텍스트 마이닝에서 근접성 데이터의 군집화 기법 연구

Author and Position

- ❑ *정민지 : 석사과정, 부산대학교 통계학과
- ❑ 정호영 : 석사과정, 부산대학교 통계학과
- ❑ 이수진 : 석사과정, 부산대학교 통계학과
- ❑ 신상민 : 조교수, 동아대학교 경영정보학과
- ❑ 최용석 : 교수, 부산대학교 통계학과

1. Abstract

- 텍스트 마이닝 분야에서 보편적으로 사용하는 문서-용어 빈도행렬(document-term frequency matrix)은 문서에서 추출된 용어의 발생빈도를 원소로 하며 행과 열에 문서와 용어 정보가 나열된 행렬을 의미한다.

- 분석하고자 하는 개체가 존재하고 개체들이 제공하는 여러 개의 문서들을 바탕으로 만들어진 문서-용어 행렬은 흔히 볼 수 있는 그룹화 된 형태의 데이터라 할 수 있다.

- 개체 정보가 존재하는 문서-용어 빈도행렬로 부터 개체에 대한 분석을 진행할 경우 일반적으로 각 개체에 속하는 문서들이 공통으로 가지고 있는 용어 만을 추출하여 핵심어(keyword)라 하고 핵심어들의 빈도 값을 더하거나 평균 낸 새로운 개체-핵심어 빈도행렬을 이용한다. 그러나 이 경우 개별 문서들이 가진 고유한 정보들이 삭제될 뿐만 아니라, 개체 내 여러 문서들에서 공통적으로 등장하진 않지만 발생빈도가 높아 중요하게 여겨야 할 핵심어들이 삭제되는 문제가 발생한다.

- 본 연구에서는 개체 정보가 존재하는 문서-용어 행렬에 TF-IDF(Term Frequency-Inverse Document Frequency) 분석을 적용하고 평균 가중점수가 높은 상위 용어들을 핵심어로 필터링하는데 네 가지 방법을 사용하고자 한다. 만들어진 문서-핵심어 가중행렬에 세 가지 거리측도를 적용하여 최종적으로 근접성 데이터(proximity data)를 생성하고 시각화함으로써 텍스트 데이터의 군집화에 적합한 근접성 데이터 생성 방법을 제안하고자 한다.

2. Proximity data

of Institutes

◆ 문서-용어 빈도행렬로부터 근접성 데이터를 생성하는 과정은 다음과 같다.

[1단계] 개체의 수를 g라 하고 개체 별 문서의 수를 $n_r,\ r = 1, \cdots, g$이라 하면 전체 문서의 수는 $n = \sum_{r=1}^{g} n_r$과 같다. n개 문서에서 적어도 한번 이상 추출된 용어의 개수를 p라 하면 크기가 $n \times p$이고 개체 정보가 존재하는 문서-용어 빈도행렬 Y는 다음과 같이 표현된다.

Document-term freq. matrix : $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_g \end{bmatrix}$

Y는 크기가 $n_r \times p$인 부분행렬 Y_r로 이루어져 있다. 여기서 r번째 개체의 l번째 문서를 나타내는 Y_r의 행벡터는 $y_{rl} = (y_{rl1}, y_{rl2}, \cdots, y_{rlp})^t$와 같이 정의할 수 있다. 따라서 문서-용어 빈도행렬은 개체 정보를 무시한 행렬 $Y = (y_{ij}), i = 1, \cdots, n$ 또는 개체 정보를 고려한 행렬 $Y = (y_{rlj}), j = 1, \cdots, p$ 로 표현이 가능하다.

[2단계] 텍스트 데이터를 보다 타당하게 분석하기 위하여 용어에 부여하는 가중치 계산법인 TF-IDF 분석법을 수식으로 표현하면 다음과 같다.

Total # of documents

Freq. of jth term in ith document

$$TFIDF(i,j) = TF(i,j) \times \log\left(\frac{n}{DF(n,j)}\right)$$

Total # of documents including jth term

용어의 중요도는 i번째 문서에서 추출한 j번째 용어의 발생 빈도인 $TF(i,j)$에 비례하고 n개의 문서 중 j번째 용어가 들어가는 문서의 총 수인 $DF(n,j)$에는 반비례한다. 즉 문서 내에서 많이 등장하되 전체 문서를 기준으로 보았을 때 해당 용어가 포함된 문서의 수는 적을수록 용어의 중요도는 높아진다.
TF-IDF를 문서-용어 빈도행렬에 적용시키는 경우 개체 정보를 무시하는지 혹은 고려하는지에 따라 두 가지 방법으로 가중치 부여가 가능하다.

Document-term weighted matrix Z :

$$z_{rlj} = y_{rlj} \times w_j, \quad w_j = \log\left(\frac{n}{DF(n,j)}\right) \qquad (1)$$

$$z_{rlj} = y_{rlj} \times w_{rj}, \quad w_{rj} = \log\left(\frac{n_r}{DF(n_r,j)}\right) \qquad (2)$$

따라서 [2단계]를 통해 문서-용어 빈도행렬 Y에서 두 가지 방법으로 가중치를 부여한 문서-용어 가중행렬 Z가 도출된다.

[3단계] 문서-용어 가중행렬 Z에서 각 용어들이 갖는 평균 가중점수를 기준으로 점수가 높은 상위 용어를 핵심어로 선정하고자 한다. 평균 가중점수는 두 가지 방법으로 산출이 가능하다.

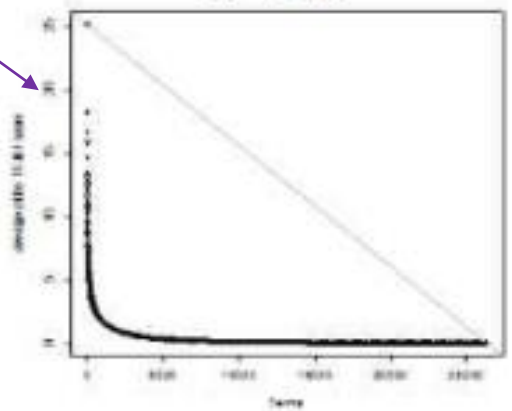
$$\bar{\mathbf{z}} = (\bar{z}_1, \bar{z}_2, \cdots, \bar{z}_p)^t = n^{-1}\mathbf{Z}^t\mathbf{1}_n \quad (3)$$

$$\bar{\mathbf{z}}_r = (\bar{z}_{r1}, \bar{z}_{r2}, \cdots, \bar{z}_{rp})^t = n_r^{-1}\mathbf{Z}_r^t\mathbf{1}_{n_r} \quad (4)$$

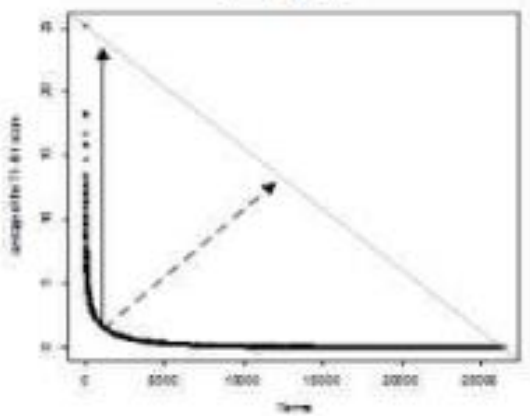
다음으로 평균 가중점수를 내림차순 한 그래프 상에서 팔꿈치 지점(elbow point)를 찾는다. 아래 그림 1과 같이 그래프의 양 끝 지점을 관통하는 직선을 만든 후에 그림 2와 같이 그래프와 직선 사이의 거리가 가장 먼 지점을 팔꿈치 지점으로 정의한다(Satopaa, Ville, et al, 2011).

$\bar{z}_{(1)} \geq \bar{z}_{(2)} \geq \cdots \geq \bar{z}_{(p)}$

[그림 1]



[그림 2]



식 (4)의 경우 각 부분행렬 Z_r에서 평균 가중점수가 높은 상위 용어를 선별하고, 용어들의 중복을 제거하여 문서-핵심어 가중행렬 X을 생성한다.
따라서 [2단계]의 식 (1), (2)와 [3단계]의 식 (3), (4)를 조합한 네 가지 방법으로 크기가 $n \times q$인 문서-핵심어 가중행렬 X를 생성한다.

Document-keyword weighted matrix X

Dissimilarity matrix between documents D

[4단계] 문서-핵심어 가중행렬 X에서 r번째 개체의 l번째 문서 $x_{rl} = (x_{rl1}, \cdots, x_{rlq})^t, l = 1, \cdots, n_r$과 s번째 개체의 m번째 문서 $x_{sm} = (x_{sm\,1}, \cdots, x_{sm\,q})^t, m = 1, \cdots, n_s$에 아래 세 가지 거리측도를 적용하여 크기가 $n \times n$인 비유사성 행렬 D를 생성한다.

$$\text{Euclidean Distance} : d_{ED}(x_{rl}, x_{sm}) = \left[\sum_{t=1}^{q}(x_{rt} - x_{sm\,t})^2\right]^{1/2},\ t = 1, \cdots, q \quad (5)$$

$$\text{Chi-square Distance} : d_{CD}(x_{rl}, x_{sm}) = \left[\sum_{t=1}^{q}(x_{rt} - x_{sm\,t})^2 / (x_{rt} + x_{sm\,t})\right]^{1/2} \quad (6)$$

$$\text{Weighted Euclidean Distance} : d_{W\,ED}(x_{rl}, x_{sm}) = \left[\sum_{t=1}^{q}\left(\frac{x_{rt}}{x_{rl\cdot}} - \frac{x_{sm\,t}}{x_{sm\cdot}}\right)^2 / \left(\frac{x_{rt}}{x_{rl\cdot}} + \frac{x_{sm\,t}}{x_{sm\cdot}}\right)\right]^{1/2} \quad (7)$$

문서 간 비유사성 행렬 D는 g^2개의 부분행렬 $D_{rs} = (d(x_{rl}, x_{sm})),\ r, s = 1, \cdots, g$로 이루어진 행렬이며 D_{rs}는 X_r과 X_s 사이의 비유사성을 의미한다.

$$D = \begin{bmatrix} D_{11} & D_{12} & \cdots & D_{1g} \\ D_{21} & D_{22} & \cdots & D_{2g} \\ \vdots & \vdots & \ddots & \vdots \\ D_{g1} & D_{g2} & \cdots & D_{gg} \end{bmatrix}$$

: Grouped Distance Matrix

[5단계] 비유사성 행렬 D에서 계산상의 편의를 위해 $\tilde{D} = \left(\tilde{d}(x_{rl}, x_{sm})\right) = \left(d^2(x_{rl}, x_{sm})/2\right)$를 정의한다. 크기가 $g \times g$인 대각행렬 $N = diag\ (n_1, \cdots, n_g)$과 크기가 $n_r \times g$인 $G_r = (g_{lr})$을 정의하는데, G_r은 l번째 문서가 r번째 개체에 속할 경우에만 1의 값을 가지고 그 외의 경우는 0의 값을 갖는 행렬 G의 부분행렬이다.
행렬 $\tilde{D}$, N, G를 이용해 행렬 $F = N^{-1}G^t\tilde{D}GN^{-1} = (f_{rs})$를 계산한다. 즉 $f_{rs} = \tilde{D}_{rs}n_r^{-1}n_s^{-1}$이며 f_{rs}는 r번째 개체와 s번째 개체 간 비유사성을 나타내는 부분행렬 $\tilde{D}_{rs}$의 평균이 된다.
r번째 개체와 s번째 개체의 평균좌표 간 거리를 ψ_{rs}라 가정하면

$$\psi_{rs}^2 = 2f_{rs} - f_{rr} - f_{ss}$$

이 성립한다. 따라서 개체 평균 사이의 비유사성을 나타내고 크기가 $g \times g$인 행렬 Ψ는 아래와 같다. 이렇게 만들어진 행렬 Ψ를 근접성 데이터라 하자.

Proximity data matrix : $\Psi = \left(\frac{1}{2}\psi_{rs}^2\right)$

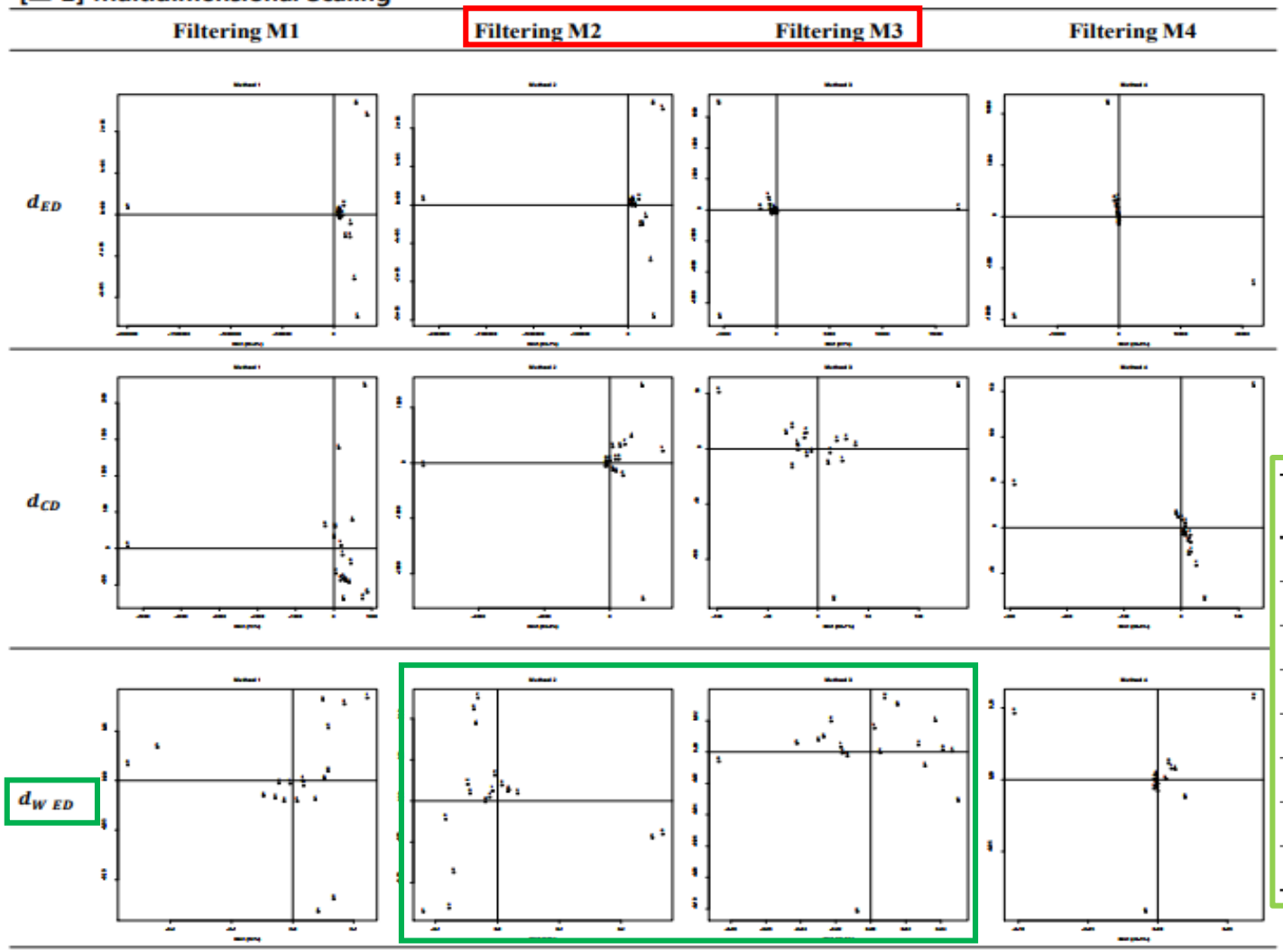
- [2~5단계]를 거쳐 생성된 12개의 근접성 데이터 Ψ를 정리한 표는 아래와 같다.

[표 1] 12개 근접성 데이터

			Filtering M1			Filtering M2			Filtering M3			Filtering M4		
Filtering Method 1~4	[2단계]	$Y \rightarrow Z$	(1)			(1)			(2)			(2)		
	[3단계]	$Z \rightarrow X$	(3)			(4)			(3)			(4)		
Distance 5~7	[4~5단계]	$X \rightarrow D \rightarrow \Psi$	(5)	(6)	(7)	(5)	(6)	(7)	(5)	(6)	(7)	(5)	(6)	(7)

Metric MDS for clustering

[표 2] Multidimensional Scaling



3. Illustrations

♦ 경제·인문 분야 정부출연연구기관 25개 중 19개 기관의 대표적인 정기간행물을 기관별로 1개씩 선정하여 2016년 한 해 동안 온라인상으로 발간된 343개 간행물에서 추출한 26,915개 명사 용어로 이루어진 텍스트 데이터를 활용하였다.

♦ 텍스트 데이터의 수집 및 전처리를 위해 파이썬(Python) 프로그램을 사용하였으며, KoNLPy 패키기에서 꼬꼬마(Kkma) 형태소 분석기를 이용하였다. 또한 데이터 전처리 과정에서 삭제 되진 않았으나 불용어로 볼 수 있는 한 글자 용어 563개를 삭제하였다.

	기관	기관별 대표 정기간행물			
		간행물명	발간주기	PDF 파일개수	발간목적
1	과학기술정책연구원	과학기술정책	매월	12	국내외 과학기술 혁신활동 관련 현안 및 동향 분석
2	대외경제정책연구원	정책연구브리핑	수시	23	연구보고서 결과 제시 및 정책 시사점 정리
3	산업연구원	KIET 산업 동향브리프	매월	12	국내외 산업 동향 제시 및 브리핑
4	육아정책연구소	육아정책포럼	연4회	4	육아관련 현안 문제와 국내·외 육아정책동향 분석
⋮	⋮	⋮	⋮	⋮	⋮
17	한국행정연구원	행정포커스	격월	6	행정 관련 정책 및 현안에 대한 정보 수록
18	한국형사정책연구원	형사정책연구	연4회	4	한국연구재단에 등재된 형사 정책 관련 전문학술지
19	한국환경정책평가연구원	환경정책	연4회	40	환경 정책 및 현안에 대한 정보를 수록한 학술지

Visual excellent clustering

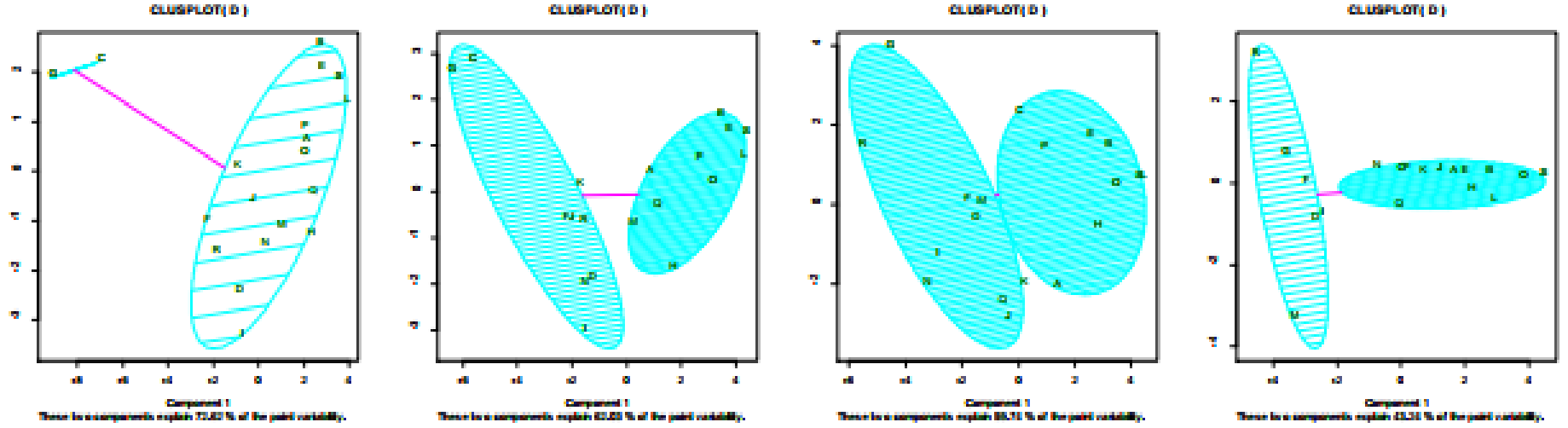
♦ 12개의 근접성 데이터 Ψ를 비교 선택하고자 계량형 다차원척도법의 대표적인 알고리즘인 토거선 알고리즘(Torgerson, 1958)에 따라 2차원으로 차원을 축소한 다차원적도법도를 표 2에 정리하였다. 다른 거리측도 보다 가중 유클리드 거리를 적용하였을 때 19개 기관에 해당하는 개체들의 시각적 군집이 극단으로 치우침이 없이 형성되어 있음을 확인하였다. 용어 필터링 방법은 개체 별 특성을 반영하는 M2 혹은 M3 방법이 권장할 만하다.

◆ 군집 내 밀집의 정도와 군집 간 분리의 정도를 나타내는 실루엣 통계량(Rousseeuw, 1987)이 1에 가까울수록 좋은 군집에 형성된다. 가중 유클리드 거리를 적용한 M1~M4 행렬에서 통계량을 산출한 결과 모든 경우에서 군집 수가 두 개일 때 최대 통계량 값을 가짐을 확인할 수 있다.

[표 3] Silhouette Statistics Value

	Filtering M1	Filtering M2	Filtering M3	Filtering M4
k=2	0.465	0.345	0.283	0.217
k=3	0.195	0.181	0.19	0.181

◆ $k = 2$로 k-means 군집 분석을 수행한 결과, M2 방법으로 용어를 필터링하고 가중 유클리드 거리로 생성한 근접성 데이터의 군집화가 가장 잘 형성됨을 확인할 수 있다.



Good Clustering