상호정보량의 정규화에 대한 연구

A Study on Relative Mutual Information Coefficients

이 재 윤(Jae-Yun Lee)*

목 차

1. 서 론	3. 1 실험 설계
2. 상호정보량과 상대적 상호정보량 계수	3. 2 실험집단에서 연관성 척도의 특성 분석
2. 1 정보량과 상호정보량	3. 3 검색성능 비교분석
2. 2 상호정보량의 정규화	4. 결 론
2. 3 상대적 상호정보량 계수의 특성	
3. 질의확장 정보검색에의 응용 실험	

초 록

상호정보량은 용어간 유사도 산출을 비롯한 다양한 분야에서 연관성 척도로 사용되어왔다. 그러나 값의 범위가 일정하지 않으며 지나치게 저빈도인 경우를 선호하는 경향이 제한점으로 지적되고 있다. 이런 점을 보완하기 위해서 상호정보량을 정규화하는 상대적 상호정보량 계수를 제안하였다. 제안된 계수의 특성을 알아본 다음, 세 실험집단을 대상으로 전역적(global) 질의확장 검색을 수행한 결과 검색 성능을 향상시킬 수 있었다.

ABSTRACTS

Mutual information, as an association measure, has been used for various purposes as well as for calculating term similarity. There are, however, some limits in mutual information. It tends to emphasize low frequency terms extremely because the marginal value of mutual information changes inversely to frequency of terms. To compensate for this limit, this study suggests relative mutual information(RMI) coefficients which normalize mutual information, and examines their characteristics in some details. The RMI coefficients also improve effectiveness of global query expansion when they are adapted to three different collections.

키워드: 상호정보량, 상대적 상호정보량 계수, 연관성척도, 정보검색, 질의확장

Mutual Information, Relative Mutual Information Coefficients, Association Measures,
Information Retrieval, Query Expansion

* 연세대학교 문헌정보학과 강사(memexlee@lis.yonsei.ac.kr)
논문접수일자 2003년 11월 22일
게재확정일자 2003년 12월 11일

1. 서론

컴퓨터에 의한 자연언어 처리 기술이 발전하면서 단어간의 관계를 언어학적 이론이나 직관적 분석에만 의존하지 않고 텍스트에 대한 통계적인 분석을 통해 자동으로 파악하고자 하는 연구가 점차 증가하고 있다. 이런 연구에서는 단어간의 유사도를 측정하기 위해 두 단어의 동시출현빈도를 이용하면서 특정한 공식 또는 계수를 사용하는데, 이때 사용하는 공식을 유사계수 또는 연관성 척도라고 한다.

코사인 계수를 비롯한 다양한 연관성 척도가 오랫동안 사용되어 왔지만, Shannon의 정보이론에서 유래한 상호정보량을 Church & Hanks(1990)가 연어(collocation) 분석에 적용한 이후부터는 많은 연구에서 상호정보량 공식을 이용하고 있다. 이런 경향은 특히 국내의 정보검색 및 자연언어 처리 분야에서 뚜렷하게 나타나고 있다(강현규 1997; 김명철 외 1992; 김정세 1996; 김판구 1994; 이공주, 김재훈, 김길창 1995; 장명길 2002; 정석원 2001).

이와 같이 널리 사용되고 있는 상호정보량 공식이지만, 연관성을 측정하는 유사성 함수로서의 기본 성질을 만족하지 못하며 지나치게 저빈도인 경우를 선호하는 것이 때로는 제한점이 되고 있다. 이 연구에서는 이런 상호정보량의 제한점을 극복할 수 있는 방안을 공식의 정규화라는 측면에서 찾아보고자 한다. 이를 위해 2장에서는 상호정보량과 그 제한점에 대해서 살펴본 다음 이를 보완한 상대적 상호정보량 계수를 제안하고 특성을 분석하였다.

제안된 상대적 상호정보량 계수의 우수성을 검증하기 위해서 3장에서는 상호정보량의 여러 응용분야 중에서 전역적 질의확장 검색 분야를 대상으로 실험을 수행하였다.

2. 상호정보량과 상대적 상호정보량 계수

2.1 정보량과 상호정보량

Shannon(1948)의 정보이론에서 제시된 정보량과 엔트로피, 상호정보량을 부호로 표기하는 방식은 이후의 관련 문헌마다 약간씩 차이가 있다. 예를 들어 엔트로피의 경우에 I, E, H, U 등의 다양한 부호가 사용된다. 여기서는 정보량을 $I(x)$, 엔트로피를 $H(X)$, 정보량에 대한 상호정보량을 $MI(x,y)$, 엔트로피에 대한 상호정보량을 $MI(X;Y)$ 로 각각 표기한다.

어떤 사건의 발생확률이 $p(x)$ 일 때, 해당 사건의 정보량은 다음과 같이 정의할 수 있다.

$$I(x) = -\log_2 p(x)$$

이산 확률변수 X 의 각 사건에 대한 정보량을 다음 공식과 같이 발생확률에 따라 가중 평균한 것을 평균정보량 또는 엔트로피라고 부른다(Shannon 1948).

$$H(X) = -\sum_x p(x) \log_2 p(x)$$

정보량의 경우에 두 사건 x 와 y 가 동시출현

할 확률과 각각 독립적으로 출현할 확률 사이의 비율을 상호정보량(Mutual Information; MI)이라고 부르며 아래의 공식으로 구한다.

$$MI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

엔트로피에 대한 상호정보량은 위의 정보량에 대한 상호정보량을 역시 동시출현 사건의 확률로 가중 평균한 값으로서 평균 상호정보량, 또는 기대 상호정보량이라고도 부른다 (Fano 1961).

$$MI(X; Y) = \sum_x \sum_y p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

두 단어의 연관성을 출현빈도에 근거해서 판단하는 상황에서 출현빈도와 동시출현빈도, 정보량과 상호정보량의 관계를 각각 도식화하면 <그림 1>과 같다.

<그림 1>의 (a)는 단어 x와 y의 빈도가 각각 $f(x)$, $f(y)$ 이고 $f(x) < f(y)$ 이며 두 단어의 동시출현 빈도가 $f(x, y)$ 일 때 이들의 관계를 벤다이어그램으로 나타낸 것이다. 이 경우

(a) 출현빈도와 동시출현빈도

$f(x)$ 와 $f(y)$, $f(x, y)$ 사이에는 다음과 같은 관계가 성립된다.

$$f(y) > f(x) \geq f(x, y)$$

이 두 단어 x와 y의 정보량 $I(x)$ 와 $I(y)$, 두 단어의 상호정보량 $MI(x, y)$, 그리고 동시출현 사건의 정보량 $I(x, y)$ 를 벤다이어그램으로 나타낸 것이 <그림 1>의 (b)이며 이들 사이에는 다음과 같은 관계가 성립한다.

$$MI(x, y) \leq I(y) < I(x) \leq I(x, y)$$

정보량은 출현빈도와 반비례하므로 빈도가 낮은 x의 정보량이 y의 정보량보다 크다. 또한 x와 y의 상호정보량은 빈도가 높은 y의 정보량보다 작거나 같다. 반면에 x와 y의 동시출현 사건의 정보량 $I(x, y)$ 는 빈도가 낮은 x의 정보량보다 크거나 같으며, x의 정보량과 y의 정보량을 더한 것에서 x와 y의 상호정보량을 뺀 값과 같다.

이와 같은 상호정보량 공식은 연관성 척도

(b) 정보량과 상호정보량

<그림 1> 빈도와 정보량의 벤다이어그램

로 널리 사용되고 있지만 일반적인 유사도 함수의 성질과는 다른 점이 있다. 일반적으로 유사도 함수 s 가 가지는 성질은 다음과 같다 (Spáth 1980, 16).

$$\begin{aligned} s(x, y) &\leq s_0 \\ s(x, x) &= s_0 \\ s(x, y) &= s(y, x) \end{aligned}$$

여기서 s_0 은 실수로서 임의의 최대값이다. 첫 번째는 일정한 최대값이 존재함을 뜻한다. 두 번째는 자신과의 유사도는 최대값에 해당한다는 뜻이다. 세 번째는 대칭성을 나타낸다.

첫째 성질인 최대값의 경우에 최대 출현빈도가 N 이라면 상호정보량은 $f(x) = f(y) = f(x, y) = 1$ 일 때 $\log_2 N$ 이 된다. 또한 셋째 성질인 대칭성의 경우에 상호정보량은 $MI(x, y) = MI(y, x)$ 이므로 만족한다.

그러나 $MI(x, x) = I(x)$ 이기 때문에 $f(x) = 1$ 일 경우에만 최대값이 되며 $f(x) > 1$ 이

면 이보다 작아지므로 둘째 성질은 만족하지 않는다. 실제로 $N=1,000$ 일 때 $f(x)$ 를 1부터 1,000까지 증가시키면 자신과의 상호정보량 $MI(x, x)$ 는 <그림 2>와 같이 $\log_2 1000 \approx 9.97$ 에서 0까지 감소하게 된다. 즉 빈도가 커질수록 상호정보량의 상한값은 낮아지는 것이다.

이와 같이 상한값이 빈도와 반비례하므로 단어간의 연관성을 측정할 경우에 상호정보량을 척도로 사용하면 주어진 단어에 대한 유사어로 저빈도어를 선호하게 된다(Chung & Lee 2001). 특히 두 단어 중 어느 하나의 출현빈도와 서로의 동시출현빈도가 같으면 동시출현빈도의 높고 낮음에 상관없이 무조건 최대값(빈도가 높은 단어의 정보량, 즉 두 단어의 정보량 중에서 낮은 값)을 가지게 되므로 분별력이 떨어지게 된다. 빈도가 높은 단어보다는 낮은 단어일 경우에 출현빈도와 동시출현빈도가 같을 확률이 높으므로 저빈도어가 높은 상호정보량을 가질 가능성이 높은 것이다. 상호

<그림 2> 빈도에 따른 자신과의 상호정보량 분포 (N=1,000)

정보량의 이런 지나친 저빈도 편향성은 정보 검색(이재운 2003)이나 자동분류(Yang & Pedersen 1997) 등의 여러 응용 분야에서 한계점으로 지적되고 있다.

2. 2 상호정보량의 정규화

2. 2. 1 정보량에 대한 상호정보량의 정규화
정보량에 대한 상호정보량의 범위를 일정하게 정규화하기 위해서 최대값인 $\log_2 N$ 으로 나눈 값을 사용하는 경우도 있다(Kim & Choi 1999). 이렇게 하는 것이 실제 적용을 할 때 편리하긴 하지만, 범위만 1에서 -1 사이로 제한될 뿐이고 앞에서 살펴본 출현빈도에 따른 최대값의 감소 현상 및 저빈도 편향성은 여전하다.

이런 한계를 극복하기 위해서는 단순히 최대값으로 정규화할 것이 아니라 <그림 1>에서 보았듯이 상호정보량의 한계값인 정보량을 이용하여 상호정보량을 정규화할 필요가 있다. 이에 따라 동시출현한 두 단어의 정보량에 대비한 상호정보량의 수준을 측정하는 연관성 척도로 상대적 상호정보량(Relative Mutual Information: RMI) 계수를 아래에 네 가지 형태로 제안한다.

우선 두 사건의 정보량 $I(x)$ 와 $I(y)$ 로 $MI(x,y)$ 의 상대적인 크기를 나타내면 다음의 두 비율이 얻어진다.

$$\frac{MI(x,y)}{I(x)}, \quad \frac{MI(x,y)}{I(y)}$$

이 값들 각각으로도 값의 정규화는 가능하지만 대칭성을 만족하지 못하게 되므로 두 값

을 합성하여 하나로 만들어야 한다. 일반적으로 두 값을 합성하는 방법으로는 산술평균과 기하평균, 조화평균을 들 수 있다.

두 비율의 산술평균을 취한 것을 RMIk 계수라고 부르기로 하며 공식으로는 다음과 같이 나타낸다.

$$\begin{aligned} RMIk(x,y) &= \frac{MI(x,y)/I(x) + MI(x,y)/I(y)}{2} \\ &= MI(x,y) \left[\frac{I(x) + I(y)}{2 I(x) I(y)} \right] \end{aligned}$$

두 비율의 기하평균은 RMIc 계수라고 부르기로 하며 공식은 다음과 같다.

$$RMIc(x,y) = \frac{MI(x,y)}{\sqrt{I(x)I(y)}}$$

두 비율의 조화평균은 RMIId 계수라고 부르기로 하며 공식은 다음과 같다.

$$RMIId(x,y) = \frac{2 MI(x,y)}{I(x) + I(y)}$$

한편 두 사건의 정보량에 대한 상호정보량의 개별적인 비율을 합성하는 대신에 두 사건이 결합된 동시출현 사건의 정보량, 즉 결합정보량 $I(x,y)$ 대비 $MI(x,y)$ 의 상대적인 크기를 나타낸 것을 RMIj 계수라고 부르기로 하며 공식으로는 다음과 같다.

$$\begin{aligned} RMIj(x,y) &= \frac{MI(x,y)}{I(x,y)} \\ &= \frac{MI(x,y)}{I(x) + I(y) - MI(x,y)} \end{aligned}$$

이 공식들에서 정보량 대신 출현빈도를, 상호정보량 대신 동시출현빈도를 사용하면 각각 기존의 연관성 척도 중 쿨친스키2 계수, 코사

인 계수, 다이스 계수, 자카드 계수가 된다. 상대적 상호정보량 계수의 개별 명칭은 형태가 유사한 기존 연관성 척도의 첫 글자(k, c, d, j)에서 따왔다.

2.2.2 엔트로피에 대한 상호정보량의 정규화

이 연구와 같이 정보량에 대한 상호정보량을 대상으로 하지는 않았지만 엔트로피에 대한 상호정보량, 즉 평균 상호정보량을 대상으로 정규화를 시도한 연구는 몇 가지 있었다. 정보량에 대한 상호정보량은 0 이하의 값이 있을 수 있으나, 엔트로피에 대한 상호정보량은 항상 0 이상의 값을 가진다. 따라서 이를 정규화한 값의 범위도 다르다.

Strehl(2002)은 클러스터링 결과를 평가하기 위해서 엔트로피의 상호정보량을 사용할 때 상한값을 1로 제한할 수 있는 정규 상호정보량(Normalized Mutual Information: NMI) 척도를 제시하였는데, 이는 앞에서 제안한 RMIC 계수와 유사한 형태이면서 0에서 1 사이의 값을 가진다.

$$\text{Strehl's NMI} = \frac{MI(X; Y)}{\sqrt{H(X)H(Y)}}$$

불확실성 계수(Uncertainty Coefficient: UC)는 SPSS(Nie et al. 1975)와 같은 통계 처리 소프트웨어에서 두 변수의 상관성을 교차 분석하는 데 이용되고 있다. 불확실성 계수에는 비대칭 공식과 대칭 공식이 있는데, 비대칭 공식은 분모에 한 쪽의 엔트로피만 있는 형태이며 대칭 공식은 X와 Y가 맞바뀐 두 비대칭 값을 평균한 것으로서 RMIC 계수와 유사한 형태이다. 역시 값은 0에서 1 사이에 분

포한다.

$$U(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)}$$

$$U(X, Y) = 2 \left[\frac{H(Y) + H(X) - H(X, Y)}{H(X) + H(Y)} \right]$$

한편 Astola & Virtanen(1982)은 통계학적인 측면에서 엔트로피를 재해석하면서 적률 상관계수와 유사한 성질을 가지도록 엔트로피 상관계수(Entropy Correlation Coefficient: ECC)를 제안하였는데 이는 불확실성 계수와 동일하다. 엔트로피 상관계수라는 이름으로는 MRI나 PET와 같은 의료 영상 기기의 영상 정합(image registration) 분야에서 두 영상을 화소단위로 비교하는데 사용된 바가 있다(Maes et al. 1997; Otte 2001).

Studholme(1997)도 3차원 의료영상의 비교에 적용할 척도로 역시 정규 상호정보량이라는 이름의 공식을 아래와 같이 제안하였다.

$$\text{Studholme's NMI} = \frac{H(X) + H(Y)}{H(X, Y)}$$

이 공식은 두 변수의 엔트로피를 더한 것을 두 변수의 결합 엔트로피로 나눈 것이다. 공식의 분자값에서 결합 엔트로피를 빼면 분자가 평균 상호정보량이 되어 RMIC 계수와 유사한 형태가 된다. 결합 엔트로피가 각각의 엔트로피보다 크거나 같고 두 엔트로피의 합보다는 작거나 같으므로 Studholme의 NMI 공식값은 1에서 2 사이의 범위를 취한다. 현재 의료영상 정합 분야에서는 이 계수가 널리 사용되고 있다(Pluim, Maintz, & Viergever 2003).

2.3 상대적 상호정보량 계수의 특성

연관성을 측정할 때 상호정보량(MI로 표기)을 사용한 경우, Kim & Choi(1999)의 방식대로 최대값 $\log_2 N$ 으로 정규화한 공식(nMI로 표기)을 사용한 경우, 그리고 이 연구에서 제안한 상대적 상호정보량 계수 네 가지를 사용한 경우를 <표 1>에서 비교해보았다. <표 1>을 통해 알 수 있는 상대적 상호정보량 계수 네 가지의 특성은 다음과 같다.

첫째, 분석대상 두 단어의 출현빈도와 동시출현빈도가 모두 동일할 때 최대값 1을 가지게 된다(<표 1>의 사례 ①, ②, ③). 상호정보량도 사례 ①과 같이 두 단어의 출현빈도와 동시출현빈도가 모두 1인 경우에 최대값을 가지지만 사례②, ③과 같이 두 단어의 출현빈도가 높아질수록 0에 가까워진다. 그러나 상대적 상호정보량 계수들은 모두 항상 1이 된다. 다만 사례 ④와 같이 출현빈도와 동시출현빈도가 전체 문헌 수와 같은 경우에는 두 단어의 정보량과 상호정보량이 모두 0이기 때문에

상대적 상호정보량이 정의되지 않는다. 이런 경우에는 사례 ③을 고려하여 1로 평활화하는 것이 바람직하다.

둘째, 두 단어 중 한 쪽이 전체 문헌 수 N 과 같은 출현빈도를 가지는 경우(<표 1>의 사례 ⑥)에는 상호정보량은 0이며 $RMId$ 와 $RMiJ$ 도 0이지만 $RMic$ 와 $RMik$ 는 정의가 안된다. 출현빈도가 전체 문헌 수와 같으면 정보량이 0이 되어 분모가 0이 되기 때문이다. 이때 $RMic$ 와 $RMik$ 는 0으로 평활화해야 할 것이다.

셋째, 최저값의 경우에 <표 1>의 사례 ⑦과 같이 상호정보량은 두 단어의 빈도가 전체 빈도의 1/2이상이며 동시출현빈도가 1이면 가장 낮은 값이 된다. 상호정보량의 최대값으로 정규화한 nMI 계수에서는 이 값이 -1보다 약간 높은 값이 되며 전체 빈도가 커질 수록 -1에 근접하지만 같게 되는 경우는 없다. $RMId$ 계수와 $RMiJ$ 계수는 상호정보량과 같은 경우에 최저값을 가지며, $RMId$ 계수의 최저값은 상호정보량의 최저값과 비슷하게 -1보다 훨씬 낮은 값이지만 $RMiJ$ 계수의 최저값은 nMI

<표 1> 상호정보량과 상대적 상호정보량 계수의 특이값 비교*

사례	f(x)	f(y)	f(x,y)	MI	nMI	$RMik$	$RMic$	$RMId$	$RMiJ$
①	1	1	1	9.9658	1.0000	1.0000	1.0000	1.0000	1.0000
②	500	500	500	1.0000	0.1003	1.0000	1.0000	1.0000	1.0000
③	999	999	999	0.0014	0.0001	1.0000	1.0000	1.0000	1.0000
④	1000	1000	1000	0.0000	0.0000	계산불가	계산불가	계산불가	계산불가
⑤	10	10	0	계산불가	계산불가	계산불가	계산불가	계산불가	계산불가
⑥	1000	10	10	0.0000	0.0000	계산불가	계산불가	0.0000	0.0000
⑦	501	500	1	-7.9687	-0.7996	-7.9802	-7.9802	-7.9802	-0.7996
⑧	999	2	1	-0.9986	-0.1002	-345.9560	-8.7777	-0.2227	-0.1002

* N=1,000인 경우

계수의 최저값과 같으며 -1보다 크다. 한편 공식의 분모에 정보량의 곱이 있는 RMIk 및 RMic 계수는 두 단어의 정보량의 차가 클수록 값이 낮아지므로 사례 ⑧과 같이 한 단어의 빈도는 N-1이고 다른 단어의 빈도가 2이면서 동시출현빈도가 1일 때 최저값을 가진다. RMic 계수의 최저값은 상호정보량의 최저값보다 약간 낮은 정도이지만, RMIk 계수의 최저값은 이보다 훨씬 낮은 값이 되므로 값의 범위를 정규화한다는 측면에서는 바람직하지 못하다고 볼 수 있다.

넷째, 동시출현빈도가 0인 경우(〈표 1〉의 사례 ⑤)에는 상호정보량을 비롯한 모든 계수를 산출할 수가 없게 된다. 이런 경우를 항상 고려대상에서 배제할 수 있는 것은 아니므로 별도의 평활화가 필요하다. 다른 연관성 척도의 경우를 보면 동시출현빈도가 0일 때 코사인 계수나 자카드 계수는 최저값 0을 가지며 동시출현빈도가 1인 경우는 이보다 큰 값을 가지게 되어있다. 따라서 상호정보량이나 상대적 상호정보량 계수도 한 번이라도 동시출현한 경우가 가질 수 있는 최저값보다 작거나 같은 값으로 평활화해야 할 것이다. 최저값이 일정한 nMI 계수나 RMIj 계수의 경우에는 -1로 정하면 무난하지만, RMIk, RMId, RMic 계수는 마땅한 고정값을 부여하기가 어렵다.

다섯째, 상대적 상호정보량 계수값이 음수인지 양수인지는 전적으로 상호정보량에 의해 좌우되므로 상호정보량과 마찬가지로 0보다 낮은 값은 부정적 관계, 높은 값은 긍정적 관계로 간주할 수 있다.

3. 질의확장 정보검색에의 응용 실험

3.1 실험 설계

질의확장 검색은 이용자가 입력한 질의와 연관성이 높은 탐색어를 질의에 추가하여 검색성능을 향상시키는 연관검색 기법의 일종이다. 특히 문헌집단에서 동시출현빈도에 근거하여 미리 구한 색인어간의 유사도를 이용하는 경우를 전역적 질의확장 검색이라고 부른다(Efthimiadis 1996). 이때 색인어간의 유사도를 상호정보량으로 구한 경우가 다른 연관성 척도를 사용한 경우에 비해서 높은 성능을 보이는 것으로 보고된 바 있다(Mandala, Tokunaga, & Tanaka 1998).

여기서는 제안된 상대적 상호정보량 계수를 이용하여 질의확장 검색을 수행한 결과를 기존의 연관성 척도를 이용하는 경우와 비교해 보고자 한다. 비교대상 연관성 척도로는 상호정보량(여기서는 최대값으로 정규화한 nMI 공식) 이외에 질의확장 검색에 자주 사용되는 코사인 계수를 채택하였다. 상대적 상호정보량 계수 중에서는 RMIk, RMId, RMIj를 이용한 질의확장 성능만 비교하였다. RMic는 사전 실험결과 다른 상대적 상호정보량 계수와 비슷한 검색 성능을 보였기 때문에 3.3절의 검색성능 비교분석에서는 단순히 도표상의 식별 문제로 제외하였다.

정확한 검증을 위해서 실험에는 〈표 2〉와 같이 규모와 주제, 언어가 다양한 3개 실험집단을 이용하였다. Medline 실험집단은 미국 국립의학도서관에서 제공한 의학논문 1,033건의 초록으로 구성되었으며, CACM 실험집단

은 *Communication of the Association for Computing Machinery*에 게재된 전산학 분야 논문 3,204건의 초록과 참고문헌을 모은 것이다(김성혁 외 1994). HANTEC-S 실험 집단은 원래 12만 건으로 구성된 HANTEC 2.0 실험집단(김지영 외 2000) 중에서 과학 기술 분야만을 대상으로 한 것이다. 원래 HANTEC 2.0 실험집단 중 과학기술 분야 문헌은 4만 건으로 발표되어 있지만 실제 배포된 CD에는 162건이 누락되어 있어서 39,838 건만을 대상으로 하게 되었다.

각 실험집단을 대상으로 질의마다 동시출현 빈도 기반 유사어를 10개부터 100개까지 10개 씩 늘이면서 추가하여 확장된 질의로 검색을 수행하였다.

검색시스템은 벡터공간모형에 근거해서 구축하였으며 색인어의 가중치 $DW(t_i)$ 는 단어 빈도에 로그를 취한 로그 TF 가중치(Buckley, Salton, & Allan 1993)에 역문헌빈도를 곱한 다음 공식으로 구하였다.

$$DW(t_i) = (1 + \log tf(t_i)) \times \log_2 \frac{N}{DF(t_i)}$$

질의어의 가중치 $QW(t_i)$ 는 로그 TF 가중치로 산출하였다.

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

그리고 문헌과 질의간 유사도 $\text{sim}(D, Q)$ 는 다음과 같이 코사인 계수(Salton 1968)로 산출하였다.

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

초기 질의어와는 별도로 추가된 질의어의 가중치는 질의어의 유사도(각 질의어와의 유사도 평균)로 할당하였다.

3. 2 실험집단에서 연관성 척도의 특성 분석

3. 2. 1 질의어별 유사어 분석

연관성 척도의 특성을 검색실험집단에서 알아보기 위하여 각 실험집단의 질의어를 대상으로 동시출현기반 유사어의 빈도 수준을 분석해보았다. 이때 질의어가 아닌 용어 중에서 문헌빈도(DF)가 1인 용어는 제외하였다. Qiu & Frei(1993)가 지적한 바와 같이 빈도 1인 용어는 질의확장에서 추가질의어로 사용될 경우 성능저하의 원인이 되기 때문이다.

각 실험집단의 질의어를 실험집단에서의 문헌빈도에 따라 각각 7개 그룹으로 나눈 다음, 각 그룹에 속한 질의어와의 동시출현기반 유

〈표 2〉 실험집단의 특성

	Medline	CACM	HANTEC-S
언어	영어	영어	한국어
분야	의학	전산학	과학기술 전반
성격	초록	초록	초록 또는 전문
문헌 수	1,033	3,204	39,838
질의문 수	30	45	30

사도가 가장 높은 용어의 평균 문헌빈도를 <표 3>과 <그림 3>에 제시하였다.

분석 결과 코사인 계수는 질의어가 고빈도어일 때 고빈도어가 유사도 상위에 속하며, 질의어가 저빈도어일 때에는 저빈도어가 높은 유사도를 가진다. 즉, 질의어와 비슷한 수준의 용어가 높은 유사도를 가지는 경향이 있다. 이와 반대로 상호정보량은 질의어의 문헌빈도에 상관없이 빈도 2.5 내외의 저빈도어를 유사어로 취하는 것으로 나타난다. 빈도 1인 용어를 제외한 것을 감안하면 사실상 최저빈도어를 유사어로 선호하는 셈이다.

이와 달리 상대적 상호정보량 공식에 의한 유사어의 문헌빈도는 코사인 계수와 상호정보량에 비해서 중간 정도임을 알 수가 있다. RMId와 RMIj는 유사도 순위를 동일하게 생성하기 때문에 같은 결과가 나타난다.

3. 2. 2 질의별 추가용어 분석

실제 질의확장 검색에서는 Qiu & Frei (1993)가 제안한 것처럼 개별 질의어가 아닌 질의에 포함된 전체 질의어와의 유사도 평균을 기준으로 추가 질의어를 선정한다. 따라서 실제로 질의에 추가되는 용어의 문헌빈도 수

(a) Medline

(b) CACM

(c) HANTEC-S

<그림 3> 질의어의 문헌빈도 수준에 따른 유사도 1위 용어의 문헌빈도

〈표 3〉 질의어의 문헌빈도 수준에 따른 유사도 1위 용어의 문헌빈도

(a) Medline

질의어의 DF	유사도 1위 용어의 DF					
	COS	MI	RMIk	RMIc	RMI _d	RMI _j
1~5	4.41	2.32	3.35	2.94	2.94	2.94
6~15	5.24	2.33	4.61	4.93	4.93	4.93
16~30	17.24	2.59	12.02	12.35	12.69	12.69
31~50	30.94	2.41	7.59	13.19	13.19	13.19
51~80	50.38	2.54	9.58	14.88	17.63	17.63
91~130	96.35	2.18	12.12	18.18	24.71	24.71
131~	194.09	2.36	7.27	24.64	31.91	31.91

(b) CACM

질의어의 DF	유사도 1위 용어의 DF					
	COS	MI	RMIk	RMIc	RMI _d	RMI _j
1~5	3.02	2.50	2.70	2.70	2.70	2.70
6~15	6.50	2.36	4.67	4.67	4.67	4.67
16~30	13.72	2.65	5.77	7.19	8.84	8.84
31~50	21.05	2.54	3.46	3.46	4.95	4.95
51~80	124.50	2.18	5.63	9.63	10.18	10.18
91~130	277.79	2.23	7.14	12.74	14.28	14.28
131~	491.44	2.08	10.33	27.69	71.33	71.33

(c) HANTEC-S

질의어의 DF	유사도 1위 용어의 DF					
	COS	MI	RMIk	RMIc	RMI _d	RMI _j
1~20	3.20	2.85	3.00	3.00	3.20	3.20
21~50	6.60	2.40	2.33	2.73	2.73	2.73
51~150	34.83	2.28	6.72	8.22	9.33	9.33
151~400	323.53	3.53	26.88	23.53	27.00	27.00
401~900	453.17	2.50	36.17	40.61	41.00	41.00
901~1700	1048.11	3.74	48.79	50.74	63.00	63.00
1701~	1703.87	2.09	163.09	164.04	292.04	292.04

준이 연관성 척도마다 어떻게 다른지를 분석해보았다. 각 질의에 대해서 100개씩 확장하되 10개씩 구간을 나누어 각 구간에 포함된 확장 질의어의 평균 문헌빈도를 〈표 4〉와 〈그림 4〉

에 제시하였다.

세 실험집단 모두에서 상호정보량에 의해서 추가되는 추가 질의어의 문헌빈도가 10 내외로 가장 낮으며, 코사인 계수의 경우에는 상당

히 고빈도인 용어를 추가 질의어로 가져옴을 알 수 있다. 상대적 상호정보량 계수에 의한 추가 질의어의 문헌빈도는 모두 상호정보량에 의한 경우보다 높으며 코사인 계수에 의한 경우보다는 낮은 수준을 나타낸다.

결국 상호정보량의 저빈도 편향성이 질의확장 과정에서 저빈도어를 추가하는 결과를 낳

는 반면에, 제안된 네 가지 상대적 상호정보량 계수는 모두 상호정보량보다 다소 높은 빈도의 용어를 질의에 추가한다는 것을 확인하였다. 한편 고빈도 편향성이 강한 코사인 계수는 추가되는 질의어도 매우 고빈도임을 알 수가 있다.

〈표 4〉 추가된 질의어의 추가 구간별 문헌빈도 평균

(a) Medline

연관성 척도	추가 구간									
	1~10	11~20	21~30	31~40	41~50	51~60	61~70	71~80	81~90	91~100
COS	69.16	65.53	64.69	61.99	63.75	55.79	55.45	60.60	58.96	52.47
MI	4.73	5.70	6.60	6.70	7.36	8.26	8.51	10.38	9.15	10.94
RMIk	8.72	9.84	12.17	14.36	14.58	11.91	14.09	14.67	13.54	14.44
RMIc	8.24	10.33	11.55	14.36	12.01	13.65	15.73	14.70	11.74	15.39
RMI _d	9.35	9.10	11.74	13.70	13.06	14.41	14.62	10.48	15.99	13.86
RMI _j	8.35	9.11	7.95	10.59	10.53	10.84	12.71	12.98	12.04	12.18

(b) CACM

연관성 척도	추가 구간									
	1~10	11~20	21~30	31~40	41~50	51~60	61~70	71~80	81~90	91~100
COS	385.2	256.4	179.6	146.9	133.3	131.3	108.8	103.7	100.9	101.7
MI	10.1	13.5	12.0	16.4	18.8	17.0	22.1	20.9	21.3	23.7
RMIk	45.7	55.5	54.5	49.7	39.8	43.1	41.1	48.5	41.9	42.0
RMIc	46.9	58.7	56.7	55.2	51.2	49.7	50.5	46.6	44.1	45.2
RMI _d	54.4	56.1	59.1	61.6	58.7	59.0	52.0	46.7	46.2	52.5
RMI _j	36.1	50.2	50.4	48.9	53.3	56.1	48.6	52.9	51.4	45.0

(c) HANTEC-S

연관성 척도	추가 구간									
	1~10	11~20	21~30	31~40	41~50	51~60	61~70	71~80	81~90	91~100
COS	834.63	1220.50	1273.40	1203.67	1300.28	1294.87	1379.94	1283.17	1244.83	1301.72
MI	5.17	8.15	7.66	8.56	7.67	6.21	9.27	9.76	11.02	12.96
RMIk	16.34	13.33	31.48	21.69	29.46	22.43	38.14	40.18	33.51	40.10
RMIc	18.37	23.74	36.75	31.67	34.96	48.99	36.70	23.41	30.31	30.60
RMI _d	35.74	32.71	38.54	28.51	43.40	43.30	46.49	39.38	60.06	42.29
RMI _j	37.77	18.90	19.42	20.13	21.16	24.81	36.93	30.12	28.61	37.32

(a) Medline

(b) CACM

(c) HANTEC-S

〈그림 4〉 추가된 질의어의 구간별 문헌빈도 평균

3. 3 검색성능 비교분석

검색 성능을 평가하기 위해서 평가 척도로 흔히 사용되는 11지점 평균정확률(11-AP) 이외에 10위내 정확률(P@10), 10위내 순위정확률(RP@10), 30위내 순위정확률(RP@30)을 함께 사용하였다. 11지점 평균정확률 척도는 전반적인 검색결과의 성능을 반영하며, 10위내 정확률과 10위내 순위정확률 척도로는 검색결과 최상위에서의 성능을 알 수 있다. 30위내 순위정확률 척도는 검색결과 중상위에서의 성능을 반영한다. 순위정확률

척도는 1위부터 각 순위에서의 정확률을 모두 더해서 평균한 척도로서 10위내 정확률 척도 보다는 적합문헌의 순위에 더 영향을 받는 척도이다(이재운 2003a).

네 가지 평가 척도를 기준으로 살펴본 세 실험집단에서의 질의확장 검색 성능을 〈표 5〉~〈표 7〉에 제시하였고, 〈그림 5〉~〈그림 7〉에는 동일한 결과를 초기 질의를 사용한 검색 성능 대비 확장 검색의 향상율로 표시하였다.

검색 실험 결과 전반적으로는 세 실험집단 모두에서 상대적 상호정보량 계수를 사용한 경우가 상호정보량을 사용한 경우보다 높게

〈표 5〉 질의확장 검색 성능 - Medline

	연관성 척도	추가용어의 수										
		0	10	20	30	40	50	60	70	80	90	100
P@10	COS	0.6333	0.6700	0.6833	0.6867	0.6900	0.6800	0.6933	0.6967	0.6967	0.6900	0.7000
	MI	0.6333	0.6633	0.6800	0.6800	0.6833	0.7033	0.7033	0.6933	0.6833	0.6833	0.6900
	RMik	0.6333	0.6667	0.6900	0.7100	0.7233	0.7267	0.7200	0.7200	0.7100	0.7000	0.6867
	RMId	0.6333	0.6700	0.6900	0.7033	0.7167	0.7100	0.7067	0.7067	0.6967	0.6933	0.6800
	RMlj	0.6333	0.6867	0.6867	0.7000	0.7100	0.7133	0.7067	0.7100	0.6967	0.6933	0.6800
RP@10	COS	0.7146	0.7501	0.7491	0.7569	0.7636	0.7580	0.7624	0.7641	0.7658	0.7574	0.7592
	MI	0.7146	0.7241	0.7354	0.7458	0.7412	0.7432	0.7391	0.7378	0.7285	0.7291	0.7299
	RMik	0.7146	0.7404	0.7436	0.7519	0.7609	0.7479	0.7458	0.7485	0.7362	0.7353	0.7360
	RMId	0.7146	0.7438	0.7442	0.7533	0.7576	0.7495	0.7397	0.7491	0.7415	0.7265	0.7231
	RMlj	0.7146	0.7549	0.7453	0.7497	0.7474	0.7430	0.7464	0.7408	0.7322	0.7347	0.7340
RP@30	COS	0.5875	0.6175	0.6205	0.6264	0.6314	0.6306	0.6320	0.6325	0.6326	0.6315	0.6338
	MI	0.5875	0.6023	0.6137	0.6193	0.6229	0.6267	0.6314	0.6306	0.6287	0.6292	0.6298
	RMik	0.5875	0.6189	0.6287	0.6328	0.6434	0.6415	0.6450	0.6469	0.6429	0.6406	0.6392
	RMId	0.5875	0.6231	0.6294	0.6379	0.6441	0.6438	0.6396	0.6452	0.6427	0.6369	0.6355
	RMlj	0.5875	0.6253	0.6273	0.6352	0.6382	0.6409	0.6431	0.6448	0.6399	0.6410	0.6382
11-AP	COS	0.5267	0.5762	0.5840	0.5929	0.5990	0.5974	0.5996	0.5988	0.6028	0.5997	0.6031
	MI	0.5267	0.5479	0.5669	0.5783	0.5857	0.5953	0.6046	0.6040	0.6050	0.6031	0.6080
	RMik	0.5267	0.5758	0.6016	0.6066	0.6187	0.6210	0.6262	0.6320	0.6270	0.6245	0.6254
	RMId	0.5267	0.5847	0.6037	0.6119	0.6215	0.6224	0.6193	0.6265	0.6265	0.6214	0.6200
	RMlj	0.5267	0.5836	0.5973	0.6035	0.6098	0.6164	0.6187	0.6237	0.6206	0.6203	0.6179

(a) P@10 향상률

(b) RP@10 향상률

(c) RP@30 향상율

(d) 11-AP 향상율

〈그림 5〉질의확장 검색 성능 향상율 - Medline

〈표 6〉질의확장 검색 성능 - CACM

	연관성 척도	추가용어의 수										
		0	10	20	30	40	50	60	70	80	90	100
P@10	COS	0.2467	0.2733	0.2778	0.2956	0.3067	0.3044	0.3022	0.3067	0.3089	0.3111	0.3133
	MI	0.2467	0.2733	0.2933	0.3067	0.3133	0.3133	0.3067	0.3178	0.3244	0.3244	0.3333
	RMIk	0.2467	0.3111	0.3111	0.3133	0.3467	0.3556	0.3600	0.3600	0.3600	0.3756	0.3711
	RMId	0.2467	0.3156	0.3289	0.3400	0.3422	0.3489	0.3511	0.3600	0.3556	0.3556	0.3622
	RMIj	0.2467	0.2933	0.3156	0.3244	0.3333	0.3378	0.3400	0.3444	0.3400	0.3467	0.3533
RP@10	COS	0.2876	0.3079	0.3197	0.3393	0.3491	0.3472	0.3474	0.3604	0.3665	0.3643	0.3657
	MI	0.2876	0.3312	0.3539	0.3667	0.3750	0.3839	0.3969	0.4042	0.4122	0.4123	0.4145
	RMIk	0.2876	0.3691	0.3946	0.4032	0.4305	0.4338	0.4392	0.4507	0.4554	0.4679	0.4704
	RMId	0.2876	0.3754	0.4034	0.4238	0.4337	0.4428	0.4441	0.4479	0.4536	0.4591	0.4605
	RMIj	0.2876	0.3496	0.3807	0.3918	0.4126	0.4122	0.4346	0.4368	0.4379	0.4379	0.4442
RP@30	COS	0.2384	0.2514	0.2579	0.2674	0.2722	0.2732	0.2746	0.2794	0.2798	0.2807	0.2817
	MI	0.2384	0.2615	0.2753	0.2806	0.2862	0.2901	0.2945	0.2987	0.3036	0.3045	0.3055
	RMIk	0.2384	0.2794	0.2935	0.2989	0.3148	0.3176	0.3234	0.3286	0.3327	0.3379	0.3379
	RMId	0.2384	0.2843	0.2988	0.3113	0.3180	0.3261	0.3274	0.3279	0.3303	0.3320	0.3312
	RMIj	0.2384	0.2717	0.2863	0.2943	0.3043	0.3068	0.3156	0.3179	0.3189	0.3191	0.3210
11-AP	COS	0.2152	0.2317	0.2379	0.2485	0.2529	0.2528	0.2527	0.2573	0.2585	0.2599	0.2604
	MI	0.2152	0.2446	0.2590	0.2682	0.2756	0.2787	0.2833	0.2890	0.2947	0.2943	0.2961
	RMIk	0.2152	0.2626	0.2856	0.2861	0.3048	0.3090	0.3131	0.3162	0.3172	0.3242	0.3276
	RMId	0.2152	0.2706	0.2911	0.2991	0.3035	0.3114	0.3122	0.3142	0.3206	0.3221	0.3283
	RMIj	0.2152	0.2572	0.2736	0.2817	0.2905	0.2992	0.3076	0.3067	0.3096	0.3088	0.3122

(a) P@10 향상율

(b) RP@10 향상율

(c) RP@30 향상율

(d) 11-AP 향상율

〈그림 6〉 질의확장 검색 성능 향상율 - CACM

〈표 7〉 질의확장 검색 성능 - HANTEC-S

	연관성 척도	추가용어의 수										
		0	10	20	30	40	50	60	70	80	90	100
P@10	COS	0.2900	0.3133	0.3200	0.3233	0.3233	0.3333	0.3267	0.3233	0.3233	0.3267	0.3367
	MI	0.2900	0.3167	0.3233	0.3367	0.3333	0.3300	0.3333	0.3367	0.3400	0.3367	0.3300
	RMik	0.2900	0.3267	0.3433	0.3300	0.3467	0.3433	0.3500	0.3433	0.3367	0.3367	0.3333
	RMId	0.2900	0.3300	0.3400	0.3400	0.3467	0.3467	0.3500	0.3500	0.3433	0.3433	0.3400
	RMij	0.2900	0.3400	0.3400	0.3467	0.3433	0.3467	0.3500	0.3400	0.3400	0.3400	0.3533
RP@10	연관성 척도	추가용어의 수										
		0	10	20	30	40	50	60	70	80	90	100
	COS	0.3328	0.3448	0.3594	0.3620	0.3587	0.3591	0.3563	0.3544	0.3584	0.3628	0.3642
	MI	0.3328	0.3714	0.3874	0.3884	0.3877	0.3829	0.3872	0.3875	0.3891	0.3964	0.3940
	RMik	0.3328	0.3601	0.3706	0.3794	0.3775	0.3800	0.3843	0.3773	0.3814	0.3770	0.3618
	RMId	0.3328	0.3629	0.3842	0.3944	0.3853	0.3762	0.3848	0.3898	0.3884	0.3959	0.3957
	RMij	0.3328	0.3729	0.3948	0.4087	0.4081	0.4066	0.4038	0.4100	0.4070	0.3982	0.4103

RP@30	연관성	추가용어의 수										
		척도	0	10	20	30	40	50	60	70	80	90
	COS	0.2649	0.2724	0.2783	0.2810	0.2816	0.2845	0.2836	0.2839	0.2860	0.2890	0.2915
MI	0.2649	0.2872	0.2948	0.3013	0.3067	0.3059	0.3080	0.3106	0.3122	0.3152	0.3144	
RMik	0.2649	0.2921	0.3020	0.3087	0.3122	0.3191	0.3259	0.3243	0.3234	0.3195	0.3137	
RMId	0.2649	0.2935	0.3102	0.3224	0.3294	0.3258	0.3301	0.3286	0.3319	0.3355	0.3343	
RMlj	0.2649	0.3009	0.3094	0.3213	0.3260	0.3279	0.3280	0.3301	0.3286	0.3284	0.3331	
11-AP	연관성	추가용어의 수										
		척도	0	10	20	30	40	50	60	70	80	90
	COS	0.1979	0.2129	0.2201	0.2216	0.2229	0.2262	0.2272	0.2289	0.2309	0.2314	0.2336
	MI	0.1979	0.2313	0.2410	0.2489	0.2666	0.2610	0.2634	0.2648	0.2672	0.2718	0.2730
	RMik	0.1979	0.2355	0.2497	0.2551	0.2651	0.2729	0.2811	0.2816	0.2842	0.2821	0.2796
	RMId	0.1979	0.2272	0.2575	0.2661	0.2751	0.2717	0.2823	0.2836	0.2899	0.2971	0.2963
	RMlj	0.1979	0.2291	0.2523	0.2719	0.2782	0.2830	0.2838	0.2890	0.2887	0.2915	0.2944

(a) P@10 향상을

(b) RP@10 향상을

(c) RP@30 향상을

(d) 11-AP 향상을

〈그림 7〉질의확장 검색 성능 향상을 - HANTEC-S

나타났으며 코사인 계수를 사용한 경우가 가장 낮은 성능을 보였다. 예외적으로 코사인 계수는 Medline 실험집단에서 RP@10 척도로만 가장 좋은 성능을 얻었다. Medline 실험집단은 질의어 자체가 저빈도어 위주로 구성되어 있기 때문에 코사인 계수를 사용한 경우의 성능이 다른 계수를 사용한 경우에 비해서 그리 나쁘지 않은 것으로 보인다.

상대적 상호정보량 계수끼리는 서로 큰 차이가 없었으나, 상대적으로 규모가 작은 Medline과 CACM 실험집단에서는 RMIk와 RMId 계수가 약간 더 좋았으며 규모가 큰 HANTEC-S 실험집단에서는 10위내 순위정확률과 11지점 평균정확률 면에서 RMIj 계수가 조금 더 나은 성능을 보였다.

결과적으로 상대적 상호정보량 계수를 사용한 경우가 상호정보량을 사용한 경우에 비해서 어느 정도 성능 개선 효과가 있는지를 알아보기 위해서 <그림 8>을 제시하였다. 이 그림은 초기 질의어 유사어를 70개 추가하였을 때의 성능을 기준으로 상호정보량을 사용한 경우 대비 향상율을 계산한 것이다. 70개를 추가한 경우를 기준으로 삼은 것은 일부 실험집단과 척도 면에서 지나치게 많이 추가할 경우에 오히려 성능이 저하되는 현상이 있었기 때문이다.

<그림 8>을 보면 HANTEC-S 실험집단에서 10위내 순위정확률을 기준으로 하는 경우 이외에는 모든 상황에서 상대적 상호정보량 계수가 상호정보량보다 높은 성능을 보이는

(a) Medline

(b) CACM

(c) HANTEC-S

<그림 8> 질의어 70개 추가시 상호정보량 대비 RMI 계수의 검색 성능 향상율

것으로 나타났다.

RMlk 계수는 Medline과 CACM 실험집단에서 가장 좋은 성능을 보이지만 HANTEC-S 실험집단에서는 상호정보량보다 10위내 순위 정확률 면에서 성능이 낮게 나타났다. 10위내 정확률은 향상되었음에도 불구하고 10위내 순위정확률이 저하된 것은 1, 2위에 적합문헌이 적게 나타났음을 뜻한다. 이에 비하면 RMId와 RMlj 계수는 상대적으로 안정적인 향상을 보였다.

4. 결 론

상호정보량은 용어간 유사도 산출을 비롯한 다양한 분야에서 연관성 척도로 사용되어왔다. 그러나 일반적인 유사도 함수와 달리 상호정보량은 상한값이 빈도와 반비례하므로 연관성을 측정할 경우에 지나친 저빈도 편향성을 보이는 점이 제한점으로 인식되어왔다.

이런 상호정보량의 제한점을 보완하기 위해서 정보량을 이용하여 상호정보량을 정규화하는 상대적 상호정보량 계수 RMI를 제안하였다. 정보량과 상호정보량의 합성 방식을 달리한 결과 RMlk, RMlc, RMId, RMlj의 네 가지 상대적 상호정보량 계수를 정의할 수 있었다.

단어간의 연관성을 측정하는 상황에서 제안된 상대적 상호정보량 계수를 적용하는 경우를 분석한 결과 다음과 같은 특성을 확인하였다.

첫째, 분석대상 두 단어의 출현빈도와 동시출현빈도가 모두 동일할 때 상대적 상호정보

량 계수는 네 가지 모두 최대값 1을 가진다. 다만 출현빈도와 동시출현빈도가 전체 문헌 수와 같은 경우에는 두 단어의 정보량과 상호정보량이 모두 0이기 때문에 상대적 상호정보량이 정의되지 않으므로 최대값 1로 평활화해야 한다.

둘째, 두 단어 중 한 쪽이 전체 문헌 수 N과 같은 출현빈도를 가지는 경우에는 RMId와 RMlj는 0이 되고 RMlc와 RMlk는 정의가 안된다. 이때 RMlc와 RMlk는 0으로 평활화할 수 있다.

셋째, 최저값의 경우에 RMId 계수와 RMlj 계수는 상호정보량과 같은 경우에 최저값을 가지며, RMId 계수의 최저값은 상호정보량의 최저값과 비슷하게 -1보다 훨씬 낮은 값이지만 RMlj 계수의 최저값은 nMI 계수의 최저값과 같으며 -1보다 크다. 한편 공식의 분모에 정보량의 곱이 있는 RMlk 및 RMlc 계수는 한 단어의 빈도는 N-1이고 다른 단어의 빈도가 2이면서 동시출현빈도가 1일 때 최저값으로 -1보다 낮은 값을 가진다.

넷째, 동시출현빈도가 0인 경우에는 모든 상대적 상호정보량 계수를 산출할 수가 없다. 이런 경우에는 한 번이라도 동시출현한 경우가 가질 수 있는 최저값보다 작거나 같은 값으로 평활화해야 한다.

다섯째, 상호정보량과 마찬가지로 0보다 낮은 값은 부정적 관계, 높은 값은 긍정적 관계로 간주할 수 있다.

이와 같은 특성을 가진 상대적 상호정보량 계수는 상호정보량과 달리 최대값을 1로 고정시키면서 자신과의 유사도가 항상 최대값이 되므로 일반적인 연관성 척도로서의 성질을

만족시킬 수 있었다. 특히 실험집단에서 유사어의 문헌빈도를 조사해본 결과 일부 응용분야에서 상호정보량의 제한점으로 지적되던 지나친 저빈도 편향성이 완화되는 것으로 나타났다.

연관성 척도의 응용분야 중에서 상호정보량이 좋은 성능을 보이는 것으로 알려진 질의 확장 정보검색 분야를 선택하여, 제안된 상대적 상호정보량 계수의 성능을 검증해보았다. 그 결과 상대적 상호정보량 계수를 사용하면 상호정보량을 사용한 경우에 비해서 향상된 검색성능을 얻을 수 있는 것으로 나타났다. 그중에서도 RMI_d와 RMI_j 계수는 안정적인 성능 향상 효과를 보였다.

제안된 네 가지 상대적 상호정보량 계수 중에서 특히 RMI_j 계수는 동시출현 사건의 정보량, 곧 결합정보량 대비 상호정보량의 비율로 구성되므로 수치의 해석 면에서 다른 상대적 상호정보량 계수에 비해 장점을 가진다. RMI_j 계수는 동시출현 사건이 원래 가지는 정보량 중에서 두 사건의 연관성을 나타내는 비율이 어느 정도인가를 알려주는 척도가 되는 것이다.

앞으로 상호정보량을 사용해온 여러 응용분야에 대해서 상대적 상호정보량 계수를 적용해 보는 다양한 추가 연구가 필요할 것으로 생각된다.

참 고 문 헌

- 강현규. 1997. 『자연언어 정보검색에서 상호정보를 이용한 2단계 문서순위 결정 방법』. 박사학위논문, 한국과학기술원.
- 김명철, 이운재, 최기선, 김길창. 1992. “시소러스 작성을 위한 개념 획득 도구”. 제 4회 한글 및 한국어 정보처리 학술대회 논문집, 39-49.
- 김성혁, 서은경, 이원규, 김명철, 김영환, 김재균. 1994. “자동색인기 성능시험을 위한 Test Set 개발”. 정보관리학회지, 11(1): 81-102.
- 김정세. 1996. 『정보 검색에서 상호 정보를 이용한 문서 순위의 재조정』. 석사학위논문, 계명대학교.
- 김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. “한국어 테스트 컬렉션 HANTEC의 확장 및 보완”. 제12회 한글 및 한국어 정보처리 학술대회 논문집, 210-215.
- 김판구. 1994. 『한국어 정보 검색을 위한 상호정보량에 기반한 복합어 자동 색인』. 박사학위논문, 서울대학교.
- 이공주, 김재훈, 김길창. 1995. “품사태깅된 말뭉치로부터 한국어 연어 추출”. 95 가을 한국정보과학회 학술발표논문집, 623-626.
- 이재윤. 2003. “유사계수에 따른 전역적 질의 확장 검색 성능 비교”. 2003 한국정보과학회 가을 학술발표논문집(I), 526-528.

- 이재윤. 2003a. "질문 유형에 따른 인터넷 검색엔진의 성능 비교". 제10회 한국정보관리학회 학술대회 논문집, 185-192.
- 장명길. 2002. 『교차언어 정보검색에서 상호정보를 이용한 사전기반 질의변환』. 박사학위논문, 충남대학교.
- 정석원. 2001. 『격 관계와 상호정보를 이용한 한국어 의존 파서에 관한 연구』. 석사학위논문, 연세대학교.
- Abramson, Norman. 1963. *Information Theory and Coding*. New York: McGraw-Hill Book Company.
- Astola, Jaakko, and Ilkka Virtanen. 1982. "Entropy correlation coefficient: a measure of statistical dependence for categorized data". *Proceedings of the University of Vaasa, Discussion Papers No. 44*, 1982. Quoted in Maes et al. 1997.
- Buckley, Chris, Gerard Salton, and James Allan. 1993. "Automatic retrieval with locality information using SMART". In D. K. Harman (ed.), *Proceedings of the First Text REtrieval Conference (TREC-1)*, (pp. 59-72). NIST Special Publication 500-207.
- 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.
- Church, K. W., and P. Hanks. 1990. "Word association norms, mutual information, and lexicography". *Computational Linguistics*, 16(1): 22-29.
- Efthimiadis, E.N. 1996. "Query expansion". In M.E. Williams (ed.), *Annual Review of Information Science and Technology*, (v.31) (pp.121-187). Information Today, Inc., Medford, NJ.
- Fano, Robert M. 1961. *Transmission of Information: A Statistical Theory of Communications*. New York: M.I.T. Press.
- Kim, M. C., and K. S. Choi. 1999. "A comparison of collocation-based similarity measures in query expansion". *Information Processing & Management*, 35(1): 19-30.
- Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Bent. 1975. *SPSS: Statistical Package for the Social Sciences*. 2nd ed. New York: McGraw-Hill Book Company.
- Maes, F., A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. 1997. "Multimodality image registration by maximization of mutual information". *IEEE Transactions on Medical Imaging*, 16(2) : 187-198.
- Mandala, R., T. Tokunaga, and H.

- Tanaka. 1998. "Query expansion using heterogeneous thesauri". *Information Processing & Management*, 36(3): 361-378.
- Otte, M. 2001. "Elastic registration of fMRI data using Bezier-spline transformations". *IEEE Transactions on Medical Imaging*, 20(3): 193-206.
- Pluim, J. P. W., J. B. A. Maintz, and M. A. Viergever. 2003. "Mutual-information-based registration of medical images: a survey". *IEEE Transactions on Medical Imaging*, 22(8) : 986-1004.
- Qiu, Y., and H. P. Frei. 1993. "Concept based query expansion". *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 160-169.
- Shannon, C. E. 1948. "A mathematical theory of communication". *Bell System Technical Journal*, 27: 379-423, 623-656.
- Salton, Gerard. 1968. *Automatic Information Organization and Retrieval*. New York : McGraw-Hill Book Company.
- Spath, Helmuth. 1980. *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*. Ellis Horwood Limited.
- Strehl, Alexander, and Joydeep Ghosh. 2002. "Cluster ensembles: A knowledge reuse framework for combining multiple partitions". *Journal of Machine Learning Research*, 3(Dec) : 583-617.
- Studholme, Colin. 1997. *Measures of 3D Medical Image Alignment*. Ph. D. thesis, University of London.
- Yang, Yiming, and J. P. Pedersen. 1997. "A comparative study on feature selection in text categorization". *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, 412-420.