https://doi.org/10.5392/JKCA.2018.18.07.449

소셜 네트워크에서 사용자 관심도를 고려한 그래프 기반 이벤트 검출 기법

Graph-based Event Detection Scheme Considering User Interest in Social Networks

김이나*, 김민영**, 임종태**, 복경수**, 유재수**
충북대학교 빅데이터학과*, 충북대학교 정보통신공학과**

Ina Kim(inakim@chungbuk.ac.kr)*, Minyoung Kim(cystitis@chungbuk.ac.kr)**,
Jongtae Lim(jtlim@chungbuk.ac.kr)**, Kyoungsoo Bok(ksbok@chungbuk.ac.kr)**,
Jaesoo Yoo(yjs@chungbuk.ac.kr)**

요약

소셜 네트워크 서비스의 사용량이 증가함에 따라 오프라인에서 발생한 이벤트 정보가 더욱 빠르게 확산되고 있다. 이에 따라 소셜 데이터를 분석하여 이벤트를 검출하기 위한 연구들이 진행되고 있다. 본 논문에서는 소셜 네트워크 환경에서 사용자 관심도를 고려한 그래프 기반 이벤트 검출 기법을 제안한다. 제안하는 기법은 사용자들이 게시한 글을 분석하여 키워드 그래프를 구축한다. 사용자의 소셜 행위로부터 관심도를 계산하고 관심도의 변화를 고려하여 이벤트 판별에 이용한다. 따라서 의미 없이 반복 게시되어 이벤트로 검출된 결과를 제거하고 결과의 신뢰성을 향상시킬 수 있다. 제안하는 이벤트 검출 기법의 우수성을 입증하기 위해 다양한 성능평가를 수행한다.

■ 중심어 : | 이벤트 검출 | 소셜 네트워크 | 키워드 그래프 | 그래프 클러스터링 | 사용자 관심도 |

Abstract

As the usage of social network services increases, event information occurring offline is spreading more rapidly. Therefore, studies have been conducted to detect events by analyzing social data. In this paper, we propose a graph based event detection scheme considering user interest in social networks. The proposed scheme constructs a keyword graph by analyzing tweets posted by users. We calculates the interest measure from users' social activities and uses it to identify events by considering changes in interest. Therefore, it is possible to eliminate events that are repeatedly posted without meaning and improve the reliability of the results. We conduct various performance evaluations to demonstrate the superiority of the proposed event detection scheme.

■ keyword : | Event Detection | Social Network | Keyword Graph | Graph Clustering | User Interest |

* 본 연구는 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단(No. 2016R1A2B3007527), 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업(IITP-2018-2013-1-00881), 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단-차세대정보 · 컴퓨팅기술개발사업(No. NRF-2017M3C4A7069432), 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A5B8059946)

접수일자 : 2018년 06월 14일
수정일자 : 2018년 07월 03일
심사완료일 : 2018년 07월 03일
교신저자 : 유재수, e-mail : yjs@chungbuk.ac.kr

I. 서 론

모바일 기기의 보편화와 무선 인터넷의 보급으로 소셜 네트워크 서비스(SNS :Social Network Service)의 이용량이 증가하고 있다. 페이스북, 트위터, 인스타그램과 같은 소셜 네트워크는 사진 및 텍스트를 포함한 짧은 글을 통해 자신의 관심사나 활동을 공유할 수 있는 플랫폼을 제공한다. 특히 대중적으로 사용되고 있는 서비스인 트위터는 세계적으로 이용자가 많은 마이크로블로깅 서비스로 간단한 기능을 제공하고 있으며 광범위한 사용자 집단을 보유하고 있다. 2017년 한 해에만 트위터를 통해 22억 명의 사용자가 매일 5억 개 이상의 트윗을 게시하며 정보를 공유하였다[1]. 사용자들은 인적 관계를 통해 정보를 습득할 뿐만 아니라 '공유', '좋아요', '리트윗' 등의 소셜 행위를 통해 정보를 전파하는데 기여하고 있다. 사용자들은 시간과 공간의 제약이 없이 모바일 기기를 이용하여 SNS를 이용하기 때문에 실시간으로 생성되는 소셜 데이터의 양은 기하급수적으로 증가하고 있다. 이러한 배경으로 인해 SNS를 통해 전파되는 데이터의 양은 더욱 빠르게 증가하고 있다.

사용자들이 작성한 소셜 데이터를 분석을 통해 사회적인 현상을 보다 빠르게 탐지할 수 있다. 예를 들어, 2011년 3월 동일본에서 발생한 대지진 사건 당시 트위터를 활용하여 국민들이 피해 상황 및 대피소 정보를 공유하였고 효과적으로 지진에 대응하였다. 이와 같이 사용자가 관심을 가지는 이슈 사항이나 사회적으로 파급력이 큰 사건, 사고 등을 이벤트라고 정의한다. 이벤트 검출은 시간과 장소를 수반하여 다수의 사람들에게 관심을 끌고 있는 일에 대한 정보를 검출하는 것이 목적이다. 특히, 사용자들은 오프라인에서 경험한 사회적인 현상이나 상황에 대해 공유하는 성향이 있기 때문에 소셜 데이터에는 이벤트와 관련된 다양한 정보를 포함하고 있다. 그러나 대량의 소셜 데이터는 구조화 되어 있지 않고 불필요한 정보를 다수 포함하고 있기 때문에 사용자가 원하는 이벤트에 관한 정보를 얻는 것은 어렵다. 따라서 정제된 정보를 제공할 수 있도록 이벤트 정보를 검출하는 연구의 필요성이 대두되고 있다.

이벤트 검출에 대한 연구는 인기를 끄는 이벤트 주제 단어(Event Topic)를 검출하는 형태로 진행되어 왔다. 초기의 연구에서는 단어의 발생 빈도수가 큰 경우를 검출하거나 TF-IDF 알고리즘을 적용하여 단어 집합을 검출하는 기법을 제안하였다[2][3]. 하지만 데이터의 특성상 단어 발생의 빈도수는 다수의 사람들이 항상 자주 언급하는 단어일 수 있기 때문에 단어를 검출할 때는 이전 시간에 비해 얼마나 증가했는지에 대한 비율과 시간의 단위를 고려해야 적절한 결과를 도출할 수 있다. 이러한 단점이 개선된 기법들이 제안되었다[4][5]. 지정한 시간의 단위에 대해 단어 빈도수의 차이를 비율적으로 고려하고 시간 단위가 겹치게 설정하는 슬라이딩 윈도우를 사용함으로써 단어의 증가량을 파악할 수 있도록 제안하였다. 하지만 이벤트 자체를 검출하기 보다는 이벤트를 나타내는 주제 단어들을 결과로 나타내기 때문에 사용자는 도출된 결과를 보고 각 단어가 같은 이벤트를 의미하는지 서로 다른 이벤트를 의미하는지 알기 위해서 데이터를 살펴보거나 사용자 스스로 검색을 통해 판별해야 한다.

이전의 연구가 가졌던 단점을 보완하기 위해 검출된 키워드를 군집화하여 사용자에게 제공하는 연구들이 진행되었다[6-9]. [9]는 키워드 그래프를 구축하여 군집화를 통해 이벤트 그래프를 검출하는 기법을 제안하였다. 그러나 사용자의 개입으로 인해 정확도 측면에서 단점을 가진다. 또한, 소셜 네트워크에서 사용자들이 공감이나 동조의 표현을 나타내는 '좋아요', '리트윗', '공유' 등의 소셜 행위를 고려하지 않았기 때문에 신뢰도가 낮은 이벤트에 대하여 분별하지 못하는 문제점이 있다.

본 논문에서는 소셜 네트워크 환경에서 사용자 관심도를 고려한 이벤트 검출 기법을 제안한다. 제안하는 기법은 사용자들이 게시한 글을 분석하여 키워드 그래프를 구축한다. 정점과 간선의 매개 중심성을 이용하여 생성된 키워드 그래프의 필터링과 클러스터링을 수행한다. 기존의 그래프 기반 이벤트 검출 기법에서 고려하는 단어의 출현 빈도 수 뿐만 아니라 사용자들의 관심도의 변화량을 이벤트를 판별하는 과정에서 고려한다. 사용자들의 소셜 행위의 변화량을 이용해 관심도 변화량을 계산하고 이벤트 판별 단계에 통하여 결과를 도출한다.

본 논문의 구성은 다음과 같이 구성된다. Ⅱ장에서 기존의 이벤트 검출기법을 설명하고 한계점을 설명한다. Ⅲ장에서는 제안하는 기법의 처리 과정과 내용을 상세히 기술하고, Ⅳ장에서는 성능평가를 통해 제안하는 기법의 우수성을 입증한다. 마지막 Ⅴ장에서는 본 논문의 결론 및 향후 연구에 대해서 기술한다.

Ⅱ. 관련 연구

이벤트를 검출하기 위한 방법으로 사용자들이 게시한 글의 감정을 분석하는 방법이 있다. 단어의 긍·부정을 분류하거나 심리학적 이론을 도입하고 다양한 감정에 속하는 단어들을 이용하여 이벤트를 검출하는 기법들이 제안되었다[10][11]. [10]에서는 7가지 감정을 분류하여 단어와 느낌표나 물음표 같은 구두점을 학습시키는 방법을 이용하였다. 각 감정에 해당하는 단어나 표현의 지역별 발생 빈도를 수집하고 지역별로 감정의 수치를 모니터링한다. 감정의 상태가 이전 시간 단위의 수치보다 급격하게 증가하면 이벤트가 발생한 것으로 추정하고 결과를 검출한다. 검출된 감정, 지역, 시간과 함께 가장 인기 있는 대표 글이 결과로 제공된다. 하지만 오직 감정에 의지하기 때문에 한 지역에 같은 감정을 의미하는 다른 이벤트가 동시에 발생할 경우 이벤트를 각각 분리하여 검출할 수 없다. 같은 이벤트가 서로 상반되는 감정을 야기하여 중복되는 이벤트를 검출하는 문제도 있다. 또한, 사용자에게 이벤트에 대한 요약 정보를 충분히 제공하지 못하는 한계를 갖고 있다.

[11]에서는 단어의 출현의 역문서 빈도(IDF)를 이용하여 이벤트를 검출하는 기법을 제안하였다. 역문서 빈도 값은 TF-IDF알고리즘에서 특정 단어가 문서에 출현한 수의 역수를 나타낸다. 즉, 어떤 문서에서 자주 나타나지 않는 단어가 IDF값이 높다. 이벤트 검출 분야에서 TF-IDF를 적용할 경우 각 트윗을 문서로 가정하고 알고리즘을 적용하고 있다. [12]에서는 트위터에서 트윗을 수집하여 이모티콘, 불용어를 제거한다. 품사 분석기를 이용하여 단어 단위로 쪼갠 후, 명사를 추출하여 이벤트 후보 집합을 구성한다. 이벤트 후보 집합은 슬라이딩 윈도우 모델을 사용하여 사용자가 설정한 시간 단위로 반복하여 구성한다. 이벤트 후보 집합에 포함된 각 단어에 대하여 IDF 값을 계산하고 이전 시간 단위에 비해 증가한 비율을 이벤트 검출의 척도로 설정한다. 현재 시간에 발생한 모든 단어의 IDF 평균 변화율 보다 높으면 해당 단어의 이벤트가 발생한 것으로 판단한다. 결과적으로 이벤트 자체를 검출한다기보다는 이벤트 키워드들을 검출하는 방식이기 때문에 사용자가 나열된 결과를 보고 이벤트 정보를 유추해야 하는 단점이 있다.

최근에는 사용자에게 보다 의미 있는 이벤트 정보를 제공하기 위해 그래프를 이용한 이벤트 검출 기법이 제안되었다[6-9]. [6]는 사전에 분류한 이벤트명과 단어 집합을 이용하여 키워드 그래프를 구축하고 이벤트를 검출하는 기법을 제안하였다. 키워드 그래프에서 각 정점은 온톨로지에 포함된 각 이벤트명과 단어들을 의미하고 이벤트명을 중심으로 함께 발생한 단어를 간선으로 연결한 후 동시 발생 수를 가중치로 부여한다. 구축된 그래프에서 이벤트명을 중심으로 가중치가 λ 이하인 경우 간선을 제거하고 초기 그래프에서 후보 이벤트 그래프를 생성한다. 후보 그래프 중 최소 발생 단어 수보다 적은 단어가 포함되어있다면 후보에서 제외하고 나머지를 결과로 제공한다. 그러나 온톨로지에 의존하기 때문에 등록되어 있지 않은 단어는 결과에서 무시되고, 중요하고 큰 이슈를 불러일으킨 이벤트라 하더라도 온톨로지에 포함되지 않다면 결과를 만들 수 없다는 한계가 있다. [9]에서는 온톨로지를 사용하지 않고 키워드 그래프를 클러스터링하여 이벤트를 검출하는 기법을 제안하였다. 하지만 검출할 이벤트의 수를 입력하는 사용자의 개입으로 인해 분리되어야할 이벤트가 병합되거나 같은 이벤트가 분리하여 결과의 정확성 측면에서 단점을 가진다.

본 논문에서는 소셜 네트워크 환경에서 사용자 관심도를 고려한 그래프 기반 이벤트 검출 기법을 제안한다. 제안하는 기법은 소셜 네트워크에 게시된 글을 분석하여 키워드 그래프를 구축한다. 정점과 간선의 매개중심성을 이용하여 그래프 필터링과 클러스터 과정을 수행하여 후보 이벤트 그래프를 도출한다. 제안하는 기

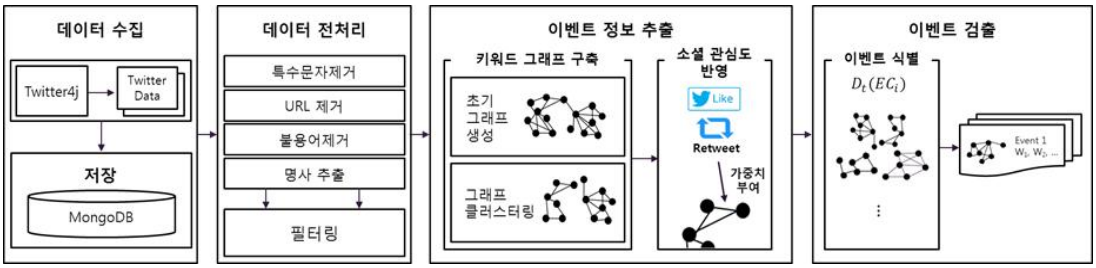


그림 1. 제안하는 기법의 전체적인 처리 과정

법은 기존의 기법에서 고려하는 단어의 출현빈도 뿐만 아니라 사용자들의 관심도를 이벤트를 판별하는 과정에서 고려한다. 사용자들의 소셜 행위의 변화량을 이용해 사용자 관심도를 계산하고 이벤트 판별에 이용하여 최종적인 결과를 도출한다. 사용자의 관심도를 반영함으로써 결과의 신뢰성을 향상시킬 수 있다.

Ⅲ. 제안하는 이벤트 검출 기법

1. 제안하는 기법의 구조

대용량의 소셜 데이터 속에서 이벤트에 대한 정보를 요약하여 사용자에게 보여주고자 하는 연구가 진행되었지만 온톨로지에 포함된 이벤트만 검출하거나 사용자의 개입으로 인해 정확도 감소의 한계점이 있었다. 따라서 사용자의 개입을 최소화하고 다양한 이벤트를 사용자에게 보여주는 기법이 필요하다. 또한, 결과 신뢰성 향상을 위해 사용자의 관심도를 반영하는 기법이 필요하다. 제안하는 기법은 이벤트와 관련된 키워드의 동시발생 빈도를 이용하여 키워드 그래프를 구축한다. 키워드만 나열하는 기존 기법과 달리 그래프를 통해 키워드의 유기적인 관계를 나타냄으로써 효율적으로 정보를 검출할 수 있다. 그래프를 기반으로 이벤트를 검출하는 기존 기법에서 단어의 출현 빈도만을 고려한 것과 달리 소셜 네트워크 서비스에서 사용자들의 관심정도를 나타내는 '좋아요', '공유' 등의 행위를 고려하여 결과의 신뢰성을 향상시킬 수 있다.

[그림 1]은 제안하는 기법의 전체적인 처리 과정이다. 제안하는 기법은 크게 네 단계로 구성된다. 수집단계에서 Twitter 데이터를 수집한 후 데이터베이스에 저장한다. 전처리 단계에서는 이벤트와 관련된 명사를 추출하고 불필요한 정보들을 제거한다. 특수문자, URL, 불용어를 제거하고 형태소 분석기를 통해 명사를 추출한다. 추출한 명사 중 우연히 발생한 것으로 추정되는 명사들은 제거한다. 이벤트 정보 추출 단계에서는 키워드 그래프를 구축하고 후보 그래프를 만드는 이벤트 정보 추출 과정을 수행한다. 키워드 그래프 구축을 위해 전 단계에서 처리된 명사를 정점으로 생성하고 동시에 발생한 경우 간선 연결하고 소셜 행위를 고려하여 사용자 관심도를 계산한 후 가중치로 부여한다. 이벤트 검출 단계에서는 클러스터링 결과를 바탕으로 후보 이벤트에 대한 검증을 수행한다. 모든 단어가 유의미한 것인지 검증하는 단계를 거쳐 사용자가 원하는 Top-k개의 이벤트 그래프를 결과로 도출한다.

2. 데이터 전처리

제안하는 기법에서는 소셜 데이터를 수집하여 분석하기 때문에 분석에 필요한 데이터를 정규화 하고, 불필요한 내용은 제거하는 과정이 필요하다. 사용자들이 글을 게시하면서 감정을 표현하는 이모티콘이나 소셜 네트워크 서비스에서 제공하는 멘션, 리트윗(@), 해시태그(#)를 표현하는 다양한 특수문자를 제거한다. 또한, 사용자들은 가독성으로 인해 게시할 내용을 요약하고 내용을 상세히 설명할 수 있는 URL을 함께 첨부하는 경우가 많기 때문에 이벤트 검출 자체에 큰 의미를 주지 않는 URL도 제거한다.

이벤트 검출에 가장 큰 의미를 보여주는 명사를 추출하여 실질적인 이벤트 검출에 이용한다. 추출된 키워드

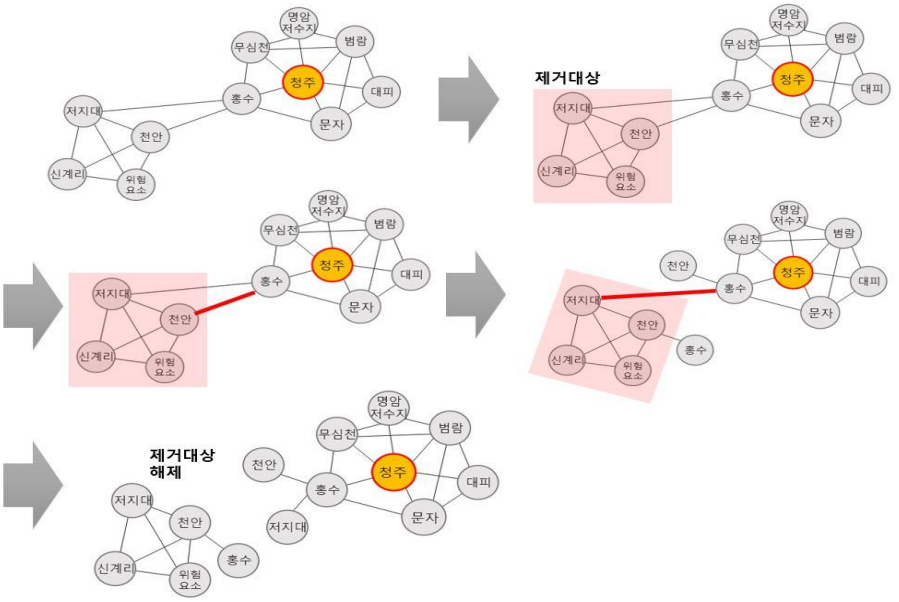


그림 2. 그래프 구축 및 그래프 클러스터링 과정

중 대다수를 차지하는 빈도수가 임계치 이하인 키워드는 우연히 발생하거나 큰 의미가 없는 것으로 판단하여 제거한다. 하루 동안 수집한 데이터의 키워드 출현 빈도수 추이를 분석할 결과 전체 데이터의 50% 가량이 5 이하로 발생하였다. 즉, 빈도수가 5 이하인 단어는 우연히 발생한 단어일 가능성이 높기 때문에 제거한다. 따라서 높은 비율을 차지하는 낮은 빈도수의 단어를 제거함으로써 이벤트 감지에 부적합한 단어를 제거하고 추후 그래프 구축 시 처리 속도를 향상시킬 수 있다.

3. 키워드 그래프 구축

기존의 이벤트 검출 기법에서 이전보다 많이 발생한 키워드를 이벤트 토픽으로 검출하는 기법들이 제안되었다. 그러나 사용자가 결과로 도출된 키워드들을 보고 이벤트 내용을 유추해야 한다. 이벤트를 표현할 수 있는 단어들은 다양하며 이벤트는 주제어와 묘사할 수 있는 다른 단어나 장소들을 포함한다. 사용자들은 소셜 네트워크 서비스에 글을 게시할 때 대체로 토픽 하나만을 게시하지 않고 그 상황을 묘사할 수 있는 다른 표현, 부수적인 상황들을 함께 언급하는 경향이 있다. 따라서 이벤트 토픽과 함께 발생한 단어를 그래프 형태로 구축하여 이벤트 정보를 추출한다.

전처리를 거친 키워드를 이용하여 초기 그래프를 생성한다. 키워드 그래프 G_t 는 한 슬라이드 간격에 해당하는 데이터를 대상으로 구축한 가중치가 잇는 무방향성 그래프이다. 정점들의 집합 V, 정점을 잇는 간선들의 집합 E, 정점 사이의 간선을 갖는 가중치들의 집합 W 로 구성된다. 각 정점 V_i 는 키워드를 나타내며, 다음 클러스터링 단계에서 활용하기 위해 키워드의 출현 빈도수와 그래프에서의 정점의 매개중심성(C_{vetex}^{b}) 값을 속성 값으로 갖는다. 정점의 매개중심성은 그래프에 존재하는 모든 최단 경로에 대하여 해당 정점이 얼마나 많이 그 경로에 속하는지에 대한 비율을 나타낸다. 따라서 매개중심성이 높다는 것은 다른 키워드들과 함께 언급된 비율이 높다는 것으로, 이벤트와 관련된 키워드 중 주제 키워드일 가능성이 높다는 것을 의미한다. 따라서 클러스터링 과정에서 정점의 매개중심성을 활용하여 이벤트 주제 키워드를 찾고 그 키워드를 중심으로 클러스터링을 진행한다. 그래프의 각 간선 E_i는 한 번 이상 동시에 발생한 두 정점을 연결한다. 가중치 W_i는

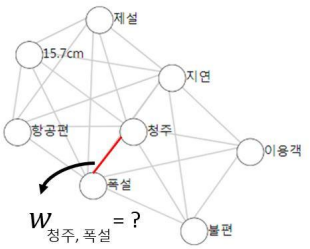
두 정점 V_i와 V_j를 잇는 간선 E_i 에 동시 발생빈 도와 소셜 행위 변화량을 고려한 사용자 관심도 값을 가중치로 갖는다.

4. 그래프 클러스터링

구축된 초기 그래프에서 후보 이벤트 그래프를 찾기 위해서 클러스터링 과정을 수행한다. 기존의 기법에서는 사용자가 찾고자 하는 이벤트 수를 입력하여 클러스터링을 수행하기 때문에 사용자의 요청에 따라 결과가 달라지는 문제가 있었다. 제안하는 기법에서는 사용자의 개입을 최소화하고 반복적인 방법을 통해 클러스터링을 수행함으로써 일관성 있는 결과를 도출한다. 이때, 중심 이벤트 토픽을 중심으로 클러스터링을 수행한다.

[그림 2]는 제안하는 기법의 그래프 구축과 이벤트 클러스터링 알고리즘의 예시이다. 먼저 초기 그래프에서 C^b_{vertex} 값이 가장 높은 중심 키워드를 찾는다. 다음으로 중심 키워드를 기점으로 2홉 이상 떨어진 정점의 경우 붉게 표시된 것처럼 제거대상에 추가한다. 중심이 되는 키워드와 한 문장에 동시 발생하지 않았기 때문에 관련이 없는 키워드일 가능성이 높기 때문이다. 다음으로 그래프에 포함된 모든 간선에 대하여 간선 매개중심성 C^b_{edge} 를 계산한다. 두 이벤트 클러스터를 잇는 간선은 최단경로를 구성할 때 불가피하게 항상 거쳐야하는 경로로 포함되기 때문에 높은 값을 가지기 때문에 C^b_{edge} 값을 통해 서로 다른 이벤트 클러스터의 존재를 확인할 수 있다. 따라서 C^b_{edge} 값이 정규분포를 따른다고 가정할 때, C^b_{edge} 값이 $(m+2\sigma)$을 벗어나는 간선이 있을 때와 없을 때로 경우를 나누어 클러스터링을 진행한다. 이때, m는 평균, σ는 표준편차를 의미한다. 만약 $(m+2\sigma)$ 범위를 벗어나는 간선이 있을 경우, [그림 3]에서 붉게 표시한 간선과 같이 가장 높은 C^b_{edge} 값을 갖는 간선을 절단한다. 절단한 간선과 연결된 두 정점을 각 정점에 새로운 간선을 생성하여 복제하고 앞 단계부터 다시 반복하여 수행한다. 다음으로 범위를 벗어나는 간선이 없을 경우 두 가지 경우로 나누어 처리한다. 첫째, 제거대상 정점들이 첫 단계에서 선택된 중심 키워드와 제거대상 정점들이 같은 클러스터에 있을 경우 대상 정점들을 제거하고 클러스터를 반환하여 후보 이벤트 그래프를 생성한다. 둘째, 첫 단계에서 선택된 중심 키워드와 제거대상 정점들이 다른 클러스터에 존재하는 경우 제거대상에 포함된 정점들을 대상에서 해제하고 각 클러스터를 반환하여 후보이벤트 그래프를 생성한다.

Tweet	Retweet	Likes
"청주 15.7cm 폭설에 항공편 지연"	+20	+10
"청주 폭설로 항공편 지연, 이용객 불편	+6	+2



$\alpha = 0.9 \quad \beta = 1.0001 \quad \mu = 10$

$$W_{청주, 폭설} = \alpha \cdot \mu\left(\beta^{s_{i,j}} - 1\right) + (1-\alpha) \cdot F_{i,j}$$
$$= 0.9 \cdot 10(1.0001^{38} - 1) + (0.1 \cdot \log(2))$$
$$= 0.3483 + 0.0301 = 0.3784$$

그림 3. 간선의 가중치 값 계산 및 부여

5. 사용자 관심도 반영

기존의 기법에서는 단어의 출현 빈도 수, 감정 관련 키워드의 변화량을 고려하고 있다. 그러나 소셜 네트워크 서비스에서 제공되는 특징적인 기능인 '좋아요', '공유' 와 같은 소셜 행위는 고려하지 않고 있다. 소셜 행위는 더 많은 유저들에게 정보를 전파할 수 있다는 특징뿐만 아니라 사용자들의 관심 정도를 의미한다. 따라서 이벤트 검출에서는 소셜 행위의 변화량을 고려하면 사용자의 관심도를 반영할 수 있으며 무분별하게 올라온 스팸문구나 악의적인 글들은 결과에서 제외시킬 수 있기 때문에 결과의 신뢰성을 향상시킬 수 있다.

식 (1)은 두 단어가 동시에 출현한 글의 사용자들의 소셜행위 변화량을 이용한 사용자 관심도 계산 수식으로 리트윗, 좋아요 수의 변화량을 나타낸 것이다. 변화

량 값을 정규화하고 변화율에 따라 가중치 양을 조절하기 위해 식 (2)와 같은 수식을 적용한다.

$$S_{i,j} = \frac{(N_{RT}^{t} + N_{Like}^{t})}{(N_{RT}^{t-1} + N_{Like}^{t-1})} \quad (1)$$

$$NS_{i,j} = \mu \cdot (\beta^{S_{ij}} - 1) \quad (2)$$

식 (3)은 두 단어가 동시에 출현한 빈도수를 의미한다. 식 (4)는 식 (2)와 식 (3)를 이용하여 간선에 부여할 최종적인 가중치를 계산하는 수식이다.

$$F_{i,j} = log(frequency_{i,j}) \quad (3)$$

$$W_{i,j} = \alpha \cdot NS_{i,j} + (1-\alpha) \cdot F_{i,j} \quad (4)$$

위의 수식을 이용하여 사용자 관심도를 산출하고, 후보 이벤트 그래프의 간선에 가중치를 부여하면 [그림 3]과 같은 결과를 도출할 수 있다. 사용자 관심도와 단어 빈도수에 따라 간선에 가중치가 부여되고 최종적인 이벤트 검출 단계에 활용하게 된다.

6. 이벤트 검출

생성된 n 개의 후보 이벤트 그래프의 이벤트 가치를 판별하는 단계를 수행한다. 후보 이벤트 그래프의 이벤트 가치를 판별하기 위해서 이벤트 감지 계수 D_t를 계산하기 위해 간선에 부여된 가중치 값을 활용한다. 따라서 단어의 출현 빈도만을 고려하는 기존의 기법과 비교하여 결과의 신뢰성을 향상시키고자 하였다. 식 (5)은 이벤트 감지 계수를 계산하는 수식이다. 후보 이벤트 그래프에 속하는 단어들의 사용자들의 관심도 변화량을 수식화한 것이다. 큰 값을 가질수록 발생량도 많을뿐더러 많은 사용자들에게 관심을 야기한 것으로 이벤트로써 가치가 있다고 판단할 수 있다. 각 후보 이벤트 그래프의 D_t에 대해 사용자가 요청한 Top-k 개의 이벤트 그래프를 결과로 검출한다.

$$D_t = \log \sum_{v_i v_j \in V} w(v_i, v_j) \quad (5)$$

Ⅳ. 성능 평가

성능 평가 환경은 [표 1]과 같다. 실험 데이터는 2017년 10월 한 달간 Twitter4j API를 통해 수집한 트위터 데이터이며 1,102,534건의 데이터를 수집하였다. Java를 이용하여 데이터를 수집하였으며 전체적인 성능평가는 Python을 이용하여 구현하였다. 제안하는 기법의 우수성을 입증하기 위해 그래프를 기반으로 하는 이벤트 검출 기법 [9]과 검출 이벤트 수에 따른 정확도와 중복 이벤트 비율을 비교 평가한다. 제안하는 기법의 특징인 사용자 관심도를 고려했을 때의 결과를 비교하여 자체 평가를 수행한다.

표 1. 성능평가 환경

항목	값
CPU	Intel(R) Core(TM) i5-6500 CPU 3.2GHz
RAM	8.00GB
OS	Windows 7 64bit
사용 언어	Java 1.8.1 , Python 3
DB	MongoDB 3.6.3

제안하는 기법과 비교대상 기법을 구현하여 Top-k 개의 이벤트 그래프를 결과로 검출하여 정확도를 비교하였다. 정확도는 이벤트 검출 기법을 통해 추출한 이벤트를 기준으로 실제 발생한 이벤트의 비율을 통해 계산한다. 오탐지한 결과가 많을수록 정확도는 낮아진다. [그림 4]는 비교대상과 정확도를 비교한 결과를 나타낸다. 기존 기법 와 k 값에 따른 정확도 비교결과 k값이 20, 25, 30, 35일 때는 정확도가 80%정도로 우수한 성능을 보였지만 k값이 높아질수록 정확도가 급격히 낮아

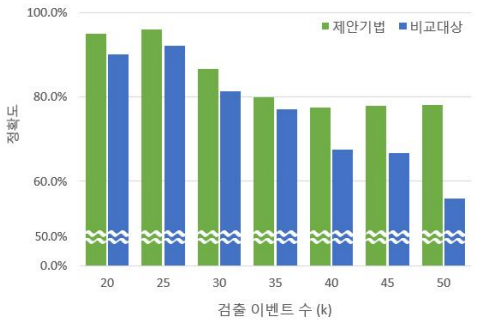


그림 4. 정확도 비교

지는 성능을 보였다. [9]는 k값에 따라 클러스터링 결과가 달라지기 때문에 하나로 검출되어야할 이벤트가 분할되거나 다른 그래프와 합병될 수 있기 때문에 성능이 저하되는 양상을 보였다. 제안하는 기법은 정확도 측면에서 평균 84.4%로 기존 기법과 비교하여 8% 가량의 성능향상을 보였다.

제안하는 기법과 비교대상 기법에 대하여 Top-k 개의 이벤트 그래프를 결과로 검출하여 중복 이벤트 비율을 비교하였다. 중복 이벤트 비율은 검출된 전체 이벤트 결과 중 같은 결과를 중복적으로 검출한 비율을 계산한다. 즉, 검출된 전체 이벤트 결과 중 동일한 의미를 나타내는 이벤트가 서로 다른 클러스터에서 반복적으로 검출된 비율을 나타낸다. [그림 5]는 비교대상과 중복 이벤트 비율을 비교한 결과를 나타낸다. 제안하는 기법은 k값이 증가하여도 10%내외의 중복 이벤트 비율을 보이지만 기존 기법은 k값의 증가에 따라 중복 이벤트 비율이 급격히 증가한다. 기존 기법은 검출하고자 하는 이벤트의 수가 증가할수록 이벤트 그래프를 분할하여 결과로 도출하기 때문에 한 이벤트를 여러 개의 그래프 결과로 검출하는 경향을 보였다. 제안하는 기법의 중복 이벤트 비율은 8.0%로 기존기법과 비교했을 때 평균 10% 가량의 성능향상을 보였다.

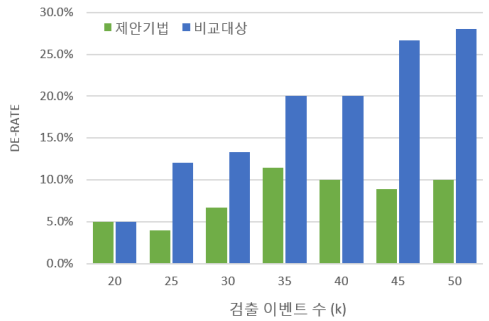


그림 5. 중복 이벤트 비율(DE-RATE) 비교

제안하는 기법에서 특징적으로 고려하고 있는 사용자 관심도를 고려한 것과 고려하지 않고 단어의 출현 빈도수만 고려한 결과를 비교하는 실험을 수행한다. [표 2]와 [표 3]는 2017년 10월 25일 14시에서 15시 사이에 검출된 Top 5 의 이벤트 검출 결과의 차이를 보여준다. 검출된 이벤트 결과의 차이는 사용자들에게 보다 많은 공감과 신뢰를 얻은 이벤트가 상위 랭크가 올라 이벤트로 검출될 수 있었음을 나타낸다.

표 2. 제안하는 기법의 검출 결과

순위	결과
1	이영학, 계부, 유서, 자살, 어금니, 자택, 영월
2	비, 김태희, 출산, 딸, 부모, 득녀, 소식, 더유닛
3	문화, 날, 공연, 영화, 행사, 미술관, 헤이즈
4	박근혜, 국선, 변호인, 비공개, 역대, 선정, 지정
5	토르, 라그나로크, 개봉, 마블, 호평, 기대, 흥행

표 3. 사용자 관심도를 고려하지 않은 검출 결과

순위	결과
1	이영학, 계부, 유서, 자살, 어금니, 자택, 영월
2	김택진, 광고, 야구, 리니지, 엔씨, 쿠폰
3	맨유, 스완지, 기성용, 무리, 풀타임, 멀티, 웨일스
4	비, 김태희, 출산, 딸, 부모, 득녀, 소식, 더유닛
5	박근혜, 국선, 변호인, 비공개, 역대, 선정, 지정

V. 결론 및 향후 연구

본 논문에서는 기존의 소셜 네트워크 환경에서의 이벤트 검출 기법의 문제점을 제시하고 사용자 관심도를 고려한 그래프 기반 이벤트 검출 기법을 제안하였다. 제안하는 기법은 관련된 단어들을 그래프로 표현하고 클러스터링을 수행하기 때문에 중복된 이벤트 발생을 감소시킨다. 또한, 소셜 네트워크 서비스에서 좋아요, 공유 등과 같은 사용자들의 관심도를 고려하여 이벤트 검출의 정확도를 향상시킨다. 성능 평가를 통해 정확도 및 중복 이벤트 검출 비율이 향상된 것을 입증하였다. 제안하는 기법은 소셜 네트워크를 통해 이슈 사항이나 관련 사건, 사고를 실시간으로 검출할 수 있으며 이를 재난 안전 서비스, 마케팅에 활용할 수 있다. 제안하는 기법은 전처리 단계에서 무의미한 키워드를 삭제하기 위한 기준을 보다 명확하게 할 필요가 있으며 이벤트 검출의 정확도를 향상시킬 필요가 있다. 향후 연구로는 무의미한 키워드를 삭제하기 위한 방안과 이벤트 검증 단계에서 정확도를 향상시키기 위한 연구를 진행할 예정이다.

참 고 문 헌

[1] A. Aldhaheri and J. Lee, “Event detection on large social media using temporal analysis,” Proc. IEEE Annual Computing and Communication Workshop and Conference, pp.1-6, 2017.

[2] J. Guzman and B. Poblete, “On-line relevant anomaly detection in the Twitter stream: an efficient bursty keyword detection model,” Proc. ACM SIGKDD Workshop on Outlier Detection and Description, pp.31-39, 2013.

[3] Y. Endo and H. Toda, “What's Hot in The Theme: Query Dependent Emerging Topic Extraction from Social Streams,” Proc. International Conference on World Wide Web, pp.31-32, 2013.

[4] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, “Twitinfo: aggregating and visualizing microblogs for event exploration,” Proc. International Conference on Human Factors in Computing Systems, pp.227-236, 2011.

[5] M. Mathioudakis and N. Koudas, “Twittermonitor: trend detection over the twitter stream,” Proc. ACM SIGMOD International Conference on Management of Data, pp.1155-1158, 2010.

[6] A. Edouard, E. Cabrio, S. Tonelli, and N. L. Thanh, “Graph-based event extraction from twitter,” Proc. International Conference Recent Advances in Natural Language Processing, pp.222-230, 2017.

[7] H. Sayyadi and L. Raschid, “A graph analytical approach for topic detection,” ACM Transactions on Internet Technology, Vol.13, No.2, pp.1-23, 2013.

[8] B. Manaskasemsak, B. Chinthanet, and A. Rungsawang, “Graph Clustering-Based Emerging Event Detection from Twitter Data Stream,” Proc. International Conference on Network, Communication and Computing, pp.37-41, 2016.

[9] S. Katragadda, R. Benton, and V. Raghavan, “Framework for real-time event detection using multiple social media sources,” Proc. Hawaii International Conference on System Sciences, 2017.

[10] G. Valkanas and D. Gunopulos, “Event Detection from Social Media Data,” IEEE Data Engineering Bulletin, Vol.36, No.3, pp.51-58, 2013.

[11] M. Thelwall, K. Buckley, and G. Paltoglou, “Sentiment in Twitter events,” Journal of the Association for Information Science and Technology, Vol.62, No.2, pp.406-418, 2011.

[12] A. Weiler, M. Grossniklaus, and M. H. Scholl, “Event identification and tracking in social media streaming data,” Proc. Workshops of the EDBT/ICDT 2014 Joint Conference, pp.282-287, 2014.

저 자 소 개

김 이 나(Ina Kim) 준회원

- 2016년 2월 : 충북대학교 생화학과(이학사)
- 2016년 8월 ~ 현재 : 충북대학교 빅데이터학과 석사과정

<관심분야> : 소셜네트워크, 빅데이터, 데이터마이닝 등

김 민 영(Minyoung Kim) 준회원

- 2017년 2월 : 충북대학교 정보통신공학과(공학사)
- 2017년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 소셜 네트워크, 데이터베이스 시스템, 분산 컴퓨팅, 그래프 분석

임 종 태(Jongtae Lim) 정회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2011년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2015년 8월 : 충북대학교 정보통신공학과(공학박사)
- 2015년 9월 ~ 현재 : 충북대학교 정보통신공학과 Postdoc

<관심분야> : 데이터베이스시스템, 이동 P2P 네트워크, 소셜 네트워크, 빅데이터 등

복 경 수(Kyoungsoo Bok) 종신회원

- 2009년 2월 : 충북대학교 수학과(이학사)
- 2000년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2005년 8월 : 충북대학교 정보통신공학과(공학박사)
- 2005년 3월 ~ 2008년 2월 : 한국과학기술원 정보전자연구소 Postdoc
- 2008년 3월 ~ 2011년 2월 : 가인정보기술 연구소 차장
- 2011년 3월 ~ 현재 : 충북대학교 전자정보대학 정보통신공학부 초빙교수

<관심분야> : 데이터베이스 시스템, 이동 객체 데이터베이스, 이동 P2P 네트워크, 소셜 네트워크 서비스, 빅데이터 등

유 재 수(Jaesoo Yoo) 종신회원

- 1989년 2월 : 전북대학교 컴퓨터공학과(공학사)
- 1991년 2월 : 한국과학기술원 전산학과(공학석사)
- 1995년 2월 : 한국과학기술원 전산학과(공학박사)
- 1995년 2월 ~ 1996년 8월 : 목포대학교 전산통계학과 전임강사
- 1996년 8월 ~ 현재 : 충북대학교 전자정보대학 정교수

<관심분야> : 데이터베이스시스템, 멀티미디어 데이터베이스, 센서 네트워크, 바이오인포매틱스, 빅데이터 등