JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 34, 1579-1598 (2018)
DOI: 10.6688/JISE.201811_34(6).0012

SocialDNA: A Novel Approach for Distinguishing Notable Articles and Authors through Social Events*

MING-HUNG WANG[1] AND CHIN-LAUNG LEI[2]
[1]*Department of Information Engineering and Computer Science*
Feng Chia University
Taichung, 407 Taiwan
[2]*Department of Electrical Engineering*
National Taiwan University
Taipei, 106 Taiwan
E-mail: mhwang@fcu.edu.tw; cllei@ntu.edu.tw

With the rapid development of online social networks, increasing amounts of user-generated content are posted online. While information is overwhelming a site, readers actually prefer fully reading some influential articles or information rather than glancing sequentially at every article on a forum. In this paper, we propose techniques for forecasting public responses to articles shortly after the articles are published. Our proposal identifies the important articles from two perspectives: frequent discussion and extreme acceptance or rejection by the online public. We also discuss approaches for distinguishing influential authors who are popular and receive consistently high ratings from online users. To verify our methodologies, we analyzed a popular social forum in Taiwan during three large-scale social events, a social movement and two national election campaigns. Our results demonstrate that our methodologies achieve high accuracy with significant time reduction and outperform previous methodologies in distinguishing notable articles and authors on social forums.

Keywords: social networks, social media, information filtering, opinion leader, public opinion

1. INTRODUCTION

In recent years, online social networks and forums have become important platforms for people to express their opinions and to discuss them with others. With the increasing number of articles posted online, determining methodologies for identifying notable information and authors is a significant issue, especially with regard to trending and controversial topics. For example, on March 18, 2014, protesters entered and occupied the Parliament of Taiwan to protest against the review process of the Cross-Strait Service Trade Agreement (CSSTA) [1]. This drastic action attracted substantial attention in Taiwan for a short time from the government as well as from citizens. Benefiting from the well-developed online social networks, heated debates occurred online as well.

When social events occur and become popular, articles temporarily tend to overwhelm social media sites and information often appears chaotic. It is not easy for readers to find notable information efficiently under these conditions. Identifying influential articles for readers assists them in understanding the trending topics among the sites com-

Received March 8, 2017; revised October 13, 2017 & February 5, 2018; accepted March 12, 2018.
Communicated by Wen-Chih Peng.
* A preliminary version of this work was presented at the 2015 IEEE International Conference on Systems, Man, and Cybernetics, Hong Kong, China, October 9-12.

prehensively. Furthermore, while news and rumors rapidly spread over the Internet, the efficient and effective dissemination of official news and clarifications in urgent situations can be essential for the government. Previous scholars have suggested that spreading news via highly credible, trustworthy opinion leaders among the platforms could promote civil mobilization and engagement [2, 3]. Hence, to achieve a fast and comprehensive understanding of articles and authors on a social forum, this paper addresses the following two major issues:

- **Extracting trending information:** Extracting trending topics is useful for the participants and the audience to understand the current development of an event comprehensively. In addition, the sooner government officials are able to identify possible trending threads, the more time they will have to make decisions or to devise strategies to handle an event during urgent situations.
- **Identifying influential authors:** Influential authors with good reputations usually play key roles in communications on the web. They deliver official or notable information to the public; moreover, they can even change peoples' minds. Therefore, identifying influential authors is advantageous for the government and for other organizations in order to spread the latest information efficiently and effectively.

To make this possible we propose SocialDNA, which constitutes a series of novel approaches for important information extraction and influential author identification. We first quantified the importance of articles from two perspectives: reply quantity (popularity) and acceptance/rejection consistency (sentiment). To achieve this objective, we observed the behavior of peer-to-peer replies and rating records of online users toward articles. To detect potential popular threads quickly, we developed regression models to predict the number of replies and rating score of an article using its early reply records, rating records, and other related metrics. For influential author identification, we proposed a hybrid approach of identifying possible influential authors on a forum with two characteristics: collection of a large number of replies (popular) and receipt of extreme rating scores (consistent acceptance/disagreement) of their articles.

To verify our proposal, we investigated user behavior on the largest bulletin board system in Taiwan during three major social events. Then we conducted in-depth user behavior analyses on these datasets and applied our proposals to the two different types of social events mentioned above. Two major contributions of our study can be summarized as follows:

1. We developed regression models to predict highly discussed and publicly accepted/rejected articles in a short time period. Based on the datasets we collected, our regression models demonstrated that they could identify influential articles effectively and efficiently.
2. We proposed a methodology to identify influential authors on a social forum. Our results indicate that our proposal outperforms the two other popular methodologies for identifying authors who receive many responses from the public and are publicly accepted/rejected.

The remainder of this paper is organized as follows. Section 2 discusses related

studies. Section 3 presents our methodologies for identifying important articles and authors on social forums. Section 4 summarizes the collected datasets and evaluates our proposal using the data collections. Section 5 summarizes the findings and concludes our research.

2. RELATED WORK

2.1 Predicting Web Content Popularity

Predicting the popularity of web content has been a challenging problem in the Web 2.0 era. Previous scholars have investigated multimedia content [4-7], news articles [8-12], and social network services [13-15]. Several preceding methodologies have employed temporal user behavior or early content popularity records [5, 8, 14, 16, 17] to predict future popularity. These prediction techniques were built over long periods of observation. For example, Szabo and Huberman [5] studied the hourly activity on Digg, a popular news aggregation website, to forecast the popularity of news articles according to a univariate regression model. Pinto *et al.* [18] improved the Szabo-Huberman method by developing multi-variate regression model using early patterns. Similarly, Cha *et al.* [4] used the first few days' access records to predict the near-future popularity of videos. These works demonstrated the possibility of using temporal and early-access behavior to predict popular content on the Internet. However, these methodologies might not be applicable to predicting the popularity of content in urgent situations for the following two reasons: (1) articles can appear suddenly and trending topics can live and die on the forum in hours [19], and; (2) observing daily or hourly records takes too long to provide near real-time predictions. Therefore, in this paper we focus on shortening the time for predicting the popularity of articles.

The followings are related scholars using regression-based models; [5, 8, 9] used historical viewing records on a particular time to predict the content popularity. To enhance the univariate models, [17, 18] leveraged temporal dynamics of viewing history to develop multivariate regression models. However, most of the abovementioned works required a relatively long observation time, for example, more than 30 minutes. While in an urgent situation, people usually tend to shorten the monitoring length and take instant actions. Thus, in this paper, we focused on monitoring less than 30 minutes and achieves a certain accuracy when predicting.

2.2 Predicting Web Content Ratings

In addition to highly discussed articles, another type of influential articles could be generally accepted/rejected ones. Due to review mechanisms on Web 2.0 sites, review ratings have been a popular reference for evaluating web objects. A product receiving a large number of consistent ratings could indicate reliable public opinion toward the product. Many previous scholars investigated the prediction of rating polarities (positive, neutral, and negative) [20-22] and the product rating scales (*e.g.*, one star to five stars) [23, 24]. These studies were built primarily on text mining and sentiment analysis of online reviews. In contrast to previous methodologies utilizing various features of reviews to develop a prediction model, this paper addresses predictions by learning from

very early review activities. To the best of our knowledge, our study is the first to attempt to predict review ratings by the early rating records collected during the very first minutes.

2.3 Social Media in Urgent Events

Social network services and social forums have become popular channels for information dissemination in recent years, especially in urgent situations. Online users leverage the power of online social networks to deliver exclusive photos, personal experiences, opinions, and official news to others. Several research groups have studied how to use social network services to detect emergent situations such as earthquakes [25, 26] and floods [27, 28]. As well as reporting natural disasters, numerous reports have addressed the capability of social media to connect potential protesters to strengthen participation in social movements [29-32].

Regarding social movements, one previous report [33] proposed that social media, such as Facebook and Twitter, could be leveraged by activists to challenge traditional mass media. However, since the number of messages might be overwhelming while new events arise [34], Kang pointed out that articles become meaningful only when others pay attention and respond to them [35]. In addition, some studies have proposed that social movement organizations might not be able to control their messages through social media because of the multitude of voices among social media, and that a movement's theme could be changed as a result [31, 36, 37]. Based on the abovementioned literature, we consider an early-detection methodology for trending articles to be beneficial for social organizations to realize quickly which articles are popular at the moment. In addition, such identification could help organizations respond quickly to the multitude of voices that could potentially fragment a movement as well as an election campaign group.

2.4 Concept of Opinion Leaders

Katz first described a two-step model [38, 39] for explaining the opinion leader concept in communications. According to this model, information from mass media is first received by those people who act as opinion leaders. Later, these leaders disseminate the information to their readers. A number of scholars have focused on this concept in real life, a widely studied example being market mavens [40-42]. According to Katz, opinion leadership can be categorized into three dimensions on the basis of the behavior and capabilities of an individual:

1. Who one is: a person's specific values or characteristics, such as an optimistic personality;
2. What one knows: a person's expertise, capabilities, and knowledge of a particular topic, and;
3. Whom one knows: the number of people an individual knows or with whom he or she maintains contact.

However, identifying online leadership is not a simple task according to the three dimensions in [38] due to the anonymity of online forums. It is not easy for observers to

distinguish accurately who is a leader from the limited information about an online user. To determine correctly who the leaders are, or more precisely, who can influence more users or even change public opinion, many methodologies leveraging structural analysis in social networks have been proposed for discovering online leaders [43-49]. However, online participants' feedback behaviors toward users have not been studied thoroughly. For example, influential authors publish articles, prompting many people to comment (structural) and receiving ratings (sentimental). This kind of leaders has a large number of connections showing his or her capabilities. However, these influential authors might not be detected using merely a single metric, such as structural position or sentimental scores, as suggested in previous studies. Therefore, this study not only considers the structural position of an author in a network by evaluating the number of received comments, but also public agreement or rejection by studying the public ratings the author has received. In this study, we considered not only the structural position of an author in a network by evaluating the number of received comments, but also public agreement or rejection by studying the public ratings the author received, and we called them "hybrid influential authors."

3. METHODOLOGY

This section proposes different approaches to identify influential articles on a social forum. In contrast to previous methodologies that primarily rely on page view counts, we identify articles from two perspectives: discussion frequency and public ratings. We leverage the peer-to-peer rating mechanism, which enables us to observe opinions of every article. From the observation of reply and rating behaviors, we developed a prediction model to identify influential articles early, according to partial reply quantities and article rating scores. For influential author identification, based on the prediction metrics, we present an approach to identifying notable authors' popularity and general acceptance/ rejection from the public. Our approach can distinguish influential authors who have written highly discussed articles and receive consistent ratings.

3.1 Article Reply Quantity Prediction Approach

In this part, we mainly employ replying history in predicting reply counts. In addition to the temporal changes of the number of replies, we added article attributes into the model to enhance the model, including the word count, keyword occurrences, and article publishing time and date. To achieve this, we initially investigated the reply quantity of each article to measure its popularity. Previous methodologies have typically measured the popularity of each article by the number of views. However, there are two defects with this. First, it is easy for a programmer to write scripts to produce page views for an article automatically. Second, since the page view count indicates only the content loaded on the user's browser, it is unknown whether the user reads the content thoroughly or just glances at the article. To avoid possible bias introduced by measuring an article's influence according to the page view count, we considered the reply quantity to be a more solid indicator of effective dissemination than the number of page views, since a reply is an action taken by a user.

Our goal was to identify those articles that received more replies from online users.

To identify possible popular threads from a large number of articles in a short time period, we built a model to predict the final number of replies from the temporal reply behavior to an article and other related metrics. By using our prediction model, it is possible to forecast the number of replies an article will receive and to identify the most discussed articles.

3.1.1 Temporal changes in reply quantities

We began from the temporal changes in the reply quantities. The following metrics were employed.

Definition 1: The total number of replies to article p is denoted as T_p. An observer monitors the reply behavior of p for s time slots, from slot 1 to slot s. The length of each time slot is t seconds. The number of replies in each time slot i of p is denoted as $D_{i,p}$.

By using this approach, we sought to use the temporal reply behavior during each time slot $D_{i,p}$ to estimate T_p. Since we wanted to shorten the time required for an accurate prediction, s and t had to be estimated carefully. More time observed might enable the acquisition of more information to predict the results, while less time observed would provide real-time identification for possible popular threads. Because our objective was to forecast the number of replies in a short period after an article was posted, we aimed to find the shortest observation time, $t * s$, for a high accuracy.

3.1.2 Development prediction model

In addition to the temporal reply behavior, we considered the following article-related metrics that could influence the reply behaviors of users and that could be incorporated into our model.

- **Text length:** Since the length of text in a social forum is not as limited as it is in Twitter, users can take more time to read and understand longer articles. We considered the length of a text to be a factor and included it into our model. Since the articles on PTT were mainly written in Chinese, we counted the number of Chinese characters in the text body of an article as the text length.
- **Work or leisure hours:** An article can be published at midnight or in the afternoon, and it can receive different amounts of attention at different times. To offer a better prediction, we investigated whether the publishing time affected the number of replies to each article. If the article was published between 8:00 AM and 8:00 PM, this metric was set to one; otherwise, it was set to zero.
- **Workday or weekend:** We also investigated whether an article being published on a workday or on a weekend would influence the prediction. If an article was published between Monday and Friday, it was categorized as a workday article; otherwise, it was a weekend post. If an article was published Monday through Friday, this metric was set to one; otherwise, it was set to zero.

In our prediction model, the dependent variable T_p is the number of replies to article p, and our regression factors include the number of replies in time slot 1 to n as $D_{1,p}$, $D_{2,p}$, …, $D_{n,p}$. We also incorporated the abovementioned article-related metrics into our

regression model. For every article p, our model can be expressed as

$$T_p = \alpha_0 + \sum_{i=1}^{n} \alpha_i D_{i,p} + \beta_1 Text_p + \beta_2 is.workhour_p + \beta_3 is.workday_p. \quad (1)$$

From the prediction model, we proposed to detect early possible influential articles with high reply quantities, while observing the reply behavior for a short time.

3.2 Article Rating Score Prediction Approach

In this section, we discuss the prediction of article rating scores to extract those articles that receive high volumes of consistent ratings (extreme rating scores) from online users. When an article receives strong positive/negative polarity, it can be influential, because many users express opinions about it, and most of these users hold the same opinion. To predict the rating scores, we initially quantified each article's rating score according to the peer-to-peer rating mechanism on social forums. We developed a regression model to predict the final rating scores of articles on a social forum based on records observed early. By our prediction model, we proposed to detect early influential articles that could receive extreme positive/negative rating scores.

3.2.1 Quantifying public opinion toward articles

To quantify public opinions toward articles, we began from a rating mechanism that included thumbs-up, thumbs-down, and neutral options for users to express their attitudes.

Definition 2: For any article p on the forum, the total number of ratings of p is denoted as R_p. The number of thumbs-up, thumbs-down, and neutral ratings of p are denoted as TU_p, TD_p, and NT_p, respectively.

From the abovementioned definition, the relationship between TU, TD, and NT can be expressed as

$$R_p = TU_p + TD_p + NT_p, \quad (2)$$

and the rating score of p can be computed as

$$Article.score_p = TU_p - TD_p. \quad (3)$$

An extremely positively or negatively rated article can imply consistent opinions from online users toward the article since there is a significant difference between TU and TD on the thread; Also, this kind of article indicates that the audience pays more attention to it, because more users express their opinions toward such articles.

3.2.2 Determining a suitable rating data size for prediction

To predict extreme rating score articles accurately in a short time period, it is necessary to determine the number of ratings n that must be observed. We defined the partial rating records of an article as follows.

Definition 3: For any article p on a forum, the numbers of thumbs-ups, thumbs-downs, and neutrals observed for p in the first n ratings are denoted as $TU_{p,n}$, $TD_{p,n}$, and $NT_{p,n}$, respectively.

Since the objectives are to shorten the observation time and to raise the accuracy, finding a minimum value of n is important. A larger n yields more rating records for predicting rating scores but requires more time, while choosing a smaller n yields less evidence to build the prediction model. However, during emergencies or developing events, people might care more about decision speed, rather than prediction accuracy. Hence, the smallest n that could help in deriving a particular prediction accuracy in a short time period is preferable.

In the proposed methodology, we consider to choose the smallest n must yield high correlation coefficients (Pearson $r >= \alpha$)) between the percentages of thumbs-ups/thumbs-downs in the first n ratings and in the complete ratings. We set the threshold α value as 0.8, a general standard for indicating a high correlation between two continuous variables [50, 51]. This threshold ($\alpha = 0.8$) indicates that the relationship is at least 64% higher than no relationship when r is over the threshold.

3.2.3 Developing a prediction model

We incorporated the partial rating records and some article-related features into the prediction model. The following factors were used in our model:

- n denotes the first n rating records used to predict the final article rating score,
- $TU_{p,n}$ and $TD_{p,n}$ denote the numbers of thumbs-ups and thumbs-downs, respectively, in the first n ratings of article p,
- $Time_{p,n}$ denotes the time required to collect n ratings in article p,
- $Text_p$ represents the number of Chinese characters in the text body of article p, and
- $Match_p$ refers to the frequency of keyword occurrences in the text body of article p[1].

Our observations demonstrate that the time difference between publishing an article and the article being rated approaches a log-normal distribution. To achieve a better fit on our linear regression model, we performed a log-transform $\log_{10}(x)$ to $Time_{p,n}$, denoted as $Time^*_{p,n}$ and included it in our model. We also observed that the positive and negative rating scores both followed exponential distributions. Hence, we conducted a polarity-preserving log-transform to $Article.score.t_p$, and the transformed score is calculated as

$$Article.score.t_p = \text{sgn}(Article.score_p) * \log_{10}(|Article.score_p + 1|), \tag{4}$$

$$\text{sgn}(x) = \begin{cases} 1, & \text{if } a \geq 1, \\ -1, & \text{otherwise.} \end{cases} \tag{5}$$

We next included all first order predictor sand second order interactions between $TU_{p,n}$, $TU_{p,n}$, and $Time^*_{p,n}$ in our regression model, since interdependences between these predictors could exist, $TU_{p,n}$ resulting in a low $TU_{p,n}$. Overall, for article p, our regression model can be expressed as

[1] Note that as most of our collected articles are in Chinese, to avoid potential mis-segmentation while doing Chinese segmentation, while counting keyword occurrences we did not apply segmentation procedure and removing stop words in our dataset; instead, for each keyword, we calculated the total occurrence of every defined keywords in the content string. We then derived the total keyword occurrence of each article.

$$Article.score.t_p = \alpha_0 TU_{p,n} + \alpha_1 TD_{p,n} + \alpha_2 Time^*_{p,n} + \alpha_3 TU_{p,n}: TD_{p,n} + \alpha_4 TU_{p,n}: Time^*_{p,n} + \alpha_5 TD_{p,n}: Time^*_{p,n} + \alpha_6 TU_{p,n}: TD_{p,n}: Time^*_{p,n} + \beta_0 Text_p + \beta_1 Match_p.$$

3.3 Hybrid Influential Author Identification Approach

This section proposes a hybrid approach to identify influential authors leveraging peer-to-peer ratings on the forum. To find possible influential authors with two opinion leader characteristics, popular and generally accepted/rejected, we defined the influential scores of authors as follows.

The rating score of article p posted by author v is denoted as $Article.score_p$. To represent the consensus among the ratings of article p, we divided the $Article.score_p$ by the number, R_p, which refers to the number of ratings of article p. Hence, the consensus score of article p can be denoted as $Consensus_p = \frac{Article.score_p}{R_p}$, where $-1 \le Consensus_p \le 1$. We next defined the consensus score of author v by calculating the average consensus scores of all articles published by author v as follows:

$$Author.score_v = \frac{\sum_{p \in v} Consensus_p}{Post_v}. \tag{7}$$

The author scores consider the permanency of the ratings that each author has received. This score can be used to measure the average consensus from the public regarding the author. However, our objective was to identify not only publicly agreement/disagreement, but also popular authors. To add the popularity characteristic to the score, we slightly alternated *Author.score* calculation and incorporated the logarithm of the number of comments, $\log_{10}(R_p)$, received by the author into the influential score calculation. The influential score of author v can be computed using

$$Influential.score_v = \frac{\sum_{p \in v} (Consensus_p * \log_{10}(R_p))}{Post_v}. \tag{8}$$

This definition of the *Influential.score* reflects the measurement of two characteristics (popular and consistently rated) of each author. However, bias can result in misidentification of influential authors, when the number of ratings is extremely small, *e.g.*, only one rating for an article. As a result of receiving very few ratings, the article consensus score of such an author will be near 1 or −1. To avoid such biases, in this study we counted only articles that had many ratings ($R_p \ge 20$), and we assigned a weight (denoted as $\log_{10}(R_p)$) to each article to reflect the difference in rating quantities; receiving more ratings resulted in a higher score. As our goal is to identify "hybrid influential authors" who can either receive many ratings and extreme scores, to demonstrate the capability of our method, we will select two popular baselines to depict that our method to identify popular and generally accepted/rejected authors on the forum.

4. EVALUATION

We evaluated our proposal using through three social events: a large-scale social movement and two major election campaigns in Taiwan. We briefly describe the events as follows; (1) The Sunflower movement was a protest that occurred in Taiwan from

March 18, 2014 to April 10, 2014 to express opposition to the content of the Cross-Strait Service Trade Agreement (CSSTA) between Taiwan and China and the review process of the agreement; (2) The 2014 Taiwan general election was focused on the mayoral elections in six special municipalities, accounting for two-thirds of the population of Taiwan (16 million); (3) The 2016 Taiwan presidential election was the most important election in Taiwan and elected the current president.

4.1 Dataset Descriptions

We focus on the PTT, the largest social forum in Taiwan. We studied the most popular board called "Gossiping," which focuses on political news discussions. We crawled all the articles, including article content, reply records, rating records, and article metadata during the Sunflower movement and the last two months of the two election campaigns in Taiwan. We preprocessed our crawled datasets according to the title and content of each article. Articles that did not contain keywords related to the movement or the names of the candidates were removed. To concentrate on highly discussed and highly rated articles, we included in our datasets only those articles that had been rated by and/or to which at least 20 users had replied. The summaries of the three datasets are presented in Table 1.

Table 1. Summaries of our 3 datasets.

	2014 Movement	2014 Election	2016 Election
Start	2014/03/18 06:00:00	2014/09/30 06:00:00	2015/11/16 06:00:00
End	2014/04/11 06:00:00	2014/11/30 06:00:00	2016/01/16 06:00:00
# Articles	8,431	7,211	3,022
# Replies	609,232	623,340	237,135
Words/post	574.3	612.5	638.8
Ratings/post	72.3	86.4	78.5
# Users	49,681	36,452	23,614
# Authors	4,518	2,483	1,235
# Raters	48,832	35,983	23,239

4.2 Predicting Reply Quantities

We set different time slots (from 1 to 6) for observation to develop a regression model; multiple time slots enabled us to incorporate temporal changes into the regression model. We also varied the length of observations from 3 to 25 minutes (180 to 1500 seconds). We evaluated the prediction accuracy of our model by changing the total observation time. In addition, for the same observation time, we applied different time slots to investigate whether more observation frames would increase the prediction accuracy. We used two metrics to measure the accuracy: the adjusted R^2 and the Spearman ρ. Based on the adjusted R^2 values, we evaluated the accuracy of our regression model. However, if a user's objective is to filter important articles on social forums, he or she might care more about the relative ranking of these articles rather than the actual number of replies to each article. Therefore, we used the rank correlation indicator, the Spearman ρ, to investigate the accuracy of our model in predicting the rankings of articles.

Our evaluation results, described in Fig. 1, reveal the following significant findings; (1) The prediction accuracies determined by using both the adjusted R^2 and ρ increase with the observation time. These results were expected because more observation time

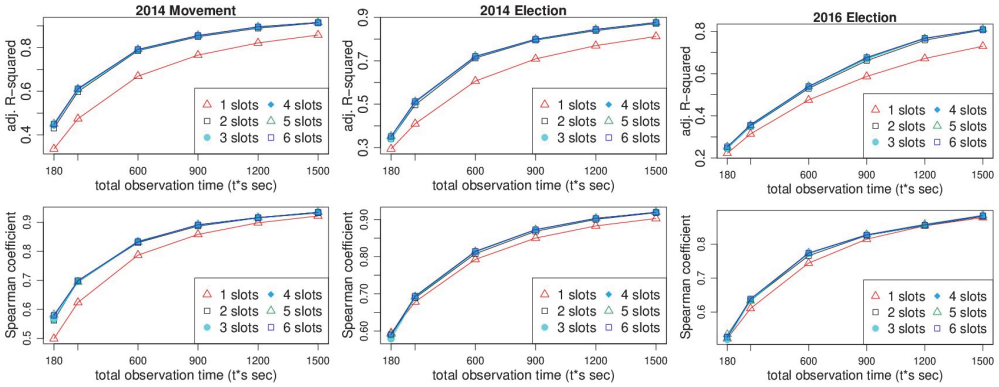

Fig. 1. Temporal behavior prediction results for both datasets.

yields more records for model building; (2) There is no significant difference in accuracy between the results obtained using different time slots (from 2-6 slots) and slot sizes with the same length of observation time. Only in the 1-slot setting, the model is not as good as in other slot settings. The results are reasonable because the 1-slot setting removes the temporal change feature in the model; (3) Our prediction model performs well (Spearman $\rho \geq 0.8$) in these events when observing more than 900 seconds. Meanwhile, an article's reply-rate rank can be predicted accurately within 15 minutes after authors published the article. The results present a significant reduction in the time spent if we use the model for prediction rather than waiting until the reply behavior becomes smooth.

We further compare the predicting performance of our model and other two popular methods, S-H model [5], and Pinto model [18]. We set the time slot from 4 to 6 and adjusted the observation time from 3 to 25 minutes. We here adopted mRSE (mean relative squared error) and MSE (mean squared error) as the evaluation metrics. From the results shown in Fig. 2, we find our model outperforms the other two previous studies. From the experiment, we consider our model could help predict popular articles while monitoring a short time on the forum during emergencies.

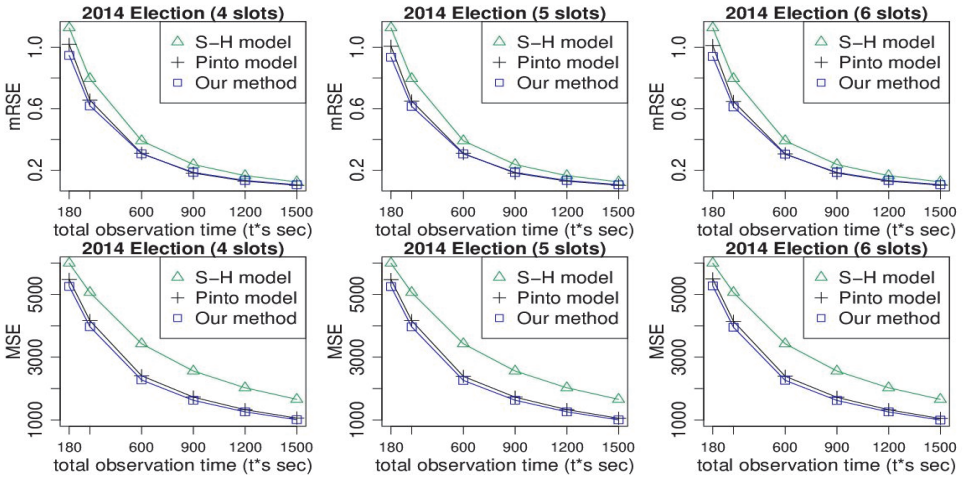

Fig. 2. Prediction accuracy for our proposal and the other two popular methods.

We also present the regression table of our proposed model (setting: 6 slots, 1500-second observation time) in the three datasets in Table 2. From the table, we can find most of the variables we included in the model are significant and these features do contribute to the model in different datasets. The adjusted R^2 is all over 0.8 in the regression. In addition, to identify potential overfitting, we employed 10-fold cross-validation to estimate the performance of our model. The RMSE of 10-fold validation has slight difference (< 3%) with the RMSE of the entire dataset and 10-fold validation, and we consider our regression models are not overfitted.

Table 2. The regression table of our proposed model in three datasets.

	2014 Movement	2014 Election	2016 Election
first_slot	0.924***	0.897***	0.794***
	$p = 0.000$	$p = 0.000$	$p = 0.000$
second_slot	1.070***	0.903***	1.100***
	$p = 0.000$	$p = 0.000$	$p = 0.000$
third_slot	1.040***	1.560***	1.450***
	$p = 0.000$	$p = 0.000$	$p = 0.000$
fourth_slot	1.880***	2.160***	3.080***
	$p = 0.000$	$p = 0.000$	$p = 0.000$
fifth_slot	2.950***	3.060***	4.620***
	$p = 0.000$	$p = 0.000$	$p = 0.000$
sixth_slot	4.070***	5.560***	5.330***
	$p = 0.000$	$p = 0.000$	$p = 0.000$
word_count	0.002***	0.010***	0.006***
	$p = 0.000$	$p = 0.000$	$p = 0.000$
post_hour	–2.290***	–6.100***	–8.770***
	$p = 0.00001$	$p = 0.000$	$p = 0.000$
Constant	2.940***	2.480***	4.370***
	$p = 0.000$	$p = 0.004$	$p = 0.008$
Observations	8,431	7,211	3,022
Adjusted R^2	0.916	0.877	0.809

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

4.3 Predicting Article Rating Scores

To establish a reliable prediction model that meets users' expectations regarding both accuracy and speed, the number of records to observe must be carefully determined. To investigate potential features, we perform exploratory data analysis (EDA) over our dataset. We observed the percentages of thumb-ups/thumb-downs (*TU*/*TD*) in the first *n* ratings correlate to the complete rating records. Thus, we first computed the portions of *TU*/*TD* of all articles in the dataset. Next, for each *n* ranging from 1 to 20, we calculated the portions of *TU*/*TD* of all articles in the first *n* ratings. Finally, for each dataset, we retrieved the correlation coefficient of the portions of *TU*/*TD* of the first *n* ratings and the complete ratings, while *n* ranges from 1 to 20. Therefore, we had 20 points for *TU*, 20 points for *TD* and derived the coefficient changes as shown in Fig. 3. The figure shows

the correlations between the percentages of thumbs-ups/thumbs-downs in the first n ratings and in the complete rating records of an article in the three datasets.

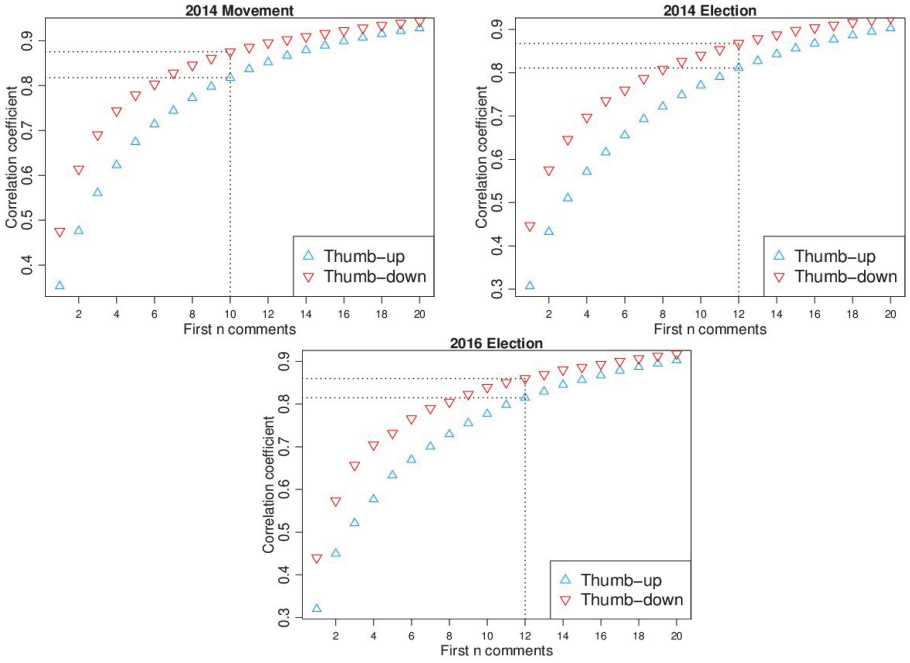

Fig. 3. Correlations between first n ratings and the complete ratings from both datasets ($1 \leq n \leq 20$).

From Fig. 3, it is evident that the correlation coefficients increase with the size of the observed ratings records. Since the threshold value α was set to 0.8, according to the methodology in Section 3.2.2, we determined the number of rating observations n = 10 for the movement dataset and n = 12 for the two election datasets because the correlation coefficients of thumbs-up and thumbs-down are both greater than 0.8. We next developed a regression model using the predictors described in Section 3.2.3. The adjusted R^2 values of the regression models with n = 10 in the movement dataset and n = 12 for the both election datasets are 0.740, 0.718, and 0.746. The average time for collecting n observations in the three datasets are 225, 365, and 468 seconds. From the results, we considered our approach only requires less than 8 minutes for observation to achieve a high prediction accuracy (adjusted $R^2 > 0.71$). Even though the experiment results exhibited high accuracy, we aimed to further improve our model. We also examined the model by changing n, the number of observations for prediction. We increased the rating records required for models from n to n+5 and investigated the improvement in prediction accuracy. The experimental results are presented in Table 3. From the table, the adjusted R^2 value of our model increases to 0.808 when n = 15 in the movement dataset. However, in the two election datasets, the adjusted R^2 increases to 0.767 and 0.786 with 15 observations, which is lower than that for the movement dataset. We considered the results show that predicting articles involving heated debates on elections is relatively difficult.

Table 3. Results of our rating score prediction model with different numbers of ratings observed.

2014 Movement				2014 Election				2016 Election			
n	Adj. R^2	RMSE	$RMSE_{cv}$	n	Adj. R^2	RMSE	$RMSE_{cv}$	n	Adj. R^2	RMSE	$RMSE_{cv}$
10	0.740	0.624	0.627	12	0.718	0.686	0.687	12	0.746	0.681	0.685
11	0.756	0.603	0.606	13	0.735	0.666	0.667	13	0.761	0.660	0.664
12	0.773	0.583	0.586	14	0.751	0.645	0.646	14	0.777	0.639	0.642
13	0.786	0.565	0.568	15	0.767	0.624	0.625	15	0.786	0.624	0.628
14	0.796	0.552	0.555	16	0.778	0.609	0.610	16	0.796	0.610	0.614
15	0.808	0.536	0.538	17	0.787	0.596	0.597	17	0.806	0.594	0.597

We also confirmed the effectiveness of our model by conducting a 10-fold cross-validation. As shown in Table 3, there is no significant difference between RMSE and $RMSE_{cv}$ in either datasets. The validation results indicate that our model can effectively predict rating scores. According to our results, our model can achieve a high accuracy (adj. $R^2 > 0.71$) with a less-than-8-minute observation. Moreover, increasing the observation time can yield better prediction results.

4.4 Identifying Influential Authors

While finding popular authors who receives many replies or extreme rating scores is quite easy, how to find authors owning both characteristics is not easy. In previous studies, most of them are focused on certain characteristics; *e.g.*, receiving the most attentions or the most endorsement from other users. However, these scholars did not focus and discuss the hybrid influential author identification method a lot yet. Therefore, in our method, we focus on this kind of "hybrid authors". To evaluate the performance of our approach with structural methodologies, we first defined one rating from author x to the article posted by author y as a directed edge $\alpha = (x, y)$, and let $\tau(\alpha)$ be the terminal vertex of edge α. Using the definition, we constructed an author-rater graph for both of our datasets. Based on the graph, we compared our proposal with the following two baseline algorithms.

1. **By in-degrees:** Since the in-degree is the number of ratings an author has received, we used it as a metric to assess author popularity. We calculated the in-degree score of every author and chose the top ones as influential authors. The in-degree of author v is defined as follows:

   $$in.degree(v) = |\alpha: \tau(\alpha) := v|.$$

2. **By PageRank:** Leveraging from the concept of PageRank, we measured each author's influences according to the link structure of the author-rater graph. We select the top authors according to their PageRank scores.

To evaluate our proposal, described in Section 3, we reviewed Katz's [38] opinion leader concept. First, Katz proposed that an opinion leader would have more network connections. However, it is unknown exactly how many people an author knows. To capture this concept more effectively for modern social media websites, we considered

that the number of replies an author has received to be another index for measuring popularity. Katz also considered that an opinion leader would have higher expertise and capability. To measure this characteristic, we used the peer-to-peer rating records as an evaluation of the capability of each author. We describe the two metrics in detail as follows.

1. **Most replied-to authors:** From the perspective of popularity, authors who receive large numbers of replies and ratings could be influential, since readers pay more attention and like to express opinions regarding the authors' articles. To measure the popularity of an author, we considered that **the average number of replies to an article** is an appropriate metric for measuring the average influence of a single article by the author of interest.
2. **Most agreed/disagreed authors:** We considered that authors who receive one-sided ratings when publishing articles are likely to be well evaluated according to the consistent opinions of users. Authors with whom the public agrees could be influential, because most of the users acknowledge the author's expertise. Authors with whom the public disagrees could also be influential, as they could represent opinion from the opposite side and might be influential on other sites. We examined **the average absolute rating score of an article** to measure authors' general levels of general acceptance/rejection.

From the objectives, we designed an experiment to demonstrate the performance author recommended by our proposal and the other two famous baselines (in-degree and PageRank). As we aim to identify top authors who have both characteristics, in the experiment, we retrieved the top n (ranging from 10 to 100) influential authors identified by the three methods and compared the performance of these top authors according to the two metrics. The results are shown in Fig. 4. From the figure, the authors selected by our proposal significantly outperform those chosen by the other two algorithms for the three datasets, especially the top 30 authors identified by our proposal. Meanwhile, the authors selected by our recommendation not only receive more replies but also receive a large number of consistent ratings from the public.

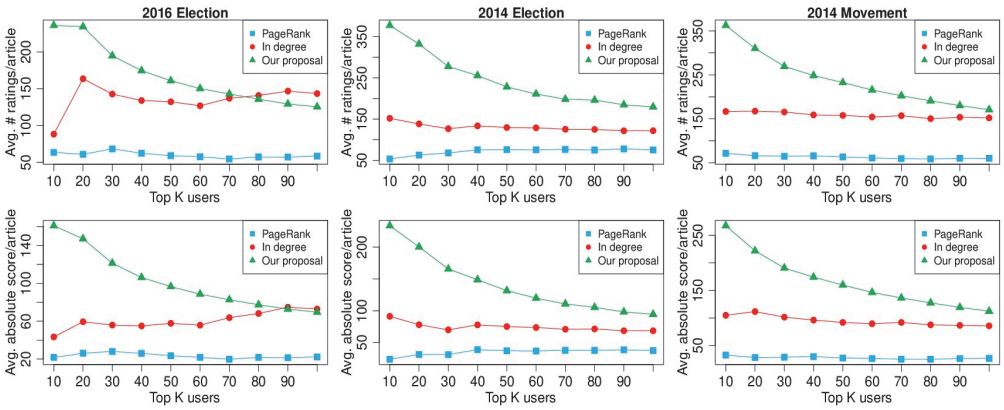

Fig. 4. Evaluation of the identified influential authors using our proposal and two previous methodologies.

5. CONCLUSIONS AND FUTURE WORK

To distinguish important articles on a social forum efficiently, we proposed methodologies to predict possible trending threads. We evaluated our proposals by studying user behavior on a popular online social forum in Taiwan during three major social events: a three-week-long social movement and two two-month-long election campaigns, including a major mayoral election and the presidential election in Taiwan. The contributions of our research are described as follows:

1. **Information filtering:** We proposed two indicators, reply quantity and rating score, by leveraging the peer-to-peer ratings records from users on a social forum. Using the indicators, we quantified the importance of articles with different characteristics.
2. **Notable thread prediction:.** We used the temporal reply behavior and early rating records to build our prediction models. We developed regression models to detect influential articles based on their early reply and rating records from online users. Our results showed that our approach can achieve a high level of accuracy within 8 minutes of observation time.
3. **Influential author identification:** We proposed identifying influential authors according to popularity and public rating consistency. Our results showed that our proposal significantly outperforms two previous link-based algorithms, *i.e.* in-degree and PageRank, in both average number of replies received and rating consistency from the public.

Besides the above contributions, this paper has some issues which could be further investigated in the future research; (1) Comparing with other baselines rather than structural methods in influential author identification can be considered; (2) In this work we focus on influential article identification; in the following studies, predicting the replying and rating behaviors on non-popular articles could be investigated.

This research has concentrated on studying social events and the social context in Taiwan. We hope that the methodologies introduced and adopted in this study for identifying notable articles on social sites will be used in the future, especially under urgent situations such as disasters or social movements. In addition, we expect that our contributions can help governments and organizations with news dissemination and the effective obtaining of the most recent public opinions.

ACKNOWLEDGEMENT

This work was supported by Ministry of Science and Technology, Taiwan, under the Grant MOST 106-3114-E-002-005, MOST 106-2221-E-002-053-MY2, and MOST 107-2218-E-035-009-MY3. We would like to thank the editor and anonymous reviewers for their valuable comments and suggestions to improve the manuscript.

REFERENCES

1. I. Rowen, “Inside Taiwan’s sunflower movement: Twenty-four days in a student-occupied parliament, and the future of the region,” *The Journal of Asian Studies*,

Vol. 74, 2015, pp. 5-21.
2. P. Norris, *Digital Divide: Civic Engagement*, *Information Poverty*, *and the Internet Worldwide*, Cambridge University Press, UK, 2001.
3. F. Li and T. C. Du, "Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs," *Decision Support Systems*, Vol. 51, 2011, pp. 190-197.
4. M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Transactions on Networking*, Vol. 17, 2009, pp. 1357-1370.
5. G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, Vol. 53, 2010, pp. 80-88.
6. S. Siersdorfer, S. Chelaru, W. Nejdl, and J. S. Pedro, "How useful are your comments?: Analyzing and predicting youtube comments and comment ratings," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 891-900.
7. F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto, "Trendlearner: Early prediction of popularity trends of user generated content," *Information Sciences*, Vol. 349, 2016, pp. 172-187.
8. K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 621-630.
9. M. Tsagkias, W. Weerkamp, and M. de Rijke, "News comments: Exploring, modeling, and online prediction," in *Proceedings of European Conference on Information Retrieval*, 2010, pp. 191-203.
10. T. B. Ksiazek, L. Peer, and K. Lessard, "User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments," *New Media & Society*, Vol. 18, 2014, pp. 502-520.
11. A. Tatar, P. Antoniadis, M. D. De Amorim, and S. Fdida, "From popularity prediction to ranking online news," *Social Network Analysis and Mining*, Vol. 4, 2014, pp. 1-12.
12. M. T. Uddin, M. J. A. Patwary, T. Ahsan, and M. S. Alam, "Predicting the popularity of online news from content metadata," in *Proceedings of International Conference on Innovations in Science*, *Engineering and Technology*, 2016, pp. 1-5.
13. J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in twitter," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 251-260.
14. O. Tsur and A. Rappoport, "What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 2012, pp. 643-652.
15. Z. Ma, A. Sun, and G. Cong, "On predicting the popularity of newly emerging hashtags in twitter," *Journal of the American Society for Information Science and Technology*, Vol. 64, 2013, pp. 1399-1410.
16. C.-T. Li, M.-K. Shan, S.-H. Jheng, and K.-C. Chou, "Exploiting concept drift to predict popularity of social multimedia in microblogs," *Information Sciences*, Vol. 339, 2016, pp. 310-331.
17. Q. Cao, H. Shen, H. Gao, J. Gao, and X. Cheng, "Predicting the popularity of online

content with group-specific models," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 765-766.
18. H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of youtube videos," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, 2013, pp. 365-374.
19. S. Nikolov and D. Shah, "A nonparametric method for early detection of trending topics," in *Proceedings of the Interdisciplinary Workshop on Information and Decision in Social Networks*, 2012, pp. 1-2.
20. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 2004, pp. 168-177.
21. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of ACL-02 Conference on Empirical Methods in Natural Language Processing*, Vol. 10, 2002, pp. 79-86.
22. P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 417-424.
23. B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, Vol. 2, 2008, pp. 1-135.
24. L. Qu, G. Ifrim, and G. Weikum, "The bag-of-opinions method for review rating prediction from sparse text patterns," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 913-921.
25. P. S. Earle, D. C. Bowden, and M. Guy, "Twitter earthquake detection: Earthquake monitoring in a social world," *Annals of Geophysics*, Vol. 54, 2011, pp. 708-715.
26. J. W. Cheng, H. Mitomo, T. Otsuka, and S. Y. Jeon, "Cultivation effects of mass and social media on perceptions and behavioural intentions in post-disaster recovery – the case of the 2011 great east japan earthquake," *Telematics and Informatics*, Vol. 33, 2016, pp. 753-772.
27. S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 1079-1088.
28. S. E. Middleton, L. Middleton, and S. Modafferi, "Real-time crisis mapping of natural disasters using social media," *IEEE Intelligent Systems*, Vol. 29, 2014, pp. 9-17.
29. M. Lim, "Clicks, cabs, and coffee houses: Social media and oppositional movements in Egypt, 2004-2011," *Journal of Communication*, Vol. 62, 2012, pp. 231-248.
30. S. Harlow, "Social media and social movements: Facebook and an online guatemalan justice movement that moved offline," *New Media & Society*, Vol. 14, 2012, pp. 225-243.
31. A. Segerberg and W. L. Bennett, "Social media and the organization of collective action: Using twitter to explore the ecologies of two climate change protests," *The Communication Review*, Vol. 14, 2011, pp. 197-215.
32. J. S. Juris, "Reflections on# occupy everywhere: Social media, public space, and emerging logics of aggregation," *American Ethnologist*, Vol. 39, 2012, pp. 259-279.
33. A. A. Olorunnisola and B. L. Martin, "Influences of media on social movements: Problematizing hyperbolic inferences about impacts," *Telematics and Informatics*,

Vol. 30, 2013, pp. 275-288.
34. M.-H. Wang and C.-L. Lei, "Modelling polarity of articles and identifying influential authors through social movements," in *Proceedings of IEEE International Conference on Systems*, *Man and Cybernetics*, 2015, pp. 1664-1669.
35. J. Kang, "A volatile public: The 2009 whole foods boycott on facebook," *Journal of Broadcasting & Electronic Media*, Vol. 56, 2012, pp. 562-577.
36. K. Gillan, J. Pickerill, and F. Webster, *Anti-War Activism: New Media and Protest in the Information Age*, Palgrave Macmillan, London, 2008.
37. N. Fenton, "Mediating hope new media, politics and resistance," *International Journal of Cultural Studies*, Vol. 11, 2008, pp. 230-248.
38. E. Katz, "The two-step flow of communication: An up-to-date report on a hypothesis," *Public Opinion Quarterly*, Vol. 21, 1957, pp. 61-78.
39. E. Katz and P. F. Lazarsfeld, *Personal Influence*, *the Part Played by People in the Flow of Mass Communications*, Transaction Publishers, NJ, 1970.
40. A. Ruvio and A. Shoham, "Innovativeness, exploratory behavior, market mavenship, and opinion leadership: An empirical examination in the Asian context," *Psychology & Marketing*, Vol. 24, 2007, pp. 703-722.
41. R. Iyengar, C. Van den Bulte, and T. W. Valente, "Opinion leadership and social contagion in new product diffusion," *Marketing Science*, Vol. 30, 2011, pp. 195-212.
42. S. J. Barnes and A. D. Pressey, "In search of the 'meta-maven': An examination of market maven behavior across real-life, web, and virtual world marketing channels," *Psychology & Marketing*, Vol. 29, 2012, pp. 167-185.
43. D. Kempe, J. Kleinberg, and É. Tardos, "Influential nodes in a diffusion model for social networks," in *Proceedings of the 32nd International Conference on Automata*, *Languages and Programming*, 2005, pp. 1127-1138.
44. J. Weng, E. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 2010, pp. 261-270.
45. K. Song, D. Wang, S. Feng, and G. Yu, "Detecting opinion leader dynamically in Chinese news comments," in *Proceedings of International Conference on Web-age Information Management*, 2011, pp. 197-209.
46. Y. Cho, J. Hwang, and D. Lee, "Identification of effective opinion leaders in the diffusion of technological innovation: A social network approach," *Technological Forecasting and Social Change*, Vol. 79, 2012, pp. 97-106.
47. Y. Li, S. Ma, Y. Zhang, R. Huang, *et al.*, "An improved mix framework for opinion leader identification in online learning communities," *Knowledge-Based Systems*, Vol. 43, 2013, pp. 43-51.
48. W. Zhang, X. Li, H. He, and X. Wang, "Identifying network public opinion leaders based on Markov logic networks," *The Scientific World Journal*, Vol. 2014, 2014, pp. 1-8.
49. N. Ma and Y. Liu, "SuperedgeRank algorithm and its application in identifying opinion leader of online public opinion supernetwork," *Expert Systems with Applications*, Vol. 41, 2014, pp. 1357-1368.
50. M. Zampieri, N. Soranzo, D. Bianchini, and C. Altafini, "Origin of co-expression patterns in e. coli and s. cerevisiae emerging from reverse engineering algorithms," *PloS One*, Vol. 3, 2008, p. e2981.

51. P. Jones, G. Harding, P. Berry, I. Wiklund, W. Chen, and N. K. Leidy, "Development and first validation of the cOPD assessment test," *European Respiratory Journal*, Vol. 34, 2009, pp. 648-654.

Ming-Hung Wang (王銘宏) received his Ph.D. degree in Department of Electrical Engineering at National Taiwan University in 2017. Before entering National Taiwan University, he received his B.S. degree in Computer Science in 2008 and M.S. degree in Communication Engineering in 2010, both from the National Tsing-Hua University. In 2018, he joined Department of Information Engineering and Computer Science, Feng Chia University, as an Assistant Professor. His research interests include network security, social media analysis, and software-defined networking.

Chin-Laung Lei (雷欽隆) received the B.S. degree in Electrical Engineering from the National Taiwan University, Taipei, in 1980 and the Ph.D. degree in Computer Science from the University of Texas at Austin in 1986. From 1986 to 1988, he was an Assistant Professor in the Computer and Information Science Department, Ohio State University, Columbus. In 1988, he joined the faculty of the Department of Electrical Engineering, National Taiwan University, where he is now a Professor. His current research interests include network security, cloud computing, and multimedia QoE management.