Journal of the Korean Data & Information Science Society
2015, **26**(1), 151–159

http://dx.doi.org/10.7465/jkdi.2015.26.1.151
한국데이터정보과학회지

토픽 모형 및 사회연결망 분석을 이용한 한국데이터정보과학회지 영문초록 분석

김규하[1] · 박철용[2]

[12]계명대학교 통계학과

접수 20140년 12월 16일, 수정 2015년 1월 2일, 게재확정 2015년 1월 10일

요 약

이 논문에서는 텍스트마이닝 (text mining) 기법을 이용하여 한국데이터정보과학회지에 게재된 논문의 영어초록을 분석하였다. 먼저 다양한 방법을 통해 단어-문서 행렬 (term-document matrix)을 생성하고 이를 사회연결망 분석 (social network analysis)을 통해 시각화하였다. 또한 토픽을 추출하기 위한 방법으로 LDA (latent Dirichlet allocation)와 CTM (correlated topic model)을 사용하였다. 토픽의 수, 단어-문서 행렬의 생성방법에 따라 엔트로피 (entropy)를 통해 토픽 추출 모형들의 성능을 비교하였다.

주요용어: 사회연결망 분석, 텍스트마이닝, 토픽 모형, 한국데이터정보과학회지.

1. 서론

스마트폰의 보급률이 높아지면서 스마트폰의 활용도를 높이기 위해 다양한 어플리케이션들이 개발되었다. 그 중 대부분의 스마트폰 유저들은 메신저와 사회연결망 서비스 어플리케이션을 사용하고 있거나 사용한 경험이 있을 것이다. 위 어플리케이션들은 스마트폰을 이용하여 서로 문자를 주고받거나, 자신의 정보 및 상태를 글과 사진을 통하여 실시간 업로드를 할 수 있는 어플리케이션이다. 이 어플리케이션을 통해 구조와 형태가 복잡하고 정형화되지 않은 글, 사진, 영상과 같은 비정형 데이터들이 생성되고 있다. 예를 들어 대표적 사회연결망 서비스 어플리케이션인 페이스북의 사용자는 2014년 6월 기준 약 13억 2천만 명이다. 한 명이 하루에 한 문장의 정보를 업로드 한다고 가정하면 일주일이면 대략 92억 개의 문장이 생성이 된다. 이렇게 생성된 92억 개의 문장에서 중요한 정보를 가진 단어들을 추출하고 추출된 단어들과 문장들 간 관계를 파악하기 위해 비정형 데이터 분석 방법 중 하나인 토픽 모형 (topic models)을 적용할 수 있을 것이다.

토픽 모형은 문서를 이루는 기반이지만 관측할 수 없는 토픽을 단어를 통해 찾고자 하는 통계적 모형의 일종이다. 다양한 토픽 모형들이 존재하지만 Blei 등 (2003)이 제안한 LDA (latent Dirichlet allocation)와 Blei와 Lafferty (2006)가 제안한 DTM (dynamic topic models), Blei와 Lafferty (2007)가 제안한 CTM (correlated topic models)이 지금까지도 많이 사용되고 있다. 본 논문에서는 LDA와 CTM을 실제 자료에 적용하여 비교분석하고자 한다. LDA의 경우 토픽 1에 포함된 단어들과 토픽 2에 포함된 단어들 간의 상관성을 가정하지 않는 반면, CTM의 경우 상관성이 있다는 가정을 하게 된다. 이러한 모형의 특징은 분포에 기인한 것으로 2.2절에서 자세한 설명을 할 것이다.

[1] (704-701) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학과, 석사과정생.
[2] 교신저자: (704-701) 대구광역시 달서구 달구벌대로 1095, 계명대학교 통계학과, 교수.
E-mail: cypark1@kmu.ac.kr

토픽 모형의 분석결과로 엔트로피 (entropy)와 각 토픽에 속한 단어들이 출력된다. 그러나 출력된 단어들과 문서가 어떤 관계를 가지는지 알 수는 없다. 관계를 파악하기 위한 방법으로 사회연결망 분석 (social network analysis)을 이용하여 시각화할 수 있다. 각 모형에서 추출된 단어 일부를 이용하여 단어들 간의 관계, 나아가 토픽들 간의 관계를 시각화 자료를 통해 해석할 수 있다. 또한 빈도가 높은 단어들을 추출하고 추출된 단어들과 문서들의 사회연결망을 통해 문서의 특징에 대해 분석하고자 한다.

이 논문은 다음과 같이 구성되어 있다. 2절에서 토픽 모형의 구조와 각 모형의 특징을 설명하고, 3절에서는 LDA와 CTM으로 토픽과 토픽에 포함된 단어를 추출하고 단어들과 논문 간의 관계를 사회연결망으로 시각화한다. 마지막 4절에서는 이 연구의 결론을 내리고 추후 연구 과제를 제시한다.

2. 토픽 모형

2.1. LDA와 CTM

Blei 등 (2003)이 제안한 LDA는 전체 문서의 단어들에 Dirichlet 분포를 적용해 토픽들을 구성시키는 토픽 모형의 한 방법론이다. 또한 계층적 구조로서 각 문서에서 토픽들을 구성하고 구성된 토픽에서 단어를 추출하게 된다. Blei와 Lafferty (2009)에 의하면 LDA는 관측된 각 문서의 단어들을 이용하여 은닉 변수 (hidden variables)로 표현된 잠재 토픽 구조 (latent topical structure)를 예측하고자 하는 것이다. Figure 2.1은 LDA를 시각화한 것이다 (Blei와 Lafferty, 2009).

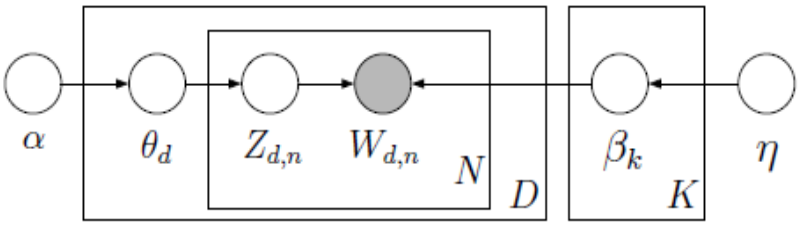

Figure 2.1 Visualization of LDA

K: 토픽의 수
V: 중복되지 않은 단어인 어휘의 수
α : 양의 K-벡터 모수, 사전확률 (prior)
η : 상수 모수, 사전확률
$\beta_k \sim Dir_V(\eta)$: k-번째 토픽의 단어 분포의 확률 β_k 추출
$\theta_d \sim Dir(\alpha)$: d-번째 문서의 토픽 분포의 확률 θ_d 추출
$Z_{d,n} \sim Mult(\theta_d), Z_{d,n} \in \{1, \cdots, K\}$: d-번째 문서의 토픽 분포에서 n번째 토픽 $Z_{d,n}$ 추출
$W_{d,n} \sim Mult(\beta_{z_{d,n}}), W_{d,n} \in \{1, \cdots, V\}$: 토픽 $Z_{d,n}$의 단어 분포에서 단어 $W_{d,n}$ 추출
Dirichlet 밀도함수(density): $p(\theta|\alpha) = \dfrac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$

Figure 2.1에서 단어 (W_d, n)는 관측된 데이터이다. α, β는 단어를 추출하기 위해 사용된 모수이고, θ, Z는 직접 관측할 수 없거나 측정되지 않는 잠재변수 (latent variable)이다. LDA에서 $Z_d, n|\theta$가 다항분포를 따른다는 정보를 가지고 있기 때문에, 공액 사전분포 (conjugate prior)로 Dirichlet 분포를 사용하게 되면 사후확률의 분포는 다항분포가 됨을 쉽게 알 수 있다.

LDA는 토픽들 사이에 상관관계가 존재할 경우 모델링이 어려운 것으로 알려져 있다. 그러나 실생활의 다양한 문서들은 서로 상관관계가 존재하는 것이 일반적인 현상일 것이다. 이와 같이 토픽들 간에 상관관계가 존재할 경우 CTM을 사용하게 된다.

LDA와 CTM의 차이점은 문서의 토픽 분포가 LDA의 경우 Dirichlet 분포를 따르고, CTM의 경우 로지스틱 정규분포 (logistic normal distribution)를 따른다는 것이다. Figure 2.2는 CTM을 시각화한 것이다 (Blei와 Lafferty, 2007).

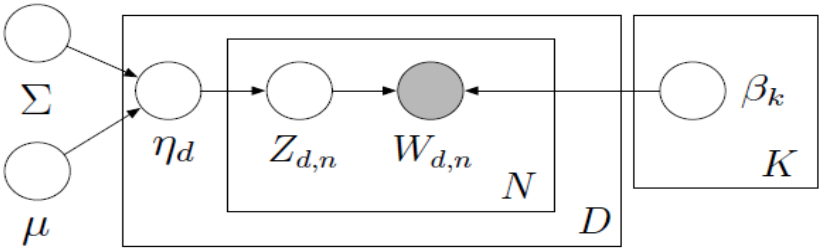

Figure 2.2 Visualization of CTM

$\{\mu, \Sigma\}$: K-차원의 평균과 공분산 행렬
β_k : k-번째 토픽의 단어 분포의 확률
$\eta | \{\mu, \Sigma\} \sim N(\mu, \Sigma)$
$n \in \{1, \cdots, N\}$:

$$Z_n | \eta \sim Mult(f(\eta))$$
$$W_n | z_n \sim Mult(\beta_{z_n})$$

여기서 $f(\eta) = \dfrac{\exp\{\eta_i\}}{\sum_j \exp\{\eta_j\}}$ 이다.

CTM의 경우 토픽들 간의 상관관계를 모형화할 수 있기 때문에 LDA보다 실제 문서를 분석하는데 더 유용하다. 그러나 CTM은 계산시간이 오래 걸리며, 특히 토픽 분포의 변동 모수 (variational parameter)의 분포에 대한 업데이트에서 기울기 기반 최적화 (gradient -based optimization)에 의한 적합이 이루어져야만 하는 문제점이 있다.

2.2. Dirichlet 분포와 로지스틱 정규분포

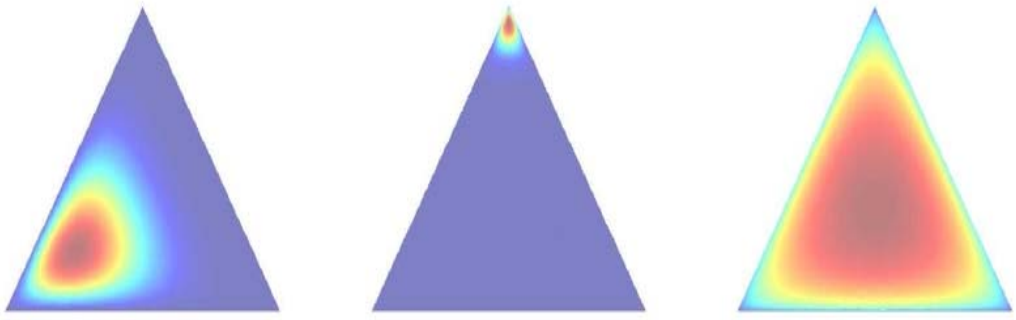

Figure 2.3 Dirichlet distributions for various parameter settings on a 2-simplex

Figure 2.3은 Dirichlet 분포의 모수 α의 값을 바꾸어 가며 나타낸 단어의 심플렉스 (simplex)와 토픽의 부-심플렉스 (sub-simplex)이다 (Huang과 Malisiewicz, 2006). Figure 2.3의 왼쪽과 가운데 그림은 토픽의 부-심플렉스가 단어의 심플렉스에 치우쳐 나타남을 의미하고, 오른쪽 그림은 토픽의 부-심플렉스가 단어 심플렉스 전체에 나타남을 의미한다. 그러므로 LDA에서 토픽들 간에 거의 독립이 되는 이유가 Dirichlet 분포의 특징에 기인함을 알 수 있다.

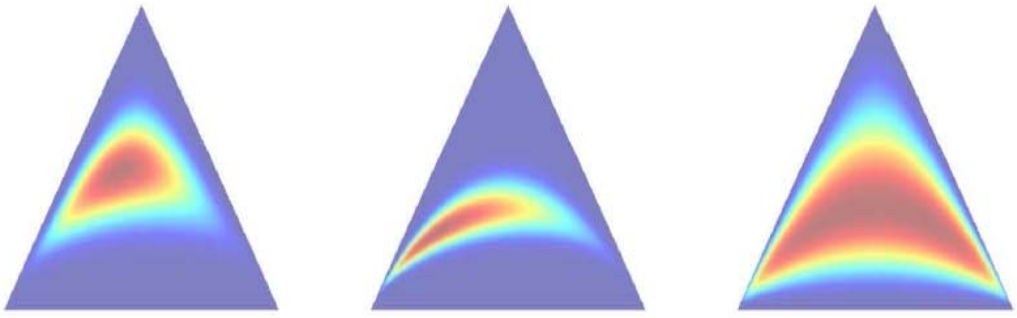

Figure 2.4 Logistic normal distributions for various parameter settings on a 2-simplex

CTM에서 μ, Σ가 주어졌을 때 문서의 토픽 분포의 확률인 η의 조건부 분포는 $\eta | \{\mu, \Sigma\} \sim N(\mu, \Sigma)$가 된다. Figure 2.4는 μ, Σ에 따른 단어의 심플렉스와 토픽의 부-심플렉스를 나타내고 있다 (Huang과 Malisiewicz, 2006). 세 그림 모두 토픽 부-심플렉스의 형태가 단어의 심플렉스에서 두 꼭지점에 걸쳐 나타남을 볼 수 있다. 그러므로 모수 μ, Σ에 따라 강한 상관을 보이는 것이 CTM의 특징임을 알 수 있다.

3. 자료 분석

이 절에서는 2절에서 설명한 LDA와 CTM을 실제 데이터에 적용하고 그 결과를 비교분석한다. 자료 분석에는 통계프로그래밍 언어인 R을 이용하였다. 본 연구에서는 단어-문서 행렬 (term-document matrix)을 만들고 LDA와 CTM을 이용하여 토픽을 추출하기 위해 tm 및 topicmodels 패키지를 사용하였고, 두 토픽 모형의 특징을 시각화하기 위해 igraph 패키지를 이용하여 사회연결망으로 나타내었다.

LDA와 CTM을 비교분석하기 위해 사용된 데이터는 2013년 한국데이터정보과학회지에 게재된 142편의 논문의 영어초록이다. 원 데이터를 이용하여 단어-문서 행렬을 만들고, 계산된 단어-문서 행렬을 이용하여 모형을 생성한 후에 토픽의 수에 따라 토픽 모형에서 추출된 단어에 어떤 차이가 있는지 확인하였다.

먼저 토픽 모형에 적용하기 전 빈도가 높은 단어들과 논문들 간의 관계를 시각화시켜 보았다.

Figure 3.1은 빈도수가 25회 이상인 단어들을 추출하여 사회연결망으로 나타내었다. 글자의 크기는 연결선 수 (degree)에 의해 크기가 결정되며, 녹색은 추출된 단어의 노드 (node)이고, 붉은색은 142편의 논문의 노드를 의미한다. 논문에서 가장 많이 출현한 단어는 data (70회)이며, 다음으로 model (60회), results (59회), method (55회) 순으로 나타났다. Figure 3.1에서 나타난 단어들이 가장 많이 쓰인 논문은 125번째 논문 (Shim 등, 2013)이다. 특이한 점은 2번째 논문 (Chung과 Han, 2013)으로 영문초록에서 위 단어가 하나도 쓰이지 않았다는 점이다.

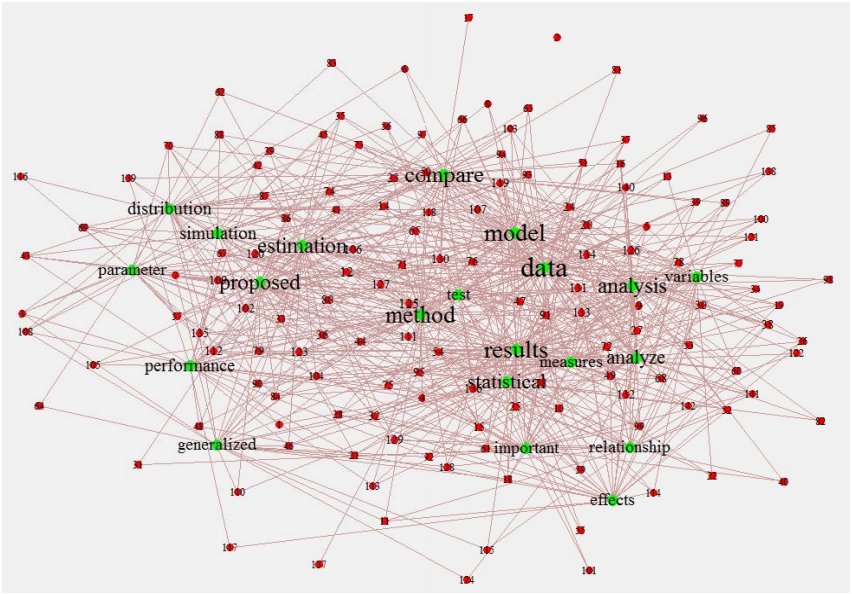

Figure 3.1 Social network analysis among the terms with frequencies of 25 or more and documents

Table 3.1은 단어-문서 행렬을 이용하여 토픽 모형에 적용한 결과이다.

Table 3.1 Entropies of LDA and CTM

# of topics\model	LDA	CTM
K=10	0.1909	0.3771
K=15	0.1414	0.1911
K=20	0.1347	0.2083
K=25	0.1114	0.1840
K=30	0.1543	0.1614

Table 3.1의 엔트로피는 문서에서 토픽의 퍼짐 정도라 할 수 있다. 토픽의 수에 관계없이 엔트로피 값이 작고 변동이 적은 모형은 LDA이다. 엔트로피 값이 가장 작게 나타난 K=25일 때의 LDA에 대해 각 토픽에 포함된 6개의 단어를 나타내면 Table 3.2와 같다.

Table 3.2 Topics and associated terms for LDA with K=25

term\topic	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
term 1	data	test	data	design	sport	prior
term 2	model	data	bat	risk	emotion	bayesian
term 3	proposed	estimation	hits	type	preterm	matching
term 4	broadcast	statistical	big	industrial	infants	order
term 5	estimation	chisquare	election	model	breast	estimation
term 6	method	bias	participation	allocation	feeding	developed

Table 3.2에 의하면 토픽 1과 토픽 2에는 일반적인 단어들이 추출되었고, 토픽 3에는 9월 특별호에 게재된 빅 데이터와 관련된 단어들이 주로 추출되었다. 토픽 4에는 위험관리와 관계된 단어와 산업공학

에서 많이 사용하는 단어가 추출되었고, 토픽 5에는 간호학 및 축산학 관련 논문 단어들과 스포츠와 관련된 논문의 단어들이 추출되었다. 토픽 6에는 베이지안 통계와 관련된 단어들이 추출되었음을 알 수 있다. Table 3.2에서 빈도가 높은 일부 단어들을 이용하여 사회연결망으로 시각화시켜 보았다.

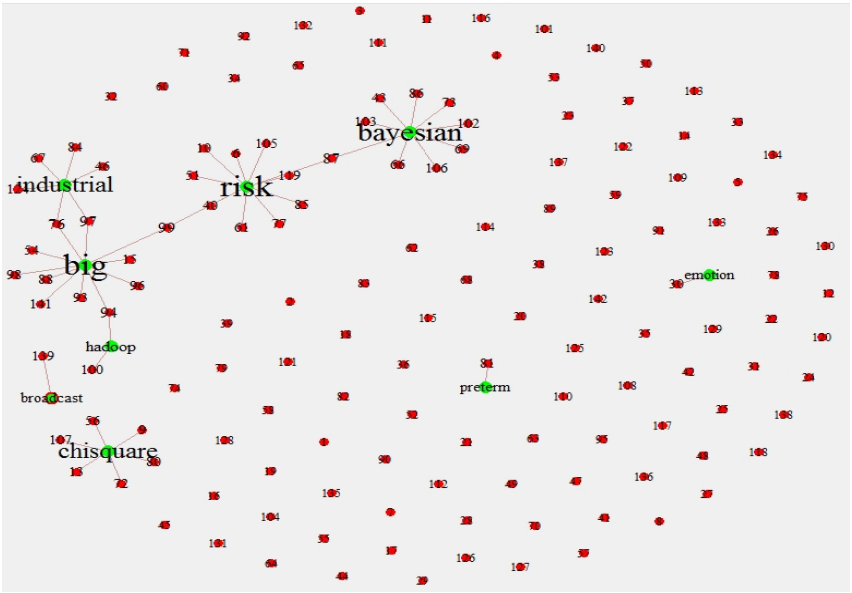

Figure 3.2 Social network analysis for some frequent terms from LDA with K=25 and documents

Figure 3.2는 선정된 단어들과 논문들 간의 관계를 나타내고 있다. emotion, broadcast, hadoop 등의 단어들은 하나 또는 두 논문에서 나타나고 있으며, bayesian, risk, big 단어는 여러 논문에 나타남을 보이고 있다. 그러나 여러 논문에서 나타나지 않고 한 논문에서만 나타나는 단어들이 일부 보인다. 이는 2.2절에서 언급한 LDA의 특징을 나타내는 것이라 생각된다.

Table 3.1을 살펴보면 CTM에서 엔트로피 값이 가장 작을 때는 K=30일 때이다. 이 때 각 토픽에 포함된 6개의 단어를 나타내면 Table 3.4와 같다.

Table 3.3 Topics and associated terms for CTM with K=30

term\topic	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
term 1	preterm	method	method	milk	statistical	estimation
term 2	infants	generalized	proposed	flow	small	confidence
term 3	breast	estimation	type	average	area	interval
term 4	feeding	process	estimation	main	estimation	model
term 5	terms	identified	regression	minute	model	distribution
term 6	allocation	multivatiate	data	phase	difficult	method

Table 3.4의 토픽 1에는 간호학과 관련된 논문에서 주로 나타난 단어들이 추출되었고, 토픽 2에는 일반적인 통계학 논문에서 나타나는 단어들이 추출되었다. 토픽 3에는 회귀분석과 관련하여 연구된 논문의 단어들이 추출되었고, 토픽 4에는 축산과 관련된 논문의 단어들이 추출되었다. 토픽 5에는 소지역 추정 기법 (small area estimation)과 관련된 단어들이 추출되었고, 토픽 6에는 분포의 신뢰구간 추정과

관련된 단어들이 추출되었음을 알 수 있다. Figure 3.3은 CTM에서 추출된 일부 단어를 이용하여 사회연결망으로 시각화시켜 보았다.

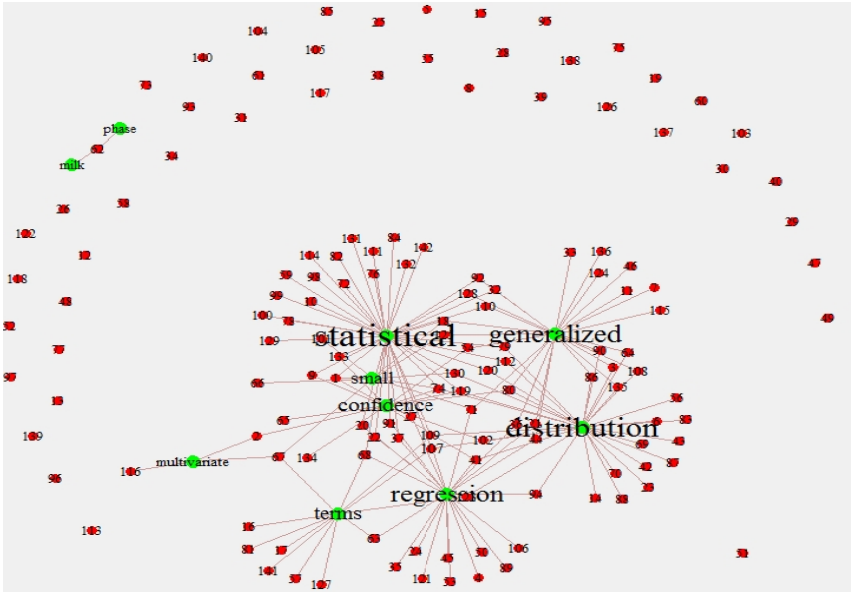

Figure 3.3 Social network analysis among some frequent terms from CTM with K=30 and documents

Figure 3.3에서 milk, phase 두 단어를 제외한 나머지 단어들은 여러 개 논문들에 연결되어 있음을 알 수 있다. 다시 말해 보통 하나의 논문이 2개 이상의 단어들로 연결되고, 연결된 단어들도 한 토픽에서 추출된 단어뿐만 아니라 다른 토픽의 단어와도 연결고리를 가지고 있는 것이다. 이 또한 2.2절에서 설명한 CTM의 특징을 시각적으로 잘 드러낸 경우라고 할 수 있다.

4. 결론

토픽 모형은 아직 활발히 연구 중인 비정형 데이터 분석의 한 분야이다. 본 연구에서는 토픽 모형에서 널리 알려진 LDA와 CTM 방법을 이용하였다. 또한 LDA와 CTM의 이론적 특징을 이해하기 쉽도록 시각화하기 위해 사회연결망 분석을 이용하였다. 두 토픽 모형을 2013년 한국데이터정보과학회지에 게재된 논문의 영어 초록에 적용한 결과 LDA에서는 일부 논문에서 사용되는 특정한 단어들이 많이 추출된 반면, CTM에서는 여러 논문에서 공동으로 나타나는 단어들이 많이 추출되었다. 그러므로 사용자의 요구에 따라 두 모형을 적절히 사용할 수 있을 것이다. 또한 사회연결망을 이용한 시각화 과정에서 특정 단어와 관계된 문서를 찾거나 수치적으로 나타낼 수 있음을 알 수 있었다. 나아가 한글 텍스트의 형태소 분석과 어근 추출에 따른 문제점이 줄어든다면, 다양한 한글 문서에 적용하여 분석 및 시각화를 할 수 있을 것이다. 또한 다년간 축적된 대용량의 데이터일 경우에도 본 연구의 결론과 동일할지는 추후 연구가 필요할 것이다.

References

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113-120.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, **1**, 17 -35.

Blei, D. M. and Lafferty, J. D. (2009). Topic models. In *Text Mining: Classification, Clustering, and Applications,* edited by A. N. Srivastava and M. Sahami, Champman and Hall/CRC, Boca Raton, 71-94.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

Chung, H. and Han, C. (2013). Conditional bootstrap confidence intervals for classification error rate when a block of observations is missing. *Journal of the Korean Data & Information Science Society*, **24**, 189-200.

Hornik, K. and Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, **40**, 1-30.

Huang, J. and Malisiewicz, T. (2006). *Correlated topic model details,* Technical Report, Carnegie Mellon University, Pittsburgh, PA.

Shim, J., Kim, Y. and Hwang, C. (2013). Generalized kernel estimating equation for panel estimation of small area unemployment rates. *Journal of the Korean Data & Information Science Society*, **24**, 1199-1210.

Journal of the Korean Data & Information Science Society
2015, **26**(1), 151–159

http://dx.doi.org/10.7465/jkdi.2015.26.1.151
한국데이터정보과학회지

Analysis of English abstracts in Journal of the Korean Data & Information Science Society using topic models and social network analysis

Gyuha Kim[1] · Cheolyong Park[2]

[12]Department of Statistics, Keimyung University

Received 16 December 20140, revised 2 January 2015, accepted 10 January 2015

Abstract

This article analyzes English abstracts of the articles published in Journal of the Korean Data & Information Science Society using text mining techniques. At first, term-document matrices are formed by various methods and then visualized by social network analysis. LDA (latent Dirichlet allocation) and CTM (correlated topic model) are also employed in order to extract topics from the abstracts. Performances of the topic models are compared via entropy for several numbers of topics and weighting methods to form term-document matrices.

Keywords: Journal of the Korean & Information Science Society, social network analysis, text mining, topic models.

[1] Master candidate, Department of Statistics, Keimyung University, Daegu 704-701, Korea.
[2] Corresponding author: Professor, Department of Statistics, Keimyung University, Daegu 704-701, Korea. E-mail: cypark1@kmu.ac.kr