Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 31 (2014) 731 – 739
Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

Estimating the Locations of Emergency Events from Twitter Streams

Ji Ao[a], Peng Zhang[a], Yanan Cao[a]

[a]*Institute of Information Engineering CAS, Beijing, China*

Abstract

Sina Weibo, the most popular microblogging service in China, plays a significance role in sharing and exchanging information on local and global level. Actually, part of the information are correlated with some events which may be local ones such as accident, protests or natural disasters to widely ones such as events concerning well-known people or political affairs. Our research interest is to detect the location where some events are occurring by analyzing only data from weibo information. In this paper, we denote three kind of location for each weibo: (1) content-based location; (2) posting location; (3) registration location of author of weibo. There are two main contributions in our work: (a) detect event-aware location of each weibo; (b) estimate accurate event location based on sets of event-aware locations. In order to detect event-aware location, we model there kinds of location data of each weibo to tackle with detection challenge. After we obtained a set of event-aware locations, we use a hierarchical clustering to estimate the accurate event location. The result showed that our method could improve the accuracy of event location estimation.

Keywords: Sina Weibo; User Profiles; Geo tags; Location estimation; hierarchical clustering

1. Introduction

Microblogging services such as Twitter, Sina Weibo allow large numbers of users to share and discuss by posting short messages whose length is limited, such as 140 characters on Sina weibo. Sina Weibo, the most popular microblogging service in China, plays a significance role in sharing and exchanging information on local and global level. Actually, part of the information are correlated with some events which may be local ones such as accident, protests or natural disasters to widely ones such as events concerning well-known people or political affairs. For example, in April 2013, online users used Weibo to disseminate their firsthand information about Yushu Earthquake within few minutes, and fast spread during a short period of time. Our research interest is to estimate the location where the objective events are taking place.

* Corresponding author. Tel.: +0-108-254-6715;
E-mail address: aoji@iie.ac.cn

doi:10.1016/j.procs.2014.05.321

Location is a valuable attribute for us to better understand and analyze the political, economic or social trends underlying online flow of information. Author[1] presented that location information is critical to understanding the impact of a disaster, including where the damage is, where people need assistance and where help is available. In wireless technology, spatial analysis mainly used GPS location of sensor and other related data. Hence, for event analysis based on social network, we can consider each weibo as social sensor. Then we can implement spatial analysis of events by analyzing the social sensors. In this paper, we investigate locating a specified event using weibo information crawled from Sina Weibo.

However, social sensors with useful rich information provided by social network do not have explicit GPS location (here denoted as event-aware location) used for describe the events. Hence, in this paper, we denote three kinds of location data to cover the shortage:

(1) Content-based location. We use the content of weibos as complementary characteristic to help us estimate event location. When users post a message, they usually attach some geographic terms to emphasize the description about the event. But there are two obvious challenges. One is the quality of content, Weibo users often use shorthand and non-standard vocabulary for informal distribution, which means that standard terms and accurate place names may not be present in the content of the weibo at all. Another one is that, when user posts a weibo via Sina Weibo, they do not always introduce the obvious location names.

(2) Posting location. Due to the increasing usage of mobile phones, such as an iPhone, equipped with Global Positioning System (GPS), a weibo may be associated with GPS data sometimes when users use it, which is denoted as weibo posting location. While these location data only represent the location of weibos posting, and cannot surely indicate occurring location of an event. Because that there may be a delay from occurring time of events to posting time of weibos. Furthermore, users may hear the news through other channels such as telephones and disseminate them via Weibo as original information.

(3) Registration location. As in Twitter[2], Sina Weibo can also permit users to enter their actual home locations in their user profiles. Hence, each weibo has a home location of author of weibo, here denoted as registration location.

After we obtain three kinds of location data, we propose a framework which combines the data to detect event-aware location for each weibo, and then uses a hierarchical clustering to estimate the accurate event location.

The main contributions of our framework are summarized as follows:

- A model is presented to determine the content-aware location of weibo.
- We model content-based location, posting location and registration to detect event-aware location.
- Use a hierarchical clustering algorithm to estimate the geographical locations of events based on sets of event-aware locations.

The rest of the paper is organized as follows. In Section 2, we introduce related work on location estimation. Then we briefly describe the background of Sina Weibo and the method of obtain location information directly in Section 3. The proposed framework for location estimation is shown in Section4, followed by the results of our experiments in Section 5. We finally present our concluding remarks and future plans in Section 6.

2. Related work

For years, microblogging as a social network has recently attracted much attention[3,4,5,6]. Analyzing location information from microblogging messages is helpful for understanding the impact of the emergency event. Location estimation on the web is studied using numerous techniques, such as by assigning weights to location names, by regarding actual locations of users as event locations, by using an Dempster-Shafer Theory or by applying Bayesian filtering methods.

Authors[7] work on automatically identifying the geographical scope of Web documents. A shared knowledge base is used to augment RDF-based descriptions of crawled Web pages with geographic meta-data. They recognize and disambiguated the geographical references over the documents and applying the PageRank algorithm to assign the given web page to a geographic location. Backstrom et al. develop a probabilistic framework for quantifying spatial variation[8]. There model is able to localize large classes of queries to within a few miles of their natural centers based only on the distribution of activity for the query. Furthermore, the model provides not only an estimate of a querys geographic center, but also a measure of its spatial dispersion, indicating whether it has highly local interest or broader regional or national appeal.

Some studies have specifically investigated collaborative bookmarking data, as Flickr provides, from a spatiotemporal perspective: Serdyukov et al. investigate generic methods for placing photos uploaded to Flickr on the World map by using a language model based entirely on the annotations provided by users and using tag-based smoothing and cell-based smoothing, and leveraging spatial ambiguity to improve the effect of this language model[9]. Large numbers of images uploaded to platforms such as Flickr do not contain GPS coordinates. Hauff et al. utilized the information extracted from a persons Twitter stream to improve the accuracy of the image location estimation[10].

Li present a Twitter-based Event Detection and Analysis System for searching[11], ranking and analyzing crime and disaster related tweets in Twitter. While plotting a tweet on a map, they first check its GPS tag. If it is not GPS annotated, they look for a place name in the tweet content. If it does not exist either, they use the location in the profile of its creator.Unakard[12] proposed a system for the early detection of emerging events by grouping micro-blog messages into events and using the message-mentioned locations to identify the locations of events. In their research they correlate user locations with event locations in order to identify the strong correlations between locations and events that are emerging. TwitterStand has been proposed by Sankaranarayanan et al.[13]. The 2,000 handpicked users of Twitter are used as seeders who are known to publish news. The online clustering method is used to group the message into the news topic. User location and content location are used to locate geographic content from each news topic. However, the results relied on handpicked users and there was no evaluation conducted.

Bayesian-filter techniques provide a powerful statistical tool to help manage measurement uncertainty and perform multisensor fusion and identity estimation[14]. Location estimation studies are often done in the field of ubiquitous computing. Estimating an objects location is arguably the most fundamental sensing task in many ubiquitous and pervasive computing scenarios. Representing locations statistically enables a unified interface for location information which produces applications independent of the sensors used such as GPS and infrared badges.

Sakaki at el.[15] devise a classifier of tweets based on features such as the keywords in a tweet and produce a probabilistic spatiotemporal model for the target event that can find the center of the event location. They regard each Twitter user as a sensor and apply particle filtering for estimating the locations of the target events.

Authors use Bayesian Filters to establish an early warning system for earthquakes in Japan[16]. They follow specific terms related to earthquakes in Twitter to capture relevant tweets. GPS annotations on tweets are used in Bayesian Filter for assigning locations to the detected earthquakes.

Dempster-Shafer Theory (an evidential reasoning technique) is based on the generalized Bayesian inference model introduced by Dempster. Ozdikis et al. applied basics of Dempster-Shafer Theory and Dempsters Rule of Combination on the location estimation problem for the events detected in Twitter[17], and it is the first study to use these methods on the location estimation problem in social network.

3. Background

Sine Weibo has received much attention recently. Since Sine Weibo was launched in August, 2009, its numbers of registered users was only 8 million. Along with the explosive development of Sine Weibo, the numbers has exceeded 500 million in 2013 and the active users have reached to 50 million every day[18]. By default, all the weibos are visible on a public timeline allows users to see their personal timeline for weibos they consider to be interesting. A user can follow other users to read their tweets. Due to the asymmetric following feature of Sina Weibo, a user who is followed by other users need not necessarily reciprocate by following them back.

Sina Weibo, along with other online social networking services such as Twitter, Foursquare and Facebook, have provided location services in their messages, either explicitly, by letting users choose their residential places in user profiles as registration location or implicitly, by enabling posting location, which is to associate messages with latitudes and longitudes.

3.1. User profiles

Sina Weibo allows users to enter their current locations in their user profiles, but the locations users select must be in the list of permissions. It is different from Twitter which location entries are entirely freeform, but limited to 30 characters[2]. So the location field in Sina Weibo may more accurate because it cannot corporate sarcastic comments that can fool traditional geographic information tools. This part (non-geographic information) accounts for about 16% of location field.

Naturally, most of registered addresses are correct on account of their little incentive to enter false information. They may be more likely to leave this field blank in comparison with deception. However, user are not always leaves in one place, he appears in this place today and may miles away tomorrow. So just using this field to predict the location of an event may a little incorrect.

For evaluating the validity of estimating the location of an event using registration location, we convert actual places into latitude/longitude coordinates by using the API interface[19] which is provided by Sina Weibo.

3.2. Geo tags

Due to the increasing use of mobile phones, such as an iPhone, equipped with Global Positioning System (GPS), users can share their location information in a more precise way as coordinates can locate where they sent weibos. Unfortunately, weibo users have been slow to adopt geospatial features because of the consideration for privacy or other reasons. Although more than over 70 percent of the users post weibos using smart phones, there are only a small part of users share their current location.

The system[20] presented by Romsaiyud only regards posting location as the locations of emergency events. Although this field can precisely represent of the location of where the tweet post, it is only the location of this tweet but not the event which this tweet described. That is posting location cannot completely correct to locate the event.

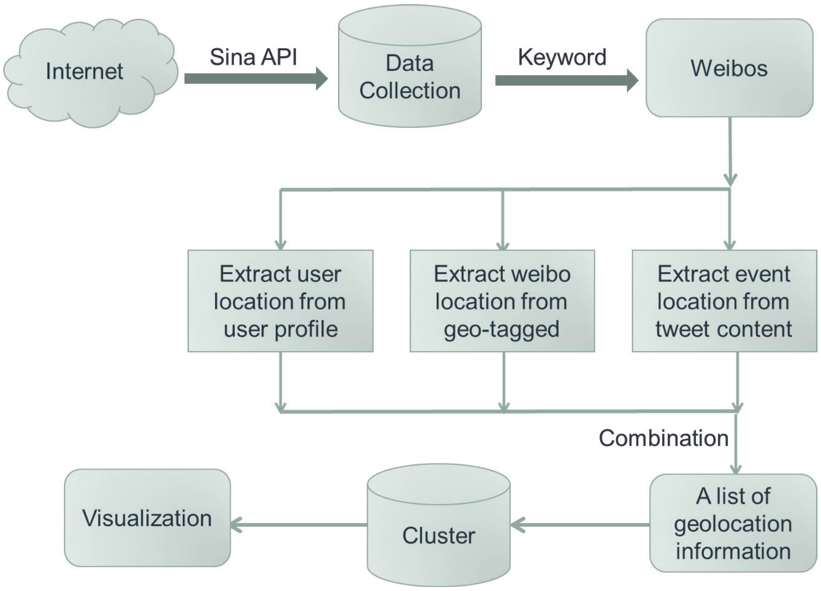
Internet
Sina API
Data Collection
Keyword
Weibos
Extract user location from user profile
Extract weibo location from geo-tagged
Extract event location from tweet content
Combination
A list of geolocation information
Cluster
Visualization

Fig. 1. Architecture of the System.

4. Technical Approach

In this section, we present our approach for location estimation as outlined by the processing flowchart in Fig. 1. For assigning locations to events, we utilized three location features in Twitter. There are the posting location, registered locations of author, and content-based location. The important novelty of our work are the framework we used combining these three three kind of location data which can enhance the correct rate compared to using only one feature. Then we obtain a set of locations relevant to the event, we used a clustering method to determine the location of the target event.

4.1. The Architecture of the Proposed Methods

In general, the system architecture comprised 4 main parts, Data collection based on keywords, training and prediction model, hierarchical clustering process and generating the web GUI. As Figure 1 illustrates the system consists of 4 steps: 1) we obtained the relevant data in a period of time via Sina Weibo API, then filtered these data based on the keyword related to specific event to obtain a serious of event-relevant weiobs. 2) This algorithm combined three kinds of location data above and we model these locations to make sure the location of each weibo presents. 3) based on a set of locations, we used a hierarchical clustering to get the final result. 4) GUI, we displayed the final results on the map to the users.

4.2. Description of Algorithm

The algorithm works on data set we have collected from Sina Weibo. It consists of 3 steps as following:

- Preprocessing of weibos; when we got a data set we filter it based on a keyword about a certain event and remove all duplicated weibos. Furthermore we removed all URLs of the weibos and automatic segmented each weibo.
- Each weibo was associated with three pair of geolocation information (latitude and longitude coordination). The first pair is obtained by converting the registered locations of the users from their user profiles to geolocation. The second pair is posting location. Because there may be more than one location appearing in a weibo, we propose an approach to choose one based on some regulations as the location of this weibo and convert it to geolocation as a third pair.
- Using a combination approach to compare the three pair in above and detect event-aware location for each weibo.

Since both field of posting location and registration location are easier to get, we focus on how to locate the event through the raw textual content of weibo. We find all terms or phrases reference to geographic location (e.g. province, city and country) from the contents of the weibos. By consideration that the location extraction from short text is one of the challenging problems of this research area, this task will be more focused in future work. In this work, we simply extract the message-mentioned locations via ICTCLAS[21], a Chinese word segmentation tool to identify locations in the messages.

If there is no location information can be found in a weibo, we cannot predict the event only analyzing the textual content of this weibo. So in our approach we discard the weibo which have no locations. If we extracted at least one location from a weibo, it needs us to choose a most relevant one to the detected event. There are two most likely to occur are: 1) there is only one location is associate with the event and others are irrelevant. 2) Although the relevant location is more than one, the scope which the location represents is not the same. So our algorithm will choose the most relevant one and its scope is the least minimum.

The presence of accurate location information on microblogs may be important since users reported in a survey that Twitter authors whose location is near to their own location are viewed as more credible[22]. So when the distance between registration location and posting location is lower than a threshold, we give a greater weight on it. Furthermore, if the location of weibo is far apart from the location user described, it means that the event may not take place around the users location. So in this case, we only consider the location in the content of weibo.

All tables should be numbered with Arabic numerals. Every table should have a caption. Headings should be placed above tables, left justified. Only horizontal lines should be used within a table, to distinguish the column headings from the body of the table, and immediately above and below the table. Tables must be embedded into the text and not supplied separately. Below is an example which the authors may find useful.

4.3. The Algorithm

Algorithm 1 The Events Model on Weibo Streams.

```
Input: Sina Weibo Streams have already crawled
        K: A query term (keyword) for a target event
        P: the locations from User profiles
        G: latitude and longitude coordinates from Geo-tags
Output: a list of pairs of geolocation information
Obtain weibos W by filtering keyword (K) from dataset we have collected.
For each weibo, applying a Chinese word segmentation tool, ICTCLAS, to segment the words and label the keywords and locations position in the weibo.
IF weibo contains location word, for each location lilocations (i=1, 2..., n), computer the distance di from li to the keyword based on their position in the weibo and other regulations.
Endow weight wi for each location.
    $W_i = \frac{\sum_{i=1}^{n} d_i/d_j}{\sum_{j=1}^{n}\sum_{i=1}^{n} d_i/d_j}$    (1)
Because each location has its own level, we modify the weight wi based on a regulation that the lower the level the greater the weight.
According to the wi we have calculated, we sort the locations and choose the maximum value to determine this location as the ultimately result of this weibo only based on the content. The locations we obtain based on this method are defined as L.
Convert L and G to latitude/longitude coordinates and combination three kinds of data with two principles: for each weibo, when the distance between lL and gG is larger than the threshold, we only concern l; when the distance between lL and pP is lower than the other threshold; we concern all the three ones.
Get a list of pairs of geographic information (latitude and longitude coordination)
```

5. Experimental Setup

5.1. Data collection

In order to understand the availability of the locations of weibos we collected weibo data randomly. We collected datasets via Sina Weibo Open API. Identifying weibos which associated with latitude/longitude coordinates. We collected Sina Weibo over a period of 20 days via Sina Weibo Open API[23]. Overall, we collected more than 480 thousands Weibos by users from November 20, 2013 to December 9, 2013. For better understanding the impact of posting location, all the weibos we collected are associated with a GPS data. For each of the retrieved messages, the following basic information was available: author, publication time, location field in user profiles, geo tags, and message content.

5.2. Metric

To evaluate the quality of our approach and baseline approach, we compare the estimated location versus the actual city location (which determined by manual label). There are two metrics:

Error distance: to each weibo, we adopt Euclidean Distance as the error distance which quantifies the distance in miles between the estimated location l1: (x1, y1) and the actual location l2 :(x2, y2). The Error Distance is defined as:

$$ErrorDistance = \sqrt{(x1 - x2)^2 + (y1 - y1)^2} \tag{2}$$

Furthermore we define average Euclidean Distance to evaluate the overall performance of locate an event. We depend on Weibos W to predict the event; Average Error Distance is defined as

$$AverageErrorDistance = \sum_{w \in W} \frac{ErrorDistanc(w)}{|W|} \tag{3}$$

The number of W is the total number of tweets in the test set.

5.3. *Results and Discussion*

We filtered weibos from the dataset which contains a keyword of an event, such as (earthquake), then we obtained 766 weibos in total. By preprocessing these weibos through removing the same one, the numbers of the remaining weibos are 637. Because there are not all weibos contain geographic terms, we filtered those which did not contain locations and remain 281 messages. In this experiment, we label the location of the specific event each weibo described based on the location of this event occurred and the raw content of weibo, there are 20 messages are not correlated with the event so we discarded them.

Table 1. Result

	megs	ED (total)	ED (Average)
Profile location	222	1267.284	5.708
Geo-tags	261	837.432	3.209
Content location	261	59.249	0.227
Combination	261	58.419	0.224

We show the comparison between our algorithms with other baselines discussed above. As shown in the table 1, we can see that the approach only concern registration location are the least relevant with event location because its Average Error Distance(AED) was the largest one and reached to 5.708.

Furthermore, the value of AED from Content location is dramatically lower than it from posting location. Although posting location can precisely represent the location of where the weibo post, it is cannot complete represent of the event which this weibo described. This shows that posting location cannot completely correct to locate the event.

Our framework can effectively locate the event with its Average Error Distance is only 0.224 which is lower than other baselines. That means our proposed approach can better detect event-aware location for each weibo.

5.4. *Cluster and map the data on the map*

Hierarchical clustering is a common method used to determine clusters of similar data points in multidimensional spaces[24].This method starts with a set of distinct points, each of which is considered a separate cluster. The two clusters that are closest according to some metric are agglomerated. This is repeated until all of the points belong to one hierarchically constructed cluster. We use this method to cluster all locations we have got above. Consider a completely connected graph where the vertices are the points we wish to cluster and the edges have a cost function that is the Euclidean distance between the points. The graph metrics determine intercluster distances according to the cost functions of the edges between the points in the two clusters.

The results after clustering are shown in Fig 2. The earthquake happened in several places in this period of time. We compare our result with the information from. We find that if the center of the earthquake is in sparsely populated region or earthquake magnitude is slower enough that people cannot feel it, it is more difficult to locate the earthquake from weibos. That result is reasonable: all other things being equal, the greater the number of sensors, the more precise the estimation will be.

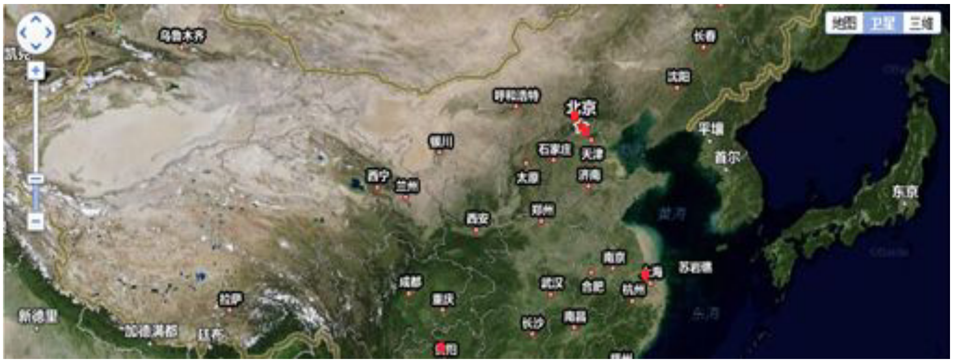

Fig. 2. showing in map.

6. Conclusion

Sina Weibo, the most popular microblogging service in China, plays a significance role in sharing and exchanging emerging events on local and global level. Location is a crucial attribute to understanding the ways in which online flow of information might reveal underlying. One of the biggest challenges is identifying the location where events are taking place. In this work, we focus on estimating the geographical locations of events that are detected in Sina Weibo.

We present an approach that incorporates two novel approaches: 1) selected one place as the geographic location of the target event which a weibo described based on the raw textual content only. 2) Combination the three kinds of locations we have get from user profiles, geo-tags and the content from weibos.

Microblogging has real-time characteristics that distinguish with other social media such as blogs. In future work, we will apply our work to deal with real-time data.

Acknowledgements

This work was supported by the NSFC (No. 61370025), 863 projects (No. 2011AA010703 and 2012AA012502), 973 project (No. 2013CB329606), and the Strategic Leading Science and Technology Projects of Chinese Academy of Sciences(No.XDA06030200).

References

1. Lingad, J., Karimi, S., Yin, J.. Location extraction from disaster-related microblogs. In: *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee; 2013, p. 1017–1020.
2. Hecht, B., Hong, L., Suh, B., Chi, E.H.. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM; 2011, p. 237–246.
3. Starbird, K., Palen, L., Hughes, A.L., Vieweg, S.. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM; 2010, p. 241–250.
4. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM; 2010, p. 1079–1088.
5. Z. Qiao P. Zhang, J.H.Y.C.C.Z., Guo, L.. Combining geographical information of users and content of items for accurate rating prediction. In: *In Proceedings of the 23rd ACM International World Wide Web Conference*. International World Wide Web Conferences Steering Committee; 2014, .
6. C. Zhou P. Zhang, W.Z., Guo, L.. Maximizing the long-term integral influence in social networks under the voter model. ????
7. Silva, M.J., Martins, B., Chaves, M., Afonso, A.P., Cardoso, N.. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems* 2006;**30**(4).
8. Backstrom, L., Kleinberg, J., Kumar, R., Novak, J.. Spatial variation in search engine queries. In: *Proceedings of the 17th international conference on World Wide Web*. ACM; 2008, p. 357–366.
9. Serdyukov, P., Murdock, V., Van Zwol, R.. Placing flickr photos on a map. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM; 2009, p. 484–491.

10. Hauff, C., Houben, G.J.. Placing images on the world map: a microblog-based enrichment approach. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM; 2012, p. 691–700.
11. Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.. Tedas: A twitter-based event detection and analysis system. In: *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE; 2012, p. 1273–1276.
12. Unankard, S., Li, X., Sharaf, M.A.. Location-based emerging event detection in social networks. In: *Web Technologies and Applications*. Springer; 2013, p. 280–291.
13. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.. Twitterstand: news in tweets. In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM; 2009, p. 42–51.
14. Fox, D., Schulz, D., Borriello, G., Hightower, J., Liao, L.. Bayesian filtering for location estimation. *IEEE pervasive computing* 2003; **2**(3):24–33.
15. Sakaki, T., Okazaki, M., Matsuo, Y.. Tweet analysis for real-time event detection and earthquake reporting system development. *Knowledge and Data Engineering, IEEE Transactions on* 2013;**25**(4):919–931.
16. 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*. ACM; 2010, p. 851–860.
17. Ozdikis, O., Oguztuzun, H., Karagoz, P.. Evidential location estimation for events detected in twitter. In: *Proceedings of the 7th Workshop on Geographic Information Retrieval*. ACM; 2013, p. 9–16.
18. http://data.weibo.com/report/detail/ 2013;.
19. http://open.weibo.com/wiki2/location/geo/addresstogeo 2013;.
20. Romsaiyud, W.. Detecting emergency events and geo-location awareness from twitter streams. In: *The International Conference on E-Technologies and Business on the Web (EBW2013)*. The Society of Digital Information and Wireless Communication; 2013, p. 22–27.
21. http://www.ictclas.org/ 2011;.
22. Morris, M.R., Counts, S., Roseway, A., Hoff, A., Schwarz, J.. Tweeting is believing?: understanding microblog credibility perceptions. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM; 2012, p. 441–450.
23. http://open.weibo.com/wiki2/statuses/publictimeline 2013;.
24. Murtagh, F.. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal* 1983;**26**(4):354–359.