2-포아송 모형의 전문검색시스템 응용에 관한 연구

Application of the 2-Poisson Model to Full-Text Information Retrieval System

문성빈(Sung-Been Moon)*

목 차

초 록

본 연구는 질문용어의 분포가 초록/표제 및 전문으로 표현된 문헌 내에서 2-포아송 분포를 따르고 있는지를 조사하였으며 질문용어의 2-포아송 분포여부가 확률이론에 기반을 둔 이진독립모형과 2-포아송 독립모형에서 초록/표제 및 전문의 검색효율성에 미치는 영향을 비교 · 분석하였다.

ABSTRACT

The purpose of this study is to investigate whether the terms in queries are distributed according to the 2-Poisson model in the documents represented by abstract/title or full-text. In this study, retrieval experiments using Binary independence and 2-Poisson independence model, which are based on the probabilistic theory, were conducted to see if the 2-Poisson distribution of the query terms has an influence on the retrieval effectiveness, particularly of full-text information retrieval system.

키워드: 2-포아송 모형, 전문검색시스템, 검색효율성, 이진독립모형, 확률검색모형

* 연세대학교 문헌정보학과 부교수
■ 논문 접수일 : 1999년 9월 1일

1 서 론

지난 몇 년 동안 정보기술의 발전은 초고속정보통신망의 구축을 촉진시켜 왔고 이미 몇몇 선진국가에서는 이러한 통신망을 이용한 학술정보의 교환이 활발하게 이루어지고 있다. 이러한 초고속통신망을 이용하여 탐색한 결과는 수초 내에 정보요구자의 퍼스널컴퓨터의 하드디스크에서 전송 받을 수 있을 뿐만 아니라, 프린터를 이용하여 출력할 수도 있다. 정보기술, 특히 네트워크기술의 끊임없는 발전은 보다 저렴하게 데이터베이스의 구축을 가능하게 하고 이들을 이용한 정보검색활동은 더욱 활발해지고 있다. 특히 정보기술을 이용한 전문데이터베이스의 구축과 상용시스템에 있어 전문데이터베이스가 차지하는 비율은 주목할 만 하다.

전문데이터베이스는 장서에 포함되어 있는 모든 문헌들의 단어, 그리고 모든 문자들을 저장시킴으로써 컴퓨터에 의해 이들을 탐색할 수 있도록 만들어진 데이터베이스로 정의되고 있다(Blair and Maron 1985, 289). 이러한 데이터베이스를 이용한 전문검색시스템은 자연언어탐색을 가능하게 하며, 표제나 초록, 그리고 통제언어집을 이용하는 기존 검색시스템의 대체물로서 인기를 모으고 있다. 전문검색시스템과 초록, 표제, 통제언어집 등을 이용하는 검색시스템의 효율성을 비교하는 논쟁은 끊임없이 이루어지고 있는데 이는 경제적 측면에서 보았을 때 전문검색시스템의 효율성에 대한 진지한 연구가 계속되어야 함을 보여주고 있는 것이다.

일반적으로 정보검색시스템의 검색효율성은 정확률(precision)과 재현율(recall)에 의해 측정된다. 전문검색시스템은 대체로 낮은 정확률과 높은 재현율을 보여주는데, 낮은 정확률은 전문검색시스템의 문제점으로 간주되고 있다.

본 연구는 확률검색모형에 기초를 둔 이진독립모형(Binary Independence Model: 이하 BI)과 2-포아송 독립모형(2-Poisson Independence Model: 이하 TPI)을 이용하여 전문검색시스템에서 이들의 검색효율성을 비교·분석하고 2-포아송 모형의 전문검색시스템에서의 적용가능성에 대해 살펴보았다.

2 배 경

지난 십 여년 동안 많은 정보학자들은 전문검색시스템의 평가에 관심을 보여 왔는데, 특히 정보기술의 발전은 전문을 이용한 문헌의 표현과 자연언어탐색을 가능하게 하는 전문검색시스템의 잠재력을 검토하게 하였다(Mckinin et al. 1991; 노정순 1988a, 1988b; Tenopir 1985; Blair and Maron 1985). 전문검색시스템은 색인의 망라성(indexing exhaustivity)을 최대화시키는 시스템이라고 할 수 있다. 색인의 망라성은 문헌에 할당된 색인어의 수에 의해서 측정될 수 있는데 이는 재현율을 향상시키는 방법으로 간주되고 있다. 데이터베이스에 있는 한 문헌에 부여되는 색인어의 수를 증가시키는 것은 그 문헌이 질문용어를 포함할 수 있는 기회를 높여주게 된다. 그러나 전문을 이용한 문헌의 표현은 검색에 유용한 색인어 뿐만 아니라 검색과정 중에 잡음(noise)의 역할을 하게 될 주제와 관련 없는 불필요한 용어를 추가하는 결과를 가져오게 된다 (Blair and Maron, 1990). 정보학자들은 이러한 정보잡음을 통제함으로써 전반적인 전문검색시스템의 검색효율성을 증진시키기 위해 노력하고 있다.

경제적 혹은 실용적인 측면에서 볼 때 전문을

이용하지 않고 초록이나 표제를 이용하여 문헌을 표현하는 것은 물론 저장 및 검색을 위한 비용을 절감하게 한다. 그러나, 이러한 요소들은 빠르고 저렴한 컴퓨터로의 접근을 가능하게 하는 미래의 초고속 정보통신시대에는 심각하게 고려할 필요가 없게 될 것이므로, 오히려 앞으로의 정보검색 시스템은 전문에 의해 동반될 수 있는 정보잡음을 제거하거나 감소시킬 수 있는 검색기법의 개발에 초점을 맞추어야 할 것이다.

Mckinin et al(1990)과 Blair and Maron (1985)은 대규모의 데이터베이스를 이용하는 전문검색시스템에서는 너무 많은 문헌이 검색되는 "Output Overload"의 문제점을 발견하고, 이를 피하기 위해 매우 제한적인 불린 탐색문을 이용하였는데, 이 문제점은 불린 모형 대신 확률모형을 이용하여 제거될 수 있다고 한다(Maron, 1988).

확률이론에 기초한 용어분산모형으로는 BI와 TPI을 들 수 있는데, 이 두 모형의 검색효율성에 대한 연구는 몇몇 학자들에 의해 이미 보고된 바 있다(Losee 1988; losee 1987; Losee, Bookstein, and Yu 1986; Raghavan, Shi, and Yu 1983). 650개의 초록으로 구성된 데이터베이스를 이용한 실험결과에 의하면, 초록 및 표제로 구성된 데이터베이스의 문헌 내에서 대부분의 용어들은 2-포아송의 분포을 따르고 있지 않는다고 보고하고 있다(Harter 1975a; 1975b). 그러므로 BI모형의 검색효율성은 항상 TPI모형을 능가하고 있다고 한다. 그러나, 2-포아송 모형을 만족시키는 실험 데이터베이스를 구축하여 BI와 TPI의 성능을 비교한 결과 TPI모형이 BI모형을 능가하고 있음을 증명하고 있다(Losee, Bookstein, and Yu 1986). Srinivasan(1990)은 용어의 분산모형을 2-포아송보다 일반화된 3-포아송 모형에 의해 설명하고자 했다. 즉, 데이터베이스 내의 문헌들을 3개의 그룹으로 나누어 이들 내에서의 용어의 분포를 3-포아송 모형에 의해 설명하고자 했으나 문헌검색에서 포아송 모형의 사용은 조심스럽게 이루어져야 한다고 보고하였다.

그러나, 아직까지 전문에서 용어의 분산 모형을 설명한 연구는 없었다. TPI모형은 BI모형에서는 고려하지 않는 문헌 내에서의 용어빈도수를 이용하는 이점이 있기 때문에 특히 전문검색시스템에서 유용할 것으로 보인다. 그러나, 전문검색시스템에서의 BI모형은 적절하지 못한 것으로 여겨진다. 왜냐하면 문헌 내의 중요한 개념을 표현하기 위해 자주 사용되는 용어와 그렇지 못한 용어들을 구분해 주지 못하기 때문이다(Fuhr 1992). 그러므로, 전문데이터베이스를 이용한 두 용어분산모형의 검색효율성의 비교 및 전문에서의 용어의 분산 모형에 대한 연구 및 논의가 필요하다.

3 확률검색모형

적합성(relevance)과 문헌의 특성사이에는 확률관계가 존재하고 있으며 전문검색시스템의 근본적인 문제점을 해결하기 위해서는 확률원칙이 전문검색시스템의 설계에 적용되어야만 한다(Maron 1988). 특히 전문에는 적합성을 예측할 수 있는 많은 단어들이 포함되어 있는데 이러한 단어를 이용하여 문헌이 적합할 확률(probability of relevance)을 계산하고 이에 의해 데이터베이스 내에 있는 모든 문헌들의 검색순위을 매길 수 있도록 한다(Robertson 1977).

정보검색시스템은 어떤 문헌의 검색여부를 결

〈표 1〉 적합성 여부에 따른 문헌의 검색과 관련된 비용

		적합성여부(Relevance Value)	
		적 합	부적합
행위(Action)	Retrieve	Cost Ret, Rel	Cost Ret, Notrel
	Not Retrieve	Cost Notret, Rel	Cost Notret, Notrel

정하기 위해 수학 모형을 이용하는데 이를 "Decision Theoretic Model" 이라고 부른다. 이 모형은 문헌의 검색 여부를 결정하기 위해서 〈표 1〉과 같이 비용이 관련되어 있다고 보고 검색을 위한 원칙을 제시하고 있다: 만약에 어떤 문헌을 검색하는데 드는 비용이 그 문헌을 검색하지 않는데 드는 비용보다 적다면 그 문헌을 검색하고 그렇지 않으면 검색하지 않는 것이다.

〈표 1〉을 참조해서 다음과 같은 수학적인 공식을 만들 수 있다.

확률(적합|d)*Cost Ret, Rel
+ 확률(부적합|d)*Cost Ret, Notrel
< 확률(부적합|d)*Cost Notret, Notrel
+ 확률(적합|d)*Cost Notret, Rel

이것은 아래와 같이 변환될 수 있다(Losee 1987).

$$\frac{\text{확률(적합}|d)}{\text{확률(부적합}|d)} > \text{Cost}$$

확률(적합|d)/확률(부적합|d)는 계산이 간단하지 않으므로 베이스의 원칙(Bayes Rule)에 근거하여 아래와 같이 변환시킬 수 있다.

$$\text{확률(적합}|d) = \frac{\text{확률(적합}|d) * \text{확률(적합)}}{\text{확률}(d)}$$

$$\text{확률(부적합}|d) = \frac{\text{확률(부적합}|d) * \text{확률(부적합)}}{\text{확률}(d)}$$

그러므로

$$\frac{\text{확률(적합}|d)}{\text{확률(부적합}|d)} = \frac{\text{확률}(d|\text{적합})}{\text{확률}(d|\text{부적합})} * \frac{\text{확률(적합)}}{\text{확률(부적합)}}$$

여기서 확률(적합)/확률(부적합)은 상수이므로 전체 문헌의 순위에 영향을 미치지 않는다.

그러므로, 확률(d|적합)/확률(d|부적합)만이 "Retrieval Status Value(RSV)"로 남아 확률검색모형에서는 이 값으로 문헌의 검색순위를 결정한다.

3.1 이진독립모형 (Binary Independence Model: BI)

BI모형은 문헌내에서 질문용어가 출현하면 '1', 그리고 출현하지 않으면 '0'이라고 문헌벡터상에 표시한다. 용어출현 및 분산의 통계적 독립성을 가정하고 위에서 언급한 검색을 위한 문헌의 RSV는 아래 공식에 의해 결정된다 (Bookstein, 1983; Losee, 1988).

$$\frac{\text{확률(적합}|d)}{\text{확률(부적합}|d)} = \frac{\prod_{i=1}^{n} p_i^{d_i} (1 - p_i)^{(1 - d_i)}}{\prod_{i=1}^{n} q_i^{d_i} (1 - q_i)^{(1 - d_i)}}$$

여기서 p_i, q_i는 i 번째 질문용어가 적합문헌이나 부적합문헌에서 출현할 확률을 각각 나타내고 d_i는 '0' 혹은 '1'로 표현되며 i번째 질문용어

가 적합문헌 및 부적합문헌에서의 출현여부를 보여주고 있다. 위의 표현은 문헌가중치를 계산하기 위해 양변에 log를 취하면 아래와 같이 변환되어진다.

$$\log \frac{\text{확률}(d|\text{적합})}{\text{확률}(d|\text{부적합})} = \sum_{i=1}^{n} d_i * log \frac{p_i / (1 - p_i)}{q_i / (1 - q_i)}$$

3.2 2-포아송 독립모형 (Two-Poisson Independence Model)

2-포아송 독립모형은 Bookstein과 Swanson (1974)에 의해 제안되어 Harter(1975a; 1975b)에 의해 연구되었다. 이 모형은 장서 내에서 주제를 표현하지 못하는 단어인 비주제용어(non-specialty words)는 전체문헌 내에서 그 분포는 단일 포아송 함수에 의해 표현되어질 수 있다고 가정하고 있다.

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

그러나, 주제를 표현할 수 있는 단어인 주제용어(specialty words)는 적합문헌들과 부적합문헌들 사이에 서로 다른 분포를 가지고 있다는 가정에 기초를 두고 있으며 그 분포는 아래와 같이 2-포아송 함수에 의해 표현된다. 즉, 주제용어는 적합문헌 및 부적합문헌 내에서 각각 서로 다른 평균 출현빈도를 갖고 포아송 분포를 따른다는 것이다.

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

여기서 π는 장서 내에 적합문헌이 차지하는 비율을 나타내며, λ_1, λ_2는 각각 적합문헌과 부적합문헌 내에서 용어의 평균출현빈도를 나타낸다. 적합문헌 내에서 용어의 평균출현빈도는 부적합문헌 내에서 용어의 평균출현빈도보다 크다고 가정하고 있다 (즉, $\lambda_1 > \lambda_2$). BI모형에서는 문헌의 가중치를 결정하기 위하여 용어의 출현여부만을 고려하는 반면에, 2-포아송 독립모형에서는 용어 출현빈도를 고려하여 문헌의 가중치를 계산한다. 아래와 같이 2-포아송 용어분산을 이용하여 문헌의 검색순위를 위한 공식은 아래와 같이 결정된다 (Bookstein, 1983; Raghavan, Shi, and Yu 1983; Losee, 1987; Losee, 1988).

$$\frac{\text{확률}(d|\text{적합})}{\text{확률}(d|\text{부적합})} = \frac{\prod_{i=1}^{n} \frac{e^{-\lambda_1} \lambda_1^k}{k!}}{\prod_{i=1}^{n} \frac{e^{-\lambda_2} \lambda_2^k}{k!}}$$

양변에 log를 취하고 나면, 위의 표현은 아래와 같이 변환된다.

$$\log \frac{\text{확률}(d|\text{적합})}{\text{확률}(d|\text{부적합})} = \sum_{i=1}^{n} d_i * log\ (\lambda_1 / \lambda_2) + \sum_{i=1}^{n} (\lambda_2 - \lambda_1)$$

결국, 우변은 상수가 되고, 아래의 공식으로 변환된다.

$$\log \frac{\text{확률}(d|\text{적합})}{\text{확률}(d|\text{부적합})} = \sum_{i=1}^{n} d_i * log\ (\lambda_1 / \lambda_2)$$

4 실험환경

정보검색시스템의 검색효율성에 대한 연구는

각기 다른 실험환경 때문에 일관성 없는 실험결과를 보여주고 있는 경우를 자주 볼 수 있다. 정보검색실험결과에 영향을 미치는 요소로써 검색시스템, 데이터베이스, 질문서, 적합성판단 등을 들 수 있다(Robertson, 1981).

본 연구의 모든 실험은 1960년대에 미국의 코넬대학에서 개발되어 지난 30년 동안 정보검색실험도구로 많이 이용되어 온 SMART 시스템에 의해 수행되었다. 문헌과 질문서들은 SMART시스템에서 쓰이고 있는 단일용어 자동색인방법에 의해 어미를 제거한 형태로 표현되었다. SMART시스템은 평균정확률(Average Precision)과 보간법을 이용하여 21개 지점의 표준재현율(standard recall : 0.00, 0.05, 0.1, 0.15, ... 0.95, 1.0)에서의 정확률을 계산해서 보여 준다.

본 연구에서 사용된 CF392 전문데이터베이스는 1974-1979 사이에 출간된 392개의 문헌으로 구성되어 있는데 이들은 미국의 국립의학도서관에 의해 MEDLINE 화일에 "Cystic Fibrosis"라는 주제용어로 색인되어 있다. 질문서는 노스캐롤라이나 주립대학의 의과대학 소아과교수인 Dr. Robert Wood에 의해 만들어졌으며, 이는 소아의 호흡기 질환인 "Cystic Fibrosis"와 관련된 문항으로 이루어져 있다. 질문서와 문헌간의 적합성판정은 Dr. Robert Wood와 몇몇 동료교수, 같은 주제 분야의 포스트박사과정에 있는 연구원, 그리고 의학 분야의 온라인 탐색경험이 풍부한 서지 전문가 등 세그룹에 의해 이루어졌다. 주제전문가들과 의학분야의 서지 전문가는 질문서에 의해 검색된 문헌의 내용을 평가하기 위해 문헌의 전문을 살펴보았다. 적합성판단은 전문을 대상으로 해서 이루어져야한다는 점을 강조하면서 텍스트뿐만 아니라 그림 및 표에 나타나는 정보만을 보고 적합성여부를 판단한 경우도 있다(Shaw, Wood, and Tibbo 1991). 질문서-문헌들간의 적합성판단을 위해서 위의 세 그룹의 주제전문가들은 "매우 적합한", "조금 적합한", "전혀 적합하지 않은"으로 구분하였으며 적합문헌은 세 그룹 중 최소한 두 그룹이 "매우 적합한"이라고 판정하였을 경우에 해당되는 것으로 정하였다. 질문서의 수는 83개이고 질문서당 최소 적합문헌의 수는 한 개인데 이는 23개의 질문서에 해당되고 최대 적합문헌의 수는 33개로 질문서 44에 해당된다. 각 질문서당 평균 적합문헌의 수는 4.66이고 평균용어의 수는 6.84이다(문성빈 1993).

5 실험결과 및 토론

CF392 전문데이터베이스 및 83개의 질문서를 이용하여 검색실험을 한 결과 〈표 2〉와 같이 83개의 질문서에 대한 평균정확률 및 21개 지점의 재현율에서의 정확률 값을 얻을 수 있었다. 특히, 전문검색에서의 검색효율성은 BI보다는 TPI와 함께 향상되었음을 보여주고 있다. 초록/표제는 BI와 TPI의 평균정확률이 각각 0.3996, 0.2811인 반면에, 전문에서는 BI의 평균정확률이 0.2179이고 TPI는 0.2810을 나타내고 있으며 이에 대한 t검증은 α=0.05수준에서 유의미한 것으로 나타났다. 이는 전문검색에서 문헌 내의 용어빈도수를 고려하는 것이 적합문헌과 부적합문헌을 구분하는데 유용하였음을 보여주고 있다. 즉, BI모형이 초록/표제를 위해서는 적절한 모형이지만 전문검색을 위해서는 부적합하다는 것을 암시하고 있다. 왜냐하면, 전문검색에서 BI모형은 문헌 내에서 중요한 개념을 표현하기 위해 자주 출현하는 용어와 한번만 출현하는 용어들을 구분해 주지 못하기

〈표 2〉 용어분산모형에 따른 83개 질문서의 검색효율성

	초록/표제		전문	
재현율	BI	TPI	BI	TPI
0.00	0.5846	0.4317	0.2791	0.3759
0.05	0.5775	0.4301	0.2774	0.3759
0.10	0.5563	0.4261	0.2753	0.3721
0.15	0.5333	0.4058	0.2730	0.3589
0.20	0.5236	0.3948	0.2628	0.3573
0.25	0.5013	0.3672	0.2512	0.3439
0.30	0.4575	0.3426	0.2381	0.3155
0.35	0.4373	0.3098	0.2317	0.3054
0.40	0.4331	0.2978	0.2264	0.3014
0.45	0.4167	0.2890	0.2227	0.2928
0.50	0.4139	0.2856	0.2218	0.2908
0.55	0.3322	0.2262	0.1865	0.2422
0.60	0.3300	0.2223	0.1862	0.2405
0.65	0.3189	0.2114	0.1841	0.2339
0.70	0.2876	0.1962	0.1823	0.2112
0.75	0.2835	0.1904	0.1807	0.2084
0.80	0.2718	0.1823	0.1761	0.1988
0.85	0.2597	0.1705	0.1667	0.1877
0.90	0.2528	0.1683	0.1657	0.1864
0.95	0.2471	0.1662	0.1651	0.1726
1.00	0.2453	0.1662	0.1650	0.1772
평균정확률	0.3996	0.2811	0.2179	0.2810

때문이다. 또한, 초록/표제 그리고 전문에서 TPI의 평균정확률 값은 거의 동일하지만, 표준 재현율에서 정확률의 값을 비교해 보면, 전문에서의 TPI는 재현율 0.5지점 이후부터 정확률 값이 초록/표제의 정확률 값을 능가하고 있는 것을 알 수 있다. 이는 TPI가 높은 재현율에서

〈표 3〉 용어분산모형에 따른 8개 질문서의 평균정확률

	초록/표제			전문		
질문서# (총용어수)	BI 평균정확률	TPI 평균정확률	TP 용어수	BI 평균정확률	TPI 평균정확률	TP 용어수
44(7)	0.5658	0.6131	2	0.3008	0.7255	0
51(15)	0.2126	0.2473	4	0.1032	0.2456	2
58(6)	0.2546	0.3282	2	0.1269	0.2627	1
49(7)	0.6926	0.4360	2	0.5500	0.7778	1
37(10)	0.6086	0.5406	4	0.1651	0.5483	0
54(7)	0.5654	0.5551	3	0.4475	0.7143	0
15(4)	0.1426	0.0489	2	0.1881	0.0697	0
91(8)	0.4373	0.3098	2	0.2317	0.3054	1
평균	**0.4349**	**0.3848**	**2.625**	**0.2641**	**0.4561**	**0.625**

정확률을 증진시키고 있음을 보여주고 있는 것이다.

본 연구에서는 적합문헌 내에서의 용어의 분포를 살펴보기 위해 가능한 한 적합문헌을 많이 가지고 있는 질문서를 선정하였다. 83개의 질문서 중 9개 이상의 적합문헌을 가지고 있는 질문서 8개를 선정하여 각 질문서에 포함되어 있는 질문용어가 초록/표제 및 전문 데이터베이스 내에서 2-포아송 분포를 따르고 있는 지를 조사하였다. 이를 위해, Kolmogorov-Smirnov 검증(이하 K-S 검증)을 이용하였다(Winkler and Hays 1975). 질문용어는 적합문헌 및 부적합문헌 내에서는 각각 포아송 분포를 따르고 적합문헌 내의 평균출현빈도(λ_1)가 부적합문헌 내의 평균출현빈도(λ_2)보다 크면(즉, $\lambda_1 > \lambda_2$) 그 질문용어는 2-포아송 분포를 따른다고 볼 수 있는 것이다. K-S 검증 결과는 〈부록〉에서 나타나 있다. 각 질문용어의 적합문헌 또는 부적합문헌 내에서 포아송 분포 여부를 "P?"의 칸에 "Y" 또는 "N"으로 명시하였으며, 이것이 2-포아송 분포의 조건을 만족시키면 "TP?"의 칸에 같은 방법으로 명시하였다. 결과는 질문용어의 대부분이 2-포아송 분포를 따르지 않는 것으로 나타났으며, 질문용어는 전문에서 보다는 초록/표제에서 2-포아송 분포에 따르는 것으로 나타났다. 전문에서는 54.7%의 질문용어가 적합문헌 내에서 포아송 분포를 따르지 않았으며, 반면에 87.5%가 부적합문헌 내에서 포아송 분포를 따르지 않는 것으로 나타났다. 그리고, 초록/표제에서는 20.3%의 질문용어가 적합문헌 내에서 포아송 분포를 따르지 않았으며, 반면에 43.7%가 부적합문헌 내에서 포아송 분포를 따르지 않는 것으로 나타났다. 결과적으로 전문에서는 92.2%의 질문용어가, 초록

/표제에서는 67.2%의 질문용어가 2-포아송 분포를 따르지 않는 것으로 나타났다. 특히, 전문에서는 질문용어가 적합문헌보다는 부적합문헌 내에서 포아송 분포를 따르지 않았기 때문에 전체 문헌에서 2-포아송 분포를 따르지 않는 결과를 초래하였다. 이는 몇몇 질문용어가 전문의 부적합문헌 내에서 임의로 분산되어 있지 않고 있음을 암시하는 것이다. 동일한 주제를 다루고 있는 문헌으로 구성된 소규모 전문데이터베이스의 부적합문헌 내에서 용어가 임의로 분산될 기회는 매우 낮을 것임을 암시하고 있다.

선정된 8개 질문서의 번호와 그 질문서에 포함되어 있는 총 용어 수, 그리고 초록/표제 및 전문에서 2-포아송 분포를 따르는 용어의 수를 〈표 3〉에서 보여주고 있다. 또한 용어분산모형에 따른 질문서의 평균정확률을 살펴봄으로써 질문용어의 2-포아송 분포 여부가 검색효율성에 미치는 영향을 조사하였다.

각 질문서에 포함되어 있는 총 용어의 수에 비해 2-포아송 분포를 따르는 용어의 수(TP 용어수)는 상대적으로 적었으며, 특히 전문에서의 그 수는 주목할 만하다. 질문서 44와 37의 경우, 전문에서 2-포아송 분포를 따르는 용어가 전혀 없음에도 불구하고 TPI의 평균정확률은 각각 0.7255, 0.7143으로 초록/표제의 BI나 TPI보다도 높은 검색효율성을 보여주고 있다.

6 결 론

본 연구에서 사용된 CF392 데이터베이스 내에서 대부분의 질문용어는 2-포아송 분포에 따르지 않고 있다는 결론을 내릴 수 있다. 그 이유로, CF392가 "Cystic Fibrosis"와 관련된 동일한 주제를 다루는 매우 유사한 문헌으로 구성된 소규모 데이터베이스라는 사실을 들 수 있을 것이다. 그러나, 대부분의 질문용어가 2-포아송 분포에는 따르지는 않았지만 전문검색에 있어서 용어의 출현빈도는 TPI의 검색효율성을 향상시키고 있으며, 이는 특히 전문검색에 있어서 유용하게 사용될 수 있음을 알 수 있다. 이를 "Two Poisson Effectiveness Hypothesis"라고 부르고 있는 데, 이는 비록 질문용어가 2-포아송 분포에 따르지는 않았지만 "0"과 "1"이 아닌 실질적인 용어의 출현 빈도수를 반영함으로써, 특히 전문검색에서는 BI모형보다 많은 정보를 제공받아 검색효율성의 증진에 기여하고 있다는 것이다(Losee, Bookstein, and Yu 1986).

본 연구는 매우 유사한 주제를 다루고 있는 소규모 문헌집단을 이용한 실험이므로, 실험결과의 보편성을 얻기 위해서는 이질적인 문헌으로 구성된 대규모 전문데이터베이스를 이용한 2-포아송 모형의 응용에 대한 연구뿐만 아니라, 유사한 문헌으로 구성된 문헌집단 내에서 적합문헌 및 부적합문헌을 식별해 낼 수 있는 검색기법의 개발을 위한 계속적인 연구의 필요성을 제시하고 있다.

참 고 문 헌

문성빈. 1993. 적합성피이드백을 이용한 전문검색 시스템의 검색효율성 증진을 위한 연구. 『情報管理學會誌』, 10(2): 43-67.

Blair, D. C. and M.E. Maron. 1990. "Full-Text Information Retrieval: Further Analysis and Clarification." *Information Processing and Management*, 26(3): 437-447.

Blair, D. C., and M.E. Maron.1985. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System". *Communications of the ACM*, 28(3): 289-299.

Bookstein, A. 1983. "Information Retrieval: A Sequential Learning Process." *Journal of the American Society for Information Science*, 34(5): 331-342.

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

Fuhr, N. 1992. "Probabilistic Models in Information Retrieval." *The Computer Journal*, 35(3): 243-255.

Harter, S. P. 1975a. "A Probabilistic Approach to Automatic Keyword Indexing. Part I. On the Distribution of Specialty Words in a Technical Literature." *Journal of the American Society for Information Science*, 26(4): 197- 205.

Harter, S. P. 1975b. "A Probabilistic Approach to Automatic Keyword Indexing. Part II. An Algorithm for Probabilistic Indexing." *Journal of the American Society for Information Science*, 26(5): 280-289.

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

Losee, Jr., R. M. 1987. "The Effect of Database Size on Document Retrieval: Random and Best-First Retrieval Models." *In ACM Annual Conference on Research and Development in Information Retrieval* (pp.164-169). New York: ACM Press.

Losee, Jr., R. M., A. Bookstein, and C. Yu. 1986. "Probabilistic Models for Document Retrieval: A Comparison of Performance on Experimental and Synthetic Databases." *In ACM Annual Conference on Research and Development in Information Retrieval* (pp.258-264). New York: ACM Press.

Maron, M. E. 1988. "Probabilistic Design Principles for Conventional and Full-Text Retrieval Systems." *Information Processing and Management*, 24(3): 249-255.

Mckinin, E. J., M. Sievert, E.D. Johnson, and J.A. Mitchell.1990. "The Medline/Full-Text Research Project." *Journal of the American Society for Information Science*, 42(4): 297-307.

Raghavan, V. V., H. Shi, and C.T. Yu. 1983. "Evaluation of the 2-Poisson Model as a Basis for using Term Frequency Data in Searching." *In ACM Annual Conference on Research and Development in Information Retrieval* (pp. 88-100). New York: ACM Press.

Ro, J. S. 1988a. "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. I: On the Effectiveness of Full-Text Retrieval." *Journal of the American Society for Information Science*, 39(2): 73-78.

Ro, J. S. 1988b. "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. II: On the Effectiveness of Ranking Algorithms on Full-Text Retrieval." *Journal of the American Society for Information Science*, 39(3): 147-160.

Robertson, S. E. 1981. The Methodology of InformationRetrieval Experiment. In Sparck Jones, K. (Ed.), *Information Retrieval Experiment* (pp. 9-31). London. Butterworths.

Robertson, S. E. 1977. "The Probability Ranking Principle in IR." *Journal of Documentation*, 33(4): 294-304.

Shaw, W. M. Jr., R. E. Wood, and H. R. Tibbo 1991. "The Cystic Fibrosis Database: Content and Research Opportunities." *LISR*, 13: 347-366.

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

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

Tenopir, C. 1985. "Full Text Database Retrieval Performance." *Online Review*, 9(2): 149-164.

Winkler, R. L. and W. L. Hays. 1975. *Statistics, Probability, Inference, and Decision*. New York: Holt, Rinehart and Winston.

〈부록〉 각 질문서에 나타난 질문용어의 K-S 검증의 결과

질문서 44

What structual or enzymatic differences are there between fibroblasts from CF patients and non-CF patients?

질의용어	전문					초록/표제				
	적합문헌		부적합문헌			적합문헌		부적합문헌		
	λ_1	P?	λ_2	P?	TP?	λ_1	P?	λ_2	P?	TP?
CF	29.03	N	16.73	N	N	1.82	N	1.28	N	N
DIFF	6.18	Y	3.40	N	N	3.40	Y	0.31	Y	Y
ENZYM	11.42	N	2.32	N	N	1.27	N	0.11	Y	N
FIBROBLAST	24.3	N	0.5	N	N	3.03	Y	0.03	Y	Y
NON	0.82	Y	0.98	N	N	0.06	Y	0.07	Y	N
PATI	11.7	N	26.01	N	N	1.91	Y	2.13	N	N
STRUC	0.79	N	0.42	N	N	0.03	Y	0.03	Y	N

질문서 51

What circulating or secreted "factors" have been described in CF patients? ("Factors" are unidentified biologically active molecules thought to play some pathogennetic role in cystic fibrosis).

질의용어	전문					초록/표제				
	적합문헌		부적합문헌			적합문헌		부적합문헌		
	λ_1	P?	λ_2	P?	TP?	λ_1	P?	λ_2	P?	TP?
ACT	17.6	N	5.98	N	N	2	N	1.28	N	N
BIOLOG	1.07	Y	0.17	Y	Y	0.2	Y	0.31	Y	Y
CF	40.77	N	15.86	N	N	2.57	N	0.11	N	N
CIRCL	0.43	N	0.59	N	N	0	N	0.03	Y	N
CYST	13.43	N	13	N	N	3.23	Y	0.07	N	N
FACT	13.27	N	2.2	N	N	1.13	Y	2.13	Y	Y
FIBROS	13	N	13.03	N	N	3.23	Y	0.03	N	N
MOLECL	3.17	N	0.59	N	N	0.33	Y	0.05	Y	Y
PATHOGEN	0.03	Y	0.08	Y	N	0	N	0	N	N
PATI	10.83	N	25.96	N	N	0.97	Y	2.2	N	N
PLAY	0.07	Y	0.24	Y	N	0	N	0.02	Y	N
ROL	0.5	Y	0.63	N	N	0	N	0.07	Y	N
SECRET	4.47	Y	2.93	N	N	0.33	Y	0.22	N	N
THOUGHT	0.23	Y	0.33	N	N	0.07	Y	0.01	Y	Y
UNIDENT	0.07	Y	0.03	Y	Y	0	N	0.01	Y	N

질문서 58
What is the immunologic response to pulmonary infection in CF patients?

질의용어	전문					초록/표제				
	적합문헌		부적합문헌			적합문헌		부적합문헌		
	λ_1	P?	λ_2	P?	TP?	λ_1	P?	λ_2	P?	TP?
CF	28.39	N	17.11	N	N	1.87	N	1.29	N	N
IMMUNOLOG	0.87	Y	0.11	Y	Y	0.09	Y	0.01	Y	Y
INFECT	12.22	N	2.34	N	N	1.65	Y	0.15	Y	Y
PATI	58.87	N	22.68	N	N	5.74	Y	1.88	N	N
PULMON	4.44	N	3.66	N	N	0.48	Y	0.26	N	N
RESPONS	11.7	N	2.34	N	N	0.96	N	0.19	N	N

질문서 49
Is RNA methylation or polyamine metabolism normal in CF patients?

질의용어	전문					초록/표제				
	적합문헌		부적합문헌			적합문헌		부적합문헌		
	λ_1	P?	λ_2	P?	TP?	λ_1	P?	λ_2	P?	TP?
CF	17.09	N	17.79	N	N	1.18	N	1.33	N	N
METABOL	3.64	Y	0.81	N	N	0.27	Y	0.07	Y	Y
METHYL	9.36	N	0.14	N	N	1.18	N	0.01	Y	N
NORM	17.82	N	10.23	N	N	2.27	Y	2.12	N	N
PATI	15.09	Y	25.08	N	N	1.91	N	0.07	N	N
POLYAM	22.27	Y	0.03	Y	Y	1.46	Y	0	Y	Y
RNA	13.09	N	0.1	N	N	1.55	N	0	Y	N

질문서 37

What techniques are available for screening of newborn infants for CF, and what factors contribute to erroneous results of these tests?

질의용어	전문					초록/표제				
	적합문헌		부적합문헌			적합문헌		부적합문헌		
	λ_1	P?	λ_2	P?	TP?	λ_1	P?	λ_2	P?	TP?
CF	30	Y	17.42	N	N	2.09	Y	1.3	N	N
CONTRIBUT	0.27	Y	0.56	N	N	0	N	0.05	Y	N
ERRON	0.00	N	0.01	Y	N	0	N	0	N	N
FACT	2.09	Y	3.08	N	N	0.09	Y	0.22	Y	N
INF	23.82	N	1.65	N	N	1.46	N	0.13	Y	N
NEWBORN	7.73	N	0.37	N	N	1	Y	0.04	Y	Y
RESULT	15	Y	5.57	N	N	0.91	Y	0.27	Y	Y
SCREEN	21.09	N	0.92	N	N	2.46	Y	0.1	Y	Y
TECHNIQU	3.18	Y	0.97	N	N	0.64	Y	0.06	Y	Y
TEST	0.07	Y	5.95	N	N	2.46	Y	0.37	N	N

질문서 54

What is the relationship of allergy or hypersensitivity to lung disease in CF patients?

질의용어	전문					초록/표제				
	적합문헌		부적합문헌			적합문헌		부적합문헌		
	λ_1	P?	λ_2	P?	TP?	λ_1	P?	λ_2	P?	TP?
ALLERG	31	N	0.6	N	N	2.5	Y	0.55	Y	Y
CF	32.1	N	17.39	N	N	3.5	Y	1.27	N	N
DISEAS	9.7	Y	6.47	N	N	0.7	Y	0.58	N	N
HYPERSENSIT	6.3	N	0.15	N	N	0.7	Y	0.02	Y	Y
LUNG	3.7	N	3.1	N	N	0.2	Y	0.22	N	N
PATI	60.7	N	28.86	N	N	5.1	Y	2.03	N	N
RELAT	1.5	Y	0.77	N	N	0.1	Y	0.06	Y	N

질문서 15
What are the hepatic complications or manifestations of CF?

질의용어	전문					초록/표제				
	적합문헌		부적합문헌			적합문헌		부적합문헌		
	λ_1	P?	λ_2	P?	TP?	λ_1	P?	λ_2	P?	TP?
CF	12.67	N	17.89	N	N	0.56	Y	1.34	N	N
COMPL	1.11	N	1.27	N	N	0	N	0.1	Y	N
HEP	8.11	Y	0.49	N	N	0.67	Y	0.02	Y	Y
MANIFEST	1.56	Y	0.49	N	N	0.11	Y	0.03	Y	Y

질문서 91
What are the unusual manifestations of CF (other than lung disease or exocrine pancreatic in sufficiency?

질의용어	전문					초록/표제				
	적합문헌		부적합문헌			적합문헌		부적합문헌		
	λ_1	P?	λ_2	P?	TP?	λ_1	P?	λ_2	P?	TP?
CF	7.35	N	18.51	N	N	0.5	Y	1.39	N	N
DISEAS	7.58	N	6.48	N	N	0.92	Y	0.56	N	N
EXOCR	0.27	Y	0.66	N	N	0	N	0.07	Y	N
INSUFFICI	0.92	Y	0.51	N	N	0	N	0.06	Y	N
LUNG	2.04	Y	3.19	N	N	0.08	Y	0.23	N	N
MANIFEST	1.15	Y	0.46	N	N	0.04	Y	0.03	Y	Y
PANCREAT	4.27	N	3.36	N	N	0.39	Y	0.29	N	N
UNUSU	0.42	Y	0.23	Y	Y	0.08	Y	0.02	Y	Y