Analyzing the Spatiotemporal Effects on Detection of Rain Event Duration

Jyun-Yu Jiang, Yi-Shiang Tzeng, Pei-Ying Huang, and Pu-Jen Cheng

Department of Computer Science and Information Engineering
National Taiwan University, Taiwan
{b98902114,r00944020}@ntu.edu.tw, {d96004,pjcheng}@csie.ntu.edu.tw

Abstract. There has been significant recent interest in using the aggregate information from social media sites to detect and predict real-world phenomena. Temporal and geographic effects are often considered as two possible impact factors on detection of rain event from microblog data. However, the actual contribution of them to rain event detection has yet to be defined. To investigate this issue, one method considering overall effects of time and geography is proposed for detecting the rain event. Our analysis implies that the way people post tweets changes dynamically during a day. The number of tweets grows from the early morning and peak at midnight. Besides, distribution of the population and user responses to the rain event are both not the same in different regions. Our findings therefore suggest that temporal and geographic effects may play an important role in the detection of rain event. We also apply our strategy to forecast the rain events. Our results show that our strategy performs well both in detecting and predicting events of rain. Comparative analysis with existing methods is also presented to demonstrate the effectiveness of our method. Our proposed scheme is therefore practical and feasible to be deployed in the real world.

Keywords: Event detection, Event duration, Microblog, Temporal, Geographical.

1 Introduction

When people plan to travel to somewhere, the first thing comes up to their mind might be how the weather is there (e.g., Is it raining now? Or will it rain in the near future?). The weather forecast for a city is summarized from different regions which belong to this city. It cannot reflect the regional variation individually. For example, in a city, it might rain in some places while it might not rain in other places. Or some places may be located at the boundary between two cities. However, most weather stations only provide weather forecast for cities or famous scenic spots. There are no weather data for local areas which lie in the city. It is therefore infeasible for people to obtain exact weather conditions for local regions.

That is now changing. Data from increasingly popular online social network sites (e.g., Twitter) allow us to study the detection of rain event duration in real

Y. Hou et al. (Eds.): AIRS 2012, LNCS 7675, pp. 506–517, 2012.

time in a way that is both fine-grained and massively global in scale [14]. Twitter, as a popular microblogging service, has become a new information channel for users to receive and share information. As of March 2012, there are more than 140 million active users and over 340 million tweets are created and redistributed a day[10]. Compared to news or blogs articles, Twitter messages(tweets) have 140-character-message limit in length, resulting in a short and ungrammatical textual feature. Twitter users tend to use abbreviations and acronyms (e.g., IC refers to I see, BTW refers to By the way). To detect the rain events by using tweets collected from Twitter is therefore a real challenge due to the heterogeneous and noisy nature of the data. On the contrary, such limitation enables users to update information instantly. With the popularity of mobile device, Twitter users worldwide act as a group of sensors, forming a social sensor network to share what is happening around (e.g., tsunami, rain, personal status), making it possible to real-time report the rain event which happened anywhere at any time.

However, except the limitations of Twitter in nature mentioned above, the rain event detection might still be potentially influenced by many external factors (e.g., geo-location, time, human behavior). By considering the time and location information, we can detect target events and estimate location of target events. As a tweet is often associated with a post time and a geo-location, we can detect when and where a rain event happens. For example, a user might make a tweet such as "Now it is raining" at 7:13pm on December 24. Consequently, if a rain event happens in an oceanic area, it is more difficult to locate it precisely from tweets. It also becomes more difficult to make good estimation in less populated areas. These two cases imply that practicability of detection of rain events mainly relies on the number and spatial dispersion of Twitter users. The way people post tweets changes dynamically during a day. The number of tweets grows from the early morning and peak at midnight. It should be noted that tweets around the time and geographically close to such areas would be considered alternatively as approximate indicators to detect rain events happened in such queried areas at a given time.

As mentioned above, we conjecture that the spatial and temporal features may play the important role in rain event detection. However, to the best of our knowledge, there are no previous research studies targeted to a spatiotemporal issue in detection of rain events. Thus, it remains unclear to what extent and in what way the effects of time and geography would be imposed on the detection of rain event. In this paper, we therefore focus on understanding the influence of time and geographic on the detection of rain event by aggregating Twitter messages (tweets). Our study provides clear evidences that the spatiotemporal feature is an essential factor in detection of rain event duration.

2 Related Work

In the literature, several approaches are proposed to detect events. Allan [1] studied the topic detection and tracking from documents. Allan et al. [3] and

Yang et al. [15] also use documents to do the on-line event detection. They calculate similarity among the existing documents and new coming one by the incremental clustering to determine a new generated event. Besides, there are many previous works studied the event detection on web pages. However, their proposed methods might not be applicable to deal with the detection of the rain events from microblogs. Microblogs are often shorter and more frequent updated than documents and blogs. Therefore, conventional statistic-based term weighting strategies like frequency might not reliable in analyzing microblog contents.

Some researches studied the event detection problem from the view of signal by analyzing the frequency of the time series data. Chen et al. [5] and Jianshu et al. [9] consider words on Twitter and tags of photos on Flickr as the energy, the events are then detected by analyzing the energy distribution with wavelet transform. He et al. [7] analyzed words in both time and frequency domain with Fourier transform. Such methods can detect well while signal alters suddenly. However, the rain events usually have the duration such that we need not only detect the beginning but the ending whose variation is not obvious. Cataldi et al. [4] introduced the aging theory to emerged terms and group them into some topics. Sakaki et al. [12] studied whether users observe the effect of earthquake and locate it in a probability-based approach. The same idea can also be applied to other short length document. Teevan et al. [13] traced the trajectory of storm by the query logs of a search engine.

Various studies have been made of the analysis of microblogging data (e.g., Twitter) from spatial and temporal perspectives. Sakaki et al. [12] introduced a concept of social sensors to detect earthquake event in real time. They examined the time-series data to create a temporal model to calculate the probability of an event occurrence. Spatial models such as Kalman filtering and particle filtering are then proposed to estimate the locations of events. Java et al. [8] considered both geographical and topological properties from twitters to analyze the distribution of users and their tweets. MacEachren et al. [11] used the geographical information of tweets to visualize the location and content of tweets.

Even different from general events, weather events are usually related to regional and local information [13]. Cox and Plale [6] used the Twitter data to improve the weather observation. Sakaki et al. [12] determined whether a user observes the earthquake. But the weather events are always with the duration such that we cannot detect only the happening of an event like first story in topic detection and tracking problem (TDT problem) [2] but all process of the event. Unlike the traditional TDT problem, the duration of rain events are usually shorter. The strength of signal is also weaker than usual events in the later stage in the duration.

3 Modeling Signals

Our goal is using the twitter data to determine whether it rains or not for anywhere and anytime. For any locations we are interested in their weather

conditions, we collect tweets nearby them and use corresponding tweets to create their own signals so as to detect the rain events. In the rest of this paper, we use L_j and T_i to represent locations and tweets, respectively.

3.1 Uniform Weighted Signal

As shown in Figure 1, we segment time line into equal size. The value in each time slot is the summation of $score(T_i)$, where T_i is tweets posted in the period and $score(T_i)$ is defined as follows:

$$socre(T_i) = 1 \ if \ T_i \ is \ related \ to \ rain, \ otherwise \ 0$$

The score function preserves all rain related tweets and views them identical. In the implementation, we construct a classifier trained by support vector machine (SVM) to filter out the tweets not talking about rain. The signal in Figure 1 now reflects the probability of the whether it rains in L_j in each time slot to a certain extent. The higher the value in a time slot, the higher the chance that it rains in L_j at that time.

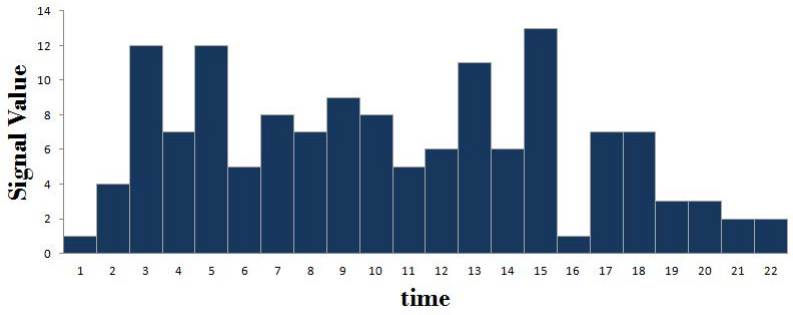

Fig. 1. Signal example

3.2 Temporal and Geographical Weighted Signal

We furthermore take time and geographic factors into consideration.

Temporal Aspect. For the same location but different time, we observe that the number of rain-related tweets in a day changes significantly (Figure 2). Many twitter users like to post tweets in the night and the number of tweets drops dramatically in the early morning. The phenomenon probably makes detecting rain event in those "inactive" period hard. Moreover, the durations of rain events are often less than a day, also highlighting the importance of daily dynamics.

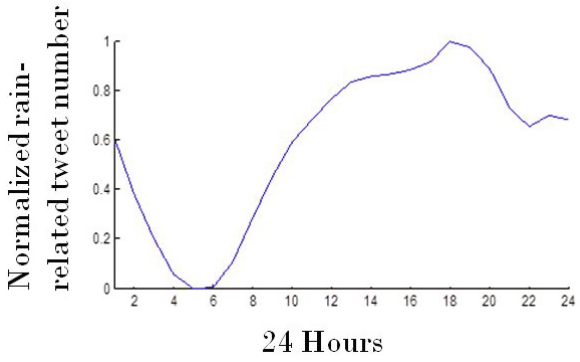

Fig. 2. Distribution of rain-related tweets in a day

Geographical Aspect. For the same time but different locations, we also find the similar situation. Figure 3 is the sampling tweets distribution of several locations. In this example, we find the distributions in these locations are not uniform. Tweets in some sub-regions are quite dense while some are very sparse. The serious unbalance may make the outcome be dominated by dense sub-regions and thus can't determine the weather condition objectively.

We also observe an interesting phenomenon that the degree of people's interest in rain($RID(L_j)$) varies in different regions. Before further discussing the relation between them, we need to define how to measure the "degree of interest" first. For each location L_j, we calculate the number of rain related tweets, divided by the number of the tweets in each raining time slot, and then set $RID(L_j)$ by averaging the values over all of the raining time slots. The measurement assumes users post more rain related tweets if they are more interested in rain events. Using normalization instead of frequency is more proper to measure $RID(L_j)$ since the population varies among regions. Figure 4 shows the relation between $RID(L_j)$ and their corresponding raining frequency. After taking logarithm, we find they have negative correlation with R^2 =0.73. Surprisingly, it quite fits power law. The finding suggests if it seldom rains in a region, the rain events tend to catch one's eye. Thus, more information about weather condition will be shared on microblogging. The experiment demonstrates how geographical factor influences the signals in Figure 1 again.

Weighted Score Function. All of above observations suggest temporal and geographical factors affect signals, we therefore tune the score function $score(T_i)$ to reflect their influences. The updated score function is named weighted score, donated as $WScore(T_i)$, and defined as follow:

$$WScore(T_i) = Tmp(T_i) \times Geo(T_i) \times score(T_i),$$

where $Tmp(T_i)$ and $Geo(T_i)$ are the temporal and geographical weights, respectively.

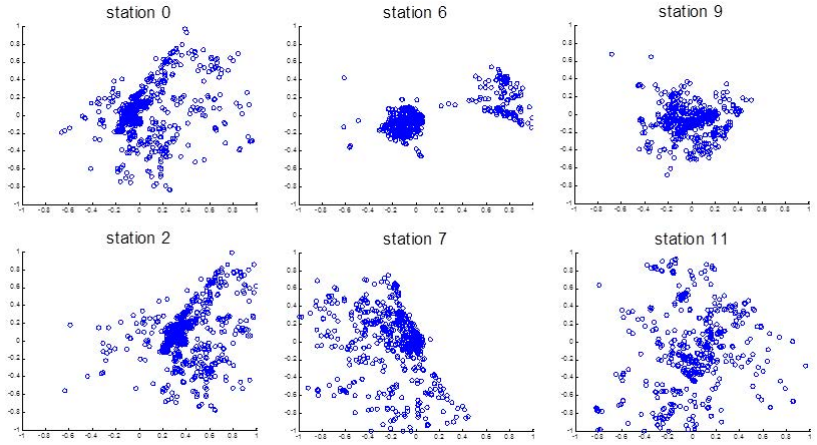

Fig. 3. The distribution of users for six stations. Each station is located at (0,0)

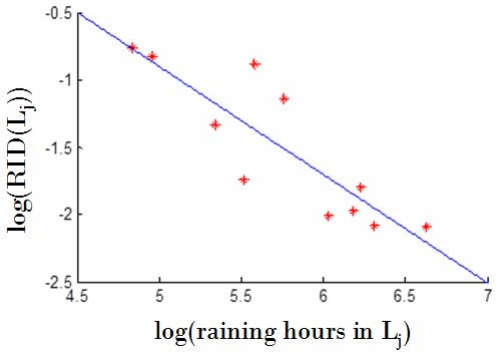

Fig. 4. The relation between the frequency of rain and the degree of users' interest in rain

$$Tmp(T_i) = \frac{1}{\begin{array}{c}\textit{the average number of the users who are interested} \\ \textit{in rain in the time slot when } T_i \textit{ is posted}\end{array}}$$

$$Geo(T_i) = \frac{(1 + dist(T_i, L_j))^{\alpha}}{\#\textit{tweets in the sub-region where tweet is posted}} \text{ ,where } \alpha \leq 0.$$

In our weighting function $Tmp(T_i)$, "the users" who are interested in rain are those users posting rain related tweets, which can be learned from training data. $Geo(T_i)$ contains two parts. In the denominator, we divide location L_j into 4 by 4 sub-regions. Then a tweet T_i is normalized according to the population of a sub-region where it is posted. The numerator is a function proportional

to probability density function of a exponential distribution. It considers the distance between L_j and T_i(i.e., the location where T_i is posted). The farther the distance between L_j and T_i is , the lower reliability of T_i is to identify the weather condition of L_j. The parameter α ($\alpha \leq 0$) is used to control the weight. The distance factor will be emphasized if we decrease α.

3.3 Event Detection and Modeling

Now we use the weighted signal to detect rain events and model their life cycles. Here we borrow the aging theory based method proposed in [14]. The method uses a wavelet based method to detect the burstiness of signal as the beginning of rain events. Since users don't always keep their interest in the rain event, the signal will decrease as time goes by. In this duration of rain, exponential function is applied to model the decay of signal. To determine when the rain ends, finally a threshold based method is proposed. The illustration of the model is shown in Figure 5.

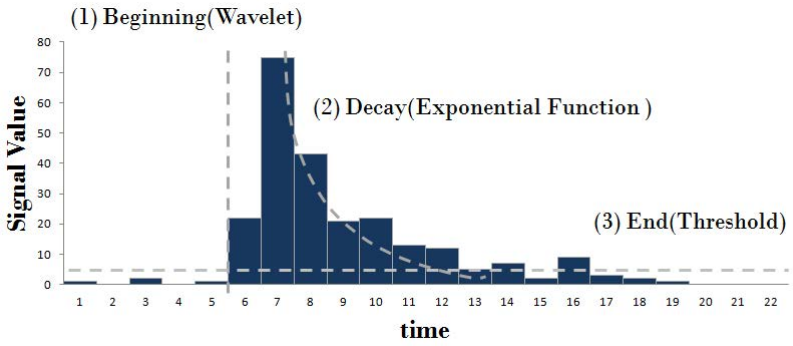

Fig. 5. Illusion of aging theory based model for rain event

4 Experiments

To evaluate the performance of our system, we gather the weather information(rain event) from thirteen weather stations in American and tweets posted around these stations(20 miles $\times$ 20 miles) over a period of three months, from November 1, 2011 to January 31, 2012. We measure the performance by three-fold cross validation on four indicators including precision, recall, F1-score and accuracy. The definitions of four indicators are defined as follows:

	Rain(Actually)	No Rain(Actually)
Rain(Reported)	A	B
No Rain(Reported)	C	D

$$precision = \frac{A}{A+B}$$

$$recall = \frac{A}{A+C}$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{A+D}{A+B+C+D}$$

We compare our performance with a threshold based method[14]. If the value of uniform weighted signal in any time slot is larger than the threshold, then the method judges the time slot as raining. We hope the system alarms everytime when it rains, but we do not expect many false alarms. Therefore, F1-score is adopted as the main indicator. Results are shown in Table 1. We can see that the performance of T-Signal is improved on recall. When we add the time factor into our model, the more desolate time slot will be more sensitive and more raining slots can be detected. Inversely, G-Signal improves uniform-Signal on precision. It is caused by weighting tweets with the geographical factor, and thus closer users have larger weights. Then the model can detect events more exactly. Overall, our experiments show that the spatiotemporal factor is helpful to detect the duration of a rain event.

Table 1. Performance of rain event detection model. G and T donates temporal and geographical weights respectively

	Threshold-Based	Uniform-Signal	T-Signal	G-Signal	T+G-Signal
Precision	0.4972	0.5610	0.5528	0.6160	0.6036
Recall	0.6253	0.7102	0.7254	0.6741	0.7059
F1-Score	0.5424	0.6239	0.6255	0.6425	0.6507
Accuracy	0.9299	0.9432	0.9424	0.9501	0.9492

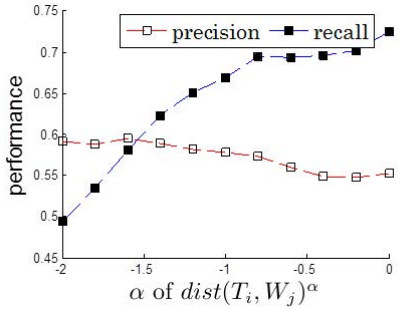

Fig. 6. The performance under various *alpha*'s

We also repeat the experiments under various α's (in $Geo(T_i)$) to see how the region size we use to crawl data affects the performance. The tendency in precision and recall is plotted in Figure 6. Smaller regions lead to lower recall and higher precision. For location L_j , merely using its nearby tweets is better in identifying its weather condition. However, not every location L_j always has sufficient nearby tweets, which will lead to the low recall.

We further categorize rain as heavy, normal and light, and then examine their properties. In our experiment, recall values for three are 0.9173(heavy), 0.7200(normal) and 0.6633(light) using T+G-Signal. The average number of rain related tweets normalized by the number of total tweets during heavy rain is 0.0598, which is greater than 0.0404 and 0.0301 for normal rain and light rain respectively. More tweets are posted during the more serious events, making it easier to detect in these periods.

As we discussed in Section 3.2 (geographical aspect), the location of users is an important factor to measure reliabilities of tweets. It also suggests that the bias of location affects performance a lot. For instance, station 0 and station 2 are near to each other, but their performance is quite different (Table 2). As shown in Figure 3, the users in station 2 are bias to upper right, namely, the direction of station 0. In contrast, the users in station 0 are closer to the center. This example may suggest the performance will be better if the users are closer to the station. In the case of station 7, it is limited by the topography(Figures 3 and 8) but it also has the good performance. The station 9 is another extreme instance. The users of station 9 are very close to the center and it results in a great performance.

Table 2. Performance of each station

station id	0	2	6	7	9	11
F1-score	0.6602	0.2100	0.5062	0.5664	0.7368	0.6399

Fig. 7. The locations of station 0 and station 2. The blue balloon means the center of station 0; the red thumbtack means station 2's.

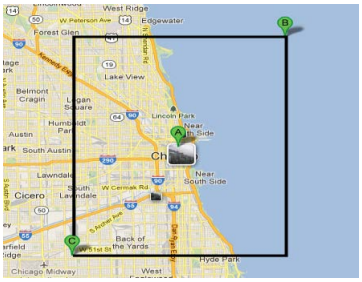

Fig. 8. The topography around station 7. Balloon A is the center, and other balloons and black frame are the boundaries for tweets gethering.

5 Rain Prediction

In the previous section, we have showed that our system is practical for detecting the rain events. Here we further want to understand whether we can predict weather conditions in the next n time slots. In Figure 9 we plot the signal during a rain event. From our observation, we believe there are two possible directions for weather prediction. First, the signal of L_j in this example starts to rise before it starts raining in L_j. A possible explanation is that before it rains in L_j, it rains near L_j so the users nearby post rain-related tweets. We therefore can detect rain events before it occurs. Second, whether it rains or not in consecutive time slots is not independent, making it possible to use previous data to infer the weather conditions in the near future.

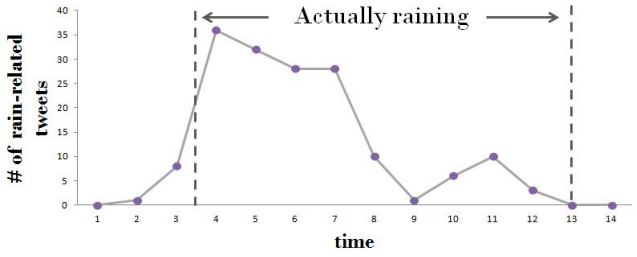

Fig. 9. A rain event example

We extend the aging based model [14] and construct a simple weather prediction system by modeling the life cycles of rain events. For a location L_j, after catching the beginning of a rain event, we start to monitor its signal. The signal generally doesn't reach maximum at once since users need time to respond the new event. To predict if it is still raining in the following time slots, we always assume the signal will decay exponentially(see Figure 5). This is because we have a few hints about future. What we know is how many tweets in the current time

Table 3. The performance of rain prediction

Time(Hour)	0.25	0.5	0.75	1	1.25	1.5	1.75	2
Precision	0.5562	0.5548	0.545	0.5401	0.4974	0.5018	0.4926	0.4868
Recall	0.7050	0.7118	0.7117	0.7089	0.7191	0.7233	0.7268	0.7171
F1-Score	0.6218	0.6235	0.6172	0.6130	0.5880	0.5925	0.5872	0.5799
Accuracy	0.9484	0.9498	0.9504	0.9508	0.9508	0.9516	0.952	0.9519

slot. If the current signal is strong enough, the decayed signal we expect in the next slot will not be lower than a given threshold. Then we predict the rain event will continue; otherwise, it will stop. Table 3 displays the results.

As we expect, the performance drops when the time we'd like to predict is farther from now, but it doesn't drop dramatically. We get 0.579 in F1-score for predicting if it still rains two hours later. One way to improve the performance is considering meteorological knowledge simultaneously. For instance, the length of rain in one location may vary with season and thus we can use an adaptive threshold instead a static one.

6 Conclusions

In this paper, we discuss the influence of spatiotemporal factor on rain events detection and prediction. The number of tweets changes in a day, making it hard to detect events on those "inactive" periods. The different distribution of population also leads to similar problem. Moreover, the degree of users' interest in rain varies with regions. We, therefore, re-weight tweets according to their spatial and temporal properties. Our experiment show its effectiveness in the detection of rain event duration. Different settings have been carefully examined. Finally, a simple prediction model is proposed to forecast weather conditions.

References

1. Allan, J.: Topic detection and tracking: event-based information organization, vol. 12 (2002)
2. Allan, J., Lavrenko, V., Jin, H.: First story detection in tdt is hard. In: Proceedings of the Ninth International Conference on Information and Knowledge Management. pp. 374–381. ACM (2000)
3. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998)
4. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining (2010)
5. Chen, L., Roy, A.: Event detection from flickr data through wavelet-based spatial analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (2009)

6. Cox, J., Plale, B.: Improving automatic weather observations with the public twitter stream (2011)
7. He, Q., Chang, K., Lim, E.P.: Analyzing feature trajectories for event detection. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007 (2007)
8. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65. ACM (2007)
9. Jianshu, W., Bu-Sung, L.: Event detection in twitter. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011)
10. Li, C., Sun, A., Datta, A.: Twevent: Segment-based event detection from tweets. In: Proceedings of the 21th ACM International Conference on Information and Knowledge Management, CIKM 2012. ACM (2012)
11. MacEachren, A.M., Robinson, A.C., Jaiswal, A., Pezanov, S., Savelyev, A., Blanford, J., Mitra, P.: Geo-Twitter analytics: Application in crisis management. In: 25th International Cartographic Conference (2011)
12. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web (2010)
13. Teevan, J., Ramage, D., Morris, M.R.: #twittersearch: a comparison of microblog search and web search. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM 2011, pp. 35–44. ACM (2011)
14. Tzeng, Y.S., Jiang, J.Y., Cheng, P.J.: Event duration detection on microblogging. In: Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2012 (2012)
15. Yang, Y., Pierce, T., Carbonell, J.: A study on retrospective and on-line event detection. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998)