Journal of the Korean Institute of Industrial Engineers
Vol. 41, No. 5, pp. 453-460, October 2015.
ISSN 1225-0988 | EISSN 2234-6457
http://dx.doi.org/10.7232/JKIIE.2015.41.5.453

<Application Research>

텍스트마이닝을 활용한 미국 대통령 취임 연설문의 트렌드 연구

조수곤[1] · 조재희[2] · 김성범[1†]

[1]고려대학교 산업경영공학과 / [2]광운대학교 경영대학

Discovering Meaningful Trends in the Inaugural Addresses of United States Presidents Via Text Mining

Su Gon Cho[1] · Jaehee Cho[2] · Seoung Bum Kim[1]

[1]Department of Industrial Management Engineering, Korea University
[2]Business School, Kwangwoon University

Identification of meaningful patterns and trends in large volumes of text data is an important task in various research areas. In the present study, we propose a procedure to find meaningful tendencies based on a combination of text mining, cluster analysis, and low-dimensional embedding. To demonstrate applicability and effectiveness of the proposed procedure, we analyzed the inaugural addresses of the presidents of the United States from 1789 to 2009. The main results of this study show that trends in the national policy agenda can be discovered based on clustering and visualization algorithms.

Keywords: Text mining, Data mining, Clustering, Locally linear embedding

1. 서 론

최근 데이터 분석은 다변화된 현대 사회를 더욱 정밀하게 관찰 및 예측하기 위하여 생산, 의료, 스포츠, 기상, 마케팅 등 거의 모든 분야에 걸쳐 활용된다. 이와 같은 현상은 인류에게 가치 있는 정보를 제공 할 수 있는 가능성을 제시하면서 그 중요성이 점차 부각되고 있다. 이에 따라, 수많은 연구자들로 하여금 기존의 숫자형태의 데이터로 대표되는 정형데이터(structured data)의 분석과 함께, 텍스트, 이미지, 동영상, 음성 등과 같은 구조화되지 않은 비정형데이터(unstructured data)의 연구에 힘을 기울이는 촉매제가 되고 있다. 특히, 텍스트 데이터는 인터넷 사용자의 폭발적인 증가와 함께 연구의 중요성이 더욱 부각되고 있다(Hu and Lu, 2012). 방대한 분량의 문서를 구조적으로 저장하고 열람할 수 있는 데이터베이스 등 컴퓨터 시스템의 발달 또한, 보다 많은 연구자들에게 역동적으로 연구를 수행할 수 있는 원동력이 되어오고 있다(Rebholz-Schuhmann *et al*., 2005).

이와 같이 텍스트 데이터로부터 의미 있는 정보의 추출 및 분석을 위한 텍스트마이닝(text mining, Chakraborty *et al*., 2013)은 문서요약(Kim *et al*., 2013), 정보검색(Pai *et al*., 2013), 감성분석(Liu, 2012), 트렌드 분석(Hung and Jang, 2012), 군집분석(Aggarwal and Zhai, 2012) 및 분류분석(Chen and Chen, 2011) 등 다양한 주제의 연구로 심화되는 현상을 보이고 있다. 이중, 텍스트의 트렌드 연구는 기존의 전문가 의견, 설문조사 등과 같은 전통적인 방법에서 벗어나, 해당 분야에서 생산되는 텍스트 자체를 분석함으로써, 데이터 기반의 객관적인 결과를 탐구한다는 데에 의의가 있다. 예를 들어, 학술지의 논문 주제어간 연관관계 연구(Cho and Kim, 2012), 국내 산업공학 연구기법의 트렌드 분석(Cho *et al*., 2014), 특허문서를 활용한 기술추이 연구(Kam *et al*., 2013; Kim *et al*., 2009, Park *et al*., 2014), 인터넷 검색추세를 활용한 주식투자전략 연구(Kim and Koo,

† 연락저자 : 김성범 교수, 02841 서울시 성북구 안암동 5가 1번지 고려대학교 산업경영공학과, Tel : 02-3290-3397, Fax : 02-929-58889, E-mail : sbkim1@korea.ac.kr
2015년 2월 25일 접수; 2015년 3월 30일 수정본; 2015년 5월 11일 게재 확정.

2013), 신문기사를 활용한 소비현상의 분석(Kim *et al*., 2012), 사회연결망서비스(social network service) 데이터를 활용한 패션 트렌드 연구(Lee *et al*., 2014) 등 다양한 영역의 텍스트를 분석함으로써 실용성이 높은 연구결과로 이어지고 있다.

특히 본 연구에서 분석대상으로 설정한 대통령의 연설문은 국가적 당면 과제와 사회적 지향점이 기록된 자료로서, 그 역사적 가치가 크다. 따라서 대통령 연설문을 활용한 트렌드 연구는 시대변화에 따른 국정운영의 추이를 살펴볼 수 있다는 점에서 그 효용가치가 높다. 현재까지 수행된 대부분의 연구는 크게 연설문에서 사용되는 특정 단어의 추이를 관찰함으로써 트렌드를 도출하거나(Lim, 2002), 연설문 내용에 포함된 국가적 주제 즉, 정치, 경제, 사회, 복지, 통일, 외교 등의 측정지표를 연구자가 설정하고, 정량화하여 그 추이를 분석하였다(Kim, 2014). 그 외에, 연설문에 포함되는 단어들의 관계를 활용하여 네트워크를 형성하고 분석하려는 시도도 있었다(Kim, 2013). 이와 같은 대통령 연설문의 트렌드를 분석하기 위한 다양한 시도는, 연설문을 통하여 한 국가집단의 지향점을 살펴보는데 큰 역할을 했다.

하지만 이들 연구는 몇몇 한계점을 내포하고 있는데 먼저, 연설문 트렌드분석의 핵심이 되는 단어의 선정을 전문가 또는 연구자가 직접 선택함으로써, 객관성이 결여되어 있다. 선정된 단어는 각각의 연설문에서 주장하려는 주제를 대표하는 변수 역할을 하는데, 연구자가 자의적으로 설정한 변수를 활용한 분석 결과는 데이터 자체에 내재되어있는 의미를 온전하게 반영할 수 없기 때문이다. 또한 선정된 단어의 어휘적 유사성에 대한 고려가 부족하여, 분석대상 단어가 중복되는 문제 또한 해결해야할 문제로 여겨진다. 예를 들어, 연설문내의 유사한 역할을 하는 두 단어, '정책'과 '시책'은 개별단어로서의 관찰이 아니라, 하나의 군집으로 통합하여 분석하는 것이 좀 더 합리적이라고 하겠다. 아울러, 대부분 기존 연구에서 시행한 문서 내 단어의 출현 빈도만을 활용한 단편적인 트렌드 분석보다는 단어와 단어 사이의 관계를 고려한 통합적인 트렌드 분석이 필요하다.

따라서 본 연구는 이와 같은 기존연구의 한계점을 극복하기 위하여, 전처리 단계 및 분석단계로 구성되는 미국 대통령 취임연설문의 분석방법을 제안한다. 먼저, 전처리 단계는 기존 연구자의 자의적 단어선정을 지양하기 위하여 객관적 절차를 수행했다. 텍스트에 포함된 모든 단어집합으로부터 불용어(stop words), 기호 등 무의미한 단어 들를 객관적인 기준에 의하여 제거하고, 어휘적으로 동일한 의미의 단어를 통합하여 단어 중복의 문제를 피하기 위한 절차로써, 어간(stemming analysis) 분석을 활용하여 데이터를 재구성했다. 이후, 문서의 특성을 대표하는 주제어 어간을 선정하고 및 모든 문서에서의 출현 정도를 기록한 문서-단어 교차표를 생성하여, 문서의 트렌드 정보를 내포한 정형 데이터를 준비했다. 분석 단계에서는 *k*-평균 군집화 알고리즘의 활용을 통한 문서 및 주제어 어간의 군집 및 패턴의 관찰과 함께, 지역선형사상(locally linear embedding)을 이용한 주제어 어간의 시각화를 활용하여 연설문에 내포된 통합 트렌드를 도출하고 관찰함으로써, 기존의 단편적 트렌드 관찰의 한계를 극복했다.

본 논문의 구성은 다음과 같다. 제 2장에서는 본 연구에서 활용된 미국 대통령 취임사의 소개와 함께 전처리과정을 서술하고, 제 3장에서는 텍스트 트렌드 분석을 위한 방법을 설명한다. 제 4장에서는 제안 기법을 적용하여 미국 대통령 취임사의 시계열 트렌드를 확인한 결과를 보인다. 제 5장에서는 본 연구의 결론 및 기대효과와 함께 본 연구의 한계점에 대하여 논의하고 향후 연구 방향을 모색하였다.

2. 데이터 및 전처리

본 논문의 연구대상으로 선정된 미국 대통령 취임사는, 1789년 초대 대통령 'George Washington' 이후 2009년 'Barack Obama' 까지 약 200여 년간 축적된 총 56개의 문서집합이다. 각각의 대통령 취임사는 발표된 시점의 국가 미래 비전과 국정운영의 청사진을 기록한 자료이며, 시대의 변화에 따른 언어 활용의 차이점을 유추할 수 있는 문서로서의 가치도 가지고 있다 (Kim, 2013).

분석대상인 56개의 취임사를 살펴보면, 총 4,905개의 문장과 145,735개의 단어가 담겨져 있으며, 중복을 제거한 단어의 개수는 43,469개이다. 각 시기별 문장의 평균 길이는 <Figure 1>에서 살펴볼 수 있듯이, 과거에서 현대로 이동함에 따라 그 양이 감소함을 알 수 있다. 이는 근대의 문서들에서 문장의 평균길이가 점차 감소하는 일반적인 현상을 반영한다(Julia and Giuliana, 2013). 이와 같은 문장의 평균길이 감소 현상은 독자들이 보다 용이하고 신속한 읽기를 할 수 있도록, 문장이 점차 문어체에서 구어체로 변화하는 현상을 대변한다(Akimoto, 2010). 따라서 본 연구의 대상인 대통령의 연설문 또한 대중에게 그 내용이 전달되는 방법, 즉 미디어가 활자로 대표되는 신문 또는 도서 등에서 라디오 또는 텔레비전 등으로 변화함에 따라 보다 평이하고, 간결한 형태를 보이고 있음을 유추할 수 있다. 그러나 본 연구는 이와 같은 단편적인 추이의 관찰을 넘어, 보다 정밀한 연설문의 트렌드 분석을 수행하기 위하여 문서의 특성을 대표하는 객관적 주제어를 추출하고, 연설문 내의 주제어의 출현 횟수를 기록한 문서-단어 교차표를 생성하기 위한 전처리 작업을 수행 하였다.

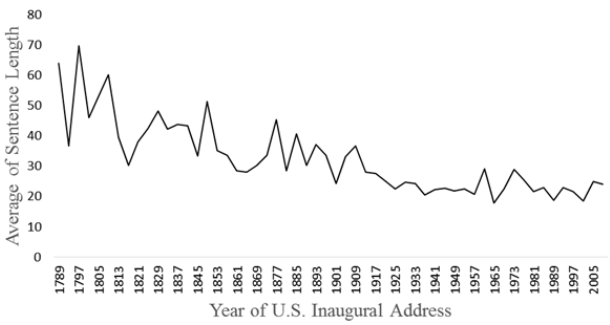

Figure 1. Average of sentence length in inaugural addresses of the presidents of the United States over the time (1789~2009)

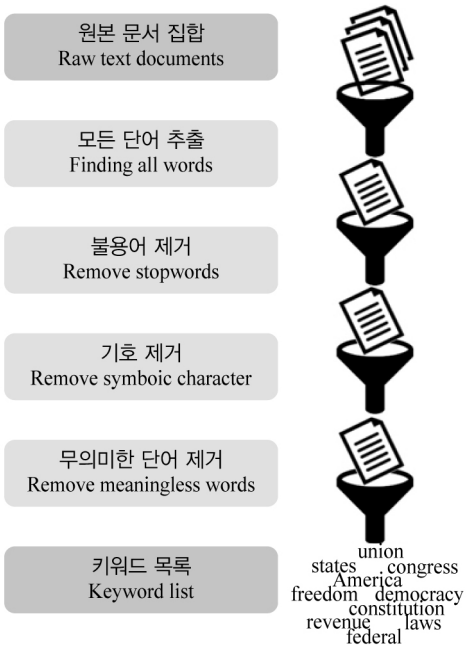

Figure 2. An overall process to find terms from text data

<Figure 2>는 텍스트 데이터로부터 주제어를 추출하는 과정이다. 자세히 살펴보면, 먼저 모든 문서 내에 포함된 단어를 추출하고, 불용어를 제거한다. 불용어는 의미 없는 단어들의 집합이며 관사, 전치사, 조사, 접속사가 그 예이다. 본 연구에서는 Natural language toolkit(Bird, 2006) corpus의 불용어 사전을 활용하여 불용어를 제거하였다. 이후, 문서 내에 포함된 기호 및 무의미한 단어는 연구자가 확인하고 삭제했는데, 이때 제거된 기호는 '-', ')', '?' 등이며, 무의미한 단어는 '0', '13', '14th', '15th' 등이다. 이와 같은 전처리 과정을 거쳐 추출된 단어는 8,837개이며, 이중 동일한 의미를 가지는 단어의 중복을 피하기 위하여 어간추출을 수행하였다(Lovins, 1968). 어간추출은 어형이 변화된 단어로부터 단어가 변하지 않는 부분 즉, 어간을 분리하여 정리하는 과정이다. 이는 인터넷 검색엔진이 질의어를 통한 검색을 수행할 때, 동일한 어간을 가지는 단어들을 동의어로 취급하여 질의어를 확장하고, 검색결과의 품질을 향상시키는 방법과 같은 맥락이라고 하겠다. 주요 어간추출 방법은 Snowball stemmer(Porter, 2001), Lacaster stemmer (Chris, 1990), Porter stemmer(Porter, 1980)가 존재하는데 본 연구에서는 이중 가장 널리 사용되는 Snowball stemmer 방법을 활용하여(Jivani, 2011), 총 5,134개의 어간을 추출하였다. 어간이 추출된 결과의 예를 살펴보면, 복수형의 경우 'people'과 'peoples'의 어간은 동일하게 'peopl'이 선정된다. 또한 동사형의 경우 'govern', 'governed', 'governing' 및 'governs'은 어간 'govern'으로 대표된다.

이렇게 어간으로 추출된 총 5,134개 단어 중, 핵심어를 추출하기 위하여 과정으로 TF-IDF(Rajaraman and Ullman, 2011)분석을 수행하였다. TF-IDF는 여러 문서로 이루어진 문서의 집합이 존재할 때, 각각의 문서에 포함된 단어의 중요도를 산출하는 통계적인 수치로써, 문서 내 단어의 출현양을 나타내는 TF와 총 문서에서의 단어의 출현비율의 역수를 취한 IDF을 활용하며, 식 (1)과 같이 계산된다.

$$TF-IDF_i = TF_i \times IDF_i = TF_i \times \log_2 \frac{N}{n_i}, \qquad (1)$$

여기서 TF_i는 단어 i가 모든 문서에서 관찰되는 양이며, IDF_i는 총 문서의 수 N 대비 단어 i가 출현한 문서의 수 n_i 출현 비율의 역수를 취한 값이다. 따라서 단어의 중요도를 나타내는 TF-IDF 값은 특정 단어의 출현양이 많을수록 증가하지만, 모든 문서에서 빈번하게 사용되는 단어는 그 정도에 따라 값이 낮게 계산된다.

총 56개의 연설문에서 전 처리 된 5,134개 단어의 전체문서 집합으로부터 중요도를 산출하기 위하여, 각 단어의 평균 TF-IDF 점수를 산출했다(Pramokchon and Piamsa-nga, 2014; Zhang *et al*., 2010). 모든 단어의 중요도를 내림차순으로 정렬한 결과는 <Figure 3>과 같다. 이때 평균 TF-IDF의 급격한 변화를 보이는 지점, 즉 팔꿈치 지점(elbow point)를 기준으로 상위 TF-IDF 단어를 선택함으로써 최종 주제어 어간을 선정하게 되는데, 본 연구에서는 평균 TF-IDF 점수가 0.34 이상인 125개의 단어를 선정하였다. 팔꿈치 지점이 명확하지 않을 경우에는 해당 분야 전문지식이 있는 사람들이 주관적으로 판단하게 된다.

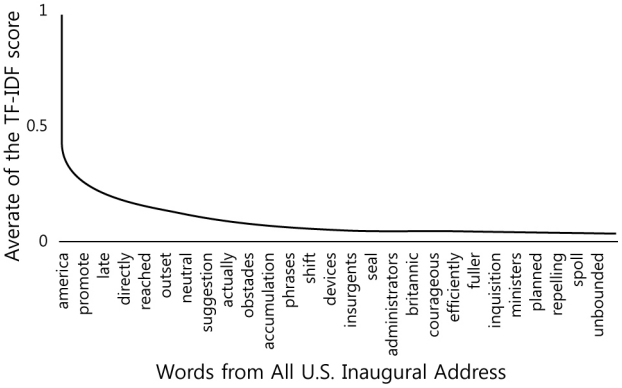

Figure 3. The average of TF-IDF score plot using all keywords in inaugural address in the president of the United States

이후, 문서-단어 교차표는 총 56개의 문서와 최종 추출된 125개의 주제어 어간이 각 문서에서 출현한 정도를 기록한 56 × 125차원의 표로 생성된다(<Table 1> 참조). 예를 들어, 1789년 'George Washington' 대통령 취임 연설문의 경우, 주제어 어간 'administr'가 2회, 2009년 'Barack Obama' 대통령의 경우 0회 관찰됨을 알 수 있다. 이때 주제어 어간 'administr'은 여러 단어를 대표하는 결과로서, 다음 5개의 단어를 의미한다. 'administration', 'administrations', 'administrative', 'administrated', 'administrators.'

이와 같이 여러 단계에 걸쳐 선정된 각각의 주제어 어간이 문서 내에 출현한 정도를 기록한 문서-단어 교차표는, 비정형

텍스트 데이터를 정형데이터로 재구성한 결과가 된다. 이는 기존의 정형데이터의 분석기법을 보다 용이하게 활용할 수 있는 장점과 함께, 주제어 어간으로 대표되는 문서의 핵심 정보를 포함하고 있다는 점에서 의의가 있다.

Table 1. An document-term matrix in inaugural addresses of the presidents of the United States over the time(1789~2009)

	administr	alway	⋯	whole
1789, Washington	2	0	⋯	0
1793, Washington	1	0	⋯	0
⋮	⋮	⋮	⋮	⋮
2009, Obama	0	1	⋯	0

3. 분석 방법

본 장에서는 문서 및 주제어 어간의 군집을 확인하고 패턴을 관찰하기 위하여 활용된 *k*-평균 군집화 알고리즘을 살펴본다. 또한, 직관적이고 용이한 트렌드의 관찰을 가능하게 하는, 지역선형사상의 시각화 방법과 그 예를 확인한다.

3.1 *k*-평균 군집화(k-Means clustering)

k-평균 군집화는 분석의 대상이 되는 데이터의 각 관측치간 최소평균거리를 활용하여, *k*개의 군집으로 분할하는 데이터 마이닝의 기법 중의 하나이다(Jain and Dubes, 1988). *k*-평균 군집화 기법은 다음과 같은 과정을 통하여 활용된다. 먼저 분석자가 선정한 군집 수 *k*개의 초기점을 임의로 생성하고, 모든 관측치와 *k*개의 초기점과의 거리를 각각 계산하여, 모든 관측치를 가장 가까운 초기점과 동일한 *k*개의 군집을 형성한다. *k*개의 군집이 생성되면 각 군집내의 평균점을 새롭게 계산하고 또다시 관측치 사이의 거리를 계산하여 새로운 군집을 형성한다. 위 단계를 반복하여 더 이상 *k*개의 평균점이 바뀌지 않아 군집의 변화가 없는 경우 최종 군집으로 설정한다(Hartigan, 1975). *k*-평균 군집화 기법을 사용하기 위해서는 먼저 연구자가 설정하는 군집수(*k*)와 거리척도를 결정한다. 다양한 군집수의 결정 방법들이 존재하나(Gordon, 1999), 방법 간의 우열을 논하기는 어려우며, 통상 문제의 배경지식을 기반으로 연구자가 결정한다. 또한 거리계산방식은 유클리드, 맨하튼, 상관관계 등 다양한 방식이 있으나 데이터의 특성과 분석 목적에 맞게 결정을 한다.

3.2 지역선형사상(Locally Linear Embedding, LLE)

지역선형사상은 고차원 데이터를 시각화하고 해석하기 위하여, 차원을 축소하는 다변량 통계 분석기법 중 하나이다(Rowie and Saul, 2000). 주성분 분석과 같은 기존의 차원축소 방법과 목적은 동일하지만, 지역선형사상은 차원축소 시 인접한 점들의 관계를 고려한다는 점에서 그 차이점이 있다.

지역선형사상의 절차는 다음과 같다. *D*차원의 실수 벡터 X_i가 *n*개가 있다고 가정한다면, 먼저 각각의 데이터 점 X_i와 가장 가까운 이웃 점들 $X_j(j = 1, \cdots, k)$ *k*개 구한다. 그리고 X_j와 가중벡터 W_{ij}의 선형방정식($\sum_j W_{ij}X_j$)과 X_i의 거리를 최소로 만드는 가중치 W_{ij}를 구한다. 즉, 식 (2)로 정의된 오차 *E*의 값을 최소로 하는 W_{ij}를 구하며, 이때 최소제곱기법을 사용한다.

$$\min_W E(W) = \sum_i \left| X_i - \sum_j W_{ij} X_i \right|^2 \tag{2}$$

이후, 고차원의 점들과 동일한 가중치를 갖도록 구성된 저차원의 벡터 Y_i를 구한다. 이 과정에서는 식 (3)의 Φ를 최소화하는 Y_i를 구하며, Y_i를 구하기 위하여 고유치문제(eigenvalue problem)를 풀어준다(Saul *et al.,* 2000).

$$\min_W \Phi(Y) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|^2 \tag{3}$$

<Figure 4>은 지역선형사상의 효과 및 과정을 보여주는 예이다. (a)는 3차원의 다양체이며, (b)와 (c)는 선과 점으로 추출된 된 데이터들이다. (d)는 점으로 추출된 (c)에 대하여 지역선형사상을 수행한 결과이다. 이는 3차원상의 데이터들의 인접관계가 유지되어, 차원의 축소가 이루어졌음을 확인할 수 있다.

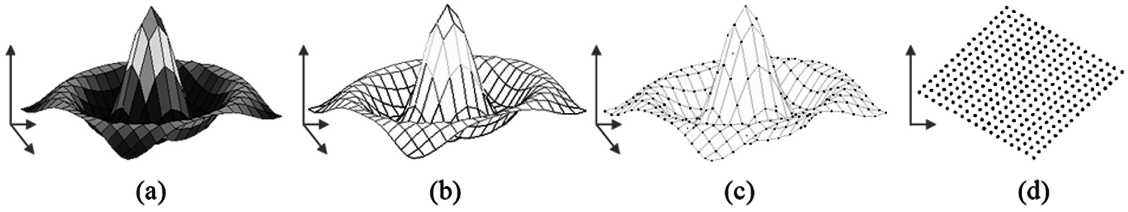

Figure 4. An example of dimensionality reduction by LLE. (a) 3-dimension manifold. (b) sampled line on manifold. (c) sampled points on manifold, (d) points in the reduced dimension by LLE

4. 결 과

본 장은 미국 대통령 취임사를 활용하여, 문서 및 주제어 어간의 군집화를 활용한 트렌드분석 결과를 보인다. 또한 주제어 어간의 시각화를 통한 관찰 내용을 제시한다.

4.1 연설문의 트렌드

각 시기별 미국 대통령 취임 연설문의 트렌드를 살펴보기 위한 *k*-평균 군집화 분석에 앞서, 본 연구에서는 적정 군집수의 측정을 시도하였는데 여러 기법 중 널리 쓰이고 있는 실루엣 통계량을 활용하였다(Rousseeuw, 1987). 실루엣 통계량은 군집 내 밀집의 정도와 군집 간 분리의 정도를 나타내며, 큰 값을 가질수록 좋은 군집이 형성된 결과로 판단한다. <Table 2>는 *k*-평균 군집화의 군집수를 2부터 6까지 변경하며, 실루엣 통계량을 산출한 결과인데, 군집수를 2개로 설정했을 때 실루엣 통계량 값이 최대가 됨을 알 수 있었다. 이 결과는 미국 대통령 취임 연설문이 세계 1차 대전이 발발하여 국가채무가 급격하게 증가하기 직전인, 1901년을 중심으로 이전과 이후로 구분됨을 알 수 있었다.

Table 2. Silhouette statistic values to determine number of clusters in inaugural addresses of the presidents of the United States

	k = 2	k = 3	k = 4	k = 5	k = 6
Silhouette Statistic Value	0.22	0.07	0.02	0.03	0.01

미국 대통령의 취임 연설문으로부터 최적의 군집수를 확인하고, 선정된 주제어 어간들의 출현빈도 데이터를 활용한 *k*-평균 군집화의 결과는 <Figure 5>와 같다. 군집수, *k*는 앞서 실루엣 통계량을 활용하여 측정된 적정 군집수인 2와 비교 군집수 3, 4를 추가로 설정했다. 이때, 거리척도는 주제어의 벡터로 표현되는 개별 문서의 유사성을 상대적으로 잘 반영하는 것으로 알려진 상관관계거리를 사용하였다(Huang, 2008). X-축은 문서의 연도, Y-축은 군집의 색인을 의미하며, 2개의 군집수로 설정하여 분석한 결과(a)는 1901년의 연설문을 중심으로 이전과 이후로 분류됨을 확인할 수 있다. 또한 군집수를 3개(b), 4개(c)로 설정한 경우도 결과 (a)와 동일한 1901년의 분류 점과 함께 1957년, 1837년을 기점으로 군집이 형성되었다. 각각의 시기는 미국의 국가채무비율이 급증하기 직전 시점, 즉 나라빚이 감소하여 안정화되는 시점과 일치한다. 따라서 주제어로 살펴본 미국 대통령 취임 연설문은, 국가채무비율이 변동되는 시점을 기준으로 군집이 나누어지는 것을 확인할 수 있었다.

본 연구는 미국 대통령 취임 연설문의 군집 확인과 함께, 문서 주제의 시대별 트렌드를 파악하기 위하여, 앞서 <Table 1>에서 제시한 문서-단어 교차표의 전치행렬(transpose matrix)을 활용하였다. <Table 3>과 같이 구성되는 전치행렬은 원래의

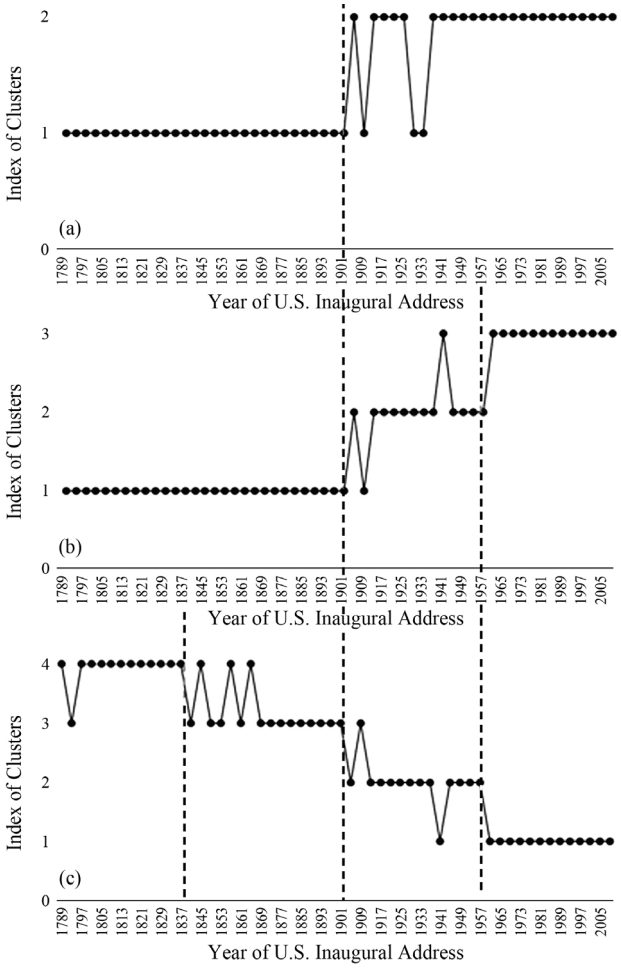

Figure 5. *k*-means clustering results of inaugural addresses of the presidents of the United States, (a) $k = 2$, (b) $k = 3$, (c) $k = 4$

행을 열로, 열을 행으로 바꾼 교차표로, 분석을 위한 관측치가 연설문에서 주제어 어간으로 변경됨을 알 수 있다. 시간 흐름에 따른 주제의 트렌드를 살펴보기 위해 *k*-평균 군집화의 수행하였다. 이때, 실루엣 통계량으로부터 구한 적절한 군집수는 3이었으며, 거리척도는 상관관계거리를 활용하였다.

Table 3. Term-document matrix of inaugural addresses of the presidents of the United States

	1789 Washington	1793 Washington	…	2009 Obama
administr	2	1	⋯	0
alway	0	0	⋯	1
⋮	⋮	⋮	⋮	⋮
whole	0	0	⋯	0

<Figure 6>은 군집화 결과를 활용하여, 각 군집에 속해있는 주제어 어간에 대한 평균출현 정도를 연도별로 보여주고 있다. <군집 1>은 출현양의 추세가 과거에서 현대로 이동하면서

감소하고 있으며, <군집 2>는 1900년대 들어 급격히 증가하는 주제어 어간들의 평균 시계열 패턴을 보여준다. <군집 3>에 속해있는 주제어 어간들은 1900년까지 증가하다가 이후 감소하는 추세를 보였다.

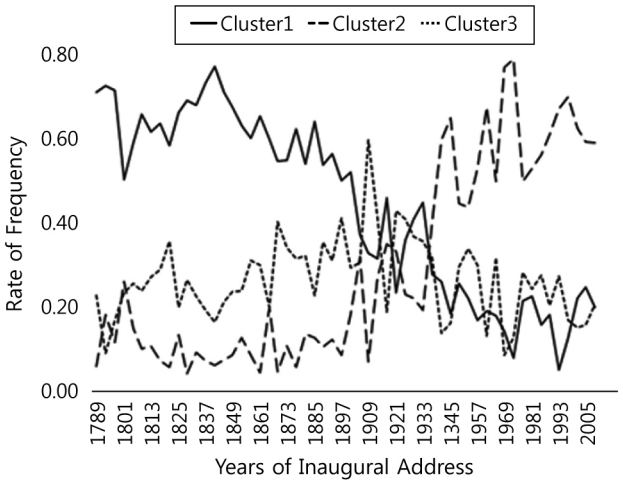

Figure 6. The frequency of keywords in each of 3 clusters of inaugural addresses of the presidents of the United States over the time (1789~2009)

<Table 4>는 *k*-평균 군집화를 통하여 구분된 3개 군집 내 주제어 어간의 목록을 보여주고 있다. 각각의 군집은 <Figure 7>에서 보여주고 있는 패턴인 '상승 트렌드', '하강 트렌드', '1900년 전후 상승 · 감소 트렌드'로 분류하였다. 각 군집에 속한 단어의 최상위 출현비율을 나타내는 3개 단어를 살펴보면, <군집 1> '하강 트렌드'의 경우 'state', 'constitute' 및 'public', <군집 2> '상승 트렌드'는 'peace', 'America' 및 'freedom' 그리고 <군집 3> '1900년 전후 상승 · 감소 트렌드'에는 'law', 'congress', 'import' 등의 단어를 확인할 수 있다. 이와 같은 결과는 미국 대통령 연설문의 주제가 1900년대 이전에는 '공공', '의무'를, 1900년 초반에는 '법', '의회'를, 1900년대 이후에는 '평화', '자유'를 강조하는 트렌드의 변화를 읽을 수 있다. 제 2장에서 밝힌 것과 같이, <Table 4>의 주제어 어간 목록은 무의미한 단어 또는 잘못 표기된 단어로 보일 수도 있으나, 이는 단어의 활용에서 변하지 않는 부분을 나타낸다.

이와 같이 *k*-평균 군집화는 연설문에 존재하는 주제어 어간의 군집을 확인하고, 그 트렌드를 보여주지만 주제어 어간들 사이의 관계를 살펴보는 데는 한계가 있다. 따라서 본 연구에서는 차원 축소를 통해 시각화를 용이하게 할 수 있는 지역선형사상 방법을 활용하여, 연설문에서 사용된 단어들의 관계를 살펴보았다.

4.2 연설문 주제어 어간의 시각화

지역선형사상으로 살펴본 주제어 어간의 시각화 결과는 <Figure 7>과 같다. 오른쪽 하단 점선영역에는 '상승 트렌드' <군집 2>가 뚜렷하게 군집을 형성하는 것을 확인할 수 있었으며, 좌측에는 '하강 트렌드' <군집 1>, 그리고 '1900년 전후 상승 · 감소 트렌드' <군집 3>은 경계선을 형성하며 또 다른 군집을 형성하고 있는 것을 확인할 수 있었다. 이와 같이 시각화의 결과로 확인되는 군집사이의 뚜렷한 구분은, 각각의 군집으로 설명되는 미국 대통령 연설문의 거시적 트렌드를 알 수 있게 해 준다. '상승트렌드' <군집 2>에 포함된 연설문의 주제어 어간들은 다른 군집에 속한 주제어 어간들의 출현 경향과 확실하게 구분되는 상승의 특성을 보인다. 하지만, '하강 트렌드' <군집 1>과 '1900년 전후 상승 · 감소 트렌드' <군집 3>은 어느 정도 경계선을 형성하고 있지만, 확실하게 구분이 되지 않는 경계를 이루고 있으므로 군집간의 성격이 유사함을 알 수 있다. 즉, '1900년 전후 상승 · 감소 트렌드' <군집 3>은 1900년 이전에 상승한다는 점에서 <군집 2>와 유사하거나, 1900년 이후 감소한 다는 점에서 <군집 1>과 비슷한 특성을 보일 수 있다. 하지만, 본 시각화의 결과를 통하여, '1900년 전후 상승 · 감소 트렌드' <군집 3>에 속한 주에 어간들은 '하강 트렌드' <군집 1>과 보다 유사하다는 결과를 얻을 수 있었다.

또한 본 시각화의 결과는 주제어 어간의 군집 형성 결과를 통한 거시적 트렌드의 분석과 함께, 각 개별 주제어 어간을 관찰함으로써 미시적 트렌드 파악을 위한 추가 정보도 얻을 수 있는 장점이 있다. 우측 하단 <군집 2>의 영역 상단에 위치한

Table 4. A list of stemmed keywords in each of 3 clusters based on their patterns (increasing, decreasing, and peaked) in inaugural addresses of the presidents of the United States

군집	주제어 어간	분류
군집 1	author, case, caus, circumst, constitute, depart, desir, duti, effect, event, execut, exercis, exist, extend, foreign, form, general, grant, hand, influenc, institut, interest, legisl, major, money, object, observ, offic, opinion, parti, patriot, perfect, principl, protect, provis, public, regard, relat, republ, respect, revenu, section, spirit, state, subject, support, system, territori, union, upon, want, whole, administer	하강 트렌드
군집 2	america, american, cannot, centuri, challeng, children, democraci, done, dream, earth, freedom, friend, generat, help, histori, ideal, know, let, live, man, million, mr, old, opportune, peac, promis, respons, thing, today, togeth	상승 트렌드
군집 3	alway, amend, busi, chang, condit, congress, defens, econom, elect, enforc, favor, feder, forc, import, increas, industri, intern, interst, law, like, made, maintain, mere, method, must, necessari, negro, ought, pass, polici, practic, product, promot, proper, question, race, reason, south, tariff, tax, trade, use	1900년 전후 상승 · 감소 트렌드

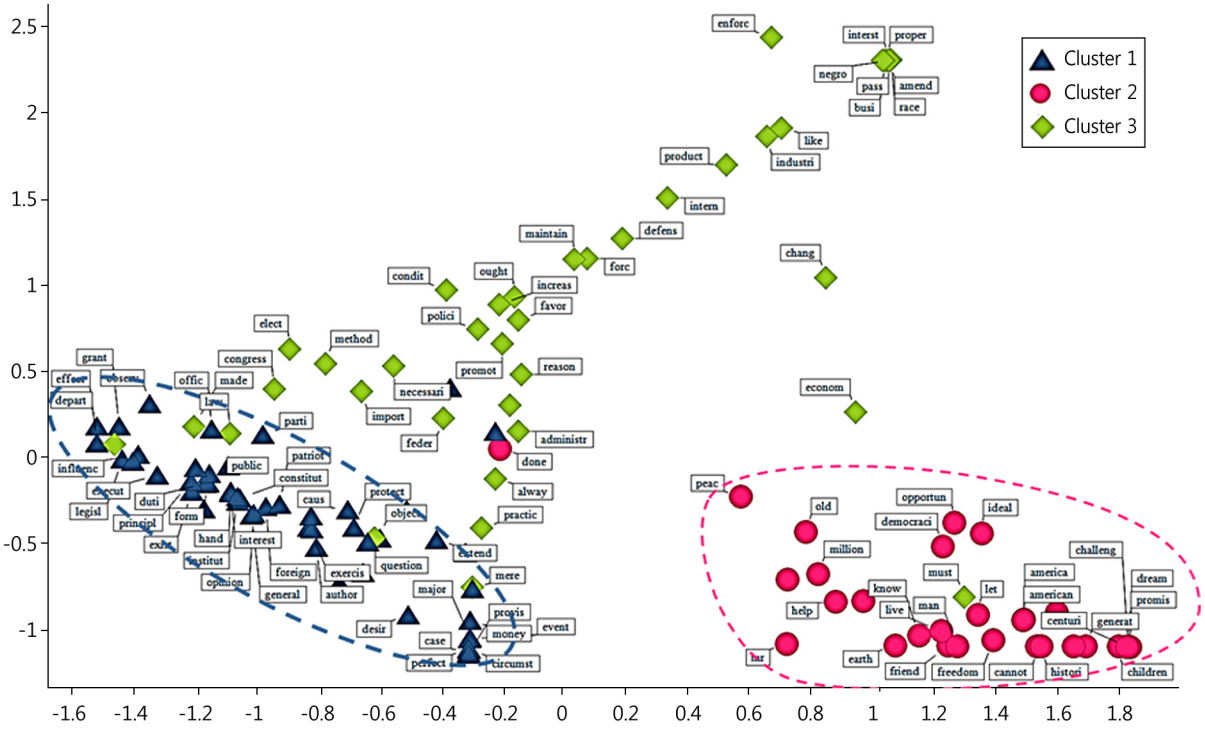

Figure 7. Clustering results of keywords in inaugural address of the president of the United States using locally linear embedding

'econom', 'chang'는 각각 '경제', '변화'를 대표하는 어간으로 '1900년 중심 상승 주제어' <군집 3>에 포함된 단어지만, 1900년대 후반에도 빈번하게 사용된 단어로 확인할 수 있었다. 이와 같이, 차원축소를 활용한 주제어 어간의 시각화는 문서의 트렌드 특성을 탐구함에 있어, 거시적 또는 미시적 관점에서의 분석을 가능하게 하는 좋은 자료가 될 수 있음을 확인했다. 또한, 개념적 사고의 접근이 필수적인 고차원데이터의 연구의 수행에서, 지역선형사상과 같은 차원축소를 활용한 시각화는 직관적이고 용이한 해석을 돕는다는 장점을 확인했다.

본 결과는 미국 대통령 연설문이 시대적 과제 및 지향점의 요약정보를 담고 있다는 점을 감안했을 때, 국정 운영의 관심정도가 시대에 따라 변한다는 점을 보여주었다. 이와 같이 시간에 따른 주제어의 군집 결과는 국가적 의제 설정의 추이를 알아볼 수 있는 매우 흥미로운 자료로도 사용될 수 있을 것이다.

참고로, 본 연구에서는 취임 연설문이 발표된 시기에 대한 분석과 함께, 각 대통령의 소속정당, 초선 또는 재선여부, 전쟁 또는 평화 시기에 대한 군집여부를 확인하였으나, 뚜렷한 차이가 없었음을 부가적으로 밝혀둔다.

5. 결 론

본 연구는 약 200여 년 동안 축적된 56개 미국 대통령 취임사의 트렌드 분석을 수행하였다. 주제어 선정을 위한 전처리 과정을 보다 객관적으로 제시하고, 분석을 위한 문서-단어 교차표와 k-평균 군집화를 활용하여 유사한 특성을 보이는 문서 및 단어들을 군집화 하였다. 또한 지역선형사상을 이용하여 주제어들을 효과적으로 시각화함으로써 텍스트의 객관적, 정량적 트렌드 분석을 위한 연구결과를 도출하였다.

본 연구는 미국 대통령 취임사라는 특정 영역의 문서를 대상으로 분석을 수행하였지만, 제시한 분석 절차 및 기법은 다른 분야에도 충분히 적용이 가능하다. 예를 들면, 특정 제품에 대한 소비자 후기를 분석함으로써 보다 고객 지향적인 제품의 생산 및 판매의 기초데이터로 활용할 수 있다. 또한, 생산현장에서 도출되는 공정의 불량 및 수리 이력을 분석함으로써, 보다 신뢰성 높은 제품 생산을 위한 지식으로 활용할 수 있을 것으로 기대된다.

본 연구에서 주요한 분석기법으로 활용된 k-평균 군집화 방법은 문서 및 주제어의 군집을 이해하는데 매우 효과적인 데이터마이닝 방법임에도 불구하고, 개별 관측치 사이의 상관관계를 살펴볼 수 없다는 한계점을 지닌다. 향후, 사회연결망분석(social network analysis) 또는 연관성분석(association analysis) 등의 방법을 활용한 개별 관측치 사이의 관계를 살펴보는 연구가 수반되어야 할 것으로 보인다.

참고문헌

Aggarwal, C. C. and Zhai, C. (2012), Mining text data, Springer.

Akimoto, M. (2010), *Language Change and Variation from Old English to Late Modern English*, Peter Lang. New York, U.S.

Bird, S. (2006), NLTK : the natural language toolkit, *In Proceedings of the COLING/ACL on Interactive presentation sessions*, 69-72.

Chakraborty, G., Pagolu, M., and Garla, S. (2013), Text Mining and Analysis : Practical Methods, Examples, and Case Studies Using SAS, SAS Institute.

Chen, Y. T. and Chen, M. C. (2011), Using chi-square statistics to measure similarities for text categorization, *Expert systems with applications*, **38**, 3085-3090.

Cho, S. G. and Kim, S. B. (2012), Finding Meaningful Pattern of Key Words in IIE Transactions Using Text Mining, *Journal of the Korean Institute of Industrial Engineers*, **38**(1), 67-73.

Cho, G. H., Lim, S. Y., and Hur, S. (2014), An Analysis of the Research Methodologies and Techniques in the Industrial Engineering Using Text Mining, *Journal of the Korean Institute of Industrial Engineers,* **40**(1), 52-59.

Chris, D. P. (1990), Another Stemmer, *ACM SIGIR Forum,* **24**(3), 56-61.

Gillani, S. A. and Kő, A. (2014), Process-based knowledge extraction in a public authority : A text mining approach, *In Electronic Government and the Information Systems Perspective*, 91-103.

Gordon, A. D. (1999), Classification, Champman and Hall, New York, USA.

Hartigan, J. A. (1975), Clustering Algorithms, John Wiley and Sons, New York, USA.

Hu, X. and Liu, H. (2012), Text analytics in social media, *Mining text data*, 385-414.

Huang, A. (2008), Similarity measures for text document clustering, *Proceedings of the sixth new zealand computer science research student conference*, 49-56.

Hung, J. L. and Zhang, K. (2012), Examining mobile learning trends 2003~2008 : A categorical meta-trend analysis using text mining techniques, *Journal of Computing in Higher Education*, **24**(1), 1-17.

Jain, A. K. and Dubes, R. C. (1988), Algorithms for clustering data, Prentice-Hall, Inc.

Jivani, A. G. (2011), A comparative study of stemming algorithms, *Int. J. Comp. Tech. Appl*, **2**(6), 1930-1938.

Julia, B., Silvia, C., and Giuliana, D. (2013), *Variation and Change in Spoken and Written Discourse : Perspectives from Corpus Linguistics*, John Benjamins publishing company, Philadelphia, U.S.

Kam, J. S., Kim, M. W., and Hyun, B. H. (2013), A Study on Analysis of Patent Information Based Biotechnology Research Trend and Promising Research Themes, *The Korea Society for Innovation Management and Economics*, **21**(2), 25-56.

Kim, H. Y. (2013), Analysis of an Inaugural Address of Korean Presidents Based on Network, *Korea Content Association*, **3**(2), 67-68.

Kim, H. Y., Kim, H. G., and Kang, B. M. (2012), A Trend Analysis of Curtural comsumption Based on Newspaper Texts, *Journal of KIIS E : Software and Applications*, **39**(3), 244-251.

Kim, H. (2014), A Study on Presidential Leadership and Policy Agenda Setting Pattern : A Content Analysis of Korean Presidential Addresses, *Journal of Korean Politics*, **23**(2), 77-102.

Kim, M. and Koo, P. (2013), A Study on Big Data Based Investment Strategy Using Internet Search Trends, *Journal of the Korean Operations Research and Management Science Society*, **38**(4), 53-64.

Kim, M., Notkin, D., Grossman, D., and Wilson, G. (2013), Identifying and summarizing systematic code changes via rule inference, *Software Engineering, IEEE Transactions on*, **39**, 45-62.

Kim, Y., Tian, Y., Jeong, Y., Jihee, R., and Myaeng, S. H. (2009), Automatic discovery of technology trends from patent text. *Proceedings of the 2009 ACM symposium on Applied Computing*, 1480-1487.

Lee, Y. J., Seo, J. H., and Choi, J. T. (2014), Fashion Trend Marketing Prediction Analysis Based on Opinion Mining Applying SNS Text Contents, *The Journal of Korean Institute of Information Technology*, **12**(12), 163-170.

Lim, E. T. (2002), Five trends in presidential rhetoric : An analysis of rhetoric from George Washington to Bill Clinton, *Presidential Studies Quarterly*, **32**(2), 328-348.

Liu, B. (2012), Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies*, **5**(1), 1-167.

Lovins, J. B. (1968), Development of a stemming algorithm, MIT Information Processing Group, Electronic Systems Laboratory.

Min, K. Y., Kim, H. T., and Ji, Y. G. (2014), A Pilot Study on Applying Text Mining Tools to Analyzing Steel Industry Trends : A Case Study of the Steel Industry for the Company "P", *Society for E-Business Studies*, **19**(3), 51-64.

Pai, M. Y., Chen, M. Y., Chu, H. C., and Chen, Y. M. (2013), Development of a semantic-based content mapping mechanism for information retrieval, *Expert Systems with Applications*, **40**, 2447-2461.

Park, H., Seo, W., Coh, B., Lee, J. and Yoon, J. (2014), Technology Opportunity Discovery Based on Firms' Technologies and Products, *Journal of the Korean Institute of Industrial Engineers*, **40**(5), 442-450.

Porter, M. (2001), Snowball : A language for stemming algorithms, http://snowball.tartarus.org/texts/introduction.html.

Porter, M. F. (1980), An algorithm for suffix stripping, *Program : electronic library and information systems*, **14**(3), 130-137.

Pramokchon, P. and Piamsa-nga, P. (2014), A feature score for classifying class-imbalanced data, *In Computer Science and Engineering Conference (ICSEC),* 409-414.

Rajaraman, A. and Ullman, J. D. (2011), Mining of massive datasets, *Cambridge University Press*.

Rebholz-Schuhmann, D., Kirsch, H., and Couto, F. (2005), Facts from text-Is text mining ready to deliver?, *PLoS biology*, **3**(2), e65.

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.

Rowie, S. T. and Saul, L. K. (2000), Nonlinear Dimensionality Reduction by Locally Linear Embedding, *SCIENCE*, **290**(5500), 2000-2326.

Saul, L. K., and Roweis, S. T. (2000), An Introduction to Locally Linear Embedding, http://cs.nyu.edu/~roweis/lle/publications.html.

Zhang, J., Kawai, Y., and Kumamoto, T. (2010), A Flexible Re-ranking System Based on Sub-keyword Extraction and Importance Adjustment, *IAENG International Journal of Computer Science*, **37**(3), 1-8.