중복을 허용하는 계층적 클러스터링 기법에서 클러스터 간 유사도 평가

전준우○, 송광호, 김유성
인하대학교, 정보통신공학과
yskim@inha.ac.kr

A Novel Linkage Metric for Overlap Allowed Hierarchical Clustering

Joon-Woo Jeon○, Kwang-Ho Song, Yoo-Sung Kim
Dept. of Information and Communication Engineering, Inha University

요 약

본 논문에서는 클러스터 간의 중복을 허용한 계층적 클러스터링(hierarchical clustering) 기법에 적합한 클러스터 간 유사도 평가방법(linkage metric)을 제안하였다. 클러스터 간 유사도 평가방법은 계층적 클러스터링에서 클러스터를 통합하거나 분해하는데 쓰이며 사용된 방법에 따라 클러스터링의 결과가 다르게 형성된다. 기존의 클러스터 간 유사도 평가방법인 single linkage, complete linkage, average linkage 중 single linkage와 complete linkage는 클러스터 간 중복이 허용된 환경에서 정확도가 낮은 문제점이 있고, average linkage는 정확도가 두 방법에 비해 높지만 계산 시간 소요가 크다는 단점이 있다. 따라서 본 논문에서는 기존의 average linkage를 개선하여 중복된 데이터에 의한 필요 계산량을 크게 줄임으로써 시간적 성능이 우수한 클러스터 간 유사도 평가방법을 제안하였다. 또한, 제안된 방법을 기존 방법들과 비교 실험하여 중복을 허용하는 계층적 클러스터링 환경에서 정확도는 비슷하거나 더 높고, average linkage에 비해 계산량이 감소됨을 확인하였다.

주제어: 계층적 클러스터링, 중복 허용, 클러스터 간 유사도, 효율성 개선

1. 서론

방대한 학술문헌들로 이루어진 문헌DB에 대한 정보검색 서비스에서는 검색성능을 증진하기 위해 여러 방법들을 사용하며, 그 중 문헌들을 주제별로 계층적 포함관계를 이용해 미리 분류하고 검색 요청 시 적절히 분류된 클러스터만을 검색 대상으로 제공하여 결과를 제공하는 계층적 클러스터링이 자주 사용된다. 이러한 계층적 클러스터링에서는 클러스터 간 유사도에 따라 클러스터를 통합하거나 분해하기 때문에 클러스터 간의 유사도 평가방법(linkage metric)은 최종 클러스터링 결과에 큰 영향을 줄 수 있는 중요한 요소이다. 유사도 평가방법에는 일반적으로 single linkage[1], complete linkage[2], average linkage[3]등이 있으며 사용된 유사도 평가방법 따라 클러스터링의 진행 과정과 최종 결과가 다르기 때문에 원하는 결과를 도출하기 위해서는 해당 환경에 적합한 유사도 평가방법을 선정하여 이용하는 것이 중요하다[4].

한편, 문헌 DB를 구축할 때 여러 주제를 동시에 가지는 문헌들은 여러 클러스터에 속할 수 있어야한다. 그로 인해 문헌들 사이의 여러 포함 관계를 나타낼 필요성이 있지만, 기존의 계층적 클러스터링은 클러스터 간 중복을 허용하지 않으므로 수직적 포함관계만을 나타낼 수 있고 수평적 포함관계는 나타낼 수 없어 그 필요성을 완벽히 충족시키기 어렵다. 따라서 클러스터 간 중복을 허용하는 환경을 제공할 수 있는 계층적 클러스터링 기법들을 연구 할 필요가 있으며 그 기법들에 적합한 유사도 평가방법을 고안 할 필요가 있다. 기존의 유사도 평가방법들은 동일 레벨의 클러스터들이 배타성을 가진다는 점을 기본적으로 전제 하기 때문에, 클러스터들 간의 중복을 허용한 환경에서는 정확도가 대체적으로 낮은 결과를 보인다. 실제로 [5]에 따르면 single linkage와 complete linkage는 데이터의 중복이 심해질수록 정확도가 감소한다. 이들과 달리 average linkage는 데이터의 중복도가 높아질수록 정확도가 증가하는 경향을 보인다. 하지만 average linkage는 계산량이 너무 많아져 시간적 성능이 많이 떨어진다. 따라서 본 논문에서는 average linkage를 변형하여 중복된 데이터들에 대해 다른 유사도 평가방법들보다 정확하고, average linkage의 단점인 시간 효율성을 개선시킨 클러스터 간 유사도 평가방법을 제안하고자 한다.

2절에서는 일반적인 계층적 클러스터링에서 사용되는 유사도 평가방법들인 single linkage, complete linkage, average linkage에 대해 설명하고 3절에서는 클러스터 간의 중복을 허용하는 계층적 클러스터링 기법에 특화된 유사도 평가방법에 대해 제안하겠다. 4절에서는 제안한 유사도 평가방법과 기존 유사도 평가방법들을 비교 실험하여 제안하는 유사도 평가방법의 타당성과 효율성을 증명하고 마지막 5절에서 결론 및 향후 연구 방향에 대해 서술하겠다.

2. 관련 연구

유사도 평가방법(linkage metric)은 계층적 클러스터링의 클러스터 간 병합 및 분리를 위한 두 유사 클러스터를 선정하기 위해 사용된다. 클러스터 간 유사도 평가 방법에는 대표적으로 세 가지 방법, single linkage, complete linkage, average linkage가 있다.

첫 번째, single linkage는 두 클러스터 간 유사도를 측정 할 때 클러스터 간의 데이터 쌍 중 가장 유사도가 높은 데이터 쌍의 유사도 값을 두 클러스터 간의 유사도라 정의하며 식은 (식 1)과 같다[1].

$$l_{single}(C_p, C_q) = \max(s(p,q)),\ p \in C_p, q \in C_q \quad \text{(식 1)}$$

$s(p,q)$는 데이터 p와 q 간의 유사도이고 C_p와 C_q는 유사도를 평가 할 두 클러스터를 의미한다. single linkage는 가장 간편하게 사용되는 방식이지만 특유의 "chaining phenomenon"으로 인해 노이즈에 많이 민감하다는 단점이 있다[6].

$$l_{complete}(C_p, C_q) = \min(s(p,q)),\ p \in C_p, q \in C_q \quad \text{(식 2)}$$

두 번째, complete linkage는 (식 2)와 같이 두 클러스터 간의 데이터 쌍 중 가장 작은 유사도를 두 클러스터 간의 유사도라 정의한다[2]. single linkage에 비해 노이즈에 덜 민감하지만, 데이터의 분포 및 실제 유사도와 상관없이 비슷한 크기의 여러 클러스터들이 생성된다는 단점이 있다[7].

마지막 세 번째, average linkage는 (식 3)과 같이 두 클러스터 간의 모든 데이터 쌍들의 평균을 두 클러스터의 유사도라 정의한다[3]. $|C_p|$는 클러스터 C_p 내의 데이터의 개수를 의미한다.

$$l_{average}(C_p, C_q) = \frac{1}{|C_p||C_q|} \sum_{p_i \in C_p} \sum_{q_j \in C_q} s(p_i, q_j) \quad \text{(식 3)}$$

average linkage는 클러스터 내의 모든 데이터들을 고려하기 때문에, single linkage와 complete linkage가 가지는 정확도 저하 문제들이 없어 비교적 정확도가 높지만, 수학적 전수조사 기법이기 때문에 시간복잡도가 높다는 단점이 있다.

실제로 [5]에서는 데이터의 중복 정도에 따라 여러 유사도 평가방법들의 성능을 테스트하였다. 실험 결과에 따르면 average linkage 방식은 데이터의 중복 비율이 높아질수록 정확도가 증가하는 추세를 보였다. 하지만, average linkage는 클러스터 간의 중복을 허용한 환경에서 중복된 데이터에 대한 불필요한 계산이 많이 일어나서 전체적인 클러스터링의 시간적 성능을 급격히 저하시킨다. 따라서 본 논문에서는 average linkage의 컨셉을 차용하여 수식을 개선시켜 클러스터 간 중복을 허용한 환경에서 다른 linkage method와 비슷하거나 더 높은 정확도를 보이고, 기존의 average linkage보다는 시간적 효율성을 높인 클러스터 간 유사도 평가방법을 제안하겠다.

3. 중복을 허용한 계층적 클러스터링 기법에서 계산량을 개선한 클러스터 간 유사도 평가방법

average linkage는 클러스터 간의 중복을 허용했을 경우 중복된 데이터에서 불필요한 계산이 많이 일어나게 된다. average linkage는 유사도 평가를 위해 모든 데이터 쌍 간 유사도들의 평균을 구하게 되는데, 이 과정에서 중복된 데이터는 자기 자신이 속해있는 클러스터 내의 데이터들과의 유사도도 계산하게 된다.

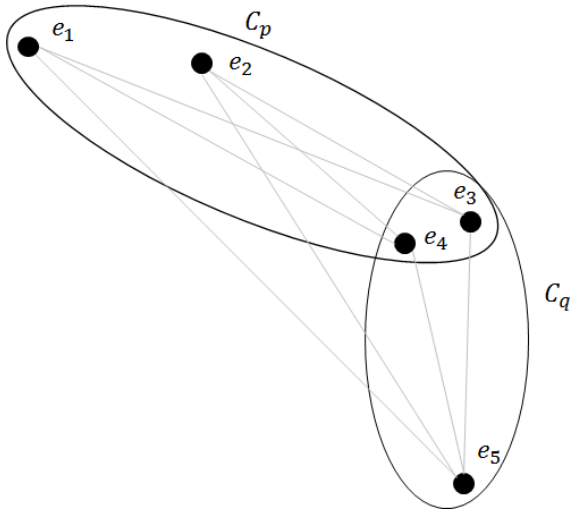


그림 1. average linkage의 중복 데이터 처리

[그림 1]은 C_p와 C_q 두 클러스터 간의 중복이 일어난 상황에서 average linkage의 계산을 도식화 한 것이다. 이때 각각의 점은 데이터를 의미하고, 점들을 포함하고 있는 타원형들은 클러스터를 의미한다. 데이터 간의 선은 두 클러스터 간의 유사도를 평가 할 때 사용되는 데이터 간 유사도이다. 이때, 선의 개수는 클러스터 간 유사도 평가에 필요한 데이터 간 유사도 계산 횟수를 뜻한다. average linkage 공식에 따르면, 중복되지 않은 부분의 데이터인 e_1, e_2, e_5는 서로 다른 클러스터와만 계산이 일어난다. e_1과 e_2는 클러스터 C_q의 데이터들과만 계산을 하게 되고, e_5는 C_p에 속한 데이터들과만 계산을 한다. 그에 반해 중복된 부분인 e_3과 e_4는 C_p, C_q 두 클러스터의 데이터들과 모두 유사도 계산을 진행하게 된다. 동일 클러스터와의 계산은 물론이고 자신과의 계산도 중복하여 일어난다. 그에 따라 시간적 효율성을 저하시키는 불필요한 계산을 제거한 중복 특화 평균 유사도 평가방법(Overlap Specialized Average linkage Metric: OSAM)을 (식 4)와 같이 제안한다.

$$l_{OSAM}(C_p,C_q)=\frac{\overbrace{\sum_{e_i\in C_p-C_q}\sum_{e_j\in C_q-C_p}s(e_i,e_j)}^{(1)}+\overbrace{(0.1713*|C_p\cup C_q|+7.1809)|C_p\cap C_q|}^{(2)}}{\underbrace{|C_p-C_q||C_q-C_p|}_{(3)}+\underbrace{|C_p\cup C_q||C_p\cap C_q|}_{(4)}} \quad \text{(식 4)}$$

(식 4)의 $s(e_i,e_j)$는 데이터 e_i와 e_j 간의 유사도(similarity)이다. 데이터 간의 유사도를 측정하는 방법들 중 코사인 유사도(cosine similarity)가 다른 유사도 측정법들에 비해 고차원 데이터에서 효과적으로 작동한다고 알려져 있으므로[8] 본 논문에서의 데이터 간 유사도 측정은 코사인 유사도를 이용했다.

제안하는 유사도 평가방법인 (식 4)는 분모와 분자가 좌우측으로 나뉘어져 있다. 그 중 좌측 부분, (1)과 (3)은 중복되지 않은 데이터들을 계산하는 부분들이고, 우측 부분 (2), (4)는 중복된 데이터들을 계산하는 부분들이다. (1)과 (3)은 기존의 average linkage와 계산 방식이 동일하게 해당하는 모든 데이터들의 평균을 계산한다. (4)는 두 클러스터에 나온 데이터들의 수와 두 클러스터에 중복되는 데이터들의 수의 곱이고, (2)은 두 클러스터에 나온 데이터들의 수의 선형식과 중복되는 데이터들 수의 곱이다.

만일 중복된 데이터가 존재하지 않으면, (2)와 (4) 가 0이 되어 (식 5)와 같이 (1)과 (3)만이 남게 된다. 게다가 $|C_p-C_q|$가 $|C_p|$로 변형되고 $|C_q-C_p|$가 $|C_q|$로 변형되어 (식 5)는 일반적인 average linkage 방법과 동일한 식으로 변하게 된다.

$$l_{OSAM}(C_p,C_q)=\frac{\sum_{e_i\in C_p-C_q}\sum_{e_j\in C_q-C_p}s(e_i,e_j)}{|C_p-C_q||C_q-C_p|} \quad \text{(식 5)}$$

한 편, [그림 1]과 같이 중복된 데이터가 있다면 (2)와 (4)를 전체 계산에 포함하여 진행하게 되는데, OSAM을 이용하여 중복된 데이터가 있는 클러스터 간의 유사도를 평가하게 되면 [그림 2]와 같이 중복된 부분에 의한 계산이 필요하지 않아, 전체적인 계산량을 크게 줄일 수 있다.

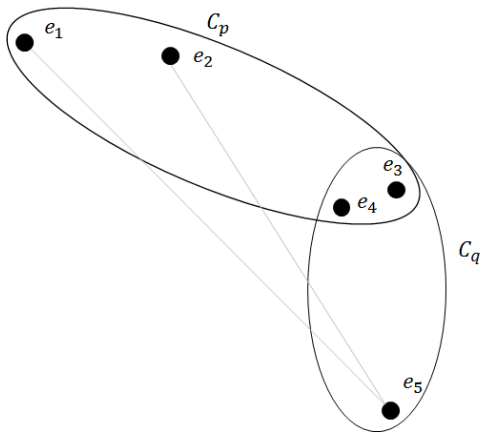


그림 2. OSAM의 중복 데이터 처리

OSAM은 average linkage 공식을 전개하여 변형시킨 것으로 공식의 변형 유도과정은 [그림 3]과 같다.

$$l_{average}(C_p,C_q)=\frac{1}{|C_p||C_q|}\sum_{e_i\in C_p}\sum_{e_j\in C_q}s(e_i,e_j) \quad \cdots\cdots ①$$

$$=\frac{\sum_{e_i\in C_p}\sum_{e_j\in C_q}s(e_i,e_j)}{(|C_p-C_q|+|C_p\cap C_q|)(|C_q-C_p|+|C_p\cap C_q|)} \quad \cdots\cdots ②$$

$$=\frac{\sum_{e_i\in C_p}\sum_{e_j\in C_q}s(e_i,e_j)}{|C_p-C_q||C_q-C_p|+|C_p\cap C_q|\times(|C_p-C_q|+|C_q-C_p|+|C_p\cap C_q|)} \quad \cdots\cdots ③$$

$$=\frac{\sum_{e_i\in C_p}\sum_{e_j\in C_q}s(e_i,e_j)}{|C_p-C_q||C_q-C_p|+|C_p\cap C_q||C_p\cup C_q|} \quad \cdots\cdots ④$$

$$=\frac{s(e_{p_0},e_{q_0})+s(e_{p_0},e_{q_1})+s(e_{p_0},e_{q_2})+\cdots+s(e_{p_0},e_{q_n})+s(e_{p_1},e_{q_0})+\cdots+s\left(+s(e_{p_m},e_{q_n})\right)}{|C_p-C_q||C_q-C_p|+|C_p\cap C_q||C_p\cup C_q|} \quad \cdots\cdots ⑤$$

$$=\frac{\sum_{e_i\in C_p-C_q}\sum_{e_j\in C_q-C_p}s(e_i,e_j)+\sum_{e_i\in C_p\cap C_q}\sum_{e_j\in(C_p\cup C_q+C_p\cap C_q)}s(e_i,e_j)}{|C_p-C_q||C_q-C_p|+|C_p\cup C_q||C_p\cap C_q|} \quad \cdots\cdots ⑥$$

$$=\frac{\sum_{e_i\in C_p-C_q}\sum_{e_j\in C_q-C_p}s(e_i,e_j)+(|C_p\cup C_q|+|C_p\cap C_q|)\times\rho\times|C_p\cap C_q|}{|C_p-C_q||C_q-C_p|+|C_p\cup C_q||C_p\cap C_q|} \quad \cdots\cdots ⑦$$

$$\rho=\frac{\sum_{e_i\in C_p\cap C_q}\sum_{e_j\in(C_p\cup C_q+C_p\cap C_q)}s(e_i,e_j)}{|C_p\cap C_q||C_p\cup C_q+C_p\cap C_q|} \quad \cdots\cdots ⑧$$

그림 3. average linkage 변형

①에서 분모 $|C_p|$는 $|C_p-C_q|+|C_p\cap C_q|$로 전개 할 수 있고 $|C_q|$ 또한 같은 방식을 이용해 ②로 전개 가능하다. ②를 전개하여 중복된 데이터들 계산에 해당하는 부분인 $|C_p\cap C_q|$로 묶어보면(③), 묶인 부분인 $|C_p-C_q|+|C_q-C_p|+|C_p\cap C_q|$는 $|C_p\cup C_q|$와 같으므로 식은 ④와 같이 변형된다. 이때, 분모의 좌측 부분은 중복이 아닌 데이터들만의 연산, 우측 부분은 중복인 데이터에 의한 연산으로 나뉘었다. 분자 부분을 ⑤로 전개하면, 분자도 분모와 같이 중복된 부분의 식과 중복되지 않은 식으로 식 ⑥과 같이 나눌 수 있다. 중복되는 부분 데이터는 양쪽 클러스터에 모두 속해있기 때문에, 모든 데이터와 유사도 계산을 하게 된다. 중복된 부분의 데이터들이 가지는 모든 유사도의 평균 ρ(⑧)을 통해 정리하면 ⑦과 같은 식을 얻을 수 있다.

$$(|C_p\cup C_q|+|C_p\cap C_q|)\times\rho \quad \text{(식 6)}$$

$$|C_p\cup C_q| \quad \text{(식 7)}$$

이 때, 중복 부분의 분자와 분모에 공통적으로 있는 $|C_p\cap C_q|$는 분자에서 (식 6), 분모에서 (식 7)과 곱해진다. 계층적 클러스터링의 특성 상, 클러스터링이 진행됨에 따라 $|C_p\cup C_q|$와 $|C_p\cap C_q|$의 크기는 커지고 ρ는 감소하게 된다. 따라서 (식 6)와 (식 7)이 어떠한 상관관계를 가질 것이라 예상된다. average linkage를 이용한 중복 클러스터링을 진행하면서 각 클러스터 간 유사도 계산마다 두 식의 값을 추출했다. 추출된 값들을 10-fold cross validation하고 선형회귀를 통해 correlation coefficient가 0.8227인 (식 8)을 도출해냈다.

$$(|C_p\cup C_q|+|C_p\cap C_q|)\times\rho = 0.1713*|C_p\cup C_q|+ 7.1809 \quad \text{(식 8)}$$

(식 8)을 이용해 [그림 3]의 ⑦을 최종적으로 제안하는 중복에 특화된 평균 유사도 평가 방법(OSAM)인 (식 9)로 변형한다.

$$l_{OSAM}(C_p, C_q) = \frac{\left(\sum_{e_i \in C_p - C_q} \sum_{e_j \in C_q - C_p} s(e_i, e_j)\right) + (0.1713 * |C_p \cup C_q| + 7.1809)|C_p \cap C_q|}{|C_p - C_q||C_q - C_p| + |C_p \cup C_q||C_p \cap C_q|} \quad \text{(식 9)}$$

OSAM의 중복된 데이터 처리는 서로 다른 주제 간의 중복이 많은 데이터들을 클러스터링 할 때 더 좋은 효율을 보인다. 서로 다른 주제 간의 중복이 적은 데이터들을 클러스터링 할 때 (식 8)를 이용해 구한 선형식의 MSE(Mean Square Error)가 높아지지만, 중복된 데이터의 양이 적기 때문에 그 부분이 OSAM과 전체 클러스터링의 결과물에 끼치는 영향도 적어서 감수 할 수 있다.

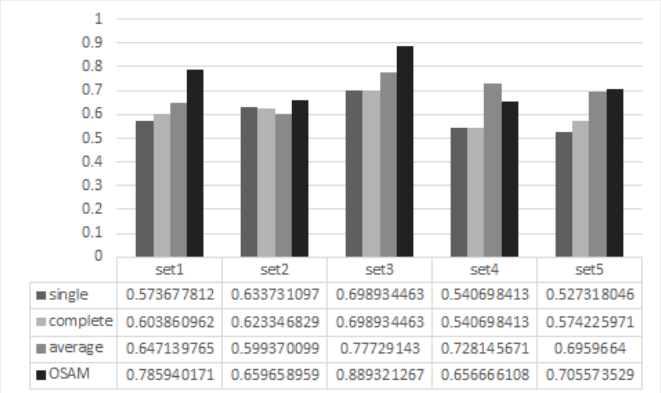
4. 실험 및 평가

성능평가 실험에는 중복을 허용하는 여러 계층적 클러스터링 기법들 중 구현이 간편하고 한글을 적용하여 클러스터링을 실시한 최근 논문인 [9]에서 제안한 HOC를 클러스터링 방법으로 사용하고, 데이터는 5개 주제로 된 문서 100개씩 총 500개의 문서를 100개씩 임의로 선택하여 bag of words로 변형한 뒤 이용하였다[그림 4].

주제	출처	문서 수	단어 수
동물	NDSL	100	138579
방위산업	NDSL	100	139768
질병	NDSL	100	188106
컴퓨터	NDSL	100	179574
물리학	NDSL	100	143513

그림 4. 실험 문서 집합 설명

OSAM의 성능평가를 위해 [그림 5]와 같이 사용된 클러스터 간 유사도 평가방법에 따른 f-measure 값을 비교하였다. OSAM을 적용한 클러스터링의 결과는 single linkage, complete linkage보다 더 높은 f-measure 값을 얻을 수 있었다.



	set1	set2	set3	set4	set5
single	0.573677812	0.633731097	0.698934463	0.540698413	0.527318046
complete	0.603860962	0.623346829	0.698934463	0.540698413	0.574225971
average	0.647139765	0.599370099	0.77729143	0.728145671	0.6959664
OSAM	0.785940171	0.659658959	0.889321267	0.656666108	0.705573529

그림 5. 유사도 평가방법에 따른 f-measure

OSAM은 기본적으로 average linkage를 변형한 유사도 평가방법이기 때문에, average linkage와 유사한 f-measure 결과를 보이게 된다. 특히, average linkage보다는 [그림 5]의 set 4번 결과와 같이 더 낮은 f-measure 값을 얻는 경우도 있다. 하지만 [그림 6]과 같이 OSAM을 이용한 클러스터링은 average linkage을 이용한 클러스터링에 비해 처리 시간이 평균적으로 절반정도 소요됐다. 중복된 데이터를 처리함에 있어 average linkage는 데이터 간 유사도 계산을 반복적으로 진행해야 했지만, OSAM은 중복된 데이터를 처리 할 때 한 번의 계산만을 요하기 때문에 이와 같은 차이가 발생했다.

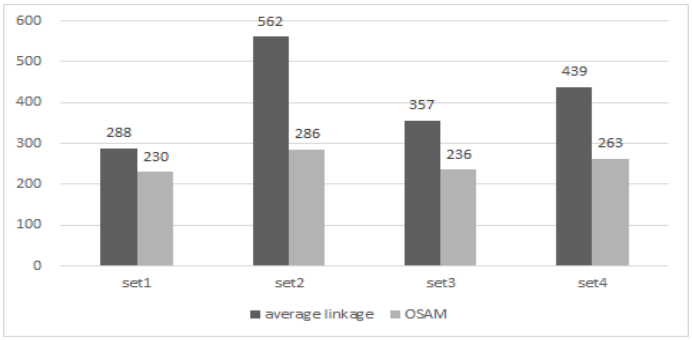


그림 6. OSAM과 average linkage의 processing time

5. 결론 및 향후연구

본 논문에서는 클러스터 간 중복을 허용한 계층적 클러스터링 기법에 특화된 클러스터 간 유사도 평가방법(OSAM)을 제안하였다. OSAM은 average linkage를 변형하여 정확성은 유지하면서도 시간적 성능은 높였다. 하지만 중복된 데이터에 대한 계산량을 줄이는 식을 만들기 위해 사용한 선형회귀 방식의 특성 상 데이터를 10-fold cross validation을 진행했음에도 불구하고 overfitting 여부에 대한 완전한 검증이 이뤄지지 않아 더 많은 데이터를 통한 검증이 필요하다. 또한 제안한 유사도 평가방법을 중복을 허용하는 여러 계층적 클러스터링 기법들에도 적용해보도록 하겠다.

참고문헌

[1] Robin Sibson, SLINK: an optimally efficient algorithm for the single-link cluster method, The computer journal 16.1: 30-34, 1973

[2] Daniel Defays, An efficient algorithm for a complete link method, The Computer Journal 20.4: 364-366, 1977

[3] Robert R. Sokal, A statistical method for evaluating systematic relationships, Univ Kans Sci Bull 38: 1409-1438, 1958

[4] Pavel Berkhin, A survey of clustering data mining techniques, Grouping multidimensional data. Springer Berlin Heidelberg, 25-71, 2006

[5] Chris Ding, and He Xiaofeng, Cluster merging and splitting in hierarchical clustering algorithms, Data Mining, ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002
[6] T. Narang, Hiearchical clustering of documents: A brief study and implementation in Matlab, Proceedings of the International Conference on Emerging Trends of Engineering, Science, Management and Its Applications, 2015.
[7] Brian S. Everitt, et al, Cluster analysis, SSRC Reviews of Current Research 11, 2011.
[8] Amit Singhal, Modern Information Retrieval: A Brief Overview, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35-43, 2001
[9] 홍수정, 최중민, 중복을 허용한 계층적 클러스터링에 의한 복합 개념 탐지 방법, 지능정보연구, 17.1: 111-125, 2011