하였다. 특히 를, 는, 가, 의, 으로 등과 같은 비주제어가 중간빈도를 갖게된 것은 문헌 1에 출현한 단어유형수가 적고

〈표 1〉 단어빈도 순위(문헌 1)

순위	빈도	단 어
1	28	에
2	20	은
3	17	회석
4	16	이, 정액
5	14	액, 시험
6	13	을
7	10	에서
8	9	가장, 대한
9	8	및, 냉동, 로, 구(區)
10	7	시간, 는, 과, 의, 방법
11	6	냉충격, 가, 제조, 같다, 냉각
12	5	온도, 돈(豚), 저항성, 으로, 증류수, 까지, 이상
13	4	충격, 를, 다음, 좋다, 냉동성, 위하다
14	3	관하다, 때, 생겼다, 삼투압, 본, 연구, 의하다, 에서는, 이다, 이었다, 처리, 강했다, 적합하였다, 생존율, 원정액
15	2	결과, 도, 등, 로서, 와, 실시하였다, 성적, 소요되다, 우수하다, 저온, 첨가, 개발하다, 가하다, 있다, 좋았다, 재료, 냉각시키다

〈표 2〉 단어빈도 순위(문헌 2)

순위	빈도	단 어
1	32	이
2	27	을
3	22	가
4	21	에
5	20	으로
6	19	의
7	18	를
8	15	말록, 에서
9	14	을
10	12	로
11	11	는
12	10	있다, 에서는, 것
13	9	지반

14	8	시공
15	7	작업, 같다, 보인다, 관입
16	6	시간, 제수문, 심도
17	5	도, 흙, 투입, 기계, 철널, 계화도, 간척지, 기초, 수
18	4	동진강, 과, 방법, 값, 까지, 트럭, 의하다, 위하다, 이다,

문헌길이 비교적 짧기 때문이 아닌가 생각된다. 문헌 2(표-2)에서는 고빈도와 저빈도의 단어들은 거의가 비주제어이고 대부분의 주제어는 중간빈도의 단어 그룹에 속해 있는 것으로 보인다.

결과적으로 이 실험결과 전환공식에 의한 색인어 선정기법은 한국어문헌에는 적용되지 않는 것으로 보이나 문이 제시한 단어빈도론의 기본원칙은 대략 적용되는 것을 볼 수 있다. 따라서 위 각 실험에서 전환점의 빈도가 낮아지면 파오의 기법을 적용할 수 있으리라고 보여진다. 파오가 사용한 고프만의 전환점공식은 지프의 제 2 법칙에 근거하여 만들어진 것이므로 위 공식이 적용되지 않는 원인으로 저빈도의 한국어단어들의 출현양상이 영어와는 다르리라는 가능성을 생각해 볼 수가 있다. 결론적으로 이러한 실험에 앞서 한국어단어의 출현빈도현상을 표집을 크게 하고 보다 깊이있게 연구하여 실제현상에 맞게끔 법칙을 수정한 다음 색인어선정에 이용해야 할 것으로 보인다.

#### 4.2 문헌빈도론의 실험

하터의 확률이론을 실험한 농학문헌집단을 대상으로 하여 문헌빈도를 측정한 결과 대부분의 주제어가 문헌빈도 1을 갖는 것으로 나타나서 실험이 불가능하였다. 즉 30문헌으로 구성된 표집이 너무 적어 대부분의 문헌이 다른 주제를 다루고 있기 때문에 각각의 주제어가 대부분 한 문헌에만 나타나는 현상을 보인 것으로 판단된다. 이 문헌빈도론은 추후 표집을 크게하여 셀톤의 문헌분리기에 의한 선정기법과 함께 실험해 볼 예정이다.

#### 4.3 하터의 확률이론의 실험

2-포아손 분포모형의 주제어와 비주제어 분리기능이 한국어단어들에 어느 정도 적용되는가를 보기 위해 하터의 색인어 선정기법을 실험하였다.

실험규모는 문헌수 30, 단어유형 350, 문헌당 평균단어출현수 70으로 설



계하였다. 실험대상 문헌집단은 농학분야의 논문으로 구성되며 주제별로는 작물, 원예, 식물병리분야에서 각각 10편씩의 논문을 임의로 선택하였다. 문헌수를 30개로 한정된 것은 통계학적으로 타당한 표본크기와 실험능력을 감안하였기 때문이며, 모든 파라미터 값의 계산은 컴퓨터에 의해 처리하였다. 단어유형을 350개로 한 것은 하터의 실험에서의 표집크기인 358단어유형에 가깝게 하기 위하였으며 문헌집단으로부터 무작위로 추출하고 한 단어당 한 장의 펀치카드를 사용하여 입력하였다.

이 태영 논문의<sup>20)</sup> 농학분야 디소오러스에 수록된 단어들을 주제어로 간주하여 358단어를 분류한 결과 주제어가 135개, 비주제어가 200개였다. 이 가운데 예로 10개 단어를 뽑아 각 단어의 문헌분포양상을 도식하면 표-3과 같다.

〈표 3〉

10개어의 문헌빈도분포

			k토큰을 갖는 문헌수																
빈도	단어유형		0	1	2	3	4	5	6	7	8	9	10	...	19				
7	감	굴	29	0	0	0	0	0	0	1	0	0	0	(전부 0)	0				
13	경	과	23	3	3	0	1	0	0	0	0	0	0		0				
12	과	실	25	2	1	0	2	0	0	0	0	0	0		0				
10	관	옥	29	0	0	0	0	0	0	0	0	0	1		0				
11	변	화	22	6	1	1	0	0	0	0	0	0	0		0				
10	비	교	22	6	2	0	0	0	0	0	0	0	0		0				
11	분생	포자	26	1	1	1	0	1	0	0	0	0	0		0				
3	배	양 토	28	1	1	0	0	0	0	0	0	0	0		0				
1	동백	나무	29	1	0	0	0	0	0	0	0	0	0	0					
20	고	추	28	1	0	0	0	0	0	0	0	0	0	0	1				

입력된 358단어의 출현빈도분포로부터 하터가 사용한 방법에 의해  $\lambda_1, \lambda_2, \pi$ 의 값을 구하고 다시 Z치를 계산하였는데 표-3에 예시한 10개 단어에 대한 세 파라미터의 값과 Z치를 표-4에 수록하였다. 또한 실험단어군을 주제어와 비주제어로 나누어 Z치에 따른 분포현상을 그림-1에 도식하였다.

20) 이태영. 농학문헌의 한국어 색인시스템에 관한 연구. 석사학위논문, 연세대학교 대학원, 1982.

〈표 4〉 10개어의  $\pi, \lambda_1, \lambda_2, Z$  값

단 어 유 형		$\lambda_1$	$\lambda_2$	$\pi$	$Z$
감	귤	6.00000	.00000	.03889	2.44949
경	과	1.38462	.00000	.31296	1.17670
과	실	2.16667	.00000	.18462	1.47196
관	옥	9.00000	.00000	.03704	3.00000
변	화	.75072	.02239	.47269	.82834
비	교	.40000	.00000	.83333	.63246
분	생 포 자	2.54545	.00000	.14405	1.59545
배	양 토	.66667	.00000	.15000	.81650
동	백 나 무	.00000	.00000	1.00000	-99.99899
고	추	17.10000	.00000	.03899	4.13521

〈그림 1〉 주제어와 비주제어의 Z치분포

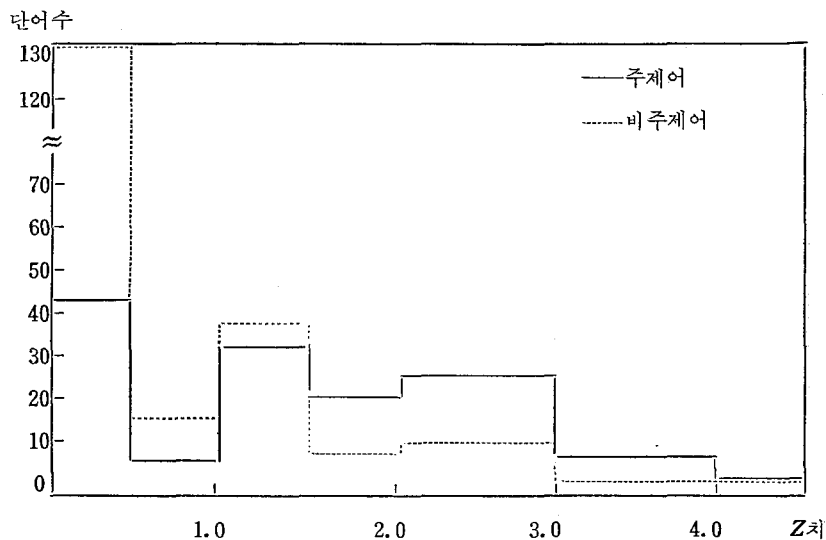


표-3에 수록된 단어들의 출현현상과 Z치와의 관계를 상세히 살펴보면 “감귤”, “과실”, “관옥”, “분생포자”, “배양토”, “동백나무”, “고추”는 주제어이고 “경과”, “변화”, “비교”는 비주제어인바 주제어중 Z치가 1이하인 것은 “배양토”, “동백나무”였고 비주제어중 Z치가 1이상인 것은 “경과”로서 모두 10개 단어중 3개가 기대치에서 벗어나고 있다. 특히 비주제어이면서 높

은  $Z$ 치를 갖는 단어들이 문제가 되는데 “경과”의 경우에는 실제빈도분포가 단일포아슨을 닮아가다가 2-포아슨으로 돌아선 것을 볼 수 있다. 이것은 하터의 실험문헌집단의 크기에 비해 본 실험집단의 크기가 작은데에 이유가 있는 듯이 보이며 실험집단이 클 경우에는 3, 3, 2, 1, 0과 같이 빈도가 분포되어  $Z$ 치가 낮아질 것을 기대할 수 있다.

그림-1을 하터의 실험결과와 비교해 보면  $Z \geq 1.5$ 에서 비주제어가 많이 출현한 현상과  $0.5 \leq Z < 1.5$ 에서 비주제어가 주제어보다 더 많이 출현한 현상은 결과가 상이하나  $Z < 0.5$ 인 단어들 가운데 25%정도가 주제어인 것은 하터의 실험결과와 유사하였다. 정보검색효율의 측정단위인 재현율과 정확율을 응용하여 이 실험결과를 다음과 같이 평가할 수 있다. 즉  $Z \geq 1$ 에 오는 주제어는 전체주제어의 66%로서  $Z=1$ 을 한계치로 정했을 때 주제어의 재현율 66%를 기대할 수 있으며  $Z \geq 1$ 에 오는 단어 가운데 29%는 비주제어로서 정확율은 대략 70%정도를 기대할 수 있다.

## 5. 결 론

1960년대 이후 자동색인에 관한 연구는 끊임없이 계속되고 있으며 자동색인 결과와 수작업색인 결과를 비교하는 실험도 다수 발표되었다. 자동색인의 목표는 수작업색인에서와 같이 주제성이 큰 색인어 추출에 있으며 많은 실험에서 자동색인결과가 수작업색인 결과보다 크게 떨어지지 않는다는 것을 입증하려고 하였다. 그러나 현재로는 주제어선정에 있어서 자동색인이 수작업색인을 대신하기는 어려우며 단지 문헌의 본문 내지는 조목이 기계가독형으로 변환되어 데이터베이스가 만들어져 있는 경우에는 수작업색인의 보조 수단으로 자동색인을 고려할 수 있을 것으로 보인다. 실제로 다이알로그와 같은 대규모 정보검색시스템에서는 색인자가 부여한 색인어와 자동색인방법에 의해 데이터베이스로부터 추출한 색인어를 기본색인에 함께 수록하여 검색에 이용시키고 있다.

본 논문에서는 현재까지 연구되어 온 자동색인기법을 종합적으로 고찰하였고 특히 통계적기법에 역점을 두어 기술하였다. 또한 통계적 색인기법에

서 사용하고 있는 색인어 선정기준 가운데 단어빈도론에 속하는 파오의 전환공식기준과 문헌·단어빈도론이라고 할 수 있는 하터의 확률이론을 한국어로 쓰여진 농학문헌을 대상으로 실험하여 보았다. 본 실험에서는 파라미터의 계산에는 컴퓨터를 사용하였으나 단어의 출현빈도는 수작업으로 계산하였기 때문에 실험집단의 크기를 제한할 수 밖에 없었다.

그러나 하터의 확률이론의 적용결과는 상당히 고무적이며 앞으로 실험집단의 규모를 크게하여 재실험해 볼 가치가 충분히 있는 것으로 보인다. 반면 파오가 사용한 전환공식에 의한 색인어선정기준은 전혀 적용되지 않았는데 이유로는 한국어단어의 구분방법과 사용패턴이 영어와는 차이가 있기 때문으로 보인다.

통계적기법은 단어의 출현빈도라는 통계적 현상에 기초를 둔 것이므로 우선 색인하고자 하는 문헌에서 사용된 특정한 언어에 관한 통계적 특성을 파악하는 일이 선행되어야 할 것이다. 앞으로 계속적인 연구를 통해서 한국어단어의 통계적특성을 파악하고 주제어와 비주제어의 분포현상을 규명한다면 자동색인뿐만 아니라 일반 색인이론을 발전시키는데 큰 공헌을 할 수 있을 것으로 생각된다.

## Statistical Techniques for Automatic Indexing and Some Experiments with Korean Documents

Young Mee Chung\*

Tae Young Lee\*\*

### Abstract

This paper first reviews various techniques proposed for automatic indexing with special emphasis placed on statistical techniques. Frequency-based statistical techniques are categorized into the following three approaches for further investigation on the basis of index term selection criteria: term frequency approach, document frequency approach, and probabilistic approach.

In the experimental part of this study, Pao's technique based on the Goffman's transition region formula and Harter's 2-Poisson distribution model with a measure of the potential effectiveness of index term were tested. Experimental document collection consists of 30 agriculture-related documents written in Korean. Pao's technique did not yield good result presumably due to the difference in word usage between Korean and English. However, Harter's model holds some promise for Korean document indexing because the evaluation result from this experiment was similar to that of the Harter's.

---

\* Associate Professor, Yonsei University.

\*\* Instructor, Soong Eui Woman's Junior College.