# 클러스터링 성능 평가를 위한 비편향적 척도의 개발

# Development of an Unbiased Measure for Clustering Performance

정영미, 이재윤, 연세대학교 문헌정보학과

Young-Mee Chung, Jae-Yun Lee, Yonsei University

클러스터링 성능 평가를 위한 척도로 여러 공식이 개발되어 사용되어왔다. 이들 평가척도는 가급적 범위가 제한되며, 가환성이 있고, 비편향적이며 단일 척도일 필요가 있다. 기존 평가 척도에 대해서 검토한 후 비편향적인 단일 척도 *WACS*를 개발하였다. 클러스터 수를 달리하는 클러스터링 결과에 대해 여러 평가척도를 적용해서 성능을 평가하는 실험을 통해서 *WACS* 척도가 평가척도로서의 요건을 만족시킨다는 것을 확인하였다.

## 1. 서론

문헌 클러스터링의 성능 평가를 위해 클러스터를 이용한 정보 검색 실험을 수행한 다음, 검색 성능을 재현율과 정확률 척도로 측정하는 경우가 많다. 이러한 접근방법은 검색 성능으로 분류 성능을 대신하고 있는 것이다.

그러나 재현율과 정확률에 근거한 척도는 "잘 된 클러스터링 결과는 정보 검색의 효율성을 증진시킨다"는 가설에 기반한 분류 성능의 간접적인 평가 척도이므로 자동분류에서는 클러스터링 결과에 대한 직접적인 해석을 위해 생성된 클러스터 자체를 평가할 필요가 있다.

클러스터링 성능의 평가는 정보 검색이나 텍스트 범주화 기법의 성능 평가에 비해 어려운 점이 있다. 정보 검색과 텍스트 범주화의 평가에서 흔히 사용되는 정확률과 재현율은 각 문헌에 대한 적합 질의나 적합 범주가 미리 판정이 되어 있는 상태이므로 객관적이고 절대적인 평가가 가능하다. 그러나 클러스터링의 경우에는 생성된 클러스터가 어느 범주에 해당하는지, 또는 특정 문헌이 어느 범주로 자동 분류되었는지를 판정하기가 어렵다.

따라서 동일한 환경에서 상대적인 평가를 하는 것이 현실적인 방안이다. 여기서 동일한 환경이란 클러스터 수와 같은 파라미터를 일치시키며 동일한 평가척도를 적용하는 것을 말한다.

클러스터링 결과를 수작업 분류 결과와 비교하기 위한 클러스터링 성능 평가척도에 요구되는 성질은 다음과 같다. 첫째, 값의 범위가 고정되어야 한다. 둘째, 가환성(commutative)이 있어야 한다. 즉, 기준 범주와 평가 대상 클러스터를 뒤바꾸더라도 결과가 같아야 한다는 것이다. 셋째, 클러스터의 수나 크기와 같은 변수에 영향을 덜 받도록 비편향적이어야 한다. 넷째, 가능하면 단일 척도이어야 한다.

## 2. 기존 평가척도

클러스터링 결과 자체를 평가하는 방법은 크게 문헌쌍을 단위로 분류하는 경우와, 개별 문헌을 단위로 분류하는 경우로 나눌 수 있다.

### 2.1 문헌쌍 단위 평가척도

클러스터링 결과의 성능을 쌍 단위로 평가하는 방법은 Rand(1971)가 제시한 이후 여러 연구에서 검토되었다. 문헌쌍을 기준으로 평가할 때에는 두 문헌이 동일한 클러스터에 속했는가의 여부에 따라서 다음과 같은 2×2 분할표에 기반한 공식을 적용하게 된다.

| | | 클러스터링 결과 B: 동일 클러스터에 속함 | 클러스터링 결과 B: 다른 클러스터에 분리 |
|---|---|---|---|
| 클러스터링 결과 A | 동일 클러스터에 속함 | $a$ | $b$ |
| | 다른 클러스터에 분리 | $c$ | $d$ |

**(1) *Rand***

Rand(1971)가 제시한 척도는 2개의 다른 클러스터 집합에 다음과 같은 단순일치계수 공식을 적용하여 성능을 평가한다.

$$Rand(C_A, C_B) = \frac{a+d}{a+b+c+d}$$

Rand 척도는 비록 0보다 크고 1보다 작거나 같은 값을 가지도록 되어 있지만, 현실적으로는 $d$항이 나머지에 비해서 상대적으로 매우 크므로 높은 수준에서 좁은 범위의 값을 가지게 된다. 또한 클러스터의 수가 많아질수록 $d$항이 커지므로 이에 비례해서 높은 값을 가지게 되는 편향성이 있다(김정하, 이재윤 2000).

**(2) 클러스터링 오류**

Roussinov & Chen(1999)은 Rand의 공식에서 분자를 불일치 쌍의 수로 설정한 클러스터링 오류(clustering error) 공식을 제안하였다. 위의 2×2 표로 클러스터링 오류 공식을 표현하면 다음과 같다. 이는 1에서 Rand 공식값을 뺀 것과 같다.

$$CE(C_A, C_B) = \frac{b+c}{a+b+c+d}$$

클러스터링 오류값은 Rand 척도와 마찬가지로 클러스터의 크기가 작은 경우(클러스터의 수가 많은 경우)를 선호하는 편향성이 있기 때문에 비교되는 두 클러스터링 결과의 클러스터 크기에 의한 영향을 줄일 수 있도록 다음과 같은 정규화 클러스터링 오류값 공식도 제안하였다.

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

**(3) *CSIM***

Chung & Lee(2001)에서는 결과 A에서 동일한 클러스터에 속한 쌍이 결과 B에서도 동일한 클러스터에 속할 확률을 나타내는 공식으로 클러스터링 유사도 *CSIM*을 제안하였다. 이 공식은 앞의 2×2 분할표에 다이스 계수 공식을 적용한 형태이다. 1에서 *CSIM*값을 빼면 *NCE*값이 된다.

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

*CSIM*은 2×2 분할표의 $d$항을 사용하지 않으므로 Rand 공식과는 달리 비편향적이긴 하지만 크기가 1인 클러스터에 적용할 수가 없다는 한계가 있다. 이는 정규화 클러스터링 오류 공식도 마찬가지로서 Roussinov & Chen(1999)은 크기가 3 이상인 클러스터만을 대상으로 하고 나머지는 평가에서 제외하였다.

### 2.2 문헌 단위 평가척도

문헌 단위 평가척도로는 $\chi^2$ 통계치(Borko et al.

1968)나 엔트로피(Boley et al. 1999)가 사용되기도 하는데, $\chi^2$ 통계치는 값이 클수록 연관성이 높음을 나타내지만, 최대값이 고정되어 있지 않으므로 성능 차이의 정도를 직관적으로 해석하기가 어려우며, 생성한 클러스터의 수가 많을수록 높은 값을 가지게 되는 편향성이 있다(김정하, 이재윤 2000). 엔트로피 공식은 최저값이 고정되어 있지 않으며 가환성이 없는 것이 가장 큰 단점이다. 엔트로피 공식에서는 클러스터의 크기가 작을수록 높은 성능이 나오게 되며 극단적으로는 클러스터의 크기가 모두 1인 경우에 최고의 성능, 즉 0이 된다.

여기서는 이 두 가지 이외의 문헌 단위 평가척도로 상호정보량과 F 척도를 살펴보았다.

### (1) 상호정보량

Vaithyanathan & Dom(1999)은 수작업 분류와 자동 클러스터링 결과 사이의 상호정보량으로 성능을 평가하였다. 이는 정보 이론을 응용하되 엔트로피와 달리 가환성을 확보하기 위한 방안이라고 할 수 있다. 수작업 분류 $M$과 자동 클러스터링 $C$ 사이의 상호정보량 공식은 다음과 같다.

$$I(M;C) = H(M)-H(M|C) = \sum_i\sum_j p(M_i,C_j)\log_2\frac{p(M_i,C_j)}{p(M_i)p(C_j)}$$

위의 공식을 클러스터의 크기와 전체 문헌수 $D$로 표현하면 아래와 같다.

$$I(M;C) = \frac{1}{D}\sum_i\sum_j |M_i\cap C_j|\log_2\frac{D\ |M_i\cap C_j|}{|M_i|\ |C_j|}$$

상호정보량은 비교 대상인 두 결과의 클러스터 수(또는 크기)의 차가 현저한 경우를 과대 평가하는 편향성이 있다. 이는 상호정보량 공식의 분모에서 수작업 범주와 자동생성 클러스터의 크기를 곱하기 때문이다. 따라서 수작업 범주가 크기가 고를 경우에는 자동생성 클러스터의 크기가 편차가 심한 경우에 유리하고, 반대로 수작업 범주의 크기가 편차가 심할 경우에는 자동생성 클러스터의 크기가 고른 경우에 유리하다.

Strehl, Ghosh, & Mooney(2000)에서는 역시 상호정보량을 이용하되, 데이터의 분포에 따른 영향을 없애기 위해서 임의로 클러스터링한 경우의 성능을 기본값(baseline)으로 설정한 다음, 각 실험 결과의 상호정보량과 이 기본값과의 비율을 "성능 향상율(performance lift)"이라는 이름의 척도로 사용하였다.

### (2) $F$ 척도

클러스터내 검색 성능이나 문헌 범주화 성능을 측정하기 위해 사용되는 $F$ 척도를 클러스터링 성능을 평가하기 위해 사용하는 경우도 있다. 클러스터링에서 $F$ 척도의 산출 방법은 클러스터-범주 대응 방식과 문헌별 정확률, 재현율 평균 방식의 두 가지가 있다.

클러스터-범주 대응 방식 $F$ 척도에서는 우선 생성된 클러스터를 수작업 분류 범주 중 가장 유사한 하나와 대응시킨 다음 범주화에서와 같이 분류 정확률과 분류 재현율을 계산하여 $F$ 값을 구한다(Sahami, Yusufali, & Baldonado 1998). 이때 한 클러스터에 대해서 대응시킬 수작업 분류 범주를 찾는 방법으로는 해당 클러스터내 소속 문헌들이 가장 많이 속한 범주를 택하거나 또는 모든 수작업 분류 범주와 비교하여 $F$ 값을 구한 다음 가장 높은 경우를 선택하는 방식이 사용된다.

이 방식은 문헌들이 속한 여러 수작업 범주 중에서 최빈 범주 이외의 것들은 모두 오류로 처리하기 때문에 정밀하지 못한 경향이 있다. 예를 들어 크기가 10인 클러스터에 속한 문헌들 중 6개는 범주 A에 속하고 나머지 4개는 모두 범주 B에 속한 경우와, 6개는 범주 A에 속하고 나머지 4개는 모두 소속 범주가 각각인 경우를 구분하지 못하는 것이다. Vaithyanathan & Dom(1999)도 이와 비슷한 문제점으로 한 클러스터에 복수 수작업 범주를 대응시킬 수 없다는 점을 지적하고 그 대신 상호정보량을 척도로 사용하였다.

위의 방법과는 달리 문헌별 정확률, 재현율 평균

방식 $F$ 척도에서는 클러스터별로 해당 수작업 범주를 판정하지 않고 각 문헌별로 정확률, 재현율을 계산하여 그 평균값을 이용한다(Aslam, Pelekhov, & Rus 1998). 이 경우에는 각 문헌을 질의로, 그 문헌이 속한 수작업 범주를 적합문헌 집합으로, 그 문헌이 속한 자동생성 클러스터를 검색된 문헌집합으로 간주하여 정확률과 재현율을 계산하게 된다.

문헌 $d_k$가 속한 수작업 범주와 자동생성 클러스터가 각각 $M_i$와 $C_j$일 때, 문헌 $d_k$의 정확률과 재현율은 다음과 같이 구한다.

$$Precision(d_k) = \frac{|M_i \cap C_j|}{|C_j|}$$

$$Recall(d_k) = \frac{|M_i \cap C_j|}{|M_i|}$$

전체 문헌의 수가 $D$이고, 수작업 범주의 수가 $m$, 자동생성 클러스터의 수가 $n$이라고 할 때, 정확률과 재현율은 각 문헌에 대해 구한 정확률과 재현율의 평균으로 정하게 되고 단일 척도로는 이 정확률과 재현율의 복합 척도를 사용한다. 이를 공식으로 정리하면 아래와 같다.

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

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

$$Clustering\ F(p,r) = \frac{2pr}{p+r}$$

문헌별 정확률, 재현율 평균을 이용한 $F$ 척도는 0에서 1사이로 값이 제한되고 가환성이 있으며 클러스터 크기의 편차에 그다지 영향받지 않고, 클러스터-범주 대응 방식의 F 척도보다 정밀하다는 장점이 있다. 다만 분류라기 보다는 검색의 측면에서 정립된 척도로서 단일 척도를 직접 산출하지 못하고 복합 척도 방식으로 산출해야 한다는 한계가 있다.

## 3. 비편향적 단일 척도 *WACS*

본 연구에서는 클러스터링 성능 평가를 위한 비편향적 단일 척도로 가중 평균 클러스터 유사도인 *WACS(Weighted Average Cluster Similarity)* 척도를 개발하였다. 기본적인 발상은 비교 대상인 수작업 분류 범주와 자동생성 클러스터를 각각 소속 문헌의 벡터로 표현한 다음 유사한 정도를 벡터 유사도 공식으로 계산하는 것이다. 모든 범주와 클러스터 사이의 유사도를 계산하여 범주와 클러스터의 크기를 고려한 가중 평균을 산출하면 전체의 유사도를 구할 수가 있다.

*WACS* 척도에서는 먼저 각 수작업 분류 범주와 자동 생성된 각 클러스터와의 유사도를 다이스 계수 공식을 적용하여 산출한 후 각 클러스터 $C_j$의 성능을 계산한다. 즉 클러스터 $C_j$에 속한 문헌이 하나 이상 속한 수작업 분류 범주를 모두 찾아서 유사도를 계산한 다음 일치하는 문헌의 수를 반영하여 가중 평균을 산출한다. 범주 $M_i$와 클러스터 $C_j$ 간의 유사도 공식과 각 클러스터 $C_j$에 대한 *WACS* 공식은 다음과 같다.

$$Sim(M_i, C_j) = \frac{2|M_i \cap C_j|}{|M_i| + |C_j|}$$

$$WACS(C_j) = \sum_{i=1}^{m}\frac{Sim(M_i, C_j)|M_i \cap C_j|}{|C_j|} = \frac{1}{|C_j|}\sum_{i=1}^{m}\frac{2|M_i \cap C_j|^2}{|M_i| + |C_j|}$$

클러스터링 기법의 전체 성능 $WACS(C)$는 각 클러스터에 대한 $WACS(C_j)$를 모두 합한 후 클러스터 크기를 반영한 가중 평균을 구하여 산출한다.

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

표 1. 수작업 분류 39범주를 기준으로 클러스터링 결과를 평가한 경우

| | *CSIM* | *F* 척도 | (정확률) | (재현율) | 상호정보량/10 | *WACS* (코사인) | *WACS* (다이스) |
|---|---|---|---|---|---|---|---|
| k=39 | 0.21265 | 0.26103 | (0.30361) | (0.22892) | 0.24668 | 0.24657 | 0.23030 |
| k=200 | 0.10287 | 0.18499 | (0.55954) | (0.11081) | 0.36675 | 0.22628 | 0.16419 |

표 2. 수작업 분류 360범주를 기준으로 클러스터링 결과를 평가한 경우

| | *CSIM* | *F* 척도 | (정확률) | (재현율) | 상호정보량/10 | *WACS* (코사인) | *WACS* (다이스) |
|---|---|---|---|---|---|---|---|
| k=39 | 0.17447 | 0.25265 | (0.15757) | (0.69237) | 0.43756 | 0.28547 | 0.21514 |
| k=200 | 0.26203 | 0.48375 | (0.42083) | (0.56878) | 0.62360 | 0.43211 | 0.38897 |

범주와 클러스터 사이의 유사도 산출을 위해 코사인 계수가 아닌 다이스 계수를 사용한 이유는 코사인 계수 공식이 분모에 벡터의 길이를 곱하도록 되어 있기 때문에 긴 벡터의 경우 분모가 커지므로 상대적으로 작은 값을 가지게 되는 편향성이 있기 때문이다. 즉, 클러스터를 크게 분할한 경우가 작게 분할한 경우에 비해서 낮은 값을 가지게 된다. 이보다 더 큰 문제는 두 클러스터 사이의 크기의 편차가 큰 경우에 곱한 결과가 작아지므로, 평가 기준이 되는 수작업 분류 범주의 수가 자동 생성한 클러스터의 수와 큰 차이가 날수록 좋은 성능으로 평가될 여지가 많다는 점이다.

## 4. 평가척도의 비교 실험

각 평가척도의 편향성을 알아보기 위해서 자동분류 연구용 신문기사 집합인 KC-KFCM 1,020건을 대상으로 k-means 클러스터링을 수행하고 그 결과를 각 척도로 평가해보았다. 클러스터링에서는 문헌을 표현하는 색인 단어 집합을 장서빈도 10이상으로 제한하여 축소하고 TF · IDF 가중치를 적용하였다. k=39일 때와 k=200일 때의 k-means 클러스터링 결과를 수작업 대분류(39개 범주) 및 수작업 소분류(360개 범주)를 기준으로 각각 5가지 방법으로 비교 평가하고 그 결과를 그림 1, 그림2와 표 1, 표 2로 나타내었다. 이때 상호정보량의 값은 다른 척도와 스케일을 맞추기 위해서 10으로 나누었다.

그림 1은 수작업 분류가 39개 범주일 때 클러스터 수를 39개 생성한 경우와 200개 생성한 경우를 비교한 것이다. *CSIM*과 *F* 척도, *WACS*(다이스) 척도는 수작업 분류와 같은 39개 클러스터를 생성한 경우가 성능이 현격하게 더 좋은 것으로 평가하는 반

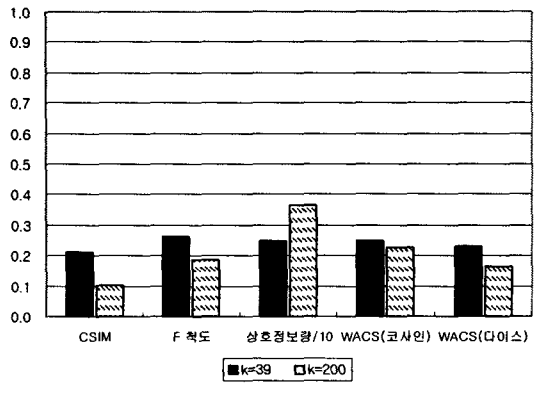


그림 1. 39범주 기준 클러스터링 성능 평가 결과

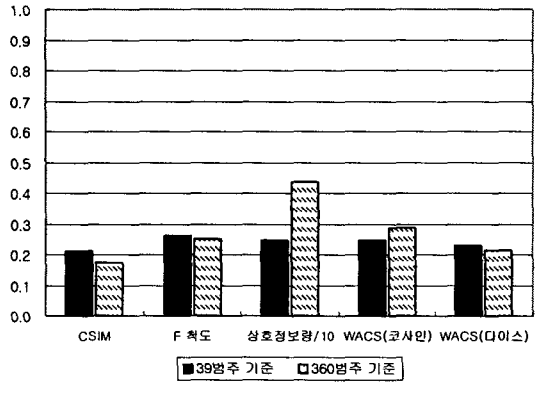


그림 2. k=39인 경우의 클러스터링 성능 평가 결과

면에, 상호정보량은 수작업 분류보다 훨씬 많은 200개 클러스터를 생성한 경우가 오히려 더 좋은 것으로 평가함을 볼 수가 있다. *WACS*(코사인)은 비록 39개를 생성한 경우가 더 좋은 것으로 나타나지만 그 차이가 그다지 크지 않다.

그림 2는 그림 1과 반대로 클러스터 수를 39개 생성하였을 때, 수작업 분류 39인 경우와 수작업 분류 360인 경우를 각각 기준으로 한 평가 결과를 비교한 것이다. 여기서도 *CSIM*과 *F* 척도, *WACS*(다이스) 척도는 클러스터 수와 같은 39개 수작업 분류 범주를 기준으로 비교한 경우가 성능이 더 좋은 것으로 평가하는 반면에, 상호정보량과 *WACS*(코사인)은 클러스터 수보다 훨씬 많은 360개 수작업 분류 범주를 기준으로 비교한 경우가 더 성능이 좋은 것으로 평가하며 상호정보량의 경우에 그 정도가 심함을 알 수가 있다.

## 5. 결론

문헌 클러스터링 결과를 평가하기 위한 기존 척도를 검토한 결과 범위 제한, 가환성, 비편향적, 단일 척도라는 네 가지 요건을 모두 만족하는 것을 찾지 못하였다. 이에 따라 가중평균 클러스터 유사도인 *WACS* 척도를 개발하고 비교 실험을 통해서 특성을 확인하였다. *CSIM*과 *F* 척도도 범위 제한, 가환성, 비편향적이라는 조건을 만족시키긴 하지만, *CSIM*은 크기가 1인 클러스터를 배제하는 문제가 있으며, F 척도는 단일 척도가 아니라는 단점이 있다. 정보 시스템에서 클러스터링을 적용하는 영역이 급증하는 현 추세에서 본 연구에서 개발한 비편향적인 단일 척도 *WACS*는 올바른 시스템 성능 평가에 기여할 것이다.

## 참고문헌


김정하, 이재윤. 2000. "문헌 클러스터링 결과의 성능 평가방법에 관한 비교 연구." 제7회 정보관리학회 학술대회 논문집: 45-50.

Aslam, J., Pelekhov, K., and Rus, D. 1998. "Static and dynamic information organization with star clusters." In *Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management* (pp.208-217).

Boley, D., et al. 1999. "Partitioning-based clustering for Web document categorization." *Decision Support Systems*, 27(3): 329-341.

Borko, H. et al. 1968. *On-line Information Retrieval using Associative Indexing.* RADC-TR-68-100. AD670195. Califonia : System Develop. Corp. Quated in Anderberg M.R. *Cluster Analysis for Applications.* (New York : Academic Press, 1973), 204-207.

Chung, Young Mee, and Lee, Jae Yun. 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.

Rand, W.M. 1971. "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical Association*, 66: 846-850.

Roussinov, D.G., and Chen, H. 1999. "Document clustering for electronic meetings: an experimental comparison of two techniques." *Decision Support Systems*, 27(1-2): 67-79.

Sahami, M., Yusufali, S., and Baldonado, M.Q.W. 1998. "SONIA : a service for organizing networked information autonomously." In *Proceedings of the Third ACM Conference on Digital Libraries* (pp. 200-209).

Strehl, A., Ghosh, J., and Mooney, R. 2000. "Impact of similarity measures on Web-page clustering." In *Proceedings of the AAAI-2000 Workshop on Artificial Intelligence for Web Search.*

Vaithyanathan, S., and Dom, B. 1999. "Model selection in unsupervised learning with applications to document clustering." In *Proceedings of the 16th International Conference on Machine Learning* (pp. 423-433).