J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr
Copyright ⓒ 2018 KSII

토픽모델링을 이용한 교육정책 키워드 기반 소셜미디어 분석☆

Social Media Analysis Based on Keyword Related to Educational Policy Using Topic Modeling

정 진 명[1] 박 영 호[2*] 김 우 주[3]
Jin-myeong Chung Young-ho Park Woo-ju Kim

요 약

정보를 전달하고 여론을 형성하는 전통적인 매스미디어의 기능이 ICT 기술의 발전으로 소셜미디어를 통해 정보와 의견을 공유하는 환경으로 급격하게 변해 왔으며, 그 영향력을 더욱 강화시키고 있다. 즉, 일반 대중들이 소셜미디어를 통해 정치·사회·경제 변화에 대한 여론을 생산하고 공유하는 여론의 영향력이 갈수록 커지고 있는 것이 확인되고 있으며, 그 변화는 선거활동과 같은 정치 분야에서 활용되고 있다. 소셜미디어를 활용해서 대중들의 의사를 파악하고, 반영하기 위한 노력은 정치 영역뿐만 아니라 공공 영역에서도 활발하게 이루어지고 있다. 본 논문은 교육분야 정책과정에서 소셜미디어 기반 여론을 활용하기 위한 가능성을 탐색하는 것을 목적으로 한다. 이를 위해 교육정책 중 소프트웨어교육에 관한 키워드를 중심으로 데이터를 수집하고, 문서의 주요 토픽과 토픽별 출현 확률, 토픽 트렌드를 분석하였다. 그 결과 '국내 컴퓨터 교육 시간' 토픽이 전체의 43.99%를 차지하였으며, '프라임 사업 선정' 토픽이 36.81%, '인공지능 프로그램' 토픽이 7.94%의 출현 확률을 나타내어, 대중의 소프트웨어교육 정책에 대한 주요 관심도를 파악할 수 있었다. 또한, 시기별 토픽 추세 및 연관성 있는 토픽간의 트렌드 비교 분석을 통하여 동일한 주제의 정책이라도 교육과정의 시기와 정책의 대상에 따라 유연한 정책수립이 필요하다는 시사점을 도출할 수 있었다.

☞ 주제어 : 교육정책, 소셜미디어 기반 여론, 토픽모델링

ABSTRACT

The traditional mass media function of conveying information and forming public opinion has rapidly changed into an environment in which information and opinions are shared through social media with the development of ICT technology, and such social media further strengthens its influence. In other words, it has been confirmed that the influence of the public opinion through the production and sharing of public opinion on political, social and economic changes is increasing, and this change is already in use on the political campaign. In addition, efforts to grasp and reflect the opinions of the public by utilizing social media are being actively carried out not only in the political area but also in the public area. The purpose of this study is to explore the possibility of using social media based public opinion in educational policy. We collected media data, analyzed the main topic and probability of occurrence of each topic, and topic trends. As a result, we were able to catch the main interest of the public(the 'Domestic Computer Education Time' accounted for 43.99%, and 'Prime Project Selection' topics was 36.81% and 'Artificial Intelligence Program' topics was 7.94%). In addition, we could get a suggestion that flexible policies should be established according to the timing of the curriculum and the subject of the policy even if the category of the policy is same.

☞ keyword : education policy, public opinion based on Social media, topic modeling

1. 서 론

인터넷과 정보통신기기의 발달로 이를 기반으로 하는 다양하고 새로운 미디어들이 등장하고 있다. 인터넷 사용자들의 소통은 블로그, 트위터, 페이스북과 같은 소셜미디어를 통해 정보와 의견을 공유하는 환경으로 변화하고 있다. 이러한 소셜미디어는 적은 비용과 직접적인 소통 수단이라는 강점을 가지고 있어 정책의제의 수립과 집행, 정책평가 등의 과정에서 활용이 가능하다[1]. 따라서 소셜미디어의 서비스 확대는 신문, 방송 등과 같은 전통적인 매스미디어의 역할을 일부 대체하고 있다고 할 수 있다. 즉 의제설정과 정보제공으로서의 전통적인 매스미디어의 역할은 소셜미디어로 전이되고 있으며, 더욱

1 School of Information Security, Kyungpook National Univ., Daegu, 41566, Korea.
2 School of Electronics Engineering, Kyungpook National Univ., Daegu, 41566, Korea.
3 Department of Industrial Engineering, Yonsei Univ., Seoul, 03722, Korea.
* Corresponding author(parkyh@knu.ac.kr)
[Received 27 March 2018, Reviewed 30 March 2018(R2 25 May 2018), Accepted 4 June 2018]

http://dx.doi.org/10.7472/jksii.2018.19.4.53

강화되고 있다[2].

이미 많은 기업들은 상품의 개발 및 판매에 소셜미디어의 의견을 적극적으로 반영하고 있으며, 공공영역에서도 소셜미디어를 활용하기 위한 노력이 시도되고 있다. 또한 미국의 상원, 주지사 선거에서 후보들의 트위터와 같은 소셜미디어 활용 실태에 대한 분석을 하는 등 다양한 연구가 이루어지고 있다[3].

소셜미디어와 정책의제의 관계에 대한 연구도 많이 이루어지고 있다. 남궁근(2008)은 신문, 라디오, TV 등 매스미디어가 정부에 대한 감시자(watchdog) 역할을 담당함으로써 정책의제(policy agenda) 설정 과정에서 중요한 기능을 담당하고 있다고 하였다(홍유정·황주성, 2015 재인용)[4]. 유현재·송지은(2012)은 자살관련 주제의 인터넷신문(대안지)과 전통적신문(연장지)의 미디어 성격별 보도 스타일을 비교 분석하여 인터넷 신문의 기사가 더 선정적이며 흥미위주의 내용으로 작성된다고 판단하였다[5]. 뉴만 외 2인(2014)은 2012년 한 해 동안 29가지의 정치적 이슈에 관한 전통 매스미디어와 소셜미디어의 주목도와 프레이밍에 대해 빅 데이터 분석을 하였다[6]. 이 연구에서 정치적 이슈들에 대한 의제설정이 한 가지 패턴이 아니라 두 미디어 간 복잡하고 역동적인 상호작용을 통해 이루어진다는 것을 밝혀냈다. 김은미・이주현(2011)은 소셜미디어서비스인 트위터를 통해서 뉴스의 유형과 트위터 저자들의 유형에 따른 뉴스 의제와 언론에서 보도되는 뉴스의 빈도수와 트위터에서 회자되는 뉴스의 빈도 경향에 대한 분석을 시도하여, 새로운 뉴스미디어로서의 트위터에 대하여 논하였다.[7]

본 논문에서는 이와 같은 미디어 변화 환경을 반영하여 소셜미디어 상의 여론의 정책과정에 활용 가능성을 탐색하기 위해 교육정책에 대한 소셜미디어의 여론을 분석해보고자 한다. 이를 위하여 소셜미디어서비스들로부터 교육정책 키워드 기반으로 데이터를 수집하고, 주제분석에 많이 사용되고 있는 토픽모델링을 이용하여 문서 내 저자의 특정 토픽에 대한 빈도와 그 관심 주제들의 트렌드 분석을 통해 정책과정에서 소셜미디어의 활용 방안을 탐색하였다.

2. 관련연구

2.1 소셜미디어

소셜미디어의 개념은 매우 포괄적으로 사용된다. 페이스북과 같이 온라인 기반으로 친구 관계를 맺고 콘텐츠를 공유하고, 소통하는 소셜네트워크서비스(SNS:Social Network Service)를 포함하여 블로그와 위키스 등 그 서비스의 종류와 내용은 계속 진화하고 있다.

소셜미디어의 주요 특성은 참여, 개방, 대화, 연결을 들 수 있으며, 특정한 주제에 대하여 같은 공감대를 공유한 사람들이 의견을 교환하고, 지식을 공유하며, 이러한 참여와 피드백 활동을 투표에 활용할 수도 있다[8]. 서론에서 살펴본 바와 같이 소셜미디어가 기존 매스미디어의 일부 기능을 담당하는 것에 대한 연구가 활발히 진행된 것으로 보아, 이미 사회의 정치·사회 환경에 영향을 미치는 여론 형성에 일부 기여를 하게 되었다고 볼 수 있다. 이런 소셜미디어의 기능은 정치적 선거활동에서의 의제설정에 활용되고 있으며, 소셜미디어의 의제설정 기능의 정치적 활용에 관한 연구도 활발히 이루어지고 있다. 최진호·한동섭(2011)은 트위터와 신문/방송 뉴스의 의제를 비교하여 의제설정 가능성을 탐색하고자 정치인 트위터를 신문/방송 뉴스와 비교하여 의제의 속성과 다른 매체와의 관계에 대한 실증 분석을 시도하였다[9]. 구교태(2002)는 2000년 미국대통령 선거기간 중 일어나는 웹사이트를 통한 캠페인이 기존의 뉴스관련 미디어가 생성하는 의제와 공중 의제에 끼치는 영향에 대해서 분석하였다. 홍유정·황주성(2015)은 광주 인화학교사건(도가니)을 사례로 정책의제설정과정에서 소셜미디어와 매스미디어의 역할을 비교하였다. 이러한 연구들을 통해서 소셜미디어가 여론 형성과, 정치적 의제 설정에도 기여하고 있음이 실증되었다. 소셜미디어에서 생산되고 공유되는 일반 대중의 의견들이 공공 의제 설정으로 이어지는 현상에 대한 연구도 이루어지고 있으며, 이는 교육관련 정책의제의 발굴에도 활용 될 수 있는 가능성이 있다고 볼 수 있다.

2.2 토픽모델링

토픽 모델링은 구조화되어 있지 않은 방대한 문서에서 중심 주제를 발견해내는 알고리즘으로 발견한 주제에 따라서 수집한 문서를 구조화할 수 있다[10][11]. 토픽 모델링은 문서의 주제를 도출하기 위해 텍스트를 분석하는 방법으로서 토픽들의 확률적인 집합으로 이루어진다. 즉, 주어진 문서를 이루고 있는 주요 토픽과, 각 토픽을 구성하고 있는 키워드들을 파악하기 위한 분석도구이다. 토픽 모델링은 주제 도출을 위한 분석에 많이 사용되고 있는데, Griffiths & Steyvers(2004)는 PNAS에 실린 논문의 초록에서 주제를 도출하여 시기별 관심 주제 등을 분석

하였다[12]. Gerrish & Blei(2010)는 논문의 시기에 따른 주제 변화를 분석하여 개별 논문의 양적 질적 영향력을 측정하였다[13]. 홍정하·최재웅(2017)은 코퍼스의 주제 구조 분석을 통해 토픽 모델링의 다양한 활용 가능성을 논의하였다[14]. 토픽모델링 기법을 활용하여 소셜미디어 데이터를 분석한 연구로는 Bae et al.(2011)의 연구가 있다. 이 연구에서는 대선에 대한 트윗 멘션 기반을 분석하여 사회적 네트워크 특성을 규명하고 있다[15]. 신문기사와 같은 매스미디어 자료에 대한 토픽모델링을 활용한 연구도 많이 이루어지고 있는데, 강범일외(2013)는 토픽모델링을 이용하여 신문자료의 오피니언 마이닝분석 통해 언론 매체별 오피니언 마이닝을 통한 매체간 발생하는 입장 차이를 밝혀내고자 하였다[16]. 토픽모델링 기반의 데이터 마이닝기법은 제품의 평판을 분석하는데에도 활용되는데, 박상민·온병원(2017)은 국내에서 생산된 자동차들의 후기글을 수집하여 토픽모델링을 실시한 후 오피니언 마이닝을 통해 제품에 대한 사용자들의 평판을 분석하였다[17].

이와 같이 토픽 모델링을 이용한 다양한 연구가 있으나, 토픽모델링기반의 데이터 마이닝 기법을 통한 제품 평판 분석을 하거나, 토픽모델링과 오피니언 마이닝을 통하여 언론 매체의 입장을 분석하는 연구, 또는 정치적 이슈와 관련한 분석 연구로서 정책과정에서의 소셜미디어와의 관계를 다루지는 않고 있다. 정책의제와 소셜미디어를 다룬 연구로서는 홍유정·황주성(2015)의 연구가 있는데 이 논문에서는 콥과 로스(Cobb & Ross)의 이론을 배경으로 시간적 선도성과 내용적 차이에 따라서 정책의제설정에서의 소셜미디어와 매스미디어의 역할 차이를 비교하는 것에 중점을 두었으며, 소셜미디어기반 데이터 분석이 정책의제 설정에 기여할 수 있는 가능성에 대하여 다루지는 못하였다. 김은미・이주현(2011)은 매스미디어에서 생산한 뉴스와 트위터에서 회자되는 뉴스의 빈도 경향에 대한 분석을 시도하여, 새로운 뉴스미디어로서의 트위터에 대하여 논하였으나, 대중들에게 회자되는 내용의 공공의제 혹은 정책의제 반영에 대하여 연구되지는 않았다. 따라서 본 논문에서는 토픽 모델링기법을 이용하여 소셜미디어 기반 여론을 분석이 정책의제 설정과정을 지원할 수 있는 가능성을 모색하고자 한다. 이를 위해 3장에서는 토픽 모델링을 수행하기 위한 데이터 수집 및 토픽 모델링 연구 방법을 기술하였으며, 4장에서는 토픽 모델링 결과를 설명하였다. 5장에서는 결론 및 시사점을 제시하였다.

3. 연구 방법

3.1 교육정책 키워드 선정

소셜미디어 기반으로 교육 정책에 대한 여론 분석을 위해 최근 교육 정책에서 주요 이슈인 소프트웨어 교육에 대해 주요 키워드를 선정하여 관련 데이터를 수집하고 분석하고자 한다.

소프트웨어 교육 정책 키워드 선정은 교육유관기관에서 발행하는 백서, 보도자료 등을 참고로 하여 후보 키워드를 추출, 교육관련 전문가 9명의 검토를 통하여 '소프트웨어 교육(SW 교육), 소프트웨어 의무(SW 의무), 코딩 교육, Computer science(컴퓨터과학), Computational Thinking(컴퓨터적 사고)'을 선정하였다.

3.2 소셜미디어 데이터 수집

교육정책에 대한 소셜미디어 기반 여론의 토픽 모델링을 실시하기 위하여 3가지 방법으로 데이터를 수집하였으며, 수집 대상 소셜미디어 서비스와 세부 수집방법은 다음과 같다.

3.2.1 네이버 블로그

네이버 블로그의 데이터를 수집하기 위하여 JAVA 프로그래밍 언어 기반의 웹크롤러를 활용하였다. 입력은 URL 클래스의 OpenStream 메서드; BufferedInputStream을 이용하였고, 출력은 FileOutputStream, mySql DB를 이용하였다.

3.2.2 네이버뉴스

네이버 뉴스의 수집 방법은 네이버 블로그와 같은 JAVA기반의 자체 웹크롤러를 이용해 네이버 뉴스 기사들을 수집하였다. 각 키워드는 일치 조건 검색을 이용하여 "키워드"로 검색된 결과를 수집하며 수집 기간과 키워드를 Input으로 받아 실행되고, 수집된 데이터는 연결된 데이터베이스에 저장된다.

3.2.3 트위터

트위터는 OpenAPI 제공을 통하여 많은 기능을 제공하고 있어서 많은 사용자들이 트위터를 사용하고 있다. 트위터 OpenAPI는 트위터에서 생산하는 데이터를 외부에

서 활용할 수 있도록 제공하는 인터페이스로서 다양한 사용자들이 OpenAPI를 통해 트위터 데이터를 수집 및 분석하거나 검색하고, 이를 이용한 앱을 만들고 있다. 본 논문에서는 트위터 데이터 수집을 위하여, 트위터가 제공하는 JAVA 라이브러리인 Twitter4J를 선택하여 활용하였다. Twitter4J는 가장 널리 사용하는 트위터 라이브러리로써 지속적인 업데이트가 수행되며, 관련 문서화가 잘 되어 있는 장점이 있다. 또한 트위터는 REST API와 Streaming API 두 가지 방식이 있는데, REST API는 단어의 검색 결과, 사용자 정보 활용과 같은 특정 작업을 수행하는데 적절하고, Streaming API는 트위터 데이터를 실시간 활용할 경우에 유용하다. 본 논문에서는 REST API의 Search API를 사용하였다.

트위터 데이터는 이 API를 이용하여 최근 10일간 수집할 수 있으며, 10일 이내에서는 일 단위로 임의로 날짜를 선택하여 수집할 수 있다. 본 논문에서는 트위터 API를 통하여 가져올 수 있는 사용자, 작성시간, 트위터 콘텐트 지리적 위치, 해쉬태그, 리트윗 관련정보 등을 가져올 수 있는데, 본 논문에서는 트위터 API 수집기를 이용하여 작성자, 시간, 콘텐트 등의 항목들을 활용하여 분석에 사용하였다.

3.3 토픽 모델링 분석

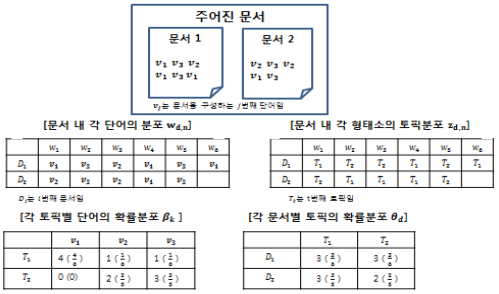
본 논문에서는 선정된 교육정책 키워드를 기반으로 수집한 소셜미디어 데이터를 분석하기 위해 토픽모델링을 이용하였으며, 추출된 문서를 분석하여 출현빈도와 트렌드 변화 분석을 수행하였다.

먼저, 수집된 소셜미디어 문서들에 대하여 토픽모델링을 수행하고 그 결과 데이터를 바탕으로 비슷한 토픽끼리 군집화하여 토픽 그룹화를 수행하였다. 그룹화된 토픽들은 AKE(Automatic Keyword Extraction)과 PMI, Page Rank 방법론을 토픽그룹의 타이틀을 선정하였고, 토픽그룹의 트렌드를 분석하는 과정으로 진행하였다.

토픽분석을 위해 R프로그램을 이용하여 LDA 토픽모델링을 수행하였으며, 사용한 기본 파라미터와 그 기능은 다음과 같다.

- α(alpha): 문서들이 토픽에 분포되는데 영향을 준다. α값이 증가하면 문서들이 특정 토픽에 편중되는 경향이 있으며, α값이 감소하면 각 문서들이 더 많은 토픽그룹에 퍼져서 분포하게 된다.
- β(beta): 토픽에 단어들이 분포되는데 영향을 준다. β값이 증가할수록 각 단어의 토픽 분포가 다양해지며, β값이 감소할수록 각 단어가 특정 토픽에 편중되어 분포하는 경향이 커진다.
- iteration : 각 단어들을 토픽 할당의 일반화를 위해 일정 횟수 이상 반복적인 샘플링을 수행한다.
- 토픽수(K): 최종 추출된 토픽의 개수를 의미한다.

본 논문에서 소셜미디어 문서를 이용한 토픽모델링의 파라미터 값은 α: 0.1, β: 0.1, iteration: 3000이며, 최종 토픽의 수는 20개로 지정하였다. 입력 데이터는 앞서 수집한 문서들로 하였으며, 1차적으로 토픽 분석을 통해 문서별 각 토픽의 비중과 토픽별 각 단어의 비중을 추출하였다. 그림 1은 v_1, v_2, v_3 3개 단어로 작성된 문서1과 문서2에 대해 2개의 토픽을 추출하는 LDA토픽모델 분석을 진행한 예시이다.



(그림 1) LDA 토픽모델 예시
(Figure 1) LDA Topic model example.

각 문서에 대하여 공백기호와 문장부호를 기준으로 나누어진 문장의 최소 단위인 형태소분석을 하고, LDA 토픽모델 분석을 진행하는데, 예시에서는 주어진 두 문서를 그림의 문서 내 각 단어의 분포 $w_{d,n}$, 그리고 토픽모델 분석결과로 문장의 각 형태소가 어떤 토픽에 분포되었는지에 대한 정보 $z_{d,n}$, 각 토픽에서 v_1, v_2, v_3 3개 단어가 각각 차지하는 비중에 대한 정보 β_k, 각 문서에서 토픽1과 토픽2가 각각 차지하는 비중에 대한 정보 θ_d를 얻을 수 있다

토픽 모델링을 통해 각 토픽에서 단어의 비중을 논하는 방법으로는 토픽 출현 확률과 스코어를 이용하는 방법이 있다. 토픽 출현 확률은 각 토픽별 단어의 확률분포인 $\beta_{k,w}$를 이용하는 방법이고, 스코어를 이용하는 방법

은 앞서 추출된 $\beta_{k,w}$데이터를 활용하여 수식계산을 통해 스코어를 계산하는 방법이다. 스코어 계산 수식은 $score_{k,w} \equiv \beta_{k,w}(\log\beta_{k,w} - \frac{1}{K}\sum_{k'}\log\beta_{k',w})$로 표현된다. 본 논문에서는 토픽모델링 결과에 따라 도출된 토픽들 중 키워드 비중을 보기 위해, 위에 설명한 두 방법 중에서 확률 분포를 이용하는 방법을 활용하여 토픽에 포함된 단어를 정렬하고, 토픽모델링 분석에 활용하였다.

토픽모델링의 결과는 크게 3가지로 구분할 수 있다. 각 문서별 토픽의 확률분포와, 각 단어별 토픽의 확률분포 그리고 각 토픽이 전체 토픽에서 출현할 확률로 구분할 수 있다.

토픽모델링 결과로 얻은 토픽들을 바탕으로 군집분석을 진행하여 유사한 토픽끼리의 군집을 구성하였다. 토픽들의 군집화는 K-means 클러스터링 방법을 이용하였으며, 클러스터링은 R프로그램과 JAVA를 이용하여 실행하였다.

K-means 클러스터링은 유클리드 거리 값을 이용해 정해진 K값만큼의 군집을 만드는 알고리즘이며, 본 논문에서는 K값은 7로 설정하였다.

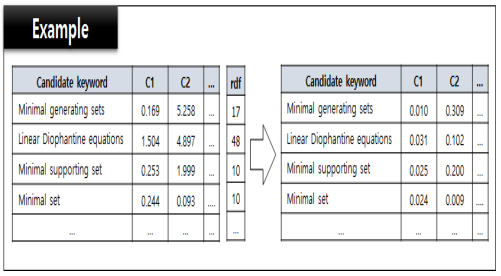
본 논문에서는 앞서 분석한 토픽 그룹의 타이틀을 선정하기 위해 기존 연구 방법인 AKE(Automatic Keyword Extraction) 방법론을 사용하였다(Rose, 2010). AKE의 절차는 다음과 같다.

첫째, 각 문서에 출현한 단어별로 다음 값을 계산한다.

- Word frequency(키워드가 사용된 횟수) : freq(w)
- Word degree(키워드가 포함된 문서에 있는 단어 수): deg(w)
- Ratio of degree to frequency(rdf, 키워드가 포함된 타이틀 후보군) : deg(w)/freq(w)

둘째, words scores를 계산한다. 추출된 단어별로 값을 계산하며 각 단어의 score는 deg(w)/freq(w)값으로 계산한다. 셋째, 최소 두 개 단어로 구성되어 있는 단어를 뽑아 "Candidate keyword" set을 만든다. 넷째, 전체 word의 점수 순서로 상위 1/3 개의 candidate keywords만 남기고 이를 최종 선정 키워드로 한다.

토픽 타이틀을 선정하기 위해 후보 키워드들에 대한 PMI를 계산한 후 이어서 Page-Rank를 계산한다. AKE방법을 통하여 추출한 토픽타이틀 후보단어의 rdf(k) 값과 토픽타이틀 후보단어를 포함하는 문서의 토픽 포함 확률을을 합산하여, 각 토픽타이틀 후보셋의 토픽 포함 확률을 계산, 각 확률값을 rdf(k) 값으로 나누어 토픽 그룹을 대표 할 수 있는 후보 키워드들을 추출한다. 이후 후보 키워드들의 PMI와 Page-Rank 계산을 통해 최종 토픽그룹에 대한 타이틀을 선정한다.



(그림 2) 토픽그룹별 타이틀 후보 키워드 확률값 계산
(Figure 2) Calculation of the probability values of the Title Candidate keywords by Topic Group.

토픽 그룹의 타이틀 선정을 위하여 단어의 중심성 값을 구하였으며, 입력 값으로는 용어간 밀접도를 나타내는 PMI(pointwise mutual information) 지수를 사용하였다. PMI 지수는 하나의 단어 쌍 (x, y)에 단어 x와 y가 각각 등장하는 확률과 동시 등장하는 확률을 사용하여 계산한다. PMI 값이 크면 클수록 단어 사이에 높은 관련성이 있다고 볼 수 있다. PMI 수식은 다음과 같다.

$$PMI(x,y) = \log\frac{P(x,y)}{P(x)P(y)}$$

위의 PMI 공식이 일반적이나 본 논문에서는 음수가 나오지 않는 것을 고려하여, 기존의 PMI공식을 바탕으로 아래의 공식을 이용하여 토픽그룹의 타이틀 선정 과정에 이용하였다. 아래 수식에서 hits(word)는 word에 대한 웹 페이지의 검색 결과 수를 의미한다.

$$PMI(Word_1, Word_2) = \frac{hits(Word_1, Word_2)}{hits(Word_1) \times hits(Word_2)} \times 100000000$$

페이지랭크(Page-Rank)는 하이퍼링크 구조를 가지는 문서에 대하여 상대적 중요도에 따라 가중치를 부여하는 방법으로서, 인용과 참조로 연결된 임의의 문서들에 적용할 수 있다. Page-Rank에 대한 수식은 다음과 같다.

$$PR(A) = [\frac{1-d}{N} + \frac{d(PR(T_1))}{C(T_1)} + ... + \frac{d(PR(T_n))}{C(T_n)}]$$

- PR(A): ‘A’ 노드의 페이지랭크 값
- N: 모든 노드 수
- d: 이탈 확률(damping factor), 일반적으로 페이지를 무작위적으로 탐색하는 임의 탐색기(random surfer)를 가정했을 때, 해당 페이지에서 다른 페이지로 이동하는 링크를 클릭할 확률로서 0과 1 사이의 값이며, 일반적으로는 0.85이 기본 값이다.
- C(T): 노드 T가 가지는 링크 개수

본 논문에서는 PMI와 Page-Rank 값을 계산하고, 이를 바탕으로 후보 키워드별 중심성 값을 계산하여 각 토픽 그룹별들에 대한 최종 타이틀을 선정하였다.

4. 연구 결과

4.1 문서별 토픽의 출현 확률 분포

다음 표는 소프트웨어 교육관련 문서의 문서별 토픽 출현 확률을 나타낸 것이다. 표에서 docNo는 각 문서의 고유번호이며 date는 문서의 게시일이다. 토픽모델링에 사용한 각 문서는 토픽 20개에 다양하게 분포 되며 각 행의 값을 모두 더하면 1이 된다. 예를 들면, A1번 문서는 네이버 뉴스 기사로 토픽1에서의 출현 확률이 0.273 (27.3%)이고, 토픽20에서의 출현 확률은 0.052(약 5.2%)인 것을 알 수 있다.

(표 1) 소프트웨어 교육 관련 문서별 토픽 출현확률
(Table 1) Probability of Occurrence of Topic per Documents in Software Education

docNo	date	Topic1	Topic2	…	Topic19	Topic20
A0000000001	20160526	0.273	0.000	…	0.000	0.052
A0000000002	20160526	0.208	0.015	…	0.162	0.000
A0000000003	20160524	0.444	0.000	…	0.000	0.272
A0000000004	20160529	0.000	0.000	…	0.141	0.000
A0000000005	20160527	0.000	0.010	…	0.000	0.000
…	…	…	…	…	…	…

토픽모델링 결과로 추출한 20개의 토픽이 전체에서 어느 정도의 확률로 출현 하는지에 대한 결과가 다음 표 2와 같다. (표 2)에서 토픽 번호 T1의 확률이 5.12%라는 것은, 토픽 1번이 전체 토픽중에서 5.12%의 비중을 차지한 것이라 할 수 있다. 결과를 살펴보면 토픽 18번이 전체의 24.93%로 가장 비중이 높고, 토픽 5번이 전체의 0.91%로 비중이 가장 낮은 것을 알 수 있다.

(표 2) 소프트웨어 교육 관련 각 토픽별 확률
(Table 2) Probability of Each Topic in Software Education(단위 : %)

토픽	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
확률	5.12	3.30	5.22	2.22	0.91	5.31	3.91	5.03	7.25	3.26
토픽	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20
확률	3.90	2.42	1.08	2.02	11.8%	1.61	4.64	24.9%	2.69	3.31

토픽모델링 결과로 얻은 각 토픽별 상위 5개 단어는 다음 표 3에 나타낸 것과 같다. 각 표에서의 순위는 각 토픽에서 단어의 출현확률(토픽에서 각 단어가 차지하는 비중)을 기준으로 내림차순으로 정리 한 결과이다. 토픽 2의 경우 ‘인공지능’, ‘미래’, ‘인간’, ‘AI’순으로 토픽 내에서 해당 단어가 차지하는 비중이 높다는 것을 의미한다. 각 토픽은 많은 단어들이 확률적으로 분포되어 있고, 이는 각 토픽들에게 특성을 갖도록 해준다. 예를 들어, 토픽2의 경우 ‘인공지능’, ‘미래’, ‘인간’ 등 미래의 인공지능과 관련된 단어들이 모이고, 토픽3의 경우 ‘대학’, ‘미국’, ‘학생’, ‘학교’ 등 대학, 교육과 관련된 단어들이 모여 있는 것을 알 수 있다.

(표 3) 소프트웨어 교육 관련 토픽1~토픽20의 상위 5개 단어
(Table 3) Top 5 words from topic 1 to topic 20 in Software Education

순위	1	2	3	4	5
T1	개발	기업	서비스	지원	클라우드
T2	인공지능	미래	인간	AI	시대
T3	대학	미국	학생	학교	대학교
T4	아시아경제	기자	코딩교육	어린이	카드・뉴스
T5	게임	포켓몬	게임・문화	부모	대통령
T6	사람	말	생각	교수	정보
T7	로봇	네이버・앱	번역	전용・뷰어	확인
T8	학생	대학	인재	소프트웨어	교육
T9	교육	소프트웨어	코딩	취업	양성
T10	교육과정	시간	과목	컴퓨터과학	머니투데이
T11	코딩교육	코딩	교육	대상	리얼·코딩
T12	아이	네이버	친구	레고	감사

순위	1	2	3	4	5
T13	텃밭·이야기	오늘	문자	옥수수	택배
T14	it	computer	science	주임·교수	thinking
T15	교육	소프트웨어	학생	SW교육	프로그램
T16	선발	단계	전형	등급	대학
T17	순위	조회수	드론	기준	개발
T18	교육	코딩	소프트웨어	컴퓨터	프로그래밍
T19	대회	팀	연세대학교	해킹	참가
T20	교육	의원	SW교육	소프트웨어	국회

4.2 토픽 클러스터링 및 타이틀 선정

본 논문에서는 토픽모델링의 결과인 20개의 토픽을 각 토픽별 단어의 출현확률을 중심으로 K-means 클러스터링을 진행하여 주제가 비슷한 토픽끼리 총 7개의 그룹을 형성하였다. 소프트웨어 교육 관련 토픽들의 군집 결과는 다음 표 4와 같다. 군집 이름은 SG1~SG7로 하였다.

(표 4) 소프트웨어 교육 관련 군집 결과
(Table 4) Topics Clustering Results of Software Education

군집번호	각 군집별 토픽번호
SG1	토픽5
SG2	토픽9
SG3	토픽13
SG4	토픽11, 토픽15, 토픽18, 토픽20
SG5	토픽14
SG6	토픽2, 토픽17
SG7	토픽1, 토픽3, 토픽4, 토픽6, 토픽7, 토픽8, 토픽10, 토픽12, 토픽16, 토픽19

소프트웨어 교육 관련 토픽 군집에 대한 타이틀 선정을 위해, AKE 방법을 활용하여 문서들로부터 타이틀 후보 키워드를 추출하였다. 소프트웨어 교육 관련 키워드로 수집한 문서 1870개에서 후보키워드를 추출하였다. 이 키워드들 중 rdf(k)가 6 이상인 키워드를 추출하고, 각 키워드가 포함된 문서들의 토픽 그룹 확률을 계산하였다.

선정된 후보 키워드들의 PMI지수를 구하고, 이를 이용하여 Page-Rank(PR)를 계산하였다. 토픽 군집의 최종 타이틀을 선정하기 위해 각 토픽 그룹이 전체에서 출현할 확률 값의 순서대로 토픽그룹에 포함된 후보 키워드를 Page-Rank 순서로 정렬하여 3개씩 추출하였다. 만약 동일 후보키워드가 서로 다른 그룹의 Page-Rank 값 3위 안에 각각 랭크되어 있다면, 토픽그룹이 전체 중에 출현할 확률 값이 큰 그룹이 A키워드를 선정하고, 출현 확률 값이 적은 토픽그룹은 A를 제외한 키워드 3개를 선정하게 된다. 따라서 각 토픽그룹은 모두 서로 다른 타이틀 후보를 갖게 된다.

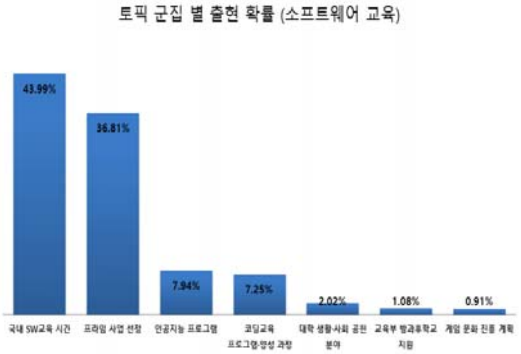
이런 방법으로 모든 토픽 그룹이 3개의 타이틀 후보 키워드를 선정하게 되고, 그 중에서 최종 그룹 타이틀을 선정한다. 다음 표 5는 각 토픽 그룹의 최종 선정 타이틀을 나타낸다.

(표 5) 토픽군집별 최종 선정된 타이틀
(Table 5) Finally Selected Group Title of Each Group

토픽 군집	최종 군집 타이틀
SG1	게임 문화 진흥 계획
SG2	코딩교육 프로그램·양성 과정
SG3	교육부 방과후학교 지원
SG4	국내 SW교육 시간
SG5	대학 생활·사회 공헌 분야
SG6	인공지능 프로그램
SG7	프라임 사업 선정

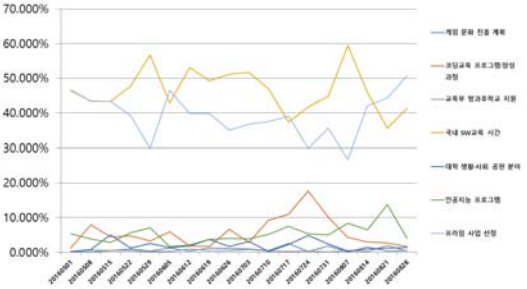
4.3 토픽의 그룹별 출현 확률 및 트렌드

다음 그림 3은 클러스터링 결과 각 토픽의 군집별 출현 확률을 확률의 크기 순서로 나타낸 히스토그램(histogram)이다. 아래의 그림에서 x축은 위의 K-means 클러스터링 결과로 나눠진 각 군집의 타이틀을 의미하며 y축은 각 그룹이 출현 할 확률을 의미한다. 예를 들어 “프라임 사업 선정” 군집은 전체 7개의 그룹 중 출현 확률이 36.81%가 되며 “인공지능 프로그램” 군집은 7.94%의 출현 확률을 나타낸다. “국내 SW 교육 시간” 군집이 43.99%로 가장 높은 출현확률을 보임을 알 수 있다.



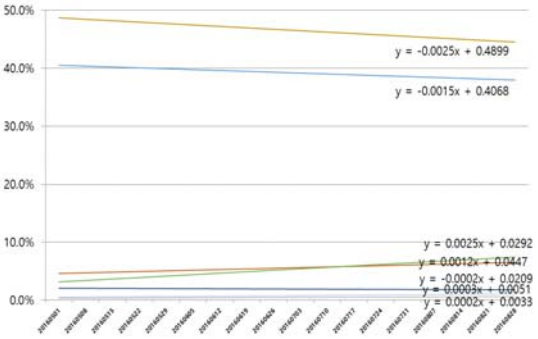
(그림 3) 토픽의 군집 별 출현 확률 (소프트웨어 교육)
(Figure 3) Probability of Occurrence of Each Topic Group(Software Education).

다음 그림 4는 소프트웨어 교육에 대한 토픽 군집 트렌트를 보여주고 있다. “코딩교육 프로그램·양성 과정” 군집의 트렌드인 주황색 선을 보면 7월에 그래프가 특이하게 높은데, 이는 7월에 “코딩교육 프로그램·양성과정” 군집에 대한 관심이 높아졌음을 유추할 수 있다. 또한 “국내 SW교육 시간”의 트렌드를 나타내는 노란색선을 보면 8월 초에 “국내 SW교육 시간”관련 군집에 관심이 높아졌음을 유추할 수 있다.



(그림 4) 소프트웨어 교육 토픽 군집 트렌드
(Figure 4) Software Education Topics Group Trends.

다음 (그림 5)는 소프트웨어 교육 토픽 트렌드에 대한 추세선을 나타낸다. 추세선은 선형회귀를 이용하여 구한 것으로 전체 기간의 트렌드를 단순화 시켜 볼 수 있다. “국내 SW교육 시간”을 나타내는 노란색 그래프를 보면 5월부터 8월까지 관심이 감소하는 추세인 것을 알 수 있고, 토픽 “인공지능 프로그램”을 나타내는 초록색 그래프 선을 보면 추세가 전체적으로 증가하고 있음을 확인 할 수 있다.



(그림 5) 소프트웨어 교육 토픽 군집 트렌드의 추세선
(Figure 5) Regression Line of Software Education Topics Group.

5. 결론 및 시사점

본 논문에서는 소셜미디어를 활용하여 정책과정 지원 가능성을 탐색하기 위해, 토픽모델링을 이용하여 소프트웨어교육 정책에 대한 소셜미디어 데이터를 분석하였다.

네이버블로그, 네이버뉴스, 트위터를 대상으로 소프트웨어교육에 대한 데이터를 4개월간 수집하여 총 1,870개의 문서를 수집하였다. 이 문서들을 대상으로 토픽분석을 통해, 주요 토픽, 토픽별 확률, 토픽 추세의 총 3가지 지표를 얻을 수 있었으며, 주요 결과는 다음과 같다.

먼저 주요 토픽모델링을 통해서 수집한 문서들로부터 주요 토픽을 도출할 수 있었으며, 각 토픽에서 중요한 비중을 차지하는 단어들의 그룹을 확인할 수 있었다. (표3)에서 보면 토픽2(T2)의 경우 인공지능, 미래, 인간, AI 등 미래사회 기술관련 단어들이, 토픽18(T18)의 경우 코딩, 교육, 소프트웨어, 컴퓨터, 프로그래밍 등 소프트웨어교육관련 단어들이 중요한 비중을 차지하고 있는 것으로 나타났다. 토픽5(T5)의 경우는 게임, 포켓몬, 게임문화 등 게임과 관련된 단어들이, 토픽16(T16)의 경우는 선발, 단계, 전형, 등급, 대학 등 대학입시와 관련된 단어들이 중요한 비중으로 출현하였다. 이는 토픽모델링을 통해서 도출된 토픽들과 그 토픽에서 출현할 확률이 높은 단어들의 그룹핑을 통하여 정책분류와 그룹핑에 활용할 수 있는 가능성을 나타낸다고 할 수 있다.

두 번째로, 토픽군집을 통하여 비슷한 토픽을 그룹핑하고 그에 대한 대표 타이틀을 선정하는 것을 기술적으로 구현함으로써 상향식 정책 유형 혹은 키워드 분류작업이 가능하였다. 본 논문에서는 k-means 클러스터링을 통하여 총 7개의 토픽군집을 만들었으며, 그 중 ‘국내 컴퓨터 교육 시간’ 이라는 토픽군집이 전체의 43.99%를 차지하였다. 이 토픽 군집에는 T11, T15, T18, T20과 같이 4개의 토픽이 그룹핑 되었으며, (표3) 토픽별 상위 비중 단어를 확인해 보면 대부분 코딩, 소프트웨어, 교육 등 SW교육과 관련된 단어들로 구성되어 있다. 또한 ‘프라임 사업 선정’ 이라는 토픽군집은 전체 토픽군집의 36.81%를 나타내었는데, 이 토픽군집에는 총 10개의 토픽이 연결되었으며, 마찬가지로 (표3) 토픽별 상위 비중 단어를 확인해 보면 이 토픽군집이 교육과정 및 대입은 물론, 산업과도 연관이 두루두루 있음을 알 수 있었다. 마지막으로 ‘인공지능 프로그램’ 군집은 7.94%를 차지하며 T2, T17 2개 그룹이 그룹핑 되었는데 해당 토픽들은 드론, 인공지능, AI 등 미래사회 기술들을 지향하는 단어들과 연결이 되어 있는 것을 알 수 있다. 이와 같이 주요 토픽군

집화의 결과를 보면 소프트웨어 교육 정책은 미래사회의 기술과 프라임사업, SW교육의 3가지 토픽과 큰 관계가 있다는 것을 알 수 있다.

마지막으로 토픽모델링 추세분석을 통하여 유의미한 결과를 도출해 볼 수 있다. 소셜미디어 상에서 주요 토픽들의 비중이 어느 정도 인지, 시간이 지남에 따라 해당 토픽이 감소하는 추세인지 증가하는 추세인지, 즉 어떠한 토픽들이 소셜미디어 상에서 관심을 받고 있으며, 관심이 변화하고 있는지에 대한 정보를 통하여, 정책 의제에 반영할 수 있는 가능성을 탐색해본다. 본 연구의 결과인 (그림 4)의 내용을 토픽 군집별로 비교 분석해보면 다음과 같다. “국내 SW교육 시간”, “프라임 사업 선정” 두 토픽군집은 전체 기간에 걸쳐 높은 비중을 차지하는 군집들인데 한 군집이 증가하면 다른 군집이 감소하고 또 그 반대 현상을 나타내는 경향이 있다. 이것은 두 군집 간에 음의 상관관계가 있는 것으로 보여진다. “인공지능 프로그램” 군집은 전체 기간 동안 꾸준히 증가하는 양상을 보이고 특히 8월에 들어 대폭 증가하는 추세를 보인다. 또한 추세선의 기울기가 높은 양의 값을 가져 토픽 군집에 관련한 사항들이 점점 이슈화되고 있고, 사람들에게 많이 언급되고 있다는 것으로 해석할 수 있다. “게임 문화 진흥 계획”, “교육부 방과 후 학교 지원” 군집은 대체로 매우 낮은 추세를 보이다가 각각 7월과 8월 이후부터 조금씩 출현하는 것을 볼 수 있다. 이는 이 토픽 군집들이 모두 초중등학교 방학기간과 연관이 있을 수 있다고 해석할 수 있다. “코딩교육 프로그램 양성 과정”, “대학 생활 사회 공헌 분야” 토픽군집들은 5월과 7월에 큰 증가를 보이지만 그 시기가 지나면 다시 감소하는 시기성을 가지고 있다고 보여진다. 따라서 이 두 토픽 군집들은 단기적으로 특정 시기에만 관심이 생기는 군집들로 해석 할 수 있다. 이와 같은 토픽 분석 결과는 정책 수립, 정책 평가 등과 같은 정책과정에 반영해 볼 수 있을 것이다. 앞에서 언급한 “게임 문화 진흥 계획”, “교육부 방과 후 학교 지원” 토픽의 트렌드는 방학기간에 대한 교육 정책을 수립할 때 고려할 수 있는 분석 결과라고 할 수 있다. 또한 크게는 소프트웨어 교육의 범위이면서도 초중등학교와 관련된 “국내 SW교육 시간” 토픽과 대학교와 관련된 “프라임 사업 선정” 토픽이 서로 음의 상관관계를 보이고 있는 것은 향후 정부의 교육기관의 소프트웨어 교육 정책 검토 시 초중등학교에서 대학으로 이어지는 교육과정의 연계성이나 대입정책 등 관련정책 분석이 필요할 것으로 판단된다.

본 논문에서는 테이터 수집이 단기간에 이루어졌다는 점과 오피니언마이닝, 네트워크 분석 등 다양한 분석 기법이 활용되지 못했다는 한계점이 있다. 향후에는 본 논문에서 실시한 토픽모델링을 통해 나온 결과와 각 토픽이 도출된 문서의 오피니언마이닝, 네트워크 분석을 병행하여 다양한 분석을 시도해 볼 필요가 있다. 또한 연구의 목적에 따라 다양한 소셜미디어를 대상으로 이와 같은 분석 연구를 추진하고, 소셜미디어의 주요 저자들 성향과 논조에 대하여 시간을 가지고 분석을 시도한다면, 소셜미디어 여론에 대한 좀 더 구체적인 분석이 가능해질 것이다. 향후 소셜미디어의 영향력이 증대되면, 이러한 데이터에 대한 분석 및 결과 활용이 무엇보다 중요해질 것이다. 또한 정책과정에서의 소셜미디어 의견 반영도 중요해지리라 판단된다.

참고문헌(Reference)

[1] Jin-Myeong Chung, Woo-Joo Kim, Chan-dong Koo, ″Social Media Big Data Analysis for ICT Policy Agenda in Education,″ pp. 4, Korea Education and Research Service(KERIS), 2016. http://lib.keris.or.kr/search/detail/CATLAB000000012076?briefLink=/searchA/lab?briefType=L?st=KWRD_A_si=TOTAL_A_q=%EB%B9%85%EB%8D%B0%EC%9D%B4%ED%84%B0

[2] Jin-hyong Lee, ″Proliferation and Trends of SNS(Social Network Service),″ Journal of Communication &Radio Spectrum, Vol. 44, 2012. https://www.kca.kr/open_content/bbs.do?act=detail&msg_no=10462&bcd=radiotrends&keyfield=bbs_title&keyword=SNS

[3] A. Livne, M. Simmons, E. Adar, and L. Adamic, “The Party is Over Here : Structure and Content in the 2010 Election,” Proceedings of 5th ICWSM(2011). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2852/3272

[4] You-Jung Hong, Joo-Seong Hwang, “Social Media vs. Mass Media in the Policy Agenda-setting Process : The Case of the Gwangju Inhwa School Incident(Dogani),” Broadcasting & Communication, Vol. 16, No. 1, 2015. http://www.dbpia.co.kr/Journal/ArticleDetail/NODE06268584

[5] Hyun-jae Yu, Ji-eun Song, “Investigation into the Styles of Articles Regarding Suicide by Different Media: Comparison of the Styles between Internet Media and Newspapers,” Health and Social Welfare Review, Vol. 32, No 2, pp. 427-467, 2012. http://www.dbpia.co.kr/Journal/ArticleDetail/NODE01899750

[6] Russell Neuman, W.Guggenheim, Lauren Mo Jang, S. Soo Young Bae, “The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data”, Vol. 64, No 2, pp. 193-214, 2014 https://doi.org/10.1111/jcom.12088

[7] Eun Mee Kim, Ju Hyun Lee ″The Diffusion of News through Twitter and the Emerging Media Ecosystem,″ Korean Journal of Journalism & Communication Studies, Vol. 55, No. 6, pp. 152-180, 2011. http://www.dbpia.co.kr/Journal/ArticleDetail/NODE01757395

[8] Jin-myeong Chung, Ki-young Yoo, Chan-dong Koo, “A Study on Social Media Sentiment Analysis for Exploring Public Opinions Related to Education Policies”, Vol 24, No 4, pp. 3~16, 2017. http://www.papersearch.net/thesis/article.asp?key=3578607

[9] Jin-Ho Choi, Dong-Sub Han, “A Study on the Correlation of Agendas between Politicians’ Twitters and traditional News Media,” Journal of Communication Science, Vol. 11, No. 2, pp. 501~532, 2011.

[10] D. Blei, ″Probabilistic Topic Models,″ Communication of the ACM, Vol. 55, No. 4, pp. 77-84, April. 2012. https://doi.org/10.1145/2133806.2133826

[11] M. Steyvers and T. Griffiths, ″Probabilistic Topic Models,″ Handbook of latent semantic analysis. Analysis. Edited by T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch. NJ: Erlbaum, 2007. http://173.236.226.255/tom/papers/SteyversGriffiths.pdf

[12] T. Griffiths, & M. Steyvers, ″Finding scientific topics,″ Proceedings of the National Academy of Sciences, Vol. 101, pp. 5228-5235, 2004. https://doi.org/10.1073/pnas.0307752101

[13] S. Gerrish, and D. Blei, “A Language-based Approach to Measuring Scholarly Impact,” The 27th International Conference on Machine Learning, pp. 375-382, 2010. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.182.4459&rep=rep1&type=pdf

[14] Jung-ha Hong, Jae-Woong Choe, “Exploring the Thematic Structure in Corpora with Topic Modeling,” Language & Information Society, Vol. 30, pp. 239~275, 2017. http://www.papersearch.net/thesis/article.asp?key=3507933

[15] Jung-hwan Bae, Ji-eun Son and Min Song, ″Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques,″ Journal of Intelligent Information System, Vol. 19, No. 3, pp. 141-156, 2013. http://dx.doi.org/10.13088/jiis.2013.19.3.141

[16] Beom-il Kang, Min Song, and Wha-sun Jho, ″A Study on Opinion Mining of Newspaper Texts based on Topic Modeling,″ Journal of the Korean Society for Library and Information Science, Vol. 47, No. 4, pp. 315-334, 2013. http://www.dbpia.co.kr/Article/NODE02321458

[17] Sang-Min Park and Byung-Won On, ″Latent topics-based product Reputation Mining,″ Journal of Intelligent Information System, Vol. 23, No. 2, pp. 39-70, 2017. http://dx.doi.org/10.13088/jiis.2017.23.2.039

◐ 저 자 소 개 ◑

정 진 명(Chung, Jin-myeong)
2001년 고려대학교 컴퓨터학과(이학학사)
2013년 고려대학교 정보보호대학원(공학석사)
2016년~현재 경북대학교 정보보호학과 박사과정
관심분야: 컴퓨터공학, 정보보호, 교육정보화
E-Mail: jinmong@keris.or.kr

박 영 호(Park, Young-Ho)
1989년 2월 경북대학교 전자공학과 학사
1991년 2월 경북대학교 전자공학과 석사
1995년 2월 경북대학교 전자공학과 박사
1996년~2008년 상주대학교 전자전기공학부 교수
2003년~2004년 Oregon State Univ. 방문교수
2008년~2014년 경북대학교 산업전자공학과 교수
2014년~현재 경북대학교 전자공학부 교수
관심분야: 정보보호, 빅데이터
E-Mail: yhpark@knu.ac.kr

김 우 주(Kim, Wo-joo)
1987년 연세대학교 경영학과(경역학학사)
1989년 한국과학기술원 경영과학과(경역학석사)
1994년 한국과학기술원 경영과학과(경역학박사)
2004년~현재 연세대학교 정보산업공학과 교수
관심분야: 인공지능, 빅데이터
E-Mail: wkim@yonsei.ac.kr