Clustering-Algorithm-Based Rare-Event Evolution Analysis via Social Media Data

Xiaoyu Sean Lu, *Student Member, IEEE*, MengChu Zhou, *Fellow, IEEE*, Liang Qi, *Member, IEEE*, and Haoyue Liu, *Student Member, IEEE*

***Abstract*— Exploration and discovery of the relationship between social media activities and rare-event evolution have been investigated by many researchers in recent years. Their investigations have revealed the existence of such relationship. Furthermore, some researchers regard finding either a temporal or spatial pattern of social media activities as a way to evaluate the evolution of rare event. However, most of them fail to deduce an accurate time point when a rare event highly impacts social media activities. This paper concentrates on the intensity of information volume and proposes an innovative data processing method based on clustering algorithms. The proposed method can characterize the evolution of a rare event in the real world by analyzing social media activities in the virtual world. This exploration contributes to study changes of social media activities in the time domain. A case study is based on Hurricane Sandy that occurred in 2012. Social media data collected from Twitter during its arrival time span are adopted to evaluate the feasibility and effectiveness of our proposed method. First, this paper confirms that a strong correlation between a rare event and social media activities does exist. Next, it uncovers that a time difference does exist between the real and virtual worlds. In general, this paper gives a novel idea that deduces a temporal pattern of social media activities during the occurrence of rare events.**

***Index Terms*— Clustering algorithms, data processing, rare events, social media data.**

I. Introduction

With the development of mobile technology and social platform, social media has received much attention so that people can post their messages anytime and anywhere [1]. Especially, when an event occurs, many event-related messages including human being's ideas, attitudes, and behaviors are posted. This leads a way that analyzes event-related information and data such as finding the relationship of mobility patterns and happiness [2], and connecting attractions and tourist origins [3]–[5]. As a rare event [6], [7], a disaster seriously threatens people's lives, property loss, and environment safety. The study in [8] states that a study on social behaviors and activities could be an important way to comprehend a disaster. Thus, [8] and [9] link actual disasters with social behaviors and activities together.

A strong relationship between the real and virtual worlds does exist. In other words, the former may impact the latter while the latter may be struck by the former. On the contrary, the latter one is also able to characterize the real event in the former one. Let us use a heavy storm as an example. When it passes, human beings may post and share texts, videos, and photos related to some phenomena and its damages. These messages may characterize how strong wind storm is according to human being's real feelings and attitudes. Thus, if temporal and spatial patterns can be found that a rare event will impact the social media and social media can be used to characterize the event, it helps relevant departments to evaluate and handle such event. Some studies model and analyze the temporal patterns of several activities in the social media network [10], [11]. This paper considers only one rare event that triggers and deduces the changes of the virtual world. Then, we analyze the relationship between the changes of both the virtual and real worlds.

As one of the most serious rare events, Hurricane Sandy 2012 has been analyzed by using social media data in some research studies. They are categorized as two major interests. The first category of studies investigates the changes of awareness and moods of users during the occurrence of Hurricane Sandy [12]–[18]. Some techniques, such as natural language processing and sentiment analysis, are utilized. The second category focuses on exploring the relationship between the changes of social media activities and the evolution of rare events [8], [19]–[21]. Its core purpose aims at investigating events either in the time or spatial domain. The work in [19] counts the number of photographs that are related to Hurricane Sandy and compares it with the changes of atmospheric pressure. The practical atmospheric pressure is used as the evolution of the hurricane in the real world. The relationship of linking the hurricane with social media activities is verified by using a correlation coefficient. A reliable metric is required to evaluate an event's impacts. As mentioned in [8] and [22], the way that only counts the volume of related instances is lack of meaning. Disaster-related ratio (DRR), a metric proposed in [8], replaces counting the number of messages. This metric

Manuscript received January 13, 2019; accepted January 31, 2019. Date of publication March 15, 2019; date of current version April 1, 2019. This work was supported in part by Fundo para o Desenvolvimento das Ciencias e da Tecnologia under Grant 119/2014/A3 and in part by the U.S. National Science Foundation under Grant CMMI-1162482. *(Corresponding author: MengChu Zhou.)*

X. S. Lu and H. Liu are with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: xl267@njit.edu; hl394@njit.edu).

M. Zhou is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA, and also with the Institute of Systems Engineering, Macau University of Science and Technology, Macau 999078, China (e-mail: zhou@njit.edu).

L. Qi is with the Department of Computer Science and Technology, Shandong University of Science and Technology, Qingdao 266590, China, and also with the Department of Computer Science, Tongji University, Shanghai 201804, China (e-mail: qiliangsdkd@163.com).

Digital Object Identifier 10.1109/TCSS.2019.2898774

divides the number of related instances by the total number. It can evaluate the relationship between a rare event and social media activities in the specific time span and area. If a topic is discussed many times and has a high percentage of attention among other topics, this denotes that many people pay much attention to it. Thus, DDR is more useful than only the number of disaster-related messages. The study in [8] uses DRRs to verify a strong connection between social media activities and Sandy evolution. However, in their work, only a few days' DRRs can be computed for specific cities, since they use a date as the time granularity. For different cities, only peak dates can be obtained thereby leading to high error rates. For example, the time granularity adopts an hour in [19]. In addition, no matter how the aforementioned studies decide time granularities, the time period is divided into some constant and equal time intervals. Thus, in the temporal domain, the intensity of message volume is easily ignored, since the intervals decided in advance are subjective and equal. For example, if there are a lot of posted messages around a time point, but some of them belong to an earlier time interval and the left ones belong to the latter time interval. In such cases, they are cut into two time intervals, and then the intensity of them is broken and reduced. Thus, finding proper time intervals, not fixed ones, are very important. Such intervals tend to be different. Thus, this paper uses clustering algorithms to study the intensity of message volume in the time domain. These messages can be automatically assigned into clusters, i.e., time intervals. More accurate time points can be revealed when the peaks of social media activities arrive during a rare event. Meanwhile, based on the intensity of message volume, the time intervals are selected automatically instead of fixed and equal ones. In addition, there exists a time difference that is obtained by the hurricane's occurrence and the peak intensity of message volume. In other words, it reveals the difference between the virtual world and the real one in a time domain. The contributions of this paper contain four parts. First, it verifies that the connection between the virtual world and the real one is strong. Second, by using our proposed method and finding the proper time intervals, we can deduce the temporal evolution of hurricane and obtain a more accurate impacted time point than the existing methods [8], [19]. Third, for our case, the selection of initial centers is more important than the utilization of global optimization toward the clustering algorithm-based data processing method. Finally, we confirm that the time difference does exist and varies for different cities.

In Section II, three clustering algorithms-based data processing methods are illustrated. The major part of the proposed method includes k-means, k-means++, and k-mussels wandering optimization (k-MWO) clustering algorithms. By showing and comparing the experimental results in Section III, the effectiveness of our method is verified. This paper is concluded and future directions are discussed in Section IV.

II. Data Processing Method

To discover the groupings of objects, points, or patterns naturally, clustering algorithms are widely used [23]. About 60 years ago, k-means was proposed as a basic one. Because of its high efficiency, simplicity, and easy-to-use, it works as one of the most common and useful clustering methods and is adopted in many studies successfully, such as [23]–[25]. The work [26] proposes an approach to select initial centers and extends the k-means as an upgraded algorithm called k-means++ for short. With swarm intelligence, k-MWO is a novel clustering method and proposed in [27].

In this section, our data processing method is proposed. One of the three clustering algorithms mentioned above constitutes the major part of our proposed method. Each of them is adopted individually. The detailed descriptions of the method are described next.

First, we define the related notations. $D = \{x_n\} \subset R^d$, $n \in \{1, 2, \ldots, N\}$, is a set d-dimensional points where R^d represents a data set with d-dimension and each element is a real number. N is a positive integer denoting the instance count of D. $r = (r_1, r_2, \ldots, r_N)$ is a $1 \times N$ vector, where $r_n \in \{0, 1\}$ is a binary indicator that indicates whether the nth instance is associated with a specific event. In other words, when $r_n = 1$, the instance is a rare-event-related one; otherwise, $r_n = 0$, and it is a rare-event-unrelated one. Indicated by vector r, D can be divided into two independent sets described by X_α and X_β, respectively, where X_α consists of all rare-event-related instances from D while X_β contains all rare-event-unrelated ones, i.e., remaining ones. A vector of binary variables $z_{nk} \in \{0, 1\}$ is associated with a data point x_n and describes that x_n is assigned to the kth cluster, where $k \in \{1, 2, \ldots, K\}$. Thus, if x_n is belong to kth cluster, then $z_{nj} = 1$, and $z_{nj} = 0$ when $j \neq k$. Accordingly, Z corresponds to an $N \times K$ matrix. Two additional sets of binary indicators α_{nk} and β_{nk} are defined to determine those rare-event-related and rare-event-unrelated ones. Let $A = \{\alpha_{nj}\}$ and $B = \{\beta_{nj}\}$. If x_n is indicated as a rare-event-related instance and belong to the kth cluster, $\alpha_{nj} = 1$; otherwise, $\alpha_{nj} = 0$ for $j \neq k$, where $j \in \{1, 2, \ldots, K\}$. Analogously, if a rare-event-unrelated instance x_n belongs to kth cluster, then $\beta_{nj} = 1$; otherwise, $\beta_{nj} = 0$ for $j \neq k$. After this, we update $Z = A + B$. Thus, we have the following objective function:

$$J = \sum_{k=1}^{K} \sum_{\substack{x_n \in C_k, \\ r_n = 1}} \|x_n - u_k\|^2. \tag{1}$$

It is the sum of all distances between a rare-event-related instance x_n and its center μ_k. We should determine each of z_{nk} and u_k such that J is minimized. All the rare-event-related ones are grouped and divided into K clusters. Afterward, we assign the remaining ones, i.e., rare-event-unrelated ones, into K clusters by searching the shortest distance with each cluster center. When a rare-event-related instance x_n is allocated into a cluster, α_{nk} is updated as

$$\alpha_{nk} = \begin{cases} 1, & \text{if } r_n = 1 \quad \text{and } k = \underset{j}{\operatorname{argmin}} \|x_n - u_j\|^2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Note that only rare-event-related instances are considered, and thus β_{nk} is not changed. Then, $Z = A + B$. The coordinates

of center C_k associated with cluster k is computed as

$$u_k = \frac{\sum_n r_n \alpha_{nk} x_n}{\sum_n r_n \alpha_{nk}}. \tag{3}$$

When the clustering is complete, β_{nk} is calculated and updated as

$$\beta_{nk} = \begin{cases} 1, & \text{if } r_n = 0 \quad \text{and } k = \underset{j}{\operatorname{argmin}} \|x_n - u_j\|^2 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Then, by using $Z = A + B$, Z is updated and DRR_k is calculated as

$$\mathrm{DRR}_k = \frac{\sum_n \alpha_{nk}}{\sum_n z_{nk}} \tag{5}$$

where DRR_k denotes the DRR of the kth cluster. Then, the curve of DRR versus K centers, i.e., DRR curve is generated. The steps of the proposed method are described as follows.

Initially, each instance is labeled by 0 or 1. An instance is labeled as 1 if it is related to a specific rare event; otherwise, 0. Then, vector r associated with binary values is obtained. In this paper, a keyword search method is used to distinguish instances. Thus, by searching predefined keywords, it identifies the rare-event-related and unrelated ones from the original data set. Next, clustering algorithms are used to group the rare-event-related ones.

A. k-Means and k-Means++Based Processing

The k-means-based data processing proceeds as follows.

1) Randomly generates K clusters and obtain an initial partition.
2) Assign every data point into its nearest cluster by (2).
3) Update the cluster centers by using (3).
4) Steps 2 and 3 until J reaches its minimum value.

Finally, when the clustering is complete, rare-event-unrelated ones can be divided into K new clusters via (4). Then, DRR is calculated and obtained via (5). Note that Steps 1–4 only focus on rare-event-related instances.

The k-means++-based data processing is similar to the one based on k-means. The only difference is that k-means++ chooses the initial centers with a probability [24].

B. k-MWO-Based Data Processing Method

The k-MWO-based data processing proceeds as follows:

1) Initialize N mussels, i.e., $Y_i = (y_{i1}, y_{i2}, \ldots, y_{ik}, \ldots, y_{iK})$ where $i \in \mathbf{N}_N = \{1, 2, \ldots, \mathbf{N}\}$ and $k \in N_K = \{1, 2, \ldots, K\}$.
2) Compute the fitness of each mussel by using (1). α_{nk} is calculated via (2). Note that Y_i corresponds to one set of centers $U = \{\mu_k\}$, where $k \in N_K$, in (1) and (2).
3) Obtain the best fitness and search the top $\eta\%$ mussels, and then compute the center y_g.
4) Update the position of mussels by calculating each mussel's Levy walk, and then update mussel's position by $y'_{ik} = y_{ik} + l_i(y_g - y_{ik})$.
5) Calculate the new mussels' fitness, search their top $\eta\%$ ones, and update y_g.
6) Check whether the termination criterion is reached or not. If yes, return the best one; otherwise, go back to Step 4 and continue to the next iteration.

Finally, when the clustering is complete, rare-event-unrelated ones are partitioned into K new centers via (4). Then, DRR is calculated and obtained via (5). Note that the Levy walk adopted here is $l_i = \gamma[1 - \lambda]^{-1/(\rho-1)}$, where $1.0 < \rho < 3.0$ is a shape parameter. The walk scale factor λ is a positive real number and randomly generated from the uniform distribution [0,1]. In fact, k-MWO generates some mussels and uses them as centers. Then, its evolutionary mechanism updates those mussels and searches the best centers that minimize the objective function.

C. Time Difference

The study of time difference plays a vital role in understanding and revealing the relationship between the virtual and real worlds in a time domain. It reflects the precedence order between the two worlds. Understanding the time difference is able to help broadcast warnings and predict the severity of an event in advance. Thus, the time difference is proposed and adopted to evaluate the approach regarding the hurricane in the time domain. In this paper, the time difference is defined as the time point associated with the peak of DRR curve minus the time of the arrival of hurricane. If it is a negative value, it represents that DRR reaches its peak a little late than air pressure does, namely, a lag time difference. Otherwise, it is called a lead time difference, which denotes that the DRR reaches its peak earlier than air pressure does. Meanwhile, the minimum air pressure and maximum wind speed are assumed as a sign of the hurricane's arrival. Thus, their corresponding time points denote the time of the hurricane's arrival.

D. K-Value Selection

Even though k-means, k-means++, and k-MWO are different, the selection of k value is the same. In the real world, users usually post more messages during the daytime and relatively fewer at deep night when very few activities are ongoing. The intensity of messages, thus, varies. By using clustering algorithms, the centers of clusters move toward the high intensity of messages. Hence, centers should be obtained during the daytime or at earlier night. Then, the cluster count corresponds to the number of days when the data are collected. However, because of the impact of rare events, the regularities may be broken, especially for those rare events that occur at deep night and last for a long time. Thus, determining a K-value is difficult. Yet this value should be around the number of days during which the data are collected.

TABLE I
COMPARISONS WITH METEOROLOGICAL DATA FOR WASHINGTON DC BY USING k-MEANS++ -BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.800	0.083	**-0.732**	0.002	**-0.700**	*0*	**-0.683**	*0*	**-0.665**	*0*
	var	*0*	*0*	*0*	*0*	0.001	*0*	*0*	*0*	*0*	*0*
Wind speed	avg.	0.527	0.333	0.288	0.364	0.354	0.107	0.278	0.105	**0.362**	0.001
	var	*0*	*0*	0.020	0.031	0.007	0.009	0.001	0.002	0.001	*0*

III. EXPERIMENTAL RESULTS

Experiments are described in this section to show the feasibility and effectiveness of proposed data processing methods. Also, they are compared with the real meteorological data obtained from the National Oceanic and Atmospheric Administration (NOAA). The data in the real world contain air pressure and wind speed. As mentioned above, the low air pressure and high-speed wind indicate a hurricane's arrival [19], and thus, we use them as comparison objects. Kendall's τ as a correlation coefficient is adopted to evaluate the feasibility and effectiveness of our method. The following section first introduces both the social media and meteorological data. Then, we analyze and discuss our experimental results. Since k-means++-based data processing method outperforms its peers, we focus on the results obtained by it in this section.

A. Data Set

Twitter, a popular social media platform, was created and launched in 2006. Users are allowed to write at most 280-character short texts that are called tweets for short. Three large cities in the United States are considered as our concerned regions. They help to verify the performance of our method. They are the capital of the United States, Washington DC, the global power city, New York City (NYC), and a large seaport, Baltimore. These cities contain many Twitter users and provide sufficient social media and meteorological data. The tweets are preprocessed by filtering information in both the spatial and temporal domains. In the former domain, the center of each city is represented by the weather station that is located at each city's geographical center. Buffer distances are set as 19.65, 8.72, and 7.51 km for NYC, Baltimore, and Washington DC, respectively. These regions are specified as the same as that in [8]. In the time domain, the range from October 27, 2012 to November 7, 2012 is specified as our time span. The former one is the date that was two days before the hurricane's arrival. The latter is associated with the time that was a week after the hurricane's arrival. In total, over 289 000 tweets are collected via Twitter's application programming interface (API). Each tweet contains five attributes. They are identifier, posting time, contents, and geographic coordinates. Also, the last one contains both the longitude and latitude. Then, by using keywords in the contents, if one of the predefined keywords is contained, it identifies that this tweet is about the hurricane and named as a rare-event-related tweet. The predefined keywords are as same as in [8], i.e., "Sandy," "hurricane," and "storm." In this step, around 27 000 rare-event-related tweets are extracted. In order to make the computation easily and take the time zone into consideration, the posted time of tweets is converted into seconds, and all time points are converted into Greenwich mean time. Since the specific time span starts from October 27, we set the starting time as 00:00:00, October 27. October 29 and 30 as two specific dates correspond to the dates when the hurricane landed the specific region and a day right after the landing, respectively. The two dates are distinguished by 190 800, 277 200, and 363 600 s.

B. Experimental Results

Kendall's τ measures the difference between two variables and is adopted here to evaluate the feasibility and effectiveness of proposed methods [19]. Mathematically, τ approaching -1 or $+1$ indicates that there is a strong correlation between the two variables; otherwise, the two variables have less correlation if τ is close to 0. In addition, a p-value is companying with each τ value and associated with a hypothesis testing. It indicates whether the two variables have a significant difference or not. It also means that even though a τ value is 1, if the p-value is greater than a significance level [28], we still need to accept that the two variables do not have any strong correlation. On the contrary, if a p-value is less than a significance level, the corresponding τ value is named as a satisfied τ value. Normally, the significance level is 0.05.

In our cases, depending on the posting time of all tweets, they are grouped into K classes. In the time domain, this helps to analyze the evolution of an event for each specific city. We select different K values and compare among them. Tables I–IX illustrate the average and variance of τ regarding our three specific cities by using the proposed methods. K values are chosen as 5, 10, 15, 20, and 50. For each city and each K value, each method is executed 200 times. Then, if τ value is close to -1, it represents that the experimental results have a correlation with the meteorological data. Then, in Tables I–IX, the satisfied τ values are put in a bold font. In order to keep these values simple in Tables I–IX, each value uses three decimal places. In other words, for example, if a value is 0.7001, it is written as 0.700. Note that if it is a value smaller than 0.001, it is written as 0 and put in italic font. In other words, it is very small but not a real value 0. We now show the results of experiments obtained and the air pressure. For each specific city, there exists at least one satisfied τ value that is slightly greater than or less than -0.6. Even some of them are less than -0.7. Note the τ value greater than -0.6 and less than -0.8 indicates that two variables have a

TABLE II
COMPARISONS WITH METEOROLOGICAL DATA FOR NYC BY USING k-MEANS++-BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.800	0.083	**-0.674**	0.009	**-0.735**	*0*	**-0.742**	*0*	**-0.676**	*0*
	var	*0*	*0*	*0*	*0*	0.001	*0*	0.001	*0*	*0*	*0*
Wind speed	avg.	0.316	0.633	0.163	0.584	0.260	0.215	0.296	0.104	**0.290**	0.010
	var	*0*	*0*	*0*	0.001	0.002	0.005	0.005	0.004	0.002	*0*

TABLE III
COMPARISONS WITH METEOROLOGICAL DATA FOR BALTIMORE BY USING k-MEANS++-BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.800	0.083	**-0.725**	0.003	**-0.657**	*0*	**-0.667**	*0*	**-0.597**	*0*
	var	*0*	*0*	*0*	*0*	0.001	*0*	*0*	*0*	*0*	*0*
Wind speed	avg.	0.600	0.233	0.419	0.139	**0.421**	0.046	**0.444**	0.011	**0.449**	*0*
	var	*0*	*0*	0.008	0.005	0.003	0.001	0.002	*0*	0.001	*0*

TABLE IV
COMPARISONS WITH METEOROLOGICAL DATA FOR WASHINGTON DC BY USING k-MEANS-BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.800	0.083	**-0.674**	0.007	**-0.630**	0.001	**-0.609**	*0*	**-0.540**	*0*
	var	*0*	*0*	0.001	*0*	*0*	*0*	0.001	*0*	0.001	*0*
Wind speed	avg.	0.748	0.125	0.524	0.068	**0.644**	0.004	**0.664**	*0*	**0.568**	*0*
	var	0.010	0.007	0.011	0.005	0.009	*0*	0.005	*0*	0.001	*0*

TABLE V
COMPARISONS WITH METEOROLOGICAL DATA FOR NYC BY USING k-MEANS-BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.800	0.083	**-0.688**	0.007	**-0.685**	*0*	**-0.695**	*0*	**-0.658**	*0*
	var	*0*	*0*	0.005	*0*	0.002	*0*	0.002	*0*	0.001	*0*
Wind speed	avg.	0.280	0.646	0.280	0.646	**0.445**	0.036	**0.498**	0.012	**0.506**	*0*
	var	0.017	0.002	0.017	0.002	0.004	0.001	0.009	0.001	0.002	*0*

TABLE VI
COMPARISONS WITH METEOROLOGICAL DATA FOR BALTIMORE BY USING k-MEANS-BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.776	0.101	**-0.644**	0.010	**-0.580**	0.002	**-0.600**	*0*	**-0.464**	*0*
	var	0.004	0.002	*0*	*0*	*0*	*0*	*0*	*0*	0.001	*0*
Wind speed	avg.	0.464	0.403	0.485	0.078	**0.457**	0.029	**0.536**	0.002	**0.446**	*0*
	var	0.009	0.014	0.005	0.006	0.004	0.001	0.001	*0*	0.002	*0*

TABLE VII
COMPARISONS WITH METEOROLOGICAL DATA FOR WASHINGTON DC BY USING k-MWO-BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.651	0.195	**-0.675**	0.008	**-0.674**	*0*	**-0.658**	*0*	**-0.579**	*0*
	var	0.008	0.004	0.002	*0*	0.001	*0*	*0*	*0*	0.001	*0*
Wind speed	avg.	0.458	0.416	0.334	0.271	0.387	0.095	0.341	0.067	**0.406**	*0*
	var	0.012	0.016	0.016	0.034	0.012	0.011	0.005	0.005	0.002	*0*

moderate correlation. For each specific city and each clustering algorithm, all highest τ values are obtained when K equals 10, 15, or 20. For each city, among three clustering algorithm-based methods, the best satisfied τ values are obtained by the k-means++-based method because its best τ values are less than those of other two methods. The k-MWO-based and

TABLE VIII
COMPARISONS WITH METEOROLOGICAL DATA FOR NYC BY USING k-MWO-BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.597	0.240	**-0.551**	0.036	**-0.588**	0.003	**-0.583**	*0*	**-0.527**	*0*
	var	0.003	0.003	0.004	0.001	0.002	*0*	0.001	*0*	0.001	*0*
Wind speed	avg.	0.257	0.778	0.462	0.104	**0.425**	0.049	**0.449**	0.016	**0.517**	*0*
	var	0.058	0.114	0.009	0.008	0.005	0.002	0.005	0.001	0.002	*0*

TABLE IX
COMPARISONS WITH METEOROLOGICAL DATA FOR BALTIMORE BY USING k-MWO-BASED METHOD

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.792	0.089	**-0.709**	0.005	**-0.695**	*0*	**-0.659**	*0*	**-0.583**	*0*
	var	0.002	0.001	0.001	*0*	*0*	*0*	*0*	*0*	*0*	*0*
Wind speed	avg.	0.580	0.261	0.406	0.152	**0.518**	0.010	**0.456**	0.009	**0.426**	*0*
	var	0.006	0.013	0.007	0.007	*0*	*0*	0.002	*0*	0.002	*0*

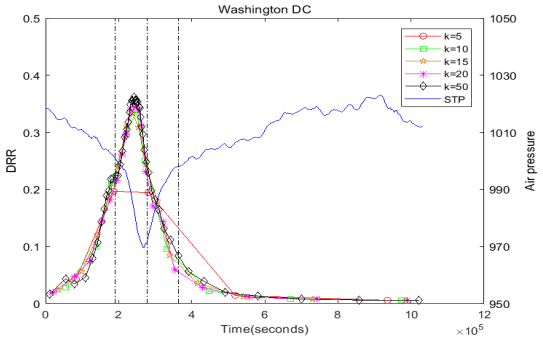


Fig. 1. Curve of DRR versus air pressure in Washington DC.

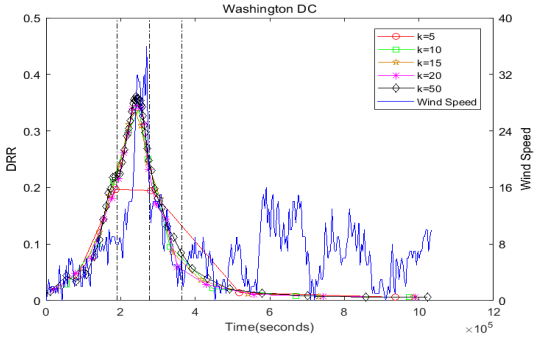


Fig. 2. Curve of DRR versus wind speed in Washington DC.

k-means-based methods obtain roughly the same results. In other words, in some cases, the k-MWO-based method performs better than k-means-based one, but in other cases, the k-MWO-based method performs worse than k-means-based one. If we focus on the comparisons with the variant of wind speed, only a few of τ values are satisfied with the constraint that p-value is less than 0.05 and they are much less than the τ values obtained by air pressure. It concludes that the social media data from the virtual world have a relationship with meteorological data in the real world. It clearly means that the social media activities are associated with the disaster in the real world. The k-means++-based method is the best. Due to the wind speed changes uncertainly and randomly, the relative stable variant of air pressure is better than wind speed for the comparisons between the virtual world and the real one.

The curve of DRR for each city is similar. Washington DC is shown here and regarded as an example. Figs. 1 and 2 show its curves of DRR by using k-means-based method versus air pressure and wind speed, respectively. For other cities and methods, Figs. 1 and 2 are similar. Three black dotted vertical lines distinguish October 29 and 30 as two specific dates. Fig. 1 shows that the air pressure reaches its peak, i.e., the minimum value, on October 29. Note that the hurricane touched these cities on that day. Furthermore, around the time when the hurricane strikes the cities, the air pressure decreases sharply and then increases. After a short period, the air pressure restores to a normal status gradually. Its tendency of variation is similar to the curve of DRR we obtained. In Fig. 1, first off, the curve of DRR increases sharply, but then gradually decreases and approaches to be 0 in a few days after the arrival of hurricane. Since many factors, such as the angle of wind, that impact the measurement of wind speed, the speed is changed more frequently and sharply than the air pressure does. Fig. 2 shows that the wind speed is changed sharply. The maximum value of wind speed is found on October 29. At the same time, the maximum DRR values appear on the same day as well. Clearly, the curves of DRRs and wind speed have a very similar tendency. In other words, both the wind speed and the DRR grow from a low value to its peak sharply and drop back to a low one gradually.

If we have a close view of Fig. 1, we discover that a short-time difference exists between the peak of DRR and the peak of air pressure. Since many rare-event-related tweets were posted a little earlier than the arrival of hurricane, the short-time difference is supposed to be derived. Tables X–XVIII concern the time differences and then do the comparisons among our meteorological data and experimental results. As we did earlier, if a τ value is a satisfied one in

TABLE X

COMPARISONS WITH METEOROLOGICAL DATA FOR WASHINGTON DC BY USING k-MEANS++BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.600	0.233	**-0.827**	*0*	**-0.810**	*0*	**-0.785**	*0*	**-0.767**	*0*
	var	*0*	*0*	*0*	*0*	0.002	*0*	*0*	*0*	*0*	*0*
	Time Difference	18428 s		**34477 s**		**24934 s**		**23197 s**		**27764 s**	

TABLE XI

COMPARISONS WITH METEOROLOGICAL DATA FOR NYC BY USING k-MEANS++-BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.600	0.233	**-0.688**	0.005	**-0.769**	*0*	**-0.745**	*0*	**-0.684**	*0*
	var	*0*	*0*	*0*	*0*	0.001	*0*	0.001	*0*	*0*	*0*
	Time Difference	-26504 s		**-2988 s**		**12575 s**		**3421 s**		**2820 s**	

TABLE XII

COMPARISONS WITH METEOROLOGICAL DATA FOR BALTIMORE BY USING k-MEANS++-BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.601	0.233	**-0.824**	0.001	**-0.758**	*0*	**-0.747**	*0*	**-0.661**	*0*
	var	*0*	*0*	0.002	*0*	*0*	*0*	0.001	*0*	0.001	*0*
	Time Difference	15900 s		**31602 s**		**20780 s**		**22456 s**		**15653 s**	

TABLE XIII

COMPARISONS WITH METEOROLOGICAL DATA FOR WASHINGTON DC BY USING k-MEANS-BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.800	0.083	**-0.830**	*0*	**-0.832**	*0*	**-0.791**	*0*	**-0.722**	*0*
	var	*0*	*0*	*0*	*0*	*0*	*0*	0.001	*0*	0.003	*0*
	Time Difference	8121 s		**24011 s**		**28305 s**		**26316 s**		**25646 s**	

TABLE XIV

COMPARISONS WITH METEOROLOGICAL DATA FOR NYC BY USING k-MEANS-BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.614	0.223	**-0.732**	0.003	**-0.704**	*0*	**-0.704**	*0*	**-0.675**	*0*
	var	0.003	0.001	0.003	*0*	0.002	*0*	0.002	*0*	0.001	*0*
	Time Difference	-22051 s		**8356 s**		**3244 s**		**1855 s**		**3889 s**	

TABLE XV

COMPARISONS WITH METEOROLOGICAL DATA FOR BALTIMORE BY USING k-MEANS-BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.800	0.083	**-0.738**	0.003	**-0.729**	*0*	**-0.685**	*0*	**-0.574**	*0*
	var	*0*	*0*	0.004	*0*	0.004	*0*	0.003	*0*	0.004	*0*
	Time Difference	15142 s		**19157 s**		**24410 s**		**15264 s**		**16002 s**	

Tables X–XVIII and is less than the corresponding value in Tables I–IX, they are put in a bold font. Let us use the comparisons in Tables I and X for Washington DC as an example. When $K = 15$ and the air pressure data are compared in Table I, τ is -0.7, and in Table X and it is -0.810 that is less than -0.7. At the same time, the p-value of τ, -0.810, is less than 0.05. Then, in Table X, -0.810 is put in a bold font.

As the time difference defined in Section II-C, it is acceptable only when the p-value is lower than 0.05. The time differences put in a bold font denote the acceptable ones in Tables X–XVIII. Note that only air pressure is paid attention by considering the unreliability and uncertainties of the wind speed. In Tables X–XVIII, most of the time differences are put in bold. It represents that most of the time differences are acceptable and most of their corresponding τ values are less

TABLE XVI
COMPARISONS WITH METEOROLOGICAL DATA FOR WASHINGTON DC BY USING k-MWO-BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.657	0.191	**-0.787**	0.001	**-0.735**	*0*	**-0.737**	*0*	**-0.681**	*0*
	var	0.008	0.005	0.001	*0*	0.001	*0*	0.001	*0*	0.003	*0*
	Time Difference	22592 s		**34622 s**		**21917 s**		**22017 s**		**25935 s**	

TABLE XVII
COMPARISONS WITH METEOROLOGICAL DATA FOR NYC BY USING k-MWO-BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.613	0.224	**-0.642**	0.012	**-0.605**	0.002	**-0.598**	*0*	**-0.552**	*0*
	var	0.002	0.001	0.002	*0*	0.002	*0*	0.003	*0*	0.007	*0*
	Time Difference	2729 s		**14580 s**		**2480 s**		**2359 s**		**7264 s**	

TABLE XVIII
COMPARISONS WITH METEOROLOGICAL DATA FOR BALTIMORE BY USING k-MWO-BASED METHOD AND CONSIDERING TIME DIFFERENCE

	K	5		10		15		20		50	
		τ	p-value	τ	p-value	τ	p-value	τ	p-value	τ	p-value
Air pressure	avg.	-0.659	0.192	**-0.792**	0.001	**-0.714**	*0*	**-0.730**	*0*	**-0.656**	*0*
	var	0.011	0.005	0.001	*0*	*0*	*0*	0.001	*0*	0.003	*0*
	Time Difference	18500 s		**31119 s**		**11883 s**		**18449 s**		**19100 s**	

than -0.7. In other words, the τ values become low values when the time differences are considered. All satisfied τ values are increased when the time differences are considered. Then, we conclude that the time difference does exist since it is able to increase the τ values. In other words, when the time difference is concerned, the relationship between the virtual world and the real one is much more correlated.

For the three cities, all satisfied time differences are greater than 0. It also represents that the time differences are lead time differences. The only difference among the three cities is that their lead time is different. The lead time of Baltimore and Washington DC is much greater than that of New York City. As studied in [29], the people located at different communities should have different responses. Let us use Indian Ocean Tsunami in 2004 as an example. Due to the tragic memory during that time, South Asia is more sensitive to tsunami than other continents. Therefore, for each specific city, the records of hurricanes in history are investigated. Because Baltimore is located at Maryland and Washington DC is adjacent to Maryland, the records of the region containing both the Maryland and Washington DC are used. Two links, [35] and [36], from Wikipedia uncover the records of hurricanes in history. Since the year of 1950, the region of Washington DC and Maryland has already been attacked by hurricanes for 111 times with 12 of them defined as deadly storms. New York State has been impacted for 61 times since the year of 1950 and 13 of them were concerned as deadly storms. For both the NY and the region of Washington DC and Maryland, the number of deadly storms is similar. However, in the region of Washington DC and Maryland, the number of deaths is 61 during the deadly storms. This number in New York State is 107. The death count in NY is 1.75 times more than that in the region of Maryland and Washington DC, but the population of New York State is triple more than that in the latter. Also, the number of hurricanes in the region of Maryland and Washington DC is 1.82 times more than that in NY. From this perspective, we conclude that the region of Washington DC and Maryland is more sensitive to hurricanes than NY. It well explains the reason that both the Washington DC and Baltimore have longer lead time differences than NY has. It is because that the residents are more sensitive to and need more time in advance to cope with hurricanes. Thus, during the hurricane, residents in the region of Washington DC and Baltimore intend to post more rare-event-related tweets or alerts much earlier than the hurricane's arrival. Meanwhile, there is a short lead time difference in New York City. Our experimental results reveal that the time differences between the virtual world and the real one definitely exist. Three clustering algorithms adopted in our work can obtain the same conclusions and results.

C. Comparisons and Impact of k

Because of uncertainty and rapid changes of wind speed, only the comparisons with the air pressure are discussed in this section. Tables I–IX reflect the correlation that is computed without the concern of time difference by using three clustering algorithm-based methods. Overall, the best τ values among the three cities are obtained via k-means++. The three best τ values for Washington DC, NYC, and Baltimore are given when $K = 10$, 20, and 10, respectively. No matter which method is used, the best τ values are obtained when $K = 10$, 15, or 20 for three cities. Thus, the proper range of K is from 10 to 20. In addition, when selecting K in this

range, the τ values change slightly only. However, the cases with $K = 5$ or 50 result in the unsatisfied τ values for each city and method. This implies that the too few or many clusters cannot lead to acceptable results for the problem in this paper.

Tables X–XVIII illustrate the correlation that is computed with the consideration of time difference. Overall, the best τ values for NYC and Baltimore are obtained via k-means++. That for Washington DC is given by using k-means, but the k-means++-based method only gives a slightly greater τ value than k-means-based one. Three best τ values for Washington DC, NYC, and Baltimore are given when $K = 15$, 15, and 10, respectively. It reflects that the proper range of K from 10 to 20 is acceptable with the consideration of time difference. In addition, in this case, no matter which city and method are concerned, $K = 10$ or 15 for the best τ values. Meanwhile, $K = 5$ gives the unsatisfied τ values and $K = 50$ yields the worst values for each city and method. The k-means++-based data processing method performs well since its τ values reach the smallest for most cases. The k-means-based method only has the best τ value for Washington DC. Furthermore, this best τ value is just slightly less than the τ value obtained via the k-means++-based method. The k-MWO-based method cannot reach the best τ value, and thus it is not good enough. In conclusion, the experimental results suggest that the number of clusters should be selected around the number of days during which data are collected. The proposed k-means++-based data processing method performs the best among three.

D. Discussion Among Adopted Clustering Algorithms

The three used clustering algorithms have some differences. First of all, k-means is the basic one. In general, it randomly selects K initial points as centers, and then it stops when the objective function reaches the local or global minimum. Initially, k-MWO randomly selects centers as k-means does. However, the former utilizes the global optimization ability of MWO and combines with k-means. That is the reason that k-MWO performs slightly better than k-means. The initial points selected by k-means++ differ from the previous two algorithms. It can start from better initial centers. We reveal that k-means++ is superior to its two peers. In addition, the intensity of posted tweets should be high in the daytime, especially at noon or afternoon, and low at night. Thus, starting from proper initial centers is more important in its performance. Furthermore, we study the posted time of tweets and cluster them in the time domain. It means that the data are clustered at a low dimension. Thus, we obtain a significant result that even though k-MWO combines the ability of global optimization and local search, it does not have superiority over k-means, implying that local search is suitable for our case. As a result, the selection of proper initial centers is more important than others like global optimization. k-means++ performs the best among three algorithms.

IV. Conclusion

A reliable and robustness temporal-spatial pattern of social media activities can reflect real impact that people are suffered and be used to evaluate impact during the arrival of a rare event. Regularities between virtual and real worlds are explored in this paper. By using the proposed clustering-algorithm-based data processing methods and analyzing the social media data in the virtual world, more precise and accurate temporal information can be obtained regarding a rare event. First, we verify that there is a strong connection between the variations of social media activities and the evolution of a rare event in a time domain. Second, it provides a more precise and believable impacted time point of a rare event. Furthermore, we reveal that time differences exist and are different for different cities. Investigating and revealing the differences are helpful to build the temporal pattern of an virtual world or social media activities during the occurrence of an event. Since social media activities are timely information, they can accurately reflect the human's behaviors, mood, and awareness in real time. The study of time differences is one important component of temporal patterns. It provides an approach to track, understand, analyze, and evaluate the evolution of a rare event precisely and rapidly in a time domain. Then, relevant departments and organizations, even individuals can start to better prepare for some extreme events in advance.

In this paper, we only deal with an event that lasts a relatively long time such as storms. A very short-time event, such as an earthquake, that occurs suddenly without any warning is not considered here. In the future, we plan to concern multiple types of rare events. Advanced clustering methods [26], [30] can be used to improve the performance of data processing. Outlier detection can be used to filter outliers and parameter-free clustering algorithms are able to group those tweets automatically. The way that distinguishes the rare-event-related tweets may be improved by some advanced methods [31] and [32]. Natural language processing, sentiment and semantic analysis, and machine learning should be adopted to mine and purify more useful and valuable contents. Also, using such data mining techniques in [33] and [34] to estimate the impacts of a rare event in advance is desired. Moreover, by analyzing semantic information and understanding the temporal evolution, we may evaluate the impacted time and predict the severity degree of a subsequent rare event ahead.

References

[1] R. Narayanam and Y. Narahari, "A Shapley value-based approach to discover influential nodes in social networks," *IEEE Trans. Autom. Sci. Eng.*, vol. 8, no. 1, pp. 130–147, Jan. 2011.

[2] M. R. Frank, L. Mitchell, P. S. Dodds, and C. M. Danforth, "Happiness and the patterns of life: A study of geolocated tweets," *Sci. Rep.*, vol. 3, Sep. 2013, Art. no. 2625.

[3] S. A. Wood, A. D. Guerry, J. M. Silver, and M. Lacayo, "Using social media to quantify nature-based tourism and recreation," *Sci. Rep.*, vol. 3, Oct. 2013, Art. no. 2976.

[4] W. Luan, G. Liu, and C. Jiang, "Collaborative tensor factorization and its application in POI recommendation," in *Proc. IEEE 13th Int. Conf. Netw., Sens., Control*, Mexico City, Mexico, Apr. 2016, pp. 1–6.

[5] E. Aslanian, M. Radmanesh, and M. Jalili, "Hybrid recommender systems based on content feature relationship," *IEEE Trans. Ind. Informat.*, 2016.

[6] E. L. Quarantelli and R. R. Dynes, "Response to social crisis and disaster," *Annu. Rev. Sociol.*, vol. 3, no. 1, pp. 23–49, 1977.

[7] R. K. Merton and R. A. Nisbet, Eds., "Disasters," in *Contemporary Social Problems*. Berkeley, CA, USA: Univ. California Press, 1961, pp. 97–122.

[8] X. Guan and C. Chen, "Using social media data to understand and assess disasters," *Natural Hazards*, vol. 74, pp. 837–850, Nov. 2014.
[9] C. Chen, D. Neal, and M. Zhou, "Understanding the evolution of a disaster—A framework for assessing crisis in a system environment (FACSE)," *Natural Hazards*, vol. 65, pp. 407–422, Jan. 2013.
[10] E. Stai, E. Milaiou, V. Karyotis, and S. Papavassiliou, "Temporal dynamics of information diffusion in twitter: Modeling and experimentation," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 256–264, Mar. 2018.
[11] V. Raghavan, G. Ver Steeg, A. Galstyan, and A. G. Tartakovsky, "Modeling temporal activity patterns in dynamic social networks," *IEEE Trans. Comput. Social Syst.*, vol. 1, no. 1, pp. 89–107, Mar. 2013.
[12] Y. Huang, H. Dong, Y. Yesha, and S. Zhou, "A scalable system for community discovery in Twitter during hurricane sandy," in *Proc. 14th IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput. (CCGrid)*, May 2014, pp. 893–899.
[13] H. Dong, M. Halem, and S. Zhou, "Social media data analytics applied to hurricane sandy," in *Proc. Int. Conf. Social Comput. (SocialCom)*, 2013, pp. 963–966.
[14] C. Caragea, A. Squicciarini, S. Stehle, K. Neppalli, and A. Tapia, "Mapping moods: Geo-mapped sentiment analysis during hurricane Sandy," in *Proc. ISCRAM*, 2014, pp. 354–358.
[15] D. Ediger, S. Appling, E. Briscoe, R. McColl, and J. Poovey, "Real-time streaming intelligence: Integrating graph and NLP analytics," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Sep. 2014, pp. 1–6.
[16] T. Lansdall-Welfare, S. Sudhahar, G. A. Veltri, and N. Cristianini, "On the coverage of science in the media: A big data study on the impact of the Fukushima disaster," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2014, pp. 60–66.
[17] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena, "Statistically significant detection of linguistic change," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 625–635.
[18] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter," *PLoS ONE*, vol. 6, no. 12, 2011, Art. no. e26752.
[19] T. Preis, H. S. Moat, S. R. Bishop, P. Treleaven, and H. E. Stanley, "Quantifying the digital traces of hurricane Sandy on Flickr," *Sci. Rep.*, vol. 3, Nov. 2013, Art. no. 3141.
[20] X. S. Lu and M. Zhou, "Analyzing the evolution of rare events via social media data and k-means clustering algorithm," in *Proc. IEEE 13th Int. Conf. Netw., Sens., Control*, Mexico City, Mexico, Apr. 2016, pp. 1–6.
[21] K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, "Characterizing and countering communal microblogs during disaster events," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 2, pp. 403–417, Jan. 2018.
[22] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Multi-resolution spatial event forecasting in social media," in *Proc. IEEE 16th Int. Conf. Data Mining*, Dec. 2016, pp. 689–698.
[23] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
[24] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, Philadelphia, PA, USA, 2007, pp. 1027–1035.
[25] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, 1988.
[26] J. Hou and W. Liu, "A parameter-independent clustering framework," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1825–1832, Aug. 2017.
[27] Q. Kang, S. Liu, M. C. Zhou, and S. Li, "A weight-incorporated similarity-based clustering ensemble method based on swarm intelligence," *Knowl.-Based Syst.*, vol. 104, pp. 156–164, Jul. 2016.
[28] H. Abdi, "The Kendall rank correlation coefficient," in *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA USA: Sage, 2007, pp. 508–510.
[29] C. Gao and J. Liu, "Network-based modeling for characterizing human collective behaviors during extreme events," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 1, pp. 171–183, Jan. 2017.
[30] Z. Zhou, M. Sellami, W. Gaaloul, M. Barhamgi, and B. Defude, "Data providing services clustering and management for facilitating service discovery and replacement," *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 4, pp. 1131–1146, Oct. 2013.
[31] K. Y. Wu, M. Zhou, X. S. Lu, and L. Huang, "A fuzzy logic-based text classification method for social media data," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Banff, AB, Canada, Oct. 2017, pp. 1942–1947.
[32] Z. Xia, X. Wang, X. Sun, and Q. Wang, "A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 2, pp. 340–352, Jan. 2016.
[33] T. T. Aye *et al.*, "Layman analytics system: A cloud-enabled system for data analytics workflow recommendation," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 1, pp. 160–170, Jan. 2017.
[34] Y. Ni, Y. Fan, W. Tan, K. Huang, and J. Bi, "NCSR: Negative-connection-aware service recommendation for large sparse service network," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 2, pp. 579–590, Apr. 2016.
[35] *List of New York Hurricanes*. Accessed: 2018. [Online]. Available: https://en.wikipedia.org/wiki/List_of_New_York_hurricanes
[36] *List of Maryland Hurricanes*. Accessed: 2018. [Online]. Available: https://en.wikipedia.org/wiki/List_of_Maryland_hurricanes_(1950%E2%80%93present)

Xiaoyu Sean Lu (S'14) received the B.S. degree from the Nanjing University of Technology, Nanjing, China, in 2011, and the M.S. degree from the New Jersey Institute of Technology, Newark, NJ, USA, in 2015, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering.

His interests include data processing, data mining, machine learning, natural language processing, semantic and sentiment analysis, power systems, power management, and smart grids.

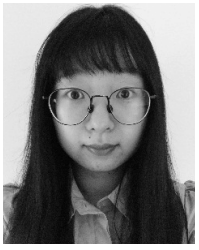
MengChu Zhou (S'88–M'90–SM'93–F'03) received the B.S. degree in control engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1983, the M.S. degree in automatic control from the Beijing Institute of Technology, Beijing, China, in 1986, and the Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He joined the New Jersey Institute of Technology, Newark, NJ, USA, in 1990, where he is currently a Distinguished Professor of Electrical and Computer Engineering Department. He has authored or co-authored more than 800 publications including 12 books, more than 460 journal papers, and 28 book chapters. He holds 12 patents. His interests include Petri nets, intelligent automation, Internet of Things, big data, web services, and intelligent transportation.

Dr. Zhou is a fellow of the International Federation of Automatic Control, the American Association for the Advancement of Science, and the Chinese Association of Automation (CAA).

Liang Qi (S'16–M'18) received the B.S. degree in information and computing science and the M.S. degree in computer software and theory from the Shandong University of Science and Technology, Qingdao, China, in 2009 and 2012, respectively, and the Ph.D. degree in computer software and theory from Tongji University, Shanghai, China, in 2017.

From 2015 to 2017, he was a Visiting Student with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ, USA. He is currently with Shandong University of Science and Technology, Qingdao, China. His interests include machine learning, intelligent transportation systems, and optimization.



Haoyue Liu (S'17) received the B.S. degree from the Kunming University of Science and Technology, Kunming, China, in 2014, and the M.S. degree from the New Jersey Institute of Technology, Newark, NJ, USA, in 2016, where she is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering.

Her current research interests include machine learning, natural language processing, sentiment analysis, and big data analytics.