2-포아송 모형을 이용한 한글 주제어 선정에 관한 연구

A Study on the Applicability of 2-Poisson Model for Selecting Korean Subject Words

정영미(Young-Mee Chung)*, 최대식(Dae-Shik Choi)**

목 차

초 록

최근 구축된 한글 실험문헌 집단을 대상으로 2-포아송 모형의 Z값의 주제어 식별력을 측정하였으며, 역문헌빈도와 2-포아송 모형간의 상관관계를 분석하였다. 이를 위해, Z와 수정 β값 및 IDF와 수정 TF · IDF 가중치를 하위 실험집단별로 각기 산출, 비교하였다. 실험 결과 Z값의 주제어 선정능력은 3개의 하위 실험집단 가운데 과학기술분야에서만 확인되었다. 2-포아송 모형의 Z값과 역문헌빈도 가중치간의 상관관계 분석에서는 전문(full text)인 여성학분야 실험집단에 비해 초록 및 신문기사와 같이 단문(short text)으로 구성된 과학기술분야 및 일반사회 분야 실험집단에서 상관관계가 더 크게 나타났다.

ABSTRACT

Experiments were performed on three subsets of a Korean test collection in order to determine whether 2-Poisson model's Z value is a good measure for selecting subject words from a document to be indexed. It was found that subject word selection based on the Z value was effective for only one subset with short texts, i.e., the Science and Technology subset. Correlation analyses between 2-Poisson model's Z and TF · IDF weight for the three subsets showed that the correlation was relatively high for two test subsets with short texts, i.e., the Science and Technology subset and the Newspaper subset.

키워드 : 자동색인, 포아송 분포모형, 2-포아송 분포모형, 역문헌빈도 가중치, 주제어 선정

* 연세대학교 문헌정보학과 교수
** (주) 오름정보 개발부
■ 논문 접수일 : 2000년 2월 23일

1 서 론

문헌 내 단어의 빈도에 근거한 확률적 자동색인 기법의 기본가정은 첫째, 단어의 출현빈도가 문헌의 주제를 정하는 기준이 되며, 둘째, 주제어는 무작위로 출현하지 않고 소수의 문헌에 집중적으로 출현한다는 것으로 요약된다. 이로써 통계적 자동색인의 기틀이 마련된 것이며, 이러한 가정에 충실한 대표적인 연구성과 가운데 하나로 하터(Harter 1975)의 2-포아송 분포모형을 꼽을 수 있다.

2-포아송 분포모형은 크게 두 가지 용도로 적용되어 온 것으로 평가할 수 있다. 하나는 정보검색 관련 이론을 보완하기 위한 것이었고, 다른 하나는 색인어 및 검색어에 대한 가중치 부여기법으로 적용하기 위한 목적이었다.

전자의 예인 북스타인(Bookstein 1977)의 연구에서는 스웨츠(Swets 1963)의 정보검색 이론의 약점이 2-포아송 모형에 의해 부분적으로 해결된다고 주장하였다. 후자의 예로는 적합성 피드백검색의 초기탐색 가중치 부여 방안으로 2-포아송이 이진독립(binary independence) 모형보다 낫다는 연구가 있다(Raghavan, Shi, and Yu 1983). 마찬가지로 문성빈(1999)은 초록이나 제목으로부터 검색하는 경우에는 2-포아송 모형에 비해 이진독립 모형이 적절하지만, 전문(full text) 검색에서는 이진독립 모형보다 2-포아송 모형이 검색효율을 증진시킨다는 점을 MEDLINE 데이터베이스를 통한 검색실험으로 입증하였다.

한편, 2-포아송 모형의 Z값을 한글 문헌의 자동색인에 적용한 정영미와 이태영(1982)의 연구에서는 농학 관련 논문을 대상으로 주제어 선정 실험을 수행하였다. 실험결과, '감귤'과 같은 주제어는 Z값이 비교적 큰 반면, 비주제어인 '변화' 등에는 1이하의 작은 값이 산출되었다. 그러나 Z값이 1.5 이상인 경우와 0.5 에서 1.5 사이인 경우 주제어보다 오히려 비주제어가 더 많이 추출되어, 2-포아송 모형이 한글문헌에서의 주제어 선정에 잘 적용되지 않음을 보여 주고 있다.

본 연구에서는 한글문헌을 대상으로 한 실험 및 다른 색인어 가중치 기법과의 비교를 통해 2-포아송 분포모형이 한글문헌의 자동색인에 활용될 수 있는지 재검토하였다.

2 포아송 분포모형

2.1 2-포아송 분포모형

문헌내에 무작위로 출현하는 단어의 분포는 단일 포아송 함수(single Poisson function)를 이용하여 정의할 수 있으며, 그 기본적인 공식은 다음과 같다(Harter 1975).

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

- $f(k)$: 문헌 내 단어가 k번 출현할 확률
- e = 2.71828. 자연대수(natural logarithm)의 밑수
- λ = 문헌클래스내 단어의 평균 출현빈도

이와 같이 포아송 함수를 이용하면 비주제어의 출현확률을 파악할 수 있다. 그러나 자동색인의 궁극적인 관심사인 주제어에 관해서는 주제어가 이 분포를 따르지 않는 정도에 대한 파악만이 가능하다. 따라서 포아송 함수를 발전시키면 비주

제어 대신 주제어의 분포를 직접 표현할 수 있는 모형을 도출할 수 있을 것이라는 판단에서 2-포아송 분포모형이 고안되었다(Bookstein and Swnason 1974; Harter 1975).

2-포아송 분포모형은 포아송 함수를 합성한 공식이며 전체문헌 클래스를 적합문헌 클래스와 부적합문헌 클래스로 임의로 구분한다.

따라서 2-포아송 분포모형에 따라 특정단어가 한 문헌에 k번 등장할 확률함수는 다음과 같이 주어진다(Harter 1975).

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

- $f(k)$: 단어가 한 문헌에 k 번 출현할 확률
- π : 문헌클래스 Ⅰ(적합문헌 클래스)에 속하는 문헌의 비율
- $1-\pi$: 문헌클래스 Ⅱ에 속하는 문헌의 비율
- λ_2 : 문헌클래스 Ⅱ(부적합문헌 클래스)에서 단어의 평균빈도
- λ_1 : 문헌클래스 Ⅰ에서 단어의 평균빈도

단어가 색인어로서의 중요성을 갖기 위해서는 문헌클래스 I과 II를 구분하는 능력이 커야 하므로 두 클래스간의 중복도를 가능한 한 줄어들게 하는 단어를 효율적인 색인어로 규정할 수 있으며, 이에 대한 기준을 제시하기 위해 하터는 브룩스(Brookes 1968) 측정법을 이용하였다.

$$Z = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

- μ_1, μ_2 : 적합, 부적합 문헌이 검색될 확률분포함수 각각의 평균
- $\sigma_1^2 + \sigma_2^2$: 적합, 부적합 문헌이 검색될 확률분포함수 각각의 분산

포아송 함수에서는 μ와 σ로 표현되는 평균과 분산이 같다고 정의된다는 성질을 이용해서 공식을 수정하면 다음과 같은 색인어 측정기준이 제시될 수 있다.

적합/부적합 문헌분리능력:

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

식에서 λ_1, λ_2값의 차이가 작다면 두 문헌클래스 사이의 중복도가 증가함으로써 Z값은 감소하고, 반대로 Z값이 크다면 두 문헌클래스 사이의 중복도가 그만큼 작다는 의미가 된다.

상기 Z값으로는 전체 문헌클래스에서 적합문헌 클래스와 부적합 문헌 클래스간 중복이 덜 되는 정도만 파악되므로, 단어에 대한 색인어로서의 가치가 β값을 통해 표현되도록 함으로써 2-포아송 분포모형이 아래와 같이 완성된다.

$$\beta = P(d \in I / K) + Z > 0$$

여기서 단어 w가 문헌 d에서 k번 발생하며 문헌클래스 I에 속할 확률은

$$P(d \in I / K) = \frac{P(d \in I / K)}{P(K)}$$

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

과 같이 정리되며, Z는 앞서 정의한 바와 같다. 즉, 특정단어에 의해 표현되는 문헌의 개념이나 주제의 취급정도는 $P(d \in I / K)$에 의해 정의되며, 이 단어가 주제어로서 지닌 가치는 Z값으로 정의될 수 있다는 것이다.

2.2 다중포아송 분포모형

다중포아송 분포모형은 포아송 모형의 항을 점차 늘리는 방식으로 생성할 수 있다. 바꾸어 말하면, n-포아송 함수의 n을 '2, 3, 4, · · · , n'으로 확장하여 n의 수만큼 항을 결합시키는 것이 얼마든지 가능하며, 엄밀한 의미로는 2-포아송 모형 역시 다중 포아송 모형의 한 종류에 포함된다.

2-포아송 모형이 성립하기 위해, 적합문헌 클래스와 부적합문헌 클래스로 전체문헌을 양분한다는 전제가 있었다. 이러한 이분법은 공식의 단순화를 기하는 데 긴요하지만, 문헌집단에 적합문헌 아니면 부적합문헌 두 클래스만 존재한다는 가정은 비현실적이므로 전체 문헌집단을 2개 아닌 3개 클래스로 다시 구분할 필요성이 제기되었으며, 이 결과 다음과 같은 3-포아송 분포모형이 제시되었다(Srinivasan 1990b).

$$f(k)=\pi_1 \frac{e^{-\lambda_1}\cdot \lambda_1^{k}}{k!} + \pi_2 \frac{e^{-\lambda_2}\cdot \lambda_2^{k}}{k!} + (1-\pi_1-\pi_2) \frac{e^{-\lambda_3}\cdot \lambda_3^{k}}{k!}$$

3-포아송 모형의 특징은 원하는 정도에 따라 하위 문헌클래스를 탄력적으로 구분할 수 있다는 것으로 집약된다. 문헌클래스를 3개로 확장한 결과, 공식의 첫 항은 매우 적합한 문헌클래스(I), 둘째 항은 보통 정도의 적합 문헌클래스(II), 그리고 마지막 항은 부적합한 문헌클래스(III)라고 정의할 수 있다는 것이다.

그렇지만 3-포아송 모형에서는 2-포아송 모형에 2개가 추가되어 파라미터가 모두 5개나 된다는 점에서 파라미터 추정에 소요되는 추가 비용을 상쇄할 수 있음을 입증하지 못하는 한, 실용화가 용이하지는 않을 것으로 보인다(〈표 1〉 참조).

〈표 1〉 2-포아송, 3-포아송, 다중포아송 모형 비교

방법 및 결과 \ 모형	2-포아송(2-P)	3-포아송(3-P)	다중포아송(n-P)
대상 자료	프로이드 저서 초록 650건	INSPEC DB 초록 59,919건, 어간 196개	Financial Times (1985년판) 기사 6,750건, On-Line News Group의 영화비평 1034건, New York Times(1991년판)기사 1,084건
대상 자료의 평균길이	223 단어	58 단어	최소 400단어 이상
파라미터	π, λ_1, λ_2	π_1, π_2, λ_1, λ_2, λ_3	nP의 n 개수에 따라 가변적, n이 커질수록 파라미터 수 증가
추정 방법	적률법	적률법	최우 추정법
실험 결과 및 유효성	유용한 색인어의 38%가 2-P 분포 자동색인에서의 2-P 적용 가능성언급	어간 196개 가운데 43%가량이 1-P나 3-P 아닌 2-P 분포 3-P보다 2-P가 우수함	70%의 단어가 nP 분포를 따르며 대부분 2P, 3P, 혹은 4P에 해당 다중포아송 분포모형의 유용성 주장
문헌길이 고려	×	×	×

이와 반대로, 다중포아송 모형이 2-포아송 및 3-포아송 모형의 문제점을 해결할 수 있다는 연구가 있다(Margulis 1993). 다중포아송 모형에 대한 이 실험결과는 첫째, 고빈도 단어 중 약 70퍼센트가 다중포아송 분포 유형이었고, 둘째, 다중포아송 모형의 단어 대부분이 2-포아송, 3-포아송, 혹은 4-포아송 모형에 해당하였으며, 셋째, 고빈도어일수록 다중포아송 모형의 n의 크기가 증가하는 특성을 보였다는 것으로 요약된다. 이 연구는 최근에 간행된 대규모 장서가 대상이었으며, 단어 대신 어간(stem)을 분석했다는 점에서 앞선 연구들과 차이가 있다. 또한, 어간에서 확장하여 n-gram이나 숙어의 분포양상이 다중포아송 모형으로 설명 가능한지 여부와 비 영어권 문헌에도 적용되는지에 관한 향후 연구의 필요성을 제안하였다.

앞에서 언급한 하터, 스리니바산, 마굴리스의 2-포아송, 3-포아송, 다중포아송 분포모형에 대한 비교는 〈표 1〉과 같다.

3 주제어 선정 실험

3.1 실험설계

3.1.1 실험문헌집단

본 연구에서 실험한 문헌집단은 연구개발정보센터(KORDIC)가 1999년 10월 현재 구축 중인 HANTEC(Hangul Test Collection)이다(맹성현 외 1999).

HANTEC은 일반사회, 사회과학, 과학기술분야에 속하는 120,000건(약 244MB)의 문헌집합이다(〈표 2〉 참조). 문헌집단은 문서집합, 적합성 판정에 이용된 30개의 질의집합, 전문가 또는 비전문가에 의해 판정이 내려진 적합한 문서집합의 세 부분으로 구성되어 있다.

대상문헌 내 단어 중에서 명사 상당어를 일차적으로 선정하였으며, 이 가운데 중간빈도어 선정기준 두 가지를 적용하여 최종적으로 826개의 단어를 가지고 실험하였다.

3.1.2 실험내용 및 방법

단어분석 도구로는 HAM(Hangul Analysis

〈표 2〉 HANTEC 문헌 통계

문헌 집합	문헌 수	최대 바이트	최소 바이트
한국일보	22,000	11,764	52
웹 문헌(.com, .gov 도메인)	18,000	361,882	82
매일경제신문	39,480	25,361	73
한국여성개발원 게재 논문	110	288,714	9,024
경북도의회 회의록	410	346,112	1,918
과기처 지원 연구보고서	10,000	5,532	206
해외 과학기술 동향	18,000	130,582	193
학술논문 서지사항	12,000	2,822	272

Module; version 4.0a)을 이용하여 과학기술, 여성학 및 일반사회 분야 문헌 500건에 출현한 단어에 대하여 불용어 제거, 형태소분석 등의 작업을 수행하였다. 문헌빈도, 장서빈도, 문헌별 단어빈도를 파악하여 가중치와 공식을 계산하는 데에는 SQL DBMS와 ACCESS를, 표와 그래프 작성에는 Excel을, 그리고 피어슨 상관계수 검증에는 SPSS를 이용하였다.

분석 대상어는 원칙적으로 명사로 한정하였으나 복합명사 가운데 색인어가 될 수 있다고 판단되는 단어도 포함시켰다. 그 예로 '핵문제', '가정폭력' 등이 있다.

실험대상으로 선정한 500개 문헌에서 중간 빈도어를 선정하기 위한 첫 기준은, 설튼 등(Salton, Yang and Yu 1975)이 제시한 것으로 문헌빈도가 n/100에서 n/10에 해당하는 단어로 제한하는 방법이다. 이로써 문헌 수가 200개인 과학기술 및 일반사회 분야에서는 문헌빈도가 2에서 20인 단어를 추출하였으며, 문헌 수가 100개인 여성학 분야에서는 문헌빈도가 1에서 10에 해당하는 단어가 우선 추출되었다.

이와 같이 일차적으로 문헌빈도를 제한한 결과, 각 문헌집단에서 선정된 단어는 과학기술 분야가 10,309개였으며 여성학 분야는 55,778개, 일반사회 분야는 11,101개였다. 이를 전부 대상으로 삼기에는 단어 수가 여전히 많다고 판단되어, 고빈도어와 저빈도어는 주제어로 적합하지 않다고 본 다메로(Damerau 1965)의 가설에 따라 장서빈도 범위를 설정하여 단어 수를 더 줄였다. 본 실험에서 적용한 방안은 각 문헌집단 내 장서빈도가 상위 25% 및 하위 25%에 해당하는 단어를 제외한 단어, 즉 장서빈도가 중간 50%에 속하는 단어를 선정하는 방식이다.

그 결과, 과학기술분야는 대상단어의 장서빈도가 14에서 38로 설정되었고 이에 해당하는 단어 수가 272개로 나타났으며, 여성학분야는 장서빈도가 30에서 96으로 설정되었고 이에 해당하는 단어 수는 407개인 반면, 일반사회 분야에서는 장서빈도가 14에서 37로 좁혀졌고 단어 수는 258개였다.

과학기술분야의 272개 단어 가운데 34개는 2-포아송 모형의 Z값 계산과정에서 에러가 발생하거나 분석의 가치가 없는 것으로 보이는 단어들로 파악되었는데, 대표적인 예로는 '연구개', '으로' 등이 있었다. 마찬가지로 여성학분야의 407개 및 일반사회 분야 258개 단어 가운데에도 'ㄱ백화점', '아동지'와 같이 형태소 분석상의 오류이거나 인명, 지명, 혹은 Z값 계산이 불가능한 단어가 분야별로 48개 및 29개씩 섞여 있었다. 이러한 오류를 제외하여 최종 선정한 단어는 〈표

〈표 3〉 실험대상 문헌 통계

분야	문헌 집합	문헌 수	최대 바이트	최소 바이트	분석단어	명사 상당어	실험 대상어
과학기술	과기처 지원 연구보고서	100	4,265	657	11,232	10,309	238
	해외 과학 기술 동향	100	2,266	256			
여성학	여성개발원 논문	100	292,829	13,044	61,126	55,778	359
일반사회	한국일보	200	5,326	204	11,283	11,101	229

3〉과 같이 과학기술 분야 229개 및 여성학 분야 359개, 그리고 일반사회 분야 258개이다.

실험을 위해 먼저 대상단어의 Z값과 IDF를 산출하였으며, 주제어로서의 최종 가중치로서 2-포아송 모형의 β를 응용한 수정 β와 IDF를 응용한 수정 TF · IDF를 산출하였다. 다음 단계에서는 단어의 Z값과 IDF 가중치간의 상관관계 및 수정 β와 수정 TF · IDF 가중치간의 상관관계를 파악하였다. 2-포아송 모형을 적용하였을 때 Z 또는 수정 β값이 큰 단어가 IDF 및 수정 TF · IDF 가중치에서도 유사한 양상을 보이는지 파악하고자 하는 것이었다.

역문헌빈도(IDF)는 스파크존스(Sparck Jones 1972)의 공식을 사용하였으며, 여기에 문헌내 단어 출현빈도(TF)를 곱해 TF · IDF를 산출하였다.

$$IDF = \log_2 N - \log_2 DF + 1$$

· N : 장서 내 문헌 수
· DF : 문헌빈도

수정 β와 수정 TF · IDF 가중치는 본 실험의 목적에 부합되도록 정규화시켜 수정, 제안한 공식이다. 수정 β를 산출하는 과정은 다음과 같다. 문헌 내 빈도(k)에 따른 개별적인 β값을 산출한 뒤 이에 대한 평균값을 계산하여 이를 수정 β값(표준화시킨 2-포아송 모형 가중치)으로 간주하였다. 예를 들어, 2-포아송 모형에 따라 특정단어의 β값이 문헌 내 빈도(k)별로 $\beta(k|1) = 2$, $\beta(k|2) = 3$, $\beta(k|3) = 1$, $\beta(k|4) = 2$와 같이 파악되었다면 이를 합산하여 문헌빈도를 갖는 경우의 수인 4로 나눈 평균값 2가 수정 β값이 된다. 즉, 수정 β 공식은 본 연구에서 아래와 같이 정의된다.

$$\text{수정 } \beta = \frac{1}{n} \sum_{k=1}^{n} \beta_k$$

한편 수정 TF · IDF는 단어의 평균 TF · IDF 값이다. 이를 최종 역문헌빈도 가중치로 간주하였으며, 본 실험에서 정의한 공식은 다음과 같다.

$$\text{수정 } TF \cdot IDF = \frac{1}{n} \sum_{i=1}^{n} \frac{TF}{P_i} \cdot IDF$$

· P_i : 단어가 속한 문헌 i의 총단어 수

3.2 실험결과

3.2.1 과학기술분야

과학기술분야 실험문헌으로부터 상기 두 가지 선정원칙에 따라 선정된 단어 238개 가운데 주제어가 103개, 비주제어는 135개였다. 이 단어들을 가, 나, 다 순으로 정렬하였을 때 상위 10위 및 하위 10위 내에 속하는 단어의 빈도분포는 〈표 4〉와 같고, 2-포아송 모형의 파라미터와 Z값 계산결과는 〈표 5〉와 같다.

〈표 5〉를 보면 과학기술분야 주제어 가운데 '가열', '화상', '효소' 등은 비주제어인 '가격', '경제'보다 높은 Z값을 보이지만 주제어로 분류된 '가스', '회로'의 Z값이 낮게 산출되었다. 예시한 단어만을 대상으로 하면 Z값 범위는 주제어의 경우 약 0.107에서 3.774까지이며, 비주제어의 경우에는 약 0.034에서 0.95이다.

〈표 6〉에서는 238개 단어 전부에 대한 Z값 분포를 주제어와 비주제어의 개수 및 비율로 제시하였다. 〈그림 1〉은 〈표 6〉을 그래프로 표현한 것이다.

〈표 6〉을 보면 1 이하의 낮은 Z값을 갖는 비주제어가 전체 비주제어 135개 가운데 120개인 89%이므로 이 실험집단에서는 Z값에 의하여 비

〈표 4〉 과학기술분야 일부 단어의 빈도분포

단 어	단어 빈도 k(0-25)를 갖는 문헌 수																장서 빈도	문헌 빈도
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…	
가 격	187	6	5	2														22	13
가 공	180	15	3	1	1													28	20
가 스	186	7	2	1	2	2												32	14
가 열	197	1	1										1					15	3
감 각	195	3	1							1								14	5
강 도	188	9	2	1														16	12
결 합	191	5	2		2													17	9
경 비	195	3	1							1								14	5
경 제	185	8	4	1	1		1											29	15
계 산	190	7	2				1											17	10
화 상	195	2	2													1		21	5
확 대	188	10	1		1													16	12
확 보	190	7	2		1													15	10
확 인	181	15	1	1	1			1										31	19
활 동	189	7	3			1												18	11
회 로	188	8	1	2	1													20	12
회 수	187	5	5			1	1	1										33	13
효 소	198			1									1					15	2
효 율	184	9	6		1													25	16
휴 대	192	6			1	1												15	8

※음영처리한 단어는 비주제어임

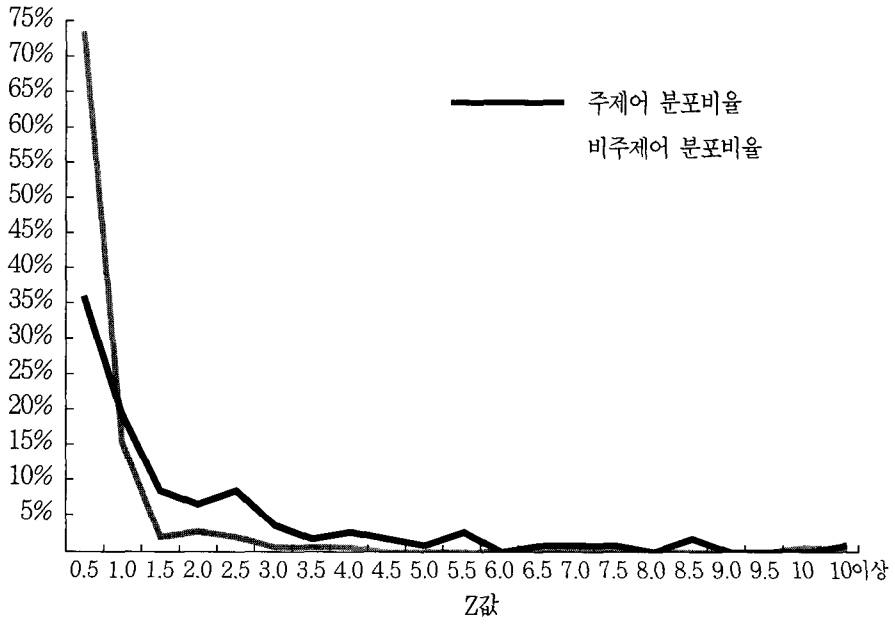
〈표 5〉 과학기술분야 일부 단어의 λ_1, λ_2, π, Z값

단 어	λ_1	λ_2	π	Z
가 격	0.00561	-0.00092	16.98770	0.09535
가 공	0.01285	0.00052	11.30864	0.10666
가 스	0.27890	0.00034	0.57316	0.52716
가 열	4.34838	0.00311	0.01654	2.08304
감 각	0.90848	0.00212	0.07490	0.94982
강 도	0.00114	-0.00001	69.68487	0.03410
결 합	0.03022	0.00006	2.81606	0.52716
경 비	0.90848	0.00212	0.07490	0.52716
경 제	0.18080	0.00292	0.79874	0.52716
계 산	0.09430	0.00101	0.90029	0.52716
화 상	14.31211	0.02393	0.00567	0.52716
확 대	0.00711	0.00017	11.51030	0.52716
확 보	0.00835	0.00013	9.10567	0.52716
확 인	0.32292	0.00651	0.46930	0.52716
활 동	0.03462	0.00053	2.62482	0.52716
회 로	0.01994	0.00010	5.03314	0.52716
회 수	0.89302	0.00777	0.17761	0.52716
효 소	4.50048	0.00151	0.01634	0.52716
효 율	0.01090	0.00006	11.53181	0.52716
휴 대	0.06264	0.00035	1.19845	0.52716

※음영처리한 단어는 비주제어임

〈표 6〉 과학기술분야 주제어와 비주제어의 Z값 분포

단 어	주제어 개수	비주제어 개수	주제어 분포비율(%)	비주제어 분포비율(%)
0.0 - 0.5 미만	37	99	35.92	73.33
0.5 - 1.0 미만	20	21	19.42	15.56
1.0 - 1.5 미만	9	3	8.74	2.22
1.5 - 2.0 미만	7	4	6.80	2.96
2.0 - 2.5 미만	9	3	8.74	2.22
2.5 - 3.0 미만	4	1	3.88	0.74
3.0 - 3.5 미만	2	1	1.94	0.74
3.5 - 4.0 미만	3	1	2.91	0.74
4.0 - 4.5 미만	2	1	1.94	0.74
4.5 - 5.0 미만	1	0	0.97	0.00
5.0 - 5.5 미만	3	0	2.91	0.00
5.5 - 6.0 미만	0	0	0.00	0.00
6.0 - 6.5 미만	1	0	0.97	0.00
6.5 - 7.0 미만	1	0	0.97	0.00
7.0 - 7.5 미만	1	0	0.97	0.00
7.5 - 8.0 미만	0	0	0.00	0.00
8.0 - 8.5 미만	2	0	1.94	0.00
8.5 - 9.0 미만	0	0	0.00	0.00
9.0 - 9.5 미만	0	0	0.00	0.00
9.6 - 10 미만	0	1	0.00	0.74
10이상	1	0	0.97	0.00
합 계	135	103	100.00	100.00



〈그림 1〉 과학기술분야 Z값에 따른 주제어와 비주제어의 비율

주제어가 대부분 식별되고 있는 것으로 보인다. 한편, 1.5 이상의 Z값을 갖는 단어 49개 중 37개(약 76%)가 주제어이므로, 하터가 설정한대로 1.5를 주제어 선정기준으로 삼는다고 가정하면 Z값이 1.5 이상인 경우 주제어가 비주제어보다 많은 현상과 Z값이 0.5보다 작은 단어의 72.8%가 비주제어로 나타난 현상은 하터의 실험과 비슷한 수치이다.

그러나 〈표 6〉과 〈그림 1〉에서 1보다 작은 Z값을 갖는 단어 177개 가운데 주제어가 57개(32.2%)로 비교적 많다는 사실은 하터의 연구와 정영미, 이태영(1982)의 실험에 이어 2-포아송 모형이 주제어 선정기반으로 확립되기에는 미흡함을 보여 준다.

2-포아송 모형 가중치와 역문헌빈도 가중치간의 상관관계를 파악하기 위해 각 주제어에 대한 Z값과 IDF간의 상관계수와 수정 β값과 수정 TF · IDF간의 상관계수를 산출하고 분석하였다.

〈표 7〉은 과학기술분야 주제어 103개 가운데 Z값이 상위 및 하위 각각 10위 이내인 단어 20개에 대해 산출한 Z, 수정 β, IDF, 수정 TF · IDF 값을 보여 준다. 실제로는 모든 주제어에 대하여 값을 산출하였으나, 이 중 일부 주제어만을 이와 같이 예시하였다.

피어슨 상관계수(Pearson's correlation coefficient)를 이용하여 두 가중치 간의 상관관계를 측정한 결과를 〈표 8〉에 수록하였다. 상관관계 정도에 대한 해석은 길포드(Guilford 1956)의 기준에 따른 것으로, 계수가 0.2 미만이면 거의 무시할만하고, 0.2에서 0.4 미만이면 낮은 상관관계, 0.4에서 0.7 미만이면 비교적 높은 상관관계, 0.9 이상이면 매우 높은 상관관계라고 해석할 수

〈표 7〉 과학기술분야 일부 주제어의 Z, IDF, 수정 β, 수정 TF · IDF값

단 어	Z	IDF	수정 β	수정 TF · IDF
용 매	0.0749070	5.6438562	1.0245742	0.1361277
자동화	0.0801133	5.6438562	1.0568182	0.1811505
정 밀	0.0828251	4.7369656	1.0938875	0.1015636
억 제	0.0953490	4.9434165	1.0578529	0.0831154
가 공	0.1066556	4.3219281	1.1167385	0.0959548
분 자	0.1243082	5.8365013	1.1227078	0.1192764
산 화	0.1260965	5.3219281	1.1250284	0.1181566
변 환	0.1374029	5.3219281	1.1383403	0.1168398
회 로	0.1402163	5.0588937	1.1412012	0.1179905
이 온	0.1478625	5.6438562	1.1459290	0.1410432
원 사	4.7995740	6.0588937	5.3054165	0.2165660
진 공	5.2237825	6.6438562	5.7237825	0.4985991
기 상	5.2438659	7.0588937	5.5771992	0.3555811
합 금	5.4475420	6.0588937	5.6975431	0.2951506
재 배	6.2807604	6.0588937	6.5307605	0.2119320
우 주	6.5169479	7.0588937	7.1836146	0.3694698
로 봇	7.1055378	6.3219281	7.4388711	0.2873825
촉 매	8.0212422	5.4739312	8.2712422	0.3535457
식 품	8.3519646	6.6438562	8.6852979	0.4546822
고 장	14.6818315	7.0588937	14.6818315	0.6485825

〈표 8〉 과학기술분야 색인어 가중치간 상관관계

	Z	IDF	수정 β	수정 TF · IDF
Z	1	0.38392	0.99779	0.60788
IDF	0.38392	1	0.39443	0.74938
수정 β	0.99779	0.39443	1	0.62655
수정 TF · IDF	0.60788	0.74938	0.62655	1

있다(정동열 1992).

〈표 8〉에서 Z와 IDF간 상관계수가 0.38392인 것으로 보아 두 가중치간에는 낮은 상관관계가 있다. 반면, 수정 β값과 수정 TF · IDF 가중치간에는 0.62655로 비교적 높은 상관관계를 나타내었다.

3.2.2 여성학분야

HANTEC의 사회과학분야 가운데 본 실험에서 사용한 논문은 여성학 논문 100건이다. 과학기술분야가 초록과 단문기사였으므로 소규모 문헌집단이었다면, 여성학은 대상단어가 61,126개인 점에서 알 수 있는 것처럼 규모가 상대적으로 크다.

이와 같이 전문(full text)인 경우, Z값 분포양상이 초록이나 단문에서와 유사한지, 아니면 상관관계가 문헌형태에 따라 다르게 도출되는지 파악하고자 하였다. 대상단어 359개의 장서빈도 범위는 30 - 96이며, 이를 주제어와 비주제어로 구분한 결과, 주제어로 여겨지는 단어가 186개, 비주제어는 173개였다.

가, 나, 다 순으로 단어를 정렬하였을 때 상단과 하단에 위치하는 단어 10개씩 모두 20개 단어의 빈도분포를 〈표 9〉에 수록하였으며, 2-포아송 모형과 적률법을 적용하여 산출한 파라미터와 Z값을 〈표 10〉에 예시하였다.

〈표 10〉에서 과학기술분야와 구별되는 여성학분야 실험집단 단어의 특징은 π값 가운데 음수가 많다는 것과, 전체단어를 대상으로 할 경우 Z값의 범위가 최저 1.4에서부터 최고 569에 이를 정도로 광범위하고 대체로 큰 값이 산출되었다는 점이다.

〈표 10〉에서 예시한 단어 20개만을 대상으로 하면 Z값의 범위는 주제어인 경우 약 16.208에서 301.897이며, 비주제어인 경우에는 10.740에서 452.300 정도인 것으로 나타났다.

〈표 11〉은 여성학분야의 주제어 186개와 비주제어 173개가 Z값에 따라 분포하는 수와 비율을 보여 주고 있으며, 이를 그래프로 나타내면 〈그림 2〉와 같다.

〈표 11〉 및 〈그림 2〉에서는 비주제어가 작은 Z값을 형성하면서 그래프의 왼쪽에 분포하거나 주제어가 큰 Z값을 형성하면서 오른쪽에 분포하는 성향이 뚜렷하지 않다.

다만, Z값이 100 이상인 단어 가운데 주제어가 비주제어보다 다소 많았다. 그러나 Z값이 약 30에서부터 50까지의 범위에서는 주제어와 비주제어의 수가 거의 같음을 알 수 있다. 이밖에도, 100 이상의 Z값을 갖는 단어 59개 가운데 14개(24%)가 비주제어인 점과 Z값이 50에서부터 70인 범위에서 주제어보다 비주제어가 오히려 더 많이 포함되어 있는 현상 등을 종합해 볼 때, 여성학분야에서는 Z값에 의한 주제어 식별능력이 저조한 것으로 평가된다.

〈표 9〉 여성학분야 일부 단어의 빈도분포

단 어	단어 빈도 k(0-81)를 갖는 문헌 수	장서 빈도	문헌 빈도
	0 1 2 3 4 5 6 7 8-11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26-29 30-34 35-38-54-63-74-		
가계	93 5 1 1	82	7
가공	92 3 1 2 1 1	30	8
가구주	90 3 4 1 1 1	41	10
가내노동 종사자	99 1	54	1
가입자	96 2 1 1	31	4
가정법원	96 1 1 1 1	47	4
가정폭력	92 3 2 2 1	35	8
가족간호 휴직제도	97 1 1 1	41	3
가족성원	93 2 2 1 1 1	37	7
가족의식	96 1 1 1 1	35	4
혈족	95 2 1 1 1	76	5
화학	93 2 1 1 1 1 1	39	7
환경문제	90 4 1 2 1 1 1	57	10
환경운동	94 3 1 1 1	36	6
활동내용	94 3 1 1 1	32	6
후견인	98 1 1	36	2
후보자	90 5 2 1 1 1	80	10
휴업	97 1 1 1	41	3
휴직기	95 4 1 1	38	5
희롱	92 1 3 2 1 1	56	8

※음영처리한 단어는 비주제어임

〈표 10〉 여성학분야 일부 단어의 λ_1, λ_2, π, Z값

단 어	λ_1	λ_2	π	Z
가계	205163.89067	196.57747	-0.00096	452.29958
가공	60.65488	0.52616	-0.00096	7.68729
가구주	124.59049	1.55100	-0.00927	10.95508
가내노동 종사자	41730.06765	-8.41939	0.00021	204.34122
가입자	458.44796	1.02624	-0.00157	21.33960
가정법원	2205.60889	6.24893	-0.00263	46.76465
가정폭력	420.83110	2.28880	-0.00463	20.34734
가족간호 휴직제도	6957.74580	3.59088	-0.00046	83.34855
가족성원	357.49835	2.15957	-0.00504	18.73690
가족의식	263.24771	0.18004	0.00065	16.20827
혈족	91620.11498	159.82927	-0.00174	301.89690
화학	117.05478	0.57654	-0.00160	10.73948
환경문제	4081.39667	24.48808	-0.00590	63.31285
환경운동	705.26195	2.51596	-0.00307	26.41496
활동내용	233.78144	0.93203	-0.00263	15.19869
후견인	4571.89658	-0.08489	0.00010	67.61768
후보자	5478.44048	25.95072	-0.00461	73.49202
휴업	739.12218	1.85109	-0.00195	27.08482
휴직기	3925.71140	3.65009	-0.00083	62.56816
희롱	1381.69336	10.02017	-0.00690	36.76848

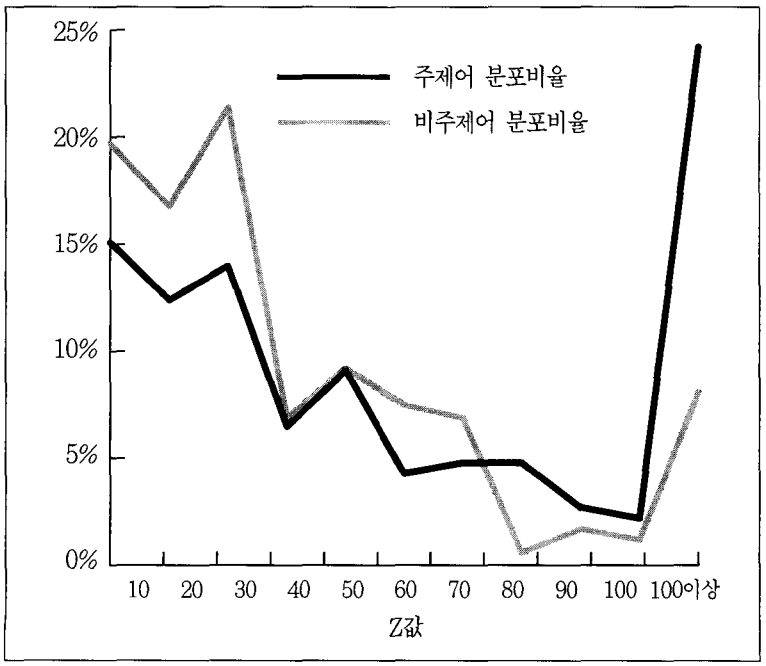
※음영처리한 단어는 비주제어임

〈표 11〉 여성학분야 주제어와 비주제어의 Z값 분포

단 어	주제어 개수	비주제어 개수	주제어 분포비율(%)	비주제어 분포비율(%)
0 - 10미만	28	34	15.1	19.7
10 - 20미만	23	29	12.4	16.8
20 - 30미만	26	37	14.0	21.4
30 - 40미만	12	12	6.5	6.9
40 - 50미만	17	16	9.1	9.2
50 - 60미만	8	13	4.3	7.5
60 - 70미만	9	12	4.8	6.9
70 - 80미만	9	1	4.8	0.6
80 - 90미만	5	3	2.7	1.7
90 - 100미만	4	2	2.2	1.2
100이상	45	14	24.2	8.1
합 계	186	173	100.00	100.00

따라서 주제어의 선정기준으로 Z값의 범위가 과학기술분야에서 '$Z \geq 1.5$'와 같이 제시되었던 것과는 달리, 여성학분야에서는 Z값에 따른 주제어 선정기준을 제시할 수 없는 것으로 보인다.

2-포아송 모형 가중치와 역문헌빈도간의 상관관계를 파악하기 위해 주제어를 대상으로 Z값과 IDF간의 상관계수 및 수정 β와 수정 TF · IDF간의 상관계수를 산출하였다.



〈그림 2〉 여성학분야 Z값에 따른 주제어와 비주제어의 비율

〈표 12〉 여성학분야 일부 주제어의 Z, IDF, 수정 β, 수정 TF · IDF값

단 어	Z	IDF	수정 β	수정 TF · IDF
수유	2.1670108	4.3219281	2.9867137	0.0125135
사회운동	2.3316251	4.4739312	3.1009131	0.0056034
모니터	3.1903851	4.3219281	3.6788673	0.0112001
교육방법	3.4208695	4.8365013	4.1707588	0.0147836
성윤리	3.5208611	4.6438562	4.0545859	0.0088704
낙태	3.6323239	4.3219281	4.2307785	0.0097855
여성사	3.7796475	4.3219281	4.0802284	0.0128297
미디어	3.8400490	4.8365013	4.5636834	0.0188844
아동복지법	3.9406927	4.3219281	4.3487363	0.0119430
남녀학생	5.1213737	4.6438562	5.2889923	0.0176464
감독관	351.9902483	5.3219281	351.9902483	0.0536145
보육교사	352.8468964	6.0588937	352.8468964	0.0995547
결혼상담원	352.9791297	7.6438562	352.9791297	0.2744579
영유아	365.4045825	4.3219281	365.4045825	0.0396905
여성담당관	373.1654069	5.3219281	373.1654069	0.0242280
매음	407.9927293	5.6438562	407.9927293	0.0623088
문화활동	421.7347276	4.3219281	421.7347276	0.0485497
부녀지도사업	453.8269632	7.6438562	453.8269632	0.3210246
여직원	499.2302626	4.3219281	499.2302626	0.0490738
텔레비젼	568.9160181	5.3219281	568.9160181	0.1019209

〈표 12〉는 전체 주제어 가운데 Z값이 상위 및 하위 각각 10위 내인 주제어 20개를 추출하여 단어별로 산출한 Z, 수정 β, IDF, 수정 TF · IDF 값을 보여 준다. 과학기술분야와 마찬가지로 여성학분야 실험집단에서도 주제어 186개 전부에 대한 가중치를 산출하였으나 그 결과를 다 보여주기는 어렵다고 판단되었으므로 이와 같이 일부 주제어의 경우만을 예시하였다.

Z값의 오름차순으로 열거된 주제어의 Z, IDF, 수정 β, 수정 TF · IDF 가중치를 보면, Z와 수정 β간에는 값의 차이가 거의 없지만 IDF와 수정 TF · IDF간에는 값의 차이가 드러나고 있다.

여성학분야 실험집단에서는 〈표 13〉과 같이 Z값과 IDF 가중치간의 상관계수는 0.29534, 수정 β값과 수정 TF · IDF 간의 상관계수는 0.43214로서 두 가지 유형의 가중치간 상관관계가 비교적 낮게 나타났다.

〈표 13〉 여성학분야 색인어 가중치간 상관관계

	Z	IDF	수정 β	수정 TF · IDF
Z	1	0.29534	0.99999	0.43225
IDF	0.29534	1	0.29511	0.79150
수정 β	0.99999	0.29511	1	0.43214
수정 TF · IDF	0.43225	0.79150	0.43214	1

3.2.3 일반사회분야

일반사회 분야의 실험문헌인 한국일보 기사 200건 가운데 실험된 단어는 모두 229개이며, 주제어가 104개, 비주제어는 125개이다. 이를 가, 나, 다 순으로 정렬한 후 상위 및 하위 10위에 속하는 단어의 빈도분포를 예시하면 〈표 14〉와 같고, 2-포아송의 파라미터와 Z 산출 결과는 〈표 15〉와 같다.

전체단어별 주제어 및 비주제어에 따른 Z값의 분포양상은 〈표 16〉과 〈그림 3〉으로 파악할 수 있다.

1.0이하의 낮은 Z값을 갖는 비주제어는 전체 비주제어의 90%가량인 점은 첫 번째 실험집단, 즉 과학기술 초록과 비슷함 분포임을 알 수 있다. 그러나 주제어의 경우도 상당수 낮은 Z값을 가지고 있으므로 주제어가 Z값에 의해 뚜렷하게 식별되지는 않는다는 사실이 파악된다.

〈표 17〉은 피어슨 상관관계 분석을 통해 2-포아송 모형과 역문헌빈도 가중치간의 상관관계를 측정한 값이다. 수정 β와 수정 TF · IDF 간의 상관관계가 0.632이며, Z와 IDF간의 상관관계가 약 0.335로 드러난 것 역시 과학기술 분야의 0.627 및 0.384와 흡사한 결과이다.

〈표 14〉 일반사회분야일부 단어의 빈도분포

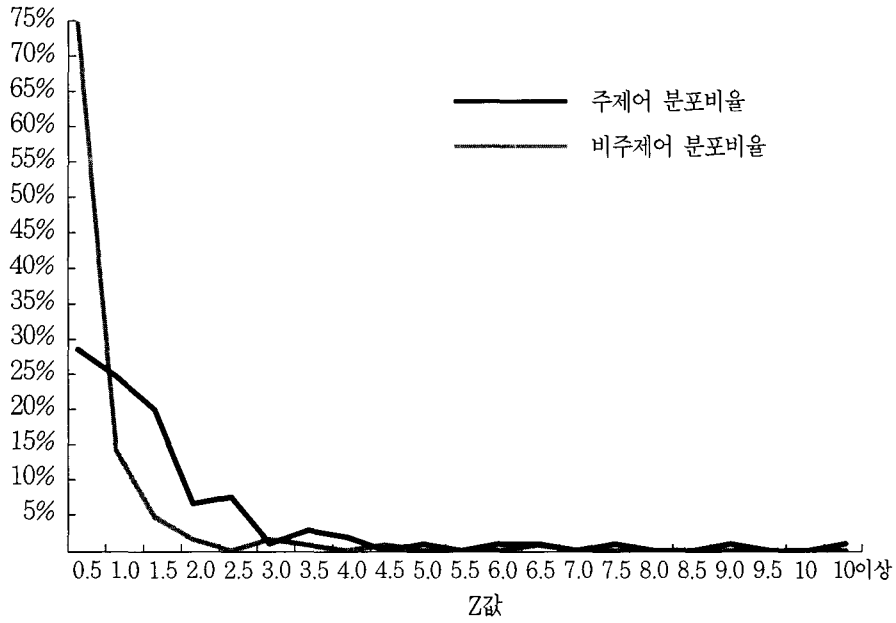
단 어	단어 빈도 k(0-25)를 갖는 문헌 수																							장서 빈도	문헌 빈도
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…	30		
가 정	196	3													1									17	14
강 조	184	11	2	2	1																			25	16
개 선	182	15	2	1																				22	18
개 설	194	3	1	1							1													18	6
개 인	189	8		2	1																			18	11
개 장	195	2		1		2																		15	5
개 최	191	2	3	2	2																			22	9
개 혁	183	10	3	2		1				1														36	17
거 래	183	9	3	3		1			1															37	17
결 혼	197	1	1										1											15	3
환 율	196	3											1											15	4
활 동	185	11	1	1	1			1																27	15
활 용	190	7	2		1																			15	10
회 담	189	6		3					1			1												34	11
회 복	187	10	1	2																				18	13
회 사	186	9	4	1																				20	14
회 의	186	12	1								1													24	14
회 장	189	8	1										1	1										35	11
효 성	198		1													1								17	2
훈 련	192	3	3							1							1							35	8

※음영처리한 단어는 비주제어임

〈표 15〉 일반사회분야 일부 단어의 λ_1, λ_2, π, Z값

단 어	λ_1	λ_2	π	Z
가 정	9.7809434	0.0075299	0.0079266	3.1238401
강 조	0.0199310	0.0002799	6.3467366	0.1382269
개 선	0.0008790	0.0000495	132.5573022	0.0272202
개 설	1.7224186	0.0056267	0.0491459	1.3059908
개 인	0.0187993	0.0001342	4.8146297	0.1356490
개 장	0.1379484	-0.0004627	0.5452067	0.3732862
개 최	0.0563259	-0.0015274	1.9277619	0.2471404
개 혁	1.4183372	0.0211084	0.1137191	1.1645815
거 래	0.9062459	0.0148242	0.1909038	0.9288313
결 혼	4.3483812	0.0031089	0.0165447	2.0830418
환 율	4.2825969	0.0038865	0.0166203	2.0666281
활 동	0.3330036	0.0049156	0.3964924	0.5643958
활 용	0.0083507	0.0001281	9.1056694	0.0892979
회 담	5.8127921	0.0360318	0.0231909	2.3886365
회 복	0.0032880	0.0000361	27.6650778	0.0564032
회 사	0.0016315	-0.0002515	53.2379952	0.0506909
회 의	1.5671193	0.0124436	0.0691825	1.2370047
회 장	21.1109058	0.0750403	0.0047519	4.5702131
효 성	14.2725275	0.0029963	0.0057468	3.7767104
훈 련	38.7571229	0.1243935	0.0013099	6.1956061

※음영처리한 단어는 비주제어임



〈그림 3〉 일반사회분야 Z값에 따른 주제어와 비주제어의 비율

〈표 16〉 일반사회분야 주제어와 비주제어의 Z값 분포

단 어	주제어 개수	비주제어 개수	주제어 분포비율%	비주제어 분포비율%
0.0 - 0.5 미만	30	94	28.85	75.20
0.5 - 1.0 미만	26	18	25.00	14.40
1.0 - 1.5 미만	21	6	20.19	4.80
1.5 - 2.0 미만	7	2	6.73	1.60
2.0 - 2.5 미만	8	0	7.69	0.00
2.5 - 3.0 미만	1	2	0.96	1.60
3.0 - 3.5 미만	3	1	2.88	0.80
3.5 - 4.0 미만	2	0	1.92	0.00
4.0 - 4.5 미만	0	1	0.00	0.80
4.5 - 5.0 미만	1	0	0.96	0.00
5.0 - 5.5 미만	0	0	0.00	0.00
5.5 - 6.0 미만	1	0	0.96	0.00
6.0 - 6.5 미만	1	1	0.96	0.80
6.5 - 7.0 미만	0	0	0.00	0.00
7.0 - 7.5 미만	1	0	0.96	0.00
7.5 - 8.0 미만	0	0	0.00	0.00
8.0 - 8.5 미만	0	0	0.00	0.00
8.5 - 9.0 미만	1	0	0.96	0.00
9.0 - 9.5 미만	0	0	0.00	0.00
9.6 - 10 미만	0	0	0.00	0.00
10 이상	1	0	0.96	0.00
합 계	104	125	100.00	100.00

〈표 17〉 일반사회분야 색인어 가중치간 상관관계

	Z	IDF	수정 β	수정 TF · IDF
Z	1	0.33521	0.99932	0.63601
IDF	0.33521	1	0.33631	0.66250
수정 β	0.99932	0.33631	1	0.63236
수정 TF · IDF	0.63601	0.66250	0.63236	1

3.3 종합 평가

과학기술분야 초록과 여성학분야 전문 및 신문기사를 대상으로 2-포아송 모형과 역문헌빈도를 이용한 주제어 선정기준을 비교한 결과는 다음과 같다.

우선, 주제어와 비주제어를 구분하는 Z값의 식별력이 본 실험에서 부분적으로 밝혀졌다. 특히, 실험집단의 특성에 따라서는 Z값에 의해 비주제어가 보다 잘 식별될 수도 있음을 보여 주었다. Z값에 따른 비주제어 식별능력이 주제어 식별능력보다 더 나은 것으로 파악된 과학기술분야와 신문기사에서 1 이하의 Z값을 갖는 비주제어가 전체 비주제어 가운데 90%에 달한다는 점이 이를

입증한다.

한편 과학기술분야 실험집단에서는 Z값이 1.5 이상인 단어의 76%가 주제어였으나, 여성학분야 실험집단과 신문기사에서는 Z값에 따른 주제어와 비주제어 분포양상간에 별다른 차이가 없었다. 따라서 Z값에 의한 주제어 식별력은 과학기술분야에서만 성과를 나타낸 것으로 분석된다.

또한 Z와 수정 β값간의 차이는 무시해도 좋을 만큼 작다는 점이 확인되었다. 세 문헌집단 전부에서 Z와 수정 β 값간의 상관관계가 0.9 이상이었기 때문이다. 따라서 색인어를 선정할 수 있다고 알려진 β값 대신 주제어를 선정할 수 있다고 알려진 Z값을 2-포아송 모형의 최종 가중치로 이용해도 무방한 것으로 보인다.

2-포아송 모형 가중치와 역문헌빈도 가중치간의 상관관계 분석 결과, 과학기술분야 및 신문기사 실험집단에서의 수정 β와 수정 TF · IDF간의 상관계수(각각 0.627 및 0.632)가 여성학분야 실험집단에서의 수정 β와 수정 TF · IDF간의 상관계수(약 0.432)보다 큰 것으로 드러났으며, Z와 IDF간의 상관성은 세 집단 각각에서 0.383, 0.335, 0.295 정도로 낮은 수준이었다.

4 결 론

한글문헌의 자동색인을 위한 2-포아송 모형의 적용성을 성격이 상이한 세 문헌집단을 대상으로 실험한 결과, 한 집단에서만 2-포아송 모형의 주제어 선정능력이 드러났으며, 두 문헌집단에서는 2-포아송 모형과 역문헌빈도와의 상관관계가 있는 것으로 나타났다.

본 연구에서는 문헌의 주제와 형태에 따라 실험대상을 구분하였으며, 2-포아송 모형의 β 및 역문헌빈도의 TF · IDF 가중치를 표준화시킨 공식을 각각 수정 β 및 수정 TF · IDF 가중치로 재정의하여 산출값의 분포를 비교하였다. 본 실험을 통해 밝혀진 사실은 다음과 같다.

첫째, 주제어 선정을 위한 2-포아송 모형의 적용성은 실험대상 문헌집단의 특성에 따라 다르게 나타났다. 즉, 2-포아송 모형은 전문(full-text)보다는 초록이나 신문기사와 같이 단어 수가 적은 문헌집단인 경우에는 효율적이지만, 여성학분야처럼 전체문헌 수는 적은데 반해 한 문헌에 동일 단어가 반복되어 출현하는 문헌에서는 효율성이 떨어졌다.

둘째, 2-포아송 모형 가중치와 역문헌빈도 가중치간의 상관관계 분석 결과, 과학기술분야 및 신문기사 실험집단에서는 Z값과 IDF 가중치간보다 수정 β값과 수정 TF · IDF 가중치간의 상관관계가 더 큰 것으로 나타난 반면, 여성학분야 실험집단에서는 Z값과 IDF 가중치간 또는 수정 β값과 수정 TF · IDF 가중치간 어디에서도 상관관계가 낮았다.

셋째, Z와 수정 β값간의 상관관계가 IDF와 수정 TF · IDF간의 상관관계보다 크게 나타났다. 즉, 2-포아송 모형의 두 가중치인 Z와 수정 β간에는 값의 차이가 거의 없으므로 수정 β 대신 Z값을 2-포아송 모형의 최종 가중치로 이용할 수 있음이 밝혀졌으나, 역문헌빈도 기반 가중치인 수정 TF · IDF와 IDF간의 대체성은 나타나지 않았다.

본 연구를 통해, 색인어 가중치의 성능이 문헌특성에 따라 다르다는 점이 확인되었으므로 다양한 분야와 문헌형태를 대상으로 2-포아송 모형의 적용성을 검토하는 후속 연구가 필요할 것으로 보인다.

참 고 문 헌

맹성현 외. 1999. 정보검색을 위한 균형테스트 컬렉션 구축. 『정보관리학회지』, 16(2): 135-148.

문성빈. 1999. 2-포아송 모형의 전문검색시스템 응용에 관한 연구. 『정보관리학회지』, 16(3): 49-63.

정동열. 1992. 『문헌정보학 연구방법론』. 서울: 구미무역.

정영미, 이태영. 1982. 자동색인의 통계적 기법과 한국어 문헌의 실험. 『도서관학』, 제 9집: 99-118.

Bookstein, A. 1977. "When the Most "Pertinent" Document Should be Retrieved - An Analysis of the Swets Model." *Information Processing and Management*, 13: 377-383.

Brookes, B. C. 1968. "The Measures of Information Retrieval Effectiveness Proposed by Swets." *Journal of the Documentation*, 24: 41-54.

Burrell, Q. L. and M. Fenton. 1993. "Yes, GIGP Really Does Work - and Is Workable!" *Journal of the American Society for Information Science*, 44(2): 61-69.

Damerau, F. J. 1965. "An Experiment in Automatic Indexing." *American Documentation*, 16(3): 283-289.

Guilford, J. P. 1956. *Fundamental Statics in Psychology and Education*. New York, McGraw-Hill. 재인용: 정동열, 『문헌정보학 연구방법론』. 서울: 구미무역, 1992, 125.

Harter, S. P. 1975. "A Probabilistic Approach to Automatic Keyword Indexing: part Ⅰ and Ⅱ." *Journal of the American Society for Information Science*, 26(4): 197-206, 280-289.

Korfhage, Robert R. 1997. *Information Storage and Retrieval*. New York: John Wiley and Sons.

Lancaster, F. W. 1998. *Indexing and Abstracting Theory and Prac-tice*. Champagne: Graduate School of Information Science, University of Illinois.

Losee, R. M. 1988. "Parameter Estimation for Probabilistic Document Retrieval Models." *Journal of the American Society for Information Science*, 39(1): 8-16.

Luhn, H. P. 1957. "A Statistical Approach to Mechanized Encoding and Searching of Library Infor- mation." *IBM Journal of Research and Development*, 1(4): 309-317.

Margulis, E. L. 1992. "N-Poisson Document Modeling." *Pro-ceedings of the 15th Inter-national ACM SIGIR Conference*: 177-189.

Margulis, E. L. 1993. "Modeling Documents with Multiple Poi-sson Distributions." *Infor-mation Processing and Mana-gement*, 29(2): 215-227.

Maron, M. E. and J. L. Khuns. 1960. "On Relevance, Probabilistic Indexing and Information Retrieval." *Journal of the Association for Computing Machinery*, 8(3): 404-417.

Oswald, V. A., Jr. et al. 1959. *Automatic Indexing and Abstracting of the Contents of Documents*. Los Angeles, Planning Research Corporation. RADC-TR-59-208. Quoted in F.W. Lancaster, Indexing and Abstracting Theory and Practice (Champaign: Graduate School of Information Science, University of Illinois, 1998), 255.

Raghavan, V. V., Hong-pao Shi, and C. T. Yu. 1983. "Evaluation of the 2-Poisson Model As a Basis for Using Term Frequency Data in Searching." *ACM Annual Conference on Research and Development in Information Retrieval. SIGIR Forum*, 17(4) Summer 83: 88-100.

Robertson, S. E. and S. Walker. 1994. "Some Simple Effective Approximation to the 2-Poisson model for Probabilistic weighted retrieval." *Proceeding of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland: 232-241.

Robertson, S. E. and K. Spark Jones. 1976. "Relevance Weighting of Search Terms." *Journal of the American Society for Information Science*, 27(3): 129-146.

Salton, G., C. S. Yang and C. T. Yu. 1975. "A Theory of Term Importance In Automatic Text Analysis." *Journal of the American Society for Information Science*, 26(1): 33-44.

Salton, G. and M. J. McGill. 1983. *Introduction to Modern Information Science*. New York: McGraw Hill.

Spark Jones, K. 1972. "A Statistical Interpretation of Term Specialty and Its Application in Retrieval." *Journal of Documentation*, 28(1): 11-20.

Srinivasan, Padmini. 1990a. "A Comparison of Two-Poisson, Inverse Document Frequency and Discrimination Value Models of Document Presentation." *Information Processing and Management*, 26(2): 269-278.

Swets, J. A. 1963. "Information Retrieval System." *Science*, 141: 245-250.

Thom, James A., Justin Zobel. 1992. "A Model for Word Clustering." *Journal of the American Society for Information Science*, 43(9). 1992: 616-627.

van Rijsbergen, C. J. 1979. *Information Retrieval*. London: Butterworths. [1999. 09.29]. In Hypertext version of the IR textbook by C. J. van Rijsbergen. 〈http://www.dcs.gla.ac.uk/Keith/Chapter.2/Ch.2.html〉

Vickery, B. and A. Vickery. 1992. *Information Science in Theory and Practice*. London: Bowker- Saur Publishers.