문헌 클러스터링을 위한 유사계수간의 연관성 측정

A Measurement of Relationship among Similarity Coefficients for Document Clustering

한승희, 이재윤 (연세대학교 대학원 문헌정보학과)

Seung-Hee Han, Jae-Yun Lee

Dept. of Library and Information Science, Graduate School of Yonsei University

자동분류나 정보검색에 주로 이용되는 문헌 클러스터링에서는 문헌간의 유사성을 측정하기 위해 다양한 유사계수를 이용하는데, 모든 유사계수가 동일한 클러스터링 결과를 가져오는 것은 아니다. 본고에서는 50건의 신문기사를 대상으로 SPSS 통계 패키지를 이용하여 다양한 유사계수에 따라 달라지는 문헌 클러스터링의 결과를 살펴본 후, 유사계수간의 연관성을 측정하였다.

1 서론

문헌의 자동분류나 정보검색에 주로 이용되는 문헌 클러스터링은 문헌에 포함된 단어와 같은 식별요소를 이용하여 유사한 문헌의 클러스터를 형성하는 것이다. 그러므로 클러스터링의 기준은 문헌과 문헌간 혹은 문헌과 클러스터간의 유사도가 된다. 유사도의 측정을 위해서는 식별요소의 벡터를 이용한 유사계수가 사용되며, 다양한 유사계수 공식이 제시되어 있다.

일반적으로, 모든 유사계수가 공통적인 특성을 나타내므로, 클러스터링에 있어 어떠한 유사계수를 선택할 것인가에 대해서는 문제가 되지 않는다고 인식되고 있다(Salton and McGill 1983).

그러나, 실제로 둘 이상의 유사계수를 이용한 실험적 연구에서는 이용된 유사계수에 따라 결과가 달라지는 현상이 나타나기도 했다. 이러한 연구들은 클러스터링 시에 유사계수를 적용하는 일이 무작위로 이루어져서는 안되고, 각 계수가 클러스터링 대상의 특성에 따라 적절하게 사용되어야함을 암시하고 있다.

따라서 이 연구에서는 유사계수에 따라 클러스터링 결과가 달라진다는 가설 하에, 다양한 유사계수를 가지고 클러스터링을 수행하여 그 결과를 분석한 후 유사계수간의 연관성을 측정함으로써 효과적인 문헌 클러스터링을 위한 자료로 삼고자 한다.

2 유사계수의 유형

소칼과 스니스는 유사계수를 거리계수, 연관계수, 상관계수, 확률적 유사계수로 구분하였다(Sokal and Sneath 1973).

거리계수(distance coefficient)는 다양하게 정의된 공간상에서의 거리를 가지고 대상간의 비유사성을 측정하는 방법이다. 대상간의 거리는 유사성의 반대 개념으로, 유사도가 높은 두 문헌간의 거리는 짧다. 대표적인 거리계수로는 유클리드 거리(euclidean distance), 민코스키 메트릭스(mincowski metrics), 시티 블록 거리(city block distance) 등이 있다.

연관계수(association coefficient)는 비교하고자 하는 두 대상을 표현하고 있는 속성간의 일치정도를 측정하는 방법으로, 특히 문헌 클러스터링에 널리 이용되고 있다. 대표적인 연관계수로는 자카드 계수(jaccard coefficient), 다이스 계수(dice coefficient), 허만 계수(hamann coefficient) 등이 있다.

상관계수(corrleation coefficient)는 비교하고자 하는 두 대상을 표현하고 있는 속성들의 벡터쌍에 대한 독립성을 측정하는 방법이다. 대표적인 상관계수로는 적률상관계수(product moment correlation coefficient), 피어슨 상관계수(pearson product moment correlation coefficient) 등이 있다.

확률적 유사계수(probabilistic similarity coefficient)는 정보량 공식에 기반하여 두 사건의 확률변수간의 의존관계를 정량적으로 나타낸 것이며, 대표적으로 새넌(Shannon)의 정보이론에 기초한 상호정보량(mutual information)이 있다.

3 문헌 클러스터링 실험

3.1 실험 대상 및 방법

유사계수간의 연관성을 측정하기 위해서 SPSS 통계 패키지를 이용하여 1997년 6월 A신문 사회면 기사 50건에 대한 문헌 클러스터링 실험을 실시하였다.

신문기사는 그 길이가 일정하지 않기 때문에 단어빈도를 이용하기 위해서는 문헌의 길이를 표준화시킬 필요가 있다. 본 실험에서는 50개 문헌에서 추출된 2,713개 단어의 문헌빈도(DF)를 나타낸 문헌×단어 행렬을 기준으로 문헌길이와 단어빈도를 표준화한 후, 역문헌빈도 가중치를 적용하였다.

우선 각 문헌을 이루는 단어의 수로 각 문헌의 길이(DL)를 측정하여 그 평균값으로 문헌길이를 표준화(WDL)한 후, 다음과 같은 공식을 적용하여 단어빈도를 표준화(WTF)하였다.

$$WTF = TF \times \frac{WDL}{DL}$$

문헌길이와 단어빈도가 표준화된 행렬에서 역문헌빈도 가중치(WIDF)를 아래와 같이 계산하여 문헌×단어 행렬을 새롭게 작성하였다.

$$WIDF = \frac{WTF}{DF}$$

SPSS 통계 패키지의 계층적 클러스터링 프로시저에서 제공하는 35개의 계수 중 33개의 계수에 가중치를 부여한 문헌×단어 행렬을 적용하였다. 2개의 계수는 문헌×단어 행렬의 요소값에 0이 포함되어 있는 경우에는 적용할 수 없는 공식이었으므로 제외하였다.

이 때 클러스터링 알고리듬은 완전연결기법을 적용하는데, 완전연결기법이란 기존의 클러스터에 포함되어 있는 모든 문헌에 대하여 일정거리 이내에 존재해야만 동일한 클러스터에 포함시키는 방법을 의미한다. 클러스터간의 거리는 각 클러스터 내에 속해있는 문헌간의 가장 먼 거리로 산정된다.

3.2 실험 결과 분석

다양한 계수를 적용하여 클러스터를 생성한 결과를 분석하기 위해서는 클러스터 생성 기준점을 결정해야 한다.

일반적으로 군집분석에서 클러스터의 수를 결정하는 절대적인 방법은 없으며, 하나의 클러스터링 결과에 대해 클러스터의 수를 결정하기 위해서는 계수가 갑자기 큰 폭으로 증가하기 전 단계에서 클러스터 형성을 종료시키는 방법을 이용한다.

그러나, 본 실험에서는 하나의 유사계수에 대한 클러스터링 결과가 아닌 다양한 유사계수에 대한 클러스터링 결과를 기준으로 유사계수간의 연관성을 살펴보아야 하므로, 각각의 클러스터링 결과에 대해 클러스터의 수를 결정할 수 있는 절대적인 기준이 필요하다.

본 실험에서는 각 클러스터링 결과에 대한 계수 값의 변화폭을 관찰하여, 덴드로그램에 나타난 계수의 표준값이 17점이 되는 지점을 클러스터 형성 기준점으로 결정하였다.

4 유사계수간의 연관성 측정

4.1 유사계수의 선정

유사계수간의 연관성을 측정하기 전에, 클러스터를 1개만 생성한 유사계수는 연관성 측정 대상에서 제외되었다. 결과적으로, <표 1>과 같이 33개의 유사계수 중 2개 이상의 클러스터를 생성한 18개의 유사계수만이 연관성 측정 대상이 되었다.

<표 1> 유사계수간의 연관성 측정 대상

유사계수	약칭	유사계수	약칭
블록 (Block)	blo	쿨친스키 공식 2 (Kulczynski 2)	kul2
카이제곱 (Chi square)	chi	러셀과 라오 공식 (Russel and Rao)	r&r
피어슨 상관계수 (Pearson correlation)	pea	랜스와 윌리엄즈 공식 (Lance and Williams)	l&w
코사인계수 (Cosine)	cos	오치아이 공식 (Ochiai)	och
분산도 (Dispersion)	dis	소칼과 스니스 공식 2 (Sokal and Sneath 2)	s&s2
크기 차이 (Size difference)	siz	소칼과 스니스 공식 4 (Sokal and Sneath 4)	s&s4
Phi 4-point correlation	phi	소칼과 스니스 공식 5 (Sokal and Sneath 5)	s&s5
자카드계수 (Jaccard)	jac	율의 Y (Yule's Y)	yuly
다이스계수 (Dice)	dic	율의 Q (Yule's Q)	yulq

<표 2>에 18개의 유사계수를 적용하여 클러스터링한 결과를 제시하였다.

4.2 유사계수간의 연관성 측정

문헌 클러스터링 결과를 가지고 계수간의 연관성을 비교하기 위해서 다음과 같이 클러스터링 유사도 C_SIM을 정의하였다.

우선 비교 대상인 두 클러스터링 결과 C_A와 C_B에 나타난 클러스터들로부터 한 클러스터에 속한 단어들의 쌍을 모두 추출하여 전체 단어쌍 집합 U를 만든다. 집합 U에 포함된 단어쌍들을 어느 클러스터링 결과의 클러스터에서 나타난 경우인가에 따라서 다음과 같은 2×2 분할표에서의 a, b, c 집합으로 나눌 수 있다.

		C_A	
		두 단어가 동일한 클러스터에 포함된 단어쌍	두 단어가 상이한 클러스터에 포함된 단어쌍
C_B	두 단어가 동일한 클러스터에 포함된 단어쌍	a	b
	두 단어가 상이한 클러스터에 포함된 단어쌍	c	

이 분할표에 대해 다이스계수 공식을 적용하면 다음과 같다.

$$C\_SIM(C_A, C_B) = \frac{2a}{2a+b+c}$$

유사계수간의 연관성 측정을 위한 이러한 C_SIM 클러스터링 유사도는 한 클러스터링 결과에서 같은 클러스터에 포함된 임의의 단어쌍이 다른 클러스터링 결과에서도 같은 클러스터에 포함될 확률을 의미한다.

<표 2> 유사계수를 달리하여 완전연결기법을 적용한 클러스터링 결과

	blo	chi	pea	cos	dis	siz	phi	jac	dic	kul2	r&r	l&w	och	s&s2	s&s4	s&s5	yuly	yulq
생성된 클러스터 수	2	4	9	7	17	2	12	11	11	10	4	12	11	12	12	11	5	8
클러스터에 포함된 전체 문헌 수	36	33	48	16	44	50	36	25	26	21	10	28	28	26	31	26	50	50
클러스터당 평균 문헌 수	18.00	8.25	5.33	2.29	2.59	25.00	3.00	2.27	2.36	2.10	2.50	2.33	2.55	2.17	2.58	2.36	10.00	6.25
클러스터에 포함된 문헌 수의 표준편차	18.38	11.18	7.09	0.49	0.94	31.11	0.74	0.47	0.50	0.32	0.58	0.49	0.69	0.39	0.79	0.50	9.80	3.69

<표 3> 유사계수간의 관계성을 나타낸 상관표

	blo	chi	pea	cos	dis	siz	phi	jac	dic	kul2	r&r	l&w	och	s&s2	s&s4	s&s5	yuly	yulq
blo	1.000	0.785	0.241	0.045	0.112	0.502	0.113	0.065	0.065	0.049	0.033	0.069	0.088	0.061	0.091	0.069	0.357	0.245
chi	0.785	1.000	0.131	0.069	0.109	0.347	0.133	0.074	0.074	0.069	0.051	0.080	0.121	0.080	0.119	0.092	0.240	0.251
pea	0.241	0.131	1.000	0.070	0.150	0.431	0.157	0.087	0.086	0.063	0.045	0.086	0.079	0.075	0.102	0.080	0.336	0.281
cos	0.045	0.069	0.070	1.000	0.377	0.011	0.400	0.571	0.533	0.609	0.737	0.516	0.514	0.593	0.308	0.600	0.042	0.095
dis	0.112	0.109	0.150	0.377	1.000	0.064	0.420	0.475	0.525	0.333	0.280	0.516	0.424	0.414	0.371	0.459	0.144	0.244
siz	0.502	0.347	0.431	0.011	0.064	1.000	0.057	0.024	0.027	0.015	0.005	0.029	0.031	0.022	0.040	0.027	0.502	0.255
phi	0.113	0.133	0.157	0.400	0.420	0.057	1.000	0.536	0.586	0.392	0.340	0.576	0.540	0.473	0.627	0.517	0.154	0.321
jac	0.065	0.074	0.087	0.571	0.475	0.024	0.536	1.000	0.944	0.552	0.480	0.919	0.683	0.909	0.489	0.778	0.074	0.163
dic	0.065	0.074	0.086	0.533	0.525	0.027	0.586	0.944	1.000	0.516	0.444	0.974	0.651	0.857	0.553	0.737	0.083	0.182
kul2	0.049	0.069	0.063	0.609	0.333	0.015	0.392	0.552	0.516	1.000	0.500	0.500	0.556	0.571	0.550	0.645	0.047	0.105
r&r	0.033	0.051	0.045	0.737	0.280	0.005	0.340	0.480	0.444	0.500	1.000	0.429	0.500	0.500	0.278	0.519	0.033	0.075
l&w	0.069	0.080	0.086	0.516	0.516	0.029	0.576	0.919	0.974	0.500	0.429	1.000	0.636	0.889	0.542	0.718	0.082	0.18
och	0.088	0.121	0.079	0.514	0.424	0.031	0.540	0.683	0.651	0.556	0.500	0.636	1.000	0.700	0.538	0.884	0.100	0.217
s&s2	0.061	0.080	0.075	0.593	0.414	0.022	0.473	0.909	0.857	0.571	0.500	0.889	0.700	1.000	0.409	0.800	0.065	0.144
s&s4	0.091	0.119	0.102	0.308	0.371	0.040	0.627	0.489	0.553	0.550	0.278	0.542	0.538	0.409	1.000	0.468	0.108	0.232
s&s5	0.069	0.092	0.080	0.600	0.459	0.027	0.517	0.778	0.737	0.645	0.519	0.718	0.884	0.800	0.468	1.000	0.078	0.172
yuly	0.357	0.240	0.336	0.042	0.144	0.502	0.154	0.074	0.083	0.047	0.033	0.082	0.100	0.065	0.108	0.078	1.000	0.601
yulq	0.245	0.251	0.281	0.095	0.244	0.255	0.321	0.163	0.182	0.105	0.075	0.181	0.217	0.144	0.232	0.172	0.601	1.000

18개의 유사계수가 생성해낸 클러스터 결과에 이 공식을 적용하여 <표 3>과 같은 상관표를 얻었다.

위의 상관표를 가지고 Ward 클러스터링 기법을 이용하여 유사계수를 분류해보면 <표 4>와 같다.

<표 4> 연관성에 따른 유사계수의 분류

클러스터 번호	클러스터에 포함된 계수
1	dic, l&w, jac, s&s2, och, s&s5
2	cos, r&r, kul2
3	phi, s&s4, dis
4	blo, chi
5	yuly, yulq, pea, siz

5 분석 및 결론

본 실험을 통해 얻게 된 결과를 요약하면 다음과 같다.

첫째, 계수가 달라짐에 따라 클러스터링 결과가 다르게 나타난다. <표 2>의 클러스터 수를 예로 살펴보면, 최소 2개에서 최다 17개까지로 유사계수에 따른 편차가 크게 나타났다.

둘째, 모든 유사계수가 문헌 클러스터링에 적절한 것은 아니다. 실험 결과, 실제로 33개의 유사계수 중 2개 이상의 클러스터를 형성한 계수는 18개에 지나지 않았다.

셋째, 대부분의 거리계수는 문헌 클러스터링에 적절하지 못했다. 실험 결과, 유클리드 거리, 제곱 유클리드 거리, 민코스키 공식 등과 같은 거리계수는 전체 문헌을 하나의 클러스터에 포함하여 문헌 클러스터링에 적합하지 않은 결과를 보여주고 있다.

넷째, 연관성에 따라 유사계수를 분류해 본 결과, 일반적으로 많이 쓰이는 지카드계수와 코사인계수간의 연관성이 낮게 나타났다.

클러스터링 대상에 적절한 유사계수를 이용할 경우 더 나은 클러스터링 결과를 얻을 수 있을 것이다. 본 연구를 바탕으로 하여 계수에 따른 클러스터링의 성능을 분석하고, 이를 정보검색 및 질의확장, 자동분류 등에 적용하는 후속 연구가 필요하다고 본다.

참고문헌

Salton, G., and M. J. McGill. 1983. *Introduction to Modern Information Retrieval.* NY: McGraw-Hill Book Company.

Sneath, P. H. A., and Robert R. Sokal. 1973. *Numerical Taxonomy: The Principles and Practices of Numerical Classification.* SF: W. H. Freeman and Company.