소셜 네트워크에서 사용자의 영향력을 고려한 신뢰성 높은 핫 토픽 검출 기법

High Reliable Hot Topic Detection Scheme Considering User Influences in Social Networks

노 연 우, 전 현 욱, 육 미 선, 한 지 은, 임 종 태,
김 연 우, 복 경 수, 유 재 수
충북대학교 정보통신공학부

Yeon-woo Noh, Hyeon-wook Jeon, Misun Yook,
Jieun Han, Jongtae Lim, Yeon-woo Kim,
Kyoungsoo Bok, Jaesoo Yoo
School of Information and Communication Engineering,
Chungbuk National University

요약

소셜 네트워크의 발달로 대량의 데이터로부터 원하는 정보를 빠르게 분석하고 유의미한 정보를 찾아내는 것이 중요해지면서 핫 토픽 검출에 대한 관심이 증가하고 있다. 본 논문에서는 단어의 출현 빈도수뿐만 아니라 사용자 영향력을 종합적으로 고려하여 이를 기반으로 트윗에 가중치를 부여함으로써 검출 결과의 신뢰성을 향상 시킬 수 있는 핫 토픽 검출 기법을 제안한다.

I. 서론

스마트 디바이스의 발달로 인해 트위터와 같은 소셜 네트워크 서비스(SNS: Social Network Service)가 급속하게 발전하고 있다. 전 세계의 수많은 사람들이 SNS를 활용하여 삶의 많은 양상에 대한 그들의 의견을 게재함에 따라 SNS는 세상의 트렌드를 파악하기 위한 중요한 데이터 소스로 활용되고 있다. 그러나 이처럼 방대하게 쏟아지는 SNS 데이터에서 원하는 정보를 찾는 작업은 점차 어려워지고 있으며 효율성 측면에서 많은 문제를 발생시키고 있다.

대량의 SNS 데이터로부터 원하는 정보를 빠르게 분석해주고 유의미한 정보를 찾아내는 것이 중요한 논제로 제기 되면서 이에 대한 관심이 증가되고 핵심 세부 연구 주제로써 핫 토픽 검출에 대한 연구가 활발하게 이루어지고 있다. [1]에서는 시간에 따른 출현 빈도수의 비율을 고려하여 급격한 변화를 보이는 단어들을 핫 토픽으로 검출한다. 그러나 사전 확인이 이루어지지 않은 불특정 다수의 글을 대상으로 평가를 수행하기 때문에 신뢰성이 저하된다.

본 논문에서는 소셜 네트워크 환경에서 사용자의 영향력을 고려한 신뢰성 높은 핫 토픽 검출 기법을 제안한다. 사용자 영향력과 신뢰성 사이에는 높은 연관성이 있으므로 사용자 영향력을 가중치로 부여하여 핫 토픽 검출 결과의 신뢰성을 부여한다. 그러므로 제안하는 기법에서는 기존 기법에서 활용했던 출현 빈도수와 사용자 영향력을 종합적으로 고려하여 핫 토픽 검출을 수행함으로써 신뢰성 높은 결과를 도출한다.

* 교신저자 : yjs@chungbuk.ac.kr

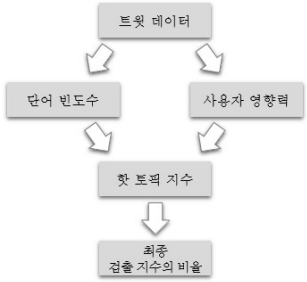
이 논문은 2012년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(2012R1A1A2A10042015)과 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발 사업의 일환으로 수행하였음[14-824-09-001, 실시간 대규모 영상 데이터 이해·예측을 위한 고성능 비주얼 디스커버리 플랫폼 개발]

Ⅱ. 제안하는 핫 토픽 검출 기법

1. 핫 토픽 검출 구조

SNS 환경에서는 신뢰성과 연관 되는 다양한 요인들이 존재하지만 이중 여론주도자의 영향력이 신뢰성과 가장 큰 연관성을 가진다[2]. 제안하는 기법은 높은 신뢰성을 갖는 핫 토픽 검출을 위해 출현 빈도수뿐만 아니라 사용자의 영향력을 이용하여 핫 토픽을 검출한다.

그림 1은 제안하는 기법의 전체적인 처리 절차이다. 핫 토픽을 검출하기 위해 단어 빈도수와 사용자 영향력을 판별한다. 단어 빈도수와 사용자 영향력을 결합하여 핫 토픽 여부를 계산하고 시간에 따른 핫 토픽 지수의 비율을 이용하여 최종 핫 토픽을 검출한다.



▶▶ 그림 1. 핫 토픽 검출 시스템 개념 모델

2. 사용자 영향력을 고려한 신뢰성 높은 핫 토픽 검출

본 논문에서는 트위터에서 사용자가 수행 가능한 다양한 활동 중에서 영향력과 높은 관계에 있는 세 가지 요소(팔로워의 수, 리트윗의 수, 멘션의 수)를 기준으로 사용자의 영향력을 도출하였다. 식 (1)은 제안하는 기법에서의 단어별 핫 토픽 지수를 나타낸다. R_t^w은 시간 t에서 단어 w에 대한 핫 토픽 지수를 의미한다. 각 단어는 핫 토픽 지수를 기준으로 순위화하여 상위 N개의 단어를 핫 토픽으로써 사용자에게 최종 추천한다.

$$R_t^w = \frac{I_t^w - I_{t-1}^w}{I_t^w + I_{t-1}^w} \quad (1)$$

식 (2)는 제안하는 기법에서의 사용자 영향력 지수를 나타낸다. 사용자 영향력 지수는 팔로워의 수, 리트윗 수, 멘션의 수를 기준으로 도출되며, 이때 각 요소들은 서로 연관성이 없으므로 각 요소별 영향력 지수를 계산한 후 합하여 특정 사용자 A에 대한 최종 영향력 지수인 I_{U_A}를 계산한다.

$$I_{U_A} = \log(I_{U_A^f}) + \log(I_{U_A^{rt}}) + \log(I_{U_A^m}) \quad (2)$$

식 (3)은 사용자 영향력을 도출하기 위한 구성 요소로써 특정 사용자의 팔로워의 수를 나타낸다. 이 식은 팔로우한 사람이 특정 사용자의 트윗에 대해 갖는 관심의 정도를 나타낸다.

$$I_{U_A^f} = \frac{\sum_{f \in U_A^f} Followers}{\alpha} \quad (3)$$

식(4)는 사용자 영향력을 도출하기 위한 구성 요소로써 특정 사용자의 리트윗의 수를 나타낸다. 특정 사용자의 트윗 당 평균 리트윗 비율 및 이것을 리트윗하는 팔로워들의 전파력을 고려하여 그 수치가 클수록 사용자 영향력이 높다고 측정하였다. 이 때 팔로워들의 전파력은 기준이 되는 특정 사용자의 팔로워들의 평균 팔로워 수를 기준으로 함으로써 특정 사용자가 트윗을 올렸을 경우 평균적으로 얼마나 많은 사용자들이 해당 트윗을 접하게 되는지를 측정하였다.

$$I_{U_A^{rt}} = \frac{\frac{\sum_{rt \in U_A^{rt}} Retweets}{\sum_{t \in U_A^f} Tweets} \times \frac{\sum_{U_f \in U_A^f} Followers}{\sum_{f \in U_A^f} Followers}}{\beta} \quad (4)$$

식 (5)는 사용자 영향력을 도출하기 위한 구성 요소로써 특정 사용자의 멘션의 수를 나타낸다. 팔로워의 수와 마찬가지로 멘션 수신 수가 많을수록 사용자 영향력이 높다고 측정하였다.

$$I_{U_A^m} = \frac{\sum_{m \in U_A^m} Mentions}{\gamma} \quad (5)$$

Ⅲ. 성능평가

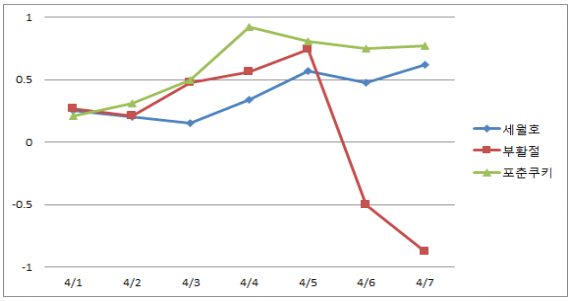
제안하는 기법의 성능을 평가하기 위하여 JAVA를 이용하여 핫 토픽 검출 시스템을 구현하였으며, 실험 데이터는 실시간으로 트위터 데이터를 모을 수 있는 Twitter Streaming API를 이용해 약 56만개의 샘플 데이터를 수집하였다. 또한 데이터베이스는 MySQL 5.6.23 버전을 사용하였다.

표 1은 제안하는 기법의 최종 핫 토픽 지수를 적용하여 검출된 2015년 4월 1일부터 7일까지의 상위 7번째까지 핫 토픽 키워드들을 나타낸다.

표 1. 제안하는 기법으로 검출된 핫 토픽 키워드 집합

날짜	검출된 핫 토픽 키워드 Top 7
15.04.01	만우절, 소녀시대, 빅뱅, KBS, 길건, 신곡, 김태우
15.04.02	흑자, 앵그리맘, 김준수, 현대, 경상수지, 핵협상, 호남고속철도
15.04.03	이란, 성완종, 수지, 이문세, 벚꽃, 세월호, 이민호
15.04.04	포춘쿠키, 무한도전, 윤승아, 강균성, 결혼, 식스맨, 개기월식
15.04.05	최시원, 포춘쿠키, 우결, 엑소, 기성용, 음악중심, 부활절
15.04.06	박효신, K팝스타, 세월호, 복면가왕, 수지, 축제, 벤드게이트
15.04.07	이문세, 봄바람, 인양, 세월호, 오드리햅번, 경남기업, 김태우

그림 2는 표 1에서 검출된 핫 토픽 키워드들 중 특정 단어에 대한 1주일 동안의 핫 토픽 지수 변화율을 나타낸 그래프이다. 부활절은 부활절 당일인 4월 5일을 기준으로 변화율의 폭인 $|R|$이 1에 가깝게 급격한 변화를 보인다. 세월호는 $|R|$은 0.5에 가깝고 꾸준하게 지수 변화율 값이 변화한다.



▶▶ 그림 2. 핫 토픽 키워드에 대한 핫 토픽 지수 변화율

Ⅳ. 결론

본 논문에서는 트위터 사용자의 영향력을 고려한 핫 토픽 검출 기법을 제안하였다. 제안하는 기법에서는 단어의 빈도수와 트위터 사용자의 영향력을 이용하여 핫 토픽 지수를 도출함으로써 이슈화가 되는 단어를 검출하고 영향력 지수의 비율을 고려하였다. 향후 연구는 트윗 데이터에서 단어의 빈도수를 측정할 때 TF-IDF 알고리즘을 사용하여 정확도를 높이고 트위터 사용자의 전문성을 고려한다.

■ 참 고 문 헌 ■

[1] H. Kim, S. Lee, and S. Kyeong, "Discovering Hot Topics using Twitter Streaming Data", Proc. ASONAM, pp.1215-1220, 2013.

[2] E. Lee, A Study on the Factors Influencing upon SNS Credibility, Sejong University, 2012. (in Korean)

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy", Proc. of the International AAAI Conference on Weblogs and Social Media, pp.10-17, 2010.