2-포아송 모형의 한국어 문헌 적용성

Applicability of Two-Poisson Model to Korean Literature

최대식, 정영미 (연세대학교 문헌정보학과)

Dae-Shik Choi, Young-Mee Chung
Dept. of Library and Information Science, Yonsei University

통계적 확률이론에 근거한 포아송 모형을 색인어 선정 기반으로 활용하고자 하는 2-포아송 함수와 3-포아송 함수 및 다중 포아송 함수에 대한 단계적 발전 과정을 살펴보았다. 아울러, 2-포아송이 한국어 문헌의 색인어 선정에 유용한지 알아보기 위해 한국어 말뭉치 데이터베이스 내 문헌 50개를 실험 대상으로 단어의 장서빈도와 문헌빈도를 이용하여 z값을 산출해 보았다.

1 서 론

자동색인의 효시로 알려진 룬은 문헌 내 단어의 출현빈도가 문헌의 의미를 결정하는 기준이 된다고 보았으며, 단어의 출현빈도에 따른 확률분포를 이용하는 확률색인은 1960년대 이래 지속적인 발전을 보여왔다. 그 기본 가정은 주제어와 비주제어의 분포 양상이 상이하다는 것이다. 확률색인 모형의 대표적인 예로 단어가 문헌 안에 무작위로 분포한다는 포아송 분포를 응용한 다중 포아송 모형이 있다. 2-포아송 모형은 결국 다중 포아송모형 중 하나로 간주될 수 있으며 가장 널리 활용되어 왔다.

2-포아송 모형에 대한 한국어 문헌 실험이 한 차례 수행되었으나 실험 결과는 이 모형이 한국어에는 실효성이 크지 않은 것으로 드러났다. 그러나 한 번의 연구로 2-포아송 모형이 한국어에는 적당하지 않다고 말할 수는 없으므로 최근 완성된 한국어 말뭉치를 대상으로 2-포아송 모형의 유효성을 재점검하기로 한다.

2 선행연구

2.1 2-포아송 모형

마론과 룬은 주제어는 문헌 안에 무작위로 등장하는 것이 아니라 소수 문헌에 집중적으로 출현한다고 주장하였으며, 북스타인과 스완슨은 주제어의 클러스터 형성 실험으로 이를 입증하였다 (Bookstein and Swanson 1977). 그러나 주제어와 달리 비주제어는 문헌 내 무작위로 나타나며, 따라서 포아송 분포에 부합한다는 것이다.

포아송 함수를 응용하여 비주제어뿐만 아니라 주제어의 분포 특성을 설명할 수 있을 것이라는 데 착안하여 2-포아송 모형이 제시되었으며, 하터는 이 모형이 궁극적으로는 문헌의 주제를 표현하는 색인어의 가치 측정에 유용하다고 보았다. 2-포아송 모형은 다음과 같이 단일 포아송 모형 두 개를 결합한 형태이다.

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

· $f(k)$: 주제어가 k 번 출현하는 문헌의 비율
즉 한 문헌 내 단어가 k 번 출현할 확률

· π : 문헌집단 I에 속하는 문헌의 비율

· λ_1 : 문헌집단 I에서 단어의 평균빈도

· λ_2 : 문헌집단 II에서 단어의 평균빈도

색인어로서의 가치는 한 단어가 하위 문헌 집단 두 개를 뚜렷하게 구분할 수 있는 능력에 의해 판별되며, 포아송 분포의 평균과 분산은 동일하므로 z값이 색인어 가치척도로 다음과 같이 도출된다(Harter 1975).

적합/부적합 문헌 분리 능력 : $Z=\frac{\lambda_1-\lambda_2}{\sqrt{\lambda_1+\lambda_2}}$

하터의 실험에서는 z가 1.5 이상인 모든 단어가 주제어로 드러난 반면, 0.5 이하의 작은 z값을 갖는 단어는 75% 가량이 비주제어였다. 따라서 2-포아송 모형을 토대로 한 z값은 주제어 식별 기능이 있다는 것이다.

2.2 3-포아송 모형

3-포아송 모형은 2-포아송 모형의 단순성을 극복하고자 한 시도였으며, 아래 식으로 표현된다(Srinivasan 1990).

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

이 모형에서는 이용자가 원하는 정도에 따라 하위문헌 집단을 탄력적으로 정의할 수 있다. 식의 세 항을 순서대로 정의하면, 첫 항은 매우 적합한 문헌 하위집단, 둘째 항은 보통정도의 적합 문헌 하위집단, 그리고 셋째 항은 부적합문헌 하위집단이라고 구분할 수 있다는 주장이다. 그러나 스리니바산은 3-포아송 모형이 2-포아송보다 우수함을 자신의 실험으로 입증하지 못했다고 밝히며 따라서 2-포아송을 더 옹호하는 입장을 취했다.

2.3 다중 포아송 모형

마굴리스의 다중 포아송 실험 결과는 첫째, 고빈도어 중 약 70%가 다중 포아송 함수 형태를 따르며 둘째, 용어 대부분은 2P, 3P, 4P 분포에 해당하고 셋째, 고빈도어일수록 n의 크기가 증가한다는 것으로 요약된다. 그러나 용어 이외 숙어 및 n-gram 등이 이 패턴을 따르는지 여부와 비영어 문헌에서의 활용 가능성에 대한 연구의 필요성이 제기되었다(Magulis 1993).

2.4 3개 모형 비교분석

2-포아송, 3-포아송, n-포아송 모형의 연구결과 비교는 [표 1]과 같다.

[표 1] 2-포아송, 3-포아송, n-포아송 모형 비교

구분 \ 모형	2-P	3-P	n-P
대상자료	초록 650건	초록 59,919건, word stems 196개	신문기사 7,784건, 영화 비평 1034건
문헌크기	223 words	58 words	최소 400words 이상
파라미터	π, λ_1, λ_2	π, π_1, λ_1, λ_2, λ_3	nP에서 n에 따라 가변적, n이 커질수록 파라미터 증가
추정방법	적률법	적률법	최우 추정법
실험결과	색인어의 38%가 2-P 분포	196개 terms 중 43%가 2-P 분포	terms의 70%가 nP 분포, 대부분 2P, 3P, 혹은 4P

3 2-포아송의 한국어 문헌 적용성 실험

3.1 선행 연구

정영미와 이태영(1982)은 2-포아송 모형을 한글 문헌에 적용하였는데, 실험 결과 '감귤' '관옥'과 같은 농학 분야 주제어는 큰 z값이 산출된 반면, '비교' '변화' 등 비주제어는 1 이하의 값이 나왔다.

그럼에도, 농학 및 원예 분야 한국어 대상 z값 산출결과가 성공적이지 못했다고 결론 내릴 수밖에 없는 이유는 z가 1.5이상일 때 및 0.5에서 1.5 사이일 때 주제어보다 오히려 비주제어가 더 많이 출현했기 때문이다. 다시 말해 한국어 문헌을 대상으로 할 경우, 하터의 영어 문헌 실험과는 관이하게, 2-포아송 기반 z값의 효용성이 뛰어나다고 단정할 수는 없음을 이 실험을 통해 알 수 있다.

3.2 한국어 말뭉치 대상 실험

3.2.1 배경 및 방법

앞서 지적한 바대로 3-포아송은 유용성이 밝혀지지 않았으며 n-포아송은 파라미터 추정이 복잡하므로 자동색인에서 실용화하기에는 문제가 있다고 판단되었다. 따라서 다중포아송 모형의 대표적 함수인 2-포아송을 한국어 문헌에 적용하는 실험이 필요하다고 보고, 선행 연구와는 다른 분야의 한국어 문헌을 대상으로 λ_1 , λ_2 , π , 그리고 z값을 산출하였다.

이번 실험에서는 한국어 문헌 가운데 비교적 최근에 작성된 문헌 전문(full text) 중에서 비교적 포괄적인 주제를 다루고 있는 문헌집단을 선정하였다. 채택된 대상은 1998년 KAIST가 개발한 '대한민국 국어정보 베이스II(CD-ROM)'로, 이 중 분류번호 300(사회과학), 400(순수과학), 500(기술과학), 600(예술) 및 700(어학)에 해당하며 크기가 61 - 76 킬로바이트에 해당하는 문헌 50개이다. 주제는 '사진'에서부터 '문화' 그리고 '과학'에 이르기까지 다양한 분포를 이루고 있으며 고등학교 국민윤리 교과서도 포함되었다.

실험대상 문헌의 특징은 지나치게 학술적이지 않으며, 문체는 수필 형식이 주류를 이루고 있다는 점이다. 단어 총 수는 42,554개이며 이들 가운데 단어 26개를 대상으로 2-포아송 모형을 적용하여 z값을 계산하였다.

형태소 분석기로는 한성대에서 개발된 '햄(HAM: Hangul Analysis Module)'을 이용하였다. 이로써 50개 문헌 내 모든 단어 각각에 대한 문헌빈도, 장서빈도, 그리고 문헌별 등장 회수를 파악했다.

대상 단어 26개는 중간빈도어로서 장서빈도가 11에서 56에 해당하는 것으로 제한하였다.

파라미터 산출 공식으로는 하터의 실험 및 정영미/이태영의 선행 연구에서 적용했던 것과 동일한 적률법(Method of Moments)을, 계산에는 Excel을 이용했다.

3.2.2 실험결과

대상단어 가운데 '도자기' '사회복지' '변증법' '도시화' '기독교' '매체' '사회주의' '우주선' 등은 실험대상 문헌의 성격상 주제어로 여겨질 만한 단어들이다. 반면, '타당성' '당연' '풍부' '이의' 등은 비주제어에 속할 것이라고 가정하였다.

실험 결과, 이들 단어의 빈도에 따라 산출된 z는 대체로 주제어와 비주제어를 구분할 수 있는 정도의 값을 보여주고 있다([표 2] 참조).

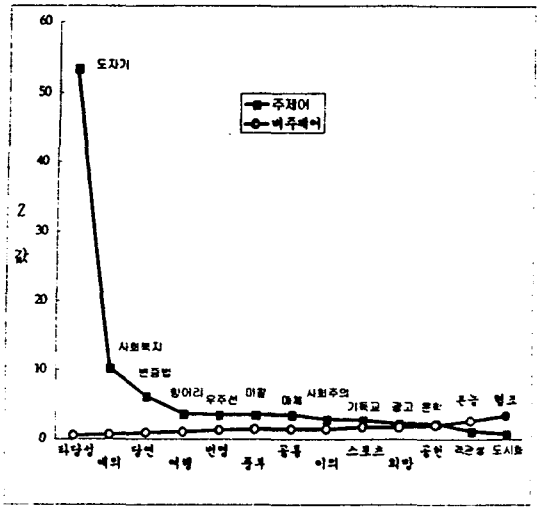
[표 2] 적률법에 의한 파라미터

단어	λ_1	λ_2	π	1-π	z
타당성	0.35362	0.02018	0.83918	0.16082	0.54539
예의	0.47970	0.02082	0.60839	0.39161	0.64862
도시화	0.56485	0.00371	0.38545	0.61455	0.74418
당연	1.35748	0.24062	0.64411	0.35589	0.88349
여행	1.59187	0.12821	0.38986	0.61014	1.11733
객관성	1.35195	0.01125	0.14079	0.85921	1.14829
변동	1.97553	0.11535	0.26054	0.73946	1.28644
풍부	2.36436	0.13236	0.34392	0.65608	1.41256
공통	3.32114	0.52281	0.21341	0.78659	1.42729
이의	2.76987	0.23396	0.13645	0.86355	1.46317
스포츠	3.09980	0.11621	0.14204	0.85796	1.66372
희망	3.19804	0.02073	0.30191	0.69809	1.77099
공헌	4.78490	0.28522	0.07440	0.92560	1.99835
문학	6.38361	0.41319	0.03129	0.96871	2.29009
광고	5.44089	0.05125	0.07213	0.92787	2.29980
본능	8.38168	0.53009	0.01655	0.98345	2.63012
기독교	8.45123	0.25123	0.05717	0.94283	2.77967
사회주의	9.22658	0.17129	0.03851	0.96149	2.95385
협조	12.76146	0.31004	0.00562	0.99438	3.44394
매체	14.94004	0.82076	0.00561	0.99439	3.55650
마찰	13.89674	0.25439	0.00774	0.99226	3.62655
우주선	13.53914	0.11462	0.01977	0.98023	3.63306
항아리	13.80668	0.11110	0.01233	0.98767	3.67109
변증법	38.48751	0.23478	0.00589	0.99411	6.14727
사회복지	105.86395	0.10236	0.00357	0.99643	10.27411
도자기	2902.77015	15.55404	-0.00514	1.00514	53.44567

※ 주제어로 여겨지는 단어는 음영 처리

[표 2]를 보면, 하위에 위치한 '도자기' '사회복지' '변증법' '항아리' '우주선' '매체' '사회주의' '기독교' 등은 예상대로 높은 z값이 나왔다. 이는 기대치의 85%(13단어 중 11단어)에 해당한다. 그러나 '도시화'는 기대와 달리 상대적으로 낮은 z값이 산출되었다.

한편, '본능'과 '협조'가 비주제어로 여겨졌음에도 불구하고 '도시화'보다 오히려 높은 z값을 보인 것은 의외이다. 비주제어 가운데 '타당성' '당연' '풍부' '공통' '이의' '희망' 은 z값의 오름차 순위 가운데 중간 이상인 상단에 위치함으로써 실험 전 예상과 일치하였다.



[그림 1] 주제어와 비주제어의 z값 분포

실험 결과 z값이 하터와 정영미/이태영의 실험에서보다 큰 수치를 나타내고 있다([그림 1] 참조). $0.5 \leq z < 1.5$ 범위 안에서 많은 주제어가 출현했던 선행 연구들과는 차이가 있다는 것이다. [표 2]와 [그림 1]을 통해 26개 단어의 z값 범위는 $0.54 \leq z \leq 53.45$ 임을 알 수 있다.

'도자기'가 이례적으로 큰 z값이 산출된 이유는 독특한 빈도 분포 때문인 것으로 여겨진다. 이것은, 문헌 50개 가운데 이 단어가 24회 출현한 문헌이 한 개, 1회 출현한 문헌이 한 개로 다른 단어에 비해 특정 문헌 하나에 지나치게 집중적으로 출현한 분포 특성을 말한다.

4 결 론

2-포아송으로 대표되는 다중 포아송 모형을 자동색인에 도입하고자 하는 시도는 적지 않았으나 색인어 선정 근거로 이 모형이 적극 활용되지는 못한 이유는 두 가지 문제점 때문인 것으로 여겨진다.

첫째, 단어가 문헌 내 독립적으로 존재한다는 가정이다. 이 모형에서는 전체 문헌집단을 적합·부적합 문헌 클래스로 분리하여 그 안에서 용어의 독립성이 유지된다는 전제가 필수적인데, 이는 제고될 필요가 있다. 비주제어와 달리 주제어는 용어간 상호의존성이 분명히 존재하기 때문이다.

둘째, 모든 문헌은 길이(크기)가 다름에도 불구하고, 문헌길이에 대한 정규화 없이 단어빈도만을 고려함으로써 모형의 신뢰성을 감소시킬 수 있다는 점 역시 간과되어왔다.

이번 실험에서는 두 번째, 즉 문헌 길이 정규화 문제를 해소하기 위해 실험대상 문헌 데이터베이스 가운데 61 - 76 킬로바이트에 해당하는 유사 크기의 문헌 50개를 대상으로 하였다. 그러나 첫 번째 문제점에 대한 고려는 향후 연구가 뒤따라야 할 것이다.

본 연구에서는 그간 한국어 문헌을 대상으로 한 2-포아송 모형 실험이 거의 없었음에 주목하여, 주제가 비교적 편향되지 않고 지나치게 학술적이지 않은 한국어 문헌을 대상으로 모형의 유용성을 검증하는 실험을 수행하였다. 실험결과는 2-포아송 모형에 의한 Z값은 대체로 주제어와 비주제어를 분리시키는 능력이 우수한 것으로 나타났는데, 이는 1982년의 연구와 다른 양상이다.

그러나 실험 문헌이 50개이며, 중간빈도어만을 대상으로 선정한 것은 실험결과의 유효성을 입증하는데 있어 다소 설득력이 부족하므로 보완이 필요하다. 또한 파라미터 추정 방법 중 적률법 대신 최우추정법을 한국어 문헌 실험에 적용해보는 연구 역시 요구된다.

참고문헌

정영미, 이태영. 1982. "자동색인의 통계적 기법과 한국어 문헌의 실험". **도서관학, 제 9집**, 99-118.

Bookstein, A. and D. R. Swanson. 1974. "Probabilistic Model for Automatic Indexing". *Journal of the American Society for Information Science*, 25(5): 312-318.

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.

Margulis, E. L. 1993. "Modeling Documents with Multiple Poisson Distributions". *Information Processing and Management*, 29(2): 215-227.

Srinivasan, P. 1990. "On Generalizing the Two-Poisson Model". *Journal of the American Society for Information Science*, 41(1): 61-66.