분포 유사도를 이용한 문헌클러스터링의 성능향상에 대한 연구*

Improving the Performance of Document Clustering with Distributional Similarities

이 재 윤 (Jae Yun Lee)**

초 록

이 연구에서는 분포 유사도를 문헌 클러스터링에 적용하여 전통적인 코사인 유사도 공식을 대체할 수 있는 가능성을 모색해보았다. 대표적인 분포 유사도인 KL 다이버전스 공식을 변형한 Jansen-Shannon 다이버전스, 대칭적 스큐 다이버전스, 최소 스큐 다이버전스의 세 가지 공식을 문헌 벡터에 적용하는 방안을 고안하였다. 분포 유사도를 적용한 문헌 클러스터링 성능을 검증하기 위해서 세 실험 집단을 대상으로 두 가지 실험을 준비하여 실행하였다. 첫 번째 문헌 클러스터링 실험에서는 최소 스큐 다이버전스가 코사인 유사도 뿐만 아니라 다른 다이버전스 공식의 성능도 확연히 앞서는 뛰어난 성능을 보였다. 두 번째 실험에서는 피어슨 상관계수를 이용하여 1차 유사도 행렬로부터 2차 분포 유사도를 산출하여 문헌 클러스터링을 수행하였다. 실험 결과는 2차 분포 유사도가 전반적으로 더 좋은 문헌 클러스터링 성능을 보이는 것으로 나타났다. 문헌 클러스터링에서 처리 시간과 분류 성능을 함께 고려한다면 이 연구에서 제안한 최소 스큐 다이버전스 공식을 사용하고, 분류 성능만 고려할 경우에는 2차 분포 유사도 방식을 사용하는 것이 바람직하다고 판단된다.

ABSTRACT

In this study, measures of distributional similarity such as KL-divergence are applied to cluster documents instead of traditional cosine measure, which is the most prevalent vector similarity measure for document clustering. Three variations of KL-divergence are investigated; Jansen-Shannon divergence, symmetric skew divergence, and minimum skew divergence. In order to verify the contribution of distributional similarities to document clustering, two experiments are designed and carried out on three test collections. In the first experiment the clustering performances of the three divergence measures are compared to that of cosine measure. The result showed that minimum skew divergence outperformed the other divergence measures as well as cosine measure. In the second experiment second-order distributional similarities are calculated with Pearson correlation coefficient from the first-order similarity matrixes. From the result of the second experiment, second-order distributional similarities were found to improve the overall performance of document clustering. These results suggest that minimum skew divergence must be selected as document vector similarity measure when considering both time and accuracy, and second-order similarity is a good choice for considering clustering accuracy only.

키워드: 분포유사도, 다이버전스, 2차 유사도, 문헌 클러스터링, 자동분류, distributional similarity, divergence, second-order similarity, document clustering, automatic classification

---

* 본 연구는 2006학년도 경기대학교 학술연구비(일반연구과제) 지원에 의하여 수행되었음.

** 경기대학교 인문대학 문헌정보학전공 조교수 (memexlee@kgu.ac.kr)

■ 논문접수일자

■

1. 서론

문헌 클러스터링은 각 문헌을 표현하는 자질들을 비교하여 문헌간 유사성을 측정한 다음 비슷한 내용의 문헌들을 동일한 집단에 속하도록 군집화하는 기법이다(정영미 2005). 최근까지 다양한 클러스터링 기법이 제안되어왔지만, 생성된 클러스터의 품질 면에서는 계층적 클러스터링 모형이 가장 뛰어난 것으로 알려져 있다. 클러스터링 모형은 대상 항목의 선정, 분류자질의 빈도행렬 작성, 유사계수의 적용, 클러스터 생성 기법의 적용 등 여러 단계로 구성된다. 각 단계마다 다양한 경우의 수가 있으므로 클러스터링 결과도 적용한 모형에 따라서 달라지게 된다(정영미, 이재윤 2001).

이 연구에서는 유사계수의 적용 단계에 특히 집중하되, 전통적으로 문헌 클러스터링에 사용되어온 코사인 계수(Salton & McGill 1983)를 비롯한 벡터유사도 방식을 벗어나서 다이버전스(divergence)를 비롯한 분포 유사도 방식을 적용하여 분류 성능을 개선하고자 한다. 분포 유사도 방식은 두 확률 분포 사이의 차이를 측정하여 거리나 유사성을 판단하는 것이다. 이미 용어 클러스터링 분야에서는 전통적인 코사인 유사도 방식을 사용하지 않고, 확률분포간의 차이를 분석하는 분포 유사도를 이용해서 좋은 성과를 얻은 연구가 여러 건 발표되었다(Dagan & Lee 1999; Lee 1999; Lin 1998; Pereira, Tishby, Lee 1993; Weeds 2003). 그러나 용어 클러스터링이 아닌 문헌 클러스터링에 분포 유사도를 적용한 연구는 없었다.

일반적으로 문헌이나 용어를 분류하는 문제에서는 개념적으로 문헌-용어 행렬을 구성하여 문헌간 유사도나 용어간 유사도를 산출하게 된다. 이는 동일한 행렬을 상이한 관점에서 보는 것에 해당한다. 따라서 용어에 대해서 적용한 분포 유사도는 동일한 방식을 관점을 바꾸어 문헌에 대해서 적용해볼 수도 있다. 문헌 클러스터링에 분포 유사도를 적용하여 기존의 코사인 유사도 방식을 적용한 경우보다 높은 성능을 얻는 것이 이 연구의 첫 번째 목표이다.

분포 유사도 방식을 이용한 문헌 클러스터링의 성능을 검증한 후에는 다음 단계로 2차 분포 유사도 산출 방식을 적용해보고자 한다. 2차 분포 유사도란 1차적으로 산출된 유사도 값으로 구성된 행렬에서 각 행(혹은 열)간의 유사도를 다시 산출한 것이다. 이는 White와 Griffith(1981)가 저자동시인용 분석에서 저자동시인용빈도행렬로부터 상관계수행렬을 산출한 것과 같은 방식이다. 문헌 클러스터링을 위해서는 문헌간 유사도를 산출하여 1차 유사도 행렬을 산출한 다음, 이 행렬에 대해서 다시 피어슨 상관계수를 적용하여 2차 유사도를 산출하게 된다. 이와 같은 2차 유사도도 두 문헌 간의 유사도 분포를 비교하므로 일종의 분포 유사도에 해당한다. 즉, 문헌 A와 문헌 B가 각자 유사하고 유사하지 않은 문헌이 서로 비슷하다면 문헌 A와 문헌 B 사이에도 깊은 연관성이 있다고 볼 수 있는 것이다. 이를 달리 생각하면, 동일한 주제의 상이한 측면을 다룬 두 문헌은 1차 유사도가 높게 나타나지는 않지만, 제 삼의 문헌과의 유사한 정도가 비슷함에 따라서 가까운 문헌이라고

판단할 수 있게 되는 것이다. 다만 2차 분포 유사도는 산출 시간이 추가되므로 설사 성능이 좋더라도 검색결과 문헌과 같은 소수의 문헌집단에 대해서 적용하는 것이 바람직하다. 이와 같이 2차 분포 유사도 산출 방식을 적용하여 문헌 클러스터링 성능을 향상시키는 것이 이 연구의 두 번째 목표이다.

이 논문의 2장에서는 다이버전스 공식으로 문헌간 유사도를 산출하는 방법을 제안하고, 3.1절에서는 제안한 방법으로 분포 유사도를 적용하는 클러스터링 실험을 수행하여 기존의 코사인 유사도를 적용한 경우와 성능을 비교해보았다. 또한 3.2절에서는 2차 분포 유사도 산출 방식을 적용하여 문헌 클러스터링 성능을 향상시킬 수 있는지 여부에 대해서 실험을 수행한 다음 4장에서 분석 결과를 정리하고 결론을 제시하였다.

2. 다이버전스와 문헌 유사도

2.1 다이버전스 공식

일반적으로 분포 유사도 척도라는 표현은 다이버전스 공식을 일컫는데 사용된다. 다이버전스는 확률분포간의 차이를 측정하는 방법으로서 정보이론에서 출발한 Kullback-Leibler 다이버전스(KL-Divergence; Kullback 1968; Kullback & Leibler 1951)가 대표적이다. Kullback-Leibler 다이버전스는 Kullback-Leibler 거리라고도 부르며(Theodoridis & Koutroumbas 2003) 흔히 KL 다이버전스라고 줄여서 표기한다. 두 이산형 확률분포 q(y)와 r(y) 사이의 차이를 구하는 KL 다이버전스 공식은 다음과 같다.

$$D(q\|r) = \sum_y q(y)(\log q(y) - \log r(y)) = \sum_y (q(y)\log\frac{q(y)}{r(y)})$$

KL 다이버전스의 값의 범위는 0 이상이며 두 확률분포가 같으면 0이 되고 다를수록 값이 커진다. 또한 KL 다이버전스는 비대칭 공식으로서 두 확률분포의 계산순서에 따라서 결과가 달라지게 된다. 대칭적인 값을 얻기 위해서는 다음과 같이 KL 다이버전스를 순서를 달리하여 두 번 구한 후 합산한 다이버전스 *J*를 사용할 수 있다(Kullback 1968).

$$J(q,r) = D(q\|r) + D(r\|q)$$

그런데 KL 다이버전스는 확률에 로그를 취하기 때문에 비교 대상 중 두 번째 확률분표 r(y)에서 확률이 0인 지점에서는 값이 정의되지 않는다. 따라서 용어 클러스터링에 적용할 경우에는 비교대상인 두 확률분포 중 어느 한 쪽이 0이더라도 다이버전스를 산출할 수 있도록 수정된 Jensen-Shannon 다이버전스나 스큐(skew) 다이버전스가 사용된다.

Jensen-Shannon 다이버전스는 두 확률분포를 직접 비교하지 않고 개별 지점마다 두 확률분포의 값을 평균하여 얻어낸 분포와 각 확률분포 사이의 KL 다이버전스를 산출한 후 두 다이버전스의 평균을 구하는 방법이다(Lin 1991). 개별 지점에서 두 확률분포의 값을 평균하면 최소한 한 확률분포의 값은 0이 아니기 때문에 평균 또한 0이 아님이 보장된다.

따라서 아래 공식과 같이 평균하여 도출한 확률분포 avg(q,r)를 두 번째 항으로 두고, 원래의 두 확률분포를 번갈아서 첫 번째 항으로 삼아서 KL 다이버전스를 두 차례 산출한 후 두 값의 평균을 구하여 최종 다이버전스를 계산한다.

$$JS(q,r) = \frac{1}{2}[D(q||avg(q,r)) + D(r||avg(q,r))]$$

이론적으로는 두 확률분포가 가까울수록 둘을 평균한 분포와 각 확률분포 사이도 가까울 것이기 때문에 Jensen-Shannon 다이버전스는 KL 다이버전스와 비례하며, 두 확률분포의 순서를 바꾸더라도 결과가 같은 대칭 공식이다.

스큐 다이버전스는 Jensen-Shannon 다이버전스와 마찬가지로 KL 다이버전스와 비례하며 모든 지점에서 값이 산출되도록 하되, KL 다이버전스처럼 비대칭 공식이 되도록 고안된 것이다. 스큐 다이버전스를 산출하는 공식은 다음과 같다(Lee 1999).

$$s_\alpha(q,r) = D(r||\alpha \cdot q + (1-\alpha) \cdot r)$$

여기서 α는 0에서 1 사이의 값을 가지는 파라미터로서 α=1이면 스큐 다이버전스는 KL 다이버전스가 된다. Lee(2001)는 α값으로 여러 경우를 실험해본 결과 α=0.99로 설정하는 경우가 가장 좋았다고 보고하였다. 스큐 다이버전스에서 α=0.99로 설정하면 거의 KL 다이버전스와 유사하면서도 확률분포에서 어느 한 쪽이 0인 지점이 있더라도 값을 산출할 수 있는 공식이 된다. 이 연구에서도 Lee(1999)와 마찬가지로 α=0.99로 설정한 스큐 다이버전스를 사용하였다.

2.2 문헌 간의 다이버전스 측정

다이버전스를 문헌 클러스터링에 적용할 경우에는 주어진 두 문헌 내에서 각 용어의 확률 분포를 구한 후 이를 비교해야 한다. 각 용어의 확률은 주어진 문헌에 출현한 모든 용어의 가중치의 합으로 특정 용어의 가중치를 나누어서 산출할 수 있다.

이 연구에서는 두 문헌 d_i과 d_j에서 용어 t_k의 가중치가 각각 $w(t_k,d_i)$, $w(t_k,d_j)$이고 문헌 d_i의 길이(출현한 용어의 가중치 합)가 $|d_i|$일 때, 문헌 d_i와 d_j 사이의 KL 다이버전스 $D(d_i||d_j)$를 다음과 같이 나타내었다.

$$\begin{aligned} D(d_i||d_j) &= \sum_k \frac{w(t_k,d_i)}{|d_i|}\left(\log\frac{w(t_k,d_i)}{|d_i|} - \log\frac{w(t_k,d_j)}{|d_j|}\right) \\ &= \sum_k \frac{w(t_k,d_i)}{|d_i|} \log\frac{w(t_k,d_i)\ |d_j|}{w(t_k,d_j)\ |d_i|} \\ &= \frac{1}{|d_i|}\sum_k w(t_k,d_i) \log\frac{w(t_k,d_i)\ |d_j|}{w(t_k,d_j)\ |d_i|} \end{aligned}$$

앞에서 살펴본 KL 다이버전스의 특성 때문에 이 공식은 문헌 d_i에는 출현하였으나 문헌 d_j에는 미출현한 용어가 있으면 산출이 불가능하다. 그런데 이는 거의 모든 경우에 해당하므로 실제 문헌간 다이버전스를 측정하기 위해서는 Jenson-Shannon 다이버전스나 스큐 다이버전스를 사용하여야 한다.

클러스터링을 위한 문헌간 거리(혹은 유사도)는 두 문헌 중 어느 쪽을 기준으로 하더라

도 같은 값이어야 하므로 대칭 공식인 Jenson-Shannon 다이버전스는 그대로 사용할 수 있지만, 비대칭 공식인 스큐 다이버전스는 대칭적인 값을 얻어내기 위한 별도의 조치가 필요하다. 이 연구에서는 두 문헌간 대칭적인 스큐 다이버전스를 산출하기 위해서 다음의 두 가지 방식을 고안하였다.

첫째, 대칭적 스큐 다이버전스 SSD(d_i,d_j)를 다음 공식과 같이 정의하여 사용한다. 이는 비교대상 두 확률분포의 계산 순서를 달리하여 두 가지 스큐 다이버전스를 구한 후 합산한 것이다. 양방향 다이버전스를 합하여 대칭적 공식을 도출하는 것은 Kullback(1968)이 양방향 KL 다이버전스를 합하여 대칭적 공식인 다이버전스 J를 도출한 것과 같은 방법이다. 따라서 대칭적 스큐 다이버전스 SSD는 비교대상인 두 확률분포 중 어느 한 쪽이 0이더라도 산출할 수 있도록 Kullback의 다이버전스 J를 수정한 공식이라고 할 수도 있다.

$$SSD(d_i,d_j) = s_{0.99}(d_i,d_j) + s_{0.99}(d_j,d_i)$$

둘째, 최소 스큐 다이버전스 MSD(d_i,d_j)를 다음 공식과 같이 정의하여 사용한다. 이는 양방향 스큐 다이버전스 중에서 작은 값을 취한 것이다.

$$MSD(d_i,d_j) = \min(s_{0.99}(d_i,d_j), s_{0.99}(d_j,d_i))$$

최소 스큐 다이버전스 MSD는 한 문헌의 관점에서 다른 문헌이 가깝게 판단된다면 다른 문헌의 관점은 무시하도록 설정한 것이다. 문헌과 문헌을 비교할 때에는 포함관계 등의 이유로 한 문헌의 내용이 다른 문헌의 일부와 매우 유사할 경우가 있으며, 이때에는 한 쪽의 관점에서 산출한 다이버전스만으로도 두 문헌이 유사한 것을 나타낼 수 있기 때문이다.

3. 분포 유사도를 이용한 문헌 클러스터링 실험

3.1 실험 설계

분포 유사도를 적용한 클러스터링 성능을 검증하기 위해서 <표 1>과 같이 실험 문헌집단을 구성하였다. 각 실험집단은 실험질의와 적합문헌이 정해진 정보검색 실험용 문헌집단에서 추출한 것으로서, 각 질의를 하나의 주제범주로 보고 질의에 대한 적합문헌을 주제범주에 소속한 문헌으로 간주하였다. 이와 같이 검색 질의와 적합문헌으로 분류범주와 소속문헌 집단을 구축한 것은, 최근 문헌 클러스터링 연구가 검색결과 문헌의 클러스터링을 주요 과제로 하고 있는 점도 감안한 것이다.

실험집단 중 MQ는 Medline 실험집단에서 추출하였다. Medline 실험집단은 의학논문 초

<표 1> 실험집단의 구성

실험집단	범주 수	문헌 수	범주당 문헌 수	범주크기 변이계수
CQ	8	263	32.9	0.331
MQ	8	276	34.5	0.150
HQ	5	351	70.2	0.367

록 1,033개로 구성되었고, 이중에서 696개의 문헌이 30개의 질의에 적합한 것으로 나뉘어져 있다. 30개 질의 중에서 적합문헌 수가 많은 순서대로 8개 질의를 대분류 범주로, 이 질의들의 적합문헌 276개를 소속문헌으로 추출하여 분류실험집단 MQ를 구성하였다.

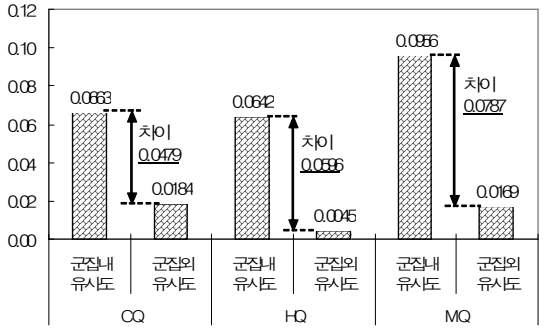
CQ는 컴퓨터과학 분야 초록으로 구성된 CACM 실험집단에서 추출한 것이다. 적합문헌 수가 많은 순서대로 8개 질의와 이 질의들의 적합문헌 263개를 각각 범주와 소속문헌으로 추출하여 분류실험집단 CQ를 구성하였다. 원래 CACM 실험집단에서 추출한 적합문헌들 중에는 둘 이상의 질의에 적합한 복수주제 문헌이 있으나 배타적 클러스터링 실험을 위해서 제거하고 CQ를 구성하였다.

HQ는 한국어 정보검색 실험집단인 HANTEC 실험집단에서 추출하였다. HANTEC에는 약 4만 여건의 문헌들을 대상으로 질의 30여개에 대한 12만 건의 적합성 판정이 이루어져있다. 적합문헌 수가 많은 순서대로 5개 질의와 이 질의들의 적합문헌 351개를 각각 범주와 소속문헌으로 추출하여 분류실험집단 HQ를 구성하였다.

세 실험집단에서 군집내 문헌간 코사인 유사도와 타 군집에 속한 문헌과의 코사인 유사도를 각각 평균해보면 <표 2> 및 <그림 1>과 같다. 실험집단에서 군집내 문헌간 유사도가 높고 타 군집 소속 문헌과의 유사도가 낮을수록 분류가 잘 될 가능성이 높다. 군집내 문헌간 유사도는 MQ가 가장 높으며 CQ와 HQ는 비슷하다. 타 군집 소속 문헌과의 유사도는 CQ가 가장 높으며 MQ가 이보다 조금 낮고 HQ는 매우 낮다. 결국 두 평균 유사도 간의 차이는 MQ가 가장 크며 HQ가 그 다음이고 CQ가 가장 작다. 이로 미루어볼 때 분류 문제로서는 세 집단 중에서 CQ가 가장 어렵고 MQ가 가장 쉬운 집단이라고 할 수 있다.

<표 2> 실험집단별 군집내/외 평균유사도

실험집단 / 평균유사도 측정 대상	CQ	HQ	MQ
① 동일 군집내 문헌간	0.0663	0.0642	0.0956
② 타 군집 소속 문헌간	0.0184	0.0045	0.0169
① - ②	0.0479	0.0596	0.0787



<그림 1> 실험집단별 군집내/외 평균유사도

실험에서 사용한 클러스터링 기법은 널리 사용되고 있는 네 가지 계층적 클러스터링 기법 중에서 단일연결 기법을 제외한 평균연결 기법, 완전연결 기법, Ward 기법의 세 가지를 적용하였다. 단일연결기법은 문헌 클러스터링에 적용하였을 때 성능이 매우 나쁘다고 알려져 있기 때문에 제외하였다(Griffith et al. 1984; Griffith et al. 1986).

비교 대상 유사도 산출 방식은 다음과 같이 다섯 가지로 구성하였다.

① COS : 전통적인 코사인 유사도 계수를 적용한 경우로서 다이버전스 공식을 사용한 경우와의 비교 기준이 된다.

② COS(ltf) : 코사인 유사도 계수를 적용하되 용어빈도에 로그를 취하고 1을 더한 로그 TF 공식을 함께 적용한 경우이다. 흔히 정보검색이나 자동분류에서 단순 단어빈도보다 로그 TF 공식을 적용하는 것이 좋은 성능을 보이는 경우가 많았으므로 이 연구에서도 이를 적용해보았다.

③ JSD : Jenson-Shannon 다이버전스 공식을 적용한 경우이다.

④ SSD : 대칭적 스큐 다이버전스 공식을 적용한 경우이다.

⑤ MDS : 최소 스큐 다이버전스 공식을 적용한 경우이다.

COS(ltf) 방식이 아닌 COS 방식과 세 가지 다이버전스 공식을 적용한 경우는 문헌 벡터의 단어 출현빈도를 그대로 가중치로 처리하였다.

단어 가중치의 중요 요소인 역문헌빈도 IDF의 적용 여부에 따라서 문헌 클러스터링 성능이 달라질 수도 있으므로 IDF 가중치를 적용한 경우와 그렇지 않은 경우를 모두 실험해보았다.

클러스터링 결과의 평가 기준으로는 여러 가지가 제시되어 있으나, 이 연구에서는 개별 문헌을 단위로 클러스터링 성능을 평가하는 단일척도인 WACS 척도(정영미, 이재윤 2001)를 기준으로 평가한 결과를 보고하였다. WACS 척도 이외에 CSIM, 엔트로피, 카이제곱 통계량, F척도 등의 주요 평가 척도를 모두 적용해보았으나 클러스터링 결과의 상대적인 성능 우열에는 거의 차이가 없는 것으로 나타났으므로 다른 평가 척도를 적용한 결과는 소개하지 않았다.

전체 문헌의 수가 D이고, 수작업 범주의 수가 m, 자동생성 클러스터의 수가 n이라고 할 때, 수작업 범주와 자동생성 클러스터를 각각 M과 C로 표기하면 특정 클러스터링 결과의 WACS 척도값은 다음 공식으로 산출한다.

$$WACS = \frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{2|M_i \cap C_j|^2}{|M_i|+|C_j|}$$

이는 평가 기준인 수작업 범주와 자동생성된 클러스터가 공유하는 문헌의 비율로 클러스터링 성능을 평가하는 방식이다.

3.2 다이버전스를 이용한 클러스터링 결과

다섯 가지 유사도 산출 방식 각각에 대해서 실험 집단 세 종류, 클러스터링 기법 세 종류, IDF 적용 여부 두 종류를 조합한 18가지 경우의 성능을 산출하여 <표 3>과 <표 4>에서 비교하였다. 다섯 가지 유사도 산출 방식 중에서 가장 좋은 결과를 보인 것은 MSD 방식으로서 18가지 경우 중에서 12가지 경우에 1위로 나타났다.

<그림 2>와 <그림 3>에서는 다섯 가지 유사도 산출 방식의 성능을 IDF 가중치를 적용하지 않은 경우와 적용한 경우로 구분하여 비교하였다. 두 가지 경우에 모두 가장 좋은 성능을 보인 것은 역시 MSD 방식인 것으로

<표 3> IDF 가중치를 적용하지 않은 경우의 클러스터링 성능 비교 (괄호 안은 순위)

실험 집단	클러스터링 기법	유사도 산출 방식				
		COS	COS(ltf)	JSD	SSD	MSD
CQ	평균연결 기법	0.2852 (2)	0.2537 (3)	0.2357 (4)	0.2332 (5)	0.3533 (1)
	완전연결 기법	0.2941 (5)	0.3001 (4)	0.3211 (2)	0.3486 (1)	0.3208 (3)
	Ward 기법	0.4115 (4)	0.4197 (3)	0.4403 (2)	0.3905 (5)	0.4755 (1)
MQ	평균연결 기법	0.6859 (3)	0.6836 (5)	0.6841 (4)	0.6995 (2)	0.7154 (1)
	완전연결 기법	0.2764 (5)	0.3316 (2)	0.2867 (4)	0.3030 (3)	0.3396 (1)
	Ward 기법	0.6383 (5)	0.7828 (4)	0.8196 (1)	0.8067 (3)	0.8171 (2)
HQ	평균연결 기법	0.7480 (4)	0.7792 (1)	0.7728 (3)	0.7783 (2)	0.6799 (5)
	완전연결 기법	0.3342 (4)	0.3395 (3)	0.3492 (1)	0.3399 (2)	0.3328 (5)
	Ward 기법	0.5607 (5)	0.7730 (3)	0.7158 (4)	0.8057 (1)	0.7751 (2)
평균		0.4705 (5)	0.5182 (3)	0.5139 (4)	0.5228 (2)	0.5344 (1)

<표 4> IDF 가중치를 적용한 경우의 클러스터링 성능 비교 (괄호 안은 순위)

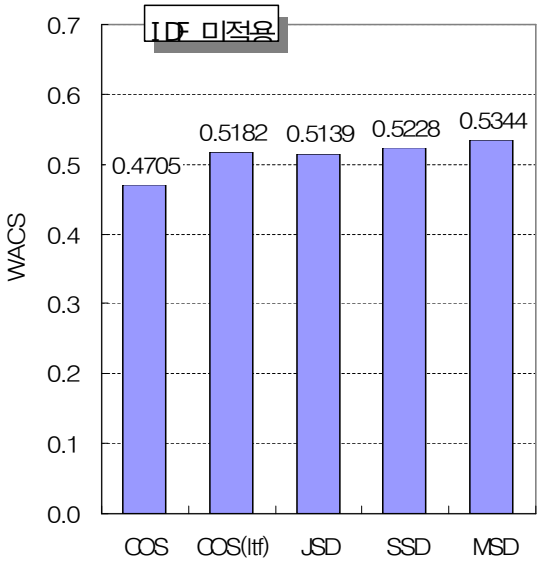
실험 집단	클러스터링 기법	유사도 산출 방식				
		COS	COS(ltf)	JSD	SSD	MSD
CQ	평균연결 기법	0.4366 (1)	0.3945 (3)	0.3253 (4)	0.3174 (5)	0.3960 (2)
	완전연결 기법	0.3002 (2)	0.2959 (3)	0.2797 (4)	0.2725 (5)	0.3576 (1)
	Ward 기법	0.4091 (4)	0.3584 (5)	0.4416 (3)	0.4818 (2)	0.5679 (1)
MQ	평균연결 기법	0.7905 (2)	0.7501 (5)	0.7552 (4)	0.7890 (3)	0.8357 (1)
	완전연결 기법	0.2923 (4)	0.2674 (5)	0.5147 (2)	0.4602 (3)	0.6019 (1)
	Ward 기법	0.6686 (5)	0.7198 (4)	0.7518 (2)	0.7377 (3)	0.7618 (1)
HQ	평균연결 기법	0.7256 (5)	0.7557 (4)	0.7944 (1)	0.7863 (3)	0.7913 (2)
	완전연결 기법	0.3293 (5)	0.3531 (4)	0.5131 (2)	0.4829 (3)	0.6296 (1)
	Ward 기법	0.6263 (5)	0.7558 (2)	0.7701 (1)	0.7518 (3)	0.7456 (4)
평균		0.5087 (5)	0.5168 (4)	0.5718 (2)	0.5644 (3)	0.6319 (1)

나타났다. 특히 전체적인 성능이 더 좋은 <그림 3>의 IDF 가중치를 적용한 경우에 MSD 방식은 COS 방식에 비해서는 24.2%, COS(ltf) 방식에 비해서는 22.3%의 높은 성능 향상율을 보였다. 다른 다이버전스 방식인 JSD와 SSD는 IDF 가중치를 적용하지 않았을 때 COS 방식에 비해서 각각 12.4%와 10.9%의 성능 향상율을 보였으나 MSD 방식의 성능과는 뚜렷한 차이가 있었다.

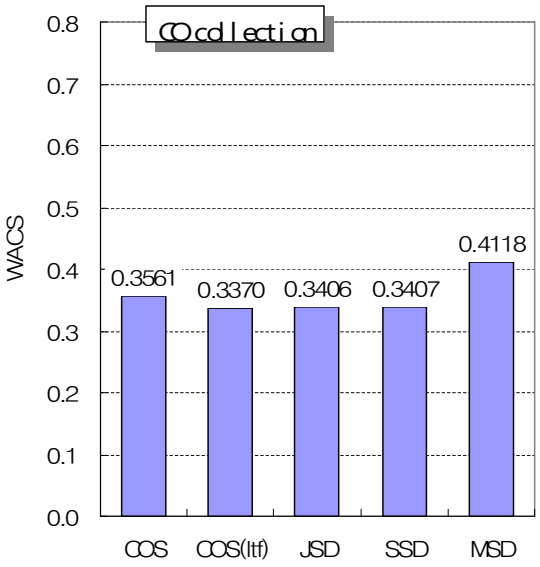
다섯 가지 유사도 산출 방식의 성능을 세 가지 실험 집단별로 평균한 결과는 <그림 4>, <그림 5>, <그림 6>에 각각 제시하였다. 세 가지 실험 집단에서 모두 MSD 방식이 가장 좋은 평균 성능을 보이는 것으로 나타났다. JSD와 SSD의 두 가지 다이버전스 공식은 MQ 실험 집단과 HQ 실험 집단에서는 코사인 유사도를 적용한 경우(COS)보다 높은 성능을 보였으나, <그림 4>의 CQ 실험 집단에서는 그렇지 못했다.

세 가지 클러스터링 기법별로 다섯 가지 유사도 산출 방식의 성능을 평균한 결과는 <그림 7>, <그림 8>, <그림 9>에 제시하였다.

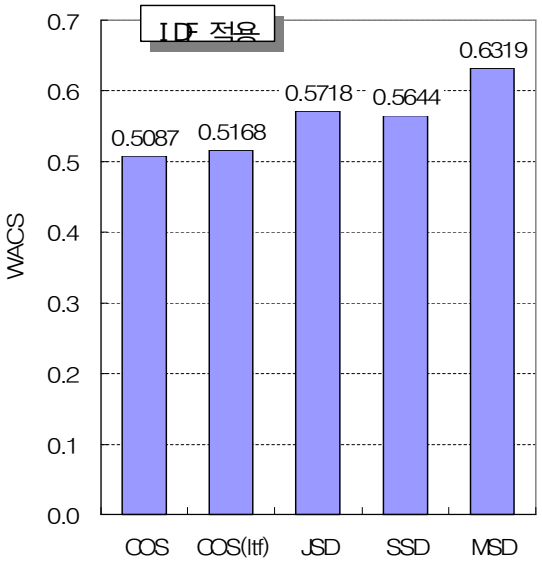
역시 MSD 방식이 세 가지 클러스터링 기법 모두에서 가장 좋은 평균 성능을 보였다. JSD와 SSD의 두 가지 다이버전스 공식은 완전연결 기법과 Ward 기법에서는 COS 및 COS(ltf) 보다 더 높은 성능을 보였으나, 평균연결 기법을 적용한 <그림 7>에서는 그렇지 못했다.



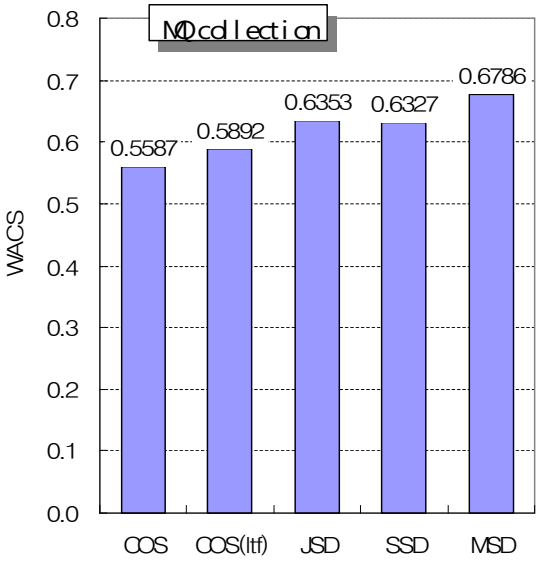
<그림 2> IDF를 적용하지 않은 경우의 클러스터링 성능 비교



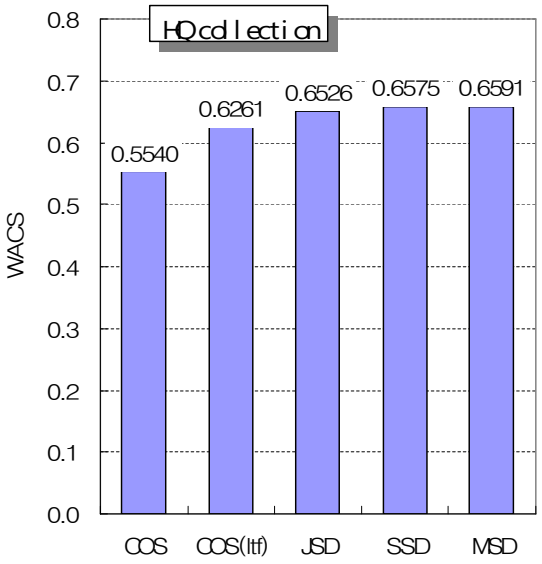
<그림 4> 실험집단 CQ에서의 클러스터링 성능 비교



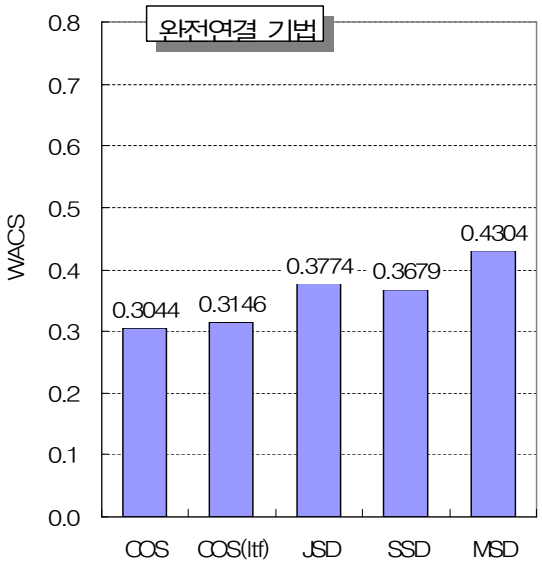
<그림 3> IDF를 적용한 경우의 클러스터링 성능 비교



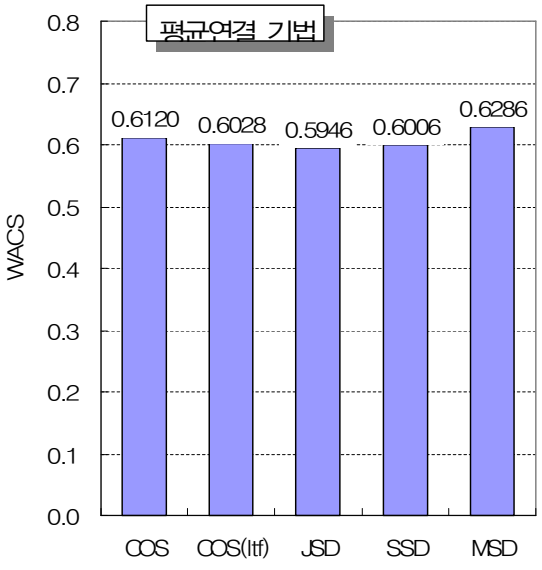
<그림 5> 실험집단 MQ에서의 클러스터링 성능 비교



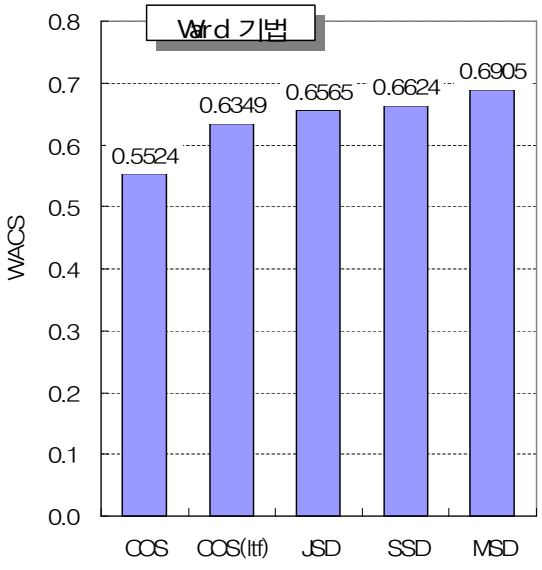
<그림 6> 실험집단 HQ에서의 클러스터링 성능 비교



<그림 8> 완전연결 기법을 사용한 경우의 클러스터링 성능 비교



<그림 7> 평균연결 기법을 사용한 경우의 클러스터링 성능 비교



<그림 9> Ward 기법을 사용한 경우의 클러스터링 성능 비교

이상과 같은 다섯 가지 유사도 산출 방식들 사이의 성능 차이가 통계적으로 유의한지 여부를 검증하기 위해 상황 변수가 조합된 18가지 경우의 성능에 대해서 비모수검증 방식인 Wilcoxon 부호순위 검증을 실시하였다. 그 결과 <표 5>와 같이 MSD는 COS 및 COS(ltf)보다는 99% 유의수준에서, 그리고 다른 다이버전스 공식인 JSD나 SSD보다는 95% 유의수준에서 클러스터링 성능이 앞선 것으로 검증되었다. 결국 MSD 방식은 코사인 유사도를 적용한 경우 뿐만 아니라 다른 다이버전스 공식을 적용한 경우에 비해서도 좋은 클러스터링 성능을 보인다는 것이 95% 유의수준에서 검증된 것이다.

다이버전스 공식 중에서 JSD와 SSD는 COS 방식보다 클러스터링 성능이 우세하다는 것이 95% 유의수준에서 검증되었지만, 로그 TF 가중치 공식을 사용한 COS(ltf) 방식보다는 통계적으로 유의한 성능 차이를 보이지 못했다.

<표 5> 클러스터링 성능 차이의 Wilcoxon 부호순위 검증 결과 (부등호 셋은 99% 유의수준, 부등호 둘은 95% 유의수준에서 각각 성능차이가 유의하다는 것을 뜻하며, 부등호의 방향이 오른쪽이면 해당 칸의 가로행에 적힌 방식이 우월함을 나타내고 부등호의 방향이 왼쪽이면 해당 칸의 세로열에 적힌 방식이 우월함을 나타낸다)

	COS	COS (ltf)	JSD	SSD	MSD
COS		−	<<	<<	<<<
COS(ltf)	−		−	−	<<<
JSD	>>	−		−	<<
SSD	>>	−	−		<<
MSD	>>>	>>>	>>	>>	

결론적으로 다이버전스 공식 중에서 이 연구에서 고안한 최소 스큐 다이버전스 공식 MSD가 모든 경우의 실험 조건에서 가장 뛰어난 평균 성능을 보였으며, 전통적인 코사인 유사도 방식을 포함한 다른 방식과의 성능 차이는 통계적으로 유의한 것으로 검증되었다.

이와 같이 다이버전스 공식 중에서도 MSD가 문헌 클러스터링에서 뛰어난 성능을 보이는 이유는 MSD가 KL 다이버전스 공식의 주요 특징인 비대칭성을 적절히 반영하기 때문이라고 추정된다. 여타 다이버전스 공식인 JSD나 SSD는 문헌간 유사도를 대칭적으로 처리하면서 비대칭성에 담긴 정보가 일부 손실되는데 반해서, MSD는 비대칭적인 두 가지 다이버전스 값 중에 하나를 그대로 취함으로써 비대칭적인 정보가 반영되기 때문이다.

3.3 2차 분포 유사도를 이용한 클러스터링 결과

다섯 가지 유사도 산출 방식을 적용하여 생성한 앞 절의 1차 문헌 유사도 행렬에서 각 행 사이의 피어슨 상관계수를 구하여 도출한 2차 분포 유사도 행렬 다섯 가지를 입력 데이터로 하여 문헌 클러스터링 실험을 수행하였다. 실험은 1차 유사도 행렬을 이용한 앞 절의 상황과 같이 IDF 가중치 적용 여부, 실험 집단, 클러스터링 기법 등을 달리하여 조합한 18가지 경우에 대해서 진행하여 성능을 얻었다. 실험 결과는 IDF 가중치의 적용 여부에 따라 구분하여 <표 6>과 <표 7>에 제시하였다.

<표 6> 2차 분포 유사도 행렬을 적용한 클러스터링 성능 비교 (IDF 가중치 미적용, 괄호 안은 순위)

실험 집단	클러스터링 기법	유사도 산출 방식				
		COS	COS(ltf)	JSD	SSD	MSD
CQ	평균연결 기법	0.3248 (5)	0.3495 (2)	0.3436 (3)	0.3379 (4)	0.3614 (1)
	완전연결 기법	0.3763 (3)	0.4006 (2)	0.3331 (5)	0.4031 (1)	0.3489 (4)
	Ward 기법	0.3872 (4)	0.4053 (3)	0.4205 (2)	0.4263 (1)	0.3867 (5)
MQ	평균연결 기법	0.6719 (5)	0.6989 (3)	0.6936 (4)	0.7242 (1)	0.7231 (2)
	완전연결 기법	0.6567 (5)	0.7913 (1)	0.6938 (4)	0.7256 (3)	0.7800 (2)
	Ward 기법	0.7691 (5)	0.7921 (2)	0.7906 (3)	0.8568 (1)	0.7712 (4)
HQ	평균연결 기법	0.7431 (5)	0.7858 (2)	0.7750 (4)	0.7867 (1)	0.7819 (3)
	완전연결 기법	0.6480 (5)	0.7857 (4)	0.8045 (2)	0.8004 (3)	0.8050 (1)
	Ward 기법	0.7769 (5)	0.8520 (3)	0.7855 (4)	0.8599 (1)	0.8573 (2)
평균		0.5949 (5)	0.6513 (2)	0.6267 (4)	0.6579 (1)	0.6462 (3)

<표 7> 2차 분포 유사도 행렬을 적용한 클러스터링 성능 비교 (IDF 가중치 적용, 괄호 안은 순위)

실험 집단	클러스터링 기법	유사도 산출 방식				
		COS	COS(ltf)	JSD	SSD	MSD
CQ	평균연결 기법	0.5862 (1)	0.4858 (3)	0.3870 (5)	0.4612 (4)	0.5160 (2)
	완전연결 기법	0.4861 (3)	0.4485 (4)	0.3656 (5)	0.5008 (1)	0.4914 (2)
	Ward 기법	0.5081 (5)	0.5334 (4)	0.5605 (2)	0.5913 (1)	0.5602 (3)
MQ	평균연결 기법	0.8145 (3)	0.8156 (2)	0.7732 (5)	0.8162 (1)	0.8084 (4)
	완전연결 기법	0.7404 (4)	0.7358 (5)	0.7999 (1)	0.7958 (2)	0.7898 (3)
	Ward 기법	0.6841 (5)	0.7912 (4)	0.8135 (2)	0.8132 (3)	0.8192 (1)
HQ	평균연결 기법	0.7702 (5)	0.7968 (3)	0.7793 (4)	0.8007 (1)	0.7989 (2)
	완전연결 기법	0.7711 (2)	0.7598 (3)	0.8159 (1)	0.6857 (5)	0.7438 (4)
	Ward 기법	0.8073 (1)	0.8032 (2)	0.7463 (5)	0.7947 (4)	0.7983 (3)
평균		0.6853 (4)	0.6856 (3)	0.6713 (5)	0.6955 (2)	0.7029 (1)

다섯 가지 방식에 대해서 각각 2차 분포 유사도 행렬을 도출하여 클러스터링한 결과에서는 다섯 가지 방식 사이의 성능 차이가 1차 유사도를 적용한 실험에서와 다소 다른 양상을 보였다. 성능이 더 좋은 IDF 가중치를 적용한 경우인 <표 7>을 보면, 평균 성능이 가장 좋은 것은 여전히 MSD 방식이고 두 번째는 SSD 방식이지만, 또 다른 다이버전스 공식인 JSD 방식이 COS 방식보다도 더 나쁜 성능을 보이는 것으로 나타났다.

또한 <표 7>에서 가장 성능이 좋은 MSD 방식의 성능 향상율도 1차 유사도 행렬을 적용하였을 때의 24.2%(COS 대비)와 22.3%(COS(ltf) 대비)에 비해서 훨씬 낮은 2.6%(COS 대비)와 2.5%(COS(ltf) 대비)에 불과하였다. 결국 2차 분포 유사도 행렬을 적용하

였을 때에는 적용 공식 사이의 차이가 크게 감소하는 것을 알 수 있다.

코사인과 다이버전스 공식 사이의 클러스터링 성능 차이는 줄어들었지만, 2차 분포 유사도 행렬을 적용한 경우의 성능을 앞 절의 1차 유사도 행렬을 적용한 경우와 비교해보면 <표 8>, <표 9>, <그림 10>, <그림 11>에서와 같이 크게 향상된 것을 알 수 있다.

성능이 더 좋은 IDF 가중치를 적용한 경우인 <표 9>를 살펴보면 대부분의 경우에 성능이 향상된 것으로 나타났다. 다섯 가지 유사도 산출 방식과 세 가지 실험 집단, 그리고 세 가지 클러스터링 기법을 조합한 45가지 클러스터링 결과 중에서 2차 분포 유사도 행렬을 적용하였을 때의 성능이 1차 유사도 행렬을 적용하였을 때보다 저하된 것은 네 경우에 불과했고 나머지 41가지 경우에는 모두 성능이 향상되었다.

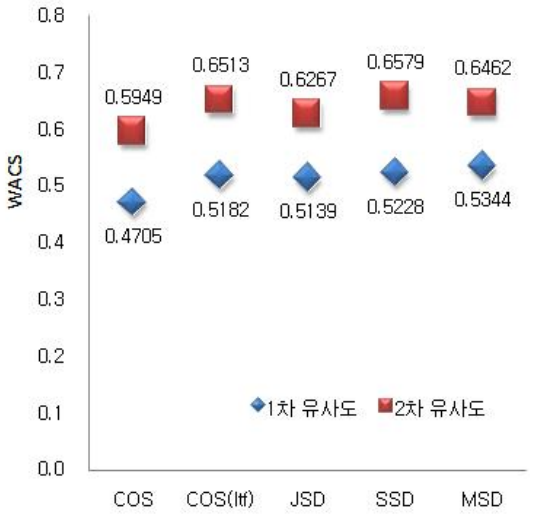
IDF 가중치를 적용하지 않은 <표 8>에서는 1차 유사도 행렬 적용 성능 대비 2차 유사도 행렬 적용 성능의 향상율이 MSD 방식의 20.9%에서부터 COS 방식의 26.4%까지 분

<표 8> 1차 유사도 적용 성능 대비 2차 분포 유사도 행렬 적용 성능의 향상율 (IDF 가중치 미적용)

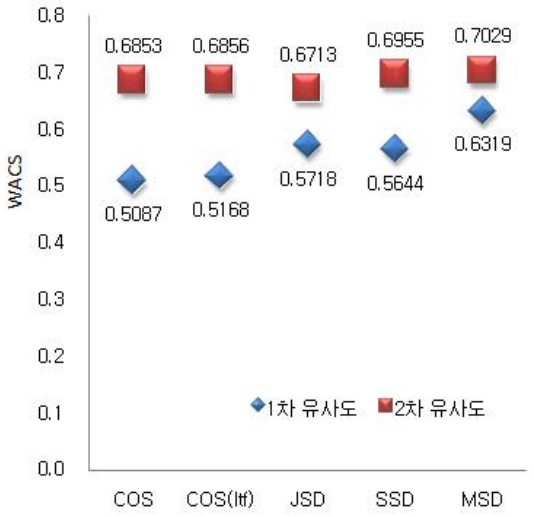
실험 집단	클러스터링 기법	유사도 산출 방식				
		COS	COS(ltf)	JSD	SSD	MSD
CQ	평균연결 기법	13.9%	37.7%	45.8%	44.9%	2.3%
	완전연결 기법	27.9%	33.5%	3.7%	15.6%	8.8%
	Ward 기법	-5.9%	-3.4%	-4.5%	9.2%	-18.7%
MQ	평균연결 기법	-2.1%	2.2%	1.4%	3.5%	1.1%
	완전연결 기법	137.6%	138.6%	142.0%	139.5%	129.7%
	Ward 기법	20.5%	1.2%	-3.5%	6.2%	-5.6%
HQ	평균연결 기법	-0.7%	0.8%	0.3%	1.1%	15.0%
	완전연결 기법	93.9%	131.4%	130.4%	135.5%	141.9%
	Ward 기법	38.6%	10.2%	9.7%	6.7%	10.6%
평균		26.4%	25.7%	21.9%	25.8%	20.9%

<표 9> 1차 유사도 적용 성능 대비 2차 분포 유사도 행렬 적용 성능의 향상율 (IDF 가중치 적용)

실험 집단	클러스터링 기법	유사도 산출 방식				
		COS	COS(ltf)	JSD	SSD	MSD
CQ	평균연결 기법	34.3%	23.1%	19.0%	45.3%	30.3%
	완전연결 기법	62.0%	51.6%	30.7%	83.7%	37.4%
	Ward 기법	24.2%	48.8%	26.9%	22.7%	-1.4%
MQ	평균연결 기법	3.0%	8.7%	2.4%	3.4%	-3.3%
	완전연결 기법	153.3%	175.1%	55.4%	72.9%	31.2%
	Ward 기법	2.3%	9.9%	8.2%	10.2%	7.5%
HQ	평균연결 기법	6.1%	5.4%	-1.9%	1.8%	1.0%
	완전연결 기법	134.2%	115.1%	59.0%	42.0%	18.1%
	Ward 기법	28.9%	6.3%	-3.1%	5.7%	7.1%
평균		34.7%	32.7%	17.4%	23.2%	11.2%



<그림 10> 1차와 2차 유사도 행렬의 클러스터링 성능 비교 (IDF 가중치 미적용)



<그림 11> 1차와 2차 유사도 행렬의 클러스터링 성능 비교 (IDF 가중치 적용)

포하여 유사도 산출 방식간에 크게 차이나지 않았다. 그런데 IDF 가중치를 적용한 <표 9>에서는 MSD 방식은 평균 11.2% 향상된 반면에 COS 방식은 평균 34.7% 향상되어 향상율의 차이가 상대적으로 크게 나타났다.

이와 같은 2차 분포 유사도 적용 효과의 차이는 <그림 10>과 <그림 11>에서도 확인할 수 있다. IDF 가중치를 적용하지 않은 경우인 <그림 10>에서는 1차 유사도 행렬을 적용하였을 때와 2차 분포 유사도 행렬을 적용하였을 때의 클러스터링 성능 차이가 다섯 가지 유사도 산출 방식 간에 크게 다르지 않게 나타났다. 그런데 IDF 가중치를 적용한 <그림 11>에서는 1차 유사도에서 성능이 좋았던 방식일수록 2차 분포 유사도를 적용하였을 때의 향상된 정도가 작게 나타나는 것을 알 수 있다.

이와 같은 현상의 원인은 1차 유사도 행렬을 적용하였을 때 IDF 가중치를 적용하지 않은 경우의 성능이 상대적으로 낮았기 때문으로 해석할 수 있다. <그림 10>과 <그림 11>에서는 전체적으로 1차 유사도 행렬을 적용하였을 때의 성능이 낮은 경우일수록 2차 분포 유사도 행렬을 적용하여서 높은 향상 효과를 얻은 것으로 나타난다.

IDF 가중치를 적용한 경우에 다섯 가지 유사도 산출 방식 중 최고 성능과 최저 성능의 차이가 1차 유사도를 적용한 실험에서는 0.1232에 달했는데, 2차 분포 유사도를 적용한 실험에서는 불과 0.0176으로 대폭 축소되었다. 처리 시간이 추가로 소요되는 단점은 있지만, 2차 유사도 행렬을 적용하는 클러스터링을 수행하면 성능은 어떤 공식을 적용하더라도 높은 수준을 달성할 수 있음을 나타

낸다.

전반적으로는 2차 분포 유사도 행렬을 사용한 경우에도 IDF 가중치를 적용한 경우의 성능이 IDF 가중치를 적용하지 않은 경우보다 더 좋게 나타났다. 또한 다섯 가지 유사도 산출 방식 중에서 가장 좋은 경우는 1차 유사도 행렬을 적용했을 때와 마찬가지로 이 연구에서 제안한 최소 스큐 다이버전스 공식을 사용한 MSD 방식이었다.

4. 결론

다이버전스 공식은 정보이론 분야에서 유래한 일종의 분포 유사도 공식으로서 최근 용어 클러스터링 분야에 성공적으로 적용되고 있다. 이 연구에서는 이와 같은 다이버전스 공식을 이용하여 문헌 사이의 유사도를 산출하는 방안을 세 가지 고안한 다음, 이를 적용하여 문헌 클러스터링 성능을 향상시킬 수 있는지 여부를 모색해보았다. 또한 2차적인 분포 유사도라고 할 수 있는 피어슨 상관계수 행렬로 문헌 클러스터링 실험을 수행하였다.

실험 결과 이 연구에서 제안한 문헌간 다이버전스 산출 방식을 사용하였을 때 전통적인 코사인 유사도 공식에 비해서 더 높은 클러스터링 성능을 얻을 수 있었다. 특히 최소 스큐 다이버전스(MSD) 공식은 코사인 유사도에 비해서 24.2%의 높은 클러스터링 성능 향상율을 보였다. 또한 최소 스큐 다이버전스 공식은 다른 코사인 유사도만 아니라 여타 다이버전스 공식들에 비해서도 우세한 성능을 보이는 것이 통계적으로 검증되었다.

1차적인 문헌간 유사도를 산출한 후 유사도행렬로부터 2차적으로 피어슨 상관계수를 구하는 2차 분포 유사도 방식을 적용한 실험에서는, 1차 유사도 산출 단계에서 어느 유사도 산출 공식을 사용하였더라도 성능 향상 효과를 얻을 수 있는 것으로 나타났다. 2차 분포 유사도 행렬을 적용하여 얻는 성능 향상 효과는 1차 유사도 행렬을 적용하였을 때의 성능에 반비례하였다. 따라서 1차 유사도 행렬을 적용한 실험에서 성능이 가장 좋았던 최소 스큐 다이버전스는 2차 분포 유사도 행렬을 적용하면 성능 향상 효과가 가장 낮게 나타났다. 그러나 여전히 평균 성능은 최소 스큐 다이버전스로 1차 유사도를 구한 후 피어슨 상관계수 행렬을 산출하여 클러스터링을 수행한 경우가 가장 좋았다.

전체적으로는 2차 분포 유사도 행렬을 적용하면 코사인이나 다이버전스 공식을 불문하고 모든 경우에 1차 유사도 행렬을 적용하였을 때의 최고 성능(MSD의 0.6319)보다 더 좋은 성능을 얻을 수 있었다. 따라서 처리 시간이 중요한 응용분야에서는 최소 스큐 다이버전스 공식을 적용하여 문헌 클러스터링을 수행하는 것이 가장 좋고, 처리 시간보다 클러스터링 성능이 더 중요한 응용분야에는 2차 유사도 행렬 방식을 적용하는 것이 바람직하다.

실제로 디지털도서관이나 검색포털의 검색결과 클러스터링 서비스에서는 실시간 처리가 필요하다. 따라서 이런 경우에는 최소 스큐 다이버전스 공식을 적용하는 것이 추가 소요 시간 없이 성능 개선 효과를 얻는 방법이 된다. 이와 달리 특정 주제에 속한 문헌들

을 더 세분하여 나누기 위한 클러스터링 처리에서는 시간이 더 소요되더라도 분류 성능을 향상시킬 수 있는 2차 유사도 행렬 방식을 적용해야 할 것이다.

다이버전스나 2차 유사도와 같은 분포 유사도가 문헌 클러스터링에 효과적임이 확인되었으므로 향후에는 또다른 자동분류인 문헌 범주화에서 분포 유사도를 적용하여 성능 향상을 모색해보는 후속 연구가 필요하다.

참고문헌

정영미. 2005.『정보검색연구』. 서울: 구미무역(주) 출판부.

정영미, 이재윤. 2001. 지식 분류의 자동화를 위한 클러스터링 모형 연구.『정보관리학회지』, 18(2): 203-230.

Dagan, Ido, Lillian Lee, and Fernando Pereira. 1999. "Similarity-based models of cooccurrence probabilities." *Machine Learning*, 34(1-3): 43-69.

Griffith, A., L. A. Robinson, and P. Willett. 1984. "Hierarchic agglomerative clustering methods for automatic document classification." *Journal of Documentation*, 40(3): 175-205.

Griffiths, A., H. C. Luckhurst, and P. Willett. 1986. "Using interdocument similarity information in document retrieval systems." *Journal of the American Society for Information Science*, 37(1): 3-11.

Kullback, S., and R. A. Leibler. 1951. "On information and sufficiency." *Annals of Mathematical Statistics*, 22(1): 79–86.

Kullback, Solomon. 1968. *Information Theory and Statistics*, 2nd ed. New York: Dover Books.

Lee, Lillian. 1999. "Measures of distributional similarity." *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 25-32.

Lee, Lillian. 2001. "On the effectiveness of the skew divergence for statistical language analysis." *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics (AISTATS-2001)*, 65-72.

Lee, Lillian, and Fernando Pereira. 1999. "Distributional similarity models: Clustering vs. nearest neighbors." *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 33-40.

Lin, Dekang. 1998. "Automatic retrieval and clustering of similar words." *Proceedings of the COLING-ACL '98*, 768-773.

Lin, Jianhua. 1991. "Divergence measures based on the Shannon entropy." *IEEE Transactions on Information Theory*, 37(1): 145-151.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. "Distributional clustering of English words." *Proceedings of the 31st Annual Meeting of the ACL*, 183-190.

Salton, Gerard, and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw Hill.

Theodoridis, S., and K. Koutroumbas. 2003. *Pattern Recognition*. 2nd ed. Oxford, UK: Elsevier.

Weeds, J. E. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph. D. diss., University of Sussex.

White, H. D., and B. C. Griffith. 1981. "Author cocitation: a literature measure of intellectual structure." *Journal of the American Society for Information Science*, 32: 163-171.