디지털산업정보학회 논문지 제14권 제4호-2018년 12월
http://dx.doi.org/10.17662/ksdim.2018.14.4.035

딥러닝을 통한 의미·주제 연관성 기반의 소셜 토픽 추출 시스템 개발*

조 은 숙**·민 소 연***·김 세 훈****·김 봉 길*****

Development of Extracting System for Meaning · Subject Related Social Topic using Deep Learning

Cho Eunsook · Min Soyeon · Kim Sehoon · Kim Bonggil

〈Abstract〉

Users are sharing many of contents such as text, image, video, and so on in SNS. There are various information as like as personal interesting, opinion, and relationship in social media contents. Therefore, many of recommendation systems or search systems are being developed through analysis of social media contents. In order to extract subject-related topics of social context being collected from social media channels in developing those system, it is necessary to develop ontologies for semantic analysis. However, it is difficult to develop formal ontology because social media contents have the characteristics of non-formal data. Therefore, we develop a social topic system based on semantic and subject correlation. First of all, an extracting system of social topic based on semantic relationship analyzes semantic correlation and then extracts topics expressing semantic information of corresponding social context. Because the possibility of developing formal ontology expressing fully semantic information of various areas is limited, we develop a self-extensible architecture of ontology for semantic correlation. And then, a classifier of social contents and feed back classifies equivalent subject's social contents and feedbacks for extracting social topics according semantic correlation. The result of analyzing social contents and feedbacks extracts subject keyword, and index by measuring the degree of association based on social topic's semantic correlation. Deep Learning is applied into the process of indexing for improving accuracy and performance of mapping analysis of subject's extracting and semantic correlation. We expect that proposed system provides customized contents for users as well as optimized searching results because of analyzing semantic and subject correlation.

Key Words : Deep Learning, Social Media, Extracting System of Social Topic, Semantic and Subject Correlation Analysis

Ⅰ. 서론

최근 인공지능 분야에서 많은 주목을 받는 딥러닝(Deep Learning)은 인공신경망에 기반을 둔 기계학습 기술의 한 종류로 최근 몇 년간 비약적인 발전을 이루고 있다[1]. 딥러닝이란, 다층 구조로 설계하여 깊어진 인공신경망이 학습이 잘 이루어지지 않는 전통적인 문제를, 학습을 위한 데이터들을 비지도 학습(Unsupervised Learning)을 통해 전처리하면 신경망이 깊어져도 학습이 잘된다는 것에서 출발한 것으로, 깊은 구조로 이루어진 인공신경망의 학습기법을 말한다. 따라서 딥러닝은 인터넷에 의해 축적된 방대한 양의 데이터에서 오는 빅데이터와 이를 처리하기 위한 컴퓨팅 능력 향상의 두 가지 요소가 없으면 이루어질 수 없는 발견이라 할 수 있다. 이는 실제 인간의 뇌가 뉴런들 간의 연결이 매우 깊은(Deep) 구조를 가지고 있다는 점에서 보다 진보된 인공지능 기술이라 할 수 있다[2].

딥러닝은 인공신경망 학습에서 발생하던 기존의 제약사항들을 부분적으로 해소함으로써 이미지, 음성, 자연어처리 등의 분야에서 뛰어난 성과를 거두며 지속적인 관심을 받고 있다. 이미 딥러닝은 인공지능의 다양한 영역에 적용되고 있으며 그 분야를 막론하고 기존의 성과들을 갱신하고 있다[3-4]. 한동안 세간의 주목을 받은 알파고(AlphaGo)는 기존의 게임 알고리즘에 딥러닝 기법을 적용하여 성능을 극대화하였으며, 페이스북은 딥러닝을 사용자 얼굴인식에 적용하여 정확도를 크게 높였다.

따라서 본 논문에서는 이러한 딥러닝 기술을 비정형 데이터 특성을 지니는 소셜 미디어 콘텐츠에 적용하고자 한다. 이를 통해 의미있는 소셜 토픽들을 추출하는 기법과 이를 적용한 시스템을 개발하고자 한다.

본 논문의 구성은 2장에서는 관련 연구로서 소셜 토픽 추출과 관련된 기존 연구들에 대해 제시하고, 3장에서는 딥러닝 기반의 소셜 미디어 콘텐츠 수집 및 추출 기법에 대해서 제시한다. 4장에서는 실험 결과에 대해 제시하고, 5장에서 마지막으로 결론 및 향후 연구과제를 제시한다.

Ⅱ. 소셜토픽 추출 관련 연구

2.1 ㈜페이스북의 소셜 토픽 추출 기법

㈜페이스북의 소셜 토픽 추출 기법은 소셜 네트워킹 시스템에서 컨텐츠 아이템으로부터 추출된 토픽을 기반으로 한 관심사 추론 기술을 적용하는 기법으로서, 사용자의 코멘트 및 페이지 '좋아요'를 활용한 사용자 관심 토픽을 추론하여 사용자 게시물로부터 토픽 자동 추출 및 사용자에 대한 추가 토픽을 식별하는 카테고리 트리를 사용한 일반화 기법이다[5-7]. 이 기법은 추출된 토픽 기초의 관심 사용자 타겟팅 기법이라고 할 수 있다. 그러나 이 기법은 현재 페이스북만의 특화된 소셜 토픽에 한정되어 추출하고 있는 기법이다. 또한 이 기법은 실시간 관심도가 높은 소셜 콘텐츠만 수집 대상이 된다. 이러한 시스템은 특정 소셜 채널에 특화된 소셜 토픽에 대해서는 적합하나, 여러 다양한 소셜 채널로부터의 데이터를 기반으로 범용적인 소셜 토픽 추출 및 분류를 목적으로 하는 경우에 있어서는 부적합하다는 한계점이 있다.

2.2 충북대의 소셜 토픽 추출 기법

충북대 산학협력단에서 특허로 개발된 이 기법은 사용자 영향력 및 시간 변화를 고려한 소셜 네트워크

* 이 논문은 2017년 중소기업청의 재원으로 2017 산학연협력 기술개발사업의 지원을 받아 수행된 연구임(과제번호: C0532698).
** 서일대학교 소프트웨어공학과 교수(교신저자)
*** 서일대학교 정보통신공학과 부교수
**** ㈜메타소프트 기술이사
***** ㈜메타소프트 대표이사

핫 토픽을 결정한 방법으로서, 복수의 소셜 네트워크 콘텐츠에 포함되는 복수의 단어의 시간 슬롯의 변화에 따른 출현 빈도의 변화를 기반으로 단어를 추출하는 기법이다[8-9]. 여기서 추출된 단어를 포함하는 소셜 네트워크 콘텐츠를 업로드한 사용자의 영향력 지수를 분석하게 되고,추출된 단어의 복수의 시간슬롯 각각에서의 출현 빈도를 기반으로 상기 추출된 단어의 복수의 시간 슬롯 각각에서의 핫 토픽 지수 결정하게 된다. 또한 핫 토픽 지수의 복수의 시간 슬롯 각각에서의 변화를 고려하여 상기 추출된 단어의 핫 토픽 지수 변화 비율을 결정하고, 핫 토픽 지수 변화 비율을 기반으로 상기 추출된 단어를 핫 토픽으로 선택 여부를 결정한다. 이 기법은 시간 슬롯 내에서 노출 빈도가 높은 단어에 대한 소셜 콘텐츠만이 대상이 되는 한계점이 있다. 본 논문에서는 보다 추출되는 소셜 토픽의 정확도를 높이기 위해서 의미와 주제 연관성에 따른 분류를 통해 관련된 소셜 토픽을 추출하고자 한다.

2.3 한국전자통신연구원의 소셜 토픽 추출 기법

한국전자통신연구원(ETRI)에서 개발한 기법은 토픽별 오피니언과 소셜 영향력자를 기반으로 토픽을 탐지하고 추적하는 시스템 및 방법으로서, 인터넷 상의 콘텐츠로부터 토픽별 오피니언을 추출한다[10]. 여기서 소셜 구성원을 행태 특성에 기초하여 순위화한 소셜 영향력자 추출 기술, 지식 데이터베이스(토픽-오피니언DB, 토픽-소셜 영향력자 DB) 설계 기술,지식 데이터베이스 기반 대상 토픽에 대한 오피니언을 탐지 기술 등이 적용된다.대상 토픽에 대한 소셜 영향력자를 검색 및 소셜 영향력자에 의해 제공 또는 생성된 콘텐츠를 수집하여 분석하는 토픽 탐지 추적장치를 개발하였으며, 토픽별 오피니언과 소셜 영향력자를 기반으로 토픽을 탐지하고 추적하는 시스템을 구축하였다. 그러나 이 기법은 소셜 영향력자가 생산하는 소셜 콘텐츠만이 대상이 된다는 한계점이 있다. 물론, 소셜 영향력자에 따른 생산 콘텐츠 역시 소셜 토픽으로 분류가 가능하지만, 본 논문에서는 소셜 영향력자는 소셜 토픽을 추출 및 분류하기 위한 하나의 파라미터이며, 기본적으로는 소셜 콘텐츠에 내포된 의미와 주제 연관성에 따른 유사 군집을 분류하여 소셜 토픽을 추출하는 것이 추출된 소셜 토픽의 정확도를 더 높일 수 있다.

2.4 현존 기법의 한계점

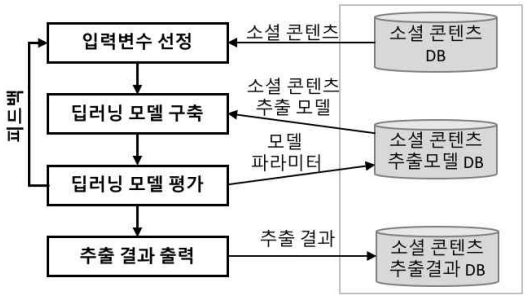
기존 소셜 토픽 추출 시스템은 통계적인 분석 모델 적용에 따른 연관 키워드 집합을 기반으로 키워드를 포함한 소셜 콘텐츠를 추출하고 있는 실정이다[10]. 종래 기술은 단일 키워드가 아닌 다수 키워드의 복합적인 포함에 대한 확률적 계산을 통해 소셜 콘텐츠 간의 유사도를 분석하지만, 일정 시점에서 구축된 통계적 모델은 추가적인 의미 혹은 주제 키워드가 고려되지 못하는 문제점이 있다. 또한, 소셜 콘텐츠는 정형화된 형식이 없는 자연어 수준의 비정형 데이터 형태로 단순히 키워드의 포함 여부를 기준으로 연관성을 판별하는 것은 소셜 토픽 추출 품질이 하락하는 문제점이 존재한다. 본 논문에서는 위와 같은 문제를 극복하기 위해 딥러닝을 적용하여 주기적인 선행학습을 통해 의미・주제 연관성 소셜 토픽 추출 모델의 생성 및 유지보수가 가능한 시스템을 제시한다.

Ⅲ. 딥러닝 기반의 의미・주제 연관성 기반 소셜 토픽 추출 시스템

3.1 소셜 토픽 추출 모델 구축 단계

소셜 콘텐츠 추출 모델 구축은 <그림 1>과 같이 크게 입력변수 선정모듈과, 딥러닝 모델 구축 모듈,

딥러닝 모델 평가 모듈, 추출 결과 출력 모듈, 관련 DB로 구성된다.



<그림 1> 의미/주제 연관성 기반 소셜 콘텐츠 모델 구축 단계

<그림1>에서 데이터베이스는 소셜 콘텐츠 DB와 소셜 콘텐츠 추출모델 DB, 소셜 콘텐츠 추출결과 DB로 구성된다. 입력변수 선정 단계에서 소셜 콘텐츠 DB로부터 각 소셜 콘텐츠 데이터와 작성기록(작성자와 작성일시), 소셜 채널 데이터를 입력 받아서 이 중 소셜 콘텐츠간의 의미 연관성과 주제 연관성을 분석하여 영향력이 높은 순으로 입력 변수를 선정한다.

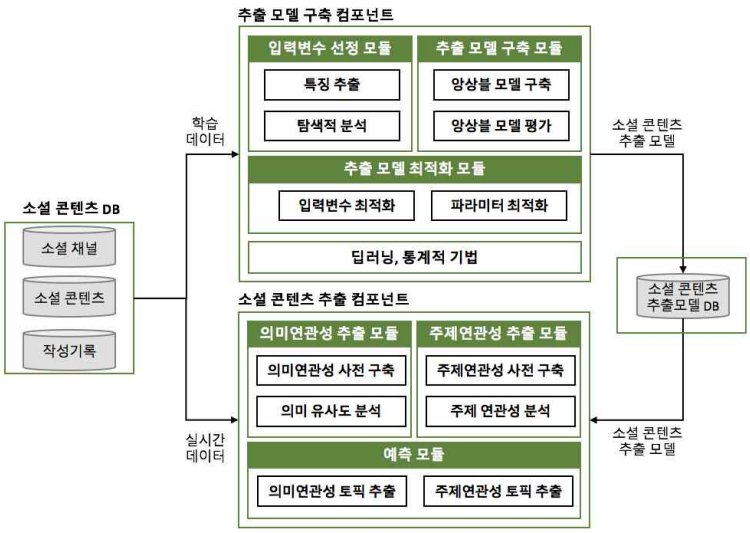
딥러닝 모델 구축 단계에서 학습 레이어의 배치를 통한 소셜 토픽 추출 결과에 대하여 의미 및 주제 연관성의 정확도를 측정하고, 시계열 분석 기법과 같은 통계 분석 기법과 랜덤 포레스트와 같은 머신러닝 기법을 혼합하여 의미 및 주제 연관성에 기반 한 소셜 토픽 추출에 최적화된 앙상블 모델을 구축한다.

구축된 앙상블 모델은 훈련 데이터를 활용하여 학습시키고, 테스트 데이터와 비교하여 추출 결과를 기반으로 모델을 평가한다.

평가 결과는 피드백으로 다시 입력변수 선정 단계로 전달되어, 입력변수 선정과 머신러닝 모델 구축 과정에 반영되어 소셜 토픽 추출모델을 점진적으로 최적화한다.

3.2 소셜 토픽 추출 시스템 설계

의미 및 주제 연관성 소셜 토픽 추출 시스템의 구성은 <그림 2>와 같이 크게 추출 모델 구축 컴포넌트와 소셜 콘텐츠 추출 컴포넌트, 소셜 콘텐츠 DB로 구성된다. 의미·주제 연관성 소셜 토픽 추출 시스템 구조는 아래와 <그림 2>와 같다.



<그림 2> 의미/주제 연관성 소셜 토픽 추출 시스템 설계

추출 모델 구축 컴포넌트와 소셜 콘텐츠 추출 컴포넌트는 소셜 콘텐츠 DB를 기반으로 머신러닝을 통한 학습과 통계적 기법을 통한 모델 평가 과정으로 수행된다[11]. 소셜 콘텐츠 DB는 소셜 채널과 소셜 콘텐츠, 작성기록으로 구성되어 있으며, 의미 연관성과 주제 연관성에 기반 한 소셜 토픽 추출을 위한 기초자료로서 활용된다.

추출 모델 구축 컴포넌트는 딥러닝과 통계적 기법을 활용하여 소셜 토픽 추출 모델을 구축하기 위해 입력변수 선정 모듈과 추출 모델 구축 모듈, 추출 모델 최적화 모듈로 구성된다. 입력변수 선정 모듈은 소셜 콘텐츠 DB로부터 입력되는 학습 데이터를 분석하여 소셜 콘텐츠 간의 의미 및 주제 연관성과 관련성이 높은 특징을 추출하여, 이 중 소셜 토픽 추출을 위한 입력변수를 선정한다. 입력변수 선정은 추출된 특징들의 가능한 조합 중에서 연산속도와 추출 정확도를 기준으로 비교하여 선정된다.

3.3 의미 연관성 식별

<그림 3>은 소셜 콘텐츠 데이터를 기반으로 의미 연관성을 식별하기 위해 군집 모델을 적용하여 의미 연관성 프로파일을 작성하는 절차를 보여준다.

군집 모델은 알고리즘에 따라 학습 데이터의 특징점을 찾고, 유사한 군집화를 수행한다.

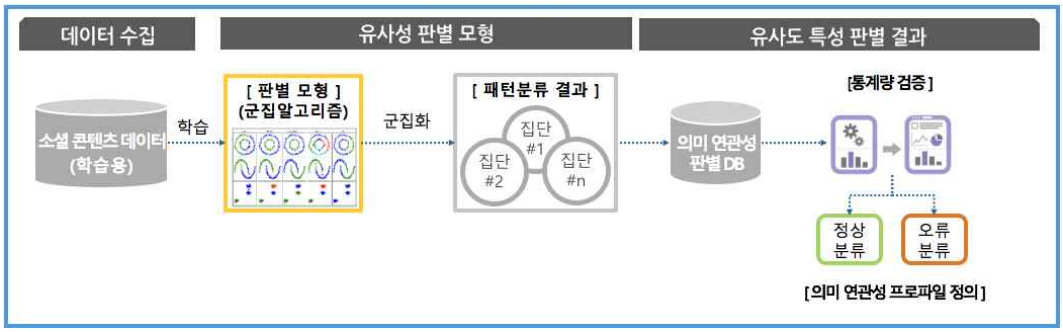
<표 1>은 군집 모델에서 활용되는 대표적인 군집 알고리즘을 보여주고 있다.

<표 1> 대표적인 군집 알고리즘

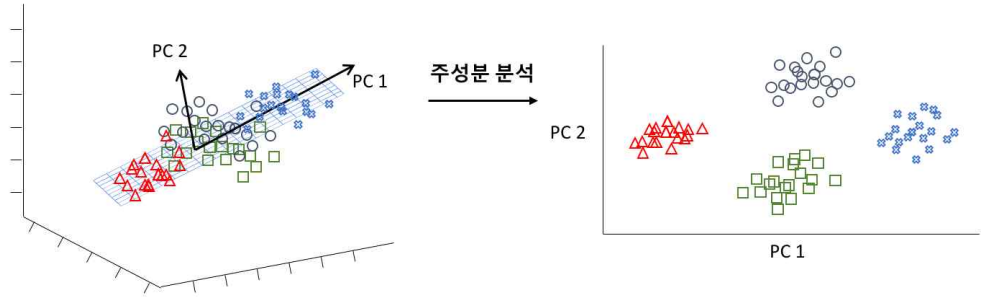
군집 알고리즘	설명
K-means Clustering	전통적인 분류기법으로 대상집단을 거리의 평균값을 기준으로 K개의 군집으로 반복 세분화하는 기법
SOM(Self-Organizing Maps)	인공신경망을 기반으로 훈련집합의 입력패턴을 가중치로 학습하여 군집화하는 기법
EM & Canopy	주어진 초기값으로 가능성이 최대인 것부터 반복 과정을 통해 파라미터 값을 갱신해 군집화하는 기법

추출 모델 구축 모듈은 앙상블 모델 구축과 앙상블 모델 평가로 구성된다. 앙상블 모델 구축은 이동평균법과 지수평활법 등의 통계적 기법 중 시계열 분석 기법과 랜덤 포레스트(Random Forest), 서포트 벡터 머신(Support Vector Machine) 등의 딥러닝 기법을 혼합한 추출 모델 구축한다[10]. 혼합될 기법은 분석 대상인 데이터와 입력변수의 특징을 기반으로 선정한다.

앙상블 모델 평가는 구축된 예측 모델을 일부 훈련 데이터 셋(Training Data Set)을 이용하여 학습시키고, 나머지 테스트 데이터 셋(Test Data Set)을 이용하



<그림 3> 군집 모델 기반 의미 연관성 판별



<그림 4> 주성분 분석을 통한 차원 감소

여 추출 결과를 평가한다.

평가 결과는 피드백 입력으로 다른 조합의 소셜 콘텐츠 추출 모델 구축 과정에서 반영하며, 추출 모델 최적화 모듈은 입력변수 최적화와 파라미터 최적화로 구성된다.

구축된 앙상블 모델은 연산속도와 정확도 등의 특성을 기준으로 최적화를 진행한다. 예측 모델의 최적화는 입력변수와 모델의 하이퍼파라미터(hyperparameter)를 대상으로 진행한다. 입력 변수 선정은 소셜 콘텐츠 DB의 데이터 중 소셜 콘텐츠 간의 의미 및 주제 연관성에 영향을 크게 미치는 요소를 추출해 사용하기 위해 데이터 차원 축소 방법인 주성분 분석(Principal Component Analysis)을 사용한다[13-14].

주성분 분석 기법은 이미지와 같은 고차원 데이터를 저차원의 데이터로 변환시키는 방법으로 [그림 4]와 같이 데이터 집합을 새로운 좌표축으로 변환한다.

데이터 손실을 최소화하면서 정보를 각각 서로 간

특징선택기법	설명
필터 (Filter)	모델의 성능을 고려하지 않고 특징 선택
	모든 특징을 척도에 따라 순위를 정하고, 가장 높은 순위의 특징들로 선택
	특징 간의 중복을 고려하지 않음
래퍼 (Wrapper)	모델이 최고의 성능을 내는 특징 선택하며, 소요 시간이 오래 걸림
	부분집합의 수가 기하급수적으로 늘어 과적합 위험 발생
	특징 선택을 위한 알고리즘과 선택기준을 결정

<그림 5> 차원 축소를 통한 소셜 콘텐츠 데이터 특징 추출

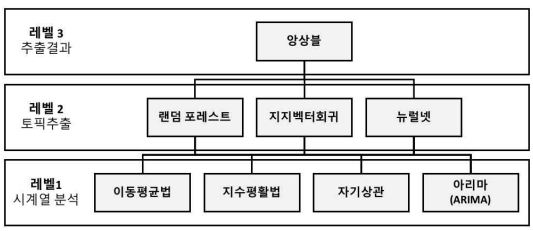
에 독립인 좌표축들로 재구성하여, 최소한의 차원으로 최대한의 설명력 획득 가능하다. 정보의 손실을 최소화하면서 데이터를 대표하는 주성분(Principal Components)을 찾아 변수의 차원(개수)을 줄여 변수에 의한 데이터의 중복(overlap)을 감소시킨다. 차원이 감소된 주성분들은 유전자 알고리즘을 이용하여 가장 식별력이 좋은 특징들을 추출한다.

3.4 소셜 콘텐츠 데이터 특징 추출

<그림 5>는 소셜 콘텐츠 데이터에 포함된 특징의 전체 집합에서 필터와 래퍼를 통한 부분 집합의 생성 및 최적 집합을 선별하는 과정을 보여준다.

차원이 감소된 주성분들은 유전자 알고리즘을 이용하여 가장 식별력이 좋은 특징들을 추출한다. 유전자 알고리즘은 자연세계의 유전과 진화 메커니즘에 기반 한 계산 모델로서, 풀고자하는 문제에 대한 가능한 해들을 정해진 형태의 자료구조로 표한 다음, 이들을 점차적으로 변형함으로써 점점 더 좋은 해들을 생성한다. 각각의 가능한 해를 하나의 유기체 또는 개체(Individual)로 보며 이들의 집합을 개체군(Population)이라 한다.

앙상블 모델은 복수개의 통계적 기법 혹은 딥러닝 기법을 동시에 적용하여 최적의 결과를 획득하기 위한 수단으로 활용한다. [그림 6]과 같이 앙상블 모델은 세 개의 레벨로 구성된다.



<그림 6> 의미·주제 연관성 분석 정확도 향상을 위한 앙상블 모델 구조

레벨 1에서는 시계열 분석으로 이동평균법, 지수평활법, 자기 상관모형, 아리마 등의 통계적 분석을 각각 수행한다. 각 통계 분석 기법을 통해 분석된 결과는 반복적인 분석 과정을 거쳐 수요예측을 위한 기초자료로 활용된다. 각 기법의 결과마다 가중치가 적용되며, 소셜 토픽 추출 결과와 추출 정확도에 따라 가중치는 지속적으로 갱신되며, 가중치의 총합은 1이다. 즉, 시계열 분석에 적용되는 n개의 통계적 분석 기법들을 TA1, TA2, TA3, ..., TAn이라 정의하면, 각각의 가중치는 W1, W2, W3, ..., Wn이고, 가중치의 총합은 $\sum_{i=1}^{n} W_i = 1$이 된다.

레벨 2의 토픽추출로 전달되는 시계열 분석의 결과 값은 $\frac{\sum_{i=1}^{n}(TA_i \times W_i)}{n}$ 이다. 시계열 분석 결과 값에 적용되는 가중치는 레벨 2의 각 토픽추출 기법의 특징에 따라 별도로 정의된다. 레벨 2는 토픽추출을 위한 머신러닝 기법이 적용되며, 레벨 1과 같이 복수개의 머신러닝 기법이 적용된다.

레벨 3은 추출 결과를 통합하는 앙상블 모델이며, 레벨 2로부터 추출 결과를 전달 받아 통합한다. 각 토픽 추출 기법의 결과마다 가중치가 적용되며, 소셜 토픽 추출 결과와 추출 정확도에 따라 가중치는 지속적으로 갱신되며, 가중치의 총합은 1이다. 즉, 수요예측 분석에 적용되는 k개의 머신러닝 기법들을 MA1, MA2, MA3, ..., MAk이라 정의하면, 각각의 가중치는 W1, W2, W3, ..., Wk이고, 가중치의 총합은 $\sum_{i=1}^{k} W_i = 1$이기 때문이다.

레벨 3의 앙상블 모델에서 통합되는 토픽 추출 결과값은 $\frac{\sum_{i=1}^{k}(MA_i \times W_i)}{k}$ 이다.

제시된 시스템은 앙상블 모델 구축을 위해 레벨 1 시계열 분석과 레벨 2 수요예측에 적용되는 통계적

분석 기법과 머신러닝 기법들을 분석 데이터의 특성을 기반으로 선정할 수 있어 적용범위와 모델 자유도가 높다.

Ⅳ. 실험 및 평가

4.1 실험

본 논문에서 제시한 딥러닝 기반을 의미·주제 연관성 소셜 토픽 추출 시스템을 주요 포털 사이트인 네이버, 다음 등에 적용하여 뉴스를 분석한 소셜 토픽 추출 실험 수행 결과는 다음과 같다. 네이버 뉴스 데이터는 총 1,933,519건을 수집하였고, 다음 뉴스 데이터 총 총 8,047,875 건을 수집하였다.

이렇게 수집된 소셜 콘텐츠들을 기반으로 딥러닝을 통한 년도별 신기술에 대한 의미/주제 연관성 기반으로 신기술에 관한 뉴스 데이터 분석을 수행하였다. 처음으로 의미 연관성을 파악하기 위해, 과학뉴스 카테고리에서 기술 키워드 사전을 작성하고 Word2Vec을 이용하여 관련 키워드에 대한 거리 값을 계산하였다. 거리값에 대한 임계값(Threshold Value)은 모델 학습에 따라 선정된다. 각 년도별 기술 키워드에 대한 의미적으로 유사한 관련 키워드가 선정되고, 이를 기준으로 해당 키워드를 포함한 소셜 콘텐츠가 분류된다. 예를 들어, 2016년도 기술 키워드 1위인 '인공지능'의 경우 'AI', '알파고', '왓슨', '인공신경망'의 네 가지 의미 연관 키워드를 포함한다. 두 번째로 분류된 소셜 콘텐츠에 Doc2Vec을 적용한 거리 값을 기반으로 주제 연관성을 분석하였다. 특정 기술 키워드는 중의적 해석이 가능하여, 주제 연관성을 통해 분류 정확도를 높여야한다. 예를 들어, '인공지능' 키워드의 'AI'는 '인공지능' 이외에 '조류독감'의 의미 역시 포함하며, 의미 연관성 분석을 통해 분류된 소셜 콘텐츠에도 '조류독감'에 해당하는 콘텐츠가 존재하였다. 주제 연관성 분석은 이와 같은 분류 오류를 필터링하는 역할을 하며, 분석 결과에 따라 기술 키워드 순위를 재평가한다. [그림 7]은 8년간 매년 선정 된 신기술 상위 10위 변화를 분석한 결과를 보여주고 있다.

<그림 7>은 8년간 매년 선정 된 신기술 상위 10위 변화를 분석한 결과를 보여주고 있다.

- 네이버 뉴스에서 경제, 정치, 과학, 사회 카테고리의 뉴스를 수집
- 수집시간 단축을 위해 본문을 제외하고 뉴스ID, 뉴스 제목, 카테고리, 뉴스출처, 시간, 뉴스URL을 수집
- 수집한 뉴스 URL을 이용해 추후 빠르게 본문 수집 가능

년도	경제 뉴스	정치 뉴스	과학 뉴스	사회 뉴스	합계
2010	18,724	17,595	11,631	16,570	64,520
2011	29,347	28,032	16,716	26,300	100,395
2012	31,327	37,719	19,094	36,412	124,552
2013	35,865	42,914	19,933	46,606	145,318
2014	38,177	41,553	21,985	49,540	151,255
2015	86,874	97,295	38,536	112,158	334,863
2016	123,765	187,624	49,348	130,845	491,582
2017	124,455	211,147	44,242	141,190	521,034
total	488,534	663,879	221,485	559,621	1,933,519

<그림 5> 네이버 뉴스 데이터 목록

- 다음 뉴스에서 사회 카테고리의 뉴스를 수집
- 수집시간 단축을 위해 사회 분야에 한정해 뉴스ID, 뉴스제목, 뉴스본문, 카테고리, 뉴스출처, 시간, 뉴스URL을 수집

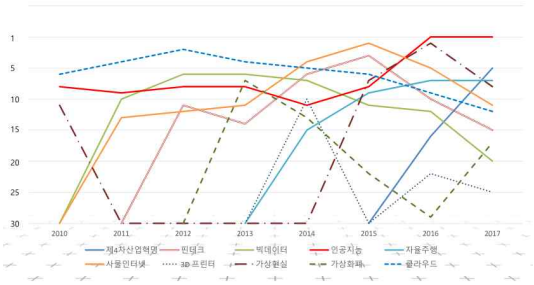
년도	사회 뉴스	합계	비고
2010	1,151,247	1,151,247	수집 완료
2011	1,240,128	1,240,128	
2012	1,377,135	1,377,135	
2013	1,554,705	1,554,705	
2014	1,739,226	1,739,226	
2015	651,376	651,376	시간적인 제약으로 수집을 100% 완료하지 못함
2016	257,327	257,327	
2017	76,731	76,731	
total	8,047,875	8,047,875	

<그림 6> 다음 뉴스 수집 목록

	2010	2011	2012	2013	2014	2015	2016	2017
1	태블릿PC	태블릿PC	우주기술	우주기술	우주기술	우주기술	인공지능	인공지능
2	우주기술	우주기술	보안	웨어러블	웨어러블	사물인터넷	가상현실	증강현실
3	SNS	SNS	클라우드	보안	보안	웨어러블	증강현실	로봇
4	보안	보안	SNS	로봇	로봇	핀테크	우주기술	랜섬웨어
5	전자책	클라우드	로봇	클라우드	사물인터넷	로봇	로봇	우주기술
6	로봇	로봇	나노기술	SNS	클라우드	보안	사물인터넷	4차산업혁명
7	클라우드	전자책	빅데이터	빅데이터	핀테크	클라우드	드론	AI비서
8	유전자기술	나노기술	유전자기술	가상화폐	빅데이터	가상현실	자율주행	자율주행
9	인공지능	유전자기술	인공지능	인공지능	재난관리	인공지능	스마트 디바이스	가상현실
10	나노기술	인공지능	양자기술	유전자기술	스마트기술	자율주행	클라우드	5G

<그림 7> 신기술 상위 10위의 연도별 변화

<그림 8>은 <그림 7>의 8년간 매년 선정 된 신기술 상위 10위 변화를 분석한 결과를 시각화 결과 화면으로 보여주고 있다.



<그림 8> 8년간 신기술 상위 10위 변화 시각화 화면

4.2 평가

본 논문에서 개발한 소셜 토픽 추출 시스템에 대해서 기존의 기법들과의 비교를 에 제시하였다.([표2] 참조)

또한 객관적인 소프트웨어의 기능, 성능, 보안성, 안전성 등에 대해 한국정보통신기술협회(TTA)가 운영하는 소프트웨어시험인증연구소의 확인 및 검증(Verification & Validation)을 통해 신뢰성을 확보하였다.([표3] 참조)

V. 결론 및 향후 연구과제

기존의 소셜 토픽 추출 시스템은 과거의 이력을 기반으로 통계적인 분석 모델을 적용하여 의미 혹은 주제 연관 키워드를 추출하여 키워드 포함 여부에 따른 소셜 토픽 추출 정확도가 낮고, 통계 분석 수식 및 변수 선택을 위해 전문 인력이 필수적으로 참여해야하

<표 2> 기존 연구들과의 비교

연구유형 / 비교항목	페이스북의 추출 기법	충북대의 추출 기법	ETRI의 추출기법	제안한 추출기법
지원소셜미디어채널	단일채널	단일채널	단일채널	다채널
수집대상선정기준	실시간 관심도가 높은 콘텐츠	노출빈도가 높은 단어에 대한 콘텐츠	소셜영향력자가 생산하는 콘텐츠	사용자의 관심주제에 따른 의미연관성이 높은 소셜콘텐츠와 피드백
의미연관성 분석여부	의미연관성 고려하지 않음	의미연관성 고려하지 않음	생산하는 콘텐츠에 대해서만 분석	단어에 내포된 의미 연관성에 따라 서로 다른 단어간의 연결관계 분석

<표 3>주요 성능지표

< 주요 성능지표 개요 >					
주요 성능지표[1]	단위	예상목표	달성 목표	가중치[4](%)	측정기관[5]
1. 딥러닝을 통한 의미 연관성 기반 소셜 토픽 추출 정확도	%	80% 이상	평균 99.4%	20	TTA
2. 딥러닝을 통한 주제 연관성 기반 소셜 콘텐츠/피드백 분류 정확도	%	80% 이상	평균 100%	20	TTA
3. 저장 대상 미디어 데이터 종류 수	종	5종 이상	5종	15	TTA
4. 딥러닝을 통한 연결 가능한 소셜미디어 콘텐츠 수집기 수	개	5개 이상	5종	15	TTA
5. 딥러닝을 통한 단어 연관성 사전 항목	개	15,000개 이상	44,703개	10	TTA
6. 분석시간	초	3초 이하	평균 0.076초	10	TTA
7. 시간당 처리량	TPS	100 이상	평균 25,638개/초	10	TTA

며, 지속적인 모델의 관리가 힘든 문제점이 있었다.

본 논문에서는 딥러닝 기술을 적용하여 학습을 통한 자율적인 의미/주제 연관성 사전의 생성 및 관리가 가능하고, 지속성 높은 소셜 토픽 추출 모델 구축이 가능한 딥러닝 기반 의미/주제 연관성 소셜 토픽 추출 시스템을 개발하였다. 이를 통해 개발된 시스템이 기존 시스템에 비해 소셜 토픽 추출 정확도가 향상되었을 뿐만 아니라 시스템의 성능 또한 향상되는 효과를 가져 오게 되었다. 향후 연구과제로는 본 논문에서 개발한 추출 시스템을 토대로 소셜 토픽 분류 시스템과 연동하여 검색 플랫폼에 통합하여 여러 다양한 분석 시스템에 활용할 방안을 제시하는 것이다.

참고문헌

[1] 김의중, (알고리즘으로 배우는) 인공지능, 머신러닝, 딥러닝 입문, 2016.

[2] 정도희, 인공지능 시대의 비즈니스 전략 : 누가 AI 환경을 지배할 것인가!, 길벗, 2018.

[3] 박영숙, 인공지능 혁명 2030 = Artificial intelligence revolution : 제4차 산업혁명과 정치 혁명의 부상, 더블북, 2016, p.357~359.

[4] 최규석, 박종진, “인공지능시스템 = Artificial intelligence system,” 21세기사, 2008.

[5] 곽기영, “소셜네트워크분석 = Social network analysis,” 2017.

[6] C. Dwyer, “Trust and privacy: A comparison of Facebook and MySpace,” 2007, p.339.

[7] 박상혁, 오승희, 성행남, “소셜 네트워크 서비스 사용 시기에 따른 사용자 이용패턴 연구: 페이스북을 중심으로,” 디지털산업정보학회, 제9권, 제5호, 2013.

[8] 육지애, 정형원, 유수미, “소셜 네트워크 서비스의 활용방안 연구 : 트위터를 중심으로,” 디지털산업정보학회, 제7권, 제7호, 2011.

[9] 노연우, “Reliable Hot Topic Prediction Scheme Considering User Influences in Social Networks,“ 충북대 석사학위논문, 2016.

[10] 박기진, “Spark 프레임워크 기반 비정형 빅데이터 토픽 추출 시스템 설계,“ 한국정보처리학회 논문지, 제5권, 11호, 2016.

[11] Marsland, Stephen, 강전형, “알고리즘 중심의 머신러닝 가이드 : 파이썬코드 기반,” 제이펍, 2016.

[12] 이상연, 이건명, “토픽 모델링을 이용한 댓글 그래프 기반 소셜 마이닝 기법,” 제24권, 6호, 한국지능시스템학회, 2014.

[13] 정진명, 박영호, 김우주, “토픽모델링을 이용한 교육정책 키워드 기반 소셜미디어 분석,” 제 19권, 4호, 2018.

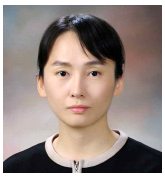
[14] 서창교, “토픽모델링과 소셜 네트워크 분석을 이용한 공급사슬관리 연구 탐색,” 한국연구재단 연구보고서, 2017.

■ 저자소개 ■

조 은 숙
(Cho Eunsook)

2005년 3월~현재
서일대학교 소프트웨어공학과
교수
2000년 2월 숭실대학교
컴퓨터공학과(공학박사)
1996년 2월 숭실대학교
컴퓨터공학과(공학석사)
1993년 2월 동의대학교 전산통계학과(이학사)

관심분야 : 소프트웨어 프레임워크 설계 및 개발, 딥러닝, 빅데이터
E-mail : escho@seoil.ac.kr



민 소 연
(Min Soyeon)

2005년 3월~현재
서일대학교 정보통신공학과
부교수
2003년 2월 숭실대학교 전자공학과(공학박사)
1996년 2월 숭실대학교 전자공학과(공학석사)
1994년 2월 숭실대학교 전자공학과(공학사)

관심분야 : 통신 및 신호처리, 임베디드 시스템, 머신러닝
E-mail : symin@seoil.ac.kr

김 세 훈
(Kim Sehoon)

2018년 4월~현재
(주)메타소프트 기술이사
2000년 2월 명지대학교
정보통신공학과(공학석사)
1998년 2월 관동대학교
정보통신공학과(공학사)

관심분야 : 데이터 분석 및 검색, 빅데이터, 딥러닝
E-mail : shkim@metasoft.co.kr

김 봉 길
(Kim Bonggil)

2014년 8월~현재
(주)메타소프트 대표이사
1997년 2월 인하대학교 금속공학과(공학석사)
1995년 2월 인하대학교 금속공학과(공학사)

관심분야 : 데이터 분석 및 검색, 빅데이터, 딥러닝
E-mail : bgkim@metasoft.co.kr

논문접수일 : 2018년 11월 6일
수 정 일 : 2018년 12월 8일
게재확정일 : 2018년 12월 17일