Journal of Korean Institute of Intelligent Systems, Vol. 22, No. 2, April 2012, pp. 205-211
http://dx.doi.org/10.5391/JKIIS.2012.22.2.205

인자점수와 자기조직화지도를 이용한 희소한 문서데이터의 군집화

Sparse Document Data Clustering Using Factor Score and Self Organizing Maps

전성해†

Sunghae Jun†

청주대학교 통계학과

요 약

통계학과 기계학습의 다양한 기법을 이용하여 문서집합을 군집화하기 위해서는 우선 군집화분석에 적합한 데이터구조로 대상 문서집합을 변환해야 한다. 문서군집화를 위한 대표적인 구조가 문서-단어행렬이다. 각 문서에서 발생한 특정단어의 빈도값을 갖는 문서-단어행렬은 상당부분의 빈도값이 0인 희소성문제를 갖는다. 이 문제는 문서군집화의 성능에 직접적인 영향을 주어 군집화결과의 성능감소를 초래한다. 본 논문에서는 문서-단어행렬의 희소성문제를 해결하기 위하여 인자분석을 통한 인자점수를 이용하였다. 즉, 문서-단어행렬을 문서-인자점수행렬로 바꾸어 문서군집화의 입력데이터로 사용하였다. 대표적인 문서군집화 알고리즘인 자기조직화지도에 적용하여 문서-단어행렬과 문서-인자점수행렬에 대한 문서군집화의 결과들을 비교하였다.

키워드 : 희소한 문서군집화, 문서-단어행렬, 문서-인자점수행렬, 인자분석, 자기조직화지도

Abstract

The retrieved documents have to be transformed into proper data structure for the clustering algorithms of statistics and machine learning. A popular data structure for document clustering is document-term matrix. This matrix has the occurred frequency value of a term in each document. There is a sparsity problem in this matrix because most frequencies of the matrix are 0 values. This problem affects the clustering performance. The sparseness of document-term matrix decreases the performance of clustering result. So, this research uses the factor score by factor analysis to solve the sparsity problem in document clustering. The document-term matrix is transformed to document-factor score matrix using factor scores in this paper. Also, the document-factor score matrix is used as input data for document clustering. To compare the clustering performances between document-term matrix and document-factor score matrix, this research applies two typed matrices to self organizing map (SOM) clustering.

Key Words : Sparse Document Clustering, Document-term Matrix, Document-factor score Matrix, Factor Analysis, Self Organizing Map.

1. 서론

문서데이터의 군집화(document data clustering)는 문서내의 존재하는 텍스트 사이의 유사성에 기반을 두어 문서를 군집화 하는 비지도학습(unsupervised learning) 기법이다. 정보검색(information retrieval), 자연어처리(natural language processing), 기계학습(machine learning) 분야에서 문서군집화는 중요한 전처리(preprocessing)기법으로 사용된다 [1]. 초기에 문서군집화는 워드넷(WordNet)과의 연계를 통하여 이루어졌다 [2-3]. 최근에는 다양한 웹 검색엔진이나 특허검색사이트를 통하여 수많은 웹페이지 또는 특허문서(patent document)를 얻고 이를 분석하게 되었다. 하지만 문서데이터는 텍스트로 이루어진 데이터구조를 갖기 때문에 수치데이터의 분석에 기반하고 있는 통계학 또는 기계학습 분석기법들을 적용하기에는 어려움이 있다 [4]. 따라서 문서데이터를 분석하기 위하여 텍스트 위주의 문서를 수치화된 데이터구조로 바꾸는 과정이 필요하다. 일반적으로 사용되는 방법은 각 문서에 나타난 단어(term) 또는 키워드(keyword)의 빈도를 구하는 것이다 [1]. 이를 통하여 문서-단어행렬(document-term matrix)을 얻게 된다. 이 행렬의 각 행(row)은 1개의 문서를 나타내고, 각 열(column)은 해당문서에 나타난 단어를 나타낸다. 행과 열이 만나는 곳(cell)의 값은 해당문서에 나타난 특정단어의 빈도를 나타낸다. 대부분의 경우 문서-단어행렬에서 문서의 수보다 단어의 수가 훨씬 크기 때문에 매우 많은 부분에서 빈도값이 0이 된다. 이와 같은 문서-단어행렬의 특성 때문에 이 행렬은 희소한(sparse) 데이터구조를 갖는다. 그러므로 문서군집화의 성능향상을 기대하기 위해서는 문서-단어행렬의 희소성문제를 해결하기 위한 노력이 필요하다.

접수일자: 2011년 11월 25일
심사(수정)일자: 2012년 2월 3일
게재확정일자 : 2012년 2월 4일
†교신저자
이 논문은 2010년도에 청주대학교 산업과학연구소가지원한 특별연구과제에 의해서 연구되었음.

본 논문에서는 문서-단어행렬이 갖는 희소성(sparsity) 문제를 해결하기 위하여 인자점수(factor score)에 의한 자기조직화지도(self organizing map; SOM)를 제안한다. 지금까지 SOM은 다양한 문서군집화 작업에 사용되어 왔다 [5-8]. 물론 문서군집화 이외에도 SOM은 다양한 군집화에 사용되고 있다 [9-10]. 그러나 기존의 연구에서는 문서-단어행렬의 희소성 문제를 해결하지 않고 이 행렬을 SOM의 입력데이터로 사용하여 문서군집화를 수행하였다. SOM 알고리즘 자체에 대한 성능향상을 위한 많은 연구도 진행되어 왔다 [11-12]. 제안방법은 다변량(multi-variate) 통계분석 기법 중 하나인 인자분석에 의한 인자점수를 이용하여 문서-단어행렬의 희소성문제를 해결한다. 즉 문서-단어행렬의 희소성 문제를 해결하기 위하여 단어들 간의 공분산(covariance) 구조를 이용한 인자점수(factor score)를 구하여 문서-단어행렬을 문서-인자점수행렬로 변형하고 이를 대표적인 문서군집화 알고리즘은 SOM의 입력데이터로 사용한다. 검색된 문서집합으로부터 문서-단어행렬을 구하고, 이것을 인자점수에 의한 문서-인자점수행렬로 변환하여 최종적으로 SOM에 의한 문서군집화를 수행한다. 제안된 문서군집화의 성능평가를 위하여 Reuters의 뉴스기사 문서집합과 USPTO(United State Patent and Trademark Office)의 특허문서집합을 이용한 실험을 수행한다 [13-14].

본 논문의 2절에서는 문서데이터의 군집화에서 나타나는 희소성문제, 이를 해결하기 위한 인자분석과 인자점수, 그리고 최종적인 문서군집화를 위한 SOM 알고리즘에 대하여 알아본다. 구체적으로 인자점수와 SOM을 이용한 희소한 문서데이터의 효율적인 군집화에 대한 제안방법은 3절에서 설명한다. 제안된 기법의 향상된 성능을 확인하기 위하여 객관적인 문서집합을 이용한 실험 및 결과는 4장에서 보이고, 마지막으로 5장에서는 본 연구의 결론과 향후 연구과제에 대하여 알아본다.

2. 관련연구

2.1 문서군집화와 희소성문제

문서군집화는 유사한 개념을 갖는 문서들을 그룹화하는 텍스트 처리과정이다 [15]. 사전에 문서들의 분류정보(preclassified information)를 알고 학습이 이루어지는 문서분류(document classification)와는 다르게 문서군집화는 이와 같은 사전정보가 없는 비지도학습이다. 문서군집화는 일반적인 연속형 데이터의 군집화에 비해 몇 가지 고려할 사항들이 있다. 즉, 문서 그 자체로는 군집화 분석이 어렵기 때문에 대개 문서로부터 단어 또는 키워드를 추출하여 다음과 같은 문서-단어 행렬을 만들고 이를 이용하여 문서군집화를 수행한다.

	단어1	단어2	...	단어m
문서1	빈도11	빈도12	...	빈도1m
문서2	빈도21	빈도22	...	빈도2m
⋮	⋮	⋮	...	⋮
문서n	빈도n1	빈도n2	...	빈도nm

그림 1. 문서-단어행렬

Fig. 1. Document-term matrix

위 그림의 (n*m) 문서-단어행렬에서 각 행은 특정문서를 나타내고 각 열은 해당문서에 포함된 특정단어를 나타낸다. 행과 열이 교차되는 곳은 해당문서에 나타난 특정단어의 빈도를 나타낸다. 예를 들어 빈도12는 문서1에 나타난 단어2의 빈도를 표시한다. 일반적으로 문서-단어행렬에서 문서의 크기 n에 비해서 단어의 크기 m이 훨씬 큰 데이터구조를 갖는다. 또한 1개의 문서에 m개의 모든 단어가 나타나지는 않기 때문에 많은 곳에서 빈도수는 0이 된다. 그러므로 문서-단어행렬은 희소한 데이터구조를 갖게 된다. 문서-단어행렬의 희소성문제는 문서군집화에서 군집화 결과의 성능에 영향을 미친다. 따라서 군집화의 성능향상을 위하여 문서-단어행렬의 희소성문제는 해결되어야 할 필요가 있다. 본 논문에서는 이와 같은 문제를 해결하기 위하여 다음 절에서 설명되는 인자분석에 의한 인자점수를 이용한다.

2.2 인자점수

본 논문에서 이용되는 인자분석와 함께 SVD(singular value decomposition)는 대표적인 차원축소 방법의 하나이며 주로 패턴인식, 전자신호, 바이오정보학 등의 데이터분석 및 압축에 사용되어 왔다 [16-21]. 예를 들어, SVD는 패턴인식에 사용되는 특징점들(features)을 나타내는 입력변수 벡터의 차원이 매우 클 경우에 이 벡터의 차원을 축소하는데 효과적으로 이용되었고, 바이오정보학에서는 관측치보다도 훨씬 많은 변인변수들의 차원축소를 위하여 사용되었다. 하지만 문서군집화 특히 희소한 특성을 지닌 문서군집화에 적용하기에는 어려움이 있습니다. 희소한 데이터는 SVD에서 요구되는 계산을 위한 행렬조건을 만족하지 못하기 때문에 본 논문에서는 SVD와는 또 다른 차원축소방법인 인자분석의 인자점수를 이용하여 희소한 문서데이터의 군집화에서 사용되는 변수들의 차원을 축소하였다. 인자분석은 많은 변수들 사이의 공분산 관계를 공통인자(common factor)인 훨씬 적은 수의 확률변수(random variable)로 축소하여 해석하고자 하는 다변량 통계분석기법이다. 즉, 전체변수와 잠재적인 공통인자 사이의 관계를 나타내는 통계적 모형을 구축하고 변수들 사이의 관계를 설명할 수 있는 관측되지 않은 잠재적인 공통인자을 찾아내어 해석하고자 하는 분석기법이다 [22-23]. m차원 확률벡터 X의 모평균벡터가 μ이고 모공분산행렬이 Σ를 가질 때 p인자모형에서 각각의 변수 $X_1, X_2, ..., X_m$ 들은 모든 변수에 공통적으로 영향을 미치는 관찰할 수 없는 공통인자 $F_1, F_2, ..., Fp$ 들과 각 변수에만 영향을 미치는 특수인자의 선형결합으로 표시하며 i번째 변수 X_i는 다음과 같이 표시한다 [24].

$$X_i - \mu_i = \lambda_{i1} F_1 + \lambda_{i2} F_2 + \cdots + \lambda_{ip} F_p + \varepsilon_i \quad (1)$$

행렬의 형태로 나타내면 다음과 같다.

$$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{\Lambda F} + \boldsymbol{\varepsilon} \quad (2)$$

여기서 선형결합에 사용한 λ_{ij}를 인자적재(factor loading)라고 하며 i번째 변수에 대한 j번째 공통인자 F_j의 중요성을 나타내는 가중계수이고 (m*p) 행렬 $\Lambda=\{\lambda_{ij}\}$는 인자적재행렬이다. (p*1) 확률벡터 F는 모든 변수에 공통으로 영향을 미치는 공통인자벡터이고, ε_i는 i번째 변수에만 영향을 미친다. 인자적재행렬을 추정하는 방법에는 주성분(principal component), 최대우도(maximum likelihood) 등 여러 가지 방법들이 있다. 인자분석은 의미있고 해석하기 쉬운 독립적

인 공통인자를 유도하는 것이 목적이므로 추정을 통하여 얻어진 인자적재행렬을 해석하는데 어려움이 있는 경우가 발생한다. 이와 같은 경우에는 인자의 수를 고정하고 인자의 축을 회전시키면서 단순한 구조로 변경하는 인자의 회전이 고려한다. 일반적으로 베리멕스회전(varimax rotation)이 사용된다. 인자분석의 목적은 인자모형의 모수인 인자적재행렬을 추정하는 것이지만 추가적으로 공통인자의 추정치인 인자점수를 구할 수 있다. 인자점수는 인자모형의 타당성을 검토하는데 사용될 뿐만 아니라 다른 통계분석에서 새로운 변수값으로 사용될 수 있다. 인자점수를 추정하는 방법에는 Bartlett에 의해 제안된 가중회귀적방법과 Thomson에 의하여 제안된 회귀적방법이 있다 [25]. 본 논문에서는 좀 더 범용으로 사용되는 회귀적방법을 사용한다. 인자모형에서 인자적재행렬 Λ와 특수분산행렬 Ψ를 알고 있고, 공통인자벡터 F와 특수인자벡터 ε가 결합정규분포(joint normal distribution)를 따른다고 가정하면 $(X-\mu)$와 F의 결합분포는 $N_{m+p}(0, \Sigma^*)$인 다변량정규분포를 따른다. 여기서 Σ^*은 다음과 같다.

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} & \boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}' & \boldsymbol{I} \end{pmatrix} \tag{3}$$

또한, $F/X=x$의 조건부분포(conditional distribution)는 다음과 같은 평균벡터와 공분산행렬을 갖는 다변량정규분포를 따른다.

$$N(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x},\ (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}+\boldsymbol{I})^{-1}) \tag{4}$$

인자점수를 계산하는 방법은 최우추정법에 의하여 표본공분산행렬 S를 이용하여 다변량정규분포의 평균을 추정한 다음의 식을 사용한다.

$$\hat{\boldsymbol{f}}_j = \hat{\Lambda}'\boldsymbol{S}^{-1}(\boldsymbol{x}_j - \overline{\boldsymbol{x}}), \quad j = 1,2,\cdots,n \tag{5}$$

위 식에서 $\hat{f}$은 추정된 인자점수행렬을 나타내고 $\hat{\Lambda}$은 추정된 인자패턴행렬을 나타낸다. 인자패턴행렬은 개개의 관측값과 각 인자와의 연관성의 정도를 나타내는 측도로 사용된다. 표본공분산행렬이 아닌 상관계수행렬 R에 대해서 인자점수를 구하면 다음과 같다.

$$\hat{\boldsymbol{f}}_j = \hat{\Lambda_z}'\boldsymbol{R}^{-1}\boldsymbol{z}_j, \quad j = 1,2,\cdots,n \tag{6}$$

여기서 z_j는 다음 식과 같이 나타낸다.

$$\boldsymbol{z_j} = \boldsymbol{D}^{-1/2}(\boldsymbol{x_j} - \overline{\boldsymbol{x}}) = \begin{pmatrix} \dfrac{x_{1j} - \overline{x_p}}{\sqrt{s_{11}}} \\ \vdots \\ \dfrac{x_{pj} - \overline{x_p}}{\sqrt{s_{pp}}} \end{pmatrix} \tag{7}$$

위 식에서 D는 공분산행렬을 나타내고 이 행렬의 제곱근과 평균벡터를 이용하여 관측값의 표준화를 수행하였습니다. 예를 들어, s_{11}은 D행렬의 첫 번째 변수의 분산이다. 이와 같은 과정을 거쳐 인자점수가 구해지고 이 값은 본 연구에서 문서군집화의 희소성문제를 해결하기 위하여 문서-단어행렬을 문서-인자점수행렬로 바꾸는데 사용된다. 본 논문에서는 이와 같은 제안방법을 대표적인 문서군집화 알고리즘인 SOM에 적용한다.

2.3 자기조직화지도

대표적인 비지도학습 신경망모형인 SOM은 입력층(input layer)과 출력층(output layer)의 2개 층으로 구성된다 [26-27]. 퍼셉트론(perceptron) 모형과는 달리 SOM에서는 출력층을 형상지도(feature map)라고 부른다. 군집화의 결과는 형상지도 위에 나타난다. 기존의 군집화 기법들과는 달리 SOM은 경쟁학습(competitive learning)을 사용한다 [28]. 하나의 입력개체에 대하여 형상지도와 연결되는 가중치 벡터는 서로 경쟁하여 입력개체와 가장 가까운 가중치가 승자(winner)가 되어 승자만이 가중치갱신을 하게 된다 [29]. 이와 같은 승자독식(winner takes all)의 원리를 사용하는 SOM은 다음과 같이 가중치갱신이 이루어진다.

$$w^{New} = w^{Old} + \alpha(x - w^{Old}) \tag{8}$$

w^{New}는 승자가 되어 갱신이 이루어진 값이고, w^{Old}는 가중치 갱신이 이루어지기 이전의 값이다. a는 학습률(learning rate)이고, 이 값에 따라 입력개체 x에 의존하는 가중치조정이 이루어진다. 형상지도의 승자노드에 대한 가중치 갱신은 승자노드 뿐만 아니라 승자노드 주위의 노드까지도 확대될 수 있고 이 범위는 사전에 이웃(neighborhood) 반경의 크기로 결정한다. 이웃의 크기가 0이 되면 승자노드 자신만 가중치갱신을 하게 된다. SOM 군집화에서 제공하는 최대 군집수(maximum number of clusters)는 사전에 정한 형상지도의 차원이 된다. 즉, 형상지도의 차원이 (3*3)이면 최대 9개까지의 군집이 형성될 수 있다. SOM은 텍스트 데이터의 군집화에도 만족할 만한 결과를 제공하지만 텍스트의 크기인 변수의 크기가 증가할수록 군집화의 성능이 떨어지는 단점을 보인다 [1]. 특히 문서-단어행렬의 군집화에서는 데이터 자체의 희소성(sparseness) 문제가 나타난다. 이 문제를 해결하기 위하여 본 논문에서는 인자분석에 의한 인자점수를 이용하여 문서-단어행렬의 희소성을 제거한 문서-인자점수행렬을 만든다.

3. 인자점수와 자기조직화지도를 이용한 문서군집화

본 논문에서는 일반적으로 문서군집화 알고리즘의 입력데이터로 사용되는 문서-단어행렬의 희소성문제를 해결하기 위하여 인자점수에 의한 문서-인자점수행렬을 구축한다. 대부분의 문서군집화 알고리즘과 마찬가지로 SOM을 이용한 문서군집화에서도 문서-단어행렬은 입력데이터로 사용된다. 그러므로 제안방법은 문서-단어행렬의 희소성을 제거한 문서-인자점수행렬을 구하고 이 행렬을 SOM의 입력데이터로 사용하여 문서군집화를 수행한다. 다음그림은 인자점수가 전체단어들의 선형결합(linear combination)의 형태로 표현됨을 나타낸다.

일반적으로 단어의 수 m보다 인자의 수 p가 매우 작다. 즉 i번째 인자는 다음과 같이 전체 단어들의 선형결합으로 표현된다.

$$\begin{aligned} &\text{인자}1 = a_{11}\text{단어}1 + a_{12}\text{단어}2 + \cdots + a_{1m}\text{단어}m \\ &\text{인자}2 = a_{21}\text{단어}1 + a_{22}\text{단어}2 + \cdots + a_{2m}\text{단어}m \\ &\quad\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad \ddots \qquad \vdots \\ &\text{인자}p = a_{p1}\text{단어}1 + a_{p2}\text{단어}2 + \cdots + a_{pm}\text{단어}m \end{aligned}$$

그림 2. 인자점수

Fig. 2. Factor scores

$$\text{인자}_i = a_{i1}\text{단어}1 + a_{i2}\text{단어}2 + \cdots + a_{im}\text{단어}m \quad (9)$$

위 식에 a_{ij}는 i번째 인자와 j번째 단어의 연관성의 크기를 나타낸다. 이와 같은 p개의 인자점수를 이용하여 다음과 같은 문서-인자점수행렬을 구한다.

	인자1	인자2	...	인자p
문서1	점수11	점수12	...	점수1p
문서2	점수21	점수22	...	점수2p
⋮	⋮	⋮	...	⋮
문서n	점수n1	점수n2	...	점수np

그림 3. 문서-인자점수행렬

Fig. 3. Document-factor scores matrix

위 행렬에서 0의 값을 갖는 곳은 없다. 예를 들어 점수12는 문서1에 대한 인자2의 점수이다. 본 연구에서는 전체데이터의 90% 이상의 설명력을 포함하는 인자의 개수 p를 결정한다. 원래 데이터가 가지고 있는 n개의 변수들이 인자분석을 통하여 n보다 훨씬 작은 p개의 인자들로 축소되면서 원래 변수들이 가지고 있는 전체 데이터에 대한 설명력이 감소한다. 이때 설명력은 전체데이터의 변동(분산)을 나타낸다. 즉 n개의 변수들의 총분산이 차원축소된 p개의 인자에 의해서 작아지게 된다. 이론적으로 n개의 인자를 선택하면 원래 데이터와 같은 설명력을 갖게 되지만 이때에는 차원축소의 효과가 전혀 없기 때문에 의미가 없다. 그러므로 차원을 축소하면서도 설명력이 급격히 떨어지지 않는 범위에서 결정되는데 보통 90%의 설명력을 갖게 되면 전체데이터에 대한 설명력을 유지하면서 차원축소의 효과도 볼 수 있게 된다 [18-19]. 단어-인자점수행렬의 각 값은 모두 연속형 수치를 포함하며, 이를 통해 문서-단어행렬의 희소성 문제가 해결된다. 다음그림은 검색된 문서집합에서부터 최종 문서군집화의 결과를 얻기까지 본 연구에서 제안하는 전체 과정을 단계별로 나타내고 있다.

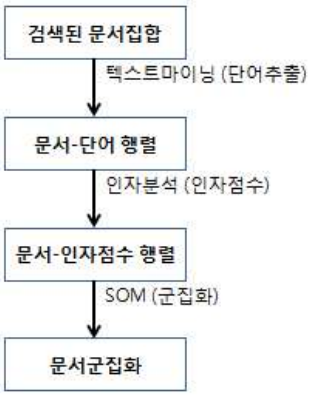

그림 4. 제안방법의 절차

Fig. 4. Process of proposed method

제안방법은 검색된 문서집합을 문서-단어행렬로 바꾸는 텍스트마이닝단계, 문서-단어행렬을 문서-인자점수행렬로 바꾸는 인자분석단계, 그리고 문서-인자점수행렬을 이용하여 군집화를 수행하는 SOM단계의 3단계로 이루어진다. 본 연구의 제안절차에 대한 구체적인 알고리즘은 다음과 같이 설명된다.

Input: Retrieved Documents, $D=\{d_1, d_2, ..., d_n\}$
n: the number of documents
Output: Documents grouping, $G=\{g_1, g_2, ..., g_k\}$
k: the number of clusters
Algorithm:
1. Constructing document-term matrix ($n*m$)
m: the number of terms
1.1 Text mining (text corpus + parsing)
1.2 Data matrix
2. Constructing document-factor score matrix ($n*p$)
p: the number of factors
2.1 Determination of the number of factors
2.2 Computation of the factor scores
3. Clustering documents using SOM
3.1 Initialization of weights, $W=\{w_1, w_2, ..., w_i\}$
3.2 Repeat until stop conditions {
for i=1 to n
selecting winner nearest to d_i
updating weight
$w^{new}=w^{old}+a(d_i-w^{old})$
3.3 Assignment of documents to similar to g_i
$(i=1,2,...k)$

제안절차의 입력은 검색된 문서들이고 최종적인 결과는 각 군집에 할당된 문서 리스트집합이다. 세 번째 단계인 SOM에 의한 경쟁학습의 정지조건(stop conditions)은 사전에 정한 반복수 또는 가중치갱신의 최소변동값 중에서 선택할 수 있다. 예를 들어 사전에 SOM 반복학습 후에 군집을 할당하는 반복수를 100으로 설정하거나, 입력층과 형상지도를 연결하는 연결가중치의 갱신폭이 0.001 보다 작을 경우 SOM 학습을 멈추고 군집을 할당하는 등의 지정이 가능하다. 문서군집화에서 군집수 k는 SOM 형상지도의 차원에 의해 결정된다. 예를 들어 $(3*3)$의 형상지도를 갖는 SOM 군집화에서는 최대 9개까지의 군집수를 가질 수 있다.

4. 실험 및 결과

제안방법의 성능평가를 위한 실험을 위하여 본 논문에서는 먼저 2개의 뉴스기사 (news articles) 문서집합을 사용하였다. Reuters-21578로부터 50개의 뉴스기사로 구성된 'topic acq' 문서집합과 20개의 뉴스기사를 포함하는 'topic crude' 문서집합을 이용하여 [13] 인자점수와 SOM을 이용한 문서군집화를 수행하였다. 실험을 위해 사용된 분석도구는 대표적인 통계계산 언어 중의 하나인 R을 사용하였다. [30]. 추가적으로 R에서 제공한 'som' 패키지 및 다변량 통계함수들도 이용하였다 [31]. 우선 다음결과와 같이 텍스트마이닝을 이용하여 문서-단어행렬을 구하였다.

표 1. acq와 crude의 문서-단어행렬의 차원
Table 1. Document-term matrix of acq and crude

문서집합	문서수	단어수
acq	50	2007
crude	20	1132

위 결과를 보면 2개의 문서집합 모두 문서의 수에 비해 단어의 수가 매우 크게 나타났다. 예를 들어 acq 문서집합에서 문서의 수는 50이지만 단어의 수는 2007이었다. 어떤 문서에 1번만이라도 나타난 단어는 모두 단어 수에 포함되기 때문에 극단적으로 1개의 문서에서만 1번 나타나고 나머지 문서에서는 모두 0이 될 수도 있다. 이와 같은 문서-단어행렬의 구조적 특성 때문에 희소성문제가 발생한다. 본 논문에서는 희소성문제를 해결하기 위한 하나의 방법으로 인자점수를 이용한 문서-인자점수행렬을 이용하였다. 물론 기존의 연구에서 이와 같은 단어 개념이 아닌 키워드개념을 적용하기도 하는 데 이때에는 의미상 중요한 단어가 키워드에 포함되지 않으면 그 만큼 정보손실을 감수해야 한다. 때문에 본 논문에서는 모든 단어를 고려한 문서-단어행렬로부터 문서군집화를 시작하였다. 인자점수를 이용하기 위해서는 먼저 인자의 수를 결정해야 한다. 인자분석은 주어진 데이터에 대한 차원축소에 해당되기 때문에 정보(설명력)의 손실이 발생한다. 선택되는 인자의 수가 많을수록 정보의 손실이 작게 되지만 너무 많은 인자의 수를 선택하면 차원축소의 효과가 떨어진다. 본 논문에서는 주어진 데이터에 대한 90% 이상의 설명력을 갖는 인자의 수로 결정하였다. 다음은 각 문서집합에 대한 선택된 인자의 수와 설명력을 나타내고 있다.

표 2. acq와 crude의 인자수와 설명력
Table 2. Number of factors and explanation of acq and crude

문서집합	인자수	설명력
acq	23	90.23%
crude	11	91.76%

acq와 crude 문서집합을 위하여 결정된 인자수는 각각 23과 11이었다. 결정된 인자수를 바탕으로 구한 인자점수를 이용하면 다음과 같이 문서-단어행렬은 문서-인자점수행렬과 변환된다.

표 3. acq와 crude의 각 행렬 데이터에 대한 차원
Table 3. Dimensions of two matrices of acq and crude

문서집합	문서-단어 행렬	문서-인자점수 행렬
acq	50*2007	50*23
crude	20*1132	20*11

문서-단어행렬의 대부분의 값은 0으로 매우 희소한 데이터구조를 갖지만 문서-인자점수행렬은 연속형 인자점수 값을 갖기 때문에 문서-단어행렬의 희소성문제를 해결하였다. 다음으로 문서-인자점수행렬을 이용하여 SOM에 의한 문서군집화를 수행하였다. 이 결과를 희소성문제를 가지고 있는 문서-단어행렬에 의한 SOM 군집화결과와 비교하였다. 문서-단어행렬과 문서-인자점수행렬의 비교를 위한 측도로는 SOM 결과의 양자화(quantization)에 대한 정확도를 계산해 주는 평균왜곡측도(average distortion measure; ADM)를 사용하였다 [31]. 이 측도는 다음과 같이 정의된다.

$$ADM_i = \sum_{i=1}^{k} \frac{\| x - m_i \|}{h_{c_i}} \quad (10)$$

이 값은 데이터 x에 대한 가중거리측도(weighted distance measure)를 계산한다. m_i와 h_{ci}는 각각 i번째 군집의 중심과 코드(code)의 이웃커널(neighborhood kernel)을 나타낸다. AMD값이 작을수록 더 유사한 개체들끼리 묶였음을 의미한다. 동일한 SOM 군집화 알고리즘에 대하여 문서-단어행렬과 문서-인자점수행렬을 각각 입력데이터로 사용한 군집화 결과는 다음과 같다.

표 4. 문서-단어행렬과 문서-인자점수행렬의 ADM 값
Table 4. AMD values of document-term matrix and document-factor score matrix

문서집합	문서-단어 행렬	문서-인자점수 행렬
acq	1035.83	966.88
crude	1192.64	1141.21

acq와 crude 2개의 문서집합에서 모두 문서-단어행렬에 비해 문서-인자점수행렬의 AMD 값이 더 작게 나타났음을 알 수 있다. 즉 인자점수에 의해 문서-단어행렬의 희소성문제를 해결한 후에 SOM 문서군집화를 한 경우가 더 좋은 결과를 제공함을 알 수 있었다.

SOM의 형상지도의 차원의 크기에 따른 군집화성능을 평가하기 위하여 지금까지 등록된 바이오기술(Bio technology)관련 특허문서집합을 이용하였다. USPTO로부터 미국에 출원된 총 50,886개의 바이오기술특허문서를 검색하였다 [14]. 이 문서집합으로부터 텍스트마이닝을 사용하여 (*50886*119458*)의 차원을 갖은 문서-단어행렬을 구하였다. 본 논문의 2번째 실험을 위하여 50886개의 특허를 포함한 문서집합에서 단순임의추출(simple random sampling)을 이용하여 각 100개의 문서들을 포함하는 5개의 샘플들에 대하여 SOM의 형상지도의 차원크기에 따른 문서군집화 결과를 비교하였다. 우선 형상지도의 차원이 (*2*2*)인 경우의 AMD 값은 다음과 같다.

표 5. 바이오특허문서 군집화 (*2*2* 형상지도)
Table 5. Bio patent document clustering (*2*2* feature map)

문서집합	문서-단어 행렬	문서-인자점수 행렬
Sample1	447.99	321.64
Sample2	442.19	403.33
Sample3	324.97	286.39
Sample4	364.92	329.12
Sample5	326.38	290.76

5개의 샘플에서 모두 문서-단어행렬에 비해 문서-인자점수의 AMD 값이 더 작게 나타났음을 알 수 있었다. 즉, 희소성문제를 가지고 있는 문서-단어행렬을 이용한 SOM 군집화결과보다 희소성문제를 해결한 문서-인자점수행렬에 의한 SOM 군집화결과의 성능이 더 우수함을 반복적으로 확인할 수 있었다. 다음은 (*3*3) 형상지도를 갖는 SOM 문서군집화 결과이다.

표 6. 바이오특허문서 군집화 (*3*3 형상지도)
Table 6. Bio patent document clustering (*3*3 feature map)

문서집합	문서-단어 행렬	문서-인자점수 행렬
Sample1	392.73	339.82
Sample2	435.80	373.95
Sample3	439.74	385.71
Sample4	472.59	409.13
Sample5	500.05	439.58

(*2*2) 형상지도를 갖는 SOM과 마찬가지로 (*3*3) 형상지도를 갖는 SOM의 문서군집화 결과도 문서-인자점수행렬이 문서-단어행렬보다 더 작은 AMD 값이 계산됨을 알 수 있었다. 마지막으로 형상지도의 차원이 (*4*4) 인 경우의 SOM 문서군집화 결과는 다음과 같다.

표 7. 바이오특허문서 군집화 (*4*4 형상지도)
Table 7. Bio patent document clustering (*4*4 feature map)

문서집합	문서-단어 행렬	문서-인자점수 행렬
Sample1	475.76	426.71
Sample2	462.54	391.74
Sample3	678.25	612.19
Sample4	386.90	327.50
Sample5	530.86	451.72

(*2*2)와 (*3*3)의 형상지도 차원크기를 갖는 SOM의 문서군집화와 같은 결과가 (*4*4) 형상지도의 SOM 군집화 결과에서도 나타나고 있음을 알 수 있었다. 그러므로 문서-단어행렬의 희소성문제를 해결한 문서-인자점수행렬의 SOM 군집화결과가 더 우수한 군집화 결과를 제공하고 있음을 반복되는 표본을 통하여 확인할 수 있었다.

5. 결론 및 향후 연구과제

본 논문에서는 문서군집화에서 발생하는 희소성문제를 해결하기 위하여 인자분석에 의한 인자점수를 이용하였다. 일반적으로 문서군집화에서는 검색된 문서집합의 분석을 위하여 텍스트마이닝의 전처리과정을 통하여 문서-단어행렬을 구축한다. 이때 문서의 크기에 비해 단어의 크기가 훨씬 크기 때문에 문서-단어행렬의 각 값에 0을 포함하는 경우가 매우 많아 희소성문제를 발생하고, 이것은 문서군집화에서도 군집결과의 성능을 저하시키는 원인이 되고 있었다. 따라서 제안연구에서는 문서-단어행렬에 대하여 인자분석을 수행하고 인자점수를 구하여 문서-인자점수행렬을 구축하였다. 새롭게 구축된 문서-인자점수행렬은 모든 값이 연속형 점수값을 나타내기 때문에 문서-단어행렬의 희소성문제를 해결 할 수 있었다. 제안방법의 성능을 평가하기 위하여 대표적인 문서군집화 알고리즘인 SOM을 이용하여 문서-단어행렬과 문서-인자점수행렬의 군집화결과를 비교하였다. Reuters-21578의 뉴스기사로 이루어진 문서집합과 USPTO의 바이오관련 기술특허를 포함한 문서집합을 이용한 실험을 통하여 제안방법의 향상된 군집화성능을 확인하였다. 또한 SOM 형상지도의 차원크기에 따른 문서군집화에서도 문서-인자점수행렬이 문서-단어행렬에 비해 더 향상된 군집화결과가 나타남을 확인할 수 있었다.

향후연구에서는 인자점수 뿐만 아니라 다양한 다변량통계분석기법을 이용하여 문서-단어행렬의 희소성문제를 좀 더 다양하게 해결하고 군집화뿐만 아니라 문서분류 등 광범위한 문서집합의 분석에 적용할 수 있는 새로운 방법에 대한 연구가 이루어 질 것이다.

참 고 문 헌

[1] N. O. Andrews, E. A. Fox, "Recent Developments in Document Clustering," *Technical Report TR-07-35*, Computer Science, Virginia Tech, 2007.

[2] S. Scott, S. Matwin, "Text Classification Using WordNet Hypernyms," *Proceeding of Workshop on Usage of WordNet in Natural Language Processing Systems*, pp. 38 - 44, 1998.

[3] M. Buenaga, J. M. Gómez-Hidalgo, B. Díaz-Agudo, "Using WordNet to Complement Training Information in Text Categorization," *Recent Advances in Natural Language Processing*, pp. 150 - 157, 1997.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer, 2001.

[5] T. W. S. Chow, M. K. M. Rahman, "Multilayer SOM With Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection," *IEEE Transactions on Neural Networks*, vol. 20, no. 9, pp. 1385-1402.

[6] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela, "Self Organization of a Massive Document Collection," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 574 - 585, 2000.

[7] C. Hung, S. Wermter, "A Dynamic Adaptive Self-Organizing Hybrid Model for Text Clustering," *Proceeding of IEEE International Conference on Data Mining* (ICDM 03), pp. 75 - 82, 2003.

[8] H. Chen, C. Schuffels, R. Orwig, "Internet Categorization and Search: A Self-Organizing Approach," *Journal of Visual Communication and Image Representation*, vol. 7, no. 1, pp. 88 - 102, 1996.

[9] 홍정표, 황승국, “SOM을 이용한 퍼지 TAM 네트워크 모델,” *한국지능시스템학회논문지*, 제16권, 제5호, pp. 642-646, 2006.
[10] 윤경배, 최준혁, "앙상블 Support Vector Machine과 하이브리드 SOM을 이용한 동적 웹 정보 추천 시스템," *한국지능시스템학회논문지*, 제13권, 제4호, pp. 433-438, 2003.
[11] S. Jun, “Improvement of SOM using Stratification,” *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 9, no. 1, pp. 36-41, 2009.
[12] S. Jun, “Improvement of Self Organizing Maps using Gap Statistic and Probability Distribution,” *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 8, no. 2, pp. 116-120, 2008.
[13] I. Feinerer, *A Text Mining Framework in R and Its Applications*, PhD Dissertation, Department of Statistics and Mathematics Vienna University of Economics and Business Administration, 2008.
[14] United State Patent and Trademark Office, http://www.uspto.gov
[15] C. Hung, S. Wermter, P. Smith, "Hybrid Neural Document Clustering Using Guided Self-Organization and WordNet," *IEEE Intelligent Systems*, vol. 19, iss. 2, pp. 68-77, 2003.
[16] Y. Yam, P. Baranyi, C. T. Yang, "Reduction of fuzzy rule base via singular value decomposition," IEEE Transactions on Fuzzy Systems, vol. 7, Iss. 2, pp. 120-132, 1999.
[17] J. J. Wei, C. J. Chang, N. K. Chou, G. J. Jan, "ECG data compression using truncated singular value decomposition," IEEE Transactions on Information Technology in Biomedicine, vol. 5, Iss. 4, pp. 290-299, 2001.
[18] S. Lee, M. H. Hayes, "Properties of the singular value decomposition for efficient data clustering," Signal Processing Letters, vol. 11, Iss. 11, pp. 862-866, 2004.
[19] P. Howland, H. Park, "Generalizing discriminant analysis using the generalized singular value decomposition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, Iss. 8, pp. 995-1006, 2004.
[20] P. Bao, X. Ma, "Image adaptive watermarking using wavelet domain singular value decomposition," IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, Iss. 1, pp. 96-102, 2005.
[21] S. K. Jha, R. D. S. Yadava, "Denoising by Singular Value Decomposition and Its Application to Electronic Nose Data Processing," Sensors Journal, vol. 11, Iss. 1, pp. 35-44, 2011.
[22] Hair, J. F., Black, B., Babin, B., Anderson, R. E., *Multivariate Data Analysis*, Prentice Hall, 1992.
[23] 김기영, 전명식, *다변량 통계자료분석*, 자유아카데미, 1994.
[24] Johnson, R. A. and Wichern, D. W., *Applied Multivariate Statistical Analysis*, 5th ed. Prentice Hall, 2002.
[25] Rencher, A. C., *Methods Of Multivariate Analysis* 2nd ed. John Wiley & Sons, 2002.
[26] T. Kohonen, *Self-Organizing Maps*, Springer, 2001.
[27] Han, J., Kamber, M., *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.
[28] T. Kohonen, “Self-Organized Formation of Topologically Correct Feature Maps,” *Biological Cybernetics*, vol. 43, pp. 59 - 69, 1982.
[29] 오일석, *패턴인식*, 교보문고, 2008
[30] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org, 2010.
[31] J. Yan, *Self-Organizing Map - Package 'som'*, CRAN www.r-project.org, 2011.

저 자 소 개

전성해(Sunghae Jun)
1993년 : 인하대 통계학과 학사
1996년 : 인하대 통계학과 이학석사
2001년 : 인하대 통계학과 이학박사
2007년 : 서강대학교 컴퓨터공학과 공학박사
2003년~현재 : 청주대학교 통계학과 부교수

관심분야 : 기술경영, 인공지능, 데이터마이닝
Phone : 043-229-8205
Fax : 043-229-8432
E-mail : shjun@cju.ac.kr