The Korean Journal of Applied Statistics (2019) DOI: http://dx.doi.org/10.5351/KJAS.2019.32.3.451
32(3), 451–462

Creation and clustering of proximity data for text data analysis

Min-Ji Jung[a] · Sang Min Shin[b] · Yong-Seok Choi[a,1]

[a]Department of Statistics, Pusan National University;
[b]Department of Management Information Systems, Dong-A University

(Received March 4, 2019; Revised April 15, 2019; Accepted April 17, 2019)

Abstract

Document-term frequency matrix is a type of data used in text mining. This matrix is often based on various documents provided by the objects to be analyzed. When analyzing objects using this matrix, researchers generally select only terms that are common in documents belonging to one object as keywords. Keywords are used to analyze the object. However, this method misses the unique information of the individual document as well as causes a problem of removing potential keywords that occur frequently in a specific document. In this study, we define data that can overcome this problem as proximity data. We introduce twelve methods that generate proximity data and cluster the objects through two clustering methods of multidimensional scaling and k-means cluster analysis. Finally, we choose the best method to be optimized for clustering the object.

Keywords: text mining, proximity data, TF-IDF, multidimensional scaling, cluster analysis

1. 서론

문서-용어 빈도행렬(document-term frequency matrix)은 행에는 문서가 열에는 문서에서 추출한 용어가 나열되고 각 용어의 발생빈도를 원소(element)로 하는 데이터이다. 분석하고자 하는 특정 개체(object)가 존재할 때 해당 개체가 제공하는 문서를 바탕으로 문서-용어 빈도행렬을 만든다. 기본적으로 한 개체 당 하나의 행렬을 구성한다. 개체가 두 개 이상인 경우 다음과 같은 두 가지 방법으로 행렬을 생성할 수 있다. 첫째, 여러 개체가 제공하는 모든 문서를 종합하여 하나의 문서-용어 빈도행렬을 만든다. 둘째, 개체들이 가지고 있는 공통 용어를 찾아내어 행에는 다수의 개체가 열에는 공통 용어가 나열된 새로운 개체-공통어 빈도행렬을 만드는 것이다. 전자는 각 개체의 특성을 반영하지 못하고, 후자는 공통어가 아닐 경우 개별 문서에서 중요한 용어라 할지라도 삭제하기 때문에 두 방법 모두 문제가 있다.

본 연구에서는 두 개 이상의 개체가 존재할 때 위와 같은 문제를 상쇄시킬 뿐만 아니라 개체 군집화에 사용할 수 있는 데이터를 근접성 데이터(proximity data)라 정의한다. 근접성 데이터는 어떤 방법으로 가중치 부여하는지 또는 어떤 거리측도를 사용하는지에 따라 12가지 방법으로 생성할 수 있다. 이어지는 2장에서 근접성 데이터 생성법을 상세히 설명한다. 3장에서는 실제 연구 자료에 근접성 데이터 생성

[1]Corresponding author: Department of Statistics, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-Gu, Busan 46241, Korea. E-mail yschoi@pusan.ac.kr

법 12가지를 적용한 후 개체 군집화에 최적화된 방법 하나를 제안한다. 개체 군집화 기법으로는 전통적으로 사용되어 왔던 다차원척도법(multidimensional scaling)과 비계층적 군집분석의 일종인 K-평균 군집분석(K-means cluster analysis)을 사용한다. 4장의 결론에서는 전체 내용을 정리 및 요약한다.

2. 근접성 데이터 생성법

근접성은 통계학에서 사용하는 개념인 유사성과 비유사성을 일컫는 단어이다. 근접성 데이터는 두 개 이상의 개체 간 근접성을 보다 효과적으로 파악하기 위해 고안된 데이터를 의미한다. 근접성 데이터라는 용어 자체는 새로운 것이 아니나 텍스트 마이닝 분야에서 개체 군집화에 최적화된 데이터로 새롭게 정의하였다는 점에서 의의가 있다.

2.1. 문서-용어 빈도행렬

개체의 수를 g라 하고 각 개체가 제공하는 문서의 수를 n_r, $r = 1, \ldots, g$라 하면 전체 문서의 수는 $n = \sum_{r=1}^{g} n_r$이 된다. i번째 문서에서 적어도 한 번 이상 추출된 용어의 개수를 p_i, $i = 1, \ldots, n$이라 하면 전체 용어의 수는 중복이 없는 경우 $\sum_{i=1}^{n} p_i$가 된다. 그러나 용어 집합에는 중복이 발생할 수 있다. 따라서 $\sum_{i=1}^{n} p_i$에서 중복을 제거한 p개의 용어를 전체 용어의 수로 정의한다. 크기가 $n \times p$이고 g개의 개체 정보가 존재하는 전체 문서-용어 빈도행렬을 $\mathbf{Y}$라 하자. 행렬 $\mathbf{Y}$를 g개의 개체별로 나눠서 표현하면 크기가 $n_r \times p$인 부분행렬 $\mathbf{Y_r}$, $r = 1, \ldots, g$로 표현할 수 있다. 부분행렬 $\mathbf{Y_r}$은 r번째 개체의 문서-용어 빈도행렬을 의미한다. r번째 개체의 l번째 문서는 $\mathbf{y_{rl}} = (y_{rl1}, y_{rl2}, \ldots, y_{rlp})^t$, $l = 1, \ldots, n_r$로 표현할 수 있다. 따라서 개체 정보가 존재하는 문서-용어 빈도행렬을 $\mathbf{Y} = (y_{rlj})$, $j = 1, \ldots, p$로 정의한다.

2.2. 문서-용어 가중행렬

Term frequency-inverse document frequency (TF-IDF)는 텍스트 데이터를 타당하게 분석하기 위하여 각 용어에 부여하는 가중치 계산법이다 (Sim과 Kim, 2016). i번째 문서에서 발생한 j번째 용어의 TF-IDF 가중치는 식 (2.1)과 같이 산출된다.

$$\text{TFIDF}(i,j) = \text{TF}(i,j) \times \log\left(\frac{n}{\text{DF}(n,j)}\right), \tag{2.1}$$

여기서 $\text{TF}(i,j)$는 i번째 문서 내에서 j번째 용어가 발생한 빈도를 뜻하며 $\text{DF}(n,j)$는 전체 n개의 문서 중에서 j번째 용어가 포함된 문서의 수를 의미한다. 이 식에 따르면 모든 문서에서 등장하기 때문에 중요도가 낮은 용어(예를 들어 '방법, 결과'와 같은 일반 명사)일 경우 $\text{TF}(i,j)$가 아무리 커도 $n/\text{DF}(n,j)$가 1이 되므로 가중치는 0이다. 즉 일반적으로 흔하게 사용되는 용어에는 낮은 가중치가 부여되고 특정 문서에서 유독 많이 사용하는 용어에는 높은 가중치가 부여되는 것이다. 그러나 단순히 TF-IDF만 사용하면 문서-용어 빈도행렬 $\mathbf{Y}$ 전체에 가중치가 일괄적으로 부여되므로 개체 정보가 무의미해지는 단점이 존재한다. 아래에서 소개할 [TF-IDF 가중치 부여 방법]에서는 단순한 TF-IDF인 [방법 W1]을 먼저 설명한 후 본 연구에서 새롭게 제안하는 개체별 용어 가중치 계산법인 [방법 W2]를 정의하고자 한다.

[TF-IDF 가중치 부여 방법]

· [방법 W1]: 문서-용어 빈도행렬 $\mathbf{Y}$의 원소 y_{rlj}에 가중치 w_j를 곱하여 가중점수 z_{rlj}를 산출한다.

$$z_{rlj} = y_{rlj} \times w_j, \quad w_j = \log\left(\frac{n}{\text{DF}(n,j)}\right). \tag{2.2}$$

행렬 $\mathbf{Y}$에 일괄적으로 용어 가중치를 부여한 문서-용어 가중행렬 $\mathbf{Z} = (z_{rlj})$가 도출된다.

· [방법 W2]: 문서-용어 빈도행렬 $\mathbf{Y}$의 원소 y_{rlj}에 개체별 가중치인 w_{rj}를 곱하여 가중점수 z_{rlj}를 산출한다.

$$z_{rlj} = y_{rlj} \times w_{rj}, \quad w_{rj} = \log\left(\frac{n_r}{\mathrm{DF}(n_r, j)}\right). \tag{2.3}$$

행렬 $\mathbf{Y}$에 개체별 용어 가중치를 부여하므로 각 개체의 특성이 담긴 문서-용어 가중행렬 $\mathbf{Z} = (z_{rlj})$가 도출된다.

2.3. 문서-핵심어 가중행렬

문서-용어 가중행렬 $\mathbf{Z}$에서 가중치 합이 0인 용어를 삭제하면 핵심어만 남는다. 삭제되는 용어의 개수가 어느 정도 보장이 된다면 TF-IDF만으로 핵심어 선별이 가능하다고 할 수 있다. 그러나 가중 합이 0인 용어보다 0이 아닌 매우 작은 값을 갖는 용어가 존재할 가능성이 더 크다. 이때에는 모든 용어를 사용하기보다 가중 합이 극히 낮은 용어는 삭제하고 핵심어만 선별하여 분석할 필요가 있다. 따라서 문서-용어 가중행렬 $\mathbf{Z}$에서 각 용어들이 갖는 평균적인 가중점수를 산출한 후 점수가 급격하게 감소하는 지점, 즉 팔꿈치 지점(elbow point)을 찾는다. 팔꿈치 지점을 기준으로 점수가 높은 상위 용어를 핵심어로 선정한다. 정리하면 가중행렬 $\mathbf{Z}$의 p개 용어로부터 q개의 핵심어를 필터링(filtering)하여 문서-핵심어 가중행렬 $\mathbf{X}$를 도출하는 것이다.

팔꿈치 지점을 활용한 용어 필터링 방법은 Cho 등 (2015)이 이미 소개한 방법이나, 연구자의 주관적인 판단으로 팔꿈치 지점을 결정하므로 연구자에 따라 선별되는 핵심어가 달라지는 문제가 발생한다. 따라서 본 연구에서는 Satopaa 등 (2011)의 Kneedle algorithm을 활용하여 그래프의 변곡점을 찾고 이를 팔꿈치 지점으로 정의한다. 변곡점은 굴곡의 방향이 바뀌는 곡선 위의 점이기 때문에 평균 가중점수가 급격하게 감소하는 팔꿈치 지점으로 간주할 수 있다. 변곡점을 활용하면 핵심어를 선정하는 방법에 대한 객관성 역시 확보할 수 있다. 이어지는 [용어 필터링 방법]에서 용어를 필터링하는 두 가지 방법을 소개하고 핵심어를 선정하는 과정을 상세히 설명하고자 한다.

[용어 필터링 방법]

· [방법 F1]: 문서-용어 가중행렬 $\mathbf{Z}$에서 각 용어의 평균 가중점수로 이루어진 평균 벡터 $\bar{\mathbf{z}}$를 계산한다.

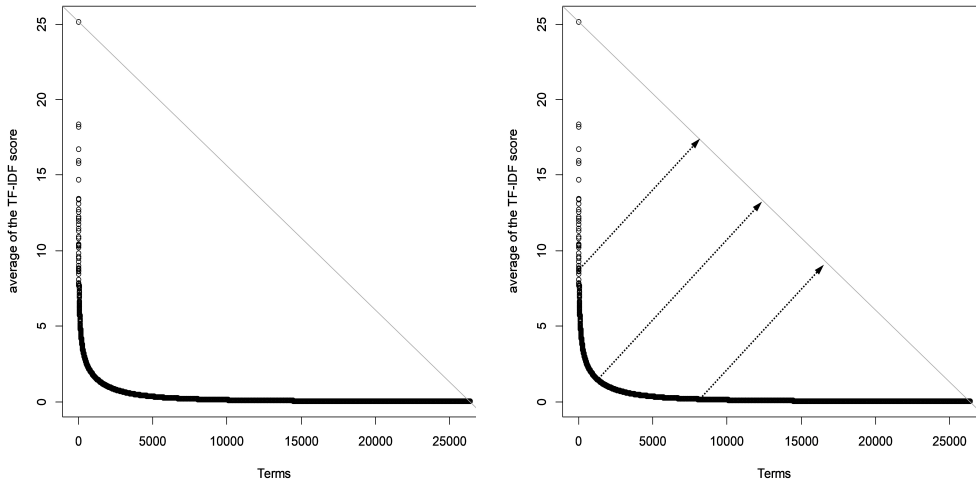
$$\bar{\mathbf{z}} = (\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_p)^t = n^{-1}\mathbf{Z}^t\mathbf{1}_n. \tag{2.4}$$

용어별 평균 가중점수의 추이를 그래프로 확인하기 위하여 p개의 용어를 x축으로 $\bar{\mathbf{z}}$의 원소를 y축으로 하는 그래프를 그린다. 여기서 그래프의 처음과 끝 지점을 관통하는 직선을 그어주면 Figure 2.1의 (a)와 같다.

Figure 2.1의 (b)와 같이 그래프에서 직선까지 이르는 거리가 최대가 되는 그래프 상의 지점을 변곡점, 즉 팔꿈치 지점으로 간주한다. 팔꿈치 지점을 기준으로 평균 가중점수가 높은 상위 용어 q개를 핵심어로 선정한다. 이를 통해 행렬 $\mathbf{Z}$로부터 문서-핵심어 가중행렬 $\mathbf{X} = (x_{rlt})$, $t = 1, \ldots, q$를 얻는다.

· [방법 F2]: 문서-용어 가중행렬 $\mathbf{Z}$ 전체를 기준으로 하는 [방법 F1]과 달리 [방법 F2]에서는 r번째 개체에 해당하는 부분행렬 $\mathbf{Z_r}$, $r = 1, \ldots, g$를 기준으로 개체별 용어의 평균 가중점수 벡터 $\bar{\mathbf{z}}_r$을 도출한다.

$$\bar{\mathbf{z}}_r = (\bar{z}_{r1}, \bar{z}_{r2}, \ldots, \bar{z}_{rp})^t = n_r^{-1}\mathbf{Z}_r^t\mathbf{1}_{n_r}. \tag{2.5}$$



(a) (b)

Figure 2.1. The process of finding one elbow point. TF-IDF = term frequency-inverse document frequency.

Table 2.1. The methods for creating a document-keyword weighted matrix

TF-IDF weighting methods	Term filtering methods
[W1] Term weighting	[F1] Term filtering
	[F2] Term filtering by objects
[W2] Term weighting by objects	[F1] Term filtering
	[F2] Term filtering by objects

TF-IDF = term frequency-inverse document frequency.

이 방법에서는 팔꿈치 지점을 각 개체별로 찾은 뒤 평균 가중점수가 높은 상위 용어를 개체별로 선별한다. 선별된 용어 집합에서 중복을 제거하면 q개의 핵심어를 선정할 수 있다. 이를 통해 핵심어로 구성된 문서-핵심어 가중행렬 $\mathbf{X} = (x_{rlt})$를 얻는다. 따라서 2.2절의 TF-IDF 가중치 부여 방법 [W1], [W2]와 2.3절의 용어 필터링 방법 [F1], [F2]를 Table 2.1과 같이 조합하면 총 네 개의 문서-핵심어 가중행렬 $\mathbf{X}$를 생성할 수 있다.

2.4. 문서 간 비유사성 행렬

문서-핵심어 가중행렬 $\mathbf{X}$에도 맹점은 존재한다. 개체 정보는 고려하였을지라도 각 문서가 가지고 있는 고유한 정보는 고려하지 않았기 때문이다. 이러한 문제를 극복하기 위하여 각 문서별 핵심어 정보가 반영된 문서 간 비유사성 행렬을 생성하고자 한다. g개의 개체 정보가 존재하며 크기가 $n\times q$인 문서-핵심어 가중행렬 $\mathbf{X}$는 크기가 $n_r\times q$, $r = 1,\ldots,g$인 부분행렬 $\mathbf{X_r}$로 이루어져 있다. r번째 개체인 부분행렬 $\mathbf{X_r}$에는 여러 문서가 존재하는데 그 중 l번째 문서를 $\mathbf{x}_{rl} = (x_{rl1}, x_{rl2}, \ldots, x_{rlq})^t$, $l = 1,\ldots,n_r$과 같이 정의한다. 마찬가지로 s번째 개체의 m번째 문서를 $\mathbf{x}_{sm} = (x_{sm1}, x_{sm2}, \ldots, x_{smq})^t$로 정의하자. 이때 다양한 거리측도를 사용하면 두 문서 간의 거리, 즉 비유사성을 계산할 수 있는데 본 연구에서는 다음 세 가지 거리측도를 사용하고자 한다. 가장 대표적인 유클리드거리(Euclidean distance, d_{ED})와 행 범주 간 거리 계산에 사용되는 카이제곱거리(chi-square distance, d_{CD}) 그리고 행 프로파일 벡터 간 거리를 계산하는 가중유클리드거리(weighted Euclidean distance, d_{WED})를 사용한다. 세 거리측도를

$\mathbf{x}_{rl}$과 $\mathbf{x}_{sm}$에 적용하면 식 (2.6)–(2.8)과 같다. 가중유클리드거리는 연구자가 어떤 방식으로 가중치를 부여하느냐에 따라 새롭게 정의될 수 있는데 본 연구에서는 식 (2.8)과 같이 정의하여 사용한다.

$$d_{\mathrm{ED}} = d(\mathbf{x}_{rl}, \mathbf{x}_{sm}) = \left[\sum_{t=1}^{q}(x_{rlt} - x_{smt})^2\right]^{\frac{1}{2}}, \tag{2.6}$$

$$d_{\mathrm{CD}} = d(\mathbf{x}_{rl}, \mathbf{x}_{sm}) = \left[\sum_{t=1}^{q}\frac{(x_{rlt} - x_{smt})^2}{(x_{rlt} + x_{smt})}\right]^{\frac{1}{2}}, \tag{2.7}$$

$$d_{\mathrm{WED}} = d(\mathbf{x}_{rl}, \mathbf{x}_{sm}) = \left[\sum_{t=1}^{q}\left(\frac{x_{rlt}}{x_{rl\cdot}} - \frac{x_{smt}}{x_{sm\cdot}}\right)^2 \Big/ \left(\frac{x_{rlt}}{x_{rl\cdot}} + \frac{x_{smt}}{x_{sm\cdot}}\right)\right]^{\frac{1}{2}}. \tag{2.8}$$

문서-핵심어 가중행렬 $\mathbf{X}$에 세 가지 거리측도를 적용하면 크기가 $n\times n$인 문서 간 비유사성 행렬 $\mathbf{D}$가 생성된다. 행렬 $\mathbf{D}$는 g^2개의 부분행렬 $\mathbf{D}_{rs}$로 이루어진 행렬이다. 부분행렬 $\mathbf{D}_{rs}$는 r번째 개체 $\mathbf{X}_r$과 s번째 개체 $\mathbf{X}_s$사이의 비유사성을 의미하므로 $\mathbf{D}_{rs} = (d(\mathbf{x}_{rl}, \mathbf{x}_{sm}))$, $r, s = 1, \ldots, g$; $l = 1, \ldots, n_r$; $m = 1, \ldots, n_s$와 같이 정의할 수 있다. 제곱근이 씌워진 거리 값을 원소로 하는 부분행렬 $\mathbf{D}_{rs}$를 Nam과 Choi (2017)이 제안한 방법을 참고하여 식 (2.9)와 같이 정의한 후 이를 비유사성 행렬 $\mathbf{D}$ 전체에 적용하여 $\tilde{\mathbf{D}}$를 생성한다.

$$\tilde{\mathbf{D}}_{rs} = \left(\tilde{d}(\mathbf{x}_{rl}, \mathbf{x}_{sm})\right) = \left(\frac{d^2(\mathbf{x}_{rl}, \mathbf{x}_{sm})}{2}\right). \tag{2.9}$$

2.5. 근접성 데이터

문서 간 비유사성 행렬 $\tilde{\mathbf{D}}$로부터 개체 간 평균적인 비유사성 정보를 담고 있는 근접성 데이터를 생성하고자 한다. 근접성 데이터는 각 문서의 고유한 정보뿐만 아니라 개체 간 관계 정보까지 담고 있어 개체 분석의 기초 데이터로 유용하게 사용할 수 있다.

Nam과 Choi (2017)의 이론에 근거하여 크기가 $g\times g$이고 개체 내 문서의 수를 대각원소로 갖는 대각행렬 $\mathbf{N} = \mathrm{diag}(n_1, \ldots, n_g)$와 크기가 $n\times g$인 행렬 $\mathbf{G}$를 생성한다. 이 행렬은 g개의 부분행렬 $\mathbf{G}_r$, $r = 1, \ldots, g$를 갖는데 크기가 $n_r\times g$인 부분행렬 $\mathbf{G}_r = (g_{lr})$, $l = 1, \ldots, n_r$은 다음과 같이 정의된다.

$$g_{lr} = \begin{cases} 1, & \text{if the } l^{th} \text{ document belongs to the } r^{th} \text{ object,} \\ 0, & \text{if the } l^{th} \text{ document does not belongs to the } r^{th} \text{ object,} \end{cases}$$

여기서 행렬 $\mathbf{F}$를 $\mathbf{F} = \mathbf{N^{-1}G^t\tilde{D}GN^{-1}} = (f_{rs})$와 같이 정의하면 $f_{rs} = 1/(n_r n_s)\,\mathbf{1}_r^t\tilde{\mathbf{D}}_{rs}\mathbf{1}_s$가 성립한다. 이는 r번째 개체와 s번째 개체로부터 파생된 부분행렬 $\tilde{\mathbf{D}}_{rs}$의 평균값이 f_{rs}가 됨을 뜻한다. 크기가 $g\times g$인 행렬 $\mathbf{F}$는 대칭행렬이지만 대각원소가 0은 아니다. 따라서 개체 간의 평균적인 비유사성을 나타내는 거리행렬로는 사용할 수 없다. 대각원소를 0으로 만들기 위해 Cox와 Cox (2001)을 참고하고자 한다. r번째 개체의 평균좌표와 s번째 개체의 평균좌표 간 거리를 ψ_{rs}라 하면 $\psi_{rs} = |2f_{rs} - f_{rr} - f_{ss}|$가 성립한다. 절대값을 취하는 이유는 ψ_{rs}를 거리 값으로 활용하기 위함이다. 따라서 개체 평균 사이의 비유사성으로 구성된 크기가 $g\times g$인 행렬은

$$\mathbf{\Psi} = \left(\frac{1}{2}\psi_{rs}\right), \quad r, s = 1, \ldots, g \tag{2.10}$$

와 같다. 이렇게 만들어진 행렬 $\mathbf{\Psi}$를 근접성 데이터로 정의한다. 2.3절에서 정의한 [W1], [W2], [F1], [F2]의 조합으로 만든 네 가지 문서-핵심어 가중행렬에 2.4절의 세 가지 거리측도를 적용한 후 2.5절의

Table 2.2. Proximity data generated by 12 methods

TF-IDF weighting	Term filtering	Distance measure	Methods	Proximity data
[W1]	[F1]	d_{ED}	Method 1	$\mathbf{\Psi}_1$
		d_{CD}	Method 2	$\mathbf{\Psi}_2$
		d_{WED}	Method 3	$\mathbf{\Psi}_3$
	[F2]	d_{ED}	Method 4	$\mathbf{\Psi}_4$
		d_{CD}	Method 5	$\mathbf{\Psi}_5$
		d_{WED}	Method 6	$\mathbf{\Psi}_6$
[W2]	[F1]	d_{ED}	Method 7	$\mathbf{\Psi}_7$
		d_{CD}	Method 8	$\mathbf{\Psi}_8$
		d_{WED}	Method 9	$\mathbf{\Psi}_9$
	[F2]	d_{ED}	Method 10	$\mathbf{\Psi}_{10}$
		d_{CD}	Method 11	$\mathbf{\Psi}_{11}$
		d_{WED}	Method 12	$\mathbf{\Psi}_{12}$

TF-IDF = term frequency-inverse document frequency. d_{ED} = Euclidean distance; d_{CD} = chi-square distance; d_{WED} = weighted Euclidean distance.

단계를 거치면 근접성 데이터 $\mathbf{\Psi}$가 생성된다. 본 연구에서는 근접성 데이터를 생성하는 12가지 방법을 Table 2.2와 같이 정의한다.

3. 활용 사례

개체 군집화에 최적화된 근접성 데이터 생성법을 제안하기 위하여, 2016년 한 해 동안 경제·인문사회연구회 소속 정부출연연구기관들이 기관별 홈페이지에 무료로 배포한 정기간행물을 연구 자료로 활용하고자 한다. 텍스트 추출이 불가능한 간행물 자료를 제공하는 건축도시공간연구소, 국토연구원, 에너지경제연구원, 한국농촌경제연구원, 한국법제연구원, 조세재정연구원, KDI 국제정책연구원은 분석 대상에서 제외하였다.

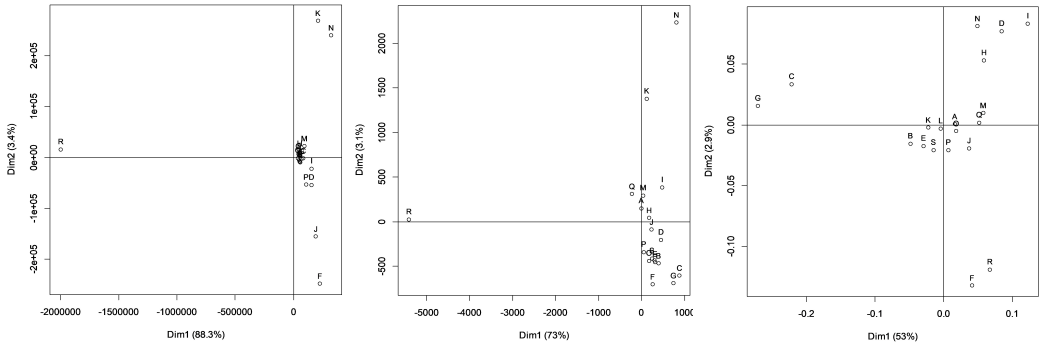
분석 대상이 되는 개체는 총 19개의 연구기관이며 알파벳 A부터 S로 표시한다. 문서-용어 빈도행렬의 행에 해당하는 문서는 각 연구기관에서 제공하는 간행물 343개이며 열에 해당하는 용어는 간행물에서 추출한 명사 용어 26,352개이다. 19개의 개체 정보가 존재하고 크기는 343×26,352인 문서-용어 빈도행렬에 Table 2.2의 12가지 방법을 적용하여 12개의 근접성 데이터를 생성한다.

3.1. 다차원척도법을 활용한 근접성 데이터의 시각적 군집화

다차원척도법이란 다차원 상에 존재하는 개체 간 유사성 또는 비유사성을 저차원 공간에 기하적으로 나타내어 개체 간 관계를 파악하는 다변량 기법이다 (Choi, 2014). 12개의 근접성 데이터는 개체 간의 평균적인 비유사성을 나타내는 거리행렬이므로 다차원척도법 적용이 가능하다.

다차원 상의 개체를 저차원 공간에 나타낸 그림을 다차원척도법도라 한다. 다차원척도법도의 차원 수는 Kruskal과 Wish (1978)의 판별 기준으로 결정할 수도 있으나 본 연구에서는 보편적으로 선호되는 이차원으로 차원을 축소한다. 공통적인 특성을 갖는 개체들은 다차원척도법도 상에서 모여 있으므로 하나의 군집으로 묶을 수 있다. 특정 군집과 멀리 떨어진 개체는 이질적인 특성을 갖는 개체로 판별할 수 있다.

다차원척도법도는 비유사성 행렬이 갖는 값에 따라 형태가 변하므로 비유사성을 측정하는 거리측도에 따라 결과는 달라진다. 거리측도만 다르게 적용한 근접성 데이터의 다차원척도법도를 살펴보자. Fig-



(a) Method 1: d_{ED} (b) Method 2: d_{CD} (c) Method 3: d_{WED}

Figure 3.1. Applications of multidimensional scaling using different distance measures. d_{ED} = Euclidean distance; d_{CD} = chi-square distance; d_{WED} = weighted Euclidean distance.

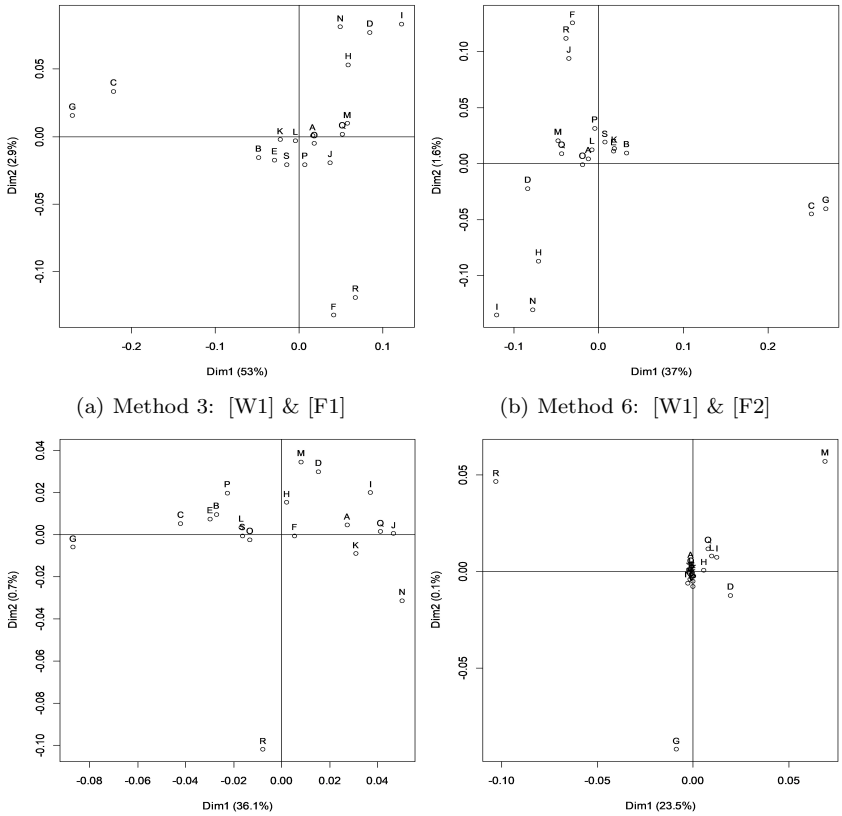
ure 3.1은 Table 2.2의 12가지 방법 중 거리측도만 다르게 적용한 Method 1부터 3, 즉 근접성 데이터 $\mathbf{\Psi}_1$(유클리드거리), $\mathbf{\Psi}_2$(카이제곱거리), $\mathbf{\Psi}_3$(가중유클리드거리)의 다차원척도법도이다. 유클리드거리를 사용할 경우 소수의 이질적인 개체들이 원점에서 과도하게 멀어져 동질적인 개체들은 원점에 몰리는 문제가 발생한다. 거리측도로 인해 이질적인 개체의 특성이 과대 추정되어 원점에 몰린 개체의 경향은 드러나지 않은 것이다. 카이제곱거리를 사용하면 문제가 다소 완화되는 듯 보이나 여전히 이질적인 개체가 부각된다. 가중유클리드거리를 사용할 경우 이질적인 개체를 과대 추정하지 않고도 구분할 수 있으며, 원점 주변에 있는 동질적인 개체들 간 군집 경향까지 파악할 수 있어 시각적 군집화에 가장 탁월하다. Method 4–6, 7–9, 10–12로 그린 다차원척도법도에서도 동일한 결과가 도출된다. 따라서 세 거리측도 중 개체 군집화에 적합한 거리측도는 가중유클리드거리임을 확인하였다.

가중유클리드거리를 활용하는 Method 3, 6, 9, 12에는 TF-IDF 가중치 부여 방법과 용어 필터링 방법이 각각 다르게 적용되는데, 네 가지 Method에 따른 다차원척도법도에는 어떠한 차이가 있는지 알아보고자 한다. Figure 3.2는 Method 3, 6, 9, 12로 생성된 근접성 데이터 $\mathbf{\Psi}_3$, $\mathbf{\Psi}_6$, $\mathbf{\Psi}_9$, $\mathbf{\Psi}_{12}$의 다차원척도법도이다. Method 3, 6, 9에 비해 Method 12는 이질적인 개체가 부각된다. Method 12를 활용하여 근접성 데이터를 생성할 경우 개체별로 용어 가중치가 부여될 뿐만 아니라(W2) 개체별로 용어의 평균 가중점수를 계산하여 필터링하므로(F2) 상대적으로 다른 방법에 비해 각 개체의 특징이 과대 추정될 수 있다.

다차원척도법을 활용한 시각적 군집 결과를 바탕으로 Method 3, 6, 9로 생성된 근접성 데이터 $\mathbf{\Psi}_3$, $\mathbf{\Psi}_6$, $\mathbf{\Psi}_9$가 상대적으로 개체 간 군집 경향을 바람직하게 형성한다는 결론은 내린다. 이어지는 3.2절에서는 세 가지 Method로 만들어진 근접성 데이터에 K-평균법을 적용하여 개체 군집화에 가장 적합한 방법 하나를 선택하고자 한다.

3.2. K-평균 군집분석을 활용한 군집화 및 자료 해석

군집분석은 유사성이 높은 개체를 하나의 군집으로 묶는 통계적 분석 방법이며 크게 계층적 군집분석과 비계층적 군집분석으로 나뉜다 (Choi, 2018). 본 연구에서 활용할 비계층적 군집분석은 연구자가 할당한 k개의 초기 군집에서 출발하여 각 군집의 중심(centroid)과 모든 개체 간의 유클리드거리를 계산한 후 가장 가까운 군집에 개체를 할당하는 방법이다. 중심에 대한 기준에 따라 K-평균법(K-means method), K-중위수법(K-median method), K-대표개체법(K-medoids method) 등과 같은 알고리즘이



(a) Method 3: [W1] & [F1] (b) Method 6: [W1] & [F2]

(c) Method 9: [W2] & [F1] (d) Method 12: [W2] & [F2]

Figure 3.2. Applications of multidimensional scaling using different weighting and filtering methods.

존재한다. 본 연구에서는 텍스트 마이닝 뿐만 아니라 보편적으로 많이 사용하는 K-평균 알고리즘을 활용하여 군집분석을 수행한다.

적정 군집 수인 k를 결정하기 위해 Rousseeuw (1987) 및 Cho 등 (2015)의 내용을 토대로 실루엣 통계량(silhouette statistics)을 계산하고자 한다. 실루엣 통계량은 각 개체를 기준으로 군집 내 밀집의 정도와 군집 간 분리의 정도를 계산하여 개체 군집이 얼마나 잘 형성되었는지를 나타내는 척도이다. 실루엣 통계량이 1에 가까울수록 각 개체가 좋은 군집을 형성한다고 본다. 나아가 k의 개수를 하나씩 늘려감에 따라 모든 개체들이 갖는 평균 실루엣 통계량 값을 계산하면, 통계량이 가장 클 때의 k를 최적 군집 수로 설정할 수 있다.

Table 3.1은 Method 3, 6, 9로 만들어진 근접성 데이터 $\mathbf{\Psi}_3$, $\mathbf{\Psi}_6$, $\mathbf{\Psi}_9$로부터 실루엣 통계량이 최대가 되는 최적 군집 수 k를 탐색한 결과이다. 모든 데이터에서 군집의 수가 2일 때 실루엣 통계량이 최대가 됨을 확인 수 있으나 비교를 위하여 군집 수 k를 2와 4로 설정한다. k가 2일 때 그리고 4일 때의 K-평균 군집분석을 근접성 데이터 $\mathbf{\Psi}_3$, $\mathbf{\Psi}_6$, $\mathbf{\Psi}_9$에 수행하여 총 6개의 결과를 도출한다.

군집분석 결과를 비교하기 위하여 군집 간 차이의 정도를 단일 값으로 표현한 통계량인 Γ를 활용한다. Γ를 도출하기 위해 Nam과 Choi (2017)의 이론을 참고한다. 먼저 개체별 문서의 수 n_r, $r = 1, \ldots, g$와

Table 3.1. The average silhouette statistics according to k

Silhouette statistics	Methods		
	Method 3($\boldsymbol{\Psi}_3$)	Method 6($\boldsymbol{\Psi}_6$)	Method 9($\boldsymbol{\Psi}_9$)
$k=2$	**0.465**	**0.345**	**0.283**
$k=3$	0.195	0.181	0.190
$k=4$	**0.204**	**0.150**	**0.178**
$k=5$	0.211	0.154	0.182
$k=6$	0.160	0.145	0.125
$k=7$	0.153	0.153	0.140
$k=8$	0.160	0.168	0.152
$k=9$	0.178	0.176	0.128
$k=10$	0.187	0.184	0.147
$k=11$	0.095	0.170	0.168
$k=12$	0.136	0.141	0.121
$k=13$	0.132	0.130	0.111
$k=14$	0.109	0.087	0.103
$k=15$	0.116	0.051	0.087
$k=16$	0.063	0.042	0.046
$k=17$	0.015	0.007	0.032
$k=18$	0.024	0.015	0.017

개체별 문서 벡터 $\mathbf{n}=(n_1,n_2,\ldots,n_g)^t$ 그리고 전체 문서의 수 n을 정의한 후 2.5절의 식 (2.9)를 활용하여 개체 구조를 분해한다. 분해 과정은 개체 간 비유사성(dissimilarity between objects), 개체 내 비유사성(dissimilarity within an object) 그리고 전체 비유사성(total dissimilarity)으로 이루어진다. Γ는 전체 비유사성에서 개체 내 비유사성이 차지하는 비율로 식 (3.1)과 같이 계산된다.

$$\Gamma = n\left(\mathbf{1}^t\tilde{\mathbf{D}}\mathbf{1}\right)^{-1}\sum_{r=1}^{g}\frac{1}{n_r}\mathbf{1_r}^t\tilde{\mathbf{D}}_{rr}\mathbf{1_r} \tag{3.1}$$

Γ값이 클수록 개체 내 비유사성은 커지고 개체 간 비유사성은 작아진다. 반대로 Γ값이 작을수록 개체 내 비유사성은 작아지고 개체 간 비유사성은 커진다. 따라서 Γ값이 작은 근접성 데이터를 개체 간의 차이가 뚜렷한 데이터로 판단할 수 있다.

Method 3, 6, 9의 군집 성능을 판단하는 기준으로 Γ와 더불어 오분류율(misclassification rate; MIR)을 함께 사용하고자 한다. 오분류율은 개체가 실제로 분류된 결과와 군집분석으로 분류된 결과가 얼마나 다른지를 측정하며 식 (3.2)와 같이 계산된다. 오분류율 값이 작을수록 실제 개체 분류결과와 유사한 데이터라 할 수 있다.

$$\text{MIR} = \frac{\text{Number of misclassified objects}}{\text{Total number of objects}}. \tag{3.2}$$

연구 자료에 대한 실제 개체 분류결과를 알기 위해 한국법제연구원 홈페이지에 게재된 공고문(2008)을 참고하여 Table 3.2를 도출한다. 이는 19개 연구기관들을 연구 분야에 따라 2개 혹은 4개 분야로 분류한 자료이다.

실제 개체 분류결과를 바탕으로 Γ 및 오분류율을 계산하면 Table 3.3과 같다. Γ와 오분류율 값이 모두 작을수록 개체 간 특성이 뚜렷하고 실제와 가깝게 군집되는 근접성 데이터 생성법이다. 계산 결과 군집수는 4일 때보다 2일 때가 더 좋고 Method 3에서 Method 9로 갈수록 결과 값이 나아짐을 확인하였다.

Table 3.2. Real classification of the nineteen government-funded research institutes

Research fields		Research institutes	Number of institutions
Economy	Economic policy	대외경제정책연구원(B), 산업연구원(C), 한국개발연구원(G)	3
	Resources / Infrastructure	정보통신정책연구원(E), 한국교통연구원(J), 한국해양수산개발원(P), 한국환경정책평가연구원(S)	4
	Public policy	과학기술정책연구원(A), 통일연구원(F), 한국행정연구원(Q), 한국형사정책연구원(R)	4
Society	Human resources	육아정책연구소(D), 한국교육개발원(H), 한국교육과정평가원(I), 한국노동연구원(K), 한국보건사회연구원(L), 한국여성정책연구원(M) 한국직업능력개발원(N), 한국청소년정책연구원(O)	8

Table 3.3. The calculation result of Γ and MIR

		Method 3($\mathbf{\Psi}_3$)	Method 6($\mathbf{\Psi}_6$)	Method 9($\mathbf{\Psi}_9$)
Γ	$k = 2$	0.028	0.007	**0.003**
	$k = 4$	0.049	0.049	0.019
MIR	$k = 2$	0.474	0.474	**0.158**
	$k = 4$	0.368	0.368	0.316

MIR = misclassification rate.

3.1절의 다차원척도법과 3.2절의 K-평균 군집분석 결과에 따라 텍스트 데이터로부터 개체 군집화에 최적화된 결과를 도출하는 근접성 데이터 생성법은 Method 9임을 확인하였다.

4. 결론

개체 정보가 존재하는 문서-용어 빈도행렬에서 개체를 분석하고자 할 경우, 일반적으로는 각 개체에 속하는 문서에서 공통으로 등장하는 용어인 공통어를 활용하여 개체-공통어 빈도행렬을 생성한다. 그러나 이 방법은 개별 문서가 갖는 고유 정보를 누락시킬 뿐만 아니라 특정 문서에서만 발생빈도가 높은 잠재적 핵심어를 제거하는 문제를 초래한다. 본 연구에서는 이러한 문제를 극복할 수 있는 근접성 데이터 생성법 12가지를 개발하였다. 12가지 방법 중 텍스트 데이터 군집에 특화된 방법을 찾기 위하여 2016년 한 해 동안 정부출연연구기관에서 발간한 정기간행물을 연구 자료로 활용하여 근접성 데이터를 생성하고 다차원척도법과 K-평균 군집분석을 적용하였다.

다차원척도법으로 분석한 결과 12가지 방법 중 가중유클리드거리를 사용하는 Method 3, 6, 9, 12를 선택하였고, 네 가지 방법 중에서는 일부 개체의 특성이 과하게 추정되지 않는 Method 3, 6, 9를 선택하였다. 바람직한 개체 군집화가 이루어졌는지를 수치로 비교하기 위해 K-평균 군집분석을 실시하였다. Γ와 오분류율을 활용하여 비교한 결과 Method 9로 생성한 근접성 데이터 $\mathbf{\Psi}_9$이 가장 우수한 결과를 도출하였다. 따라서 개체별로 용어 가중치를 부여하고 전체 용어를 기준으로 핵심어를 필터링한 후 가중유클리드거리를 적용하는 근접성 데이터 생성법 Method 9를 최종 제안한다.

텍스트 데이터 생성 시 본 연구에서 제안한 방법을 활용한다면 각 문서별 용어의 특이성을 반영할 뿐만 아니라 개체 군집화 역시 바람직하게 이루어지는 데이터를 생성할 수 있을 것이다. 그러나 하나의 연구 자료만을 활용하였기 때문에 결과를 일반화하여 단정 짓기엔 다소 무리가 있을 것이다. 따라서 다양한 연구 자료에 해당 방법을 적용하고 더 나은 결과를 도출하는 기법을 찾는 연구가 필요하다고 생각한다.

References

Cho, S. G., Cho, J. H., and Kim, S. B. (2015). Discovering meaningful trends in the inaugural addresses of United States presidents via text mining, *Journal of the Korean Institute of Industrial Engineers*, **41**, 453–460.

Choi, Y. S. (2014). *Walk in Multidimensional Scaling*, Free Academy, Gyeonggi-do.

Choi, Y. S. (2018). *Multivariate Data Analysis with R*, Kyungmoon, Seoul.

Cox, T. F. and Cox, M. A. A. (2001). *Multidimensional Scaling*, Chapman & Hall/CRC, London.

Kruskal, J. B. and Wish, M. (1978). Multidimensional Scaling, *Sage University Paper Series on Quantitative Applications in the Social Sciences*, 07-011, Sage Publications, Beverly Hills and London.

Nam, S. C. and Choi, Y. S. (2017). Non-parametric approach for the grouped dissimilarities using the multidimensional scaling and analysis of distance, *The Korean Journal of Applied Statistics*, **27**, 567–578.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a "kneedle" in a Haystack: Detecting Knee Points in System Behavior, *Distributed Computing Systems Workshops (ICDCSW) 2011 31st International Conference on, IEEE*, 166–171.

Sim, Y. S. and Kim, H. B. (2016). A study of destination image and measurement using text mining, *Journal of Tourism Sciences*, **40**, 221–245.

텍스트 데이터 분석을 위한 근접성 데이터의 생성과 군집화

정민지[a] · 신상민[a] · 최용석[a,1]

[a]부산대학교 통계학과

(2019년 3월 4일 접수, 2019년 4월 15일 수정, 2019년 4월 17일 채택)

요 약

문서-용어 빈도행렬은 텍스트 마이닝 분야에서 보편적으로 사용되는 데이터의 한 유형으로, 여러 개체들이 제공하는 문서를 기반으로 만들어진다. 그러나 대다수의 연구자들은 개체 정보에 무게를 두지 않고 여러 문서에서 공통적으로 등장하는 공통용어 중 핵심적인 용어를 효과적으로 찾아내는 방법에 집중하는 경향을 보인다. 공통용어에서 핵심어를 선별할 경우 특정 문서에서만 등장하는 중요한 용어들이 공통용어 선정단계에서부터 배제될 뿐만 아니라 개별 문서들이 갖는 고유한 정보가 누락되는 등의 문제가 야기된다. 본 연구에서는 이러한 문제를 극복할 수 있는 데이터를 근접성 데이터라 정의한다. 그리고 근접성 데이터를 생성할 수 있는 12가지 방법 중 개체 군집화의 관점에서 가장 최적화된 방법을 제안한다. 개체 특성 파악을 위한 군집화 알고리즘으로는 다차원척도법과 K-평균 군집분석을 활용한다.

주요용어: 텍스트 마이닝, 근접성 데이터, TF-IDF, 다차원척도법, 군집분석

[1]교신저자: (46241) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과. E-mail yschoi@pusan.ac.kr