연관성 척도의 빈도수준 선호경향에 대한 연구

A Study on the Frequency Level Preference Tendency of Association Measures

이 재 윤(Jae-Yun Lee)*

초 록

연관성 척도는 정보검색 및 데이터마이닝을 비롯한 다양한 분야에서 사용되고 있다. 각 연관성 척도가 높거나 낮은 빈도 중에서 어떤 쪽을 선호하는가를 나타내는 빈도수준 선호경향은 척도의 적용 결과에 중요한 영향을 미치므로 이에 대한 면밀한 조사가 필요하다. 이 연구에서는 주요 연관성 척도들의 빈도수준 선호경향을 가상의 데이터를 통해 분석하고 그 결과를 제시하였다. 또한 코사인 계수를 비롯한 대표적인 연관성 척도에 대해서 빈도수준 선호경향을 조절할 수 있는 방법을 제안하였다. 이 조절 방법을 동시출현 기반 질의확장 정보검색에 적용해본 결과 그 유용성이 확인되었다. 마지막으로 분석 및 실험 결과가 관련 분야에 시사하는 바를 논하였다.

ABSTRACT

Association measures are applied to various applications, including information retrieval and data mining. Each association measure is subject to a close examination to its tendency to prefer high or low frequency level because it has a significant impact on the performance of applications. This paper examines the frequency level preference(FLP) tendency of some popular association measures using artificially generated cooccurrence data, and evaluates the results. After that, a method of how to adjust the FLP tendency of major association measures such as cosine coefficient is proposed. This method is tested on the cooccurrence-based query expansion in information retrieval and the result can be regarded as promising the usefulness of the method. Based on these results of analysis and experiment, implications for related disciplines are identified.

키워드: 연관성척도, 동시출현분석, 정보검색, 데이터마이닝
Association Measures, Cooccurrence Analysis, Information Retrieval, Data Mining

* 경기대학교 문헌정보학과 교수(memexlee@kyonggi.ac.kr)
■ 논문접수일자 : 2004년 11월 19일
■ 게재확정일자 : 2004년 12월 13일

1. 서 론

통계적 분석을 위한 연관성 척도는 정보검색이나 데이터마이닝 분야에서 널리 활용되어 왔다. 연관성 척도의 특성은 연관성 분석 대상이 고빈도이냐 저빈도이냐에 따라 상당히 좌우되며, 연관성 척도를 이용한 실험의 성능도 이런 빈도수준에 영향을 받게 된다. 예를 들어 주어진 용어와의 관련어를 용어간의 동시출현 빈도를 근거로 통계적으로 파악하려는 경우에, 자카드 계수를 연관성 척도로 사용하면 주로 고빈도어가 관련어 순위 상위에 나타난다. 반면에 상호정보량을 연관성 척도로 사용하면 극단적으로 빈도가 낮은 용어를 관련어로 판정하게 된다. 이와 같이 주로 분석 대상이 고빈도일 때 높은 연관도를 가지게 되는 척도는 고빈도 선호 연관성 척도, 그 반대의 경우는 저빈도 선호 연관성 척도라고 할 수 있으며, 연관성 척도의 이런 특성을 빈도수준 선호경향으로 부르기로 한다.

정보검색에 대한 연구가 비교적 제한되어 있던 1970~80년대에는 연관성 척도의 차이가 검색 성능에 미치는 영향은 크지 않다고 보았다(정영미 1983, 182; van Rijsbergen 1979, 38). 그러나 연관성 척도의 적용 영역이 넓어지고 검토 대상이 되는 척도의 종류도 다양해지면서 각 영역마다 다양한 연관성 척도의 적용결과를 비교하는 시도가 점차 증가하였다. 자동 분류나 클러스터링 분야에서는 수십 종의 연관성 척도를 망라하여 적용해보는 시도도 있었다(Forman 2003; Meyer et al. 2004).

연관성 척도의 적용분야 중에서 질의확장 검색의 경우에는 자카드 계수와 같은 고빈도 선호 연관성 척도에 비해서 상호정보량과 같은 저빈도 선호 연관성 척도가 좋은 성능을 보이는 것으로 나타났다(Chung & Lee 2004). 이에 비해 자동분류를 위한 자질선정에 있어서는 저빈도 수준을 선호하는 상호정보량을 이용하는 것이 좋지 못한 결과를 가져온다고 알려져 있다(Yang & Pederson 1997).

이와 같이 빈도수준 선호경향이 연관성 척도의 적용 결과에 큰 영향을 미침에도 불구하고, 여러 연관성 척도의 빈도수준 선호경향을 객관적으로 판단할 수 있는 기준은 제시된 바가 없다.

각 연관성 척도의 빈도수준 선호경향이 어떤지를 미리 파악할 수 있다면, 연관성 척도를 사용하는 연구에서 불필요하게 많은 종류의 척도를 적용해보는 시행착오 및 시간낭비를 줄일 수 있을 것이다.

이 연구에서는 가상의 데이터를 대상으로 분석하여 주요 연관성 척도의 빈도수준 선호경향을 알아보았다. 그리고 이 분석 결과에 근거하여 코사인 계수를 비롯한 대표적인 연관성 척도의 빈도수준 선호경향을 조절할 수 있는 방안을 제안하였다.

2. 연관성 척도의 빈도수준 선호경향 분석

2.1 연관성 척도

연관성 척도의 종류는 수십 종 이상인 것으로 알려져 있다(사공철 외 2003). 대표적인 통계처리 프로그램인 *SPSS for Windows*에서는

현재 이진 값에 대한 연관성 척도로 26종을 사용할 수 있도록 되어 있다. 그러나 이들 중에서 실제로 흔히 사용되는 척도는 일부에 불과하다. 이 논문에서 빈도수준 선호경향을 분석할 연관성 척도로는 통계학 분야의 권위있는 백과사전인 *Encyclopedia of Statistical Sciences*에서 Gower(1985)가 소개한 여러 척도 이외에 정보학 분야에 응용되어 좋은 성능을 보이는 것으로 알려진 상호정보량, 상대적 상호정보량, GSS계수, 카이제곱 계수, 로그승산비를 선정하였다. 상호정보량과 카이제곱계수는 Yang and Pederson(1997)이, 로그승산비는 Mladeni'c(1998)이, GSS 계수는 Galavotti 등(2000)이 자동분류를 위한 자질선정 기준으로 사용해서 좋은 성능을 얻었으며, 상대적 상호정보량은 이재윤(2003)이 제안한 것으로 질의확장 정보검색에서 좋은 성능을 보이는 것으로 나타났다.

분석대상 연관성 척도 15종을 2×2 분할표에 적용한 형태로 나타낸 공식은 〈표 1〉과 같다. 이 표에서 *a*, *b*, *c*, *d*는 2×2 분할표의 네 항으로서 각각은 동시출현빈도, 첫 번째 대상의 빈도, 두 번째 대상의 빈도, 동시미출현빈도(co-absence frequency)를 뜻한다. 그리고 *N*은 전체 관찰빈도이다. 각 척도의 명칭은 일반적인 것을 채택하고 다른 명칭은 괄호안에 표기하였다. 단, 코사인 계수는 2×2 분할표에서 연관성을 측정하는 경우에 별도로 Ochiai 계수라는 이름으로도 불리지만 친숙한 코사인 계수라는 명칭을 채택하였다. 이후 본문에서는 꼭 필요한 경우가 아니면 〈표 1〉에 제시한 약호를 사용하여 각 척도를 표기하였다.

2.2 빈도수준 선호경향 분석 방법

연관성 척도의 빈도수준 선호경향을 분석하기 위해서는 Chung and Lee(2001)에서처럼 실제 데이터를 대상으로 연관성 척도를 적용하는 것도 한 가지 방법이다. 그러나 이런 실험 결과는 어디까지나 해당 실험 데이터에 의존해서 해석할 수밖에 없으므로, 분석 대상에 따라서 다른 결과를 얻게 될 가능성이 있다. 따라서 이 연구에서는 특정한 실험집합을 사용하지 않고, 이론적으로 가능한 경우의 수를 모두 포함하도록 연관성 측정 대상들의 출현빈도 조합을 가상으로 만들어 분석하였다.

분석 대상의 가능한 최고 출현빈도를 N이라고 한다면 통계적인 연관성을 판단하는 두 대상의 출현빈도 조합은, 둘 다 1인 경우(1,1)에서부터 둘 다 N인 경우(N,N)에 이르기까지 총 N(N-1)/2 가지가 있게 된다. 분석 대상 x와 y가 있을 때, 각각의 빈도 f(x)와 f(y)를 1에서 최고 N까지 변화시키되 f(x)가 f(y)보다 작거나 같도록 하면 다음과 같은 조합을 얻을 수 있다.

(1,1)
(1,2) (2,2)
⋮ ⋮ …
(1,N) (2,N) … (N,N)

어느 한 조합쌍 (f(x),f(y))에 대해서 x와 y의 동시출현빈도 f(x,y)는 둘 중에 빈도가 낮은 f(x)보다 작거나 같다. 예를 들어 출현빈도의 조합이 (1,100)이면 동시출현빈도는 1인 경우밖에 없으며, (20,30)이면 동시출현빈도는 1에서부터 20까지 나타날 수 있다. 엄밀히 말하면 빈도가 0인 경우도 포함해야 하지만 대

〈표 1〉 분석 대상 연관성 척도

명 칭	약 호	공 식
러셀과 라오 Russel & Rao	R&R	$\frac{a}{N}$
자카드 Jaccard	JAC	$\frac{a}{a+b+c}$
다이스 Dice	DIC	$\frac{2a}{2a+b+c}$
코사인 Cosine (Ochiai)	COS	$\frac{a}{\sqrt{(a+b)(a+c)}}$
쿨친스키 2 Kulczynski 2	KUL2	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
단순 일치 Simple Matching	SM	$\frac{a+d}{N}$
카이제곱 Chi-square	CHI	$\frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$
GSS GSS (Dispersion)	GSS	$\frac{ad-bc}{N^2}$
피어슨의 파이 Pearson's PHI	PHI	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
소칼과 스니스 4 Sokal & Sneath 4 (Anderberg)	SS4	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$
상대적 상호정보량 J Relative Mutual Information J	RMIJ	$\frac{log_2N+log_2a-log_2(a+b)(a+c)}{log_2N-log_2a}$
소칼과 스니스 5 Sokal & Sneath 5 (Ochiai 2)	SS5	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
로그 승산비 Log Odds Ratio	LOR	$\log\frac{ad}{bc}$
율의 Y Yule's Y	YULE	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
상호정보량 Mutual Information	MI	$\log_2\frac{Na}{(a+b)(a+c)}$

부분의 경우에 동시출현하지 않은 쌍은 분석에서 제외하므로 여기에서도 동시출현빈도 1 이상인 경우만 포함하였다.

이상과 같이 두 분석대상 각각의 출현빈도와 둘의 동시출현빈도가 가질 수 있는 경우를 모두 산출한 다음, 이들에 대해서 여러 연관성 척도를 적용하면 가능한 모든 경우를 고려한 비교분석이 될 수 있다.

그런데 실제 상황에서는 각각의 출현빈도가 최고빈도 N에 크게 못 미치는 경우도 많다. 예를 들어 문헌집단에서 색인어의 문헌빈도는 전체 문헌 수의 절반도 안되는 것이 일반적이며, 문헌분리가 이론(Salton & Yang 1973)에서는 색인어의 문헌빈도가 전체 문헌 수의 1/10 내지

1/100 정도 수준이 바람직하다고 제시되어 있다. 따라서 여러 상황을 고려하기 위해 최고빈도 N에 비해서 분석대상 각각의 출현빈도 상한이 100%인 경우에서부터 점차 상한을 낮추어서 83.3%, 50%, 25%, 12.5%, 6.25%인 경우까지 여섯 가지 상황을 설정하고 각각의 경우에 연관성 척도의 빈도수준 선호경향을 분석하였다.

각 연관성 척도가 분석대상의 출현빈도에 따라서 어떤 성향을 보이는가를 판단하는 기준으로는 분석대상의 출현빈도와 연관성 척도값 사이의 상관계수를 이용하였다. 빈도수준 선호경향은 출현빈도의 높고 낮음에 따라서 연관성 척도값이 어떻게 달라지느냐를 반영하는 성질이므로 상관계수로 측정하는 것이 바람직하다. 이 연구에서는 값의 상관을 측정하는 피어슨 적률상관계수(Pearson's r)와 순위 상관을 측정하는 스피어만 상관계수(Spearman's rho)를 모두 산출하였지만, 각 연관성 척도의 값의 범위가 크게 다른 경우도 있으므로 분석은 주로 스피어만 상관계수를 대상으로 하였다.

연관성 측정에서는 분석 대상이 두 개이므로 출현빈도도 둘인데, 연관성 척도와의 상관계수 산출을 위해서는 둘 중에서 빈도가 낮거나 같은 값을 채택하였다. 둘 중 낮은 빈도가 동시출현빈도까지 제한하는 역할을 하기 때문이다.

2. 3 빈도수준 선호경향 분석결과

분석대상 x와 y의 출현빈도 f(x), f(y)의 상한을 50으로 설정하였을 때 이 값이 최고빈도 N의 100%라면 최고빈도 N도 50이며, 출현빈도 상한의 비율이 83.3%, 50%, 25%, 12.5%, 6.25%이면 최고빈도 N은 60, 100, 200, 400, 800이 된다. 최고빈도가 800이면서 출현빈도 상한의 비율이 6.25%인 경우에는 저빈도인 경우만 포함되어 있을 경우를 분석하는 결과가 된다. 출현빈도 상한을 50으로 설정하고 최고빈도의 비율을 6가지 경우로 달리하여 가능한 모든 빈도 조합에 대해서 각 연관성 척도값을 계산하였다. 〈표 2〉에 출현빈도 상한의

〈표 2〉 출현빈도 상한의 비율이 6.25%인 경우의 연관성 척도값 (중간 생략)

f(x)	f(y)	f(x,y)	COS	JAC	DIC	KUL2	R&R	SM	SS4	SS5	CHI	GSS	PHI	YULE	RMIJ	MI	LOR
1	1	1	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	800.00	0.00	1.00	1.00	1.00	9.64	20.05
1	2	1	0.71	0.50	0.67	0.75	0.00	1.00	0.87	0.71	399.50	0.00	0.71	1.00	0.90	8.64	13.37
1	3	1	0.58	0.33	0.50	0.67	0.00	1.00	0.83	0.58	266.00	0.00	0.58	1.00	0.84	8.06	12.67
1	4	1	0.50	0.25	0.40	0.63	0.00	1.00	0.81	0.50	199.25	0.00	0.50	1.00	0.79	7.64	12.27
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	25	24	0.96	0.92	0.96	0.96	0.03	1.00	0.98	0.96	735.30	0.03	0.96	0.99	0.98	4.94	9.83
25	25	25	1.00	1.00	1.00	1.00	0.03	1.00	1.00	1.00	800.00	0.03	1.00	1.00	1.00	5.00	23.24
25	26	1	0.04	0.02	0.04	0.04	0.00	0.94	0.50	0.04	0.05	0.00	0.01	0.06	0.03	0.30	0.22
25	26	2	0.08	0.04	0.08	0.08	0.00	0.94	0.52	0.08	1.85	0.00	0.05	0.25	0.15	1.30	1.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	50	47	0.94	0.89	0.94	0.94	0.06	0.99	0.97	0.94	700.88	0.05	0.94	0.97	0.96	3.91	8.27
50	50	48	0.96	0.92	0.96	0.96	0.06	1.00	0.98	0.96	733.19	0.06	0.96	0.98	0.97	3.94	9.10
50	50	49	0.98	0.96	0.98	0.98	0.06	1.00	0.99	0.98	766.23	0.06	0.98	0.99	0.99	3.97	10.51
50	50	50	1.00	1.00	1.00	1.00	0.06	1.00	1.00	1.00	800.00	0.06	1.00	1.00	1.00	4.00	23.90

비율이 6.25%인 경우의 계산 결과를 일부 제시하였다.

각 연관성 척도값과 분석대상의 출현빈도 f(x)와의 피어슨 상관도를 산출하여 〈표 3〉에 스피어만 순위상관을 산출하여 〈표 4〉에 나타내었다.

〈표 3〉과 〈표 4〉에서는 빈도수준 상한을 최고빈도 N의 6.25%로 설정하였을 때의 스피어만 순위상관값이 높은 순서대로 연관성 척도를 배열하였다. 또한 각 연관성 척도가 빈도수준과 가지는 상관도를 빈도 상한 설정 별로 평균한 값과 그 표준편차도 제시하였다. 15종의 연관성 척도 중에서 평균 상관도가 양의 상관을 나타낸 것이 11종, 음의 상관을 나타낸 것이 4종이다. 그중에서도 상관도 평균이 가장 높은 것은 R&R이고 가장 낮은 것은 MI 이다. 즉, R&R이 고빈도수준 선호경향이 가장 강한 반면, MI는 저빈도수준 선호경향이 가장 강한 것이다. 한편 RMIJ는 상관도의 평균이 0에 가장 가까우면서 표준편차도 매우 낮게 나타났다. 이는 RMIJ가 분석대상의 빈도 수준에 거의 영향받지 않는 척도라는 것을 뜻한다.

각 척도의 빈도수준 선호경향을 시각적으로 살펴보기 위해서 분석대상 빈도와 연관성 척도값 간의 스피어만 순위상관값을 연관성 척도별로 〈그림 1〉에 나타내었다.

〈표 3〉 분석대상 빈도와 연관성 척도값 간의 피어슨 상관값

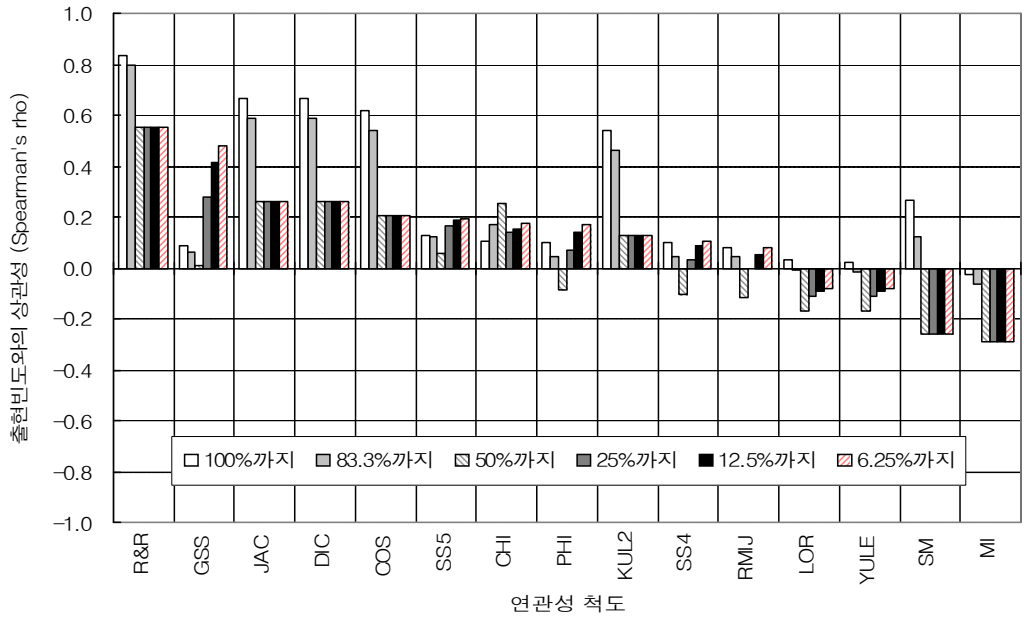
척도 / 빈도 상한	R&R	GSS	JAC	DIC	COS	SS5	CHI	PHI	KUL2	SS4	RMIJ	LOR	YULE	SM	MI
100%	0.871	0.100	0.687	0.680	0.634	0.170	0.105	0.127	0.549	0.125	0.098	0.041	0.020	0.302	-0.015
83.3%	0.829	0.080	0.602	0.604	0.556	0.169	0.158	0.069	0.474	0.063	0.064	-0.024	-0.013	0.144	-0.039
50%	0.577	-0.007	0.283	0.276	0.219	0.097	0.251	-0.107	0.137	-0.124	-0.120	-0.173	-0.190	-0.302	-0.287
25%	0.577	0.329	0.283	0.276	0.219	0.183	0.186	0.081	0.137	0.036	-0.024	-0.121	-0.134	-0.302	-0.287
12.5%	0.577	0.467	0.283	0.276	0.219	0.205	0.211	0.155	0.137	0.093	0.021	-0.112	-0.120	-0.302	-0.287
6.25%	0.577	0.525	0.283	0.276	0.219	0.213	0.228	0.188	0.137	0.116	0.047	-0.112	-0.117	-0.302	-0.287
평 균	0.668	0.249	0.404	0.398	0.344	0.173	0.190	0.086	0.262	0.051	0.014	-0.084	-0.092	-0.127	-0.201
표준편차	0.129	0.203	0.172	0.174	0.178	0.038	0.048	0.096	0.178	0.084	0.071	0.071	0.072	0.251	0.123

〈표 4〉 분석대상 빈도와 연관성 척도값 간의 스피어만 상관값

척도 / 빈도 상한	R&R	GSS	JAC	DIC	COS	SS5	CHI	PHI	KUL2	SS4	RMIJ	LOR	YULE	SM	MI
100%	0.830	0.088	0.664	0.664	0.619	0.128	0.105	0.100	0.542	0.098	0.080	0.035	0.023	0.263	-0.027
83.3%	0.795	0.063	0.590	0.590	0.542	0.122	0.168	0.047	0.464	0.042	0.042	-0.012	-0.012	0.121	-0.065
50%	0.550	0.007	0.261	0.261	0.204	0.057	0.253	-0.088	0.128	-0.103	-0.116	-0.168	-0.168	-0.262	-0.288
25%	0.550	0.278	0.261	0.261	0.204	0.162	0.141	0.071	0.128	0.032	-0.005	-0.108	-0.108	-0.262	-0.288
12.5%	0.550	0.414	0.261	0.261	0.204	0.187	0.154	0.140	0.128	0.084	0.049	-0.090	-0.090	-0.262	-0.288
6.25%	0.550	0.482	0.261	0.261	0.204	0.196	0.175	0.173	0.128	0.107	0.080	-0.083	-0.083	-0.262	-0.288
평 균	0.638	0.222	0.383	0.383	0.330	0.142	0.166	0.074	0.253	0.043	0.022	-0.071	-0.073	-0.111	-0.207
표준편차	0.124	0.181	0.174	0.174	0.179	0.047	0.045	0.083	0.178	0.071	0.068	0.066	0.063	0.218	0.115

〈그림 1〉을 보면 앞의 상관값 표준편차에서도 알 수 있듯이 RMIJ나 SS4처럼 순위상관값의 변화가 적은 척도도 있지만, GSS, JAC, SM 등의 척도는 변화가 심하다는 것이 드러난다. 연관성 척도별 순위상관값의 변화는 주로 출현빈도의 상한이 최고빈도의 50%인 경우를 중심으로 나타나고 있다. 이에 착안하여 출현빈도의 상한이 50% 초과인 경우와 50% 이하인 경우에 대해서 각 연관성 척도가 선호하는 빈도수준을 '고', '중', '저'의 3단계로 나누고, '중' 단계를 다시 '중상', '중', '중하'로 나누어 보면 〈표 5〉와 같이 연관성 척도를 7가지로 나눌 수 있다.

〈표 5〉에서 볼 수 있듯이 대부분의 척도는 선호하는 빈도수준이 최대빈도 상한에 크게 좌우되지 않고 고-고, 중상-중상, 중하-중하,



〈그림 1〉 분석대상 빈도와 연관성 척도값 간의 스피어만 순위상관값

〈표 5〉 빈도수준 선호경향에 따른 연관성 척도의 구분

구 분	해당 연관성 척도	빈도수준 선호경향
저-저	MI, YULE, LOR	거의 항상 저빈도 수준을 선호함
중-저	SM	중립에 가깝다가 빈도 상한이 50% 이하일 때 저빈도 수준을 선호함
중하-중하	PHI, SS4, RMIJ	중립이되 저빈도 선호에 가까움
중상-중상	SS5, CHI	중립이되 고빈도 선호에 가까움
중-고	GSS	중립에 가깝다가 빈도 상한이 50% 이하일 때 고빈도 수준을 선호함
고-중상	JAC, DIC, COS, KUL2	고빈도 수준을 선호하다가 빈도 상한이 50% 이하이면 다소 중립에 가까워짐
고-고	R&R	거의 항상 고빈도 수준을 선호함

저-저와 같이 일관되게 나타난다. 이와 달리 SM과 GSS는 빈도수준 선호경향이 특이한데, 빈도 상한이 50%를 넘을 때에는 둘 다 중간빈도 수준을 선호하다가 빈도 상한이 50% 이하이면 SM은 저빈도 선호경향을, GSS는 고빈도 선호경향을 나타낸다. 이 두 척도는 적용되는 상황에 따라서 빈도수준 선호경향이 달라질 수가 있으므로 사용할 때 주의가 필요하다.

3. 빈도수준 선호경향의 조절

3.1 빈도수준 선호경향의 조절방안

앞에서 살펴본 연관성 척도들 중에서 고빈도 선호경향을 가진 척도 R&R, JAC, DIC, COS, KUL2는 모두 공식의 분자에 동시출현빈도 a항만 있는 것을 볼 수 있다. 반면에 공식의 분자에 동시출현빈도와 함께 동시미출현빈도 d항이 있는 척도들은 상대적으로 선호하는 빈도수준이 낮게 나타났다. 결국 동시미출현빈도 d항의 채택여부가 선호하는 빈도수준에 영향을 미치는 것이다.

이런 점에 착안하여 고빈도 선호경향을 가진 네 척도의 공식에 동시미출현빈도 d항을 추가로 반영하면 이들 공식이 선호하는 빈도수준을 낮출 수 있을 것이다. 예를 들어 고빈도 선호경향을 가진 R&R 공식에서 a항을 d항으로 치환한 형태의 공식을 원래의 R&R 공식과 더하면, 선호하는 빈도수준이 R&R보다 낮은 SM 공식이 된다.

$$\frac{a}{N}+\frac{d}{N}=\frac{a+d}{N}$$

역시 KUL2 공식에서 a항을 d항으로 바꾼 형태의 공식을 원래의 KUL2 공식과 더하면, 선호하는 빈도수준이 낮은 SS4 공식에 2를 곱한 형태가 된다.: ①

이와 같이 고빈도 선호경향을 가진 공식에서 동시출현빈도에 해당하는 a항을 동시미출현빈도 d항으로 치환한 공식을 원래의 공식과 결합하면 선호하는 빈도수준이 낮은 공식이 만들어지게 된다. 동일한 방법을 COS와 JAC, DIC 척도에 적용하면 기존에 존재하지 않았던 새로운 공식들이 다음과 같이 도출된다.: ②

①	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)+\frac{1}{2}\left(\frac{d}{d+b}+\frac{d}{d+c}\right)$ $=2\times\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$
②	$\frac{a}{\sqrt{(a+b)(a+c)}}+\frac{d}{\sqrt{(d+b)(d+c)}}$ $\frac{a}{a+b+c}+\frac{d}{d+b+c}$ $\frac{2a}{2a+b+c}+\frac{2d}{2d+b+c}$

이렇게 만들어진 공식은 원래의 공식에 비해서 저빈도 선호경향을 가지며 범위가 0에서 2까지이므로 2로 나누면 원래의 척도와 범위가 같아진다.

위에서 도출된 새로운 공식에서 *a*항이 있는 부분과 *d*항이 있는 부분을 결합할 때, 단순히 더해서 2로 나누는 방식은 두 부분의 반영비율을 동일하게 50%씩으로 설정하는 것이다. *d*항이 있는 부분이 원래 척도의 빈도수준 선호경향을 낮은 쪽으로 끌어내리는 역할을 하므로, *a*항과 *d*항의 반영비율을 조절하면 새로운 공식의 빈도수준 선호경향을 조절할 수 있을 것이다. 예를 들어 *d*항이 있는 부분의 반영비율을 10%로 하고 나머지 90%를 *a*항이 있는 부분에 할당하면 원래 척도보다 고빈도 선호경향이 약간 덜한 척도가 된다. 반대로 *a*항이 있는 부분의 반영비율을 10%로 하고 *d*항이 있는 부분의 반영비율을 90%로 하면 저빈도 선호경향이 강한 척도가 된다. 결국 *a*항이 있는 부분의 반영비율이 새로운 척도가 선호하는 빈도수준과 비례한다.

이를 감안하여 앞에서 도출한 새로운 공식에 *a*항이 있는 부분의 반영비율(1에서 0 사이)을 나타내는 파라미터 *alpha*를 도입하고 척도값의 범위를 0에서 1 사이로 제한하기 위해 2로 나누어서 만들어진 COS_{alpha}, JAC_{alpha}, DIC_{alpha} 척도를 다음과 같이 제안한다.: ③

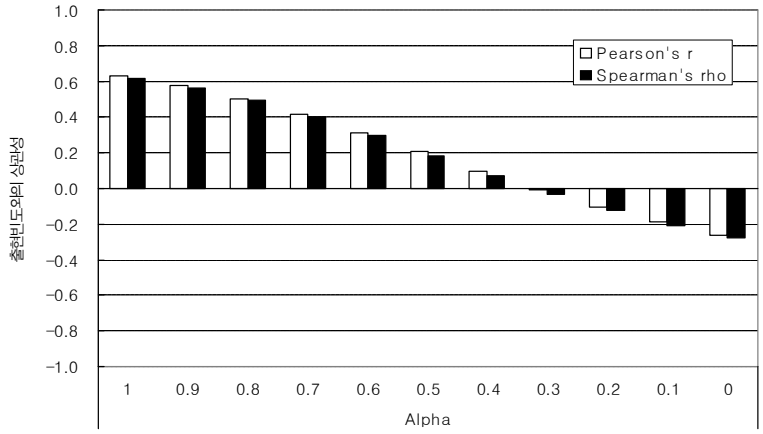
COS 척도를 변형한 COS_{alpha} 척도에 대해서 2장에서처럼 최대빈도 N이 50인 경우에 가능한 모든 조합에 대해서 분석대상의 출현빈도와 연관성 척도값 사이의 상관계수를 구하면 〈표 6〉과 〈그림 2〉와 같다. 이 표와 그림을 보면 *alpha*값이 1에서 낮아질 수록 선호하는 분석대상의 빈도수준도 비례하여 낮아짐이 분명하다.

이 COS_{alpha} 척도는 적용할 영역에 따라서 높게 판정하고 싶은 빈도수준을 파라미터 *alpha*를 통해 임의로 결정할 수 있다는 장점이 있다.

③	$COS_{alpha} = \alpha \frac{a}{2\sqrt{(a+b)(a+c)}} + (1-\alpha)\frac{d}{2\sqrt{(d+b)(d+c)}}$ $JAC_{alpha} = \alpha \frac{a}{2(a+b+c)} + (1-\alpha)\frac{d}{2(d+b+c)}$ $DIC_{alpha} = \alpha \frac{a}{2a+b+c} + (1-\alpha)\frac{d}{2d+b+c}$

〈표 6〉 alpha값에 따른 COS_{alpha} 척도값과 출현빈도와의 상관

	alpha										
	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
피어슨 상관	0.634	0.575	0.502	0.414	0.314	0.205	0.096	-0.010	-0.106	-0.190	-0.263
스피어만 순위상관	0.619	0.562	0.490	0.402	0.298	0.182	0.068	-0.036	-0.127	-0.206	-0.275



〈그림 2〉 alpha값에 따른 COS_{alpha} 척도값과 출현빈도와의 상관

3. 2 빈도수준 조절이 적용 결과에 미치는 영향 실험

이 연구에서 제안한 COS_{alpha} 척도를 실제로 적용할 때 파라미터 *alpha*를 조절하여 얻어지는 결과를 알아보기 위해서 질의확장 정보검색 실험을 수행하였다. 질의확장 정보검색에서는 입력된 질의어와의 연관성이 가장 높은 용어를 선택하여 질의에 추가한다. 이때 문헌집단에서 얻어진 용어간의 동시출현빈도를 이용하는 방법을 전역적 질의확장 검색이라 한다. 일반적으로 질의확장을 위한 연관성 척도로는 MI가 좋은 것으로 알려져 있으며 COS 척도를 적용한 경우에는 이보다 성능이 낮게 나타난다(Chung & Lee 2004).

용어간 연관성 측정을 위해서 COS 척도를 수정한 COS_{alpha} 척도를 적용하되 *alpha*값을 1에서 0.5까지 0.1씩 낮추면서 질의확장 검색실험을 수행하였다.

실험집단으로는 〈표 7〉과 같이 규모와 주제, 언어가 다양한 3개 실험집단을 이용하였다. 이중에서 CACM 실험집단은 원래 질의문이 64건이지만 적합문헌이 1건 이하이거나 서지사항을 질의로 사용한 경우 등을 제외하고 45건의 질의만을 채택하였다. 그리고 HANTEC-S 실험집단은 원래 12만 건으로 구성된 HANTEC 2.0 실험집단(김지영 외 2000) 중에서 과학기술 분야만을 대상으로 한 것이다.

〈표 7〉 검색실험집단의 특성

	CACM	HANTEC-S
언 어	영 어	한 국 어
분 야	전산학	과학기술 전반
성 격	초 록	초록 또는 전문
문헌 수	3,204	39,838
질의문 수	45	30

색인어 가중치 $DW(t_i)$와 질의어 가중치 $QW(t_i)$, 그리고 문헌 D와 질의 Q간의 유사도 $sim(D,Q)$는 각각 다음 공식으로 산출하였다.

$$DW(t_i) = 1 + \log tf(t_i)$$

$$QW(t_i) = (1 + \log tf(t_i)) \times \log \frac{N}{df(t_i)}$$

$$sim(D, Q) = \frac{\sum DW(t_i) QW(t_i)}{\sqrt{\sum DW(t_i)^2 \times \sum QW(t_i)^2}}$$

질의와의 연관성이 높은 순으로 질의에 추가하는 용어의 수는 안정된 성능을 보이는 40개로 결정하였으며 검색 성능의 평가척도로는 10위내 정확률(P@10)과 11지점 평균 정확률(11AP)을 사용하였다.

두 실험집단에 대해서 COS_{alpha} 척도를 적용하여 질의확장 검색한 것과 COS 척도, MI 척도를 이용하여 질의확장 검색한 결과를 〈표 8〉에 나타내었다. MI 척도를 적용한 결과도 제시한 이유는 이 척도가 질의확장 정보검색에서 좋은 성능을 보이는 것으로 알려져 있기 때문이다.

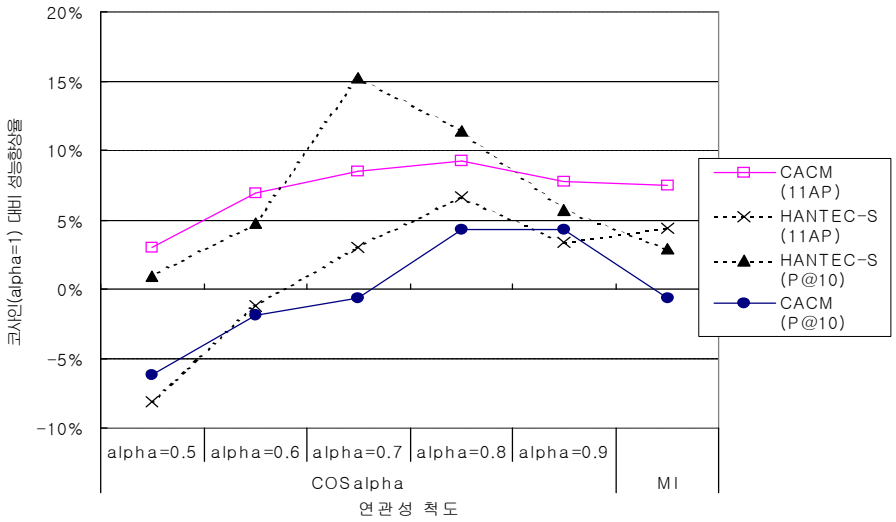
〈그림 3〉에서는 빈도수준 선호경향을 조절하지 않은 원래의 COS 척도를 적용한 검색성능을 기준으로 COS_{alpha} 척도와 MI 척도를 적용한 경우의 성능향상율을 제시하였다.

〈표 8〉과 〈그림 3〉을 보면 *alpha*가 0.8일 때와 0.9일때 두 실험집단과 두 평가척도 모두에서 COS 척도를 적용한 경우보다 COS_{alpha} 척도를 적용한 경우가 검색성능이 좋다는 것을 알 수 있다. 특히 *alpha*를 0.8로 하면 MI 척도를 적용한 경우보다도 모든 면에서 성능이 좋게 나타났다. *alpha*를 0.6 이하로 낮추어서 저빈도 선호경향을 크게 강화하면 질의확장 검색성능은 향상되지 않고 오히려 COS 척도를 적용한 경우보다도 낮아지는 현상도 보인다.

결국 흔히 사용되는 COS 척도를 변경한 COS_{alpha} 척도를 적용하면서 파라미터 *alpha*를 통해 선호하는 빈도수준을 조절하면 질의확장 검색성능이 달라지는 결과를 얻었다. 더욱이 파라미터 *alpha*를 0.8이나 0.9 정도로 지정하면 기존에 성능이 좋은 것으로 알려진 MI 척도를 적용한 것보다 더 좋은 결과가 나타났다.

〈표 8〉 빈도수준 선호경향을 조절하여 적용한 질의확장 검색 결과

실험집단 및 평가척도 / 연관성 척도		CACM		HANTEC-S	
		P@10	11AP	P@10	11AP
COSalpha	*alpha* =0.5	0.3400	0.3004	0.3533	0.2450
	alpha =0.6	0.3556	0.3118	0.3667	0.2636
	alpha =0.7	0.3600	0.3164	0.4033	0.2747
	alpha =0.8	0.3778	0.3187	0.3900	0.2845
	alpha =0.9	0.3778	0.3143	0.3700	0.2758
COS		0.3622	0.2916	0.3500	0.2668
MI		0.3600	0.3134	0.3600	0.2786



〈그림 3〉 alpha 값에 따른 질의확장 검색 결과의 성능향상율

4. 결 론

통계적 분석에 흔히 사용되는 연관성 척도에 대해서 가상으로 생성한 데이터를 대상으로 각 척도의 빈도수준 선호경향을 분석해보았다. 또한 관찰결과를 바탕으로 고빈도수준 선호경향이 있는 연관성 척도를 수정하여 선호하는 빈도수준을 낮출 수 있는 방안을 개발하였다. 분석 및 실험을 통해 얻어진 결과는 다음과 같다.

첫째, 분석대상 15개 연관성 척도 중에서 R&R이 고빈도수준 선호경향이 가장 강한 반면, MI가 저빈도수준 선호경향이 가장 강한 것으로 나타났다. 이 밖에 저빈도수준 선호경향이 강한 척도로는 LOR, YULEP이 있었다. 이들 척도를 사용하면 연관도 순위를 산출할 때 매우 고빈도이거나 매우 저빈도인 대상이 상위에 집중된다는 것을 예측할 수 있다.

둘째, RMIJ는 비슷한 성향을 보이는 PHI, SS4과 함께 분석대상의 빈도 수준에 거의 영향받지 않는 중립적인 척도임이 드러났다.

셋째, 대부분의 척도는 선호하는 빈도수준이 최대빈도 상한에 크게 좌우되지 않고 일정하게 높거나 낮지만, SM과 GSS는 빈도 상한이 50%를 넘을 때에는 둘 다 중간빈도 수준을 선호하다가 빈도 상한이 50% 이하이면 SM은 저빈도 선호경향을, GSS는 고빈도 선호경향을 나타낸다. 따라서 이 두 척도를 사용할 때에는 적용 데이터의 최대빈도 상한 비율에 따라 다른 결과가 나타날 수 있으므로 주의해야 한다.

넷째, 연관성 척도 공식의 분자에 동시출현빈도를 뜻하는 a항만 있는 공식은 고빈도수준 선호경향을 가지며, 그렇지 않은 공식은 이보다 선호하는 빈도수준이 낮게 나타났다.

다섯째, 고빈도수준 선호경향이 있는 연관

성 척도 공식의 *a*항을 동시미출현빈도 *d*항으로 치환하여 원래의 공식과 결합하는 방법을 통해, 선호하는 빈도수준이 낮은 새로운 연관성 척도로 COS_{alpha}, JAC_{alpha}, DIC_{alpha}를 제안하였다. 이때 두 부분의 결합비율을 조절하는 파라미터 *alpha*를 통해 새로운 척도의 빈도수준 선호경향을 조절할 수 있다.

여섯째, COS 척도를 변형하여 빈도수준 선호경향을 조절할 수 있도록 제안한 COS_{alpha} 척도를 질의확장 검색에 적용해본 결과 파라미터 *alpha*에 따른 선호경향의 차이가 검색성능에 영향을 미치는 것으로 나타났다. 특히 파라미터 *alpha*를 0.8 정도로 설정하여 원래의 COS 척도와의 차이를 크지 않게 하면 기존에 성능이 좋다고 알려진 MI 척도를 사용한 것보다 더 좋은 성능을 얻을 수 있었다.

이 연구에서 각 연관성 척도의 빈도수준 선호경향을 분석한 결과 개별 척도의 경향을 파악함과 동시에 비슷한 경향을 지닌 척도를 구분할 수 있었다. 또한 새롭게 제안한 연관성 척도를 이용한 검색 실험을 통해서 빈도수준 선호경향의 차이가 응용 결과의 성능에 영향을 미치는 것이 입증되었다. 선호하는 빈도수준이 비슷한 연관성 척도는 적용 결과도 크게 차이나지 않으므로 향후 연관성 척도를 적용하는 실험에서는 척도를 선택적으로 적용함으로써 실험이나 분석의 효율을 높일 수 있을 것이다. 특히 흔히 사용되는 COS나 JAC 척도가 선호하는 빈도수준을 필요에 따라서 저빈도수준으로 바꿀 수 있는 방안이 제시되었으므로, 이들 척도를 적용해온 여러 관련 분야에서 제안된 새로운 척도를 적용하는 후속 연구가 필요할 것이다.

참 고 문 헌

김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. 한국어 테스트 컬렉션 HANTEC의 확장 및 보완 『제12회 한글 및 한국어 정보처리 학술대회 논문집』, 210-215.

사공철 외. 2003. 『정보학 사전』. 서울: 문헌정보처리연구회.

이재윤. 2003. 상호정보량의 정규화에 대한 연구. 『문헌정보학회지』, 37(4): 177-198.

정영미. 1987. 『정보검색론』. 서울: 정음사.

Chung, Young Mee, and Jae Yun Lee. 2001. "A corpus-based approach to comparative evaluation of statistical term association measures." *Journal of the American Society for Information Science and Technology*, 352(4): 283-296.

Chung, Young Mee, and Jae Yun Lee. 2004. "Optimization of some factors affecting the performance of query expansion." *Information Processing and Management*, 40(6): 891-917.

Forman, George. 2003. "An extensive empirical study of feature selection metrics for text classification." *Journal of Machine Learning Research*,

3(Mar): 1289-1305.

Galavotti, L., F. Sebastiani, and M. Simi. 2000. "Experiments on the use of feature selection and negative evidence in automated text categorization." In *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries* (Lisbon, Portugal, 2000): 59-68.

Gower, J. C. 1985. "Measures of similarity, dissimilarity, and distance." In *Encyclopedia of Statistical Sciences*, Vol. 5, eds. S. Kotz and N.L. Johnson: 397-405.

Meyer, A., A. A. F. Garcia, A. P. de Souza, and C. L. de Souza Jr. 2004. "Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L)." *Genetics and Molecular Biology*, 27(1): 83-91.

Mladeni'c, D. 1998. "Feature subset selection in text-learning." In *Proceedings of the Tenth European Conference on Machine Learning* (Chemnitz, Germany, 1998): 95-100.

Salton, G., and C. S. Yang. 1973. "On the specification of term values in automatic indexing." *Journal of Documentation*, 29(4): 351-372.

van Rijsbergen, C. J. 1979. *Information Retrieval*. 2nd ed. London: Butterworths.

Yang, Yiming, and Jan O. Pedersen. 1997. "A comparative study on feature selection in text categorization." *Proceedings of the 14th International Conference on Machine Learning: 412-420.*