소셜 네트워크에서 사용자 관심도를 고려한 이벤트 검출 기법

Event Detection Scheme Considering User Interests in Social Networks

김이나*, 임종태**, 복경수**, 유재수(교신저자)**+
충북대학교 빅데이터학과*,
충북대학교 정보통신공학과**

Ina Kim*, Jongtae Lim**, Kyoungsoo Bok**,
Jaesoo Yoo(corresponding author)**
Department of Big Data, Chungbuk National University*,
School of Information and Communication Engineering, Chungbuk National University**

요약

소셜 네트워크 서비스의 사용량이 폭발적으로 증가함에 따라 실생활에서 발생한 이벤트에 관한 정보가 온라인을 통해 급속도로 확산되고 있다. 이에 따라 소셜 데이터를 통해 이벤트를 검출하기 위한 연구들이 진행되고 있다. 본 논문에서는 소셜 데이터를 이용하여 키워드 그래프를 구축하여 사용자 관심도를 고려한 이벤트 검출 기법을 제안한다. 사용자의 소셜 행위로부터 관심도를 계산하고 관심도의 변화를 고려하여 이벤트 판별에 이용한다. 기존의 기법과 달리 관심도를 반영함으로써 결과의 신뢰성을 향상시킨다.

Ⅰ. 서론

소셜 네트워크 서비스(SNS : Social Network Service)는 시간과 공간의 제한이 없기 때문에 실시간으로 생성되는 소셜 데이터는 이벤트와 관련된 다양한 정보를 포함하고 있다. 그러나 불필요한 정보가 많기 때문에 대량의 데이터에서 사용자가 이벤트에 대한 정보를 얻는 것은 어렵다. 따라서 자동화된 방법으로 이벤트를 검출하고 정보를 제공하기 위한 연구들이 진행되고 있다.

[1]에서는 TF-IDF를 이용하여 특정 시간에 빈번히 출현한 이벤트 키워드들을 검출하는 기법을 제안하였다. 하지만 이벤트 키워드들을 검출하는 방식이기 때문에 사용자가 나열된 결과를 보고 동일한 이벤트를 유추해야만 하는 단점이 있다. [2]에서는 사전에 입력한 이벤트명을 바탕으로 이벤트를 검출하는 기법을 제안하였다. 하지만 사용자가 사전에 입력한 이벤트명과 일치하는 결과만 검출하기 때문에 예기치 못한 이벤트는 검출할 수 없다는 한계가 있다. [3]에서는 이벤트 그래프를 만들고 클러스터링하여 이벤트를 검출하는 기법을 제안하였다. 하지만 검출할 이벤트의 수를 입력하는 사용자의 개입으로 분리되어야할 이벤트가 병합되거나 같은 이벤트가 분리하여 결과의 정확성 측면에서 단점을 가진다. 기존 기법은 소셜 네트워크의 특징인 좋아요, 공유 등의 사용자의 행위는 고려하지 않고 단어 출현 빈도만을 고려하기 때문에 사용자들의 관심을 반영하지 않았다.

본 논문은 소셜 네트워크 환경에서 사용자의 관심도를 고려한 이벤트 검출 기법을 제안한다. 제안하는 기법은 사용자들의 게시 글을 분석하여 키워드 그래프를 구축한다. 정점과 간선의 매개 중심성을 이용하여 그래프 필터링과 클러스터링을 수행한다. 기존의 그래프 기반 이벤트 검출 기법에서 고려하는 단어의 빈도 수 뿐만 아니라 사용자들의 관심도를 이벤트를 판별하는 과정에서 고려한다. 사용자들의 소셜 행위의 변화량을 이용해 관심도 변화량을 계산하고 이벤트 판별에 이용하여 결과를 도출한다.

+ 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원(No. 2016R1A2B3007527)과 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No.2015R1D1A3A01015962)과 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2017-2013-0-00680).

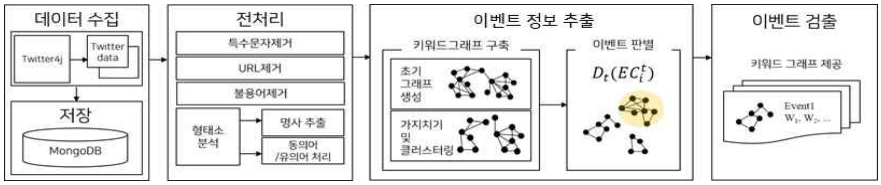
Ⅱ. 제안하는 이벤트 검출 기법

1. 특징

대용량의 소셜 데이터 속에서 이벤트에 대한 정보를 요약하여 사용자에게 정확하게 보여주고자 하는 연구가 진행되었지만 사전에 등록한 이벤트만 검출할 수 있는 문제, 사용자의 개입으로 인한 정확도 감소의 한계점이 있었다. 따라서 사용자의 개입을 최소화하고 다양한 이벤트를 사용자에게 보여주면서 사용자의 관심도를 반영함으로써 결과의 신뢰성을 향상시킬 수 있는 기법이 필요하다.

제안하는 기법은 이벤트와 관련된 키워드의 동시 발생 빈도를 토대로 키워드 그래프를 구축한다. 키워드만 나열하는 기존 기법과 달리 그래프를 통해 키워드의 유기적인 관계를 보여줌으로써 효율적으로 정보를 전달할 수 있다. 그래프를 기반으로 이벤트를 검출하는 기존 기법에서는 단어의 출현 빈도만을 고려한 것과 달리 소셜 네트워크 서비스에서 사용자들의 관심정도를 나타내는 좋아요, 공유 등의 행위를 고려하여 결과의 신뢰성을 향상시킨다.

그림 1은 제안하는 기법의 전체적인 처리과정이다. 제안하는 기법은 4단계로 구성된다. 수집단계에서 Twitter4j를 이용하여 Twitter 데이터를 수집한 후 데이터베이스에 저장한다. 전처리 단계에서는 이벤트와 관련된 명사를 추출하고 불필요한 정보를 제거한다. 이벤트 정보 추출 단계에서는 키워드 그래프를 구축하고 클러스터링을 수행한다. 이벤트 검출단계에서는 클러스터링 결과를 바탕으로 의미 있는 이벤트를 판별하고 결과를 도출한다.



▶▶ 그림 1. 제안하는 이벤트 검출기법의 처리과정

2. 데이터 전처리 및 키워드 그래프 구축

특수문자와 URL등 불필요한 문자열은 전처리 단계에서 제거되며 명사만 추출하여 사용한다. 키워드 그래프 Gt 는 t 시간 간격에 생성된 가중치가 있는 무방향 그래프이다. 정점들의 집합 V, 정점을 잇는 간선들의 집합 E, 정점사이의 간선이 갖는 가중치들의 집합 W로 구성된다. 각 정점 Vi는 트윗에 출현한 단어를 나타내며, 출현빈도와 그래프에서의 매개 중심성(Betweenness Centrality, C^b)값을 갖는다. 각 간선 E_i는 한 번 이상 동시에 발생한 두 정점을 잇는다. E_i는 두 단어의 동시발생빈도와 소셜 네트워크에서의 관심도 변화량을 고려한 가중치 값을 갖는다.

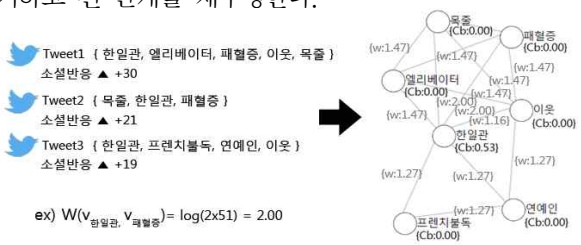
식 (1)은 이전과 현재 시간단위에 발생한 소셜 행위의 총합의 차이고 관심도 변화량을 나타낸다. 식(2)는 관심도 변화량과 동시출현 빈도의 곱이며 간선의 가중치가 된다. 그림 2는 키워드 그래프 구축과정을 간략히 그림으로 나타낸 것이다. 전처리한 단어집합과 해당 트윗의 관심도 변화량을 나타냈으며 키워드 그래프를 만들고 관심도 변화를 이용하여 가중치를 부여하는 예시이다.

$$S_{i,j} = |\sum_{t-1}(N_{rt}+N_{like}+N_{share}) - \sum_{t}(N_{rt}+N_{like}+N_{share})| \quad (1)$$

$$w(v_i,v_j) = \log(freq_{i,j}^{t} \times S_{i,j}) \quad (2)$$

3. 필터링 및 클러스터링

초기 키워드 그래프는 우연히 발생한 단어, 의미 없는 단어를 포함하고 있기 때문에 결과의 정확도와 속도를 위해 필터링 과정을 수행한다. 키워드 그래프에서 생성된 각 클러스터를 찾는다. 정점의 C^b값이 가장 높은 정점을 기준으로 2홉 이상 떨어진 정점의 경우 한 문장에 출현하지 않았으므로 중심 키워드와 관련이 없을 확률이 높다. 따라서 2홉 이상 떨어진 정점은 제거대상에 추가된다. 그래프에 포함된 모든 간선에 대해 매개 중심성(C^b)을 계산한다. 가장 높은 C^b값을 갖는 간선을 절단하고, 두 정점을 각 클러스터에 복제한다. 각 클러스터의 모든 간선에 대해 C^b값 재계산을 수행한다. 모든 간선의 C^b값이 $(m \pm 2\sigma)$을 벗어나는 값이 있다면, 그 간선을 제거하고 전 단계를 재수행한다.



▶▶ 그림 2. 키워드 그래프 구축과정 예시

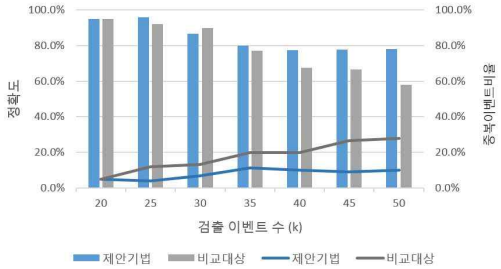
4. 이벤트 판별

생성된 n개의 이벤트 클러스터의 이벤트 가치를 판별하는 단계를 수행한다. 이벤트 클러스터의 소셜 관심도 변화를 고려한 이벤트 감지 계수 D_t를 계산한다. 기존의 기법에서는 단어의 출현 빈도만을 고려하지만 소셜 네트워크 서비스에서 사용자들의 관심정도를 나타내는 좋아요, 공유 등의 행위를 고려하여 결과의 신뢰성을 향상시키고자 하였다. 식(3)은 클러스터에 속하는 단어에 대하여 소셜 관심도의 변화량을 정량화한 수식이다. 즉, 큰 값을 가질수록 이전보다 많이 발생하고 관심을 불러일으킨 것으로 이벤트로 가치가 있다고 판단한다. 각 클러스터의 Dt에 대해 Top k개의 이벤트 클러스터를 결과로 검출한다.

$$D_t = \log \sum_{v_i,v_j \in V} w(v_i,v_j) \quad (3)$$

Ⅲ. 성능평가

제안하는 기법과 유사하게 그래프 기반으로 이벤트를 검출하는 [3]과 비교 평가한다. Intel i5, 8GB 메모리의 환경에서 2017년 10월 한 달간 트위터에서 수집한 데이터로 진행하였다. 정확도는 검출된 이벤트로써 의미가 없는 결과를 도출하는 오탐지 비율을 측정하고 중복 이벤트비율은 같은 이벤트를 다른 클러스터로 검출한 비율을 측정한다. 측정 결과 제안하는 기법이 비교대상에 비해 정확도는 약 7%, 중복이벤트 비율은 10% 가량 성능 향상을 보였다.



▶▶ 그림 3. 기존 기법과 정확도 및 중복 이벤트비율 비교

Ⅳ. 결론

본 논문에서는 소셜 네트워크에서 사용자의 관심도를 고려한 그래프 기반 이벤트를 검출하는 기법을 제안하였다. 제안하는 기법은 수집한 글을 이용하여 키워드 그래프를 구축하고 매개 중심성을 이용하여 클러스터링을 수행하였다. 클러스터링된 이벤트 그래프를 대상으로 사용자들의 소셜 관심도 변화를 활용하여 이벤트 판별을 하여 결과를 도출하였다. 사용자들의 관심도 변화와 출현빈도의 변화폭을 통해 결과를 도출했기 때문에 보다 정확하고 신뢰성 있는 이벤트를 검출할 수 있다.

■ 참 고 문 헌 ■

[1] Endo, Y., Toda, H., "Query Dependent Emerging Topic Extraction from Social Streams.", Proceedings of the 24th International Conference on WWW., pp. 31-32 ACM, 2015.

[2] LI Rui, "Tedas: A twitter-based event detection and analysis system." Data engineering, 28th international conference on IEEE, pp.1273-1276, 2012.

[3] Kateagadda, S. Ryan; Raghavan,. Framework for real-time event detection using multiple social media sources. 2017.