JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 1, pp. 51-58, Jan. 2018

WV-BTM: SNS 단문의 주제 분석을 위한 토픽 모델 정확도 개선 기법

송 애 린 · 박 영 호*
숙명여자대학교 공과대학 IT공학과

WV-BTM: A Technique on Improving Accuracy of Topic Model for Short Texts in SNS

Ae-Rin Song · Young-Ho Park*
Department of IT Engineering, Sookmyung Women's University, Seoul 04310, Korea

[요 약]

SNS의 사용자와 데이터량이 폭발적으로 증가함에 따라, SNS 빅 데이터를 기반으로 한 연구들이 활발히 진행되고 있다. 특히 소셜 마이닝 분야에서는 비 분류된 대용량 SNS 텍스트 데이터로부터 각 텍스트 별 유사성을 파악하고, 그로부터 트렌드를 추출하기 위해 대표적인 토픽 모델 기법인 LDA를 사용한다. 그러나 LDA는 단문 데이터에 대하여 비 빈발 단어 출현으로 인한 의미 희박성(semantic sparsity)으로 인해 양질의 주제 추론이 어렵다는 한계를 가진다. BTM 연구는 이와 같은 LDA의 한계점을 두 단어의 조합을 통해 개선하였으나, BTM 또한 조합된 단어 중 높은 빈도수의 단어에 더 큰 영향을 받아 각 주제와의 연관성을 고려한 가중치 계산이 불가능하다는 한계점을 지닌다. 본 논문은 단어 간의 의미적 연관성을 반영함으로써 기존 연구 BTM의 정확도를 개선하는 방안을 모색한다.

[Abstract]

As the amount of users and data of NS explosively increased, research based on SNS Big data became active. In social mining, Latent Dirichlet Allocation(LDA), which is a typical topic model technique, is used to identify the similarity of each text from non-classified large-volume SNS text big data and to extract trends therefrom. However, LDA has the limitation that it is difficult to deduce a high-level topic due to the semantic sparsity of non-frequent word occurrence in the short sentence data. The BTM study improved the limitations of this LDA through a combination of two words. However, BTM also has a limitation that it is impossible to calculate the weight considering the relation with each subject because it is influenced more by the high frequency word among the combined words. In this paper, we propose a technique to improve the accuracy of existing BTM by reflecting semantic relation between words.

색인어 : 소셜 네트워크 서비스, 자연어 처리, 텍스트 마이닝, 토픽 모델, 클러스터링,

Key word : Social Network Service, Natural Language Processing, Text Mining, Topic Model, Clustering

http://dx.doi.org/10.9728/dcs.2018.19.1.51

Received 18 December 2017 ; **Revised** 23 January 2018
Accepted 29 January 2018

***Corresponding Author; Young-Ho Park**

Tel: +82-2-2077-7297
E-mail: yhpark@sm.ac.kr

Ⅰ. 서 론

스마트폰의 보급 및 정보통신 기기 기반의 환경이 발달함에 따라, 2015년 1월을 기준으로 인터넷 이용자 수 30억 명, 모바일 사용자가 36억 명으로 증가하였다. SNS(Social Network Service) 또한 활동 중인 소셜 미디어 계정이 21억 개로 나타남으로써 활발히 이용 되고 있다[1]. SNS 중에서도 트위터(Twitter), 페이스북(Facebook), 인스타그램(Instagram)과 같은 서비스들은 등장 초반부터 사용자들의 일상 및 관심사에 대한 직접적인 의견을 신속하게 공유할 수 있다는 점에서 각광을 받았다[2]. 그림 1은 2010년부터 2017년 간 전 세계 SNS 사용자 추이에 대한 그림으로, 해당 기간의 사용자의 수가 지속적인 성장세를 보임을 알 수 있다 [3].

사용자들 간의 네트워크가 면밀해짐에 따라 SNS 미디어 콘텐츠 또한 방대한 양으로 증가 되고 있다. Intel사의 2013년 SNS 데이터 및 사용자 활동에 대한 조사[4]에 따르면, 트위터는 하루 약 1억 4천만 데이터가 생성 되며, 페이스북 사용자는 86억 건 이상의 데이터를 읽는 것으로 나타난다. 이와 같은 SNS 미디어 콘텐츠 데이터도 빅 데이터(Big data)로써 마케팅 및 서비스의 고도화를 위한 원료로 귀중한 분석의 대상이 되고 있다.

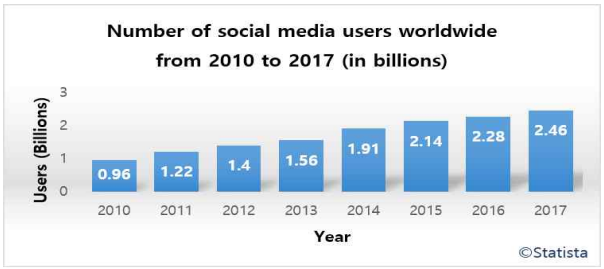

그림 1. 전 세계 SNS 사용자 추이 (2010-2017)
Fig. 1. Number of Social Media Users Worldwide (2010-2017)

SNS의 데이터는 대부분 단문 텍스트 데이터로 구성된다[5]. 대표적인 서비스 중 하나인 트위터의 경우, 140자 내로 구성된 텍스트 전송 규격인 트윗(tweet)과 트윗을 재공유하는 리트윗(retweet)의 개념을 통해 간결하게 사용자의 의견 및 소식을 쉽게 전파한다[6]. 이로 인해 대두된 사회적 이슈 및 트렌드 변화를 파악하는 소셜 마이닝(Social Mining) 분야의 연구가 활발하게 진행 되고 있다. 소셜 마이닝 분야에서는 사전 정보가 없이 비 분류된 대용량 SNS 텍스트 데이터로부터 각 텍스트 별 유사성을 파악하고, 그로부터 트렌드를 추출하기 위해서 주로 토픽 모델 기법을 사용한다.

대표적인 토픽 모델 알고리즘은 LDA(Latent Dirichlet Allocation)[7]이며, 다수의 SNS 트렌드 분석 연구가 LDA를 기반으로 이뤄졌다[8-12]. 그러나 140자 이내로 구성된 짧은 텍스트, 즉 단문은 전체 텍스트 중 불용어(Stop-words)를 제외한 의미를 내포한 단어의 희소성으로 인해 주제 추론에 대한 정확도가 낮은 한계점을 가진다.

본 논문에서는 BTM의 성능을 개선 시킨 WV-BTM을 제안함으로써 단문의 주제 추론의 정확성을 보다 높인 토픽 모델 기법을 제안한다. 본 논문의 공헌은 다음과 같다.

1) SNS의 단문에 적합한 토픽 모델 기법의 정확도 향상을 통하여 보다 정확한 트렌드 변화 추이를 확인할 수 있는 텍스트 마이닝 연구를 진행한다.

2) 해당 연구는 주제 추론의 정확도를 개선시키기 위하여 추가 데이터에 의존적인 기존의 SNS 토픽 모델 연구와 달리, 추가 데이터를 사용하지 않고 토픽 모델 개선 연구를 진행한다.

본 논문의 구성은 다음과 같다. 2장에서는 이론적 배경인 자연어 처리, 토픽 모델링에 대해 설명한 뒤, 주요 토픽 모델 알고리즘인 LDA 및 BTM의 비교를 통해 한계점을 살핀다. 3장에서는 BTM의 정확도를 개선한 토픽 모델 기법인 WV-BTM에 대하여 서술한다. 4장에서는 데이터 수집 및 전처리 방안, 실험 방안에 대하여 서술한다. 5장에서는 설계한 실험을 통해 해당 알고리즘이 SNS 단문에서 기존의 알고리즘 대비 향상된 점 및 기대효과에 대하여 기술한다. 마지막으로 6장에서는 본 논문의 결론 및 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

본 장에서는 첫째, 자연어 처리에 대하여 논한다. 둘째, 토픽 모델 기법에 대하여 설명 후, 토픽 모델 기법 중에서 SNS텍스트 데이터에서부터 트렌드를 추출하기 위하여 주로 사용된 기존의 토픽 모델 연구들을 비교한다. 셋째, 분석을 통하여 발견한 기존 연구들의 한계성 및 제안 연구의 차별성에 대해 논의한다.

2-1 자연어 처리

자연어 처리(Natural Language Processing, NLP)란 인간의 발화 언어 즉 자연어의 현상을 컴퓨터가 이해 가능한 형태로 처리 및 분석하고, 이를 다시 이해 가능한 자연어적 형태로 표현하는 기술을 의미한다. 그림 2는 자연어 처리 과정에 대한 단계 별 흐름을 나타낸다.

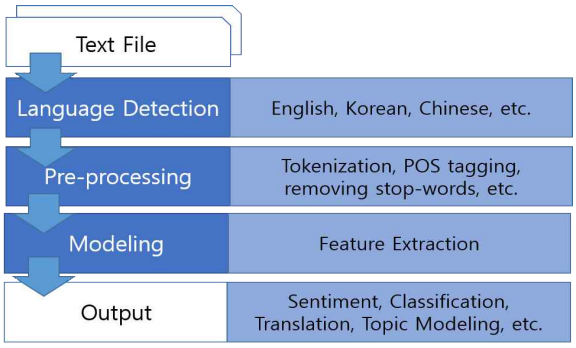

그림 2. 자연어 처리의 흐름
Fig. 2. Flow of Natural Language Processing

2-2 토픽 모델(Topic Model)

토픽 모델(Topic Model)은 광범위한 문서 집합에서 잠재적인 의미 구조를 발견하기 위한 통계적 알고리즘이다[13]. 특정 주제에 대한 문서에서는 그 주제에 관한 단어가 다른 단어에 비해 더 자주 등장할 것이므로, 함께 자주 등장하는 유사한 의미를 지니는 단어들을 하나의 주제로 묶을 수 있다. 즉 토픽 모델은 문서 내의 단어 빈도수를 계산을 통해 문서의 주제를 추출함으로, 유사 문서의 군집을 파악할 수 있도록 하는 소프트 클러스터링(Soft Clustering) 알고리즘이다. 토픽 모델은 다음의 생성모델에 대한 가정을 기반으로 한다.

1) 각 문서는 여러 개의 주제로 이루어져있다고 가정: 즉 각 문서는 주제 분포를 가지며, 문서를 구성하는 단어들은 주제 분포에 따라 주제를 배정 받는다.

2) 문서에 등장하는 각 단어는 주제로부터 생성한다고 가정: 즉 각 주제는 단어 분포를 가지며, 배정 받은 주제의 단어 분포에 따라 단어가 결정된다.

이와 같은 가정에 따라 데이터로부터 주제 분포와 단어 분포를 찾는 방법론이 '확률론적 생성과정(Probabilistic Generative Process)'이다. 그림 3는 해당 방법론을 시각화한 그림이다.

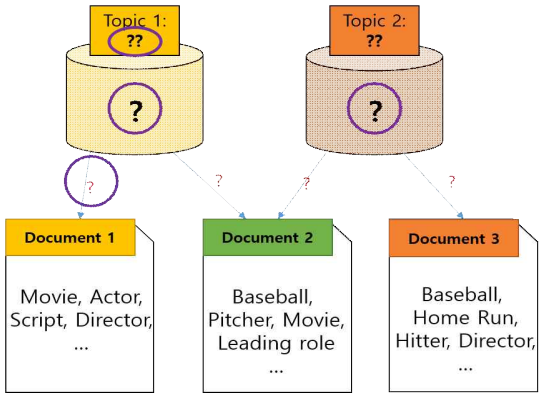

그림 3. 확률론적 생성과정
Fig. 3. Probabilistic Generative Process

그림 3에 따르면 문서 1, 2, 3은 각각의 주제를 가지고 있으며, 단일 주제 혹은 복합 주제로 구성되어 있다. 각 단일 주제는 주제를 표현하는 단어들로 구성 되어 있으며, 복합 주제는 단일 주제의 단어들을 복합적으로 작성자가 설정하는 비율에 따라 사용함으로 표현된다.

그러나 실제 텍스트 분석 시, 독자 및 분석가는 그림 3과 같이 완성된 문서만 관찰할 수 있다. 이는 완성된 문서만으로는 문서 별 주제 분포, 주제 별 단어의 분포, 주제-단어 할당에 대한 정보를 파악할 수 없음을 의미하며, 그림 4와 같이 통계적 추론(Statistical Inference) 계산을 통해 해당 정보들을 잠재 변수로써 찾는 것이 문서의 주제를 추론하는 알고리즘의 목표이다.

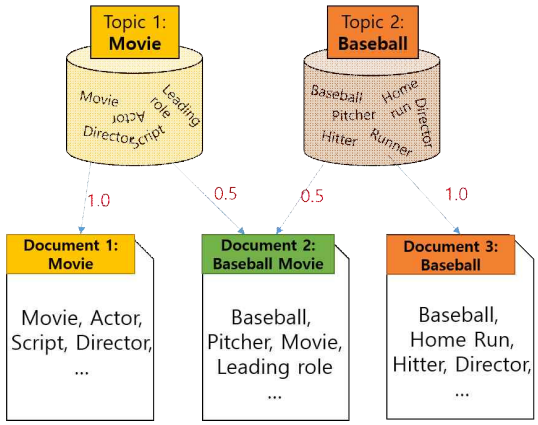

그림 4. 통계적 추론
Fig. 4. Statistical Inference

이와 같은 생성 모델을 기반으로 문서 집합의 주제를 분석하기 위한 추론 알고리즘은 다양한 방법이 존재하는데, 대표적인 추론 기법은 EM(Expectation-Maximization)을 이용한 변분 추론(Variational Inference) [14]과 깁스 샘플링(Gibbs Sampling) [15]이 있다.

2-3 LDA

LDA(Latent Dirichlet Allocation)는 Blei et al. [4]에 의해 제안된 연구로써, 디리클레 분포(Dirichlet Distribution)을 기반으로 텍스트 문서의 주제를 파악한다. 즉, 이는 각 문서 내의 단어들이 어떤 특정 주제에 포함될 확률을 계산함을 의미한다. 소셜 마이닝의 SNS 트렌드 분석 연구 중 다수의 연구는 토픽 모델 기법 중에서도 효과적인 LDA를 기반으로 이뤄졌다 [8-12].

그러나 트위터의 트윗을 비롯한 SNS 텍스트 데이터는 140자~200자 내외의 짧은 텍스트로 구성되어 있다. 즉 이와 같은 단문(Short Texts)은 관사, 전치사, 대명사와 같이 특정 주제와 상관없이 공통적으로 사용되는 고빈도 단어인 불용어(Stop-words)를 제외할 시, 특정 주제의 의미를 내포한 단어가 희소하다는 특징을 가진다. 주제를 추론하고자 하는 문서로써 장문에서의 높은 정확도를 보이는 LDA는 희소성을 지닌 단문을 대상으로 할 때, 낮은 정확도를 나타내는 한계점을 가진다 [16-18].

2-4 BTM

Yan *et al.* [16]에 의하여 제안된 BTM(Biterm Topic Model)은 LDA의 한계를 개선하는 연구이다. BTM은 특정 주제의 의미를 내포한 단어가 한번 이상 재출현 하는 경우가 드문 단문을 대상으로, biterm이라는 두 단어의 조합을 통해 주제를 추론하는 방안을 제시한다. biterm이란, 단어의 순서를 고려하지 않고 단문 내에서 발생한 두 단어의 조합을 의미한다. 즉 BTM은 biterm을 통하여 문서 별이 아닌, 전체 코퍼스(Corpus)를 대

상으로 주제와 단어의 분포를 추정한다. 예를 들어, '영화, 배우, 야구, 주연, 홈런'이라는 5개의 단어로 구성된 문서가 있다고 가정 했을 시, 각 단어를 $\{w_0, w_1, w_2, w_3, w_4\}$로 표현하면, 표 1과 같은 biterm 조합이 가능하다.

표 1. Biterm 조합의 예시
Table. 1. Example of Biterm

Transaction ID	Biterms	Transaction ID	Biterms
1	w_0, w_1	6	w_0, w_3
2	w_1, w_2	7	w_0, w_4
3	w_2, w_3	8	w_1, w_3
4	w_3, w_4	9	w_1, w_4
5	w_0, w_2	10	w_2, w_4

SNS 단문 데이터를 대상으로 한 BTM 알고리즘 개선 연구는 Lim et al. [18]과 Chen et al. [19]의 연구가 있다. 해당 연구들은 SNS의 특성에 따라 트윗 작성자인 회원(User)의 정보 및 해시 태그(Hash tag), 회원 네트워크와 같은 데이터를 이용함으로써 데이터 희소성(Data Sparsity)을 극복하여 정확도를 개선하고자 하였다.

그러나 이와 같은 SNS 사용자 계정에 따른 계산은 사용자별로 축적된 데이터의 양이 다르기 때문에, 사용자 코퍼스에 대한 참고 데이터 확보의 가능 여부가 불투명하다. 따라서 사용자별 일정한 추론의 질 보장할 수 없다는 한계점을 가진다.

BTM은 깁스 샘플링(Gibbs Sampling)을 통해 주제를 추론하며, BTM에 적용되는 깁스 샘플링의 공식은 식 (1)과 같다.

$$P(z|z_{-b}, B, \alpha, \beta) \propto (n_z + \alpha)\frac{(n_{w_i|z} + \beta)(n_{w_j|z} + \beta)}{(\sum_{w=1}^{W} n_{w|z} + D\beta)^2} \quad (1)$$

식 (1)은 Biterm이 특정 주제 z에 할당될 확률을 의미한다. Biterm은 문서 집단에 출현하는 각각의 i번째 단어 w_i 와 j번째 단어 w_j로 표현 된다. n은 개수를 의미하며, n_z은 주제 개수, $n_{w_i|z}$은 주제 z에 할당된 w_i의 개수, D는 문서수를 의미한다. 수식의 α와 β는 전체 문서 집단 내에서 정해지는 변수로써 초모수(Hyper-parameter), 즉 매개변수이다. α는 주제 분포를 추정하기 위한 매개변수로 문서 내에서 특정 주제가 할당될 사전 확률(prior probability)이다. β는 주제-단어 분포를 추정하기 위한 매개변수로, 각 단어가 특정 주제에 할당될 사전 확률이다.

앞선 BTM의 깁스 샘플링에 대한 식 (1)에서 가장 큰 영향력을 나타내는 부분은 단어 별 빈도수에 따라 변화되는 식 (1)의 분자 부분인 다음의 식 (2)이다.

$$(n_{w_i|z} + \beta)(n_{w_j|z} + \beta) \quad (2)$$

식 (2)로 인해 biterm의 두 단어 중 하나의 빈도가 높고 다른 하나의 빈도가 낮을 때, 빈도가 높은 단어의 영향을 따르기 때문에 각 주제와의 연관성을 고려한 가중치의 계산이 불가능하다. 다음의 그림 4는 주제 수 k=2일 때, '야구, 감독, 요리'라는 단어로 조합된 biterm의 주제를 추론하기 위한 상기 식 (2)의 계산 예시를 보여준다.

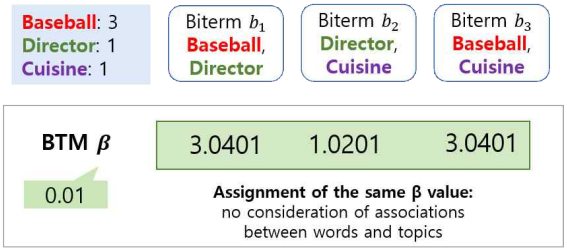

그림 5. BTM의 주제-단어 분포 계산 방법 예시
Fig. 5. Calculation of Topic-word Distribution for BTM

그림 5의 예시에 따르면, b_1과 b_3은 빈도수가 높은 단어 '야구'의 영향을 받아 동일한 값을 가진다. 고빈도 단어 외에는 단어의 등장 빈도수가 낮은 단문에서는 예시와 같은 현상이 빈발하게 발생한다. 따라서 본 논문에서 제안하는 알고리즘은 이와 같은 한계점을 개선하기 위하여 워드 임베딩(Word Embedding) 기법인 Word2Vec [20]을 이용한 개선 방안을 제안한다.

단어 임베딩(Word Embedding) 기법이란 자연어를 실수 차원의 벡터 공간으로 대응하여 단어 간의 의미를 표현하는 방법이다. Word2Vec은 단어 임베딩 기술의 높은 복잡도를 효율적으로 낮춘 계산 방법을 제안한 예측 모델이다. Word2Vec은 단어의 주변을 보면 그 단어를 안다'는 언어학자 J. R. Firth의 배분 가설(Distributional Hypothesis) [21]를 기반으로 한다. 즉 주변 단어(window)를 통해 타겟 단어(target word)를 파악하는 방법이다. 입력인 주변 단어가 유사할 경우, 출력인 단어의 벡터 또한 유사하게 된다. 벡터가 유사함은 벡터 간의 거리가 짧음을 의미한다.

Ⅲ. 제안 연구

본 절에서는 단문에서의 주제 추론 정확도를 개선한 WV-BTM 알고리즘에 대하여 서술하고자 한다.

3-1 WV-BTM

WV-BTM은 Yan *et al.* [16]의 BTM을 기반으로, Word2Vec을 이용하여 BTM의 한계점을 개선함으로써 주제 추론의 정확도를 향상시킨다. 제안 방법은 다음과 같으며, 그림 6는 제안 방법에 대한 예시이다.

(1) Word2Vec의 단어 임베딩 기법을 통해 전체 문서의 단어를 벡터화 한다. 벡터 공간에 투영 된 Biterm의 두 단어 간의 거리를 코사인 유사도(cosine similarity)를 통해 계산한다.

(2) 토픽 모델 계산 시, 코사인 유사도를 각 Biterm 별 β값으로써 사용한다. 이를 통해 두 단어 간의 의미적 연관성을 반영한 주제 추론이 가능하다.

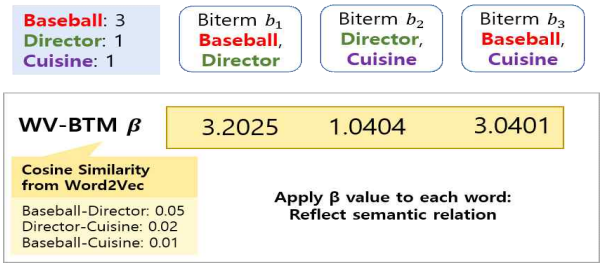

그림 6. WV-BTM의 주제-단어 분포 계산 방법 예시
Fig. 6. Calculation of Topic-word Distribution for WV-BTM

코사인 유사도는 다음의 가정 하에 표 2의 의사코드의 조건부를 기반으로 적용된다.

(1) Biterm의 유사도가 음수로 낮은 연관성을 가졌을 시, 동일 문서상에서 두 단어가 동시에 발생할 수 있는 확률은 0이 아님

(2) Biterm의 유사도는 동일 주제에 속할 확률과 비례함

표 2는 WV-BTM의 깁스 샘플링 의사코드이다. 입력값은 주제 수 K, 하이퍼 파라미터 α, β, 그리고 Biterm의 집합인 B 다. 출력값은 다항변수인 ϕ와 Θ이다. ϕ는 단어 w의 주제 할당 확률, Θ는 해당 문서에서 각 주제의 가중치를 의미한다. 출력값에 대한 계산은 깁스 샘플링의 식을 따라 계산된다.

표 2. WV-BTM 깁스 샘플링 의사코드
Table. 2. Pseudo Code of Gibbs Sampling for WV-BTM

Algorithm: Gibbs sampling algorithm for WV-BTM

Input: the number of topics K, hyperparameters α, β, biterm set B

Output: multinomial parameter ϕ and Θ

(1) initialize word embeddings of Document using Word2Vec

(2) Calculate cosine similarity of all the biterms

(3) initialize topic assignments randomly for all the biterms

(4) for iteration = 1 to iteration = N do

for $b \in B$ do

if $\beta_{w_{i,j}} < \beta$, $\beta_{w_{i,j}} = \beta$;

else if $1 > \beta_{w_{i,j}} > \beta$, $\beta_{w_{i,j}} = \beta_{w_{i,j}}$;

draw z_b from $P(z|z_{-b_{i,j}}, B, \alpha, \beta_{w_{i,j}})$

update n_z, $n_{w_i|z}$, and $n_{w_j|z}$

(5) compute the parameters ϕ and Θ

Ⅳ. 실험 데이터

본 절에서는 제안 알고리즘인 WV-BTM의 주제 추론 성능을 검증하는 실험용 수집 데이터에 대해 서술하고자 한다.

4-1 데이터 수집

본 논문의 연구 대상은 SNS 단문 데이터이다. 대표적인 SNS 단문 데이터인 트위터의 트윗(tweet) 데이터를 실험 데이터로 선정하여, 2017년 12월 14일부터 2017년 12월 19일까지 총 5일 동안의 데이터를 수집한다.

해당 데이터는 표 3의 주제 별 키워드를 기준으로 수집 및 분류 된다. 데이터는 5개의 주제 별 6개의 키워드로 구성되며, 총 45,000건의 데이터를 수집하고, 이를 주제 별로 분류한다. 각 주제 별 키워드가 한 단어 이상 포함 될 시 해당 주제로 분류하며, 이외의 선정 주제의 키워드를 포함할 경우 제외한다.

또한 트윗은 복합 주제일 수 있으나 제한된 길이로 인해 단일 주제만을 포함하는 경우가 많다. 해당 실험을 위해 한국어를 사용하는 한국인 계정의 단일 주제만을 포함한 데이터를 대상으로 제한하며, 표 3의 주제 별 키워드에 대하여 동일 의미를 가진 유사 단어 (ex. 인공지능, AI)를 포함한다. 데이터 수집 및 전처리를 위해 Jupyter notebook 기반 Python 3.5를 이용하여 웹 크롤러 및 전처리 프로그램을 구현한다.

4-2 데이터 전처리

수집한 데이터의 전처리는 다음의 4단계로 구성된 과정을 따라 수행된다.

1) 정규표현식을 통해 한글 및 영문을 별도로 감지하여 추출한 후, 해당 언어별 전처리를 수행함

2) 한글의 경우, KoNLPy 한글 형태소분석 패키지[22] 중 Komoran 태거(Tagger)를 이용하여 명사 및 동사를 판단 후 선별하여 전처리를 수행함. 동사의 경우 "어근+VV"의 형태로 추출, 명사 어미의 형태를 지닌 동일 의미 단어들 간의 구별을 방지함.

3) 영문의 경우, NLTK 자연어 처리 패키지를 이용함. 해당 패키지를 이용하여 웹사이트 링크 및 이모티콘 등의 특수문자를 제외하는 불용어 제거의 전처리를 수행함.

4) 언어별 전처리 수행 후, 동일 문서 내의 순서로 재결합하여 저장함.

Ⅴ. 실 험

본 절에서는 실험 데이터를 통해 WV-BTM의 성능을 검증한다. 성능 비교 알고리즘은 LDA, BTM, WV-BTM이며, 해당 알고리즘은 IntelliJ 기반 개발 환경에서 Java 언어를 통해 구현한다. 실험은 Intel i7-6700HQ-2.60GHz Core 16GB RAM Windows 10 기반 PC를 사용하여 수행한다.

알고리즘의 예측 주제와 수집 시 분류한 주제에 대한 정확도를 비교한다. 분류에 대한 정확도의 공식은 식 (3)과 같다.

$$\frac{\sum_{k=0}^{K} |\{relevant\ Doc\ of\ Topic_{k}\} \cap \{retrieved\ Doc\ of\ Topic_{k}\}|}{|No.\ of\ Total\ Documents|} \quad (3)$$

그림 7은 주제 수 k=2, 3, 5, 주제 추론 반복 수 i=500 일 때의 비교 알고리즘 별 정확도 비교 실험에 대한 결과를 나타낸다. 그림 7에 따르면, LDA는 문서 별 주제 추론에 대한 분류 정확도가 가장 낮게 나타나며, 제안 알고리즘인 WV-BTM은 BTM보다 높은 정확도를 보이는 것을 알 수 있다.

그림 8은 주제 수 k=5, 주제 추론 반복 수 i=500 일 때의 주제별 확률분포 결과를 누적 그래프로 나타낸다. 그림 8에 따르면, WV-BTM은 BTM보다 전체 주제에 대한 추론 정확도 분포가 편향되지 않고 고르게 나타나는 것을 알 수 있다.

표 3. 수집 트윗 데이터 주제-키워드
Table. 3. Topic-Keywords of Tweet Data Collection

Topic	6 key-words
Politics (정치)	Politics(정치), President(대통령), National Assembly(국회), Constitution(헌법), Trial(재판), National Security(국가안보)
Literature (문학)	Literature (문학), Classical Literature(고전) Author(작가), Reader(독자), Novel(소설), Poetry(시)
IT (정보통신 기술)	Information Technology(IT, 정보통신기술), Artificial Intelligent(AI, 인공지능), Deep Learning(딥 러닝), Robot(로봇), Fourth Industrial Revolution(4차 산업 혁명), Virtual Reality (VR, 가상현실)
Sports (스포츠)	Sports(스포츠), Olympic(올림픽), Baseball(야구), Soccer(축구), Basketball(농구), Athlete(운동선수)
Entertainment (연예)	Korean wave(한류), K-pop(케이 팝, 가요), Idol(아이돌), Girl Group(걸 그룹), Singer(가수),Concert(콘서트)

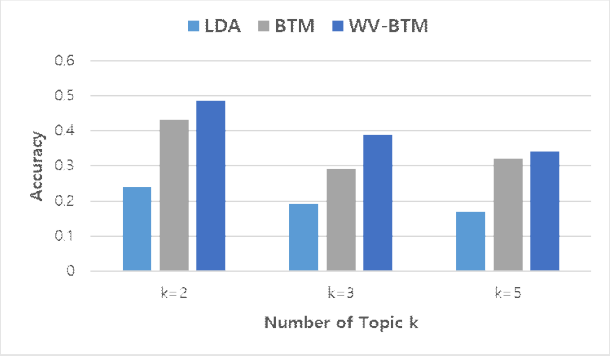

그림 7. 정확도 비교 (k=2,3,5, i=500)
Fig. 7. Comparison of Accuracy (k=2,3,5, i=500)

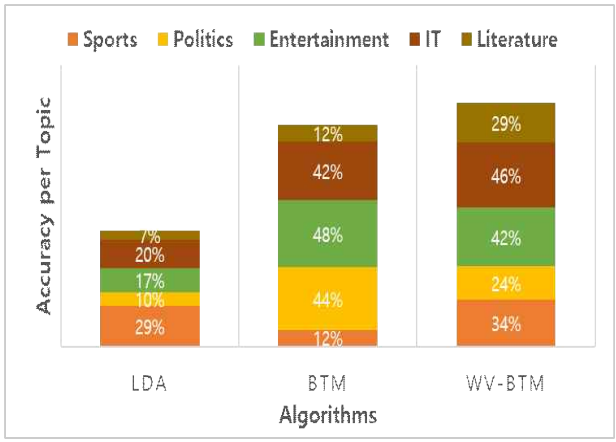

그림 8. 주제 별 정확도 비교 (k=5, i=500)
Fig. 8. Comparison of Accuracy per Topics (k=5, i=500)

Ⅴ. 결 론

본 논문에서는 단문 데이터를 대상으로 한 토픽 모델 알고리즘의 주제 추론 정확도를 향상 시킨 WV-BTM 알고리즘의 SNS 단문에서의 정확도에 대한 실험을 진행하였다. 본 실험은 모델의 수정을 통하여 SNS에 대한 추가 데이터를 사용하지 않고 입력 데이터만으로 정확도를 향상하였다. 따라서 본 연구와 관련 연구와의 결합이 가능한 확장성을 지니므로, 주제 추론의 정확도 향상에 대한 긍정적인 결과를 나타낸다.

이와 같은 토픽 모델의 정확도 향상은 SNS 사용자의 트렌드를 보다 정확하게 파악함으로써 콘텐츠 마케팅 및 광고 브랜딩과 같은 상업적 영역에서 의미 있는 결과를 가져오는 기대효과가 있다. 또한 학문적 영역에서도 SNS를 통한 연구 정보 공유가 늘어가는 가운데, 연구자들의 SNS 데이터를 통해 분야 별 연구 트렌드를 파악함으로써 연구 현황 파악 및 연구의 질을 높이는 기대효과가 있다.

그러나 본 연구는 다음과 같은 한계점을 가진다. 사용자의 의견을 나타내는 단문 자체에만 국한 되어 있으며, 한글과 영문만을 대상으로 하여, 대상 언어 이외의 다국적 언어를 포함한 문서를 제외하였다. 또한 형태소 중에서도 명사와 동사만을 추출하였다는 한계를 가진다. 따라서 SNS 단문 중 해시태그 및 링크를 분별하여 가중치를 부여하는 방안 및 전체 추론 계산에 있어 β값의 영향력에만 국한되지 않고 보다 높은 영향력을 발휘하는 방안을 고려한 연구를 진행한다면, 주제 추론의 품질이 보다 개선될 것으로 기대 한다.

사사문구

This Research was supported by Sookmyung Women's University Research Grants(과제번호 1-1503-0078).

참고문헌

[1] Korea National Statistical Office. Analysis of Consumer Propensity using SNS Data. 2015.

[2] H. Shim and K. Lim. "Research on the Effect of Different motivations on the Participation in SNSs," *Journal of Digital Contents Society*, Vol. 12, No. 3, pp. 383-390, 2011.

[3] Statista. Number of Social Media Users. [Internet] Available:https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[4] Intel. What Happens in an Internet Minute? 2013. Available: https://newsroom.intel.com/press-kits/big-data-intelligence-begins-with-intel/

[5] A. Oulasvirta, E. Lehtonen, E. Kurvinen,, and M. Raento, "Making the ordinary visible in microblogs," *Personal and ubiquitous computing*, Vol. 14, No. 3, pp. 237-249, 2010.

[6] S. H. Na, J. I. Kim, E. J. Lee, P. K. Kim, "A Study on the Short Text Categorization using SNS Feature Informations," *The Journal of Korean Institute of Information Technology*, Vol. 14, No. 6, pp. 159-165, June 2016.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, pp.993-1022, January 2003.

[8] M. C. Yang and H. C. Rim, "Identifying interesting Twitter contents using topical analysis," *Expert Systems with Applications,* Vol. 41, No. 9, pp.4330-4336, July 2014.

[9] C. Xing, Y. Wang, J. Liu, Y. Huang, and W. Y. Ma, "Hashtag-Based Sub-Event Discovery Using Mutually Generative LDA in Twitter," *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence,* pp. 2666-2672, February 2016.

[10] M. J. Paul and M. Dredze, "Discovering health topics in social media using topic models," *PloS one*, Vol. 9, No. 8, 2014.

[11] D. Y. Kim, D. H. Kim, S. W. Kim, M. H. Jo, and E. J. Hwang, "SNS-based issue detection and related news summarization scheme," *In Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication,* No. 114, January 2014.

[12] J. Ito, J. Song, H. Toda, Y. Koike, and S. Oyama, "Assessment of tweet credibility with LDA features," *In Proceedings of the 24th International Conference on World Wide Web*, pp. 953-958, May 2015.

[13] Wikipedia, Topic Model, [Internet] Available: https://ko.wikipedia.org/wiki/%ED%86%A0%ED%94%BD_%EB%AA%A8%EB%8D%B8

[14] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian analysis,* Vol. 1, No. 1, pp. 121-143, 2006.

[15] J. S. Liu, "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem," *Journal of the American Statistical Association,* Vol. 89, No. 427, pp. 958–966, September 1994.

[16] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," *In Proceedings of the 22nd international conference on World Wide Web,* pp. 1445-1456, 2013.

[17] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," *In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval,* pp. 889-892, July 2013.

[18] K. W. Lim, C. Chen, and W. Buntine, "Twitter-network topic model: A full Bayesian treatment for social network and text modeling," *NIPS 2013 Topic Models: Computation, Application, and Evaluation*, arXiv preprint arXiv:1609.06791, 2016.

[19] W. Chen, J. Wang, Y. Zhang, H. Yan, and X. Li, "User Based Aggregation for Biterm Topic Model," *In ACL*, Vol. 2, pp. 489-494, 2015.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[21] J. R. Firth, "A synopsis of linguistic theory," *Studies in linguistic analysis*, pp. 1930-1955, 1957.

[22] S. Bird, "NLTK: the natural language toolkit," In Proceedings of the *COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics*, pp. 69-72, July 2006.

송 애 린 (Ae-Rin Song)

2016년 : 서울신학대학교 신학/영어학과 (문학사)
2016년~현 재 : 숙명여자대학교 IT공학과 (공학석사 과정)

※관심분야 : 데이터 마이닝(Data Mining), 자연어 처리(Natural Language Processing), 스마트 교육(Smart Education), 추천 시스템(Recommendation System), IT 융합(IT Convergence)

박 영 호 (Young-Ho Park)

1990년 : 동국대학교 컴퓨터공학과 (공학사)
1992년 : 동국대학교 컴퓨터공학과 (공학석사)
2005년 : 한국과학기술원 전산학과 (공학박사)

1993년~1999년: 한국전자통신연구원 교환전송연구단 선임연구원
1999년~2006년: COSMO 책임연구원
2002년~2005년: 한국기술대학교 겸임교수
2005년~2006년: 동국대학교 겸임교수
2005년~2006년: 한국전산기술원 AITrc 연구원
2006년~현 재: 숙명여자대학교 공과대학 IT공학과 교수
※관심분야 : 데이터 분석 (Data Analytics), 정보검색(Information Retrieval), 감성공학(Emotional Computing), 기계학습(Machine Learning), 데이터베이스 관리 시스템(DBMS), IT 융합(IT Convergence), XML