Discover the Spatio-temporal Process of Typhoon Disaster Using Micro blog Data

Chunyang Liang[1], Guangfa Lin[1,2,3,a], and Junchao Peng[1]

[1]Institute of Geography, Fujian Normal University, 350007 Fuzhou, China
[2]Fujian Provincial Engineering Research Centre for Monitoring and Assessing Terrestrial Disasters,350007 Fuzhou, China
[3]Research Center for National Geographical Condition Monitoring and Emergency Support in the Economic Zone on the West Side of the Taiwan Strait, 350007 Fuzhou, China

Abstract. When a disaster occurs, a large number of images and texts attached geographic information often flood the social network in the Internet quickly. All these information provide a new data source for timely awareness of disaster situations. However, due to the regional variation in the number of social media users and characteristics of information propagate in cyberspace, new problems arose in the pattern analysis of spatial point process represented by the check-in data, such as the correlation between check-in points density and disasters events density, the spatial relation between check-in points, the spatial heterogeneity of point pattern and associated influences. In this study, we took the No. 201614 Typhoon as an example and collected Sina Weibo data between September 14 and September 17,2016 using keywords "Typhoon" and "Meranti". We classified the Weibo texts using Support Vector Machine(SVM) algorithms, and constructed a disaster database containing relevant check-in information. In addition, considering the spatial heterogeneity of Weibo users, we proposed a weighted model based on user activity at the check-in points. Using Moran's I of the global autocorrelation statistics, we compared the check-in data before and after adding weights and discovered obvious spatial autocorrelation of the check-in data in real geographical locations. We tested our model on Weibo data with keyword "rain" and "power failure". The results show that series map generated by our model can reflect the typhoon disaster spatio-temporal process trends well.

1 INTRODUCTION

Sina Weibo is a micro-blogging service that counts with millions of users from all over the china. It allows users to post and exchange 140-character-long messages, Sina Weibo is used through a wide variety of clients[1], from which a large portion-about 90% of active users correspond to mobile users. Besides, Sina Weibo facilitates real-time propagation of information to a large group of users. This makes it an ideal environment for the dissemination of breaking-news directly from the news source and geographical location of events. Therefore, make it one of the research hotspots for disaster emergency management.

For instance, after typhoon "Haiyan" was landed on 30th, 2013, many first-hand information and statistical information came from social media, including the release of news information and rescue information and the release of disaster emotions[2]; In 2013, the Lushan earthquake in Ya'an city, Weibo became the main communication platform for various kinds of information, and the government official also used Weibo as an important information release platform. In addition, the response of Weibo information to the earthquake is almost instantaneous, far faster than the release of authoritative earthquake monitoring agencies[3].

Many scholars explored the mechanism and role of disaster events in social media. They pointed out that social media can be a public engagement platform for citizens, cooperation and NGO to participate in the government-led disaster emergency management in a quickly way[4,5].

2 RELATD WORK

The literature on application of social media is still on increasing. So in this section our coverage of review is by no means complete. We just provide an outline of the research that is most closely related to disaster emergency management. For example, Bai Hua et al(2016) used the support vector machine to identify earthquake-related social media data and built a Sina Weibo Incident Monitor System to detect earthquake[6]. Murzintcev et al(2017) used hash tags to filter social media data that was not related to events, which reduced the cost of training classifier and enabled rapid collection of disaster-related information[7]. Chen et al(2014) analyzed the relationship between the messages and healthy states, and inferred the biological status of social media users based on the Latent Dirichlet Allocation to predict the outbreak time of influenza[8].

[a] Corresponding author: GuangfaLin@QQ.COM

The above studies have focused on the use of NLP (natural language processing) to collect and identify disaster-related messages, and have not fully utilized the spatio-temproal attributes of social media to conduct fine-grained disaster mining[9,10]. For this reason, many scholars use the spatial analysis of GIS to prove the relevance between the distribution pattern of disaster-related check-in data and the distribution of actual disaster situations. For example, Wang Y D et al(2016) analyzed the association between the the social media check-in data and the actual disaster situation based on DBSCAN[11]. Xu J H et al(2015) assigned the social media data weights according to the disaster description words, and interpolated the check-in records to generate a distribution map of disaster intensity[10]. Chen Z et al(2016) used the hotspot analysis tool to analyze the social media data related to Typhoon "Damrey" and found that the hotspots of check-in data were obviously concentrated in areas with heavy rainfall[12]. Bakillah et al(2015) used VDBSCAN (variable density cluster algorithm) to conduct geo-located detection on typhoon "Haiyan" related tweet data and used it to identify and locate post-disaster events[13]. However, because such studies do not take into account the Spatial heterogeneity of user distribution, the results are biased toward region which have larger number of social media user[9-13].

In order to eliminate the spatial heterogeneity of social media user's distribution, and mining the correlation between the disaster-related check-in data and typhoon disaster. In this study, we evaluated the impact of spatial heterogeneity of Weibo user ,propose a weighted model based on user activity at the check-in points, and combine the disaster-related data to verify our model.

3 DATA

This section describes how we collected a set of messages related to Typhoon "Meranti" from Sina Weibo and how to determine the area affected by typhoon.

3.1 Acquisition of Weibo data

For the web structure of Sina Weibo, we built a corresponding web crawler system. When the data is stored in the database, user ID and Weibo's transmission time constitute a primary key to prevent data redundancy. Based this collection system, we collected 170013 typhoon-related Weibo data from September 14 to September 18, 2016,of which 27218 messages contains location information. We will discuss how to classify Weibo data with location information in Section 3.4.

3.2 Determined by the typhoon-affected area

In order to determine the typhoon-affected areas, we manually collated the disaster information and news report showing that Typhoon "Meranti" impact areas were distributed in Fujian, Zhejiang, north eastern Jiangxi, south-central Jiangsu, south eastern Anhui and Shanghai.

3.3 Acquisition of POI check-in data

Sina Weibo official will update POI's check-in data daily. Also, based on the web crawler system, we regularly collect the POI check-in times each day, obtain the observation sequence of the POIs within a certain period of time in the study area, and provide data support for the calculation of the user activity of the POI.

3.4 Classification of Weibo data

First, We manually selected 800 Weibo data and grouped them into four categories including "warning information" , "disaster information" , "irrelevant information" and "rescue information" and labelled them manually. Next, we use the Chinese word segmentation component of ICTCLAS[14] to segment the Weibo text. Since the accuracy of word segmentation affects the performance of the classifier, we supplement the vocabulary of typhoon disaster features, establish a dictionary for typhoon disasters, and remove advertisement words and punctuation marks etc.Finally, because Weibo text is short and contains rich emoticons. In view of this phenomenon, we have retained the features of the emoji in Weibo, increased the dimensions of short texts, and made up for the sparsely of Weibo texts.

For labelled documents, We use the chi-square test[15] to filter the feature vocabulary of each category and quantify them using the TF-IDF[16] (Term Frequency - Inverse Document Frequency) algorithm. After that, We trained a supervised classifier by SVM[17] which is a machine learning algorithm proposed by Corinna Cortes et al. Different kernel functions can build different SVM, and their recognition performance is also different. In this study, We use a linear kernel function that is more suitable for text classification to train the classification model and perform also a k-fold cross validation[18] strategy to determine the value of the penalty term which is an essential parameter for the linear kernel function. The supervised classifier achieves an accuracy equal to 87.2% in sample data. This shows that the SVM algorithm can effectively classify Weibo texts, and finally get 13088 disaster Weibo texts with check-in location information.

4 ANALYSIS

The spatial distribution of check-in data is affected by the distribution of Weibo users[10-12].On this section we discuss how to build user weighted model to solve the spatial heterogeneity of Weibo users distribution,and verify the effectiveness of the model through spatial analysis.

4.1 The build of user weighted model

The spatial distribution of Weibo users will affect the distribution of Weibo check-in points. In order to eliminate the influence of user distribution , we uses the daily POI check-in amount updated by Sina official to describe the distribution differences of social media users between regions. When a disaster event occurs, because the information spreads through social media, the check-in points are not independent. Its joint probability is shown in formula 1.However, in this case, due to the lack of a priori data, the conditional probability between the check-in points cannot be calculated and the dataset of this study is the original Weibo and does not include the re-Weibo. Therefore, in this study we assumes that the check-in points are independent of each other(formula 2). $P(C_i)$ in formula 2 is the probability of position i being check-in(formula 3).N_i is the check-in amount of position i under normal conditions, and T is the sum of check-in amount of each check-in position(formula 4).

$$P_{checkin}=P(C_1)\cdot P(C_2 \mid C_1)\ldots P(C_n \mid C_1,C_2\ldots C_{n-1}) \quad (1)$$

$$P_{checkin}=\prod_{i=1}^{n}P(C_i) \quad (2)$$

$$P(C_i)=\frac{N_i}{T} \quad (3)$$

$$T=\sum_{i=1}^{n}N_i \quad (4)$$

When the disaster occurs,some users generate information either by providing first-person observations or by bringing relevant knowledge from external sources into sina Weibo.users will send Weibo with location information in the same place, so we merge them(formula 5).Finally, put formula 4 into formula 5 and take the logarithms on both sides to get formula 6.

$$P_{checkin}=P(C_1)^{n_1}\cdot P(C_2)^{n_2}\ldots P(C_i)^{n_i} \quad (5)$$

$$-\ln(P_{checkin})=\sum_{i=1}^{m}[n_i\cdot\ln(T/N_i)] \quad (6)$$

In formula 6, n_i and N_i both correspond to the attribute values of a certain check-in point, but due to the wide range of typhoon disasters, this study use city-level cities as the research granularity, so the meanings of n_i, N_i and T have changed.n_i is the sum of disaster-related Weibo in city i, N_i is the city's check-in number in normal, and T is the sum of all city's check-in number in the study area. Calculating the N_i in formula 6, we can get user weights in different areas.

In order to calculate the N_i, we collected the city's daily check-in number that were updated by sina Weibo from July 10 to July 16,2017.The number of check-in in some cities is shown in Figure 1.Because the check-in behavior of social media users can be regarded as an independent event[19], and a large number of check-in points are distributed in social networks. Therefore, the behavior of a certain Weibo user to check in at a certain POI is a small probability event celled P_c .Besides, Weibo users are larger groups celled N_{user} (formula7).The above conditions are consistent with the features of Poisson probability distribution and the limit as shown in formula 8.In this study, we use the method of MLE(maximum likelihood estimation) to calculate the likelihood function $l(\lambda)$(formula 9) to obtain an estimate of λ which is MLE value of the check-in number of daily updates in city i and x_{id} is the observation of the N_i term in the d-day.

$$\lim_{N_{user}\to\infty,P_c\to 0} C_{N_{user}}^{k}P_c^{\,k}(1-P_c)^{N_{user}-k} \quad (7)$$

$$f(X=k)=\frac{\lambda^k}{k!}e^{-\lambda} \quad (8)$$

$$l(\lambda)=\prod_{d=1}^{n}\frac{\lambda^{x_{id}}}{x_{id}!}e^{-\lambda} \quad (9)$$

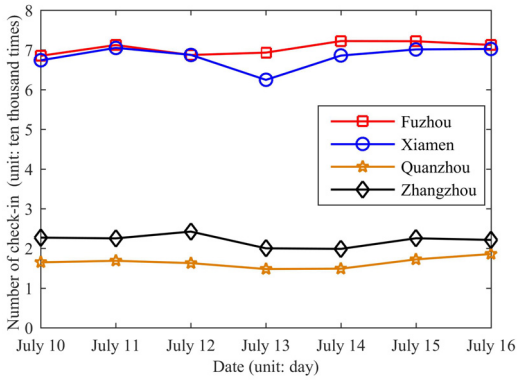

Fig.1 The statistics of daily check-in times in some cities

According to the hypothesis of this study and the above derivation process, we record the $\ln(T/N_i)$ item(formula 6) as the Weibo user's activity of city i, and use this item to eliminate the spatial heterogeneity of Weibo user's distribution. Finally, we use the natural break method to classify the user's check-in activity(Fig.2)

.As shown in the figure 2,the lowest cities are Nanjing, Hefei, Suzhou, Shanghai, Hangzhou, Guangzhou and Shenzhen. These cities have relatively high level of information and consistent with the facts. In addition, the n_i item in formula 6 is the number of disaster-related check-in records in city i and it's spatial distribution is shown in Fig3.

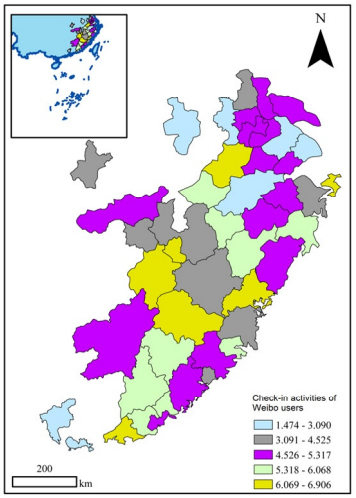

Fig.2 The distribution of microblog users check-in activities

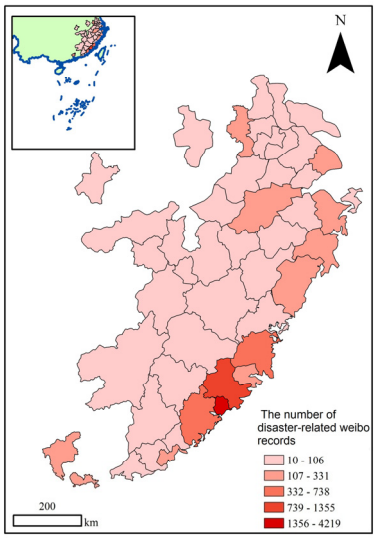

Fig.3 The distribution of disaster-related microblog's records with location information

4.2 Spatial autocorrelation

Spatial autocorrelation is an important indicator to test whether the attribute value of a geographical entity is significantly related in the adjacent space[20].

In this study, we use spatial autocorrelation analysis to explore the distribution pattern of disaster-related check-in data generated in the low-space friction social network, and compared the check-in data before and after adding user weights in IDW(Inverse distance weight) and KNN(K neighbours weight), as shown in Fig4 and 5. Besides, the comparison table between average neighbors and threshold distance in inverse distance weight in Tab1.

Tab.1 The comparison table between average neighbors and threshold distance in IDW.

Minimum number of neighbors (unit:each)	Average number of neighbors (unit:each)	Threshold diatance (unit:km)
1	5.87	178.6
5	6.41	369.1
10	10.04	488.3
15	15	637.8
…	…	…

From figures 4 and 5, it can be found that Moran's I shows a decreasing trend as the average number of neighbors increases. Among them, the downward trend of Moran's I in KNN is more obvious than under the IDW, which means that the disaster-related check-in data is sensitive to the spatial distance. Besides, after weighting ,the Moran's I value was stable at around 0.28 with a
mean neighbour of 5.87<N<7.48(threshold distance θ: 178.6km<θ<401.1 km) in IDW and compare to unweigh-ted data which is also show significantly positive spatial autocorrelation. This shows that with the distance decrea-
ses, the check-in data became more similar.

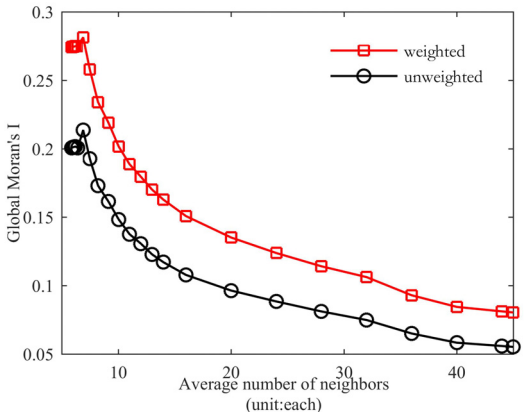

Fig.4 The average neighbors and Global Moran's I in IDW

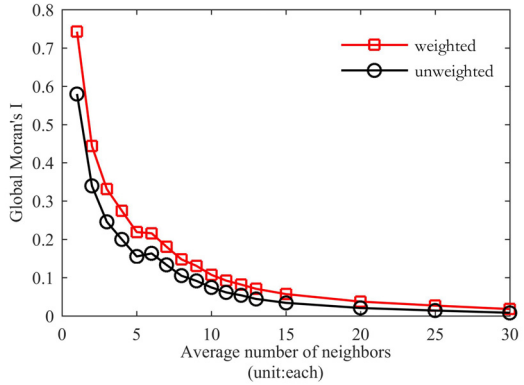

Fig.5 The average neighbors and Global Moran's I in KNN

4.3 Spatial-temporal analysis

From the results of Section 4.2, although the Weibo data is generated in the network with low space friction, it's not randomly distributed in geographic space. This section will select disaster Weibo based on the disaster feature words to discuss spatial-temporal process of the disaster-related check-in data and explore it' s relevance to the distribution of facts disaster. If they are related, we can quickly perceive the disaster situation according to the spatial distribution and content of the disaster-related Weibo data, and even predict the direction of the disaster, which has great significance for disaster emergency management.

The disaster-related Weibo texts includes different disaster features vocabulary such as rain ,power failure and landslides. According to statistical results of word frequency, the "rain" and "power failure" terms are high-frequency vocabularies of typhoon "Meranti" and can reflect situation of the typhoon at that time. Thus, we uses them as the search conditions to query on the disaster-related Weibo database, including "rain" and "power failure",the disaster-related Weibo texts were 3042 and 1236 respectively. According to the time of Weibo generation, we divided the microblogs containing "rain" into six time periods which use quantile method to ensure the number of check-in data is

the same in each time period. Finally, we obtain the center point position of each time period by formula 10.

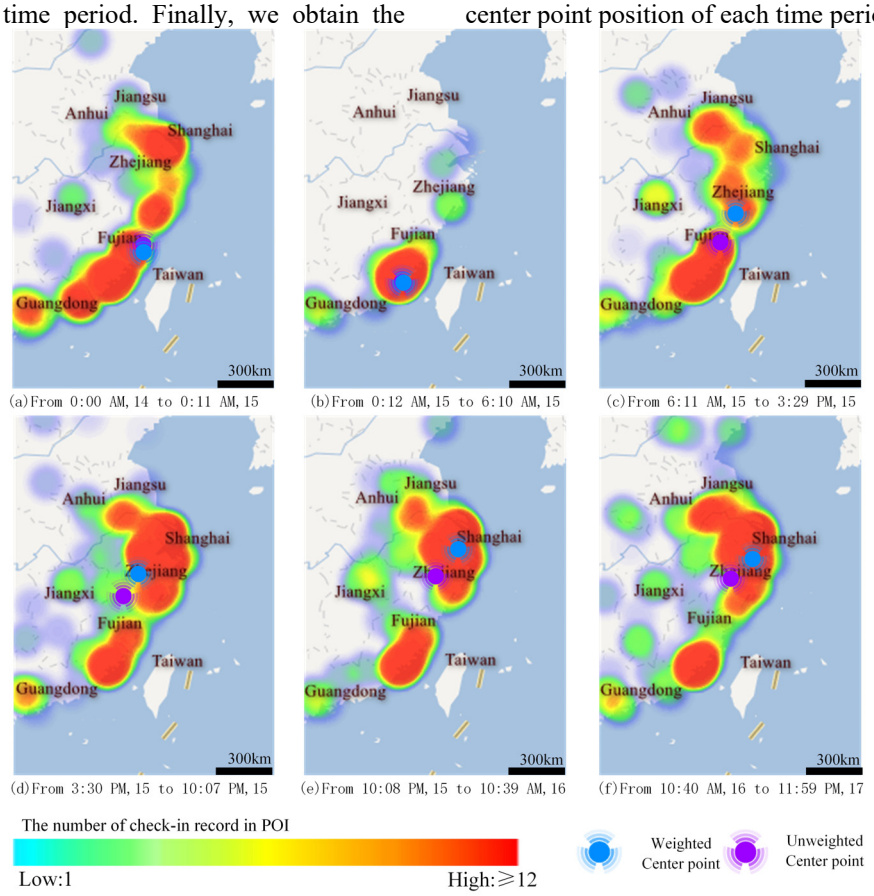

Fig.6 The time series of maps with fuzzy query using "rain" as keyword

$$\overline{X}(T_a)=\frac{\sum_{i=1}^{n}x_i}{n},\overline{Y}(T_a)=\frac{\sum_{i=1}^{n}y_i}{n} \quad (10)$$

Figure 6 shows the spatio-temporal distribution of check-in records obtained by "rain" keyword before and after adding user weights.

From 0:12 AM,15 to 6:10 AM,15, although the period is early morning, the frequency of generating disaster-related check-in data is the highest and its center is located near Xiamen whether or not it's weighted(Fig.6b)

.In addition, over time, the weighted central location of the disaster-related check-in data gradually moved to Zhejiang and Jiangsu, as shown in Figures6c ,6d and 6e.The disaster situation reflected by the map is agreement with the rainfall warning information and the actual disaster report released by Zhejiang Meteorological Observatory, Jiangsu Meteorological Observatory and China Meteorological Broadcasting Platform. By contrast,the un-weighted center location of check-in data have a tendency to move in Zhejiang province but it's not obvious. Besides, according to a report from the China Weather Broadcasting Platform at 07:51AM 17 September, typhoon "Meranti" degenerated into an extra tropical cyclone in the southern Yellow sea in the early morning of the 17th.This report confirmed our conjecture.

Figure 7 shows the spatio-temporal distribution of check-in records obtained by "power failure" keyword before and after adding user weights.

The weighted and unweighted center location of disaster-related check-in data are all concentrated in the Xiamen area, indicating that typhoon "Meranti" has a greater impact on Xiamen than other areas, and we have verified it through Fujian Provincial Climate Bulletin 2016.

5 CONCLUSIONS

In this study, we explores a method for processing and analyzing disaster-related check-in data. Based on this, we use social media data to analyze the spatio-temporal evolution features of typhoon No.14 in 2016 and compare it with actual disaster data and reports. In addition, considering the spatial heterogeneity of Weibo users, we proposed a weighted model based on user activity at the check-in points and the results of the spatio-temproal analysis show that the check-in data processed by user weighted model can better reflect spatial-temporal variations of typhoon disaster; The spatial autocorrelation analysis shows that Moran's I can be stabilized around 0.28 with a threshold distance of 178.6 km< θ <401.1 km using user weighted model, and the level of significance is less than 0.05.Although check-in data is generated in social networks with low spatial friction, it shows a significantly positive spatial autocorrelation. Thus,this study provides a theoretical basis for the perception of disasters through the spatial distribution pattern of data.

However, spatial data has multi-scale characteristics, and the attribute data of different spatial units often changes with different spatial scales and unit partitioning methods[21].In this study, we use spatial autocorrelation analysis only on the spatial scale of municipal cities. The

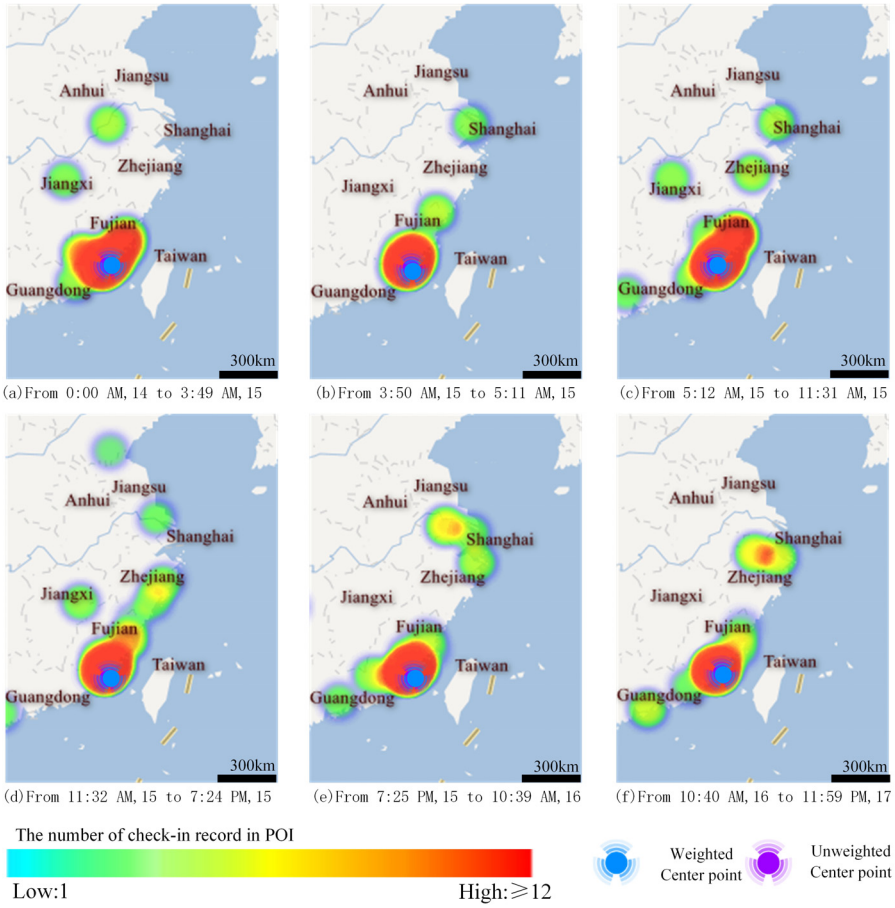

Fig.7 The time series map with fuzzy query using "power failure" as keyword

next study will conduct comprehensive analysis of the check-in data on various spatial scales to find the most appropriate expression scales and discuss changes under different spatial scales. Besides, although social media data provides a new approach to quickly sense the situation of the disaster, it also contains a large amount of redundant information. Compared with data obtained through rigorous scientific experiment, social media data is biased data, so mining it requires more robust algorithms. We will combine the Word2vec model[22] based on the deep learning framework to measure lexical similarity to improve text classification and disaster-related data retrieval accuracy.

ACKNOWLEDGEMENTS

This work was support by the Key Projects of National Key R&D Plan(2016YFC0502905), the Projects for Public Welfare Scientific Institutions of Fujian Province(2015R1034-1) and the Project of Department of Education of Fujian(JA15118).

References

1. Peng M,Guang C Y,Zhu J H,et al.A survey of topic detection and tracking technology for social media texts[J].Journal of Wuhan University(Science Edition),2016,62(3):197-217.
2. Cool C T, Claravall M C, Hall J L, et al. Social media as a risk communication tool following typhoon Haiyan[J]. Western Pacific Surveillance & Response Journal Wpsar, 2015, 6 (Suppl 1):86-90.
3. Zeng D J, Cao Z D.Big data solutions for emerging situation awareness and decision [J].China Emergency Management,2013(11):15-23.
4. Terpstra T, Vries A D, Stronkman R, et al. Towards a realtime twitter analysis during crises for operational crisis management[C]. Proceedings of the 9th International ISCRAM Conference,2012(4):1-9.
5. Vieweg S, Hughes A L, Starbird K, et al. Microblogging during two natural hazards events:what twitter may contribute to situational awareness[C].Proceedings of the SIGCHI

Conference on Human Factors in Computing System. New York,USA:ACM.

6. Bai H, Lin X G. Sina weibo disaster information detection based on chinese short text classification[J]. Journal of Catastrophology, 2016,31(2):19-23.
7. Murzintcev N, Cheng C. Disaster hashtags in social media[J]. International Journal of Geo-Information, 2017, 6(7):204.
8. Chen L, Hossain K S M T, Butler P, et al. Flu Gone Viral: Syndromic surveillance of flu on Twitter using temporal topic models[C].IEEE International Conference on Data Mining. IEEE Computer Society, 2014:755-760.
9. Wang Z, Ye X, Tsou M H. Spatial, temporal, and content analysis of Twitter for wildfire hazards[J]. Natural Hazards, 2016, 83(1):523-540.
10. Xu J H, Chu J X, Nie G Z, et al. Earthquake disaster information extraction based on location microblog[J]. Journal of Natural Disasters,2015(5):12-18.
11. Wang Y D, Li H, Wang T, et al. The mining and analysis of emergency information in sudden events based on social media[J].Geomatics and Information Science of Wuhan University,2016,43(3):290-297.
12. Chen Z, Luo N X, Gao T.Research of typhoon disaster assessment based on VGI[J]. Geomatics & Spatial Information Technology,2016(10):33-34.
13. Bakillah M, Li R Y, Liang S H L. Geo-located community detection in Twitter with enhanced fast-greedy optimization modularity: the case study of typhoon Haiyan[J]. International Journal of Geographical Information Science,2015,29(2):258-279.
14. Zhang H P, Yu H K, Xiong D Y, et al. HHMM-Based chinese lexical analyzer ICTCLAS[C].Proceedings of the 2nd SigHan Workshop.2003.184-187.
15. Meesad P, Boonrawd P, Nuipian V. A Chi-Square-Test for word importance differentiation in text classification[C].International Conference on Information and Electronics Engineering,2011.
16. Jones K S.A statistical interpretation of term specificity and its application in retrieval[J].Journal of Documentation,1972,28(1):11-21.
17. Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995,20:273-297.
18. Bengio Y, Gr Y. No unbiased estimator of the variance of K-Fold cross-kalidation[J]. Journal of Machine Learning Research, 2003,5(22):1089-1105.
19. Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users;real-time event detection by social sensors[C].Proceedings of the 19th International Conference on World Wide Web. New York,USA:ACM,2010:851-860.
20. Chen Y G. Reconstructing the mathematical process of spatial autocorrelation based on Moran's statistics[J].Geographical Research,2009,28(6):1449-1463.
21. Qi Y, Wu J G. Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices[J].Landscape Ecology,1996,11(1):39-49.
22. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science,2013(1):28-36.