http://dx.doi.org/10.5392/JKCA.2015.15.08.024

소셜 네트워크에서 사용자의 영향력을 고려한 핫 토픽 예측 기법

Hot Topic Prediction Scheme Considering User Influences in Social Networks

노연우, 김대윤, 한지은, 육미선, 임종태, 복경수, 유재수
충북대학교 정보통신공학부

Yeon-woo Noh(ywnoh@chungbuk.ac.kr), Dae-yun Kim(kdy0573@chungbuk.ac.kr),
Jieun Han(jieun24@chungbuk.ac.kr), Misun Yook(misun@chungbuk.ac.kr),
Jongtae Lim(jtlim@chungbuk.ac.kr), Kyoungsoo Bok(ksbok@chungbuk.ac.kr),
Jaesoo Yoo(yjs@chungbuk.ac.kr)

요약

최근 실시간으로 생성되는 대용량의 SNS 데이터로부터 유의미한 정보를 찾아내고 분석하는 것이 중요해지면서 핫 토픽 검출에 대한 관심도 크게 증가하고 있다. SNS 특성상 사전 확인이 이루어지지 않은 불특정 다수의 글들을 대상으로 하기 때문에 이 글들을 대상으로 핫 토픽을 예측했을 때 결과의 신뢰성이 저하된다는 문제점이 있다. 이를 해결하기 위하여 본 논문에서는 소셜 네트워크에서 사용자의 영향력을 고려한 신뢰성 높은 핫 토픽 예측 기법을 제안한다. 트위터를 기반으로 변형된 TF-IDF 알고리즘을 통하여 순간적으로 많이 이슈화되는 키워드 후보 집합을 추출하고, 트윗에 사용자 영향력을 가중치로 부여함으로써 핫 토픽 예측 결과의 신뢰성을 높인다. 제안하는 기법의 우수성을 보이기 위해 기존 기법과 제안하는 기법의 성능평가를 수행한다. 성능평가 결과, 제안하는 기법은 기존 기법에 비해 정확도, 재현율 모두 향상됨을 확인하였다.

■ 중심어 : | SNS | 트위터 | 핫 토픽 | 예측 |

Abstract

Recently, interests in detecting hot topics have been significantly growing as it becomes important to find out and analyze meaningful information from the large amount of data which flows in from social network services. Since it deals with a number of random writings that are not confirmed in advance due to the characteristics of SNS, there is a problem that the reliability of the results declines when hot topics are predicted from the writings. To solve such a problem, this paper proposes a high reliable hot topic prediction scheme considering user influences in social networks. The proposed scheme extracts a set of keywords with hot issues instantly through the modified TF-IDF algorithm based on Twitter. It improves the reliability of the results of hot topic prediction by giving weights of user influences to the tweets. To show the superiority of the proposed scheme, we compare it with the existing scheme through performance evaluation. Our experimental results show that our proposed method has improved precision and recall compared to the existing method.

■ keyword : | Social Network Services | Twitter | Hot Topics | Prediction |

* 이 논문은 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업(IITP-2015-H8501-15-1013), 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No.2015R1D1A3A01015962), 2014년도 산업통상자원부 재원으로 한국에너지기술평가원(KETEP)의 지원을 받아 수행한 연구과제임(No.20144030200450)
* 본 논문은 한국콘텐츠학회 2015 춘계 종합학술대회 우수논문입니다.
접수일자 : 2015년 07월 23일
수정일자 : 2015년 08월 05일
심사완료일 : 2015년 08월 05일
교신저자 : 유재수, e-mail : yjs@chungbuk.ac.kr

I. 서 론

최근 SNS(SNS: Social Network Service)가 발달함에 따라 수많은 사람들이 스마트 디바이스나 웹을 통해 SNS을 활용하여 의견을 게재하고 정보를 공유하고 있다. SNS는 인터넷을 기반으로 사용자들이 인적 네트워크를 형성할 수 있게 해주고 정보 공유, 인맥 관리, 자기 표현 등을 통해 타인과의 관계를 관리하기 위한 서비스이다. 초기에는 주로 친목도모 용도로 활용이 되었으나 이후 점차 SNS가 단순히 사용자들 사이의 인맥 관계를 통한 정보 공유 뿐만 아니라 새로운 정보를 생성하고 소비하는 형태로 전환되었다. 또한, 일반적인 웹 검색을 통해 정보를 찾는 것보다 친구의 추천으로 공유하는 정보가 신뢰성이 높고 간결하게 전달되기 때문에 일반적인 인터넷 검색보다는 SNS를 통하여 최신 정보를 찾고 이를 활용하는 이들이 많아지면서 기하급수적으로 재생산되고 공유되는 대용량의 정보로부터 최근 이슈가 되는 정보를 찾아내는 기법이 요구되고 있다[1].

트위터, 페이스북, 라인, 미투데이, 구글+ 등은 대표적인 SNS로 그 중 트위터는 간결한 인터페이스를 통해 사용자들이 인터넷 상에서 다른 사람과의 네트워크를 쉽게 형성하게 해주기 때문에 급속히 성장하고 있는 서비스이다. 2006년 이후 계속 성장하여 현재 월간 실사용자 수가 3억명 이상이며 일일 트윗 수가 5억 건 이상으로 활발한 이용 실태를 보여주고 있다. 또한, 트위터는 140자의 제한 수를 가지기 때문에 실시간으로 이슈화되는 글들을 찾기에 용이한 서비스로 관심 있는 상대방을 뒤따르는 팔로우(follow)라는 독특한 기능을 중심으로 소통한다. [그림 1]은 사용자 A가 B를 팔로우한 상황을 나타낸 것으로 이러한 관계에서 A는 B의 팔로워(follower)라고 하고 B는 A의 팔로위(followee)라고 한다. 이는 다른 SNS의 친구 맺기와 비슷한 개념이지만 상대방이 허락하지 않아도 일방적으로 팔로워로 등록할 수 있는 점이 가장 큰 차이점이다. 웹에 직접 접속하지 않더라도 휴대전화의 문자메시지나 스마트폰 같은 휴대기기 등 다양한 방법을 통하여 글을 올리거나 받아볼 수 있으며, 댓글을 달거나 특정 글을 다른 사용자들에게 퍼트릴 수도 있다. 트위터의 글을 트윗(tweet)이라고 하며, 팔로위가 작성한 트윗을 자신의 팔로워에게 전파하는 기능을 리트윗(retweet)이라고 한다. 멘션(mention)은 특정 사용자에게 트윗을 보내는 기능이다.

그림 1. 트위터에서의 팔로우 관계

트위터는 언제 어디서나 정보를 실시간으로 교류할 수 있는 빠른 소통이 가장 큰 특징으로 세계적 뉴스채널로 속보를 장점으로 하는 CNN을 앞지를 정도로 신속한 정보 유통망으로 주목받고 있다. 미국의 첫 흑인 대통령이 된 버락 오바마가 대통령 선거에서 승리하는데 트위터를 이용한 홍보효과를 톡톡히 본 것으로 알려져 있으며[2], 기업들도 홍보나 고객 불만 접수 등 다양한 방법으로 활용하고 있다. 연구 분석가들은 트윗들을 이용하여 서비스나 상품에 관해 시장조사를 하기 위한 목적으로 사용한다.

사용자들은 트윗을 이용하여 자신의 상태를 표현하거나 다양한 정보를 공유할 수 있다[3]. 이처럼 방대하게 생성되는 트윗들로부터 실제 원하는 정보를 찾는 것은 어려운 작업이며 그 효율성 면에서 여러 가지 문제를 발생시키고 있다. 최근 소셜 네트워크에서 이슈가 되거나 핵심 주제로 부각되고 있는 핫 토픽을 검출하기 위한 연구들이 진행되고 있다. 기존의 대표적인 핫 토픽 검출 연구는 단어 출현 빈도수의 비율을 이용한 기법이다[4]. 단어 출현 빈도수를 이용하여 핫 토픽을 검출할 때 실제 핫 토픽이 아닌 일상적으로 많이 쓰이는 단어들이 핫 토픽으로 검출된다는 문제점을 보완하기 위해 시간에 따른 단어 출현 빈도수의 비율을 이용한 기법을 제안하였다. 하지만 소셜 네트워크 서비스의 특성을 고려하지 않고 단어의 출현 빈도수만을 기준으로 핫 토픽을 검출하였기 때문에 검출 결과에 있어서 효율성이 높다고 할 수 없다. 또한, 소셜 네트워크 특성상 트위터는 사전 확인이 이루어지지 않은 불특정 다수의 글들을 대상으로 하기 때문에 검출 결과의 신뢰성이 저

하된다는 문제점이 있다.

본 논문에서는 소셜 네트워크 환경에서 사용자의 영향력을 고려한 신뢰성 높은 핫 토픽 예측 기법을 제안한다. 실제 트렌드를 분석하여 핫 토픽을 예측하는 기법으로 전체 트윗들을 대상으로 시간을 고려한 변형된 TF-IDF 알고리즘을 사용하여 특정 시간에 순간적으로 많이 발생하는 키워드의 집합을 추출한다. 단어의 출현 빈도수와 사용자의 영향력을 종합적으로 고려하여 핫 토픽 지수를 도출하고 이를 랭킹화하여 핫 토픽을 예측한다. 사용자의 영향력과 검출 결과의 신뢰성과 효율성 사이에는 높은 연관성이 있으므로 사용자 영향력을 가중치로 단어에 부여하면 기존의 단어 출현 빈도수만을 기준으로 고려된 핫 토픽 검출 방법보다 검출 결과에 있어서 정확도와 신뢰성이 더 높아진다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구와 기존 기법들이 가지고 있는 문제점을 기술한다. 3장에서는 본 논문에서 제안하고 있는 기법을 상세히 기술하고, 4장에서는 성능평가를 통해 제안하는 기법의 우수성을 입증한다. 마지막으로 5장에서는 결론 및 향후 연구에 대해 기술한다.

II. 관련연구

J.Haziq[5]는 트위터에서 기계 학습 알고리즘을 이용하여 트윗을 긍부정 분류를 하는 기법을 제안하였다. 긍부정 분류는 연구 분석가들이 서비스나 상품에 관해 시장 조사를 위해 사용된다. 긍정의 단어를 포함하는 트윗은 긍정으로 분류하고 부정의 단어를 포함하면 부정으로 분류하며 그 외의 일반 동사가 들어가면 중립으로 분류를 한다. 기존 대부분의 연구에서는 중립의 트윗이 분류가 어렵다고 하여 긍부정 트윗 외에 따로 연구하지 않았었으나 이렇게 되면 문맥상 긍부정을 나타낼 수 있는 중립의 단어를 고려하지 않게 되므로 잘못된 분류로 이어질 수 있다. 우선 기계 학습 알고리즘으로 훈련될 데이터집합으로부터 이모티콘은 모두 제외시키고 불용어를 제외시키는 등 전처리 과정을 거친다. 그리고 가중치가 부여된 유니그램인 특성 벡터를 사용하여 데이터 집합의 크기를 줄인다. 유니그램 특성 벡터는 트윗으로부터 특성을 추출하는 가장 간단한 방법인데 유니그램과 가중치가 부여된 유니그램을 사용하여 분류의 정확성을 성능평가 한 결과 3개의 기계 학습 알고리즘에서 가중치가 부여된 유니그램을 사용했을 때가 더 높은 정확성을 나타내었다. 소셜 네트워크에서 핫 토픽은 사용자들의 의견을 강하게 반영하기 때문에 긍정 또는 부정으로 분류된 트윗은 핫 토픽을 포함할 확률이 높다. 하지만 이 분류 기법은 크기가 작은 데이터 집합을 이용하여 평균적인 성능을 조사하였으므로 크기가 큰 데이터 집합에 이 기법을 적용했을 때 성능의 효율성을 보장하기 어렵다. 따라서 대용량의 데이터 집합을 대상으로 한 분석 기법이 요구된다.

RuiGuo Yu[6]은 뉴스와 토픽들 사이의 유사도를 측정하기 위한 시간적 거리 인자를 이용하여 토픽 검출 기법을 제안한다. 크롤러를 이용하여 얻은 뉴스 제목, 출처, 출판일, 내용 등의 정보를 토대로 증분 TF-IDF 알고리즘과 시간적 거리 인자를 고려하여 토픽을 검출하게 되는데, 뉴스와 토픽들 사이의 유사도가 클수록 해당 뉴스 문서는 토픽의 일부분으로 나타난다. 여기서 유사도는 뉴스 문서의 게시 시간이 최근 업데이트된 토픽의 시간과 얼마나 비슷한가를 통해 얻어내며 토픽과 비슷하게 나타난 문서는 토픽의 핫함 정도에 기여하게 된다. 이 유사도를 측정하기 위한 증분 TF-IDF 알고리즘은 특정 단어를 포함하는 과거와 현재의 뉴스 문서 빈도수를 계산한다. 이 계산을 통하여 단어와 문서 사이의 가중치를 계산하게 되는데 가중치를 통해 토픽과 뉴스 문서 사이가 얼마나 유사한가를 판단할 수 있다. 기존의 TF-IDF(Term Frequency - Inverse Document Frequency)[7]는 정보 검색과 텍스트 마이닝에서 이용하는 가중치로 여러 문서로 이루어진 문서군이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 통계적 수치이다. TF(Term Frequency)는 특정한 단어가 문서 내에 나타나는 빈도를 나타낸 것으로 값이 높을수록 문서에서 중요하다고 생각할 수 있다. 하지만 단어 자체가 문서군 내에서 자주 사용되는 경우, 이것은 그 단어가 흔하게 등장한다는 것을 의미한다. 이것을 DF(Document Frequency)라고 하며, 이 값

의 역수를 IDF(Inverse Document Frequency)라고 한다. IDF 값은 문서군의 성격에 따라 결정된다. 예를 들어, '원자'라는 단어는 일반적인 문서들 사이에는 잘 나오지 않기 때문에 IDF 값이 높아지고 문서의 핵심어가 될 수 있지만, 원자에 대한 문서를 모아놓은 문서군의 경우 이 단어는 상투어가 되어 각 문서들을 세분화하여 구분할 수 있는 다른 낱말들이 높은 가중치를 얻게 된다. TF-IDF는 TF와 IDF를 곱한 값이다. 문서 d 내에서 단어 t의 총 빈도를 f(t,d)라 할 경우, 가장 단순한 tf 산출 방식은 tf(t,d) = f(t,d)로 표현된다. |D|는 문서 집합 D의 크기 또는 전체 문서의 수이며, $|\{d \in D : t \in d\}|$는 단어 t가 포함된 문서의 수이며, 단어가 전체 말뭉치 안에 존재하지 않을 경우 이는 분모가 0이 되는 결과를 가져온다.

$$idf(t, D) = \log \frac{|D|}{|d \in D : t \in d|} \quad (1)$$

기존 기법들은 최근에 추가된 뉴스 문서의 수를 고려할 수 없게 되므로 증분 TF-IDF 알고리즘을 이용한다. 시간적 거리 인자 값은 뉴스 문서의 게시 시간과 토픽에 대해 클러스터링된 최근 뉴스의 게시 시간을 기준으로 도출되며 뉴스와 토픽의 시간 차가 작을수록 시간적 거리 인자 값이 크게 나타난다. 수동으로 직접 분류한 토픽을 기준으로 기존의 다른 토픽 검출 기법과 [6]에서 제안하는 기법을 재현율, 정확률, F-measure를 통해 성능 평가한 결과 [6]에서 제안하는 기법이 짧은 용어의 토픽을 더 정확하게 잘 검출해냈다는 것을 알 수 있다. 하지만 [6]은 토픽이 새롭게 계속해서 생성되는데 비해 뉴스 문서의 개수는 토픽의 개수만큼 비례해서 많이 생성되지 않는다. 따라서 이 기법은 빠르게 증가하는 대용량의 정보로부터 토픽을 검출하기에는 적합하지 않으므로 실시간으로 토픽을 검출할 수 있는 연구 분석이 필요하다.

H.Kim[4]에서는 기존 기법에서의 특정 단어의 출현 빈도수만을 고려하여 핫 토픽을 검출하였을 때의 문제점을 보완하여, 출현 빈도수의 비율을 활용한 기법을 제안하였다. 단순히 특정 단어의 출현 빈도수만을 고려했을 경우, 실제 핫 토픽이 아닌 일상적으로 빈번히 활용되는 단어들이 핫 토픽으로 검출되는 문제점이 발생한다. 하지만 [4]와 같이 시간에 따른 출현 빈도수의 비율을 고려할 경우, 일상적으로 자주 활용되는 단어들은 비율의 변화가 크지 않지만 실제 이슈가 되는 단어들은 급격한 비율의 변화를 보이므로 이러한 단어들을 핫 토픽으로 도출한다. 이러한 접근은 매우 간단하면서도 합리적인 방법이지만 사전 확인이 이루어지지 않은 불특정 다수의 글을 대상으로 하는 만큼 왜곡된 결과의 도출 가능성으로 인해 신뢰도가 높다고 말하기 어렵다. 따라서 분석 결과의 신뢰성 향상 및 보장할 수 있는 기법의 연구가 필요하다.

Ⅲ. 제안하는 핫 토픽 예측

1. 제안하는 기법 구조

기존 핫 토픽 검출 기법에서는 단순히 특정 단어의 출현 빈도수만을 기준으로 핫 토픽 검출하기 때문에 최종 도출된 결과의 높은 신뢰도를 보장하는 것이 불가능하였다. SNS 환경에서는 신뢰도와 연관 지을 수 있는 다양한 요인들이 존재하지만 사용자의 영향력이 신뢰도와 가장 큰 연관을 가진다. 즉, 사용자의 영향력이 클수록 신뢰도가 높다고 할 수 있다.

본 논문에서는 변형된 TF-IDF 알고리즘을 이용하여 소셜 네트워크 사용자의 영향력을 고려한 신뢰성 높은 핫 토픽 예측 기법을 제안한다. 기존 기법과 같이 단어 출현 빈도수만을 고려하여 핫 토픽을 검출했을 때 검출 결과의 신뢰성을 보장할 수 없고 불특정 다수가 올린 검열되지 않은 글들을 다루는 만큼 검출 결과의 신뢰성을 향상시킬 필요가 있으므로 사용자의 영향력을 추가적으로 고려한다. 본 논문에서 핫 토픽은 순간적으로 발생하는 키워드의 집합이므로 트윗 데이터들로부터 이러한 집합을 추출하기 위하여 시간적 속성을 고려한 TF-IDF 알고리즘을 이용하여 단어에 가중치를 부여한다. 변형된 TF-IDF 알고리즘은 과거 시간 슬롯에 대한 현재 시간 슬롯의 역 문서빈도수의 변화량을 이용하여 시간에 따라 단어별로 해당 단어가 출현한 트윗의 수를 고려한다. 이를 통해 트윗 데이터들 사이에서 해당 단어가 얼마나 언급되었는지를 판단하고, 이를 통하여 순

간적으로 발생하는 단어들의 집합을 추출할 수 있다. 전체 트윗 데이터 집합을 기준으로 의미가 있는 핫 토픽을 모두 추출한다. 또한, 트위터 사용자의 영향력을 고려하여 각 단어에 대한 핫 토픽 지수를 추가적으로 부여한 다음 지수 값의 랭킹으로 핫 토픽을 예측한다.

[그림 2]는 제안하고 있는 전체적인 핫 토픽 예측 시스템의 처리 과정을 나타낸다. 트위터 사용자 영향력을 기반으로 트윗에 가중치를 부여하는 방식으로 이루어져 있다. 우선 트윗 데이터로부터 변형된 TF-IDF 알고리즘을 이용하여 순간적으로 발생하는 단어들의 집합을 추출한다. 트윗에 나온 단어의 출현 빈도수와 사용자의 영향력을 함께 고려하여 핫 토픽 지수를 계산하고 시간에 따라 단어별 핫 토픽 지수가 변화하였는지를 계산한다. 그 비율 값을 기준으로 단어를 랭킹화한 후 Top N까지의 핫 토픽을 예측한다. 이렇게 트위터 사용자 영향력을 추가적으로 고려함으로써 보다 신뢰성 높은 결과를 도출할 수 있게 된다.

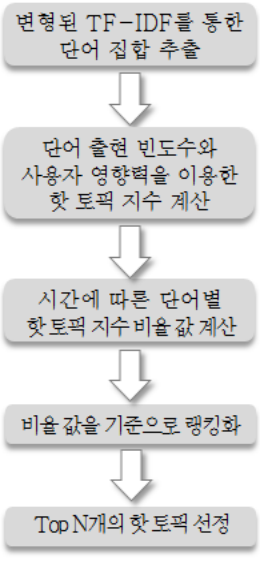


그림 2. 순서도

2. 단어 출현 빈도수 계산

핫 토픽은 순간적으로 많이 언급되는 단어들의 집합으로 정의된다. 핫 토픽을 예측하기 위해서는 우선 단어의 출현 빈도수를 고려해야하는데 이를 고려함으로써 순간적으로 이슈화되는 단어들의 집합을 추출할 수 있다. 단어 출현 빈도수를 계산하기 위한 일반적인 방법은 TF-IDF 알고리즘을 이용하는 것인데 기존의 알고리즘을 사용할 경우 순간적으로 나타나는 단어들을 추출하기 위한 시간적 속성을 무시하게 된다. 따라서 단순히 단어들을 추출하기 위한 TF-IDF 알고리즘을 변형하여 순간적인 키워드들을 추출하기 위한 TF-IDF 알고리즘을 변경한다. 제안하는 기법에서의 TF 산출 방식은 단어 w가 트윗에 한번이라도 출현한다면 1 값을 부여하고, 아니면 0 값을 부여하는 불린 빈도 방식을 이용한다. IDF는 시간에 따른 IDF의 변화량을 측정하여 도출이 되며 식 (2)와 같다. 특정 시간의 슬롯을 i라고 하였을 때 idf_i는 현재 시간 슬롯 i에서의 idf 값을 나타내고 $idf_{0,i-1}$는 0부터 i-1까지 시간 슬롯의 idf 값을 나타낸다. 식 (2)는 과거 시간 슬롯에 대한 현재 시간 슬롯의 idf 값 변화량을 나타내며 이를 통하여 순간적으로 나타나는 단어들의 집합 문서를 추출할 수 있다. 이를 통해 전체 트윗 데이터 집합을 기준으로 하여 의미가 있는 모든 핫 토픽을 추출한다.

$$IDF = \frac{idf_i}{idf_{0,i-1}} \quad (2)$$

기존의 TF-IDF 알고리즘을 통해 문서에 나타난 단어를 추출할 경우 시간적인 속성을 고려하지 않아 순간적으로 발생하는 키워드를 추출할 수 없게 된다. 본 논문에서 핫 토픽은 순간적으로 발생하는 단어들의 집합이므로, 이 집합을 추출하기 위해서는 변형된 TF-IDF 알고리즘을 사용하여 시간적인 속성을 고려해야 한다.

[그림 3]과 같이 시간 슬롯 당 트윗이 다음과 같다고 가정할 때 $idf_{0,i-1}$는 00시에서 특정 단어가 출현한 트윗 수의 역수를 말하며 idf_i는 01시에서 특정 단어가 출현한 트윗 수의 역수를 나타낸다. 예를 들어, '무한도전'이라는 단어는 00시에 '무한도전'이 한 번 등장하므로 $idf_{0,i-1}$은 1의 역수 값인 1이고, 01시에 '무한도전'이 2번 등장하므로 idf_i은 2의 역수 값인 1/2로 계산된다. 따라서 '무한도전'의 IDF 값은 idf_i를 $idf_{0,i-1}$로 나눈 1/2 값을 나타낸다.

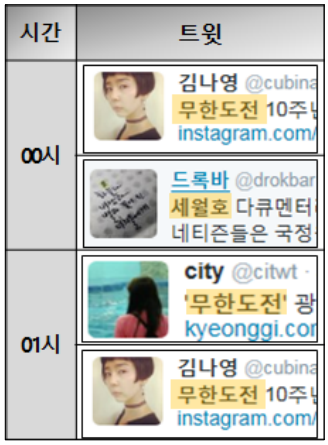


그림 3. 시간 슬롯 당 트윗

3. 사용자 영향력 판단

트위터에는 다양한 사용자들의 트윗이 등록된다. 그 중에서 많은 사람들에게 유명하고 영향력이 높은 사용자에 의한 트윗도 있고, 상대적으로 그렇지 못한 사용자에 의한 트윗도 존재한다[8]. 그러므로 기존 기법에서와 같이 단순히 특정 단어의 출현 빈도수만을 기준으로 핫 토픽을 검출할 경우 각 사용자의 영향력을 무시하고 동일한 가중치를 부여하는 것이므로 도출된 결과에 신뢰도이 저하된다. 사용자 영향력을 측정하여 영향력이 높은 사용자가 등록한 트윗에 가중치를 부여한다면 핫 토픽 검출의 신뢰도를 향상 시키는 것이 가능하다. SNS 환경에서는 신뢰도와 연관 지을 수 있는 다양한 요인들이 존재한다. 하지만 사용자들이 최근 SNS을 이용하는 목적은 사용자들이 올린 글들을 이용하여 정보를 공유하고 검색하는 데에 있으므로 보다 영향력 있는 사용자가 남긴 글에 주목을 하게 된다. 따라서 사용자의 영향력이 신뢰도와 가장 큰 연관을 가지므로 사용자의 영향력이 클수록 정보 검색의 신뢰도가 높다고 할 수 있다.

본 논문에서는 트위터 상에서 사용자가 수행 가능한 다양한 활동 중에서 영향력과 높은 상관관계에 있는 세 가지 요소(팔로워의 수, 트윗의 수, 멘션의 수)를 기준으로 사용자의 영향력을 도출한다. [그림 4]는 사용자 영향력을 판단하는 처리 절차를 나타낸 것이다. 사용자 영향력을 판단하기 위한 기준인 팔로워의 수, 멘션의 수, 리트윗의 수를 이용하여 각요소의 지수 값들을 우선적으로 계산한 후 세 요소의 지수 값들마다 각각 로그 값을 취해준다. 그리고 이 세 값들을 모두 합하여 최종적인 사용자 영향력 지수 값을 도출한다.

그림 4. 사용자 영향력 판단

식 (3)은 제안하는 기법에서의 사용자 영향력 지수를 나타내며 팔로워의 수, 리트윗 수, 멘션의 수를 기준으로 각 영향력 지수를 모두 합한 값으로 도출된다. 이때, 각 요소들은 서로 연관성이 없으므로[9] 각 요소별 영향력 지수를 합하여 최종 영향력 지수를 계산한다. 각 요소들의 분포는 지수 분포 형태를 나타내므로 한 요소에 대해 영향력이 치우치지 않도록 각 요소들에 대해 로그 값을 취하였다.

$$I_{U_A} = \log(I_{U_A^f}) + \log(I_{U_A^t}) + \log(I_{U_A^m}) \quad (3)$$

식 (4)는 사용자 영향력을 도출하기 위한 구성 요소로써 특정 사용자의 팔로워의 수를 나타내며 팔로우한 사람이 특정 사용자의 트윗에 대해 갖는 관심의 정도이다. 사용자 A의 전체 팔로워의 수를 정규화 상수 α로 나눈 값으로 계산된다. 특정 사용자의 팔로워의 수가 많을수록 사용자 영향력이 높다고 측정하였다.

$$I_{U_A^f} = \frac{\sum_{f \in U_A^f} Followers}{\alpha} \quad (4)$$

식 (5)는 사용자 영향력을 도출하기 위한 구성 요소로써 특정 사용자의 리트윗의 수를 나타내며 특정 사용자의 트윗 당 평균 리트윗 비율 및 이것을 리트윗하는 팔로워들의 전파력을 고려하였다. 이 값은 사용자 A의 전체 트윗에 대한 리트윗의 비율과 전체 팔로워의 수에

대한 팔로워의 팔로워 수의 값을 곱한 후 정규화 상수 β로 나눈 값으로 계산된다. 이 수치가 클수록 사용자 영향력이 높다고 측정하였으며 이 때 팔로워들의 전파력은 기준이 되는 특정 사용자의 팔로워들의 평균 팔로워 수를 기준으로 한다. 이에 따라 특정 사용자가 트윗을 올렸을 경우 평균적으로 얼마나 많은 사용자들이 해당 트윗을 접하게 되는지를 측정하였다.

$$I_{U_A^{rt}} = \frac{\frac{\sum_{rt \in U_A^{rt}} Retweets}{\sum_{t \in U_A^t} Tweets} \times \frac{\sum_{U_f \in U_A^f} Followers}{\sum_{f \in U_A^f} Followers}}{\beta} \quad (5)$$

식 (6)은 사용자 영향력을 도출하기 위한 구성 요소로써 특정 사용자의 멘션의 수를 나타낸다. 특정 사용자의 멘션 수신 수가 높다는 것은 그만큼 많은 사람들에게 관심을 받고 있다는 것을 나타낸다. 이 수식은 사용자 A의 전체 멘션 수를 정규화 상수 γ로 나눈 값으로 계산된다. 팔로워의 수와 마찬가지로 수신 멘션의 수가 많을수록 사용자의 영향력이 높다고 측정하였다. α, β, γ는 각 요소를 정규화하기 위한 정규화 상수들이다.

$$I_{U_A^m} = \frac{\sum_{m \in U_A^m} Mentions}{\gamma} \quad (6)$$

정규화 상수 α, β, γ가 1000이라고 가정했을 때 [표 1]을 이용하여 사용자 A의 세 가지 영향력 지수를 계산할 경우 식 (7)~식 (9)와 같이 계산한다. 따라서 사용자 A의 총 영향력 지수는 이 세가지 요소를 모두 합한 0.6 값으로 도출된다. 식 (7)은 전체 팔로워의 수 100을 α로 나눈 0.1 값이 도출되었으며, 식 (8)은 전체 트윗 수 150에 대한 전체 리트윗의 수 300을 나눈 값을 전체 팔로워의 수 100에 대한 팔로워의 팔로워 수 2000을 나눈 값과 서로 곱하여 β로 나눈 0.4 값을 도출하였다. 식 (9)는 전체 멘션 수인 100을 γ로 나눈 값인 0.1 값을 도출해내었다.

$$I_{U_A^f} = \frac{100}{1000} = 0.1 \quad (7)$$

$$I_{U_A^{rt}} = \frac{\frac{300}{150} \times \frac{2000}{100}}{1000} = 0.4 \quad (8)$$

$$I_{U_A^m} = \frac{100}{1000} = 0.1 \quad (9)$$

표 1. 영향력 지수를 도출하기 위한 사용자 A의 요소

팔로워의 수	100
트윗 수	150
리트윗 수	300
팔로워의 팔로워 수	2000
멘션 수	100

4. 최종 핫 토픽 지수 계산 및 랭킹

팔로워의 수, 리트윗 수, 멘션의 수를 기준으로 각 요소별 영향력 지수를 모두 합하여 최종 핫 토픽 지수를 도출이 되면, 이를 이용하여 시간에 따라 핫 토픽 지수가 얼마나 변화하였는지를 이용하여 변화량이 큰 단어들을 핫 토픽으로 예측한다. 식 (10)은 제안하는 기법에서의 단어별 핫 토픽 지수를 나타내며, 단어 w에 대한 시간 t에서의 사용자 영향력 지수와 시간 t-1에 대한 사용자 영향력 지수의 합에 대한 차의 비율로써 도출한다. 기존 기법에서는 단순히 각 단위 시간 동안 특정 단어의 출현 빈도만을 고려한 것에 반해, 제안하는 기법에서는 특정 단어의 출현 빈도뿐만 아니라 해당 단어를 포함한 트윗을 작성한 사용자의 영향력을 가중치로 할당하고 그 전체 합을 특정 단어의 핫 토픽 지수로 계산한다.

$$R_t^w = \frac{I_t^w - I_{t-1}^w}{I_t^w + I_{t-1}^w} \quad (10)$$

[그림 5]는 사용자의 영향력 지수를 고려하여 핫 토픽 지수를 계산하는 과정을 나타낸 것이다. 사용자의 팔로워 수, 리트윗 수, 멘션 수를 고려하여 사용자별 영향력 지수를 계산하고 단어 출현 빈도와 사용자 영향력을 이용하여 키워드별 토픽 지수를 계산한다. 계산된 키워드별 토픽 지수를 시간에 따른 비율로 계산하여 최종 핫 토픽 지수를 산출한다.

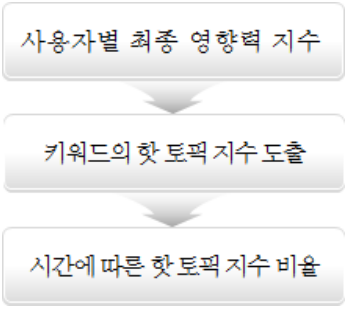


그림 5. 최종 핫 토픽 지수 도출 과정

[그림 6]은 사용자 영향력 지수를 이용하여 단어별 최종 핫 토픽 지수를 도출하는 과정을 예로 나타낸 그림이다. 시간에 따른 트윗 데이터와 각 사용자의 영향력이 다음과 같다고 가정했을 시, 00시 트윗에 출현한 '무한도전'이라는 단어가 1이 카운트될 때, 그 트윗의 사용자 영향력 지수 0.6을 곱한 0.6이라는 핫 토픽 지수가 나오게 된다. 00시 두번째 트윗에 출현한 '세월호'라는 단어의 경우도 1이 카운트되면서 그 트윗의 사용자 영향력 지수인 0.4가 곱해진 0.4라는 지수 값이 나오게 된다. 시간별 단어의 지수 값을 모두 계산한 후 시간에 따른 단어별 최종 지수의 비율을 계산하게 되는데 '무한도전'의 경우 01시의 핫 토픽 지수인 0.2값과 00시에 핫 토픽 지수인 0.6값을 이용하여 합 분의 차로 비율 값을 도출하게 된다. 시간에 따른 핫 토픽 지수의 변화량인 비율 값이 클수록 핫 토픽에 가깝다. 그러므로 비율 값을 랭킹화하여 Top N까지의 핫 토픽 키워드를 예측한 후, 이 상위 N개의 단어를 핫 토픽으로써 사용자에게 최종 추천한다.

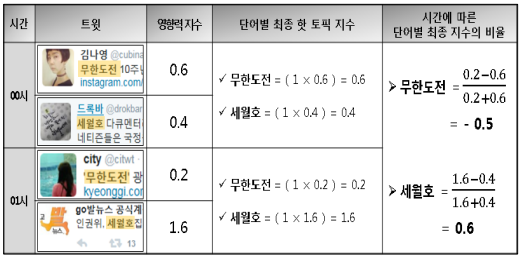


그림 6. 사용자 영향력 지수를 이용한 최종 핫 토픽 지수 도출 과정

Ⅳ. 성능평가

본 장에서는 제안하는 핫 토픽 검출 기법과 기존의 기법과의 성능 비교 평가를 통해 제안하는 기법의 우수성을 기술한다. 제안하는 기법에서 사용한 실험 환경은 다음과 같다. 실험 데이터는 실시간으로 Twitter Streaming API [10]를 통해 2015년 4월 1일부터 2015년 5월 31일까지 1,215,342개의 샘플 데이터를 수집하였다. 수집된 트위터 데이터를 통해 시간에 따른 IDF 값을 도출하기 위한 트윗들과 사용자의 영향력을 고려하기 위한 사용자의 팔로워 수, 리트윗 수, 멘션 수 등의 정보를 추출하였다. JAVA를 이용하여 전체적인 핫 토픽 검출 시스템의 성능을 평가하였으며, 데이터베이스는 MySQL 5.6.23 버전을 사용하였다.

표 2. 성능평가 환경

항목	값
CPU	Intel(R) Core(TM) i5-4440 CPU 3.10GHz
RAM	6.00 GB
사용 언어	Java(TM) SE Runtime Environment (build 1.8.0_31-b13)
데이터베이스	MySQL 5.6.23

제안하는 기법을 적용하여 Top N까지의 핫 토픽을 예측하였고 예측된 핫 토픽들 중 특정 핫 토픽에 대하여 시간에 따른 핫 토픽 지수의 변화율을 기존 기법과 비교 실험하였다. 또한 현재 시점에서 핫 토픽을 검출하여 나온 결과와 제안하는 기법을 통해 예측된 핫 토픽이 얼마나 일치하는가를 통해 제안하는 기법의 성능을 평가하였다. 현재 시점의 핫 토픽을 검출하기 위하여 TF-IDF 알고리즘을 통한 단어의 출현 빈도수와 이를 통해 우선적으로 검출된 토픽들을 대상으로 불용어와 일상적으로 많이 쓰이는 단어를 최종적으로 제외하여 최종 핫 토픽을 검출하는 방식을 이용하였다. 현재 시점의 핫 토픽을 검출하는 기법을 기준으로 제안하는 기법을 통해 예측된 핫 토픽 검출 결과가 성능 평가하기 위해 [표 3]을 이용하여 정확률, 재현율, F-measure 비교 평가한다. 정확률은 식(11)과 같이 현재 시간대의 핫 토픽을 검출하는 기법을 통해 얻은 핫 토픽들을 기준으로 했을 때 예측 기법을 통해 얻은 핫 토픽이 얼마

나 일치하는지를 나타내는 비율이며, 현재 시간대의 핫 토픽 검출 기법을 통해 얻은 실제 핫 토픽인 a와 c에 대한 예측 기법을 통해 얻은 실제 핫 토픽인 a의 비율로 구한다. 재현율은 식 (12)과 같이 예측 기법을 통해 검출된 핫 토픽들을 기준으로 했을 때 현재 시간대의 기법을 통해 얻은 핫 토픽들과 얼마나 일치하는지를 나타내는 비율이고, 현재 시간대의 핫 토픽 검출 기법을 통해 얻은 전체 결과 a, b에 대한 예측 기법을 통해 얻은 실제 핫 토픽인 a의 비율로 도출한다. F-score는 정확률과 재현율의 조화 평균으로 계산되며, 식 (13)과 같다.

표 3. Precision, Recall, F-score를 측정하기 위한 지표

구분		예측하는 제안/기존 기법	
		True	False
현재 시간대의 핫 토픽을 검출하는 방법	True	a	b
	False	c	d

$$\text{정확률(Precision)} = a / (a+c) \quad (11)$$

$$\text{재현율(Recall)} = a / (a+b) \quad (12)$$

$$F-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

[표 4]와 [표 5]는 각각 기존 기법과 제안하는 기법을 적용하여 나온 결과 중 일부인 2015년 5월 1일부터 7일까지의 상위 10번째 핫 토픽 키워드들의 집합을 나타낸다. 이 결과 데이터 집합을 토대로 위 방법대로 성능평가를 비교 분석하게 된다.

표 4. 기존 기법으로 예측된 핫 토픽 키워드 집합

날짜	예측된 핫 토픽 키워드 Top 10
15.05.01	네팔, 구호단, 새정치연합, 시내버스, 폭행, 김나영, 배터리, 삼성폰, 제2롯데월드, 박원순
15.05.02	오바마, 시내버스, 김정은, 러시아, 장윤정, 하이마트, 양미라, 유가족, 불법시위, 노동절
15.05.03	김정은, 삼성전자, 메이웨더, 마리텔, 서유리, 어벤져스, 네팔, 강진, 생존자, 탈북
15.05.04	서유리, 박준형, 빈지노, 황금락카, 연말정산, 정주리, 손흥민, 복면가왕, 리버풀, 문재인
15.05.05	어린이날, 정주리, 황금락카, 쓰나미, 인종차별, 오승환, 이연복, 후아유, 냉장고, 이재용
15.05.06	후아유, 연평해전, 조권, 연말정산, 손현주, 국민연금, 쿠바, 식스틴, 프로듀사, 홍준표
15.05.07	수요미식회, 봉태규, 하시시박, 조권, 윤건, 장서희, 컴백, 전효성, 추신수, 데이터

표 5. 제안하는 기법으로 예측된 핫 토픽 키워드 집합

날짜	예측된 핫 토픽 키워드 Top 10
15.05.01	네팔,구호단,새정치연합,지진,갤럭시,김나영,배터리,삼성폰,제2롯데월드,박원순
15.05.02	북한, 시내버스, 김정은, 러시아, 장윤정, 하이마트, 양미라, 유가족, 불법시위, 노동절
15.05.03	김정은, 삼성전자, 메이웨더, 마리텔, 서유리, 김수현, 네팔, 강진, 생존자, 탈북
15.05.04	서유리, 박준형, 빈지노, 황금락카, 연말정산, 정주리, 손흥민, 복면가왕, 안철수, 문재인
15.05.05	어린이날, 정주리, 황금락카, 쓰나미, 인종차별, 복면가왕, 이연복, 후아유, 냉장고, 이재용
15.05.06	후아유, 연평해전, 조권, 연말정산, 박진영, 국민연금, 쿠바, 식스틴, 프로듀사, 홍준표
15.05.07	수요미식회, 봉태규, 하시시박, 조권, 윤건, 장서희, 요금제, 체스, 추신수, 데이터

[그림 7]은 [표 5]에서 검출된 핫 토픽 중 특정 단어 '세월호'에 대한 핫 토픽 지수 변화율을 나타낸 그래프이다. 분석 기간 내에 지속적으로 세월호에 대한 많은 트윗이 발생하여 핫 토픽으로 선정되었으며, 지속적으로 언급되는 단어에 대해 기존 기법에 비해 검출 결과의 신뢰성이 최대 22% 향상하였다. [그림 8]은 '부활절'에 대한 핫 토픽 지수 변화율을 나타내며, 부활절(4월 5일)까지 점차 증가하던 관련 트윗이 해당일 이후 기존 기법에 비해 핫 토픽 지수의 비율이 감소하여 핫 토픽에서 제외되었다는 것을 알 수 있다. 특정 일이나 사건에 따라 급격하게 언급되는 단어에 대해서 기존 기법에 비해 검출 결과의 신뢰성이 최대 39% 향상하였다. [그림 7]과 [그림 8] 모두 기존 기법보다 제안하는 기법의 성능이 우수함을 나타내었는데, 제안하는 기법에서 추가적으로 고려한 사용자 영향력 때문에 결과의 신뢰성 측면이 높아진 걸 알 수 있다.

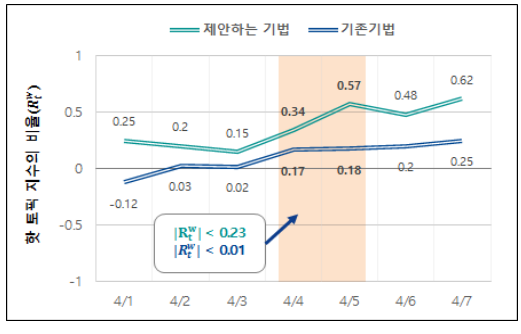


그림 7. '세월호' 에 대한 핫 토픽 지수 변화율

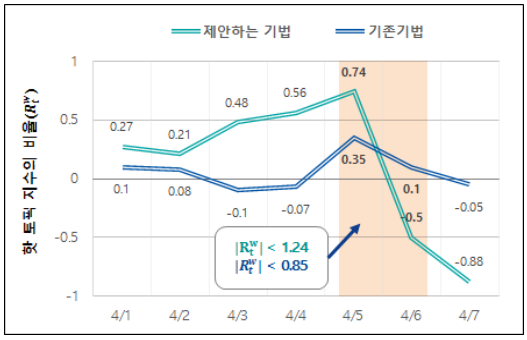
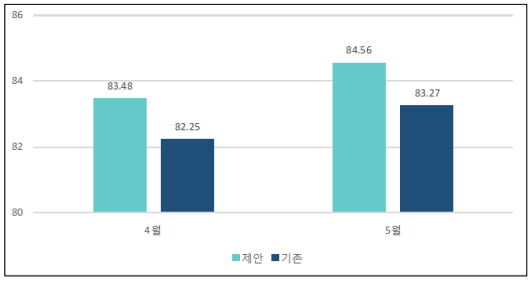


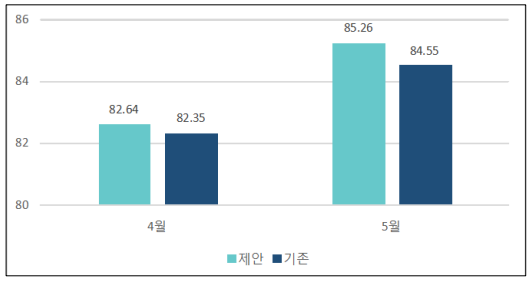
그림 8. '부활절'에 대한 핫 토픽 지수 변화율

현재 시간의 핫 토픽을 검출하는 방법에 기준하여 제안하는 기법과 기존의 기법과의 성능 비교를 위해 검출 정확률(Precision), 검출 재현율(Recall), F-measure[11]을 사용하였다. 현재 시간 t에서 핫 토픽을 검출한 후, 제안하는 기법과 기존 기법을 통해 시간 t의 핫 토픽을 예측한 결과를 비교 분석하였다.

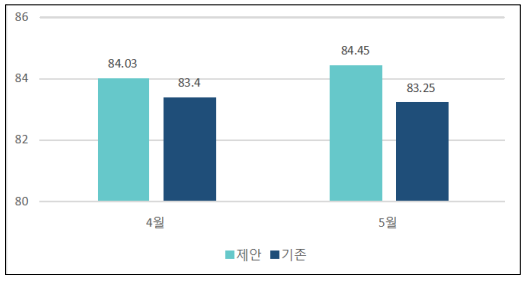
[그림 9]는 현재 시점에서 핫 토픽을 검출하여 나온 결과와 제안하는 기법을 통해 예측된 핫 토픽이 얼마나 일치하는가를 4월과 5월을 기준으로 평가한 실험이다. 측정 결과 제안하는 기법이 (a)의 경우 4월에는 83.48%, 5월에는 84.56%로 기존 기법에 비해 높은 성능을 보였다. 마찬가지로 (b)와 (c)의 경우도 제안하는 기법을 통한 검출 결과가 신뢰성이 더 높게 나왔다.



(a) Recall



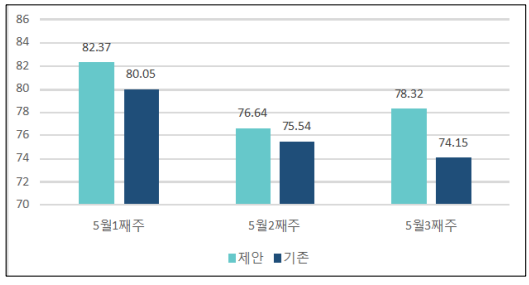
(b) Precision



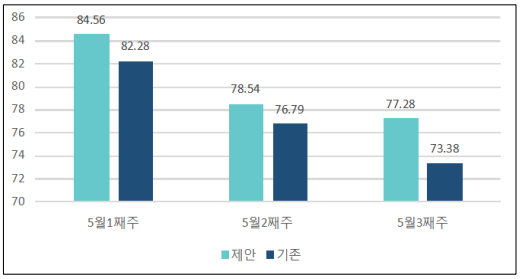
(c) F-score

그림 9. Recall, Precision, F-measure를 이용한 성능 비교

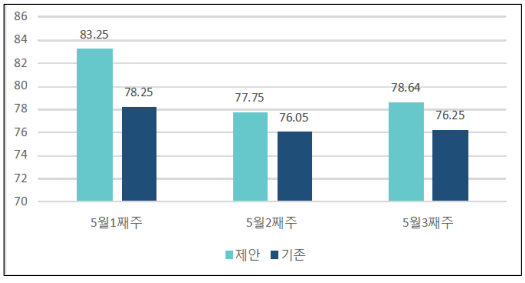
[그림 10]은 5월 3주에서 5월 1주까지 제안하는 기법을 통해 해당 날짜로부터 예측한 핫 토픽과 해당 날짜에서 실제 검출된 핫 토픽이 날짜가 가까워질수록 얼마나 일치하는지를 정확률, 재현율, F-measure 측정을 통하여 비교 분석한 결과이다. 제안하는 기법을 통한 결과가 기존 기법의 결과보다 대체적으로 높은 성능을 보였으며, 3주에서 1주로 오면서 해당 날짜에서 현재 검출된 핫 토픽과 유사하다는 점을 알 수 있다.



(a) Recall



(b) Precision



(c) F-score

그림 10. 시간에 따른 Recall, Precision, F-score의 비교

기존의 예측 기법을 기반으로 할 경우 단어의 출현 빈도수만을 고려하기 때문에 결과를 도출하는 과정에서 그 결과의 신뢰성을 보장하기 어렵다. 하지만 제안하는 예측 기법의 경우 변형된 TF-IDF 알고리즘을 통해 순간적으로 발생하는 단어들의 집합을 추출하고 사용자의 영향력을 추가적으로 고려한다. 이로 인해 전체적으로 정확도, 재현율이 모두 향상되었다.

V. 결론 및 향후 연구

본 논문에서는 기존에 제안된 핫 토픽 검출 기법의 문제점을 제시하고 변형된 TF-IDF 알고리즘을 사용하여 소셜 네트워크에서 사용자의 영향력을 고려한 핫 토픽 검출 기법을 제안하였다. 기존의 핫 토픽 예측 기법의 경우 단어 출현 빈도수만 고려하여 결과를 도출하기 때문에 결과적으로 신뢰성을 보장할 수 없다는 문제점을 가지고 있다. 본 논문은 변형된 TF-IDF 알고리즘을 기반으로 사용자의 영향력을 추가적으로 고려하여 핫 토픽을 예측한다. 이 과정에서 결과의 신뢰성을 보장할 수 없는 문제를 해결하고, 결과의 신뢰성을 높임으로써 전체적인 예측 기법의 성능이 향상되었다. 제안하는 기법은 선거 홍보와 서비스나 상품에 관한 시장조사, 그리고 기업의 홍보나 고객 불만 접수 등에 활용 가능하다. 순간적으로 발생하는 키워드의 집합을 핫 토픽이라고 정의하고, 변형된 TF-IDF 알고리즘을 통해 시간적인 속성을 고려함으로써 실제로 실시간 이슈화되는 단어들의 집합을 추출하게 하였다. 또한, 사용자의 영향력을 추가적으로 고려함으로써 단어 출현 빈도수만을 고려했을 때보다 검출 결과의 신뢰성과 정확도가 높아지도록 수행하였다. 성능평가 결과, 기존 기법에 비해 핫 토픽 지수의 비율 폭이 22~39% 증가했으며 검출 결과의 신뢰성 향상을 확인할 수 있었다. 또한 정확률, 재현율, F-measure 측정 결과에서 기존 기법보다 우수한 성능을 보였다. 향후 연구로 사용자의 전문성과 카테고리화를 통해 검출 결과의 정확도를 더 향상시킬 예정이다.

참 고 문 헌

[1] 하일규, “소셜 네트워크 서비스의 연구경향 분석:Twitter 관련 연구 중심”, 한국콘텐츠학회논문지, 제14권, 제9호, pp.567-581, 2014.

[2] 민정식, “트위터 이용과 정치 참여”, 한국지역언론학회, 제12권, 제2호, pp.274-303, 2012.

[3] 허상희, “트위터에서 트윗의 특징과 유형 연구”, 한민족어문학회, 제61집, pp.455-494, 2012.

[4] H. Kim, S. Lee, and S. Kyeong, “Discovering Hot Topics using Twitter Streaming Data,” Proc. International Conference on Advances in Social Networks Analysis and Mining, pp.1215-1220, 2013.

[5] J. Haziq and S. Khushal, “‘Good’ versus ‘Bad’ Opinion on Micro Blogging Networks: Polarity Classification of Twitter,” International Journal of Computer Science and Mobile Computing,

Vol.3, No.8, pp.49-56, Aug. 2014.

[6] RuiGuo Yu, ManKun Zhao, Peng Chang, and MuWen He, “Online hot topic detection from web news archive in short terms,” Proc. IInternational Conference on Fuzzy Systems and Knowledge Discovery, pp.919-923, Aug. 2014.

[7] https://ko.wikipedia.org/wiki/TF-IDF

[8] 이은혜, *SNS 신뢰도에 영향을 미치는 요인에 관한 연구*, 세종대학교, 2012.

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

[10] https://dev.twitter.com/streaming/overview

[11] J. Markus, H. Rainer, and D. Andreas, “On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy,” Proc. International Conference on Document Analysis and Recognition, pp.713-716, 1999

저 자 소 개

노 연 우(Yeon-woo Noh) 준회원

- 2014년 2월 : 충북대학교 정보통신공학과(공학사)
- 2014년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

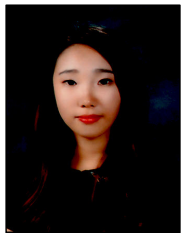
<관심분야> : 데이터베이스 시스템, 빅데이터 등

김 대 윤(Dae-yun Kim) 준회원

- 2015년 2월 : 청주대학교 경영학과/컴퓨터공학과(공학사)
- 2015년 3월 ~ 현재 : 충북대학교 빅데이터학과 석사과정

<관심분야> : 데이터베이스 시스템, RDF, 맵-리듀스, 빅데이터 등

한 지 은(Jieun Han) 준회원



- 2014년 2월 : 충북대학교 정보통신공학과(공학사)
- 2014년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 빅데이터, 프러버넌스 데이터, RDF 등

육 미 선(Misun Yook) 준회원

- 2014년 2월 : 충북대학교 정보통신공학과(공학사)
- 2014년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 빅데이터, 분산처리시스템, 소셜 네트워크 서비스 등

임 종 태(Jongtae Lim) 정회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2011년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2011년 3월 ~ 현재 : 충북대학교 정보통신공학과 박사과정

<관심분야> : 데이터베이스 시스템, 시공간 데이터베이스, 위치기반 서비스, 모바일 P2P 네트워크, 빅데이터 등

복 경 수(Kyungsoo Bok) 종신회원

- 1998년 2월 : 충북대학교 수학과 (이학사)
- 2000년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2005년 2월 : 충북대학교 정보통신공학과(공학박사)
- 2005년 3월 ~ 2008년 2월 : 한국과학기술원 전산학과 Postdoc
- 2008년 3월 ~ 2011년 2월 : ㈜가인정보기술 연구소
- 2011년 3월 ~ 현재 : 충북대학교 정보통신공학과 초빙부교수

<관심분야> : 데이터베이스 시스템, 위치기반서비스, 모바일 P2P 네트워크, 소셜 네트워크 서비스, 빅데이터 등

유 재 수(Jaesoo Yoo) 종신회원

- 1989년 2월 : 전북대학교 컴퓨터공학과(공학사)
- 1991년 2월 : KAIST 전산학과 (공학석사)
- 1995년 2월 : KAIST 전산학과 (공학박사)
- 1995년 3월 ~ 1996년 8월 : 목포대학교 전산통계학과(전임강사)
- 1996년 8월 ~ 현재 : 충북대학교 정보통신공학부 및 컴퓨터정보통신연구소 교수
- 2009년 3월 ~ 2010년 2월 : 캘리포니아주립대학교 방문교수

<관심분야> : 데이터베이스 시스템, 빅데이터, 센서네트워크 및 RFID, 소셜 네트워크 서비스, 분산 객체컴퓨팅, 바이오인포매틱스 등