연관성 척도의 빈도수준 선호지수 개발

A Frequency Level Preference Index of the Association Measures

이재윤, 경기대학교 문헌정보학과, memexlee@kyonggi.ac.kr

Jae-Yun Lee, Dept. of Library and Information Science, Kyonggi University

연관성 척도값은 연관성 분석 대상이 고빈도인지 저빈도인지 여부에 따른 영향을 받는데, 연관성 척도마다 주로 높은 연관성으로 판정하는 대상의 빈도수준이 다양하게 나타난다. 이런 연관성 척도의 빈도수준 선호경향을 수치로 나타낼 수 있다면 연관성 척도를 사용하는 실험이나 분석에서 시행착오나 시간낭비를 줄일 수 있을 것이다. 이를 위해서 연관성 척도의 빈도수준 선호지수(FLPI)를 개발하였다. 개발된 빈도수준 선호지수는 연관성 척도와 출현빈도 사이의 상관성을 이용하는 것으로서 연관성 척도를 적용하는 실험이나 분석의 효율을 높이는데 기여할 것으로 기대된다.

1. 서론

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 동시출현빈도를 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 선호 경향으로 부르기로 한다.

연관성 척도의 적용분야 중에서 질의확장 검색의 경우에는 자카드와 같은 고빈도 선호 연관성 척도에 비해서 상호정보량과 같은 저빈도 선호 연관성 척도가 좋은 성능을 보인는 것으로 나타났다(이재윤 2003). 반면에 자동분류를 위한 자질선정에 있어서는 상호정보량을 척도로 이용하는 것이 좋지 못한 결과를 가져온다고 알려져 있다(Yang & Pederson 1997).

이와 같이 연관성 척도의 빈도수준 선호 경향이 적용 결과에 큰 영향을 미침에도 불구하고, 여러 연관성 척도의 빈도수준 선호 경향을 비교할 수 있는 기준은 제시된 바가 없다. 그러다보니 일부 연구에서는 다양한 연관성 척도를 망라적으로 적용해보는 시도도 있었다(한승희, 이재윤 1999; Meyer et al. 2004).

각 연관성 척도의 빈도수준 선호 경향이 어떤지를 미리 파악할 수 있다면, 연관성 척도를 사용하는 연구에서 불필요하게 많은 종류의 척

도를 적용해보는 시행착오 및 시간낭비를 줄일 수 있을 것이다. 이를 위해서는 빈도수준 선호경향을 간단한 수치로 표현하여 비교하게 하는 것이 가장 바람직하다.

이 연구에서는 실험을 통해 여러 연관성 척도의 빈도수준 선호경향을 분석한 다음, 이를 바탕으로 연관성 척도의 빈도수준 선호지수(Frequency Level Preference Index; FLPI)를 제안하였다.

2. 연관성 척도의 빈도수준 선호경향 분석

2.1 실험 방법

연관성 척도의 빈도수준 선호경향을 분석하기 위해서는 Chung and Lee(2001)에서처럼 실제 데이터를 대상으로 연관성 척도를 적용해 볼 수도 있다. 그러나 이런 실험 결과는 어디까지나 해당 실험 데이터에 의존해서 해석할 수밖에 없으므로, 분석 대상에 따라 달라지면 안되는 지수로 이용하기가 어렵다. 따라서 이 연구에서는 특정한 실험집합을 사용하지 않고, 이론적으로 가능한 경우의 수를 모두 포함하도록 연관성 측정 대상들의 출현빈도 조합을 인위적으로 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 a와 b가 있을때, 각각의 빈도 f(a)와 f(b)를 1에서 최고 N까지 변화시키되 f(a)가 f(b)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(a),f(b))에 대해서 a와 b의 동시출현빈도 f(a,b)는 둘 중에 빈도가 낮은 f(a)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, 대부분의 경우에 동시출현하지 않은 쌍은 분석에서 제외하므로 여기에서도 동시출현빈도 1 이상인 경우만 포함하였다. 설사 포함하더라도 동시출현빈도가 0인 조합쌍은 연관성이 최하위로 판정되기 마련이므로 분석의 의미가 없다.

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.

각 연관성 척도가 분석대상의 출현빈도에 따라서 어떤 성향을 보이는가를 판단하는 척도로는 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.

빈도수준 선호경향을 분석할 연관성 척도로는 Gower(1985)에 소개된 여러 척도 이외에 정보학 분야에 응용되어 좋은 성능을 보이는 것으로 알려진 상호정보량, 상대적 상호정보량,

<표 1> 분석 대상 연관성 척도

명칭	약호	공식
Russel & Rao	R&R	$\frac{a}{N}$
Jaccard	JAC	$\frac{a}{a+b+c}$
Dice	DIC	$\frac{2a}{2a+b+c}$
Cosine (Ochiai)	COS	$\frac{a}{\sqrt{(a+b)(a+c)}}$
Kulczynski 2	KUL2	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
Simple Matching	SM	$\frac{a+d}{N}$
Chi-square	CHI	$\frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$
GSS (Dispersion)	GSS	$\frac{ad-bc}{N^2}$
Pearson's PHI	PHI	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Anderberg (Sokal & Sneath 4)	AND	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$
Relative Mutual Information J	RMIJ	$\frac{\log_2 N+\log_2 a-\log_2(a+b)(a+c)}{\log_2 N-\log_2 a}$
Sokal & Sneath 5 (Ochiai 2)	SS5	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Log Odds Ratio	LOR	$\log\frac{ad}{bc}$
Yule's Y	YULE	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
Mutual Information	MI	$\log_2\frac{Na}{(a+b)(a+c)}$

GSS계수, 카이제곱 계수, 로그승산비를 선정하였다. 각 연관성 척도를 2×2 분할표에 적용한 형태로 나타낸 공식은 <표 1>과 같다.

2.2 포괄적인 빈도수준 선호경향 분석

가능한 최고 출현빈도 N의 절대적인 크기는 빈도수준 선호경향과는 상관이 없다. 연관성 척도의 빈도수준 선호경향은 상대적인 빈도수준에 좌우되기 때문이다. 따라서 굳이 N을 몇 천이나 몇 만과 같이 큰 값으로 설정할 필요가 없다. 이 연구에서는 N을 30으로 정하였다.

다만 일부 연관성 척도의 경우에 출현빈도가 둘 다 최고값인 N이면 분모가 0이 되어 계산이 안되는 상황이 발생하므로 분석 대상의 출현빈도를 1에서 29까지만 변화시키면서 연관성 척도를 적용하였다.

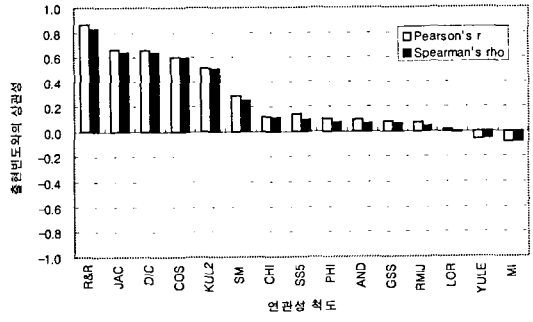
분석 대상 둘의 출현 빈도 중 낮은 값과 각 연관성 척도의 상관계수를 구한 결과는 표 2와 같으며 이를 그림으로 나타낸 것이 그림 1이다.

순위상관 Spearman's rho를 기준으로 보았을 때, 동시출현빈도를 최고 빈도로 나눈 R&R

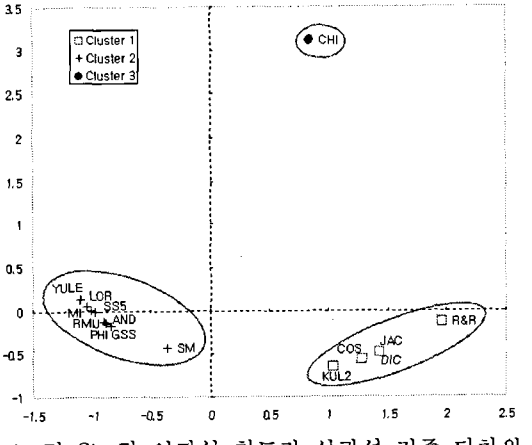
<표 2> 각 연관성 척도의 출현빈도와의 상관성 (순위상관성이 높은 것을 왼쪽에 둠)

	R&R	JAC	DIC	COS	KUL2	SM	CHI	SS5	PHI	AND	GSS	RMU	LOR	YULE	MI
Pearson's r	0.868	0.660	0.656	0.606	0.514	0.288	0.114	0.142	0.099	0.095	0.080	0.069	0.014	-0.062	-0.083
Spearman's rho	0.829	0.642	0.642	0.593	0.509	0.251	0.112	0.090	0.070	0.067	0.066	0.044	-0.012	-0.053	-0.083

이 가장 높은 상관성을 보이며 JAC, DIC, COS, KUL2가 0.5 이상의 상관성을, SM이 0.2대, CHI가 0.1대의 상관성을 보였다. 나머지는 0.1 이하로 나타나서 출현빈도와 연관도 사이의 상관성이 거의 없는 것으로 볼 수 있다. 심지어 LOR과 YULE, MI는 0이하의 부정적 상관성을 미미하게나마 보이고 있다.



<그림 1> 각 연관성 척도의 출현빈도와의 상관성



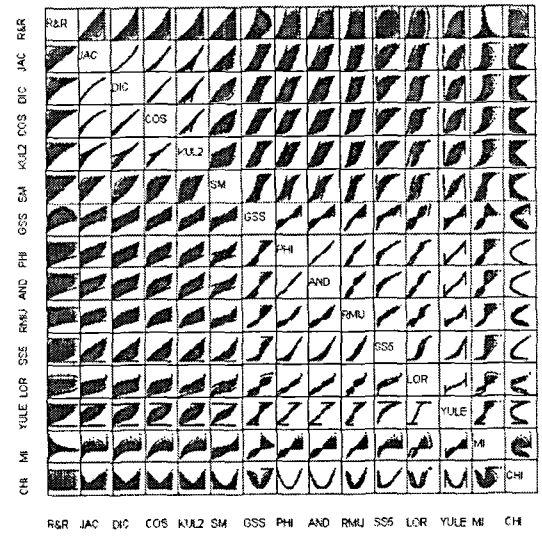
<그림 2> 각 연관성 척도간 상관성 기준 다차원 척도법 및 군집분석 결과

이로 미루어보면 출현빈도를 근거로 연관성을 판단할 때 R&R이 가장 큰 고빈도 선호경향을 보이며 반대로 MI가 가장 큰 저빈도 선호경향을 보이는 것으로 판단할 수 있다.

출현빈도와의 상관성이 비슷한 연관성 척도가 서로간의 상관성도 높은지를 검토하기 위해서, 각 연관성 척도 사이의 순위상관을 모두 구한 다음 이를 기준으로 각 연관성 척도를 완전연결기법으로 군집화하면 그림 2와 같다.

전체적으로 출현빈도와의 상관성이 0.5 이상인 군집 1과 0.3 이하인 군집 2로 나누어지지만, CHI의 경우에는 상관성이 낮은 군집과도 함께 묶이지 않음을 알 수가 있다. 그 이유를 확인하기 위해서 각 연관성 척도간의 상관성을 산포도로 그려보면 그림 3과 같다.

그림 3의 왼쪽 위에 있는 JAC, DIC, COS, KUL2 끼리는 거의 선형적인 상관성을 보여주



<그림 3> 빈도 30이내 조합쌍에 대한 연관성 척도값을 상호 비교한 산포도

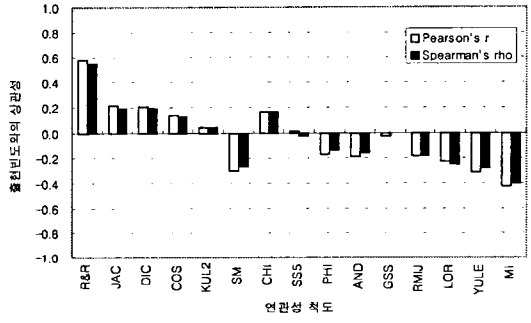
고 있으며, 오른쪽 아래에 있는 PHI, RMIJ, SS5, LOR 끼리도 어느 정도 선형적 상관성이 나타난다. MI의 경우에는 R&R 값(즉 출현빈도)이 중간까지는 증가하는 양의 상관성을 보이다가, 이를 넘어서면서 다시 감소하는 음의 상관성을 나타내고 있다. CHI는 이와 달리 다른 연관성 척도와 비교할 때, 다른 척도값이 증가함에 따라서 중간까지는 CHI 척도값은 감소하여 음의 상관성을 보이다가 이후부터는 증가하는 양의 상관성을 보이고 있다. CHI나 MI의 경우에는 특이하게도 척도값의 중간 지점을 기준으로 성향이 달라지는 것이다. 이는 선형적이지는 않지만 비선형적인 상관성을 가진다는 것을 의미한다.

이 점을 고려해볼 때, 출현빈도에 따른 연관성 척도의 특성을 분석하기 위해서는 분석 단위를 빈도 수준별로 나누어볼 필요가 있다고 여겨진다.

2.3 출현빈도가 N/2 이하인 경우의 빈도수준 선호경향 분석

일반적인 상황이라면 전체 집단의 반 이상에 분석 대상이 출현하는 경우는 드물다고 할 수 있다. 예를 들어 문헌 집단에서 색인어가 출현한 문헌은 대개 절반에 훨씬 못 미치는 것이 보통이다. 따라서 분석대상의 출현빈도가 최고빈도의 절반 이하인 경우만을 대상으로 분석하는 것은 오히려 실제 응용에 더 가까운 분석이 될 수 있다.

여기서는 최고빈도가 30인 집단에서 출현빈도가 1에서 15 사이인 경우만을 대상으로 출현빈도와 연관성 척도값 사이의 상관성을 산출하여 그림 4에 제시하였다.

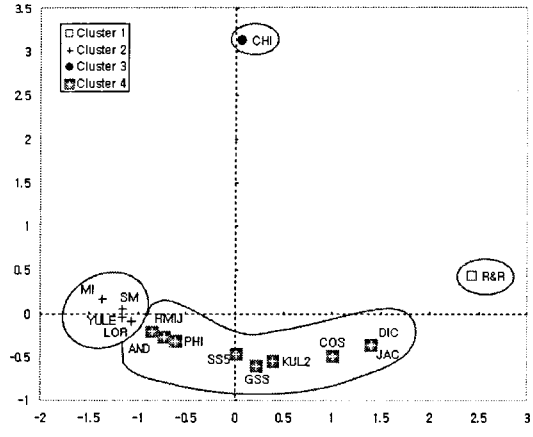


<그림 4> 출현빈도를 N/2 이하로 제한했을 때 출현빈도와의 상관성

출현빈도를 제한하지 않았을 경우에 비해서 출현빈도와의 상관성은 대부분의 연관성 척도가 낮아진 것으로 나타났다. 반면에 CHI의 경우에는 오히려 상관성이 높아져서 다른 연관성 척도와 매우 다른 성질을 가졌음을 보여준다.

앞에서와 마찬가지로 연관성 척도간의 순위상관성을 기준으로 완전연결 군집분석과 다차원척도법을 실행한 결과는 그림 5와 같다.

출현빈도가 최고빈도의 절반 이하인 경우에서는 R&R만이 출현빈도와의 상관성이 0.5 이상으로 높아서 별도의 군집을 이루며, JAC, DIC, COS, KUL2는 R&R로부터 멀어져서 저



<그림 5> 출현빈도를 N/2 이하로 제한했을 때 연관성 척도의 다차원척도법 및 군집분석 결과

<표 3> 각 연관성 척도의 빈도수준 선호지수 (N=30일때 출현빈도와 연관성 척도값의 순위상관성)

최고빈도 제한	R&R	JAC	DIC	COS	KUL2	SM	CHI	SS5	PHI	AND	GSS	RMIJ	LOR	YULE	MI
제한 안함	0.829	0.642	0.642	0.593	0.509	0.251	0.112	0.090	0.070	0.067	0.066	0.044	-0.012	-0.053	-0.083
N/2 으로 제한	0.551	0.193	0.193	0.126	0.039	-0.262	0.163	-0.027	-0.138	-0.158	-0.006	-0.180	-0.254	-0.282	-0.402

빈도 선호경향이 높은 계수들과 한 군집을 이루는 것으로 나타났다. 출현빈도와의 상관성이 -0.2 이하로 매우 낮아진 LOR, YULE, SM, MI는 별도의 군집을 이루었다. CHI는 여전히 다른 연관성 척도와 완전히 떨어진 위치에 자리를 잡고 있다.

출현빈도를 제한하지 않았던 앞의 분석 결과와 비교하기 위해서 두 경우의 순위상관을 함께 나타낸 것이 표 3이다.

결국 고빈도 선호경향이 가장 큰 R&R은 (0.83, 0.55)의 상관성을 가지며, 저빈도 선호경향이 가장 큰 MI는 (-0.08, -0.40)의 상관성을 가진다. 이 두 가지 수치가 특정 연관성 척도의 빈도수준 선호경향을 나타낸다는 점을 고려하여 연관성 척도의 빈도수준 선호지수 (Frequency Level Preference Index; FLPI) 로 사용할 수 있을 것이다.

3. 결론

연관성 척도값이 분석대상의 출현빈도와 가지는 순위상관성을 이용하여 연관성 척도의 빈도수준 선호지수 FLPI를 제안하였다. 이 지수는 분석대상의 빈도분포가 최저빈도에서 최고빈도 가까이에 이르도록 넓은 경우와, 중간빈도 이하로 제한되는 경우를 각각 반영하는 두 수치로 구성되어 있다.

제안된 FLPI 지수를 이용하면 적용 결과가 좋지 않을 연관성 척도를 사전 실험을 통해 미리 걸러낼 수 있으므로 실험이나 분석의 효율을 높이는데 기여할 것으로 기대된다.

참고문헌

이재윤. 2003. "유사계수에 따른 전역적 질의확장 검색 성능 비교." 『2003년도 한국정보과학회 가을 학술발표논문집(I)』, 30(2): 526-528.

한승희, 이재윤. 1999. "문헌클러스터링을 위한 유사계수간의 연관성 측정." 『제6회 한국정보관리학회 학술대회 논문집』, 25-28.

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*, 52(4): 283-296.

Gower, J. C. 1985. "Measures of similarity, dissimilarity, and distance." In *Encyclopedia of Statistical Sciences*, Vol. 5, eds. S. Kotz and N.L. Johnson (Wiley-Interscience, 1985), pp. 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.

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