This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2647823, IEEE Access

Identification of Potential Collective Actions using Enhanced Gray System Theory on Social Media

Wei Yang[1], Xiaohui Cui[1], Jin Liu[2,3], Yancheng Liu[1]

[1]International School of Software, Wuhan University, Wuhan, China, 430072

[2]State Key Laboratory of Software Engineering, Computer School, Wuhan University, Wuhan, China, 430072

[3]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, China, 541004

Corresponding author: Xiaohui Cui (xcui@whu.edu.cn), Jin Liu (jinliu@whu.edu.cn)

***ABSTRACT*—A collective action that considerably affects government management and public security, e.g., a mass demonstration, usually experiences a long development period, originating from small and uncertain variations called weak signals on social media. Researchers generally identify collective action by small changes in communication frequency, emerging key words, sentiment, etc. However, most studies only consider the present environment, which may not evolve into a collective action, or conduct a short-term prediction in which significant damage is already done when the collective action is identified. This paper proposes a predictive framework to identify potential collective actions, considering the future evolution as well as the present situation, and providing a reference for early decision-making. In the framework, a future sign to describe events is improved and the enhanced gray system theory is used to predict the evolution of a future sign. Mentions of events surrounding the Arab Spring—using over 300,000 different open-content web sources crawled from social media in seven different languages—are analyzed, which suggests that the predictive framework can more precisely identify the weak signals of collective actions.**

***Index Terms*—Future sign, Weak signal, Collective action, Gray system theory.**

I. INTRODUCTION

The manifestation of collective actions, e.g., mass demonstrations, often involves collective reinforcement. The expanding effect of collective actions affects public security. In the information age today, much of the public consciousness has a significant influence online, where issues of concern are discussed [1,2,3]. With the popularization of social media, public ideas are easily disseminated, promoting the formation and development of events, which may evolve into collective actions when opinions are rapidly spread via social media. Social media has been proven as a very powerful tool for self-organizing collective actions, social events, and leisure gatherings, without any formal organizations or arrangers [4]. The Arab Spring is a typical example, revealing the importance of social media during the formation of collective actions [5]. After the protest outbreak in Tunisia, one example of the Arab Spring, someone incited 350,000 people to participate via Facebook in March 2011.

Social media can have such a considerable effect on the development of collective actions is because of the lack of public-data analysis, which leads to no corresponding early-warning. The government can take measures in time to prevent a situation from worsening and maintain public security, if social media data can be analyzed before an outbreak of events. Consequently, to receive early-warning measures before an outbreak, the government must use public data and recognize the breaking point of collective actions. Fortunately, social media can help provide situational awareness and inform predictions because people tend to share their opinions and interact with others online [6]. Due to the social-media interaction, opinions can attract more and more attention and make it possible to predict the tendencies of events. In this paper, we consider a predictive function that uses social media to predict potential collective actions, to avoid or weaken their negative effects.

Apparently, collective actions usually include a three-period cycle: sprout period, boom period, and decay period. Normal events may evolve into collective actions if they attract more attention and negatively affect society.

Normal events usually exit during the sprout period and collective actions are formed during the boom period. During the cycle, the number of news or Tweets on a collective action is the most intuitive scale parameter to describe the public's attention.

In this paper, public attention is quantized by the number of Tweets on certain events, especially as the number peaks at the end of the boom period and falls into the decay period. The aim of this research is to identify and forecast the evolution before the flashpoints of collective actions. Thus, the sprout period and the boom period are emphasized. During the sprout period, there is less attention on events that have not yet evolved into collective actions and are usually ignored by the public. However, events may attract more and more attention and become increasingly discussed on social media as the situations develop. During the boom period, public attention reaches the peak phase and relevant events are widely discussed.

This paper studies using future signs to quantify and describe the evolution of events. In this paper, a future sign is described with signals, based on the number of Tweets. Two forms of signal, weak and strong, are mainly used to describe normal events and collective actions, respectively; i.e., events can be regarded as collective actions if weak signals have evolved into strong signals described in Section 3.

Weak signals of collective actions reflect aspects of a threat or an opportunity or a sign of future change [7]. This enables a timely identification of future events or developments that are relevant for a decision maker. From the view of semiotics, a weak signal can be regarded as a future sign. Weak signals may be defined as advanced indicators of change phenomena, but they are never obvious pointers. As events evolve, weak signals attract little attention from the public and are usually ignored. The weak signal was first introduced by Ansoff [8] and deepened by later researchers. Weak signals have been described with the triadic model, which can be drawn in a three-dimensional space [9]. In this paper, the forms of weak signals are developed according to the characters of collective actions. From the signal perspective, the sprout period is defined as having weak signals and may evolve into strong signals, which are the indicators of collective actions [9]. The details are presented in Section 3.

Accordingly, to identify and analyze potential collective actions, a predictive framework is proposed. In the framework, weak signals of events are discovered firstly and predicted whether they could be evolved into strong signals.

In summary, this work's contributions are as follows:

1. A productive framework is proposed, which can identify and forecast whether events can evolve into collective actions.
2. A form of future sign is improved to describe and quantize the evolution process of collective actions.
3. Gray system theory enhanced by adding piecewise polynomial fitting and Lagrange's interpolation method is applied to predict the change of future signs in the future.

The remainder of the paper is organized as follows. Section 2 briefly describes the complete approach of the predictive framework. Section 3 introduces the improved form of future signs for collective actions. Section 4 explains how to predict the change of weak signals of collective actions. Then, Section 5 presents experiments, based on 300,000 different open-content web sources from 18 countries, which were conducted to verify the performance of the predictive framework. Section 6 describes related work. Finally, Section 7 concludes the paper with a discussion of future work.

II. APPROACH

This section briefly describes the complete idea of this study. The predictive framework proposed in this paper consists of two phases: identification and prediction, as shown in Fig. 1. In the identification phase, the framework evaluates events based on the sentiment and number of related Tweets. The sentiment is analyzed using a sentiment analysis tool, SentoStrength. Thus, the number and sentiment are comprehensively transformed into the Interpretation. Note that, if the entire sentiment is negative and the value of the Interpretation is higher than a threshold, the related event is regarded as a collective action; otherwise, the signal related to the event is regarded as a weak signal in the identification phase. However, the identification phase only considers the current situation, ignoring the future change of the number. For the prediction phase, gray system theory is used to judge whether the weak signal can evolve into a strong signal based on analyzing the sentiment and predicting the change of the number before the flashpoint of a collective action. Then, the Interpretation is used to judge whether the event has become a collective action.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2647823, IEEE Access

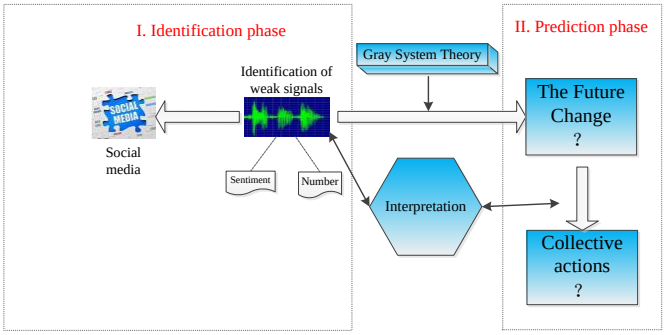

Fig. 1. The predictive framework of potential collective actions

III. FUTURE SIGN

A future sign mainly consists of two forms of signal, weak and strong. As mentioned before, a weak signal indicates a normal event and a strong signal indicates a collective action. In addition, weak signals may evolve into strong signals if the related events attract sufficient attention from the public and negatively affect society. The formalizations of future sign is described in formula (1). The weak signal was first introduced by Ansoff [8], which contains premature and imperfect information for future problems. In our view, weak signals are interesting and useful because they are full of information about new permanencies and definitive transitions [10]. Actually, the public usually pays little attention to weak signals of events because they have hardly any noticeable effect on society at the moment.

The first major contributors to defining a future sign in semiotics were Ferdinand de Saussure (1857–1913) and Charles Sanders Peirce (1839–1914). Saussure offered the 'dyadic' model for the future sign. He defined the sign as being composed of the signifier (e.g., the form the sign takes) and the signified (e.g., the concept it represents) [11]. Peirce, on the other hand, provided the triadic model of the sign, consisting of the representamen, the interpretant, and the object. The representamen stands for the form that the sign takes, e.g., newspapers, websites, pictures, and videos. The interpretant is the receiver's understanding of the future sign's meaning, and the object is that to which the sign refers [12].

To deepen understanding of the triadic model, Elina Hiltunen developed the future sign in three-dimensional space (Issue, Signal, and Interpretation)—based on the triadic model proposed by Peirce—which could graphically describe the evolution process of future signs, from weak signals to strong signals [9]. Hiltunen proposed a universal framework that focused on explaining an idea. However, it did not study a specific model on how to obtain an Interpretation based on a Signal and an Issue; i.e., she did not build the functional relation between the Interpretation and two independent variables. Thus, to study certain problems or events, the universal framework should be improved individually.

Based on Hiltunen's research, this paper improves the triadic model in three-dimensional space to a quaternionic model (Issue, Signal, Interpretation, and Sentiment). Moreover, this paper studies the specific functional relation between the Interpretation and the three other independent variables, and quantizes the Interpretation, which describes the probability that a normal event will evolve into a collective action. That is, the future sign of the event evolves from weak signals to strong signals.

Based on Hiltunen's triadic model, the dimensional units of the improved form are the following:

The Issue (I): Unlike the studies of Peirce and Hiltunen, whose Issues are all topics, which may contain several events and can predict the trend of a new phenomenon, the Issue of this paper represents a signal event or specific collective action, i.e., |Issue| = 1 in the quaternionic model. Thus, this paper mainly studies the evolution of one event from a simple situation. In future work, the situation where |Issue| >= 1 will be studied. Because |Issue| = 1, the quaternionic model can be simplified to a triadic model consisting of Signal, Interpretation, and Sentiment.

The Signal (S) is the number of forms the future sign takes. The forms are visible media, e.g., social media, newspapers, and pictures. In this paper, the Signal is represented by the number of Tweets about the discussed event.

The Sentiment (E) is positive or negative. This paper focuses on events whose sentiment is negative, which is obtained by analyzing the content of users' Tweets with Semto-Strength. Actually, some users may take a positive attitude and others may take a negative attitude. If the sentiment is negative, E = -1; otherwise, E = 1.

The Interpretation (θ) represents the receiver's understanding of the future sign's meaning. It explains the probability of the event happening, and is described by a decision parameter ($\theta = f(t, S)$), which indicates whether the event has become significant and has attracted more attention from the public, as measured by the number of Tweets. For the theory of θ, the research mainly improves Nathan Kallus's idea,

where the threshold value θ' is = 2.875 (which is also nearly the 94th percentile of the standard exponential distribution) [1]. That is, there are massive reports on the events. On this basis, this study considers the sentiment of Tweets. It suggests that if $|\theta| > \theta'$ and $\theta < 0$, the corresponding weak signal has become into a strong signal. Note that, $\theta < 0$ *means* E = -1.That is, the weak signal has become a strong signal, as shown as in formula (1).

$$future\ sign \begin{cases} weak\ signal, & \theta \le \theta' \\ strong\ signal, & |\theta| > \theta' \text{and } \theta < 0 \end{cases} \quad (1)$$

Moreover, the study defines the first day of $\theta > \theta'$ as the flashpoint of a certain collective action.

Besides, the paper develops a model F shown as formula (2)

$$\theta = f(i,\ S) \quad (2)$$

$$= E \times \frac{1}{3}\sum_{j=i-1}^{i+1} \frac{S_c'(j)}{\overline{S_c'}}$$

where $S_c'(j) = \frac{S_c(j)}{\frac{1}{|Countries|*90}\sum_{c' \in Countries}\sum_{j=i-90}^{i-1} S_{c'}(j)}$, and the explanations of every symbol is as follows. $\overline{S_c'} = \frac{1}{|Train|}\sum_{i \in Train} S_c'(j)$. $S_c(j)$ denotes the number of events in country c published on day j in Twitter. Countries denote the set of countries where the relevant events are reported. Since new media are being added daily to the Recorded Future source bank, there is a heterogeneous upward trend in the event data. To remove this trend the average volume in the trailing three months is used to normalize the mention number. $S_c(j)$. Next, $\overline{S_c'}$ is training-set average number reported on mainstream media. Train denotes the set of days in the training set.

IV. PREDICTION: EVOLUTION OF FUTURE SIGN

In Section 3, the Interpretation is introduced to judge whether a future sign has evolved into a collective action; however, it simply analyzes the current situation and does not consider the future evolution. Collective actions have many internal factors and complex relations. In addition, the mechanism of how the factors influence the development of events is not always clear to the public. Fortunately, gray system theory can analyze similar social systems [13]. This study tries to apply gray system theory to predict the variety of Signals, which describe the social system at a macro level. Then, the Interpretation is obtained, based on the predictive Signals, to predict whether weak signals (normal events) can evolve into strong signals (collective actions). Moreover, the reasoning process of introducing a gray system will be presented to explain why the gray system theory is appropriate.

To enhance the effectiveness of our framework, the study predicts a variety of Signals about the discussed events before the special flashpoints (denoted as t_0). The special flashpoints mean that the time when the corresponding Interpretation of Signal becomes higher than θ'.

A. Gray System Theory

Gray system theory has been used to investigate uncertain problems that lack data and information. The main idea of gray system theory is to correctly describe the systems' evolution law and effectively monitor their running behavior. Moreover, gray system theory can also reveal the long process of the continuing development of events. After a long development, the basic model of gray system theory has formed, including a Gray Matrix, Gray Algebra, and Gray Equation. The gray model (GM) may also implement the analysis, evaluation, prediction, and control decision of uncertain data systems by using spatial association rules and a sequence generation method [14]. Several forms of gray model exist, e.g., GM (1, 1), GM (1, N), GM (N, 1). Because only one variable, the Signal, needs to be predicted, this paper uses GM (1, 1) to analyze the Signal's variation.

GM (1, 1) usually consists of three steps: Initializing the data, building the GM (1, 1) model, and solving the GM (1, 1) model; these will be described in the following section.

1) Initializing Data

GM does not need analyze what a certain distribution the original data follows, but adopts the method of accumulation that can weaken the influence of randomness of the data. In this paper the original data is the number of tweets every day during a period, defined as Signal(S) in formula (3).

$$S^{(0)} = (\ S^{(0)}(1),\ S^{(0)}(2),\ \cdots S^{(0)}(k),\ \cdots, S^{(0)}(t_0)) \quad (3)$$

Where, t_0 is the special flashpoint. Moreover, this study applies two operators to the original data series. First, because of the data missing problem, this study uses piecewise polynomial fitting and Lagrange's interpolation method to add the real data set. For the fitting method, the study divides the data series into three subseries and uses MATLAB to fit the polynomial function for every subseries. Thus, Lagrange's interpolation method is used to add the missing data at the certain time-step. Thus, the updated data series is described as for-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2647823, IEEE Access

mula (4).

$$S^{(1)}=(\ S^{(1)}(1),\ S^{(1)}(2),\ \cdots,\ S^{(1)}(l)\cdots\ S^{(1)}(t_0)) \quad (4)$$

Where the polynomial function is $f(k)=\sum a_n * k^n$. n is tunable and a_n is obtained by MATLAB. The Lagrange's interpolation function is $L_n(k)=\sum_{i=1}^{n} f(k)\ (\prod_{\substack{j=1\\ j\neq i}}^{n}\frac{k-k_j}{k_i-k_j})$. $S^{(1)}(l)=L_n(l)$ if the data is missing at time-step l in the original data series.

Then, addition operator is also applied to the new data series, reducing the noise data, and the cumulative Signal series is shown in (5)

$$S^{(2)}=(\ S^{(2)}(1),\ S^{(2)}(2),\ \cdots,\ S^{(2)}(t_0)) \quad (5)$$

Where $S^{(2)}(i)=\sum_{j=1}^{i} S^{(1)}(j)$, i=1,2,$\cdots$, t_0.

*2***)** *Building Gm*(1,1) *Model*

This phase describes the reasoning process of building GM(1,1) model. In this study, building GM(1,1) model depends on the growth rate of the number of the tweets (Signal). Moreover, the growth rate is closely related to the followers of disseminators in Twitter. However, not all followers must retweet the information. So the real growth rate is decided by the actual retweeted tweets. In this paper, the growth rate is denoted by $F(S^{(2)}(i))$. Then, the general model is described by a formula in (6)

$$\lim_{\Delta t\to 0}\frac{S(i+\Delta t)-S(i)}{\Delta t}=F(S^{(2)}(i)) \quad (6)$$

In formula (6), $F(S^{(2)}(i))$ is not concrete functional expression and it is impossible to solve the formula. So the study starts with a simple situation that we suppose there is a linear relation between the growth rate and the Signal. In the future, we will further study the more complex situation. Hence, this study sets $F(S^{(2)}(i))$=-a* $S^{(2)}(i)$+b. Combing with the discussion in the phase of initializing data, the final model based on (6) is shown as in formula (7)

$$\lim_{\Delta t\to 0}\frac{S^{(2)}(i+\Delta t)-S^{2}(i)}{\Delta t}=F(S^{(2)}(i))$$

$$\lim_{\Delta t\to 0}\frac{S^{(2)}(i+\Delta t)-S^{2}(i)}{\Delta t}=\text{-}a*S^{(2)}(i)+b$$

$$\frac{dS^{(2)}(i)}{di}+a*S^{(2)}(i)=b \quad (7)$$

Actually, formula (7) is just GM(1,1) model, a differential equation of first order with two parameters(a,b). –a represents the variation trend of Signal and b is used to control the intercept that represents the initial growth rate of $S^{(2)}(i)$. To predict the variation of the Signal, this study uses the least square method to solve the two parameters.

3*) Solving The Model*

This study uses the least square method to solve the approximate solutions of the two parameters (a,b) and then obtain the functional expression of $S^{(2)}(i)$. Hence, we can predict the variation of Signal.

● *Approximating the Parameters*

In this study, the data is time series that is discrete values. Then, we set $dS^{(2)}/di\approx S^{(2)}(i+1)-S^{(2)}(i)=S^{(1)}(i)$. Combing with formula (7), we get equation shown in (8)

$$S^{(1)}(i)+a*S^{(2)}(i)=b \quad (8)$$

Let i=1,2$\cdots t_0$, we get the following equations shown in (9)

$$\begin{cases} S^{(1)}(1)+a*S^{(2)}(1)=b \\ S^{(1)}(2)+a*S^{(2)}(2)=b \\ \cdots\ \cdots \\ S^{(1)}(t_0)+a*S^{(2)}(t_0)=b \end{cases} \quad (9)$$

Using the least square method, we get $\theta=(a,b)^T=(X^TX)^{-1}X^TY$. The process is shown as follows.

Let Y=($S^{(1)}(1)$, $S^{(1)}(2)$, $\cdots$, $S^{(1)}(t_0)$)^T, θ=$(a,b)^T$,

$$X=\begin{bmatrix} -S^{(2)}(1) & 1 \\ -S^{(2)}(2) & 1 \\ \vdots & \vdots \\ -S^{(2)}(t_0) & 1 \end{bmatrix}=\begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(t_0)} \end{bmatrix},$$

As an initial choice, we decide to approximate $S^{(1)}(i)$ as the linear function of a and b: $S^{(1)}(i)\approx F(S^{(2)}(i))$ =-a* $S^{(2)}(i)$+b. Thus, $F(S^{(2)})=X\theta$

$$X\theta\text{-}Y=\begin{bmatrix} -S^{(2)}(1)a+b \\ -S^{(2)}(2)a+b \\ \vdots \\ -S^{(2)}(t_0)a+b \end{bmatrix}-\begin{bmatrix} S^{(1)}(1) \\ S^{(1)}(2) \\ \vdots \\ S^{(1)}(t_0) \end{bmatrix}$$

Thus, using the fact that for a vector z, we have that $z^Tz=\sum_i z_i^2$ and the cost function is:

$$J(\theta)=\frac{1}{2}(X\theta\text{-}Y)^T(X\theta\text{-}Y)=\frac{1}{2}\sum_{i=1}^{t_0}(F(S^{(2)}(i))-S^{(1)}(i))^2 \quad (10)$$

Finally, to minimize $J(\theta)$, we find its derivatives with respect to θ. Combing with matrix operations, we find that

$$\begin{aligned}\nabla_{A^T}\mathrm{tr}ABA^TC&=(\nabla_A\mathrm{tr}ABA^TC)^T=(CAB+C^TAB^T)^T\\&=B^TA^TC^T+BA^TC\end{aligned} \quad (11)$$

Hence,

$$\nabla_\theta J(\theta)=\ \nabla_\theta\frac{1}{2}(X\theta\text{-}Y)^T\text{-}(X\theta\text{-}Y)$$

$$=\frac{1}{2}\nabla_\theta(\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta\text{-}\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}Y-Y^{\mathrm{T}}\mathrm{X}\theta+Y^{\mathrm{T}}Y)$$

$$=\frac{1}{2}\nabla_\theta tr(\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta\text{-}\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}Y-Y^{\mathrm{T}}\mathrm{X}\theta+Y^{\mathrm{T}}Y)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2647823, IEEE Access

$$=\frac{1}{2}\nabla_\theta(tr\theta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}\theta-2\mathrm{tr}Y^{\mathrm{T}}\mathrm{X}\theta$$

$$=\frac{1}{2}(X^TX\theta+\ X^TX\ \theta-2X^TY)$$

$$=\ X^TX\theta-\ X^TY$$

In the third step, we used the fact that the trace of a real number is just the real number; the fourth step used the fact that tr$A$=trA^T, and the fifth step used equation (11) with A^{T}=,B=B^{T}=$X^{\mathrm{T}}X$, and C=I, and the fact that $\nabla_A\mathrm{tr}AB=B^T$. To minimize J, we set its derivatives to 0, and obtain the equations: $X^TX\theta=X^TY$

Thus, the value of θ that minimizes $J(\theta)$ is given in closed form by the equation (12)

$$\theta=(a,b)^T=(X^TX)^{-1}X^TY \tag{12}$$

- *Solution of $S^{(2)}(i)$*

This phase introduces how to obtain the functional equation of $S^{(2)}(k)$ shown in formula(7). The process is shown as follows.

$$dS^{(2)}(i)/di+a*S^{(2)}(i)=b$$

$$[1/(\ b/a-\ S^{(2)}(i))]\ dS^{(2)}(i)=adi$$

$$\int[1/(\ b/a-\ S^{(2)}(i))]\ dS^{(2)}(i)=\int a\ di$$

$$-ln|b/a-\ S^{(2)}(i)|=ai+\mathrm{C}$$

Let the growth rate - $a*\ S^{(2)}(i)+b$=0, we get the upper bound value of $S^{(2)}$, that is $S^{(2)}(i)=b/a$. Hence, $b/a-\ S^{(2)}(i)$>=0. Thus,

$$Ln(b/a-\ S^{(2)}(i))=-ai-\mathrm{C}$$

$$S^{(2)}(i)=\ b/a-\ e^{-ai}e^{-C}$$

With the fact that when i=1, $S^{(2)}(1)=\ S^{(1)}(1)$, described as equation (5), $e^{-C}=(\ b/a-\ S^{(1)}(1))e^a$. Thus,

$$S^{(2)}(i)=\ b/a-\ e^{-ai}[b/a-\ S^{(1)}(1)]e^a$$

$$=\ b/a-[b/a-\ S^{(1)}(1)]\ e^{-a(i-1)}$$

Hence,

$$S^{(2)}(i)=[S^{(1)}(1)-\frac{b}{a}]e^{-a(i-1)}+\frac{b}{a},\ i=1,\cdots,t_0 \tag{13}$$

- *Predictive Signal*

Because $S^{(2)}(i)$ is the cumulative data series, then we need use subtraction operator to obtain the Signal, the predictive value $\hat{S}^{(1)}(i)$ of $S^{(1)}(i)$, which is shown in equation (14)

$$\hat{S}^{(1)}(i)\ =\ S^{(2)}(i)-S^{(2)}(i-1)$$

$$=[\frac{b}{a}-S^{(1)}(1)]*[\ e^{-a(i-1)}-\ e^{-a(i-2)}$$

$$=(1-e^a)[\frac{b}{a}-S^{(1)}(1)]*e^{-a(i-1)},\ i=1,\cdots,t_0 \tag{14}$$

B. The Interpretation

After predicting the time series of Signal, the Interpretation θ is obtained based on formula (2) $\theta=f(i,S)$. Thus, we can make sure whether the normal event has evolved into a collective action based on (1). That is, if $|\theta|>\theta'$ and$\theta<0$, the normal event has evolved into a collective action and it can provide reference for decision-making.

V. EXPERIMENTS AND ANALYSIS

This section discusses the performance of the predictive framework to identify the weak signal of potential collective action (PFDWS) proposed in this study. The experiments mainly consist of two parts, evaluating the accuracy of PFDWS for predicting the variety of Weak signals for the mass protests and comparing with other two methods with respect to finding weak signals proposed by other researchers[15].

A. Data Sets

This study implemented experiments on the 300,000 different open content web sources in Twitter from 18 countries such as Afghanistan, India, Egypt and Italy. We obtain the data sets from the website (*www.recordedfuture.com*) collected by Recorded Future [1]. The Data Sets involves mass protests in different country, which are about the Arab Spring Starting from Tunisia. Every piece of data consists of author, location, category, fragment, time.

B. Evaluation of PFDWS

To evaluate the performance of PFDWS we proposed, this study compared it with other researches, the works in [15](VTID).The method, VTID, proposed a predictive model consisted of a variety of techniques for the detection of weak signals in social media. These techniques include (i) keyword analysis, (ii) geo-spatial analysis, (iii) frequency analysis, (iv) semantic analysis and (v) sentiment analysis. Actually, VTID, depends on the variety of weak signals in social media.

Firstly, we extracted every signal event (contains the normal events and collective actions) from 300,000 different open content web sources in Twitter. By comparing the keywords of every record, there were 113 events altogether. To enhance the efficiency, we first use SemtoStrength to analyze the sentiment of tweet about related events and mainly study the events whose sentiment is negative. Besides, we count the number, regarding as the Signal, of every event from the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2647823, IEEE Access

sprout period to flashpoints every day. According to our manual statistics, there are 34 collective actions. In the experiments, the PFDWS adopted gray System Theory to predict the variety of Signal of every event. Hence, the Interpretation was used to judge whether the event could become a collective action. And the predicting error of the Signals, recall and precision were regarded as important evaluate indexes for the performance of the PFDWS.

I. approximating the parameters

This section focuses on evaluating the accuracy of predicting the Signal for every event. For every event, the experiment chose different sizes of subset of the Signal series, from four days before of the size to the whole, to observe the errors of PFDWS. For example, if the Signal series of one event is S=(S_1, $S_2 \cdots S_k, \cdots$ Sn), the experiment chose subsets, (S_1, $S_2 \cdots S_k$), (S_1, $S_2 \cdots S_{k+1}$), (S_1, $S_2 \cdots S_{k+2}$), (S_1, $S_2 \cdots S_{k+3}, \cdots$ Sn), to approximate the parameters (a,b) and evaluate every predicting error of the method.

Because there were 113 pairs of estimated values for a and b. By experiments, we got the values for every event and they were listed in TABLE I.

TABLE I

THE OPTIMAL PARAMETERS AND THE CORRESPONDING ACCURACY FOR THE 113 EVENTS

a	*b*	*a*	*b*	*a*	*b*
0.12562567	46.8611518	-0.1693223	35.3277514	-0.0015402	20.2805391
0.08426172	13.9577368	-0.0303955	35.4262725	0.00455158	20.510654
0.04146661	34.3019285	-0.2435111	35.5460102	0.05420156	22.06811
-0.314216	5.11043938	0.00749664	35.676788	-0.0736505	24.1743982
-0.3717634	12.5787305	0.30397196	35.7855824	-0.0417354	26.1727423
-0.4502763	5.91411384	-0.0655452	35.8540575	0.03142392	27.9903007
0.22805497	47.7814912	0.05802241	35.9171816	-0.0659203	26.2790862
0.35061175	48.0128324	0.04399174	36.0249604	-0.1386283	24.1893572
0.10582092	17.0885702	0.04153392	36.0806233	-0.0458104	25.562298
0.26917795	84.2183681	0.1283957	36.0891826	-0.1094934	27.3029748
0.32795682	39.3749262	0.0545359	36.1422734	-0.0521172	29.3229194
0.03456788	55.5605478	-0.0071103	36.2333672	-0.0911861	31.4342671
0.08015967	24.0979717	-0.0167601	36.3533173	-0.0398006	32.7028094
-0.1344766	30.4787788	-0.1150064	36.4535042	-0.1982099	33.7209292
0.31702814	67.8066498	-0.0712706	36.5325087	-0.0070324	35.3289579
-0.0694797	25.645171	-0.0113679	36.6218736	0.3176165	37.0922922
-0.1678638	42.4509987	0.00791731	36.7229864	-0.0957405	38.8848345
0.10342751	68.5547384	-0.03226	36.8026795	0.04945618	40.5855115
0.2283737	-1.8448261	-0.033878	10.2366842	0.04311905	42.066276
-0.1074952	56.6374186	0.00228125	20.8550968	0.04640935	43.5448459
0.13565709	15.9157369	-0.1239357	20.8478487	0.12714788	39.3259396
-0.0039506	48.6818793	0.02311909	14.5186634	0.06069283	34.5029338
0.18485953	14.7188415	-0.0690852	21.4444816	0.0188739	36.0080842
0.06581319	35.9390318	-0.01701	15.1684041	-0.02557	38.2340427
0.1250407	14.7188415	0.28227236	1.8062283	0.02904921	36.9725544
0.10053244	12.7271801	-0.0246108	22.4887901	-0.0552137	35.7789017
-0.1889366	8.86856132	0.03594963	6.64699358	-0.0054525	38.4854409
-0.1017516	7.36573415	0.06440066	3.18996675	-0.0065568	41.1897503
-0.2649937	6.33660169	0.01704609	27.6964081	0.10249684	28.3282838
0.13513283	6.12568857	0.11018345	6.65159014	-0.0470823	28.6277589
-0.1445471	7.04426013	0.05504714	9.92290016	-0.0610345	35.9390318
0.09247967	8.19806246	-0.0275141	29.0898413	-0.0271211	40.9052518
0.29115038	9.21155028	-0.1187729	17.820691	-0.0897103	44.7481545
0.27671462	9.77316657	-0.021619	6.73396415	-0.0537729	48.0669465
0.05825041	10.1541923	-0.0267863	12.0212995	-0.113671	48.981832
0.0972142	10.3039544	-0.0905414	7.26972584	-0.1154133	49.213391
-0.0371337	10.1644552	-0.1266682	20.512833	0.06865219	49.895738
0.14854613	10.0762945	0.00093754	14.8259074		

And we measured the predicting error for every size of training set with equation (15).

$$Error = \frac{\sum_{i=k+1}^{t_0} |\hat{S}^{(1)}(i) - S^{(0)}(i)|}{S^{(0)}(i)} * 100\% \qquad (15)$$

where $\hat{S}^{(1)}(i)$ is the predictive Signal.

To display the variation of predicting error for different sizes of subsets, we randomly select 113 events in Afghanistan and set k as 4. The experimental result is shown in Fig. 2.

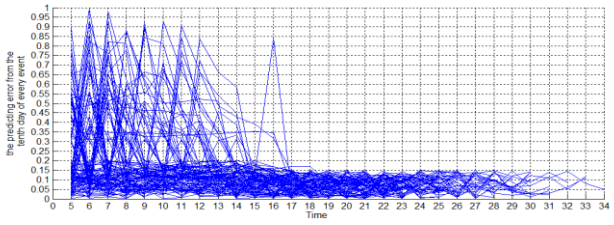

Fig. 2. The variation of predicting error for different sizes of sub-Signal series from the tenth day of every event

In Fig. 2, "5" of horizontal axis means we chose the Signals of the first four day of a certain events as the training set to approximate parameters and predict the variation of Signals during the fifth day to the special flashpoints. Thus, the labels from "6" to "34" indicate the similar meanings. As Fig. 2 shows, the predicting errors mostly are lower than 20% at

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2017.2647823, IEEE Access

the beginning and then almost vary between 0 and 15% , starting from fifteenth day. It indicates that our PFDWS can predict the variation of Signals of events accurately, which just need analyze the Signals of the first fourteen day.

II. Performance of the PFDWS

This section focuses on comparing the precision and recall of PFDWS and VTID [15]. We conducted the two methods to identify the weak signal of collective actions. The experimental results were shown in TABLE II and TABLE III. From TABLE II, it shows that 32 collective actions are identified correctly and the 2 ones are not identified for PFDWS. And 72 normal events are identified correctly, but the others are regarded as collective actions falsely. TABLE III shows that 30 collective actions are identified correctly and the 4 ones are not identified for PFDWS. And 67normal events are identified correctly, but the others are regarded as collective actions falsely.

TABLE II

THE IDENTIFICATION RESULT OF PFDWS

	Collective action	Normal event
Identify	32	5
Not identify	2	72

TABLE III

THE IDENTIFICATION RESULT OF VTID

	Collective action	Normal event
Identify	30	10
Not identify	4	67

Based on TABLE II and TABLE III, the precisions and recalls of the two methods are shown in TABLE IV.

TABLE IV

THE COMPARISON BETWEEN PFDWS AND VTID

	Precision	Recall
PFDWS	86.49%	94.12%
VTID	75%	88.24%

From TABLE IV, it reveals that the VTID is better than PFDWS from the views of recall and precision. In general, the high communication frequency means the collective actions may have formed in VTID. However, the corresponding events may not attract more public attention and affect the society destructively. If the tweets' data is graphics with line chart, collective actions correspond to the peaks. In terms of the math, the peaks correspond to local maximums. Then VTID may regards most local maximums as the formation of collective action, while not all the relevant events with local maximums have evolved into collective actions. However, our PFDWS considered the situation that it applied the threshold to judge whether the local maximums is higher than it. What's more, a few collective actions may experience a long time to outburst. The communication frequency in social media may increase slowly. Thus, VTID may not detect this kind of collective actions. However, our PFDWS discovers collective actions based on the Interpretation that doesn't ignore the collective actions.

VI. RELATED WORK

This study focuses on predicting the evolution of future sign for potential collective actions. This section will introduce related work on future sign and collective actions. For one fundamental form of future sign, weak signals pay the most important role in predicting future trend. Many relevant researches on mining and analysis in online social media have focused on weak signals. Weak signal can be used to identify the potential collective action, which may be ignored in the sprout period. The concept of weak signals has been introduced as early warning system to advance strategic planning[8,16]. Some researchers have studied weak signals from the perspective of Social Network Analysis(SNA), which is used to analyze the relationships and the information exchanged between actors((i.e., individuals or groups)). This method can reveal groups of actors within a network after examining the relationships existing between them based on the number and frequency of information exchanged [17]. Besides, text classification is already used for identification of weak signals theoretically[16]. Examples are the k nearest neighbor classification, simple probabilistic algorithms (e.g.naive Bayes), decision tree models (e.g.C4.5), and support vector machine algorithms[18-24]. C Macrae analyzed organisational and cultural challenges to identify, interpret, integrate and act on the early warnings and weak signals of emerging risks-before those risks contribute to a disastrous failure of care, and then healthcare organisations and their regulators can improve safety and address emerging risks [25]. Ponomareva et al. explored the key stages and methods for the identification of weak signals (WS) in foresight methodology [26]. In this work, key characteristics and fea-

tures of signals were identified. The key groups of methods consisted of scanning and monitoring; data analysis; modelling, clustering, interpretation; expert procedures.

On the other hand, there are many related researches on collective actions. Nathan Kallus [1] proposed a criterion to judge whether an event has become one collective action based on the amount of tweet, and this study adopts the criterion. Lei Hou et al. theoretically proposed a heterogeneous model for opinion formation dynamics where they used conviction to measure an agent's ability to insist his opinion. Results revealed that, the collective opinion of steady-state was closely related with the initial bias of the opinion leaders [27]. Markos Avlonitis et al. proposed a stochastic transformation of the aggregated users' interaction signal into a space defined by its correlation to the bell-shaped reference patterns that was shown to offer significant amelioration as to the percentage of users' interaction required in order to achieve comparable results to the original users' interaction space. Based on the users' interaction, this method can predict the video users' collective activity [28]. MV Anauati conducted a laboratory experiment to test the comparative statics predictions of a new approach to collective action games based on the method of stability sets. They found robust support for the main theoretical predictions and subjects tend to upgrade their prior beliefs as to the expected share of cooperators, as the payoff of a successful collective action increase [29].

VII. CONCLUSIONS

Collective actions considerably harm the people's safety and the social stability. It is necessary to forecast the evolution of the weak signal of collective actions. If the weak signal of collective actions can be identified in the early period, it can help the decision-makers realizing early warnings. To reach the goal, this paper proposes the predictive framework to identify weak signals of potential collective actions considering the future evolution, evaluating whether the events can evolve into collective actions based on gray system theory. What's more, the form of Signals is improved to describe and quantize the evolution process of collective actions. However, this study just analyzes the present sentiment of events, which may changes. Thus, in the future work, the study will focus on improving the forecasting technique to enhance the precision and predicting the variation of sentiment. Though this paper mainly used GM(1,1) to conduct the prediction of the evolution of signals, the predictive framework we proposed is almost universal, which can transform into different differential forms considering the actual data distribution environment. What's more, we will study how to predict future trend about a potential topic instead of a certain events based on future evolution.

ACKNOWLEDGEMENTS

This work was supported in part by National Nature ScienceFoundation of ChinaNo.61440054, Fundamental ResearchFunds for the CentralUniversities of China No. 216-274213, and Nature Science Foundation of Hubei, China No.2014CFA048,2015BAA052.Outstanding Academic Talents Startup Funds of Wuhan University, No. 216-410100003.

This work was partially supported by the National Natural Science Foundation of China (Grand No. 61572374, U163620068, U1135005), Guangxi Key Laboratory of Trusted Software (No.kx201607) and the Academic Team Building Plan for Young Scholars from Wuhan University.

REFERENCES

[1] Kallus N. Predicting crowd behavior with big public data[C]// Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, 2014:625-630.

[2] Xu Z, Zhang H, Hu C, et al. Building knowledge base of urban emergency events based on crowdsourcing of social media[J]. Concurrency and Computation: Practice and Experience, 2016.

[3] Xu Z, Zhang H, Sugumaran V, et al. Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media[J]. EURASIP Journal on Wireless Communications and Networking, 2016.

[4] Hintikka K A. Communication structure and collective actions in social media[C]// International Academic Mindtrek Conference: Envisioning Future Media Environments. ACM, 2010:201-204.

[5] González-Bailón S, Borge-Holthoefer J, Rivero A, et al. The dynamics of protest recruitment through an online network.[J]. Scientific Reports, 2011, 1(7377).

[6] Xu J, Schaar M V D, Liu J, et al. Forecasting Popularity of Videos Using Social Media[J]. IEEE Journal of Selected Topics in Signal Processing, 2015, 9(2):330-343.

[7] Thorleuchter D, Poel D V D. Weak signal identification with semantic web mining[J]. Expert Systems with Applications, 2013, 40(12):4978-4985.

[8] Ansoff H I. Managing Strategic Surprise by Response to Weak Signals[J]. California Management Review, 1975, 18(2):21-33.

[9] Hiltunen E. The future sign and its three dimensions[J]. Futures, 2008, 40(3): 247-260.

[10] Mendonça S, Cardoso G, Caraça J. The strategic strength of weak signal analysis[J]. Futures, 2012, 44(3): 218-228.

[11] D. Chandler, Semiotics for Beginners—Criticism of Semiotic Analysis, /http://www.aber.ac.uk/media/Documents/S4B/sem11.htmlS, opened 30 August 2006.

[12] http://www.bartleby.com/65/re/realism3.html, The Columbia Encyclopedia, sixth ed., 2001–05.

[13] Dejamkhooy A, Dastfan A, Ahmadyfard A. Modeling and Forecasting Non-Stationary Voltage Fluctuation Based on Grey System Theory[J]. IEEE Transactions on Power Delivery, 2015:1-1.

[14] Darong H, Jianping T, Ling Z. A Fault Diagnosis Method of Power Systems Based on Gray System Theory[J]. Mathematical Problems in Engineering, 2015, 2015:1-11.

[15] Charitonidis C. Weak Signals as Predictors of Real-World Phenomena in Social Media[C]// Ieee/acm International Conference on Advances in Social Networks Analysis and Mining. ACM, 2015:864-871.

[16] Tabatabaei N, Tabatabaei N. Detecting Weak Signals by Internet-Based Environmental Scanning[J]. 2011.

[17] C. Haythornthwaite, "Social network analysis: An approach and technique for the study of information exchange," Library and Information Science Research, vol. 18, no. 4, pp. 323 – 342, 1996.

[18] Buckinx W, Moons E, Poel D V D, et al. Customer-adapted coupon targeting using feature selection[J]. Expert Systems with Applications, 2004, 26(4):509-518.

[19] Xu Z, Luo X, Zhang S, et al. Mining temporal explicit and implicit semantic relations between entities using web search engines[J]. Future Generation Computer Systems, 2014, 37: 468-477.

[20] Poel D V D, Schamphelaere J D, Wets G. Direct and indirect effects of retail promotions on sales and profits in the do-it-yourself market[J]. Expert Systems with Applications, 2004, 27(1):53-62.

[21] Lee C H, Wang S H. An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery[J]. Expert Systems with Applications, 2012, 39(10):8954-8967.

[22] Xu Z, Liu Y, Mei L, et al. Generating temporal semantic context of concepts using web search engines[J]. Journal of Network and Computer Applications, 2014, 43: 42-55.

[23] Shi L, Setchi R. User-oriented ontology-based clustering of stored memories[J]. Expert Systems with Applications, 2012, 39(10):9730-9742.

[24] Xu Z, Wei X, Liu Y, et al. Building the search pattern of web users using conceptual semantic space model[J]. International Journal of Web and Grid Services, 2016, 12(3): 328-347.

[25] Macrae C. Early warnings, weak signals and learning from healthcare disasters.[J]. Bmj Quality & Safety, 2014, 23(6):440-445.

[26] Ponomareva, Julia V, Sokolova A. The Identification of Weak Signals and Wild Cards in Foresight Methodology: Stages and Methods[J]. Molecular Microbiology, 2015, 29(3):851–858.

[27] Lei Hou, Jianguo Liu, Xue Pan, et al. Prediction of collective opinion in consensus formation[J]. International Journal of Modern Physics C, 2014, 25(4):222-237.

[28] Avlonitis M, Karydis I, Sioutas S. Early prediction in collective intelligence on video users' activity[J]. Information Sciences, 2015, 298(C):315-329.

[29] Anauati M V, Feld B, Galiani S, et al. Collective Action: Experimental Evidence[J]. Social Science Electronic Publishing, 2016.