The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 1-11

Classification Analysis in Information Retrieval by Using Gauss Patterns[1)]

Jung Jin Lee[2)], Soo Kwan Kim[3)]

Abstract

This paper discusses problems of the Poisson Mixture model which is widely used to decide the effective words in judging relevant document. Gamma Distribution model and Gauss Patterns model as an alternative of the Poisson Mixture model are studied. Classification experiments by using TREC sub-collection, WSJ[1,2] with MGQUERY and AidSearch3.0 system are discussed.

Keywords : Poisson Mixture Model, Gauss Pattern, Information Retrieval

1. 서론

정보화 사회에서 기하급수적으로 증가하는 정보의 양은 우리에게 '수많은 정보 속에서 어떻게 효율적으로 신속하고 정확하게 정보를 찾을 수 있느냐?'라는 문제를 제기하고 있다. 효율적인 정보검색(information retrieval)은 각종 의사결정에 매우 중요하여 그 결과에 따라 개인이나 기업, 그리고 국가의 성패가 달라질 수 있다. 최근에는 효율적인 정보검색을 위해 여러 가지 통계학 기법이 많이 이용하고 있다(Lee & Kantor, 1991, 1998). 정보검색 연구 중에서도 전문(full text)을 이용하여 검색하는 방법은 자연어를 이용한 검색을 가능하게 하는 것으로, 자료의 내용을 분석 가공하여 특정한 개념을 포함하는 자료를 검색할 수 있는 기능을 제공한다. 이러한 방법은 인터넷 검색 및 디지털 도서관(digital library) 구축에 있어 비정형데이터의 처리와 대규모 데이터베이스에서 너무 많은 자료들이 검색되는 문제점을 피할 수 있다.

전문을 이용한 정보검색에 관한 연구를 가장 활발하게 발표하고 있는 학술대회는 TREC(Text REtrieval Conference)으로 현재 9회에 이르고 있다. TREC에서 다루고 있는 데이터베이스에는 Wall Street Journal, Federal Register, Associated Press 등 100만개 이상의 문서들이 있으며 실험을 위한 질문 주제(topic)는 현재 총 450가지이다. <표 1-1>은 TREC에서 활용하는 데이터베이스에 대한 질문 주제의 예(Topic 001)로서 이를 근거로 전문가들이 실험문서에 대해 적절(relevant), 부적절(non-relevant) 판정을 한다. TREC에 참가하는 연구자는 주제와 적절성이 판정된 실험문서 데이터를 이용해 다양한 검색모형을 실험한다. 최근에는 텍스트 이외에도 음성과 영상에 대한 검색도 활발한 연구가 진행 중에 있다.

---

1) This research is supported in part by the 1999 Soong Sil University Research Fund.
2) Professor, Department of Statistics, SoongSil University, Seoul, 156-743, Korea.
E-mail : jjlee@stat.soongsil.ac.kr
3) Researcher, Parole Science Inc., 2F 882-5 Bongcheon4-dong Gwanak-gu Seoul, 151-716, Korea
E-mail : skymaru@hotmail.com

<표 1-1> 주제 001에 대한 적절성 판정에 대한 설명

<num> Number: 001
<dom> Domain: International Economics
<desc> Description:
Document discusses a pending antitrust case.
<narr> Narrative:
To be relevant, a document will discuss a pending antitrust case and will identify the alleged violation as the government entity investigating the case. Identification of the industry and the companies involved is optional. The antitrust investigation must be a result of a complaint, NOT as part of a routine review.
<con> Concept(s):
1. antitrust suit, antitrust objections, antitrust investigation, antitrust dispute
2. monopoly, bid-rigging, illegal restraint of trade, insider trading, price-fixing
3. acquisition, merger, takeover, buyout
4. Federal Trade Commission (FTC), Interstate Commerce Commission (ICC), Justice Department, U.S. Securities and Exchange Commission (SEC), Japanese Fair Trade Commission
5. NOT antitrust immunity
<fac> Facter(s):
<def> Definition(s)

<표 1-2>는 MGQUERY 검색시스템을 이용하여 주제 001에 대한 80개의 스템(stem)된 단어를 보여주고 있다. 효율적인 문서의 적절성 판정을 위해서는 이들 단어 중에서 어떠한 단어들이 더 도움이 되는지 연구하여야 한다. 본 논문의 2절에서는 전문 정보검색을 위한 유용한 판별단어선택에 제일 많이 이용되는 포아송 혼합(Poisson mixture) 모형을 소개하고 그 문제점을 TREC 서브 컬렉션 WSJ[1,2](Wall Street Journal disk 1, 2)를 대상으로 MGQUERY와 AidSearch 3.0 시스템을 이용하여 실험한 결과를 분석하였다. 3절에서는 포아송 혼합모형의 대안으로 감마분포모형과 가우스 패턴(pattern) 모형을 모형을 제안하였다. 4절에서는 단어들의 관련성에 근거한 가우스 패턴을 이용하여 문서의 적절성을 판별하는 실험을 소개하고, 5절에서 결론 및 제안을 한다.

<표 1-2> MGQUERY 검색시스템을 이용하여 주제 001에 대한 80개의 스템(stem)된 단어

acquir	complain	exchang	insid	object	routin
action	concept	fact	interest	option	sec
alleg	control	fair	intern	part	secur
antitrust	corpor	feder	interst	pend	stock
bid	defin	fix	investig	practic	suit
busi	depart	friend	involut	pric	take
buyout	describ	ftc	japan	protect	takeov
call	discuss	govern	justic	relev	topic
cas	disput	ident	law	restrain	trad
case	doc	identif	merg	result	u.s
commerc	docu	illeg	monopo	retain	unfair
commit	econom	immun	narrat	retir	unfriend
compan	entit	industr	numb	review	unlaw
				rig	violat

2. 판별단어 선택을 위한 포아송 혼합모형

정보검색을 위하여 제일 많이 이용되는 모형은 단어의 출현성 여부를 이용하여 검색하는 불린(boolean) 모형이다. 하지만 정보의 양이 많아지면서 이 불린모형은 너무 많은 문서를 찾아주기 때문에 그 효용성에 문제가 제기되면서 전문(full text)을 이용한 새로운 모형 연구에 많은 노력을 기울이고 있다. 전문검색에서 전문 전체를 질문으로 이용하는 것은 힘들어 전문중 문서의 적절성을 판단하는데 유용한 단어들을 취사 선택하는데 여러 가지 통계적 기법이 이용되고 있다.

이러한 단어선택에 제일 많이 이용되고 있는 것이 포아송 혼합(Poisson mixture) 모형으로 Bookstein과 Swanson(1974)에 의해 제안되어 Harter(1975a; 1975b)에 의해 연구되어 최근 많은 정보검색시스템에 실제로 이용되고 있다.

어느 문서에 나타나는 한 단어의 수를 확률변수 X라 하자. 포아송 혼합모형이란 장서 내에서 주제를 표현하지 못하는 단어인 비주제단어(non-specialty words)는 전체문헌 내에서 단일 포아송분포에 의해 표현되어질 수 있다고 가정하고, 주제를 표현할 수 있는, 즉 적절성을 판별할 수 있는 단어인 주제단어(specialty words)는 적절한 문서들과 부적절한 문서들 사이에 서로 다른 분포를 가지고 있다는 가정을 한다. 즉, 주제단어는 적절한 문서 및 부적절한 문서 내에서 각각 서로 다른 평균 출현빈도를 갖는 다음과 같은 포아송 혼합 모형으로 표현될 수 있다.

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

여기서 π는 장서 내에 적절한 문서들이 차지하는 비율을 나타내며, λ_1, λ_2는 각각 적절한 문서와 부적절한 문서 내에서 단어의 평균출현빈도를 나타낸다. 일반적으로 적절한 문서 내에서 단어의 평균출현빈도는 부적절한 문서 내에서 단어의 평균출현빈도보다 크다고 가정한다(즉, $\lambda_1 > \lambda_2$).

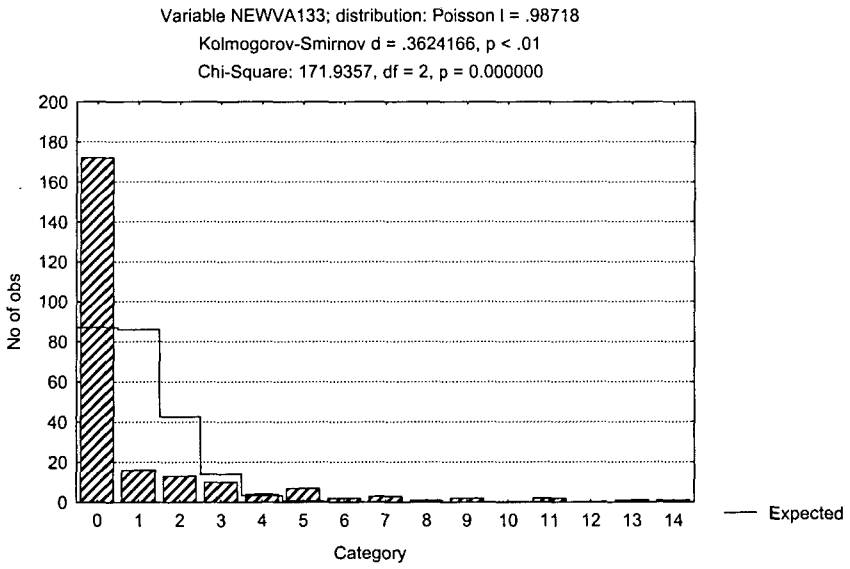
따라서 전문중 문서의 적절성을 판단하는데 유용한 단어들을 찾기 위해서는 전문가들에 의해 적절하다고 판정된 실험문서와 부적절하다고 판정된 실험문서들 중 λ_1이 λ_2보다 월등히 큰 단어들을 이용한다.

2.1 포아송 혼합모형의 문제점

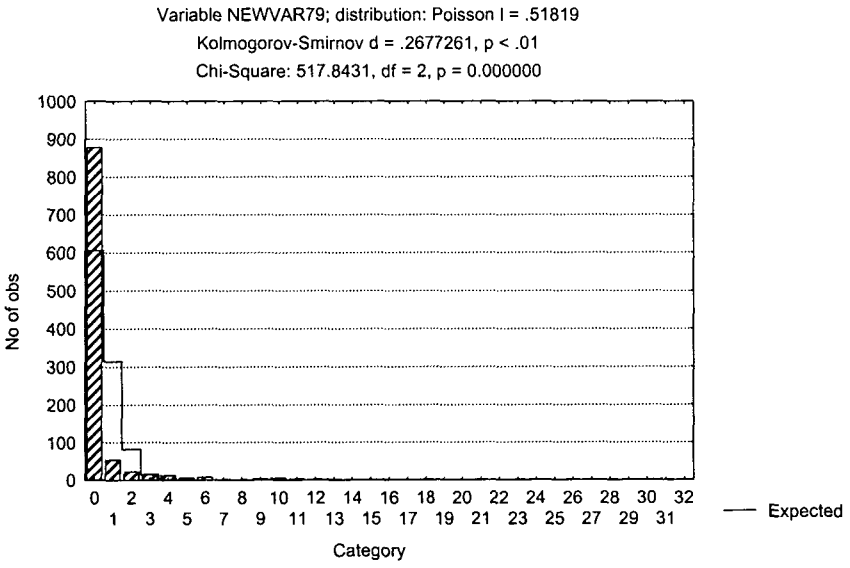
포아송 혼합 모형은 실제 검색시스템에 많이 이용되고 있는데 최근 이 모형의 가정에 대하여 여러 가지 의문이 제기되고 있다. Church(1995)는 주제단어들의 대부분은 포아송 분포보다 더 두꺼운 꼬리를 갖는다고 말하고 있다. 실제로 TREC 문서집합의 WSJ[1,2](Wall Street Journal disk 1, 2) 문서 내에서 질문주제 001에 대한 적절한 문서와 부적절한 문서에 대해 선택한 대부분의 단어들은 Kolmogorov-Smirnov 검정결과 포아송 분포를 따르지 않고 있다(예: <그림 2-1> 및 <그림 2-2>).

<표 2-1>은 여섯 개의 주제에 나타나는 단어들을 포아송분포로 Kolmogorov-Smirnov 적합성검정을 한 결과이다. 주제에 따라 약간씩의 차이는 있지만 대부분 10% 미만의 단어들이 포아송분포에 적합한 것으로 나타나 포아송분포라는 가정에 문제가 있음을 볼 수 있다.

<그림 2-1> 주제 1의 적절한 문서집합에 대한 단어 'antitrust' 분포의 K-S 검정 결과



<그림 2-2> 주제 1의 부적절한 문서집합에 대한 단어 'antitrust' 분포의 K-S 검정결과



<표 2-1> 여섯 개의 주제에 나타나는 단어들의 포아송분포로 적합된 단어들의 수

	문서집합		
주제	적절문서	부적절문서	총 단어수
주제 001	3(3.8%)	1(1.3%)	80
주제 002	3(2.8%)	0(0.0%)	106
주제 013	3(6.0%)	0(0.0%)	50
주제 027	2(2.0%)	0(0.0%)	97
주제 033	3(4.5%)	0(0.0%)	66
주제 044	0(0.0%)	4(10.8%)	37

포아송 혼합 모형의 또 다른 문제점은 모든 문서의 길이가 같다는 가정이다. 하지만 <표 2-2>에서 보듯이 문서의 길이는 매우 차이가 난다. 문서의 길이가 모형에 어떠한 영향을 주었는지 살피기 위해 각 단어의 출현빈도(term frequency)를 문서의 길이로 나누어 준 후 포아송 분포의 적합성 여부를 살펴보았지만 단어가 출현하지 않은 문서의 수가 많아 포아송 분포를 따르지 않는 것을 알 수 있다. 이와 같은 실험결과를 보면 여러 정보검색 시스템에서 사용하고 있는 포아송 혼합모형은 그 효용성이 의문시 된다.

<표 2-2> 각 주제에 속하는 문서들의 길이 (단위 byte)

주제		최소	일사분위수	평균	삼사분위수	최대
주제 001	적절문서	436	1607	4646	5872	32291
	부적절문서	336	1262	3986	5566	52532
주제 002	적절문서	386	1183	3858	5858	15467
	부적절문서	227	1147	3929	5444	80422
주제 013	적절문서	405	1333	4111	5986	22218
	부적절문서	243	1262	4654	6260	80422
주제 027	적절문서	502	1359	5052	7348	12043
	부적절문서	384	1365	4358	6274	15629
주제 033	적절문서	371	929	2973	4393	19864
	부적절문서	502	1362	6322	5918	80422
주제 044	적절문서	441	1274	3638	5281	18269
	부적절문서	350	1304	4201	5512	80422

3. 적절성 판별을 위한 용어선택 연구

2절에서 살펴보았듯이 판별단어 선택을 위한 포아송 혼합모형은 실제 데이터에 대해 적용할 때 여러 가지 문제가 있음을 알 수 있다. 이 절에서는 포아송 혼합모형의 대안으로 감마분포 모형과 가우스패턴 모형을 제안한다.

3.1 감마분포 모형의 적합성

포아송분포가 적합하지 않은 이유중의 하나는 한 단어(unique term)가 한 번도 출현하지 않는 문서의 수가 대부분의 경우를 차지하기 때문이다. 따라서 본 연구에서는 포아송분포 대신에 분포의 꼬리가 길고 급격한 감소를 하는 감마분포를 문서의 길이가 고려된 적절성 판별을 위한 단어를 구별하기 위한 분포모형으로 실험하여 보았다. 감마분포는 b를 척도 모수, c를 형상모수라 할 때 다음과 같이 표시된다.

$$f(x) = \frac{(x/b)^{c-1}\exp(-x/b)}{b\Gamma(c)}$$

감마분포 적합성 실험에 사용된 문서들은 TREC에서 이미 적절 또는 부적절한 문서로 구분되어진 문서집합으로, 예를 들면 주제 001에 대해 234개의 적절한 문서와 1017개의 부적절한 문서가 있다. 주제 001의 문서들을 대상으로 MGQUERY를 이용하여 불용어를 제거한 단어들(<표 1-2>)을 가지고 각 문서 집합들에 대해 감마분포 적합을 하였다. 계산의 편의를 위해 각각의 적절한 문서와 부적절한 문서에서의 단어의 출현빈도(term frequency)를 각각의 문서 길이로 나눈 값에 1000을 곱해서 나온 값을 분포 적합에 사용하였다. <표 3-1>은 감마분포에 대한 Kolmogorov-Smirnov 적합성 검정결과이다. 결과를 살펴보면 적절한 문서이든 부적절한 문서이든 대개 10%에서 20%의 단어만이 감마분포를 따르고 있어 모형으로 사용하기에는 불충분함을 알 수 있다. 따라서 감마분포도 판별단어 선택을 위한 모형으로는 적절치 못한다는 것을 알 수 있다.

<표 3-1> 여섯 개의 주제에 나타나는 단어들의 포아송분포로 적합된 단어들의 수

	문서집합		
주제	적절문서	부적절문서	총 단어수
주제 001	11(13.8%)	14(17.5%)	80
주제 002	28(26.4%)	18(17.0%)	106
주제 013	9(18.0%)	11(22.0%)	50
주제 027	7(7.2%)	20(20.6%)	97
주제 033	8(12.1%)	16(24.2%)	66
주제 044	3(8.1%)	7(18.9%)	37

3.2 가우스 패턴을 이용한 용어선택모형

3.2.1 가우스 패턴

앞 절에서 포아송분포나 감마분포 모두 단어 선택에 적합한 모형이 아님을 살펴보았다. 이 절에서는 이들 모형의 대안으로 단어들의 관련성을 이용하는 가우스 패턴(Gauss pattern)을 이용한 모형을 제안한다.

모든 문서들의 집합을 R 이라 하고, 이중에서 주어진 질문에 '적절하다'고 판정된 문서들의 집합과 '부적절하다'고 판정된 문서들의 집합을 각각 R^+ 와 R^- 로 표시하고, 집합 $T_1, T_2, \cdots, T_p$ 를 단어의 개수(차수)가 동일한 판별을 위한 단어들의 집합이라 하자. 예를 들어, 단어집합 T_i 가 두 개의 단어 A와 B를 포함(차수가 2)하고 있다면 전체 문서들(R)은 단어의 출현성 여부에 따라 다음과 같이 네 개의 불린(boolean)집합으로 표시할 수 있다.

$$R = \overline{A}\overline{B} \cup \overline{A}B \cup A\overline{B} \cup AB$$

적절하다고 판정된 문서들이 각 불린집합에 속하는 확률분포를 $f_i(x)$라 하고, 부적절하다고 판정된 문서들이 각 불린집합에 속하는 확률분포를 $g_i(x)$라 하자. 만일 두 분포 $f_i(x)$와 $g_i(x)$가 서로 다르다면 단어집합 T_i는 한 문서가 적절한지 아닌지를 판별하는 정보를 지니고 있는 패턴(pattern)이라 정의하자. T_i 가 패턴인지 아닌지를 구별하기 위해서는 두 분포를 비교하는 전통적인 방법인 카이제곱 검정이나 Kolmogorov-Smirnov 검정을 생각할 수 있다. 하지만 이러한 경우 너무 많은 패턴이 생성될 수 있기 때문에, 이때는 전체 패턴 중 두 분포함수의 가우스거리(Gauss discrepancy measure)

$$d(f_i, g_i | T_i) = \sum_x [f_i(x) - g_i(x)]^2$$

가 큰 패턴들만을 이용하여 실험할 수 있는데 이를 가우스패턴이라 하자.

3.2.2 가우스패턴의 생성

패턴을 생성하는 가장 직관적인 방법은 모든 단어에 대한 조합을 다 조사하는 것이다. 하지만 TREC에는 450여개의 주제가 있고 각 주제에 해당하는 문서들의 집합에서 가우스패턴을 찾아내는 일은, 특히 차수가 높아질수록, 수 많은 계산을 요구한다. 본 논문에서는 시간가 장비의 제약상 차수가 1 또는 2인 경우에 시험적으로 가우스패턴을 찾아 모형의 타당성을 조사하였다.

주제 1인 경우에 차수가 1인 한 단어의 경우의 수는 80가지이고, 차수가 2인 경우의 두 단어 조합의 경우의 수는 316가지가 된다. 본 연구에서는 각 차수의 전체 단어집합 중에서 가우스거리가 큰 상위 25개의 가우스 패턴을 이용하여 실험을 하였다. 몇 개의 패턴을 이용하는 판별을 하는 것이 효과적인가 하는 문제와, 차수가 높을 때 효율적으로 패턴을 찾는 알고리즘은 향후 더 연구하여 볼 주제이다.

<표 3-2>는 주제 001에서 선택한 80개의 단어들(<표 1-2>)에 대하여 한 단어와 두 단어집합에 대한 가우스거리가 큰 상위 25개의 가우스패턴을 보여주고 있다. 선택된 단어들을 살펴보면 주제와 밀접한 단어

들이 선택되었음을 알 수 있다.

예를 들면 한 단어 패턴에서는 monopo, violat, investig, antitrust 등이 상위에 나타난다. 한 단어 패턴은 두 단어의 가우스 패턴에서도 역시 나타나는 것을 알 수 있어 선택된 단어들이 적절문서를 판별하는데 유용할 가능성을 보여 준다.

<표 3-2> 한 단어와 두 단어의 가우스패턴 (주제 001)

	한 단어 가우스패턴		두 단어 가우스패턴		
순위	단어	가우스 측도	단어 1	단어 2	가우스측도
1	concept	0.822	narrt	violat	0.0886
2	monopo	0.059	violat	restrain	0.0883
3	violat	0.043	numb	violat	0.0879
4	investig	0.033	violat	doc	0.0868
5	feder	0.030	violat	interst	0.0856
6	alleg	0.028	violat	fair	0.0854
7	case	0.022	violat	unfair	0.0852
8	involut	0.020	violat	defin	0.0851
9	u_s_	0.020	monopo	violat	0.0847
10	law	0.016	violat	rig	0.0841
11	antitrust	0.016	violat	concept	0.0838
12	govern	0.014	violat	unfriend	0.0835
13	illeg	0.013	violat	topic	0.0831
14	depart	0.012	violat	object	0.0815
15	part	0.011	violat	investig	0.0798
16	trad	0.010	violat	retir	0.0795
17	docu	0.010	japan	violat	0.0794
18	suit	0.010	violat	immun	0.0790
19	identif	0.010	violat	buyout	0.0779
20	fact	0.010	violat	commerc	0.0779
21	numb	0.009	violat	identif	0.0777
22	call	0.009	describ	violat	0.0772
23	justic	0.008	violat	unlaw	0.0772
24	doc	0.008	investig	ftc	0.0770
25	insid	0.008	violat	relev	0.0765

4. 문서의 적절성 판별연구

4.1 문서의 적절성 판별모형

3절에서 전문(full text)을 이용하는 정보검색에서 문서의 적절성을 판별하기 위한 단어선택에 가우스패턴 모형을 연구하였다. 이 절에서는 가우스 패턴을 이용한 문서의 적절성 판별모형을 알아보자. 한 가우스 패턴 T_i 가 적절하다고 판정된 문서들에 나타나는 확률분포를 $f_i(x)$라 하고, 부적절하다고 판정된 문서들이 각 불린집합에 속하는 확률분포를 $g_i(x)$라 하고 각각의 패턴 T_i는 서로 독립이라고 가정하자. 그러면 p 개의 패턴 $T_1, T_2, \cdots, T_p$ 가 적절하다고 추정되는 한 문서에 $t_1, t_2, \cdots, t_p$ 로 나타날 수 있는 확률은

$$\prod_{i=1}^{p} f_i(t_i)$$

이고, 부적절하다고 추정되는 문서에 나타날 수 있는 확률은

$$\prod_{i=1}^{p} g_i(x)$$

이 된다. 따라서 문서의 적절성을 판별하는 식은

'만일 $\prod_{i=1}^{p} f_i(t_i) \geq \prod_{i=1}^{p} g_i(x)$ 이면 적절한 문서로, 그렇지 않으면 부적절한 문서' 이다.

4.2 판별 실험

판별 실험을 위해 TREC WSJ[1,2] 데이타베이스를 이용하였는데 100만개가 넘는 데이터를 모두 실험하기에는 시간 및 장비의 제약이 있어 3절에서 패턴을 생성할 때 실험한 질문 주제 001에 대한 문서를 대상으로 검색모형의 판별식을 만들고, 그 문서들을 대상으로 적절성 여부를 판별식을 이용하여 실험하였다.

한 단어 패턴을 이용한 실험(<표 4-1>)에서 적절(부적절)한 문서 집합에서는 12%(22%)내외의 문서들이 패턴이 없어 판별불능한 것으로 나타났다. 판별이 불가능한 문서를 제외한 206개의 적절문서 중에서는 53%(110)내외의 적중률을 보여 주고 있으며, 792개의 부적절문서 중에서는 67%(528)내외의 적중률을 보여 주고 있다.

<표 4-1> 차수 1(한 단어)의 판별 결과 (주제 001)

		실제	
		적절	부적절
판별	적절	110(47%)	264(26%)
	부적절	96(41%)	528(52%)
	불능	28(12%)	223(22%)
		234(100%)	1015(100%)

두 단어 패턴을 이용한 실험(<표 4-2>)에서는 적절(부적절)한 문서 집합에서는 28%(44%) 내외의 판별불능률을 보여 패턴의 차수가 높아질 수록 그 패턴이 문서에 있을 가능성이 낮아져 불능률이 높아짐을 보여 주고 있다. 판별불능인 문서를 제외하고는 적절한 문서(168개) 중에서는 54%(91)내외의 적중률을 보여 주고 있으며, 부적절 문서(568) 중에서는 71%(406)내외의 적중률을 보여 주고 있다. 즉 차수가 높아질 수록 판별불능인 문서들의 수는 늘어날 수 있지만 판별적중률은 높아진다는 사실을 말해 주고 있다. 두 단어 패턴인 경우 용어의 종속성을 고려하기 때문에 판별성공률은 높지만 판별불능률이 30%에 이르는 것은 이 모형이 해결하여야 할 과제라고 생각된다. 고려해볼 수 있는 방법은 먼저 상위차수의 패턴을 이용하여 판별을 실시한 후 그 차수에 대한 판별 불능 문서에 대해서는 하위차수의 패턴을 이용하여 단계적으로 판별을 실시할 수도 있을 것이다.

<표 4-2> 차수 2(두 단어)의 판별 결과 (주제 001)

		실제	
		적절	부적절
판별	적절	91(39%)	162(16%)
	부적절	77(33%)	406(40%)
	불능	66(28%)	447(44%)
		234(100%)	1015(100%)

5. 결론 및 제안

본 논문에서는 전문 정보검색에 많이 이용되는 포아송 혼합(Poisson mixture) 모형의 문제점을 연구한 후, 대안으로서 감마분포 모형을 검토하였으나 역시 만족스럽지 않음을 살펴보았다. 다른 대안으로 가우스 패턴(pattern)을 이용한 적절성 판별용어 모형을 제안하고 이를 이용한 문서의 적절성을 판별하는 실험을 실시하였다. 이 방법은 기존의 포아송 혼합 모형이 갖고 있는 문제의 대안으로 사용 될 수 있고 단어의 종속성을 고려할 수 있다는 측면에서 판별을 위한 단어의 결정에 이용할 수 있다. 그리고 이러한 가우스패턴은 각 문서집단에서의 패턴 확률을 구하여 문서의 판별에 직접 사용할 수 있다.

하지만 본 연구에서는 시간과 장비의 제약상 단지 한 단어 또는 두 단어의 패턴에 대한 실험만 실시하였다. 초고속 대용량 컴퓨터를 이용하여 모든 차수의 패턴을 구하고 적절한 패턴들의 결합에 대한 확률적 모형의 연구가 차후에 필요하다고 생각된다. 그밖에 유용한 단어결정에 있어서 한 문서 안에서 단어들 간의 거리를 포함시키던가, 문서별 단어의 빈도수를 고려하는 방법 등 복잡한 방법들을 생각할 수 있다.

참고문헌

[1] Bookstein, A., & Swanson, D.R. (1974). Probabilistic Models for Automatic Indexing. *Journal of the American Society for Information Science*, 34, 331-342

[2] Church KF. Gale WA. (1995) Inverse Document Frequency (IDF): A measure of deviation from Poisson. *Proceedings of the Third Workshop on Very Large Corpora.*

[3] Harter, S.P. (1975). 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, 197-205.

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

[5] Jung Jin Lee and Paul B. Kantor. "A Study of Probabilistic Information Retrieval Systems in the Case of Inconsistent Expert Judgments." *Journal of American Society for Information Science*, V42, pp 166-172, 1991.

[6] Paul B. Kantor and Jung Jin Lee. "Testing the Maximum Entropy Principle for Information Retrieval." *Journal of American Society for Information Science*, Vol 49, 6, pp557-566, 1998.

[7] Robertson, S.E. & Walker, S., Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. *Proceedings of the 17th Annual International ACM SIGIR Conference in Research and Development in Information Retrieval*, pp. 232-241, 1994

[2001년 3월 접수, 2001년 11월 채택]