트위터에서의 연관어 군집화를 이용한 이벤트 지역 탐지 기법

하현수, 우승민, 임준엽, 황병연
가톨릭대학교 컴퓨터공학과
e-mail:{hss0924, simter, junyeob1205, byhwang}@catholic.ac.kr

A Method for Detecting Event-location using Relevant Words Clustering in Tweet

Hyunsoo Ha, Seungmin Woo, Junyeob Yim, Byung-Yeon Hwang
Dept. of Computer Science and Engineering, The Catholic University of Korea

요 약

최근 스마트폰의 보급으로 소셜 네트워크 서비스를 이용하는 사용자들이 급증하였다. 그 중 트위터는 정보의 빠른 전파력과 확산성으로 인해 현실에서 발생한 이벤트를 탐지하는 도구로 활용하는 것이 가능하다. 따라서 트위터 사용자 개개인을 하나의 센서로 가정하고 그들이 작성한 트윗 텍스트를 분석한다면 이벤트 탐지의 도구로써 활용할 수 있다. 이와 관련된 연구들은 이벤트 발생 위치를 추적하기 위해 GPS좌표를 이용하지만 트위터 사용자들이 위치정보 공개에 회의적인 점을 감안하면 명확한 한계점으로 제시될 수 있다. 이에 본 논문에서는 트위터에서 제공하는 위치정보를 이용하지 않고, 트윗 텍스트에서 위치정보를 추적하는 방법을 제시하였다. 트윗 텍스트에서 키워드간의 관계를 고려하여 이벤트의 사실여부를 결정하였으며, 실험을 통해 기존 매체들보다 빠른 탐지를 보임으로써 제안된 시스템의 필요성을 보였다.

1. 서론

최근 스마트폰의 보급으로 인한 웹 접근성의 확대로 인해 소셜 네트워크 서비스(Social Network Service; SNS)를 이용하는 사용자들이 급증하고 있다. 그 중 트위터는 다른 SNS와 구별되는 여러 가지 특징을 가지고 있다. 우선, 개발자들에게 다양한 API(Application Program Interface)를 제공하기 때문에 연구적 활용가치가 매우 높다. 또한, 트위터 사용자들은 140자 제한의 단문텍스트를 작성함으로서 비교적 가벼운 내용의 트윗을 간편하게 작성하도록 한다. 이와 더불어 트위터는 개방적인 네트워크 구조를 지니고 있다. 트위터는 사용자간의 관계형성을 위해 팔로워(Follower)-팔로윙(Following)이라는 개념이 이용되는데, 이는 한쪽 사용자의 단 방향적인 요청만으로 관계가 성립된다. 따라서 보다 유동적이며 넓은 범위의 정보 확산을 유도해 낸다.

앞서 언급한 특징들로 인해 트위터를 이용한 연구는 매우 다양하다[1]. 그 중 이벤트를 탐지하고 전파하기 위한 시도들이 존재한다[2, 3]. 이와 같은 대다수의 이벤트 탐지 방법들은 미리 지정된 키워드를 이용하여 관련 트윗들을 수집하였으며, 이벤트 관련 트윗의 발생 위치를 추적하기 위해 트위터 사용자의 GPS 좌표를 이용한다. 하지만 이와 같은 경우 미리 입력하지 않은 키워드에 관한 이벤트는 탐지할 수 없고, 트윗을 작성한 사용자가 위치정보를 공개하지 않으면 정확한 트윗의 발생위치를 알아내는 것은 쉽지 않다. 이러한 문제를 해결하기 위해 본 논문에서는 사용자들이 작성하는 트윗 내용을 직접 분석하여 이벤트가 발생한 위치를 추적하는 방법을 제안한다. 이와 더불어 제안된 기법의 실효성을 검증하기 위해 실제 발생한 이벤트에 대한 실험을 진행하였다.

이 논문의 구성은 다음과 같다. 2장의 관련 연구에서는 제안하는 기법과 관련된 연구를 살펴보고 3장에서 제안된 기법의 구조와 실험 방법을 소개한다. 이후 4장에서는 제안된 기법의 성능을 평가하고 5장에서 결론과 향후 연구 과제에 대해 기술한다.

2. 관련연구

[4]에서는 한국어로 된 문서 내의 핵심 키워드를 추출하기 위해 비사전기반의 키워드 추출 기법을 이용했다. 우선 문장의 형태소 분석을 통한 명사 추출을 수행하고 우선순위에 따라 핵심 키워드를 선정하였으며, 실험을 통해 이를 입증하였다. 이러한 연구는 자연어 처리가 필요한 대다수의 연구에서 활용가치가 높으며, 관련 연구들의 실험을 위한 선행 연구로 볼 수 있다.

한편, [2]에서 제안한 Torretter 시스템은 이벤트의 위치를 탐색하기 위해 급증한 트윗들의 위치좌표를 이용하기

※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2011-0009407).

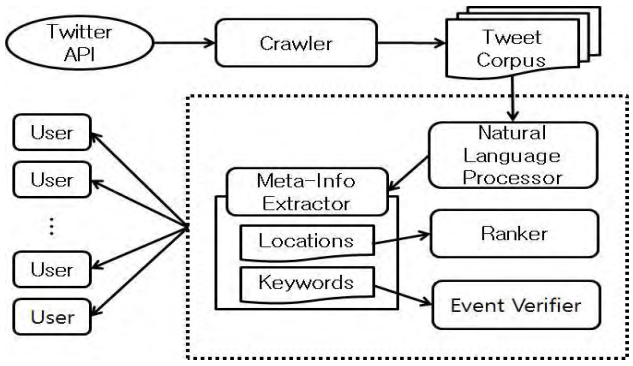
때문에 명확한 한계를 보인다. 이를 개선하기 위해 [5]에서는 트윗에 포함된 위치좌표를 이용하지 않고 트윗 텍스트 자체를 분석하여 위치 정보를 판단하였다. 하지만 많은 노이즈로 인해 탐지된 이벤트의 정확도가 낮았으며 이에 대해 개선이 시급함을 언급하였다.

3. 이벤트 지역 탐지 기법

3.1. 전체 시스템

이 논문에서는 이벤트가 발생된 지역을 탐지하기 위해 <그림 1>과 같은 형태의 시스템을 구축하였다. 우선 트위터에서 제공하는 Streaming API[6]를 이용하여 실시간으로 발생되는 트윗 데이터를 수집한다. 이때 크롤러(Crawler)를 통해 한국어로 작성된 트윗만을 수집한 후, 데이터베이스에 저장한다. 이후 [7]을 이용하여 자연어 처리를 수행한다. 이를 통해 트윗 텍스트에서 다수의 명사를 추출하였고, 추출된 명사를 지명과 그 외의 단어들로 구분하였다. 지명의 판단 기준은 [8]에서 소개된 자료를 토대로 수행되었다.

이러한 과정을 거치면 다수의 트윗 코퍼스에서 각각의 지명들이 언급된 횟수를 알 수 있다. 언급 빈도가 급증한 지역은 최근 트위터 사용자들에게 이슈가 된 지역이며 이벤트 지역으로 판단한다. 이때 시스템에서 이벤트 지역으로 판단한 지역을 검증하기 위해 트윗 문장에서 유사한 키워드를 추출하여 이벤트와 연관된 단어를 군집화 한다. 이와 관련된 내용은 3.2절에 기술하였다.



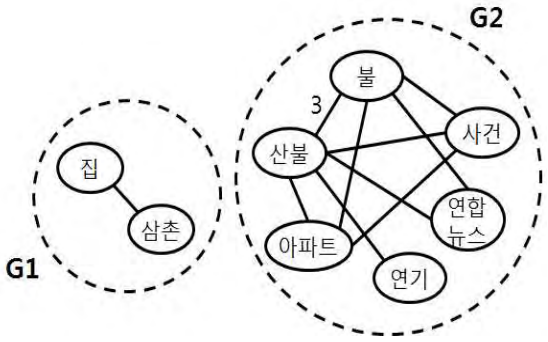
(그림 1) 트위터 이벤트 지역 탐지 시스템

3.2. 연관어 군집화를 이용한 주제별 클러스터링 기법

이 절에서는 트윗 텍스트에서 유사한 키워드를 추출하여 이벤트를 검증하는 방법을 소개한다. 유사한 키워드를 추출하기 위해 트윗 내의 연관어들을 군집화 하고 각 연관어들의 개념적 거리를 측정하였다.

연관어를 군집화 하기 위해서는 같은 트윗에 함께 포함된 키워드를 연결하는 방식으로 수행한다. 이는 동시 다발적으로 발생된 어휘들이 연관성이 높다고 판단되기 때문이다. 따라서 <그림 2>와 같이 각각의 키워드를 그래프 형태로 표현한다. 예를 들어 <표 1>과 같이 '포항'이라는 지명을 포함한 5개의 트윗에서 각각의 명사들이 추출되었다고 가정하자. 여기서 지명을 제외한 각각의 추출된 키워드들은 하나의 노드로 표현하고, 같은 트윗에서 발생된 키워드일 경우 간선으로 연결한다.

이와 더불어 각 연관어들의 개념적 거리를 측정하는 작업이 필요하다. 따라서 순차적으로 트윗이 발생될 때 마다 간선을 그려가며 이미 간선으로 연결된 노드일 경우 가중치를 두어 값을 1씩 증가시킨다. 이후 <그림 2>에서와 같이 G1, G2등으로 간선으로 연결된 노드를 군집화 하고 생성된 군집 중 가장 높은 가중치를 가지는 그룹을 최종 이벤트 키워드 집합으로 판단한다.



(그림 2) 연관어 군집화

<표 1> 예시 키워드

트윗 번호	추출된 명사
트윗 1	포항 / 산불 / 불 / 연합뉴스
트윗 2	포항 / 산불 / 불 / 사건 / 아파트
트윗 3	포항 / 산불 / 불
트윗 4	포항 / 산불 / 연기
트윗 5	포항 / 삼촌 / 집

최종 이벤트 키워드 집합이 정해지면 각 지역별로 임계값을 정하여 가중치의 합이 임계값을 넘는지를 살펴본다. 지역별 임계값은 실험을 통해 학습된 값이며, 임계값의 최대값과 최소값을 비교하여 시스템이 주기적으로 갱신한다. 이러한 이벤트 판별 방식은 단순히 발생빈도 등을 측정하는 것이 아니라 발생된 형태 및 연관 분포를 살펴보게 되므로 [7]의 한계점이었던 노이즈가 자동으로 제거되는 효과가 있다. 또한 이벤트가 발생된 지역뿐 아니라 이벤트의 핵심 키워드를 알 수 있으므로 시스템의 사용자들은 이벤트가 발생된 지역과 이벤트 키워드를 수신하여 해당 이벤트에 따른 올바른 대처를 할 수 있을 것이다.

4. 실험 및 실험 결과

실험을 위해 2013년 3월부터 13개월간 수집한 트윗 데이터를 시스템에 입력하여 이벤트 탐지 여부를 확인하였다. 탐지할 대상 이벤트는 실제 발생된 다수의 이벤트 중 KBS 뉴스 속보에서 보도된 지역 관련 이벤트로 정하였으며, 자세한 내용은 <표 2>와 같다.

시스템에서는 대부분의 속보를 평균적으로 1시간가량 빠르게 탐지하였다. 특히 속보 1의 경우 8시간 41분 빠른 탐지를 보였는데, 이는 뉴스 보도의 특성상 대규모 사건으로 확대되기 전까지는 보도되기 어렵다는 특성 때문이다. 특히, 탐지된 이벤트 중 속보 1, 7, 9는 빠른 이벤트 탐지와 그에 따른 초동대처로 인명피해를 최소화 할 수 있는 성격의 이벤트들이다.

한편, 속보 2, 4, 5, 6, 8은 제안된 시스템에서 탐지하지 못했다. 속보 2, 6은 '중부지방', '수도권' 등 전국적인 이벤트의 경우로 광역적인 지역 명칭에 대한 예외 처리가 필요할 것이다. 이와 더불어 속보 4, 8도 탐지하지 못했는데 속보 4의 동해상과 속보 8의 백령도 인근은 인구밀도가 매우 적은 지역이므로 트위터 사용자가 적기 때문에 시스템에서 탐지하기 어렵다. 마지막으로 속보 5 또한 시스템에서는 탐지하지 못하였는데, 이는 기존에 '삼성'이라는 단어가 자주 언급되어 발생 빈도의 변화가 미비했기 때문이다. 이와 같은 동음이의어에 따른 문제점들을 해결하기 위해서는 보다 정확한 형태소 분석과 문맥 내에서의 의미 판별을 위한 추가적인 작업이 필요할 것으로 보인다.

<표 2> 실험에 이용된 속보 내용

	속보 보도 일시	속보 내용	탐지여부
속보 1	2013.03.10. 02:50	전국 산불 - 울산, 포항, 공주 등	O
속보 2	2013.07.13. 09:30	중부지방 폭우	X
속보 3	2013.08.31. 10:30	KTX 대구 부근 탈선 후 추돌	O
속보 4	2013.10.08. 13:58	태풍 다나스 한반도 접근 및 동해상 이동	X
속보 5	2013.11.16. 09:15	삼성동 아이파크 아파트 헬기	X
속보 6	2014.01.20. 05:00	중부지방 많은 눈	X
속보 7	2014.02.17. 22:50	경주 마우나 리조트 붕괴	O
속보 8	2014.03.31. 13:35	백령도 북 사격 훈련 및 대응 사격	△
속보 9	2014.04.16. 10:04	진도 해상 여객선 침몰	O

5. 결론 및 향후 과제

이 논문에서는 트위터에서 유사한 키워드의 연관어 군집화를 이용한 이벤트 지역 탐지 방법에 대해 소개하였다. 제안된 시스템에서는 2013년 3월부터 13개월간 수집한 트윗 데이터를 이용하였으며, 실험 결과 탐지율은 낮았으나 실제 발생된 다수의 이벤트를 탐지한 다른 매체들보다 빠르게 탐지할 수 있었다. 한편, 시스템의 정확도와 관련하여 여러 한계점들이 발견되었는데 주된 문제점은 지명과 동음이의어 관계에 있는 단어에 의한 것이었다. 따라서 향후 과제로 단문 텍스트 내의 동음이의어를 판별하기 위한 방안이 연구되어야 할 것이다. 또한, 본문에서 언급한 광역적인 지역 명칭에 대한 명확한 정의가 필요할 것이다.

참고문헌

[1] S.-Y. Park, Y.-H. Ha, Y.-H. Kim, "Recent Studies on Twitter in the Field of Information Retrieval," Proc. of KIISE Fall Conference, pp.25-29, 2010.

[2] T. Sakaki, M. Okzaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," Proc. of the 19th Int'l Conf. on World Wide Web, pp.851-860, 2010.

[3] R. Li, K. H. Lei, R. Khadiwala, and K. Chang, "TEDAS: a Twitter Based Event Detection and Analysis System", Proc. of the IEEE 28th International Conference on Data Engineering, pp.1273-1276. 2012.

[4] M. Ku, D. Min, "Study on Keyword Extraction Method using Recursive Extracted Word Division," Proc. of KIISE Fall Conference, pp.329-334, 2009.

[5] J. Yim, J. Yoon, B. Lee, B.-Y. Hwang, "Designing of Event Decision Module using Twitter," Proc. of KIPS Spring Conference, pp.680-683, 2014.

[6] Twitter. (2012, Sep. 24). The Streaming APIs|Twitter Developers, https://dev.twitter.com/docs/streaming-apis

[7] S. Lee. (2008, Oct. 18). Lucean Korean Morph Analyzer, http://cafe.naver.com/korlucene

[8] Republic of Korea National Statistical Office, "Population and Housing Census 2010", http://www.kostat.go.kr.