응용통계연구 제 13권 1호, 2000년, pp. 1-10

논리적 패턴을 이용한 확률화 정보검색 시스템의 연구 *

이윤오[1] 이정진[2]

요 약

정보화사회에서 효율적인 정보검색(information retrieval)은 각종 의사결정에 매우 중요하다. 주어진 정보검색 문제가 있을 때 과거에 검색되었던 자료는 그 적절성여부에 대한 평가를 데이타베이스에 첨가하여 지식베이스(knowledge base)화 할 수 있다. 본 연구는 이 지식베이스에 대한 논리적 패턴을 분석하여 새로운 정보의 '적절성(relevance)' 여부를 판별하는 확률화 정보검색 모형을 만들고 이에 대한 실험을 하였다.

주요용어: 정보검색, 논리적 패턴, 지식베이스.

1. 서론

인터넷과 같은 각종 네트워크가 보편화되는 미래의 정보화사회에서 효율적인 정보검색(information retrieval)은 각종 의사결정에 매우 중요하여 그 결과에 따라 개인이나 기업, 그리고 국가의 성패가 달라질 수 있다. 이를 위해 더 빠르고 대용량인 컴퓨터 하드웨어와, 더 효율적인 소프트웨어의 발전을 위해 컴퓨터 전문가들이 많이 노력하고 있다. 정보검색에 관한 이론은 컴퓨터과학의 데이타베이스 분야에서 많이 연구되고 있지만 최근에는 통계학의 여러 기법을 이용하여 정보를 효율적으로 검색하는 연구가 많이 진행되고 있다.

본 연구에서는 TREC6 자료를 적절성(relevancy) 판단 정보를 갖는 복잡한 불린(boolean) 표현으로 변환한 후 Cramer 와 Hammer 등(1988, 1994)에 의해서 제안된 패턴을 이용한 논리적 자료분석(Logical Analysis of Data: LAD)으로 자료에 대한 적절성을 판별하는 실험이 시도되었다. 논리적 자료분석 외에 이항자료(binary data)에 대한 판별분석은 많은 방법들이 제안되어 왔다(Moore 1973, Hand 1981). 가장 단순한 방법은 n개의 변수를 갖는 표본들을 $2^n - 1$ 모수를 갖는 다항분포(multinomial distribution)로 간주하는 것이다. 그러나 이 방법은 이해하기는 쉽지만 너무 많은 모수를 예측하여야 하고, 따라서 표본의 수가 많이 필요한 단점이 있다. 이밖에도 급수를 이용하여 각 그룹의 조건부 확률을 구하거나, 베이지안 방법을 이용하여 판별하는 방법이 있으나 자료에 따라 장단점이 있기 때문에 어느 방법이 좋다고 단언할 수가 없다. 하지만 논리적 자료분석은 우리의 직관에 부합되는 합리적인 새로운 판별분석이라 생각할 수 있어 정보검색 자료를 이용하여 실험을 하여 보고자 한다.

* 이 연구는 99년도 숭실대 교내학술연구비의 일부 지원을 받았음.

1) (156-743) 서울시 동작구 상도동 1-1, 숭실대학교 정보통계학과 교수
E-mail: yolee@stat.soongsil.ac.kr
2) (156-743) 서울시 동작구 상도동 1-1, 숭실대학교 정보통계학과 교수
E-mail: jjlee@stat.soongsil.ac.kr

적절성 판별을 위한 단어를 구별하기 위해서는 훈련 자료 중 적절하다고 판단된 자료(편의상 양의 자료라 부름)와 부적절하다고 판단된 자료(편의상 음의 자료라 부름)를 Church(1995)의 비포와송성(non-Poissonnicity) 측도를 이용하여 분석하였다. 데이타베이스의 문서들에 나타나는 모든 단어들의 스템(stem)에 대하여 엠지퀘리(mgquery) 검색시스템(Witten, 1994)을 이용하여 통계량을 구한 후, 양의 자료와 음의 자료 중 각각 처치의 측도가 큰 상위 25개 단어를 판별을 위한 단어로 선택하였다. 이렇게 선택된 단어를 이용하여 각 문서를 불린(boolean)벡터로 만들어서 양의 자료와 음의 자료를 구별하는 논리적 패턴을 LAD 프로그램을 이용하여 찾고, 각 패턴들이 발생하는 확률을 구하였다. 양의 자료와 음의 자료를 확률적으로 잘 구분할 수 있는 패턴들의 적절한 선형결합으로 두 자료를 판별할 수 있는 판별식을 생성하였다.

2장에서는 논리적 자료분석 모형에 대한 소개를 하고, 3장에서는 논리적 자료분석을 이용한 실험에 대한 설명과 그 결과를 보여 주고, 4장에서 결론 및 토의를 한다.

2. 논리적 자료분석

2.1. 패턴의 정의

두개의 값(binary value)을 갖는 n개의 변수에 대한 관찰 값을 n차원 벡터로 표시하자. 그리고 각각의 관찰 값은 두 그룹, '적절하다고 판단되는 자료들의 집합'(편의상 '양'의 집합이라 하자)과 '부적절하다고 판단되는 자료들의 집합'(편의상 '음'의 집합이라 하자)중의 하나에 속한다고 하자. 모든 관찰 값들의 집합을 S라 하고, 이중에서 '양'의 집합과 '음'의 집합을 S^+와 S^- 로 표시하자. $\{S^+, S^-\}$ 는 다음과 같은 일종의 부분불린함수(partially defined Boolean function) ϕ로 볼 수 있다. 즉,

$$\phi : S \to \{0,1\},$$

여기서 S는 $\{0,1\}^n$의 부분집합이다. 의 모든 원소를 포함하는 전체불린함수(completely defined Boolean function), $\{0,1\}^n \to \{0,1\}$, 중에서 집합 $\{S^+, S^-\}$와 분류가 일치되는 함수를 ϕ의 확장(extension)이라 부른다. 집합 $\{S^+, S^-\}$에 정의된 부분불린함수 ϕ에 대한 확장불린함수 중에서 두 가지 극단적인 것에는 다음과 같은 것이 있다.

$$\phi^+(p) = \begin{cases} 1 & \text{if } p \notin S^- \\ 0 & \text{if } p \in S^- \end{cases}$$

$$\phi^-(p) = \begin{cases} 1 & \text{if } p \notin S^+ \\ 0 & \text{if } p \in S^+ \end{cases}$$

ϕ 에 근거한 모든 확장불린함수는 다음의 식을 만족한다.

$$\phi^- \geq \phi \geq \phi^+$$

따라서 지식베이스에 있는 자료에 대한 정보를 이용하여 새로운 자료에 대한 적절성여부를 검색하고자 하는 문제는, 어떻게 합당한 기준을 가지고 수많은 확장불린함수 중의 하나를 찾느냐 하는 문제로 요약될 수 있다.

집합 S^+에서 적어도 한 점을 포함하고 S^-에는 전혀 나타나지 않는 ϕ^+를 양의 패턴(positive pattern)이라 부르고, 그 반대로 집합 S^-에서 적어도 한 점을 포함하나 S^+에는 전혀 나타나지 않는 ϕ^- 를 음의 패턴(negative pattern)이라 부른다. 논리적 패턴을 이용한 자료검색은 적절한 자료와 부적절한 자료를 구분할 수 있는 직관적으로 설명이 가능한 여러 가지 형태의 패턴을 이용하여 합리적인 기준의 확장불린함수를 찾는 것이다.

2.2. 패턴의 생성

패턴을 생성하는 가장 직관적인 방법은 모든 변수에 대한 조합을 다 조사하는 것이다. 하지만 이러한 방법은 차수가 k인 패턴을 찾기 위해 $O(n^k)$ 만큼의 반복검색이 필요하다. 효율적인 검색을 위한 패턴의 생성방법은 크게 하향식(top-down approach)과 상향식(bottom up approach)으로 나눌 수 있다.

하향식 생성방법은 먼저 양(음)의 관찰값 모두가 양(음)의 패턴이라는 사실에서 출발하여 변수 하나씩 제거하여 가면서, 제거하고 남은 나머지 변수들의 관찰값이 아직도 패턴인가를 조사하는 방법이다. 그리하여 변수가 하나가 될 때 멈춘다.

상향식 생성방법은 반대로 먼저 양(음)의 관찰값에서 변수 하나씩 양(음)의 패턴인지 아닌지 조사하여 점차로 하나씩 차수를 증가하면서 모든 변수가 포함될 때 멈춘다.

2.3. 판별식

적절한 자료에 대한 패턴을 양의 패턴(P_k)이라 하고 부적절한 자료에 대한 패턴을 음의 패턴(N_l)이라 하며 새로운 자료의 적절성을 판별하여 주는 식은 이들 패턴의 선형 결합이라 할 수 있다.

$$\sum_{k=1} W_k^+ P_k + \sum_{l=1} W_l^- N_l$$

여기서 W_k^+와 W_l^-은 각각 양의 패턴과 음의 패턴에 대한 가중치로서 일반적으로 각 패턴에 대한 확률에 의해 주어질 수 있다.

3. 논리적 패턴을 이용한 검색 실험

3.1. 실험 자료

논리적 패턴을 이용한 검색 실험을 위해 TREC6 데이타베이스를 이용하였다. 여기에는 50개의 주제가 있고 주어진 주제에 대한 '적절' '부적절'성 여부가 판별된 문서들이 있다. 예를 들어 표 3.1은 주제 번호 1에 대한 적절성 판정에 대한 설명이고 이 실험에 이용된 문서는 월스트리트져널, 뉴욕타임지, 와싱톤포스트지 등 신문기사 자료이다.

표 3.1: 주제 번호 1에 대한 적절성 판정 설명의 예

<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)

3.2. 실험 자료의 트레이닝

앞 절에서 소개된 자료에 대하여 논리적 패턴을 이용한 검색을 실험하기 위하여 다음과 같은 트레이닝을 실시하였다.

1) 각각의 질문 주제에 대하여 적절하다고 판정된 트레이닝 자료를 엠지쿼리 (mgquery) 검색시스템을 이용하여 모든 단어-스템에 대한 인덱싱(indexing)을 하고 전체 단어들의 도수와, 각 문서에서의 단어들의 도수 등 통계량을 계산하였다.

2) 단어들의 도수를 이용하여 처치의 측도가 큰 상위 25개의 판별에 의미 있는 단어를 선택하였다.

3) 질문 주제에 대하여 부적절하다고 판정된 자료에 대해서도 위와 비슷한 방법으로 25개의 판별을 위한 단어를 선택하였다. 표 3.2는 주제 번호 53에 대해 의미 있다고 선택된 50개의 단어의 리스트이다.

4) 적절하다고 판정된 자료(190개)에서 추출된 25개의 단어와, 비적절하다고 판정된 자료(274개)에서 추출된 25개의 단어를 합한 50개의 단어에 대한 존재여부를 엠지쿼리 검색시스템을 이용하여 모든 문서에 대해 실시하였다. 이 결과를 이용하여 모든 트레이닝 자료의 각 문서를 0과 1로 이루어진 불린벡터로 표시하였다. 표 3.3은 주제 번호 53의 문서에 대해 불린벡터로 표시된 자료의 예이다.

5) 불린 벡터를 이용하여 각각의 질문에 대하여 LAD 프로그램을 이용하여 적절하다고 판정된 자료에는 있고 부적절하다고 판정된 자료에는 없는 양의 패턴(pattern) P_k와, 그

표 3.2: 주제 번호 53에 대해 의미 있다고 판정된 50개의 스템된 단어

abov	descript	identif	offer	shar
acquir	docum	intern	pens	stag
asses	dol	led	pric	sum
borrow	domain	leverag	priv	tak
buyout	econom	loan	propos	takeover
cash	expres	manag	relev	targes
cit	fact	ment	repay	tipster
compan	firm	million	sal	top
concept	flow	nar	secur	valu
definit	fund	number	serv	wil

표 3.3: 주제 번호 53의 문서에 대해 불린벡터로 표시된 자료의 예

적절하다고 판정된 자료
1 10110001111101000000000110000110101001001000100100110000 1
2 10110001111101000000000110000010101001001000110100010000 1
3 10110101101101010000000000000001010101100100010010011001 10
4 10100001001101000000100100010011001001000000010100100010 1
5 10100001111101010000000111000010001001001000100000000000 1
⋮
190 10110101001001000000000100000010000000000000010010001001 01
no =190 nv =56
부적절하다고 판정된 자료
1 00110001011101000010000110000010000001000000110100010000 1
2 00110000001001000000000001000011001001000000000100010001 1
3 00110001101001000000100100000010001001000000000100110000 1
4 00000001101011000000000100000110011000000000000000000100 01
5 00110001101011000010101000000010001010000000010100110001 1
⋮
274 00100000101001000000000101000010000010000000000100110011 0
no =274 nv =56

표 3.4: 주제 번호 53에 대해 LAD 프로그램으로 패턴을 찾은 예 (차수=3)

Perc-	sample-size		#pat-gene-		#pat-kept	
-Iter	difP -	difN	Pos	Neg	Pos	Neg
50%001	95	136	1070	3913	29	24
50%002	94	136	880	3080	31	23
50%003	95	137	1269	3528	30	22
50%004	95	136	1096	3046	27	23
50%005	95	136	1358	3555	37	24
50%006	95	137	1366	3468	32	25
50%007	95	136	1145	2800	30	25
50%008	95	137	1139	3224	31	24
50%009	95	135	1178	3549	32	24
50%010	95	135	1211	3028	29	23
50%011	95	137	1242	3315	30	23
50%012	95	136	1042	3004	34	23
50%013	95	136	812	3232	31	19
50%014	95	136	1366	3344	30	25
50%015	95	137	1096	3548	31	26
50%016	95	137	1056	3198	32	22
50%017	94	137	843	3079	30	21
50%018	94	137	871	8416	31	24
50%019	95	137	1075	3500	29	23
50%020	95	136	914	2854	34	26
mean	94.8	136.3	1101.5	3284.1	31.0	23.7
std	0.4	0.7	173.2	279.9	2.2	1.7
min	94	135	812	2800	27	19
max	95	137	1366	3913	37	26

반대로 부적절하다고 판정된 자료에는 있고 적절하다고 판정된 자료에는 없는 음의 패턴 N_l를 찾는다. 차수가 m인 패턴을 찾기 위한 컴퓨터의 속도는 $O(n^m)$이므로 실험컴퓨터(워크스테이션급)의 한계상 차수가 5이하인 패턴으로 제한하였다. 표 3.4는 주제번호 53에 대하여 차수가 3인 패턴을 찾은 결과를 보여주는 것으로 표 3.3의 자료에서 50%를 20번 단순확률 추출하였을 때 최종 생성된 양과 음의 패턴수(#pat-kept)를 보여 준다. 각각의 표본에서 평균적으로 31개와 23.7개의 양과 음의 패턴이 생성되었다.

6) 트레이닝 자료중에 나타나는 패턴의 빈도수를 각 패턴의 가중치로 하는 선형판별식으로 실험자료의 적절성 여부를 판별한다.

3.3. 검색모형의 실험결과

TREC6 실험자료의 50%자료를 이용해 검색모형의 판별식을 만들고, 나머지 50%의 자료에 대해 적절성 여부를 3차, 4차, 5차의 패턴을 이용하여 20회 시뮬레이션 실험하였다. 표 3.5, 표 3.6, 표 3.7는 주제 53에 대한 실험결과이다. 차수가 3일 경우에는 전반적으로 19%내

표 3.5: 주제 번호 53에 대해 차수 3인 LAD 프로그램으로 실험한 결과

Perc-	train-size-		-test-size		#patterns		%err-on-training-			-%err-on-testing-undecidible				
-Iter	difP-	difN-	-difP	-difN	-Pos	-Neg-	total-	-pos-	-neg-	-total-	-pos-	-neg-	-train-	test
50%001	95	136	94	137	29	24	0	0	0	31.5	20.3	11.2	9.9	15.1
50%002	94	136	95	137	31	23	0	0	0	27.6	14.2	13.4	12.1	15.5
50%003	95	137	95	136	30	22	0	0	0	24.1	14.7	9.5	9.1	15.9
50%004	95	136	94	137	27	23	0	0	0	28.4	15.1	13.4	9.1	20.3
50%005	95	136	95	137	37	24	0	0	0	27.2	16.4	10.8	9.5	14.2
50%006	95	137	95	136	32	25	0	0	0	29.7	16.4	13.4	9.5	16.8
50%007	95	136	94	137	30	25	0	0	0	27.6	15.5	12.1	11.2	19.0
50%008	95	137	94	135	31	24	0	0	0	25.0	11.6	13.4	11.2.	20.7
50%009	95	135	95	137	32	24	0	0	0	26.7	12.9	13.8	7.8	22.8
50%010	95	135	94	137	29	23	0	0	0	29.7	15.5	14.2	12.1	23.7
50%011	95	137	95	136	30	23	0	0	0	30.6	12.9	17.7	9.5	19.0
50%012	95	136	94	137	34	23	0	0	0	32.3	18.5	13.8	9.9	19.4
50%013	95	136	95	137	31	19	0	0	0	26.7	14.2	12.5	8.6	18.1
50%014	95	136	95	137	30	25	0	0	0	25.9	14.2	11.6	9.9	15.1
50%015	95	137	95	136	31	26	0	0	0	28.9	14.2	14.7	10.3	14.2
50%016	95	137	95	135	32	22	0	0	0	28.4	12.9	15.5	6.5	22.0
50%017	94	137	95	136	30	21	0	0	0	27.2	12.9	14.2	10.8	20.7
50%018	94	137	95	136	31	24	0	0	0	26.7	15.1	11.6	11.2	22.0
50%019	95	137	94	135	29	23	0	0	0	24.6	13.8	10.8	11.2	19.4
50%020	95	136	94	137	34	26	0	0	0	28.4	16.4	12.1	11.6	19.8
mean	94.8	136.3	94.6	136.4	31.0	23.4	0.0	0.0	0.0	27.9	14.9	13.0	10.0	18.7
std	0.4	0.7	0.5	0.8	2.2	1.7	0.0	0.0	0.0	2.2	2.0	1.9	1.4	2.9
min	94	135	94	135	27	19	0	0	0	24.1	11.6	9.5	6.5	14.2
Max	95	137	95	137	37	26	0	0	0	32.3	20.3	17.7	12.1	23.7

표 3.6: 주제 번호 53에 대해 차수 4인 LAD 프로그램으로 실험한 결과

Perc-	train-size-		-test-size		#patterns		%err-on-training-			-%err-on-testing-undecidible				
-Iter	difP-	difN-	-difP	-difN	-Pos	-Neg-	total-	-pos-	-neg-	-total-	-pos-	-neg-	-train-	test
50%001	95	136	94	137	23	19	0	0	0	26.3	15.9	10.3	0.4	16.8
50%002	94	136	95	137	22	20	0	0	0	29.3	11.2	18.1	0.4	11.6
50%003	95	137	95	136	22	21	0	0	0	26.3	12.5	13.8	0	13.8
50%004	95	136	94	137	19	20	0	0	0	30.2	16.4	13.8	0.4	15.1
50%005	95	136	95	137	23	20	0	0	0	27.6	16.8	10.8	0	15.1
50%006	95	137	95	136	24	21	0	0	0	30.2	13.4	16.8	0.4	11.2
50%007	95	136	94	137	25	19	0	0	0	24.6	11.2	13.4	0	13.8
50%008	95	137	94	135	23	17	0	0	0	28.4	14.7	13.8	0	15.1
50%009	95	135	95	137	24	19	0	0	0	34.5	18.5	15.9	0	7.8
50%010	95	135	94	137	25	20	0	0	0	26.4	14.2	12.1	0	11.6
50%011	95	137	95	136	21	17	0	0	0	30.6	12.9	17.7	0.4	12.1
50%012	95	136	94	137	20	19	0	0	0	31.0	16.8	14.2	0	9.5
50%013	95	136	95	137	21	18	0	0	0	27.6	12.9	14.7	0.4	10.8
50%014	95	136	95	137	25	21	0	0	0	28.4	11.2	17.2	0.4	11.2
50%015	95	137	95	136	22	19	0	0	0	30.6	12.5	18.1	0	12.1
50%016	95	137	95	135	17	18	0	0	0	28.9	12.9	15.9	0	15.5
50%017	94	137	95	136	24	19	0	0	0	29.7	10.3	19.4	0	8.6
50%018	94	137	95	136	23	18	0	0	0	29.3	12.9	16.4	0	12.9
50%019	95	137	94	135	23	22	0	0	0	25.9	14.7	11.2	0.4	9.9
50%020	95	136	94	137	21	19	0	0	0	34.1	15.5	18.5	0	19.8
mean	94.8	136.3	94.6	136.4	22.4	19.3	0.0	0.0	0.0	29.0	13.9	15.1	0.2	12.0
std	0.4	0.7	0.5	0.8	2.1	1.3	0.0	0.0	0.0	2.6	2.2	2.7	0.2	2.8
min	94	135	94	135	17	17	0	0	0	24.6	10.3	10.3	0	6.5
Max	95	137	95	137	25	22	0	0	0	34.5	18.5	19.4	0.4	16.8

표 3.7: 주제 번호 53에 대해 차수 5인 LAD 프로그램으로 실험한 결과

Perc-	train-size-		-test-size		#patterns		%err-on-training-			-%err-on-testing-undecidible				
-Iter	diffP-	difN-	-diffP	-difN	-Pos	-Neg-	total-	-pos-	-neg-	-total-	-pos-	-neg-	-train-	test
50%001	95	136	94	137	19	21	0	0	0	28.9	18.5	10.3	0	7.3
50%002	94	136	95	137	15	18	0	0	0	25.4	11.6	13.8	0	13.4
50%003	95	137	95	136	19	20	0	0	0	30.6	13.4	17.2	0	14.2
50%004	95	136	94	137	16	20	0	0	0	26.7	12.1	14.7	0	19.8
50%005	95	136	95	137	20	21	0	0	0	30.2	15.9	14.2	0	9.5
50%006	95	137	95	136	18	18	0	0	0	28.4	12.9	15.5	0	11.2
50%007	95	136	94	137	18	19	0	0	0	31.5	11.2	20.3	0	11.6
50%008	95	137	94	135	19	17	0	0	0	27.6	11.2	16.4	0	12.1
50%009	95	135	95	137	16	19	0	0	0	35.3	20.3	15.1	0	8.6
50%010	95	135	94	137	18	18	0	0	0	28.0	14.2	13.8	0	7.8
50%011	95	137	95	136	20	17	0	0	0	30.2	13.4	16.8	0	15.1
50%012	95	136	94	137	17	18	0	0	0	31.5	15.9	15.5	0	8.2
50%013	95	136	95	137	17	18	0	0	0	31.0	15.9	15.1	0	6.5
50%014	95	136	95	137	21	24	0	0	0	29.3	13.8	15.5	0	8.6
50%015	95	137	95	136	22	19	0	0	0	30.6	12.5	18.1	0	12.1
50%016	95	137	95	135	19	19	0	0	0	28.9	11.6	17.2	0	16.8
50%017	94	137	95	136	22	18	0	0	0	25.9	10.3	15.5	0	18.1
50%018	94	137	95	136	18	19	0	0	0	30.2	14.7	15.5	0	10.8
50%019	95	137	94	135	20	21	0	0	0	24.1	12.5	11.6	0	8.6
50%020	95	136	94	137	18	20	0	0	0	33.2	14.2	19.0	0	8.6
mean	94.8	136.3	94.6	136.4	18.6	19.2	0.0	0.0	0.0	29.4	13.8	15.6	0.0	11.4
std	0.4	0.7	0.5	0.8	1.9	1.7	0.0	0.0	0.0	2.7	2.5	2.3	0.0	3.8
min	94	135	94	135	15	17	0	0	0	24.1	10.3	10.3	0	6.5
Max	95	137	95	137	22	24	0	0	0	34.5	20.3	20.3	0.0	19.8

외의 판별불능률을 보였으며 판별가능한 문서 중에서는 72%내외의 적중률을 보여 주고 있다. 차수가 4일 경우에는 12% 내외의 판별 불능률을 보였으나 판별가능한 문서 중에서는 위와 별 차이없이 71%내외의 적중률을 보여 주고 있다. 차수가 5일 경우는 차수가 4일 경우와 판별 불능률이나 적중률에 별 차이가 없다. 즉 차수의 증가가 반드시 판별불능률을 줄이지는 않는다는 것을 보여 주고 있다.

지면관계상 본 논문에 보여주지 못하지만 다른 주제에 대한 실험에서도 판별 적중률은 비슷한 결과를 보여 주고 있다. 이러한 높은 적중률은 이 모형이 실제 검색에 이용될 수 있는 가능성을 보여 주고 있다. 하지만 판별불능률이 10%에서 20%에 이르는 것은 이모형이 해결하여야 할 과제라고 생각된다.

4. 결론 및 토의

본 연구는 일반인들이 정보검색시 선호하는 불린 형태의 자연적 정보검색 방법을 근간으로 하여 지식베이스를 이용한 확률화 정보검색을 하는 시도이다. 실험자료는 어느 정도 만족스러운 결과를 보이고 있으나, 최종적인 결론은 실제 시스템에서 본격적인 검증을 한 후에야 내릴 수 있을 것이다. 본 연구는 여러 가지 방법으로 확장을 생각할 수 있는데, 예를 들어 불린으로 표시된 단어중 관련성 여부를 판별식에 포함시키든가, 고차수의 패턴을 조사한다든가, 패턴을 결합시켜 판별식을 만드는 새로운 방법들을 생각할 수 있다. 구체적으로,

1) 자료의 불린화에 이용된 스템은 분포함수에 기준을 두어 선택된 것이다. 이 스템들

은 로그 오드스 비율(log odds ratio)이나, 카이제곱 측도를 이용하여 선택될 수 있다.

2) 패턴의 차수를 5이하로 제한한 것은 컴퓨터의 속도능력 때문이었다. 빠른 컴퓨터를 이용하면 더 고차수의 패턴에 대한 실험을 계획할 수 있다.

3) 양의 패턴이나 음의 패턴에 대한 정의를 엄격히 하기 보다는 퍼지(fuzzy)화 하여 더 일반화 할 수 있다. 하지만 이런 경우에 패턴을 찾는 데 걸리는 시간은 더욱 소요된다.

4) 모든 불린변수를 같은 비중으로 취급하지 않고 의미있는 단어에 가중을 두는 방법도 고려할 수 있다.

참고문헌

[1] Crama, Y., Hammer, P. and Ibaraki, T. (1988). Cause-Effect Relationships and Partially Defined Boolean Functions. *Annals of Operations Research*, 16, 299-326.

[2] Boros, E., Hammer, P.L. and Hooker, J.N. (1994). Predicting Cause-Effect Relationship from Incomplete Discrete Observations. *SIAM J. of Discrete Mathematics*, 7, No. 4, 531-543.

[3] Church K.F. and Gale W.A. (1995). Inverse Document Frequency (IDF): A measure of deviation from Poisson. *Proceedings of the Third Workshop on Very Large Corpora.*

[4] Witten I.H., Moffat A. and Bell T.C. (1994). <Managing Gigabytes>. New York. Van Nostrand Reinhold.

[5] Moore, D.H. (1973). Evaluation of Five Discrimination Procedures for Binary Variables. Journal of American Statistical Association, 68, No. 342, 399-404.

[6] Hand, D.J. (1981). Statistical Pattern Recognition of Binary Variables. *Pattern Recognition Theory and Application*, 19-33. J. Kittler, K.S. Fu, and L.F. Pau (eds).

[1999년 11월 접수, 2000년 2월 채택]

A Study of Probabilistic Information Retrieval System Using Logical Pattern

Yoon Oh Lee[1] Jung Jin Lee[2]

ABSTRACT

In the era of information society, an efficient information retrieval will play very important role for our decision making. Retrieved data from a database can be judged either relevant or non-relevant to a given problem. This judgment can be added to the database for building a knowledge base. Logical analysis of the knowledge database enables to judge relevancy of new data. This paper describes a probabilistic information retrieval system based on Logical Analysis of Data (LAD). Combinations of such patterns are used for developing general classification procedures. An experiment by using TREC6 data is discussed.

Keywords: Information Retrieval; Logical Pattern; Knowledge Base.

1) Professor, Dept. of Statistics, Soong Sil University. E-mail: yolee@stat.soongsil.ac.kr
2) Professor, Dept. of Statistics, Soong Sil University. E-mail: jjlee@stat.soongsil.ac.kr