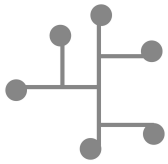


소셜 네트워크에서 연관 문서 분석을 통한 지역 이벤트 검출 기법

Local Event Detection Scheme by Analyzing Relevant Documents in Social Networks

박수빈(Soobin Park)[1] 최도진(Dojin Choi)[2]
복경수(Kyoungsoo Bok)[3] 유재수(Jaesoo Yoo)[4]

요 약

최근 소셜 네트워크 데이터를 활용하여 지역 이벤트를 검출하기 위한 연구들이 활발하게 진행되고 있다. 본 논문에서는 이벤트 검출의 정확도을 향상시키기 위해 연관 문서 분석을 통한 지역 이벤트 검출 기법을 제안한다. 지리 정보를 최대한 활용하기 위해서 지리 정보 사전을 이용하여 지리 정보를 임베딩하고 소셜 네트워크 특성을 이용하여 가중치를 부여한 키워드 그래프를 생성한다. 사용자들이 남기는 정보의 형태는 포스팅 뿐만 아니라 댓글과 스레드와 같은 연관 문서의 형태로도 나타난다. 제안하는 기법은 이와 같은 연관 문서 분석을 통해 구축한 키워드 그래프를 바탕으로 이벤트를 검출한다. 제안하는 기법은 지리 정보 사전을 이용하여 사용자가 지리 정보를 태그하지 않아도 지역 관련된 정보를 임베딩한 연관 문서를 분석함으로써 지역 이벤트 검출의 정확도를 높일 수 있다. 제안하는 기법의 우수성을 입증하기 위해서 기존의 이벤트 검출 기법과 다양한 성능 평가를 수행한다.

주제어: 소셜 네트워크 서비스, 이벤트 검출, 연관 문서, 키워드 그래프

1 충북대학교 빅데이터협동과정, 박사과정.
2 충북대학교 정보통신공학부, 박사후연구원.
3 원광대학교 SW융합학과, 교수.
4 충북대학교 정보통신공학부, 교신저자.

+ 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원(NRF-2017S1A5B8059946), 산업통상자원부와 한국산업기술진흥원의 "R&D재발견프로젝트"의 지원(과제번호: P0010202, 과제명: 소셜 빅데이터 기반의 개인맞춤형 취업 콘텐츠 추천(큐레이션) 및 디지털 증명서 발급 시스템). 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보ㆍ컴퓨팅기술개발사업의 지원(No. NRF-2017M3C4A7069432) 및 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(No.B0101-15-0266, 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발)
+ 논문접수: 2020년 03월 17일, 최종 심사완료: 2020년 04월 13일, 게재승인: 2020년 04월 18일.

Abstract

Recently, studies have been done to detect events by utilizing vast amounts of social network data. In this paper, we propose a local event detection scheme by analyzing relevant documents in social networks to improve the accuracy of event detection. To detect local events by using geographical data, the proposed scheme embeds them using the geographical data dictionary and generates a weighted keyword graph using social network characteristics. The data left by users in social networks include not only postings but also related documents such as comments and threads. Therefore, the proposed scheme detects the local event based on a keyword graph that is constructed through the analysis of the relevant documents. It can improve the accuracy of local event detection by analyzing relevant documents embedded with region-related information without requiring users to tag geographic data using the geographical data dictionary. In order to verify the superiority of the proposed scheme, we compare it with the existing event detection schemes through various performance evaluations.

Keywords: social network service, event detection, relevant documents, keyword graph

1. 서 론

네트워크 기술의 발전과 함께 모바일 스마트 기기의 대중화로 인해 사용자의 인적 네트워크를 기반으로 의사 소통과 정보를 공유하는 소셜 네트워크 서비스(SNS : Social Network Service)가 활발하게 활용되고 있다[1, 2, 3]. SNS는 인맥 관계을 기반으로 한 정보 공유 뿐만 아니라 이벤트 발생시 이를 사용자들에게 전달하는 도구로 사용되고 있다. 2012년 허리케인 샌디가 미국 동부를 강타했을 때 당시 송전탑 피해로 인해 전력 공급이 어려워져 각 가구의 발전기를 가동하기 위해 기름 확보를 위한 주유 대란이 발생했다. 이때, 페이스북, 트위터 등과 같은 SNS을 통해 주유소의 기름 보유 상태, 연락처, 대기 시간 등을 공유하였다. 이러한 SNS는 특정 시간과 장소에서 발생하는 지역 이벤트들을 빠르게 전파할 수 있는 도구로 이용될 수 있으며 신뢰성 높은 집단 지성을 활용할 수 있다는 장점이 있다[4].

지역 이벤트는 통상 대규모 시위, 스포츠 게임과 같은 큰 규모의 행사에서부터 지역의 맥주 축제나 동네 슈퍼마켓의 할인 행사까지 주변에서 다양한 크기로 나타날 수 있다. 여러 사람들이 공통된 목적을 가지고 같은 시간과 장소에 모인 행사 또는 특정 지역에서 발생한 사회적으로 파급력이 큰 이슈 등을 '지역 이벤트'로 정의할 수 있다. 지역 이벤트를 빠르게 검출하여 해당 이벤트에 대한 정보를 얻음으로써 다양한 곳에 활용할 수 있다. 또한, 자연재해가 발생했을 때 실시간 이벤트 탐지를 통해 사람들에게 대피 알람을 보내 인명피해나 경제적인 피해를 줄일 수 있다. 따라서 SNS 데이터를 활용하여 지역 이벤트를 검출하는 연구가 필요하다.

소셜 네트워크 데이터 분석을 통해서 다양한 방법을 통해 이벤트를 검출하고자 하는 연구가 진행되고 있다. 이벤트 검출에 있어서 사람들이 직접적으로 작성한 글에는 다양한 정보를 포함하고 있다. 따라서 게시글이나 해시태그를 분석하여 키워드를 추출하고 이를 통해 키워드의 빈도, 중요도를 고려하여 이벤트를 검출하는 기법이 제안되었다[5, 6, 7, 8, 9]. 텍스트 기반 이벤트 검출 연구는 주로 [5]와 같이 TF-IDF 알고리즘을 이용하여 키워드를 추출하는 기법을 사용하였으나 최근에는 Word2Vec와 같은 기계학습 알고리즘을 사용하기도 한다[9]. 소셜 네트워크 데이터에서 추출한 키워드를 그래프로 생성하여 다양한 방법으로 클러스터링 하여 이벤트를 검출하는 연구가 제안되었다[10, 11].

최근 소셜 네트워크 서비스에서 제공하는 기능인 Geo-tag 정보를 이용한 지역 기반 이벤트 검출 기법이 연구되고 있다[12, 13, 14]. [13]에서는 게시글에 포함되어 있는 지리적 정보와 타임스탬프를 이용하여 지역 이벤트를 검출하는 기법을 제안하였다. [14]에서는 Geo-tag를 활용한 지역 기반 이벤트 검출 기법을 제안하였다. 그러나 [13]에서 제안된 기법은 텍스트 기반 이벤트 검출 기법을 수행하기 때문에 게시글이 짧은 경우 획득할 수 있는 정보가 적고 지역 이벤트와 관련된 키워드 추출하는 과정에서 노이즈가 커져 정확도가 떨어지는 한계점을 가지고 있다. 또한, 실제 소셜 네트워크 데이터의 대부분이 Geo-tag를 가지고 있지 않은 데이터이기 때문에 [14]에서 제안된 Geo-tag만을 사용한 이벤트 검출 기법은 정확도가 저하되는 문제점이 있다.

본 논문에서는 소셜 네트워크 환경에서 연관 문서 분석을 통한 지역 이벤트 검출 기법을 제안한다. 제안하는 기법은 소셜 네트워크 서비스 데이터에서 키워드를 추출하여 키워드 기반 그래프를 생성하고 정점과 간선에 소셜 네트워크 특성을 반영하여 가중치를 부여한다. Geo-tag 정보와 더불어 지리 정보 사

전을 생성하여 생성된 키워드 그래프의 정점 중 지역 정보를 가지고 있는 정점을 지역 노드로 분류한다. 기존의 Geo-tag를 활용한 이벤트 검출 기법이 가지고 있는 실제 소셜 네트워크 데이터의 대부분은 Geo-tag가 없다는 한계점을 해결하기 위하여 지리 정보 사전을 사용한다. 지리를 대표하는 명사에 맵핑되는 위치 정보를 가지고 있는 지리 정보 사전을 사용하여 부족한 지리적 정보를 보충한다. 키워드 그래프를 가중치에 따라 클러스터링을 수행한 후 연관 문서 분석 과정을 통해 제안하는 기법의 정확도를 향상시킨다. 클러스터 내부와 외부 간선 가중치를 이용하여 클러스터를 병합 또는 분리함으로써 지역 이벤트를 추출하여 결과를 도출한다. 다양한 성능평가를 통해 제안하는 기법이 기존 지역 이벤트 검출 기법에 비해 성능이 우수함을 보인다.

본 논문의 나머지 구성은 다음과 같다. 2장은 관련 연구에 대하여 기술한다. 3장에서는 제안하는 지역 이벤트 검출 기법의 구조 및 처리 과정에 대해서 설명하고 4장에서는 제안하는 기법의 성능 평가에 대해 기술한다. 마지막으로 5장은 본 논문의 결론을 제시한다.

2. 관련 연구

소셜 네트워크 서비스의 활성화로 소셜 네트워크 데이터를 활용하여 이벤트를 검출하는 기법에 대한 연구가 활발하게 진행되고 있다. 그 중 사용자가 작성한 게시글에서 추출한 키워드를 활용하는 방법이 있다. [6]에서는 이벤트와 관련된 일련의 단어 쌍을 이용하여 이벤트를 검출하고 더 나아가 시간 분석을 통해 미래의 이벤트를 예측하는 기법을 제안하였다. 이벤트와 가장 연관되어 있는 단어를 찾기 위해서 자카드 유사도를 이용한다. 그리고 동일한 형태의 단어이지만 의미가 달라서 발생하는 동형이의어에 대한 노이즈를 줄이기 위해서 단어가 포함된 글의 문맥을 분석하여 같이 발생한 단어를 쌍으로 묶어서 분석한다. 예를 들어, ‘strike’ 라는 영어 단어는 같이 오는 단어에 의해서 다양한 뜻을 가질 수 있다. ‘baseball’과 함께 발생하면 야구의 규칙과 연관된 뜻을 가지고, ‘lightning’과 연관되면 ‘번개가 치다’라는 뜻을 가지게 된다. 텍스트 기반 이벤트 검출 기법은 소셜 네트워크 서비스의 특성상 대부분이 비격식체를 사용하고 텍스트에 유행어, 특수문자 등의 비문이 다수 포함되어 있다. 따라서 데이터를 분석 할 때 노이즈가 크게 나타나 정확한 이벤트 검출에 어려움이 있다.

텍스트 마이닝만 사용하여 이벤트 검출에 사용한다면 데이터의 양이 많아질수록 시간이 오래 걸리고 노이즈가 커져 정확도가 떨어지는 단점이 있다. 데이터의 크기가 크고, 단어들 간의 상관관계를 보다 효과적으로 분석하기 위해서 [10]에서는 소셜 네트워크 데이터를 활용하여 EventGraph라는 방향성 키워드 그래프를 생성하여 이벤트를 검출하는 기법을 제안하였다. 그래프의 노드는 단어들로 구성하고 동시에 출현한 단어들에 대해서 간선으로 이어준다. 그리고 두 단어의 상관관계를 계산하여 간선에 가중치를 부여한다. 그러나 사용자들이 남긴 게시글만을 이용해서는 지역에 대한 정보를 특정하기 힘들기 때문에 이벤트 검출의 정확도가 떨어지는 문제점이 있다.

대부분의 이벤트는 특정 지역에서 특정 시간에 발생하기 때문에 지역과 밀접한 연관을 가지고 있다. 그리고 소셜 네트워크 서비스들은 사용자들이 게시글에 원하는 지역을 태그 하여 다른 사용자들에게 자신의 위치 또는 게시글과 연관된 위치를 알려주는 Geo-Tag 기능을 제공하고 있다. [13]에서는 소셜

네트워크 서비스에서 제공하는 Geo-Tag 기능을 활용하여 지역 기반의 이벤트를 검출하는 기법을 제안한다. 지도 인터페이스를 활용하여 지도 공간을 동일한 크기의 정사각형으로 나누고 이를 타일로 정의한다. 각각의 타일은 해당하는 트위터 데이터에서 추출한 시공간 정보를 가지고 있다. 그리고 STExNMF 기법[15]을 사용하여 키워드를 추출하여 검출한 이벤트에 대하여 시각화하였다. Geo-Tag 기능을 제공하는 소셜 네트워크 서비스 중 하나인 트위터의 데이터들 중 Geo-Tag를 가지고 있는 트윗은 약 1% 밖에 되지 않는다[16]. 따라서 Geo-Tag가 있는 데이터만 활용하여 이벤트를 검출하는 지역 기반 이벤트 검출 기법은 소셜 네트워크 서비스 데이터의 대부분인 Geo-Tag가 없는 데이터를 고려하지 않기 때문에 정확도가 떨어지는 한계점을 가지고 있다.

3. 제안하는 지역 이벤트 검출 기법

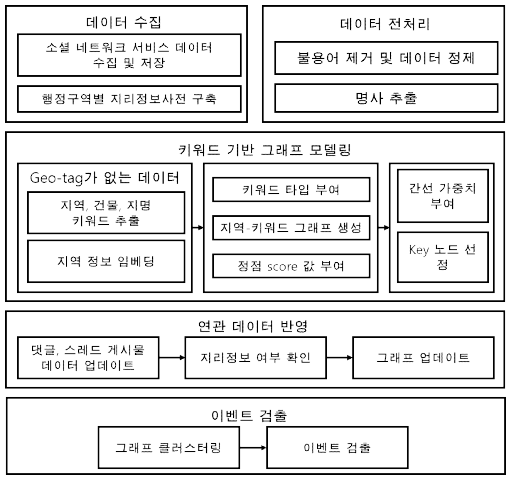
3.1 제안하는 기법의 구조

기존의 소셜 네트워크 환경에서의 이벤트 검출 기법은 게시글만을 분석하는 텍스트 마이닝을 사용하여 정확도가 떨어지거나 이미 태그 되어 있는 지리적 정보만을 이용하였다. 그러나 기존 기법들이 활용하는 Geo-Tag가 존재하는 데이터는 전체 데이터의 약 1%밖에 안 되기 때문에 나머지 데이터의 대부분을 차지하는 Geo-Tag가 없는 데이터에서도 지리적인 정보를 추출하여 이벤트 검출에 사용할 필요가 있다. 또한, 기존 이벤트 검출 기법의 경우 추가적으로 발생한 정보를 반영하지 못하기 때문에 이벤트 검출의 정확성에 한계가 있다.

본 논문에서는 기존 이벤트 검출 기법의 한계점을 해결하기 위해서 연관 문서 분석을 통한 지역 이벤트 검출 기법을 제안한다. 전체 데이터에 대해서 Geo-Tag가 없는 데이터는 텍스트 마이닝 기법을 사용하여 지리적인 정보를 임베딩하여 사용하고 연관 문서가 지역 정보를 가지고 있으면 이를 활용하여 이벤트 검출에 사용한다. 연관 문서란 소셜 네트워크 서비스에서 제공하는 댓글과 스레드 기능을 의미한다. 사용자들은 본인 또는 다른 사람들이 남긴 게시물에 댓글을 작성하여 자신의 의견을 표출한다. 그리고 소셜 네트워크 서비스 중 트위터의 경우 게시글인 트윗을 한 개 발행 할 때 140자 이내로 작성해야 하는 제한이 있다. 트위터를 이용하여 2개 이상의 트윗을 올려야 하는 경우 스레드라는 기능을 사용하여 글을 작성할 수 있다. 스레드는 하나의 주제와 관련된 여러 개의 트윗을 연결하여 조회할 수 있고 게시물에 대한 추가적인 정보나 업데이트 된 정보를 게시할 때 사용한다.

그림 1은 제안하는 지역 이벤트 검출 기법의 전체 시스템 구조를 나타낸다. 우선 데이터 수집 모듈에서는 소셜 네트워크 서비스 데이터를 수집하여 저장하고 행정 구역별 지리 정보 사전을 구축한다. 그리고 데이터 전처리 과정을 통해서 소셜 네트워크 raw data로부터 불용어를 제거하고 명사를 추출한다. 소셜 네트워크 서비스 데이터에는 특수문자, 줄임말, 은어 등의 비격식체를 사용하며 비문 표현이 많기 때문에 전처리 과정을 통해서 데이터를 정제하는 과정이 필요하다. 정제된 데이터를 이용하여 키워드 기반 그래프를 모델링하는 과정을 거친다. 키워드 기반 그래프 모델링 단계에서는 위치 정보인 Geo-Tag가 없는 데이터에 대해서 지리 정보 사전을 이용하여 지역에 대한 정보를 임베딩하고 키워드 그래프를 생성한 뒤 각 정점과 간선에 가중치를 부여한 뒤 중심 노드가 되는 키 노드(key node)를 선

정한다. 연관 데이터 반영 모듈에서는 타임 윈도우 내의 연관 문서를 분석하여 그래프를 업데이트 하는 과정을 거친다. 마지막으로 그래프 클러스터링을 통해 소셜 네트워크 데이터로부터 최종적으로 지역 이벤트를 검출한다.



[그림 1] 제안하는 지역 이벤트 검출 기법의 시스템 구조

3.2 데이터 수집 및 전처리

소셜 네트워크 서비스 환경에서 이벤트를 검출하기 위해서는 온라인에서 사용자들이 생성한 데이터를 수집해야한다. 그리고 지역 이벤트 검출에 필요한 데이터를 수집하고 그대로 사용하면 불필요한 정보가 많이 포함되어 정확하지 않은 이벤트가 검출될 수 있으므로 데이터를 전처리하는 과정이 필요하다. 제안하는 기법에서는 소셜 네트워크 데이터를 수집하고 지리 정보 사전을 구축하며 전처리 단계에서는 불용어 제거 및 데이터 정제 과정을 통해 명사를 추출한다. 데이터 수집 단계에서는 타임 윈도우 내에 발생한 소셜 네트워크 데이터인 게시물과 댓글, 리트윗 그리고 연관 게시물을 수집한다. 본 논문에서 제안하는 기법은 Geo-Tag가 없는 데이터에 대해서 텍스트 마이닝을 수행하여 게시글에 지리적인 정보가 포함되어 있으면 해당 지리정보를 임베딩하여 활용한다. 따라서 텍스트 마이닝을 수행하기 위해서 지명, 지역에 관한 정보가 있는 지리 정보 사전을 구축할 필요가 있다.

지리 정보 사전은 국가 행정상의 목적에 따라 나눈 행정구역을 기준으로 정의한다. 현재 국내 행정구역은 특별시, 광역시, 도, 시, 군, 구, 읍, 면, 동, 리로 나누어져 있으며 이를 크게 3개로 분류한다. 특별시, 광역시, 도를 상위 행정 구역으로, 시, 군, 구는 중위 행정 구역 그리고 마지막으로 읍, 면, 동, 리는 하위 행정 구역으로 나눈다. 하위 행정 구역으로 갈수록 구역의 범위가 좁아지고 하위 행정 구역을 이용하여 상위 행정 구역을 알 수 있도록 계층화 되어 있는 특징이 있다. 예를 들어, '용산구'는 중위 행정 구역이지만 계층화된 지리 정보 사전을 이용하면 상위 행정 구역은 '서울특별시'라는 것을 알 수 있다. 이렇게 정의된 지리 정보 사전을 이용하면 지리정보를 풍부하게 만들 수 있는 장점이 있다.

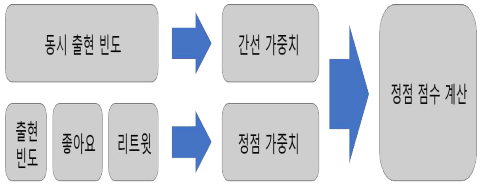
소셜 네트워크 서비스는 사용자들이 자신의 의견을 자유롭게 표현 할 수 있는 공간이기 때문에 소셜 네트워크 데이터를 수집하게 되면 필요한 정보와 불필요한 정보가 섞여 있다. 이러한 데이터를 그대로 사용하게 된다면 이벤트를 검출하는데 있어서 이상치나 무작위의 오류가 발생하는 노이즈가 커질 수 있고 정확도가 떨어진다. 따라서 노이즈를 줄이고 정확도를 높이기 위해서 데이터를 전처리 해주는 과정이 필요하다. 수집한 소셜 네트워크 데이터는 특수 문자와 줄임말, 사진, 링크 등이 포함되어 있기 때문에 이를 제거해주어야 한다. 그리고 데이터 전처리 과정에서 Geo-Tag가 포함되어 있는 게시글의 경우 Geo-Tag 정보를 텍스트와 별도로 저장하여

추후 키워드 그래프를 생성할 때 활용한다. 그리고 제안하는 기법에서는 키워드 그래프를 생성하여 이벤트 검출에 활용하므로 불용어 제거를 거친 데이터에서 형태소 분석을 실행하여 명사를 추출하여 사용한다.

3.3 키워드 기반 그래프 모델링

소셜 네트워크 데이터 전처리 과정을 거치면 키워드 집합이 생성된다. 키워드 집합만으로는 현재 이슈가 되고 있는 이벤트를 알기 어렵고 불필요한 정보들을 많이 포함하고 있다. 따라서 각 단어의 중요도와 언급량과 같은 화제성을 고려한 지역 이벤트 검출을 위해 키워드 기반 그래프 모델링을 수행한다. 키워드 기반 그래프는 간선과 정점의 가중치 부여를 통해 키워드의 유사도와 중요도를 파악할 수 있다.

그림 2는 동시 출현빈도와 좋아요, 리트윗과 같은 소셜 네트워크에서 사용자들의 명시적인 관심을 반영하여 가중치를 부여하고 정점 점수를 계산하는 과정을 나타낸다. Geo-Tag가 있는 데이터에 대해서는 Geo-Tag의 정보를 그대로 이용하여 키워드 그래프 생성 시 지역 노드로 사용하고, Geo-Tag가 없는 데이터에 대해서는 지리 정보 사전을 이용한 텍스트 마이닝을 통해서 추출한 지리정보를 지역 노드로 구분하는 방식의 임베딩을 사용한다. 그리고 동시에 출현한 단어들을 키워드 기반 그래프로 생성하고 소셜 네트워크 서비스의 특성을 반영하여 그래프의 각 정점과 간선에 가중치를 부여해준다. 그리고 키워드 그래프에서 중심이 되는 키 노드를 선정하여 키워드 기반 그래프 모델링을 수행한다.



[그림 2] 가중치 부여 및 정점 점수 계산 과정

생성된 키워드 그래프를 이벤트 검출 목적에 맞게 사용하기 위해 각각의 정점과 간선들에 대해 가중치를 부여한다. 이를 통해 단어 사이의 유사도와 좋아요와 리트윗과 같은 소셜 네트워크 특성에 따른 단어의 중요도를 알 수 있다. 동시 출현한 단어를 기반으로 연결된 간선의 가중치는 두 단어 간의 유사도를 나타낸다. 예를 들어, '고양이'와 '반려동물'이라는 단어는 두 단어 사이에 연관이 있기 때문에 가중치를 높게 주어야 하고, '고양이'와 '달력' 같은 단어는 가중치를 낮게 주어야한다. 따라서 간선에는 단어의 동시 출현 빈도를 사용하여 가중치를 부여한다. 동시 출현 빈도는 특정 단어를 기준으로 사용자가 설정한 window의 크기에 따라 달라진다. 윈도우의 크기가 1이면 해당하는 단어의 바로 앞, 뒤 단어만 동시 출현 횟수에 포함시킨다. 그리고 동시 출현 횟수 값을 0에서 1사이의 값으로 정규화를 해준다. 수식 (1)은 간선의 가중치에 대해서 정규화를 하는 수식이다. 이때, w_{ij}는 정점 V_i와 V_j두 단어 사이의 동시 출현 빈도를 나타내고 $w_{\min}$, $w_{\max}$는 전체 그래프의 간선들 중에서 가장 가중치가 작은 간선과 큰 간선을 의미한다.

$$w_{ij} = \frac{w_{ij} - w_{\min}}{w_{\max} - w_{\min}} \quad (1)$$

타임 윈도우 t 시간 내에 발생한 키워드들의 정점에 대해서 모두 같은 가중치를 부여한다면 어떤 키

워드가 이벤트와 관련 있는지 알기 어렵다. 사람들이 많이 언급한 단어와 좋아요 또는 리트윗과 같이 관심을 명시적으로 표현한 게시글에 있는 단어일수록 중요한 단어일 가능성이 높으므로 이를 키워드 정점에 반영해주어야 한다. 제안하는 기법에서는 변형된 TF-IDF 알고리즘을 사용하여 각각의 정점에 가중치를 부여하여 점수를 계산한다. 수식 (2)의 $S(V_i)$는 단어의 중요도에 따라 계산된 키워드 i에 대한 정점 V_i의 점수를 나타낸다. 초기 정점 V_i는 모두 1로 초기화하고 시간 속성을 고려한 TF-IDF를 계산한다. 그리고 좋아요($like$) 수와 리트윗($retweet$) 수를 합한 뒤 log를 사용하여 곱해준다. 좋아요와 리트윗은 사람들이 클릭 한번으로 자신의 의견을 표현할 수 있는 수단이기 때문에 중요한 단어일수록 값이 커지게 된다. 따라서 이를 그대로 곱해주게 된다면 결과 값이 좋아요와 리트윗 수에 많은 영향을 받기 때문에 log를 사용하여 조절한다. tf_i와 $idf_{i,t}$는 각각 단어 i의 출현 빈도(TF : Term Frequency)와 역문서 빈도(IDF : Inverse Document Frequency)를 나타낸다. 역문서 빈도는 현재의 타임윈도우 t 시간 값과 바로 이전 시간의 값의 비율을 사용한다.

$$S(V_i) = V_i * tf_i * \frac{idf_{i,t}}{idf_{i,t-1}} * \log(like_i + retweet_i) \quad (2)$$

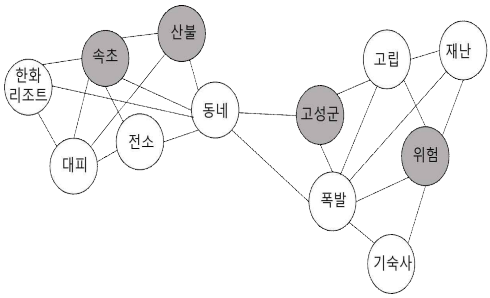
생성한 키워드 그래프 정점의 점수를 계산하면 각 키워드의 중요도를 알 수 있다. 점수가 높은 정점일수록 이벤트 검출에 있어서 핵심 단어가 된다. 각 정점에 시간 속성을 고려한 TF-IDF를 이용하여 가중치를 부여하고 TextRank 알고리즘을 사용하여 핵심 단어를 추출하기 위한 점수를 계산하여 부여한다. 수식 (3)의 $TR(V_i)$는 TextRank를 이용하여 각 정점에 부여되는 점수를 나타낸다. 초기 계산에 사용되는 $TR(V_i)$는 수식 (2)의 $S(V_i)$ 결과를 이용한다. V_i는 지역 타입 정점과 일반 타입 정점을 모두 포함하는 정점 집합 중 점수를 계산하고자 하는 정점이다. V_j는 V_i와 간선으로 연결된 키워드 정점을 나타낸다. w_{ij}는 V_i의 간선들 중 V_j와 연결된 간선의 가중치를 나타내며 w_{jk}는 V_j와 연결된 간선의 가중치를 나타낸다. d(Damping Factor)를 이용하여 랜덤 확률 변수를 조정해 줄 수 있으며 본 논문에서는 일반적으로 사용하는 0.85를 채택하여 사용하였다.

$$TR(V_i) = (1-d) + d * \sum_{j \in V_i} \frac{w_{ij}}{\sum_{k \in V_j} w_{jk}} TR(V_j) \quad (3)$$

핵심단어를 기준으로 그래프 클러스터링을 통해 이벤트 검출을 하면 해당 이벤트와 연관 있는 단어들을 알기 쉽고 전체 키워드 정점에 대해서 클러스터링 하는 것보다 처리 시간을 줄일 수 있다. 따라서 키워드 그래프 정점의 점수 값을 이용하여 핵심 단어를 의미하는 키 노드로 선정해준다. 키 노드는 정점의 점수 값을 내림차순으로 정렬하여 사용자 설정에 따라 Top-k개를 선정한다. 그리고 지역 이벤트 검출에 있어서 지역 정보는 중요한 역할을 하므로 Geo-Tag 또는 지리 정보 사전을 이용하여 지역 정점으로 라벨링한 정점도 키 노드에 포함한다.

그림 3은 키워드 그래프 생성 및 키 노드 선정 과정을 나타낸다. 전처리한 데이터를 이용하여 명사집합들에 대해서 키워드 기반 그래프를 생성하면 (a)와 같은 결과가 나타난다. '속초'와 '고성군'은 지역 정점으로 라벨링이 되어 있는 상태이다. 그리고 각 정점과 간선의 가중치에 따라서 (b)와 같이 정점 점수를 계산하여 내림차순으로 정렬하고 상위 4개의

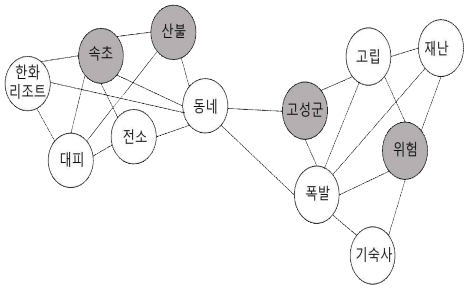
결과 값인 '산불', '속초', '고성군', '위험'을 키 노드로 선정한다. 따라서 (c)에서는 이러한 결과가 반영되어 키 노드가 선정된 그래프가 생성된다.



(a) 키워드 그래프 생성 및 지리 정보 임베딩

순위	키워드	점수
1	산불	31.36
2	속초	17.67
3	고성군	17.10
4	위험	15.32
5	대피	14.84
...		
12	고립	9.91

(b) 정점 점수 계산



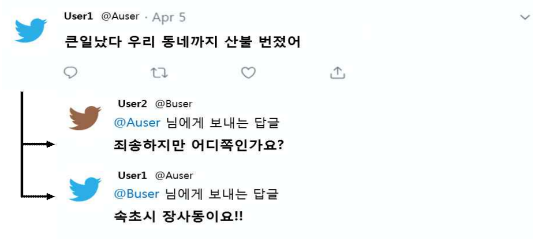
(c) 키 노드 선정

[그림 3] 키워드 그래프 생성 및 키 노드 선정 과정

3.4 연관 문서 분석

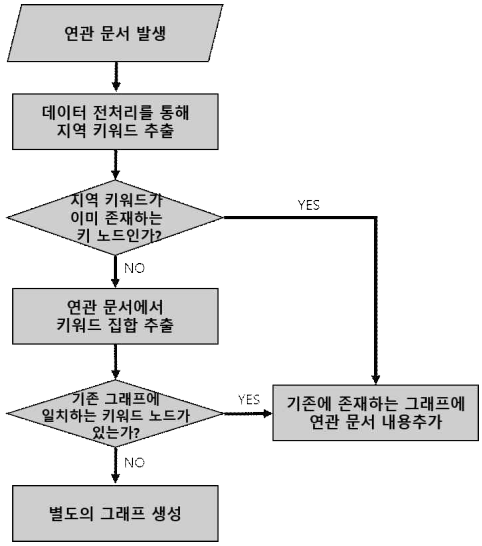
소셜 네트워크 서비스 데이터를 이용한 이벤트 검출에 연관 문서를 분석하여 고려할 필요가 있다. 지역 이벤트가 발생하고 정보 전달에 있어서 처음에는 정보가 부족했지만 소셜 네트워크 서비스를 이용하는 사용자들은 다른 사용자의 게시글이나 본인의 게시글에 유용한 정보를 연관 문서로 추가할 수 있다. 소셜 네트워크 서비스의 특성상 공개 설정 여부에 따라서 친구 관계뿐만 아니라 다수가 볼 수 있도록 게시할 수 있기 때문에 해당 게시글에 대해서 자신의 의견을 표출하거나 알고 있는 정보를 추가할 때 사용자들은 댓글 또는 스레드와 같은 연관 문서를 이용한다. 특히, 해당 정보가 지역과 관련이 되어 있는 경우 지역 이벤트 검출에 있어서 중요한 정보가 되기 때문에 활용하여 정확도를 향상시킬 수 있다.

그림 4는 연관 문서 예시를 나타낸다. User1이 이벤트에 대한 정보를 트윗으로 게시하였지만 해당 게시글에는 지리적인 정보가 포함되어 있지 않아서 지역 이벤트로 검출하기 어렵다. 그러나 User2의 댓글에 대한 답변으로 User1이 댓글을 통해서 지리정보가 추가하였다. 기존의 기법에서는 연관 문서에 포함되어 있는 정보들이 무시되었지만 제안하는 기법에서는 연관 문서 분석을 통해서 추가적으로 제공되는 정보를 그래프에 반영한다.



[그림 4] 연관 문서 예시

그림 5는 연관 문서 분석 내용을 그래프에 추가하기 위한 순서도를 나타낸다. 키워드 기반 그래프가 구축된 상태에서 연관 문서가 발생하면 데이터 전처리를 통해서 해당 연관 문서에 지역 키워드가 있는지 검사한다. 제안하는 기법에서는 지역 정보를 가지고 있는 연관 문서만을 사용한다. 이벤트는 주로 특정 지역, 장소와 밀접한 연관이 있기 때문에 지역 정보는 지역 이벤트 검출에서 중요한 역할을 한다. 모든 연관 문서에 대해서 그래프에 추가하려면 처리 시간이 증가되고 필요없는 정보가 다수 포함되어 정확도가 떨어질 수 있기 때문에 지역 키워드 여부를 통해서 해당하는 연관 문서만 분석 과정을 수행한다. 만약 지역 키워드가 존재하면 해당 키워드와 이미 구축되어 있는 키 노드를 비교한다. 이전 단계인 키 노드 선정에서 키 노드는 지역 정점을 반드시 포함하므로 일치하는 키 노드가 있으면 기존에 존재하는 그래프에 연관 문서 내용을 추가하여 부족한 정보를 보완한다. 그러나 해당 연관 문서에서 추출된 지역 키워드가 기존 그래프에 존재하지 않는 정점이라면 나머지 키워드를 기존의 그래프와 비교하여 일치하는 정보가 있는지 여부를 판단해야 한다. 만약 일치하는 정보가 있다면 기존에 존재하는 그래프에는 지역 정보가 없었지만 연관 문서 정보를 반영하여 정보를 추가하고 그렇지 않으면 새롭게 검출된 이벤트일 가능성이 높기 때문에 별도의 그래프를 생성하여 연관 문서를 분석한 정보를 추가하고 추가된 정점과 간선에도 가중치를 부여한다.



[그림 5] 연관 문서 분석 과정

3.5 그래프 클러스터링 및 이벤트 검출

연관 문서 분석을 마친 키워드 그래프를 통해서는 그래프의 연결 관계들 때문에 이벤트를 쉽게 파악하기 힘들다. 제안하는 기법에서는 키 노드와 간선의 가중치를 이용하여 그래프를 클러스터링하고 각 클러스터 간의 관계를 기반으로 유사한 이벤트를 병합하거나 별도의 지역 이벤트로 검출할 수 있다. 키워드 그래프의 각각 간선에는 키워드 간의 유사도에 따라 가중치가 부여되어 있다. 가중치가 높을수록 연관성이 높은 단어를 나타낸다. 이를 반영하여 키 노드를 기준으로 연결된 간선 중 가중치 α보다 크거나 같은 간선으로 연결된 노드를 하나의 클러스터로 묶어준다. 이때, 기준이 되는 가중치 α는 그래프 내 노드들의 클러스터링 정도를 의미하는 네트워크 모듈성을 계산하여 선정하였다.

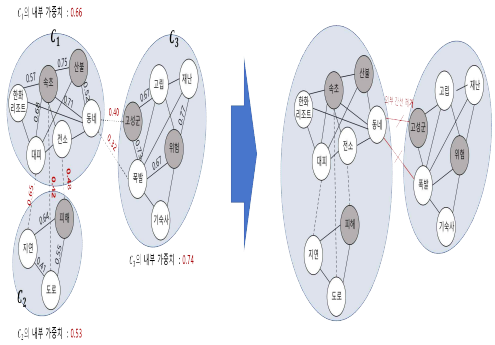
식 (4)는 네트워크 모듈성 NM을 계산하는 수식을 나타낸다. m은 전체 간선 수, n은 전체 노드 수

를 의미하며 w_{ij}는 정점 V_i와 V_j사이를 연결하는 간선의 가중치를 나타낸다. k_i는 V_i와 연결된 모든 간선의 가중치 합이고 $\delta(c_i, c_j)$는 V_i와 V_j가 같은 클러스터에 있으면 1, 아니면 0을 반환하는 불리언 함수이다.

$$NM = \frac{1}{2m}\sum_{ij}^{n}\left[w_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (4)$$

사용자가 원하는 최종 k개의 이벤트를 검출하기 위해서는 클러스터 사이의 연관 관계를 통하여 서로 관련이 없는 클러스터라면 간선을 끊어 독립된 클러스터로 생성하거나 두 개의 클러스터가 연관이 있다면 병합하여 하나의 클러스터로 만드는 과정이 필요하다. 제안하는 기법에서는 간선 가중치를 이용하여 결과로 도출된 각각의 독립적인 클러스터를 하나의 이벤트로 검출하여 제공한다.

그림 6은 제안하는 기법의 이벤트 검출 과정 예시를 나타낸다. 왼쪽은 키 노드와 간선의 가중치를 이용하여 키워드 그래프가 클러스터링 과정을 거쳐 3개의 클러스터로 나누어진 것을 알 수 있다. 클러스터로 묶인 내부 정점 사이에 연결된 간선의 평균을 내부 가중치로 정의하고, 서로 다른 두 클러스터 사이에 연결된 간선의 가중치 합을 외부 가중치로 정의한다. C_1과 C_2의 내부 가중치는 각각 0.66과 0.53이고 외부 가중치는 외부로 연결된 간선의 가중치 합인 1.55이다. 두 개의 클러스터 각각의 내부 가중치 합과 외부 가중치의 값을 비교한다. 두 클러스터의 내부 가중치의 합은 1.19로 외부 가중치의 값이 더 크다. 따라서 두 클러스터는 서로 밀접한 연관이 있기 때문에 클러스터를 하나로 병합해준다. 이와 같이 C_1와 C_3을 비교 했을 때 외부 가중치가 내부 가중치의 합보다 작기 때문에 연관성이 떨어지는 각각의 독립된 이벤트일 가능성이 높다. 그러므로 서로 연결된 간선을 제거하여 각각의 독립된 클러스터로 구성하여 이벤트 검출에 사용한다.



[그림 6] 이벤트 검출 과정 예시

4. 성능 평가

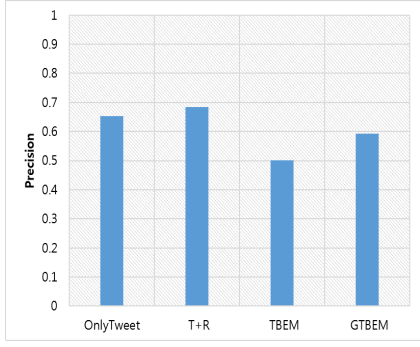
제안하는 지역 이벤트 검출 기법의 우수성을 입증하기 위하여 성능 평가를 수행하였다. 표 1은 성능 평가를 진행한 실험 환경을 나타낸다. 성능 평가는 Intel(R) Core(TM) i5-4440 CPU 3.10GHz 프로세서, 16GB 메모리를 가지는 시스템에서 Windows 7 Ultimate K 64 비트 운영 체제 환경에서 Python을 이용하여 구현하였다. 성능 평가를 수행하기 위해 사용된 데이터는 Twitter Scraper API를 사용하여 대표적인 소셜 네트워크 서비스인 트위터 데이터를 수집하였다. 수집 항목으로는 트윗 게시글, 게시한 시간, 리트윗 수, 좋아요 수 그리고 댓글 및 스레드가 있다. 2019년 4월 1일부터 2019년 4월 30일까지 수집한 982,114건의 트윗과 117,942건의 연관 문서 데이터를 사용하였다.

[표 1] 성능 평가 환경

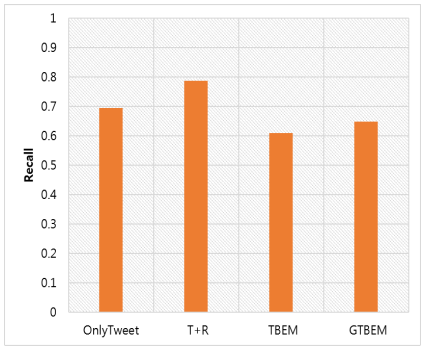
구 분	내 용
프로세서	Intel(R) Core(TM) i5-4400 CPU 3.40GHz
메모리	16GB
운영체제	Windows 7 Ultimate K 64 bit
프로그램 언어	Python

기존의 이벤트 검출 기법과 성능을 비교하기 위해서 두 가지 기법을 선정하여 평가를 수행하였다. 제안하는 기법은 연관 문서 분석을 반영 여부에 따라서 두 가지로 나누었다. 그리고 성능 비교를 위하여 키워드 쌍을 이용하여 텍스트 기반으로 이벤트를 검출하는 [6]에서 제안된 기법을 선정하였다. 또한, Geo-Tag 데이터를 활용하여 이벤트를 검출하는 기법들 중 [14]에서 제안하는 기법은 키워드 그래프를 구축하여 이벤트를 검출하는 단계가 포함되어 있어 [12]보다 비교 평가에 적합하기 때문에 선정하였다. 기존 기법과 성능 비교를 위한 평가 지표로는 정밀도(precision), 재현율(recall), F-measure를 사용하였다. 본 논문에서 연관 문서를 반영하지 않고 트윗만 사용한 기법을 OnlyTweet이라 정의하고 제안하는 기법인 트윗과 연관 문서를 모두 반영한 것을 T+R, 텍스트 기반의 이벤트 검출 기법인 [6]의 기법을 TBEM, Geo-Tag 데이터를 이용한 키워드 그래프 이벤트 검출 기법인 [14]의 기법을 GTBEM으로 정의한다.

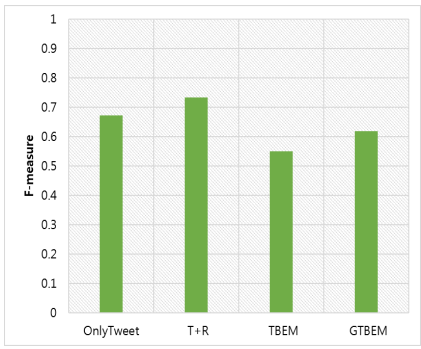
그림 7은 기존 기법과 제안하는 기법의 정밀도에 대한 성능 비교 결과를 보여준다. 제안하는 기법은 정밀도에서 기존 기법보다 약 14% 높은 결과가 측정되었다. TBEM[6]의 경우 텍스트 마이닝 기반으로 지역 이벤트 검출에 적합하지 않기 때문에 전체적으로 낮은 정확도를 보였고, GTBEM[14]의 경우 Geo-Tag 데이터만 사용하여 제안하는 기법에 비해 상대적으로 정보가 부족하여 정밀도가 낮게 측정되었다. 그림 8은 재현율에 대한 성능 비교이며 기존 기법에 비해 약 16% 높은 결과가 나타났다. 기존의 기법은 텍스트 마이닝에 의존하여 지역 이벤트와 관련이 없어도 단순한 언급량 증가만으로도 이벤트로 검출 될 수 있다. 하지만 제안하는 기법은 연관 문서 분석을 통해 정보량을 증가하여 지역 이벤트를 검출하기 때문에 기존 기법과 비교하여 우수한 성능을 보인다. 그림 9는 기존 기법과 제안하는 기법의 F-measure 측정 결과를 나타내며 약 15% 높은 결과로 제안하는 기법이 기존 기법과 비교하여 우수함을 확인하였다. 제안하는 기법은 Geo-Tag와 함께 지리정보가 부여되지 않은 데이터에 대해서도 지리정보 사전을 이용하여 지역 정보를 임베딩하고, 연관 문서 분석을 통해 부족한 정보를 추가하여 이벤트를 검출하는 방법으로 기존 기법과 비교하여 정밀도, 재현율, F-measure 3가지의 지표에서 모두 성능 향상을 보였다.



[그림 7] 정밀도 성능 비교

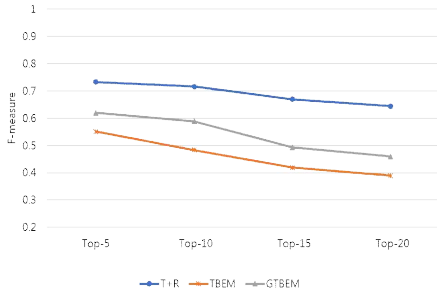


[그림 8] 재현율 성능 비교



[그림 9] 기존 기법과 F-measure 성능 비교

그림 10은 제안하는 기법과 기존 기법의 지역 이벤트 검출 개수에 따른 F-measure 값의 변화를 비교하여 나타낸 결과이다. 지역 이벤트 검출 개수를 5개씩 늘려서 실험 평가를 진행하였다. 제안하는 기법과 기존 기법 모두 이벤트의 개수를 5개로 했을 때 가장 높은 F-measure 값을 가졌으며 이벤트의 개수를 늘릴수록 성능이 저하되는 변화를 보였다. 이는 실제로는 이벤트가 아니지만 검출해야 하는 이벤트의 개수가 커지면서 이벤트로 포함되어 정밀도가 낮아지는 경우가 발생하기 때문이다. 또한, 기존 기법인 GTBEM[14]의 경우 Geo-Tag 기반의 이벤트 검출로 지역과 연관이 없는 이벤트는 검출하지 못하기 때문에 이벤트 개수의 증가에 따라 성능이 급격하게 낮아지는 것을 알 수 있다.



[그림 10] 지역 이벤트 검출 개수에 따른 F-measure 비교

5. 결 론

본 논문에서는 소셜 네트워크 환경에서 연관 문서를 분석을 반영한 지역 이벤트 검출 기법을 제안하였다. 제안하는 기법에서는 수집한 소셜 네트워크 서비스 데이터 중 Geo-Tag가 없는 데이터에 대해서 지리 정보 사전을 통하여 지리적인 정보를 추출하여 이벤트 검출에 사용하였다. 또한, 소셜 네트워크에서 많이 활용되는 댓글과 연관 문서 분석을 통해서 필요한 정보를 키워드 그래프에 추가하였다. 성능 평가를 수행한 결과 제안하는 기법은 기존 기법보다 정밀도, 재현율, F-measure에서 각각 14%, 16%, 15% 향상된 성능을 보였다. 향후 연구에서는 미리 구축되어 있는 지리 정보 사전에 의지하는 것이 아니라 지도 API를 사용하여 건물의 이름과 같이 더욱 구체적이고 정확한 연구를 수행할 계획이며 실시간 처리를 통한 이벤트 검출을 구현하여 실제 재난 알림 시스템 등에 적용 가능한 형태의 연구를 수행할 예정이다.

참고 문헌

[1] Facebook. https://www.facebook.com

[2] Instagram. https://www.instagram.com

[3] Twitter. https://twitter.com

[4] S. Choi and B. Bae, “The Sensing Model of Disaster Issues from Social Bigdata,” Journal of KISE, Vol. 20, No. 5, pp. 286-290, 2014. (in Korean)

[5] Z. Zhu, J. Liang, D. Li, H. Yu, and G. Liu, “Hot Topic Detection Based on a Refined TF-IDF Algorithm,” IEEE Access, Vol. 7, pp. 26996-27007, 2019.

[6] A. H. Hossny, and L. Mitchell, “Event detection in Twitter: A keyword volume approach,” Proc. IEEE International Conference on Data Mining Workshops, pp. 1200-1208, 2018.

[7] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, “Discover Breaking Events with Popular Hashwords in Twitter,” Proc. ACM international conference on Information and knowledge management, pp. 1794-1798, 2012.

[8] S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose, “Spatio-Temporal and Events Based Analysis of Topic Popularity in Twitter,” Proc. ACM International Conference on Information and Knowledge Management, pp. 219-228, 2013.

[9] W. Cui, P. Wang, Y. Du, X. Chen, D. Guo, J. Li, and Y. Zhou, “An algorithm for event detection based on social media data,” Neurocomputing, Vol. 254, pp. 53-58, 2017.

[10] J. He, Y. Liu, and Y. Jia, “EventGraph Based Events Detection in Social Media,” Proc. International Conference of Pioneering Computer Scientists, Engineers and Educators, pp. 150-160, 2018.

[11] H. Genc, and B. Yilmaz, “Text-Based Event Detection: Deciphering Date Information Using Graph Embeddings,” Proc. International Conference on Big Data Analytics and Knowledge Discovery, pp. 266-278, 2019.

[12] C. Zhang, L. Liu, D. Lei, Q. Yuan, H. Zhuang, T. Hanratty, and J. Han, “Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams,” Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 595-604, 2017.

[13] M. Choi, S. Shin, J. Choi, S. Langevin, C. Bethune, P. Horne, and J. Choo, “Topicontiles: Tile-based spatio-temporal event analytics via exclusive topic modeling on social media,” Proc. CHI Conference on Human Factors in Computing Systems, p. 583, 2018.

[14] S. Zhang, Y. Cheng, and D. Ke, “Event-Radar: Real-time Local Event Detection System for Geo-Tagged Tweet Streams,” arXiv preprint, 2017.

[15] S. Shin, M. Choi, J. Choi, S. Langevin, C. Bethune, P. Horne, and J. Choo, “STExNMF: Spatio-Temporally exclusive topic discovery for anomalous event detection,” Proc. IEEE International Conference on Data Mining, pp. 435-444, 2017.

[16] H. Abdelhaq, C. Sengstock, and M. Gertz, “Eventweet: Online localized event detection from twitter,” VLDB Endowment, Vol. 6, No.12, pp. 1326-1329, 2013.

박 수 빈

2018년 한남대학교 산업경영공학과, 컴퓨터공학과 학사

2020년 충북대학교 빅데이터협동과정 석사

2020년~현재 충북대학교 빅데이터협동과정 박사과정

관심분야 : 빅데이터, 분산 처리, 소셜 네트워크, 상품 추천 등

최 도 진

2014년 한국교통대학교 컴퓨터공학과 학사

2016년 한국교통대학교 컴퓨터공학과 석사

2020년 충북대학교 정보통신공학과 박사

2020년 ~ 현재 충북대학교 정보통신공학과 박사후과정

관심분야: 연속 질의 처리, 그래프 스트림, 빅데이터 처리 등

복 경 수

1998년 충북대학교 수학과 학사

2000년 충북대학교 정보통신공학과 석사

2005년 충북대학교 정보통신공학과 박사

2005년~2008년 한국과학기술원 정보전자연구소 Postdoc

2008년~2011년 가인정보기술 연구소 차장

2011년~2019년 충북대학교 전자정보대학 정보통신공학부 초빙교수

2011년~현재 원광대학교 SW융합학과 조교수

유 재 수

1989년 전북대학교 컴퓨터공학과 학사

1991년 KAIST 전산학과 석사

1995년 KAIST 전산학과 박사

1995년~1996년 목포대학교 전산통계학과 전임강사

2009.3 ~ 2010.2 캘리포니아 주립대학 방문교수

1996년~현재 충북대학교 전자정보대학 정보통신공학부 교수

관심분야 : 데이터베이스 시스템, 멀티미디어 데이터베이스, 센서 네트워크, 바이오 인포메틱스, 빅데이터 처리 등