




# Hot topic prediction considering influence and expertise in social media

Kyoungsoo Bok<sup>1</sup> · Yeonwoo Noh<sup>1</sup> · Jongtae Lim<sup>1</sup> · Jaesoo Yoo<sup>1</sup> 

Published online: 1 January 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

## Abstract

The hot topic detection designed to identify the recent issues and trends employs the analysis of real-time social media activities. The existing schemes suffer from low precision because they focus on keyword occurrence frequency in documents written by the unspecified majority. The existing schemes are incapable of predicting near-future hot topics as they are intended to detect hot topics at a particular time. We propose a new hot topic prediction scheme considering users' influence and expertise in social media. The proposed scheme detects expected near-future hot topics by extracting a set of candidate keywords from social-media posts using the modified TF-IDF. The hot topic prediction index is calculated for each candidate keyword based on the influence and expertise of users who include it in their posts and hot topic predictions are performed based on the change rate over time. Finally, a comparison between existing and proposed hot topic detection schemes demonstrates the proposed scheme's superiority.

**Keywords** Hot topic · Social media · TF-IDF · Influence · Expertise · Prediction

---

✉ Jaesoo Yoo  
yjs@chungbuk.ac.kr

Kyoungsoo Bok  
ksbok@chungbuk.ac.kr

Yeonwoo Noh  
ywnoh@chungbuk.ac.kr

Jongtae Lim  
jtlim@chungbuk.ac.kr

<sup>1</sup> Department of Information and Communication Engineering, Chungbuk National University, Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk 28644, Korea

## 1 Introduction

As web service users' online activities have expanded, the amount of information they generate and share online has also increased. In addition, social media is used as a communication tool for interactions between individuals and groups and to create interdependent relationships [1, 2]. Social media refers to activities, practices, and behaviors of exchanging and sharing information, knowledge, and opinions. Social media is a form of communication based on Web 2.0 that individuals use to share opinions, experiences, and information to create and expand their relationships with others [3, 4]. Leading social media forms include blogs, social networks, message boards, podcasts, Wikis, and Blogs.

With the growing use of social media, social network services (SNSs) have drawn attention as tools for information sharing, building connections, and expressing one's ideas and tastes [5, 6]. SNSs have evolved from information-sharing through social networks to generating and consuming new information [7, 8]. Their structure has evolved to generate and share various types of information while reprocessing them for further sharing [9–13]. Companies use SNS gathered from platforms such as Twitter and Facebook to encourage customer participation, sharing, and conversation [14–17]. In the marketing field, SNS is used to promote products or to identify the reputation of products. SNS analysis is used to develop new products or to predict the performance of products in the manufacturing field. In the CRM field, SNS is used to analyze customer requirements or to identify trends of customers. As information is generated and shared exponentially on SNSs, we require schemes to selectively provide information to individuals and groups [18–20]. Therefore, research on human network analysis, influencer identification, and personalized suggestions is underway [14, 21–24].

Hot topic detection has been studied to identify public opinion or customer trends in various industries [25–28]. A hot topic is an event or a core theme that becomes an issue or interest at a particular time [29–35]. Jeelani and Singh [34] proposed a scheme for detecting hot topics using a machine-learning algorithm on Twitter for the classification of positive and negative tweets. Yu et al. [35] proposed a topic detection scheme that applies temporal distances to measure similarities between news and topics. However, insignificant or unreliable keywords can be identified as hot topics using existing schemes because they focus on the keyword occurrence frequency at specific times and use documents created by the unspecified majority. Moreover, the existing schemes cannot identify what keywords will become hot topics near future because they use data generated at the time.

If we can predict hot topics in the near future, we can keep up with future issues and problems. For example, a company can promote or sell related products through user trend changing derived from the hot topic prediction. In disaster safety, it is possible to identify the aftereffects of events and accidents and to work on countermeasures to minimize losses. In this paper, we propose a new scheme for predicting future hot topics in social media. The proposed scheme incorporates user influence and expertise to identify what keywords will become

hot topics in the near future. Since the documents written by users with a high level of influence and expertise are continuously propagated, the keywords contained in the document can be hot topics in the near future. We consider user influence and expertise to determine the propagation of documents written by the user. We extract candidate keywords using a modified Term Frequency-Inverse Document Frequency (TF-IDF) algorithm to determine changes in keywords across different time intervals. We incorporate user influence and expertise for keywords identified using the modified TF-IDF to increase the hot topic prediction accuracy. Finally, we predict near-future hot topics using the change rate over time.

The rest of this paper is organized as follows. Section 2 describes existing schemes for hot topic detection. Section 3 provides a detailed description of the scheme proposed for hot topic prediction. Section 4 demonstrates the proposed scheme's performance over existing schemes. Section 5 provides the conclusion.

## 2 Related works

TF-IDF is used to identify major keywords in a specific document in information search and text mining [36, 37]. Term Frequency (TF) indicates the frequency at which a particular keyword appears in a document. The higher the frequency, the more important the keyword is in the document. DF indicates the number of documents that include a specific keyword; its reciprocal is the inverse document frequency (IDF). TF-IDF is a value obtained by multiplying TF by IDF. A high TF-IDF for a keyword indicates that the keyword appears frequently in the document of interest but infrequently in other documents.

Yang et al. [25] presented to identify emerging rumor for social media with hot topic. A hot topic detection combining bursty term identification and sentence modeling is performed for rumor identification. To determine the bursty term, skewness score, timeless score and periodicity score are used. The sentence modeling uses a bursty term vector and named entity vector to calculate the similarity between sentences. The bursty term vector is composed of bursty terms identified by the bursty term identification and the named entity vector is made up of named entities contained in a sentence.

Zhu and Yu [31] presented a prerecognition model for detecting hot topic. The prerecognition model finds potential hot topics during the period. The prerecognition clusters the original microblog messages to get topics and their amount, and calculates the velocity and acceleration of the topic. To classify microblog messages into different topics, the topic clustering is performed. Three factors such as topic amount, topic hot velocity, and hot acceleration, is used to detect hot topics. To extract the periodic characteristic of hot topics, the topic life cycle is defined.

Yu et al. [35] proposed a hot topic detection scheme based on the similarity between news and topics. Noting that users want to get information quickly, this scheme detects hot topics with sudden, frequent mentions by taking the following steps: capturing the title, source, publication date, and content of news, removing stopwords, and applying incremental TF-IDF; calculating the cosine similarity

between the news content and a topic to determine the relationship between the news and the topic; determining the news with higher cosine similarity between the news' publish time and the topic's updated time as a part of the topic; calculating the temporal distance between the news' publish time and the topic's updated time; determining the news with a higher temporal distance between the news' publish time and the topic's updated time as a part of the topic; determining a topic's status based on the combination of the cosine similarity and temporal distance.

Kim et al. [32] proposed a hot topic detection scheme based on the change in the keyword occurrence frequency over time on Twitter. Geographic information is used to classify geographic communities, because the geographic communities appear the similar fluctuation patterns of word frequency. To detect hot topics of each day that are not tweeted in the previous day, the rate of word frequency is calculated. However, considering only the keyword appearance frequency results in the problem of identifying frequent everyday keywords as hot topics. They addressed this problem by calculating the change rate of the keyword occurrence frequency over time. They identified keywords with a high change rate as hot topics because frequent everyday keywords have low change rates.

### 3 The proposed hot topic prediction scheme

#### 3.1 Overall procedure

The existing hot topic detection schemes do not guarantee result precision because they detect hot topics based on the frequency of keyword occurrence. Moreover, they are incapable of predicting future hot topics because they detect hot topics at a specific time. This paper presents a hot topic prediction scheme based on user influence and expertise in social media. We use Twitter, a representative service of social media for predicting hot topics. The proposed scheme identifies a set of candidate keywords using modified TF-IDF that incorporates a temporal factor. Documents written by influential users with expertise on social media are more likely to be continuously shared and reprocessed by other users. Therefore, the hot topic prediction involves the determination of user reliability and expertise based on the analysis of various user activities and networks on social media. The hot topic prediction indices of candidate keywords are calculated by considering user reliability and expertise and hot topic are made based on changes in the hot topic prediction indices.

Figure 1 shows the overall procedure of the proposed hot topic prediction. A data collector collects social documents generated in real time, human network, and user activities. A candidate keyword extraction select candidate keywords that are suddenly start being mentioned frequently at a specific time from collected documents using a modified TF-IDF. A user analysis analyzes human network and social media activities and determines user influence and expertise. A hot topic prediction calculates the hot topic prediction index by applying weights to candidate keywords based on influence and expertise and identifies hot topic based on comparisons of change rates of indices over time.

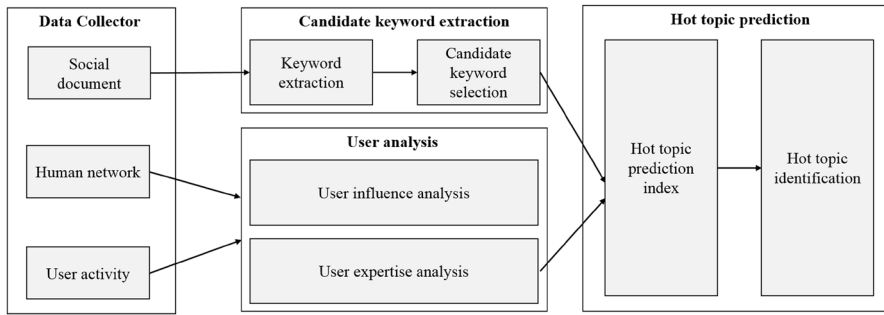


Fig. 1 Overall procedure

### 3.2 Candidate keyword extraction

The first step in hot topic prediction is to extract keywords from documents generated on Twitter using a morphological analyzer, followed by creating a set of meaningful keywords because all extracted keywords are not useful as hot topics. A set of meaningful keywords for hot topic detection are typically generated using TF-IDF. However, TF-IDF cannot extract keywords that are suddenly mentioned frequently because it does not consider the temporal factor. Therefore, the proposed method generates a set of keywords by modifying TF-IDF.

The modified TF-IDF is capable of extracting a set of keywords that are suddenly mentioned frequently because it considers the temporal factor. The modified TF-IDF extracts keywords with high occurrence frequency for a specific time-span from all data on Twitter. The modified TF-IDF extracted a set of keywords that are suddenly mentioned frequently using  $MTF_{t,w}$  and  $MIDF_{t,w}$  as shown in Eq. (1);  $MTF_{t,w}$  is obtained using Eq. (2).  $TF_{t,w}$  denotes keyword  $w$ 's occurrence frequency at time  $t$ , and  $MIDF_{t,w}$  denotes the change rate in IDF between time points as shown in Eq. (3),  $IDF_{t,w}$  denotes the IDF for keyword  $w$  at time  $t$ , and  $IDF_{t-1,w}$  denotes IDF for keyword  $w$  at time  $t - 1$ .

$$MTFIDF_{t,w} = MTF_{t,w} \times MIDF_{t,w} \tag{1}$$

$$MTF_{t,w} = \log(TF_{t,w} + 1) \tag{2}$$

$$MIDF_{t,w} = \frac{IDF_{t,w}}{IDF_{t-1,w}} \tag{3}$$

### 3.3 Influence and expertise

The proposed scheme predicts near-future hot topics rather than identifying present hot topics. Predicting hot topics utilizes user-written documents' propagation on social media. Most activities on social media are made by users. Therefore, the influence and expertise of users who write documents on social media are determined as indicators of documents' propagation because documents written by users with a high level of influence and expertise will likely be shared and reprocessed continuously, becoming

a part of hot topics. Tweeter user activities are highly correlated with influence. User influence is determined by considering followers, retweets, and mentions.  $IF_{t,u}$ , the influence of user  $u$  at time  $t$ , is obtained using Eq. (4).  $FR_{t,u}$  denotes the follower-based influence index;  $RT_{t,u}$  denotes the retweet-based influence index;  $MT_{t,u}$  denotes the mention-based influence index;  $NFR$ ,  $NRT$ , and  $NMT$  respectively denote the normalization constants for  $FR_{t,u}$ ,  $RT_{t,u}$ , and  $MT_{t,u}$ .

$$IF_{t,u} = \frac{FR_{t,u}}{NFR} + \frac{RT_{t,u}}{NRT} + \frac{MT_{t,u}}{NMT} \quad (4)$$

Documents written by users with many followers will likely be frequently shared and reprocessed by followers. In other words, users with more followers are assumed to have greater influence because tweets written by users with many followers can be propagated.  $FR_{t,u}$  denotes the influence index for a user based on the number of followers at time  $t$  as shown in Eq. (5).  $NFR_{t,u}$  denotes the number of followers for user  $u$ ;  $MNFR_t$  denotes the maximum number of follows.

$$FR_{t,u} = \log \left( \frac{NFR_{t,u}}{MNFR_t} + 1 \right) \quad (5)$$

A large number of retweets for a tweet means that the user who writes the tweet is receiving a lot of attention from other users. In addition, when a user has a large number of followers, a tweet may be continuously retweeted to other users. Therefore, the retweet-based influence index considers both the numbers of retweets and followers.  $RT_{t,u}$  is the retweet-based influence index at time  $t$  calculated by Eq. (6).  $NT_{t,u}$  denotes the total number of tweets generated by user  $u$ ,  $NRT_{t,u}$  denotes the number of retweets of tweets written by user  $u$ ,  $MNFR_u$  denotes the maximum number of followers, and  $NFFR_{t,u}$  denotes the number of followers of user  $u$ .  $RT_{t,u}$  considers the average number of retweets per tweet for a user and the propagation of their retweeting followers.

$$RT_{t,u} = \log \left( \frac{NRT_{t,u}}{NT_{t,u}} \times \frac{NFFR_{t,u}}{MNFR_u} + 1 \right) \quad (6)$$

A mention refers to comments for a specific tweet or sending a tweet to a specific user. The large number of mentions means that the tweets will continue to spread and other users will use them. As with the number of followers, a large number of mentions indicate a user's large influence.  $MT_{t,u}$  is the mention-based influence index at  $t$  calculated by Eq. (7).  $NT_{t,u}$  denotes the total number of tweets written by user  $u$  and  $NMT_{t,u}$  denotes the number of mentions for tweets by user  $u$ .

$$MT_{t,u} = \log \left( \frac{NMT_{t,u}}{NT_{t,u}} + 1 \right) \quad (7)$$

A document written by experts in a field will likely be continuously used by users with an interest in that field. A document written by a user with expertise will

also likely have be continuously spread on social media. Moreover, the number of related documents will increase as many users will share and reprocess them. In other words, documents written by experts will likely be part of a hot topic. User expertise indicates users' expertise in the content of the documents they create and calculated using the number of embedded tweets, reliability, and expertise in the user's preferred topic.  $PF_{t,u,c}$  denotes the expertise of user  $u$  at time  $t$  calculated by Eq. (8).  $c$  denotes the preferred topic's category,  $ET_{t,u}$  denotes the expert index based on tweet embedding,  $ST_{t,u}$  denotes the expert index based on user reliability, and  $CE_{t,u,c}$  denotes the expert index based on the user's preferred topic.  $NET$ ,  $NST$ , and  $NCE$  denote the respective normalization constants for  $ET_{t,u}$ ,  $ST_{t,u}$ , and  $CE_{t,u,c}$ .

$$PF_{t,u,c} = \frac{\frac{ET_{t,u}}{NET} + \frac{ST_{t,u}}{NST} + \frac{CE_{t,u,c}}{NCE}}{3} \quad (8)$$

Embedding a tweet refers to the act of quoting a user-written tweet and is considered a proactive user activity. A large number of embedded tweets would indicate that a tweet is both trusted and high quality.  $ET_{t,u}$  is the expert index based on the number of embedded tweets for user  $u$  at time  $t$  calculated by Eq. (9).  $NT_{t,u}$  denotes the total number of tweets written by user  $u$  and  $NET_{t,u}$  denotes the number of embedded tweets for the user.

$$ET_{t,u} = \log \left( \frac{NET_{t,u}}{NT_{t,u}} + 1 \right) \quad (9)$$

Malicious users interfere with normal users' information acquisition by disseminating incorrect information or including the URLs of malicious sites on social media. Therefore, user expertise determination involves determining user reliability for selecting malicious users. Generally, users are connected in social networks through follows and followings. However, malicious users tend to have fewer followers due their weak social network resulting from the dissemination of incorrect information. In other words, malicious users have relatively fewer followers than the number of users they follow. Therefore, the expert index based on user reliability considers the numbers of follows and followings.  $ST_{t,u}$  is the expert index based on user  $u$ 's reliability at time  $t$  calculated by (10).  $NFR_{t,u}$  denotes the number of followers for user  $u$  and  $NFG_{t,u}$  denotes the number of the users that user  $u$  follows.

$$ST_{t,u} = \log \left( \frac{NFR_{t,u}}{NFR_{t,u} + NFG_{t,u}} + 1 \right) \quad (10)$$

Experts are interested in specific topics and generate and share relevant information on social media; therefore, determining expertise involves determining whether a user often mentions a specific topic on Twitter.  $CE_{t,u,c}$ , the expert index for user  $u$ 's preferred topic at time  $t$ , is obtained using Eq. (11).  $c$  denotes the preferred topic category,  $NKW_{t,u}$  denotes the total number of keywords extracted from documents

written by the user on social media, and  $CKW_{t,u,c}$  denotes the number of keywords extracted from documents created on each topic.

$$CE_{t,u,c} = \log \left( \frac{CKW_{t,u,c}}{NKW_{t,u}} + 1 \right) \quad (11)$$

### 3.4 Hot topic prediction

When the user impact and expertise are determined, we predict hot topics for candidate keywords. Figure 2 shows the algorithm of hot topic prediction. Here,  $m$  denotes the number of candidate keywords. The hot topic prediction index is calculated by applying weights for user influence and expertise to candidate keywords extracted using the modified TF-IDF. The hot topic value for each keyword is calculated by comparing the change rates for hot topic prediction indices over time. When hot topic value is sorted from highest to lowest, the top  $k$ -th keywords are predicted as near-future hot topics.

The hot topic prediction index value indicates the likelihood that candidate keywords becomes hot topics.  $HTP_{t,w}$ , the hot topic prediction index for keyword  $w$  at time  $t$ , is calculated by Eq. (12).  $MTFIDF_{t,w}$  denotes the modified TF-IDF for keyword  $w$  at time  $t$  and  $KW_{t,w}$  denotes the keyword weight based on user influence and expertise.  $KW_{t,w}$  is obtained using Eq. (13). When there are  $n$  tweets that include keyword  $w$  at time  $t$ ,  $KW_{t,w}$  is the average influence and expertise of the  $n$  users who tweeted.

$$HTP_{t,w} = MTFIDF_{t,w} \times KW_{t,w} \quad (12)$$

$$KW_{t,w} = \frac{\sum_{u=1}^n \alpha IF_{t,u} + (1 - \alpha) PF_{t,u,c}}{n} \quad (13)$$

Finally, we predicts hot topics by comparing the change rates for hot topic prediction indices over time. Equation (14) indicates is the hot topic value  $HT_{t,w}$  for the

---

#### *Algorithm hot topic prediction*

---

```

{
  for(each candidate keyword w)
    calculate the keyword weight  $KW_{t,w}$  for keyword w at time t ;
    calculate the hot topic prediction index  $HTP_{t,w}$  for keyword w at time t ;
  for(each candidate keyword w)
    calculate hot topic value  $HT_{t,w}$  for the keyword w at time t ;
  sort hot topic value  $HT_{t,w}$ ;
  extract the top k-th keywords with the highest hot topic value  $HT_{t,w}$ ;
}

```

---

**Fig. 2** The algorithm of hot topic prediction



keyword  $w$  at time  $t$ .  $HTP_{t,w}$  denotes the hot topic prediction index for keyword  $w$  at time  $t$  and  $HTP_{t-1,w}$  denotes the hot topic prediction index for keyword  $w$  at time  $t - 1$ .

$$HT_{t,w} = \frac{HTP_{t,w} - HTP_{t-1,w}}{HTP_{t,w} + HTP_{t-1,w}} \quad (14)$$

## 4 Performance evaluation

The proposed hot topic prediction scheme's performance is demonstrated through comparison with the performance of an existing hot topic detection scheme [32]. Experimental evaluation was conducted on 1,215,342 data points collected April 1–May 31, 2015 using the Twitter Streaming API [38]. To determine user influence and expertise, information such as Twitter users' social network and the number of retweets, mentions, and embedded tweets was gathered. Keywords were extracted from Twitter using the HanNanum Korean Morphological Analyzer [39]. Table 1 shows the performance evaluation setup. To show the superiority of the proposed method, performance such as precision, recall, and F-Measure were compared. Equations (15), (16), and (17) are the precision, recall and F-Measure respectively, where  $N_{pt}$  is the number of the prediction hot topics and  $N_{rt}$  is the number of real topics in current time.

$$Precision = \frac{N_{pt} \cap N_{rt}}{N_{pt}} \times 100 \quad (15)$$

$$Recall = \frac{N_{pt} \cap N_{rt}}{N_{rt}} \times 100 \quad (16)$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100 \quad (17)$$

Hot topics at the present time are detected by excluding stopwords and frequent everyday keywords from the detected hot topics based on the keyword occurrence frequency using the TF-IDF algorithm. Tables 2 and 3 show sets of top-10 hot topic keywords for May 1–7, 2015 detected using existing and proposed schemes, respectively. Performance evaluation is conducted by comparing these sets of hot topic keywords.

**Table 1** Performance evaluation setup

Item	Value
CPU	Intel® Core™ i5-4440 CPU 3.10 GHz
RAM	6.00 GB
Language	Java (TM) SE runtime environment (build 1.8.0_31-b13)
Database	MySQL 5.6.23

**Table 2** Hot topic keywords predicted by the existing scheme

Date	Top-10 predicted hot topic keywords
15.05.01	“Nepal”, “Aid Organization”, “New Politics Alliance for Democracy”, “City Bus”, “Assault”, “Kim Na-young”, “Battery”, “Samsung Phones”, “Lotte World”, “Park Won-sun”
15.05.02	“Obama”, “City Bus”, “Kim Jong-un”, “Russia”, “Jang Yoon-jeong”, “Hi-Mart”, “Yang Mi-ra”, “Grieving Families”, “Unauthorized Rallies”, “Labor Day”
15.05.03	“Kim Jong-un”, “Samsung Electronics”, “Mayweather”, “Maritel”, “Seo Yu-ri”, “Avengers”, “Nepal”, “Gangjin”, “Survivor”, “North Korean Defectors”
15.05.04	“Seo Yu-ri”, “Park Joon-hyeong”, “Binzino”, “Golden Lacquer”, “Year-end Settlement”, “Jeong Ju-ri”, “Sohn Heung-min”, “King of Mask Singer”, “Liverpool”, “Moon Jae-in”
15.05.05	“Children’s Day”, “Jeong Ju-ri”, “Golden Lacquer”, “Tsunami”, “Racial Discrimination”, “Oh Seung-hwan”, “Lee Yeon-bok”, “Who Are You”, “Refrigerator”, “Lee Jae-yong”
15.05.06	“Who Are You”, “Northern Limit Line”, “Cho Kwon”, “Year-end Settlement”, “Sohn Hyeon-ju”, “National Pension Fund”, “Cuba”, “Sixteen”, “Produsa”, “Hong Jun-pyo”
15.05.07	“Wednesday Food Talk”, “Bong Tae-kyu”, “Hasisi Park”, “Cho Kwon”, “Yoon Geon”, “Jang Seo-hee”, “Comeback”, “Jeon Hyo-seong”, “Chu Shin-su”, “Data”

**Table 3** Hot topic keywords predicted by the proposed scheme

Date	Top 10 predicted hot topic keywords
15.05.01	“Nepal”, “Aid Organization”, “New Politics Alliance for Democracy”, “April Fool’s Day”, “Galaxy”, “Kim Na-young”, “Battery”, “Samsung Phones”, “Lotte World”, “Park Won-sun”
15.05.02	“North Korea”, “City Bus”, “Kim Jong-un”, “Russia”, “Jang Yun-jeong”, “Hi-Mart”, “Yang Mi-ra”, “Grieving Families”, “Unauthorized Rallies”, “Labor Day”
15.05.03	“Kim Jong-un”, “Samsung Electronics”, “Mayweather”, “Maritel”, “Seo Yu-ri”, “Kim Suhyeon”, “Nepal”, “Gangjin”, “Survivor”, “North Korean Defectors”
15.05.04	“Seo Yu-ri”, “Park Joon-hyeong”, “Binzino”, “Golden Lacquer”, “Year-end Settlement”, “Jeong Ju-ri”, “Sohn Heung-min”, “King of Mask Singer”, “Ahn Cheol-su”, “Moon Jae-in”
15.05.05	“Children’s Day”, “Jeong Ju-ri”, “Golden Lacquer”, “Tsunami”, “Racial Discrimination”, “King of Mask Singer”, “Lee Yeon-bok”, “Who Are You”, “Refrigerator”, “Lee Jae-yong”
15.05.06	“Who Are You”, “Northern Limit Line”, “Cho Kwon”, “Year-end Settlement”, “Park Jinyoung”, “National Pension Fund”, “Cuba”, “Sixteen”, “Produsa”, “Hong Jun-pyo”
15.05.07	“Wednesday Food Talk”, “Bong Tae-kyu”, “Hasisi Park”, “Cho Kwon”, “Yoon Geon”, “Jang Seo-hee”, “Payment Plan”, “Chess”, “Chu Shin-su”, “Data”

The comparison results for the hot topic value for specific keywords obtained using the existing and proposed schemes demonstrate the superiority of our scheme based on temporal factor and user influence over the existing scheme in terms of detection result reliability. The proposed scheme addresses the problem of the existing scheme incorrectly identifying commonly used everyday words through the change rates of the hot topic prediction indices. In Fig. 3a, b, the negative value means that the hot topic prediction index at the current time point is decreased. That is, the keywords indicating the negative value indicate that the hot topic is meaningless. Figure 3a shows the hot topic value in hot topic prediction indices for “Easter”.

In the figure, the keyword “Easter” is no longer a hot topic after Easter day (April 5th) according to indices based on the proposed scheme, which were far below those based on the existing scheme, as its hot topic prediction indices drastically decreased unlike those based on the existing scheme after the gradual increase leading up to Easter day. For keywords that are suddenly mentioned around a specific event, the proposed scheme’s detection result reliability was up to 39% higher than those of the existing scheme. Figure 3b shows a graph of the change rates in hot topic prediction indices for the keyword “Sewol Ferry.” The scheme identified “Sewol Ferry” as a hot topic when it was heavily tweeted continuously during the analyzed period, and the detection result reliability was up to 22% higher than the results for the existing scheme. Figure 3c shows a graph of the change rates for the hot topic prediction indices of the keyword “April Fool’s Day.” This keyword was detected as a hot topic when it was most frequently mentioned on April Fool’s Day, and the change rate was 26% higher for the proposed scheme’s results than for those of the existing scheme. The performance evaluation results suggest that the proposed scheme outperforms the existing scheme because the additional consideration of user influence increases the results’ reliability.

Figures 4, 5 and 6 show the precision, recall, and F-measure between predicted hot topics using the proposed scheme and the observed hot topics for the first, second, and third weeks of May. As the proposed scheme was designed to predict hot topics, prediction precision is assessed using the degree of correspondence between predictions and observations for each week. The proposed scheme generally outperformed those obtained using the existing scheme, and the predicted hot topics became more similar to those observed as the days passed from the first to the third week. Based on how the hot topic changed over the 3 weeks, the precision and recall improved by 3% on average and the F-measure improved by 4% for the proposed scheme.

Figures 7, 8 and 9 show the precision, recall, and F-measure between hot topics using the proposed scheme and the observed hot topics for April and May. The existing scheme detects hot topics based the rate of keyword frequency over time. It is detected as hot topics when the frequency of keywords increases rapidly. However, the existing scheme detect hot topics in current time but has limitations in predicting the near future hot topics. Since the documents written by users with a high level of influence and expertise are continuously propagated and shared among users, we consider the propagation of documents based on user influence and expertise. In the proposed scheme, the keywords contained in the document with a high level of influence and expertise are predicted as hot topics. Therefore, we increase the hot topic prediction accuracy in near future. The results demonstrate that the proposed scheme outperformed the existing scheme with 83.48% recall in April and 84.56% recall in May. Regarding precision and F-measure, the proposed scheme that incorporates the temporal factor and user influence showed higher detection-result reliability than the existing scheme.

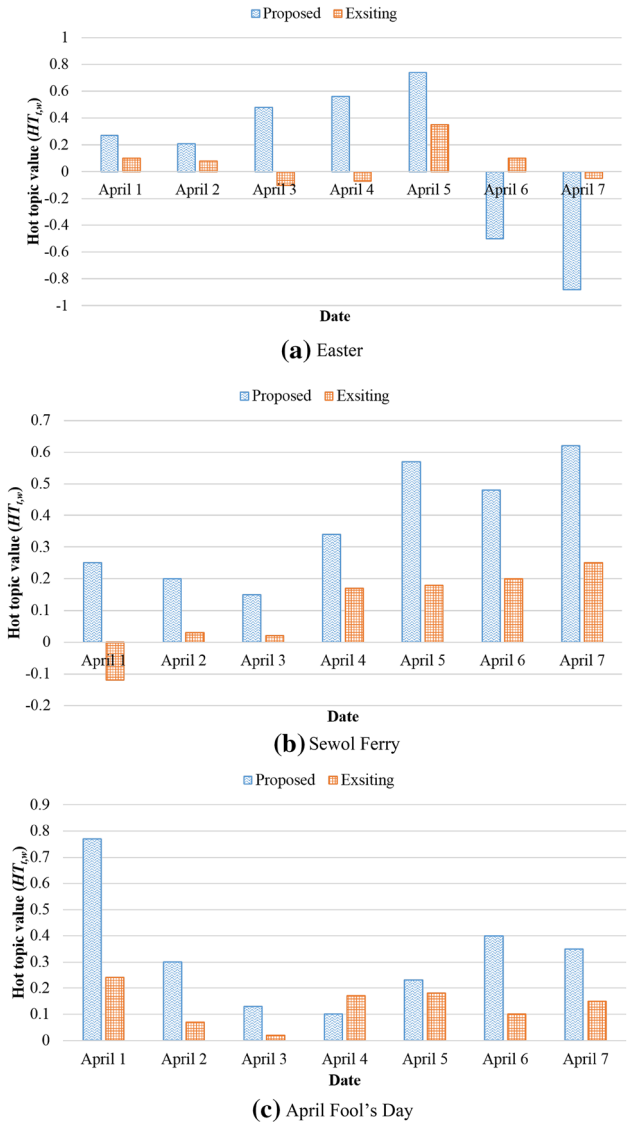


Fig. 3 Change rates in hot topic prediction indices

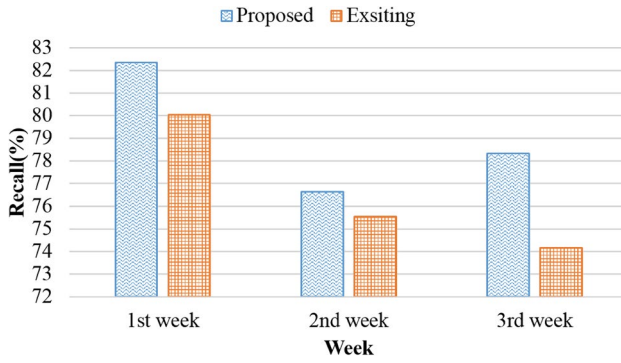


Fig. 4 Recall in May

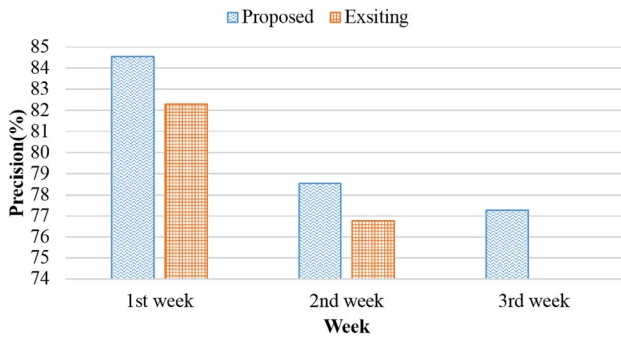


Fig. 5 Precision in May

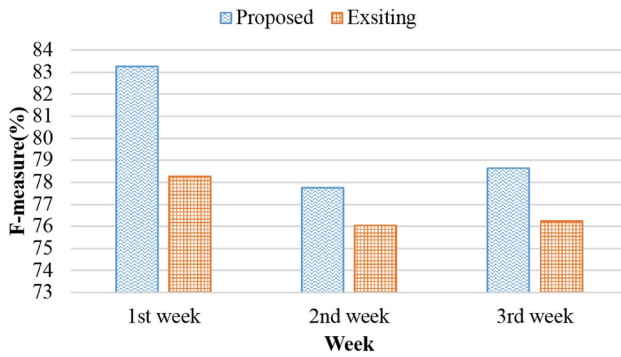


Fig. 6 F-measure in May

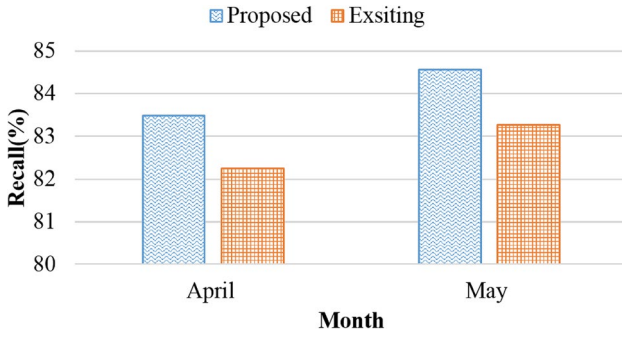


Fig. 7 Recall in April and May

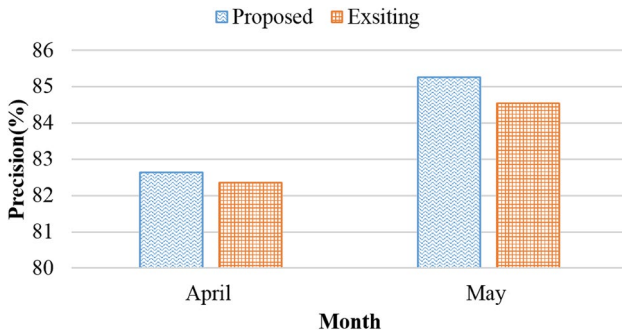


Fig. 8 Precision in April and May

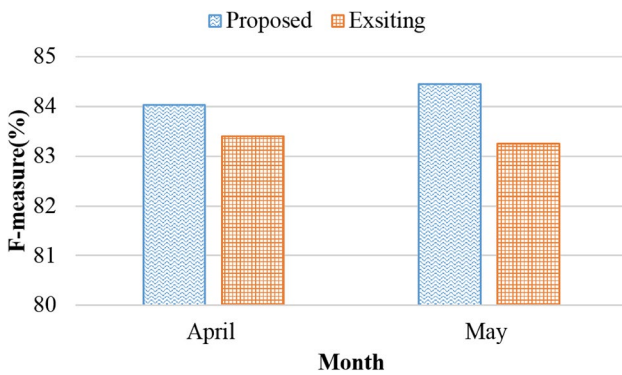


Fig. 9 F-measure in April and May

## 5 Conclusion

This paper proposed a hot topic prediction scheme based on user influence and expertise. The proposed scheme extracts a set of keywords that suddenly occur using a modified TF-IDF algorithm. The scheme incorporates user influence and expertise for those who create documents on social media to predict near-future hot topics. User influence is determined using the number of followers, retweets, and mentions; user expertise is determined using the number of embedded tweets, reliability, and preferred topic. Hot topic predictions are made by applying the weight-the average of expertise and influence indices for users who tweeted with a candidate keyword extracted using modified TF-ID to the keyword. Future research plans include research on grouping interrelated keywords around specific events.

**Acknowledgements** This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2016R1A2B3007527), by “Human Resources Program in Energy Technology” of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea. (No. 20164030201330), and by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (No. NRF-2017M3C4A7069432).

## References

1. Pee, L. G. (2018). Affordances for sharing domain-specific and complex knowledge on enterprise social media. *International Journal of Information Management*, *43*, 25–37.
2. Derczynski, L., Yang, B., & Jensen, C. S. (2013). Towards context-aware search and analysis on social media data. In *International conference on extending database technology*, March 18–22, Genoa, Italy.
3. Anandhan, A., Shuib, N. L. M., Ismail, M. A., & Shaikh, G. M. (2018). Social media recommender systems: Review and open research issues. *IEEE Access*, *6*, 15608–15628.
4. Sperli, G., Amato, F., Mercurio, F., Mezzanzanica, M., Moscato, V., & Picariello, A. (2018). A social media recommender system. *International Journal of Multimedia Data Engineering and Management*, *9*(1), 36–50.
5. Duchateau, F. (2011). Who can best answer a query in my social network? In *International conference on data engineering workshops*, April 11–16, Hannover, Germany.
6. Ehrlich, K., & Shami, N. S. (2008). Searching for expertise. In *Conference on human factors in computing systems*, April 5–10, Florence, Italy.
7. Sibona, C., Cummings, J., & Scott, J. (2017). Predicting social networking sites continuance intention through alternative services. *Industrial Management and Data Systems*, *117*(6), 1127–1144.
8. Li, H., Bok, K. S., & Yoo, J. S. (2015). A mobile social network for efficient contents sharing and searches. *Computers & Electrical Engineering*, *41*, 288–300.
9. Said, A., Luca, E. W. D., & Albayrak, S. (2010). How social relationships affect user similarities. In *Workshop on social recommender systems*, February 7–10, Hong Kong, China.
10. Mislove, A., Viswanath, B., Gummadi, P. K., & Druschel, P. (2010). You are who you know: Inferring user profiles in online social networks. In *ACM international conference on Web search and data mining*, February 4–6, New York, USA.
11. Phethean, C., Tiropanis, T., & Harris, L. (2015). Engaging with charities on social media: Comparing Interaction on Facebook and Twitter. In *International conference on internet science*, May 27–29, Brussels, Belgium.
12. Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *International conference on World Wide Web*, April 26–30, Raleigh, North Carolina, USA.

13. Dong, T., Cheng, N., & Wu, Y. J. (2014). A study of the social networking website service in digital content industries: The Facebook case in Taiwan. *Computers in Human Behavior*, *30*, 708–714.
14. Guo, C., Li, B., & Tian, X. (2016). Flickr group recommendation using rich social media information. *Neurocomputing*, *204*, 8–16.
15. Oh, S., & Syn, S. Y. (2015). Motivations for sharing information and social support in social media: A comparative analysis of Facebook, Twitter, Delicious, YouTube, and Flickr. *Journal of the Association for Information Science and Technology*, *66*(10), 2045–2060.
16. Faruque, S. A., Khatun, M. A., & Rahman, M. S. (2016). Modelling direct marketing campaign on social networks. *International Journal of Business Information Systems*, *22*(4), 422–435.
17. Khater, S., Gracanin, D., & Elmongui, H. G. (2017). Personalized recommendation for online social networks information: Personal preferences and location-based community trends. *IEEE Transactions on Computational Social Systems*, *4*(3), 104–120.
18. Keib, K., Himelboim, I., & Han, J. (2018). Important tweets matter: Predicting retweets in the #BlackLivesMatter talk on twitter. *Computers in Human Behavior*, *85*, 106–115.
19. Ramaswami, C., Murugathan, M., Narayanasamy, P., & Khoo, C. S. G. (2014). A survey of information sharing on Facebook. *Information Research*, *19*(4).
20. Jia, R., Liu, T., Zhang, J., Fu, L., Gan, X., & Wang, X. (2018). Impact of content popularity on information coverage in online social networks. *IEEE Transactions on Vehicular Technology*, *67*(8), 7465–7474.
21. Zheng, Y., Xie, X., & Ma, W. (2010). GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Engineering Bulletin*, *33*(2), 32–39.
22. Jiang, F. (2014). A uniform framework for community detection via influence maximization in social networks. In *IEEE/ACM international conference on advances in social networks analysis and mining*, August 17–20, Beijing, China.
23. Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, *54*(3), 1442–1451.
24. Das, R., Kamruzzaman, J., & Karmakar, G. C. (2018). Modelling majority and expert influences on opinion formation in online social networks. *World Wide Web*, *21*(3), 663–685.
25. Yang, Z., Wang, C., Zhang, F., Zhang, Y., & Zhang, H. (2015). Emerging rumor identification for social media with hot topic detection. In *Web information system and application conference*, September 11–13, Jinan, China.
26. Yu, Q., Weng, W., Zhang, K., Lei, K., & Xu, K. (2014). Hot topic analysis and content mining in social media. In *International performance computing and communications conference*, December 5–7, Austin, TX, USA.
27. Zhang, C., Liu, L., Lei, D., Yuan, Q., Zhuang, H., Hanratty, T., & Han, J. (2017). TrioVecEvent: Embedding-based online local event detection in geo-tagged tweet streams. In *ACM SIGKDD international conference on knowledge discovery and data mining*, August 13–17, Halifax, NS, Canada.
28. Katragadda, S., Benton, R. G., & Raghavan, V. V. (2017). Framework for real-time event detection using multiple social media sources. In *International conference on system sciences*, January 4–7, Hilton Waikoloa Village, Hawaii, USA.
29. Lin, S., Kong, X., & Yu, P. S. (2013). Predicting trends in social networks via dynamic activeness model. In *ACM international conference on information and knowledge management*, October 27–November 1, San Francisco, CA, USA.
30. Han, Y., Fang, B., & Jia, Y. (2014). Predicting the topic influence trends in social media with multiple models. *Neurocomputing*, *144*, 463–470.
31. Zhu, T., & Yu, J. (2014). A prerecognition model for hot topic discovery based on microblogging data. *The Scientific World Journal*, *2014*, 1–11.
32. Kim, H., Lee, S., & Kyeong, S. (2013). Discovering hot topics using Twitter streaming data. In *International conference on advances in social networks analysis and mining*, August 25–29, Niagara, ON, Canada.
33. Ruan, Y., Purohit, H., Fuhry, D., Parthasarathy, S., & Sheth, A. (2012). Prediction of topic volume on Twitter. In *International ACM conference on web science*, June 22–24, Evanston, IL, USA.
34. Jeelani, H., & Singh, K. (2014). ‘Good’ versus ‘bad’ opinion on micro blogging networks: Polarity classification of Twitter. *International Journal of Computer Science and Mobile Computing*, *3*(8), 49–56.



35. Yu, R., Zhao, M., Chang, P., & He, M. (2014). Online hot topic detection from web news archive in short terms. In *International conference on fuzzy systems and knowledge discovery*, August 19–21, Xiamen, China.
36. Aizawa, A. N. (2003). An information-theoretic perspective of TF-IDF measures. *Information Processing and Management*, 39(1), 45–65.
37. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (pp. 109–133). Cambridge: Cambridge University Press.
38. Twitter Streaming API. <https://dev.twitter.com/streaming/overview>. Accessed 12 August, 2014.
39. HanNanum Korean Morphological Analyzer. <http://semanticweb.kaist.ac.kr/home/index.php/HanNanum>. Accessed 24 November, 2014.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.