Where Are You Settling Down: Geo-locating Twitter Users Based on Tweets and Social Networks

Kejiang Ren, Shaowu Zhang, and Hongfei Lin

Information Retrieval Lab, School of Computer Science and Technology
Dalian University of Technology, Dalian, China
renkj@mail.dlut.edu.cn,
zhangshaowu@gmail.com, hflin@dlut.edu.cn

Abstract. In this paper, we investigate the advantages of taking two dimensions of tweet content and social relationships to construct models for predicting where people settle down as their profiles reveal city- and town-level data. Based on the users who voluntarily reveal their locations in their profiles, we propose two local word filters - Inverse Location Frequency (ILF) and Remote Words (RW) filter - to identify local words in tweets content. We also extract separately the place name mentioned in tweets using the Named Entity Recognition application and then filter them by computing the city distance. We consider users' friends and 2-hop of followings. In our experiment, we finally combine these two dimensions to estimate user location and achieve an Accuracy of 56.6% within 100 miles in city-level and 45.2% within 25 miles in town-level of their actual location which outperforms the single dimension prediction and the baseline.

Keywords: Geo-location, Social Network, Twitter, Location-based Services, Location Prediction, Text Mining.

1 Introduction

Twitter's open and succinct service allows it to gather vast amounts of data and updates by users who come from different places. The user always inadvertently leaks some dialect words and place names of his/her residence in the process of adding updates. Understanding the geographic features of those update statuses enables the system to push better local advertising, highlight points of interest, show local news, create recommendations for friends living in the vicinity, and even help search engines understand users' search intentions better. In this paper we build textual models of local words and place names based on pure tweets to estimate a user's place of residence, even when the user does not explicitly reveal the place name, or his/her geographic coordinates in the profile.

Living in close geographical proximity may enable people to share common characteristics, provide real-time information and eyewitness updates about events of local interest [12], and recommend local friends to their extended friends. Furthermore, people who have reciprocal relationships are more likely to be geographically

Y. Hou et al. (Eds.): AIRS 2012, LNCS 7675, pp. 150–161, 2012.

close [2]. In this paper, we build social network models using users' followers and followings to predict their location. We find that people prefer to follow others who live in close geographic proximity.

In this paper, we propose hybrid probabilistic models combining textual models with social network models to estimate a user's location, and propose two local word filters. In the social networks model, we also consider the 2-hop of a use following's followers to estimate the user's location excepting the user's immediate friends. Finally we predict a user's location using a hybrid probabilistic model by combining the two dimensions.

The remainder of this paper is organized as follows: In Section 2, we review related works. Section 3 introduces the dataset of textual and social networks used in the experiments and estimation metrics in the paper. We introduce our models as well as the estimation algorithm, filter algorithm, and smoothing method in Section 4. We present the experimental results in Section 5. Finally, conclusions and future work are discussed in Section 6.

2 Related Work

Twitter has quickly become the premier platform for sharing real-time information since it first arose. Its convenience lies in a user's ability to post what he/she oberserves and hears in a local place. Lee et al. [18] proposed a geo-social event detection method to monitor the geographical regularities of local crowd behavior. Yardi et al. [12] examined the relationship between online social network structure and physical geographic proximity and verified that local events are of most interest to local citizens. Vieweg et al. [19] and Lee et al. [12] both discussed event broadcasting by local people.

Ye et al. [8], who used two features of places from explicit patterns of individual places and implicit relatedness among similar places for a binary SVM algorithm, developed a semantic annotation technique for Whrrl to automatically annotate all places with category tags. Lin et al. [9] investigated the factors that influence people to refer to a location and applied machine learning to model people's place-naming preferences. Anastasios et al. [20] analyzed the geo-temporal dynamics of collective user activity on Foursquare and showed that checkins provide a means to uncover people's daily and weekly patterns, urban neighborhood conditions, and recurrent transitions between different activities.

Amital et al. [4] employed the gazetteer approach to identify all geographic mentions within Web pages and to assign a geographic location and confidence level to each geo-locate Web content instance. Fink et al. [10] used both place names and organizational entities in blogs to determine an author's location. Serdyukov[13], Crandall[14], Hays[15], and Gallagher[17] et al. all attempted to predict where photographs originated using user tags and image-textual content. Popescu et al. [16] estimated users' home location by analyzing textual metadata associated with Flickr photos.

The aspects of geo-location users using tweet content in Twitter posts have become an active and promising area of research in the past two years. The most relevant works include Cheng et al. [1], Hecht et al. [6] and Kinsella et al. [5]. Cheng et al. proposed a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets. Hecht et al. studied user behavior in typing information into location field of user profiles, and then used simple machine learning techniques to guess users' locations on the country and state levels. Kinsella et al. [5] created language models of locations using coordinates extracted from geotagged tweets and model locations on varying levels of granularity from ZIP code to country level to geo-locate user and single tweets.

Backstrom et al. [3] used the social network structure of Facebook to predict location. Scellato et al. [7] described a supervised learning framework which exploits two linked features of friends-of-friends and place-friends to predict new links among friends-of-friends and place-friends. Li et al. [11] and Kwak et al. [2] studied the geographic features in Twitter.

3 Predicting Location

3.1 Textual Model

The textual model is a probabilistic estimator based on a user's tweets to estimate the location where the user settles down. Typically, the tweets posted by the user contain a great quantity of irrelevant information for location prediction. We next describe our local words filter algorithm and smoothing method.

Local Words Filter Algorithm. In Twitter, many words which appear in tweets have the similar probability in all locations and are distributed consistently with the population across different locations. Since the distribution means that most words provide very little power at distinguishing the location of a user, this even provides a lot of complications for prediction. In addition, some locations have a sparse set of words in their tweets because of the small population of the registered Twitter users or the small number of people updated their status in these locations. In order to improve the estimation accuracy, we must identify these local words in tweets and overcome the tweets' scarcity. Afterward, we are committed to the local words' identification and, at the same time, to overcome the data sparseness.

Before using a filter algorithm to identify local words, we preprocessed the content of the tweets. First, we eliminated the repeat tweets by string matching since we observed that most of these tweets are advertising. Then we removed the link in the tweet using the regular expression, eliminated all occurrences of a standard list of 429 stop words, as well as screen names which start with @ and single-letter words. Finally, we excluded punctuation in the tweets using Lucene Tool[1] and stemmed the word using Snowball[2]. By calculating word frequency, we only considered words that occur

[1] http://lucene.apache.org/

[2] http://snowball.tartarus.org/

at least 15 times in order to reduce the impact of incidental words (e.g.,yeeeeeees). Through the above processing, about 46,369 original words were generated from a base set of 521,103 distinct words.

Inverse Location Frequency: The first filter algorithm is the Inverse Location Frequency, which we called the ILF filter. It reflects the importance of the word in the collection of locations. The more locations that a word occurs in, the less discriminating the word is between locations, and consequently, the less useful it will be in location estimation. The form of ILF filter is defined as follow:

$$ILF_w = \log \frac{N}{n_w} \tag{1}$$

where ILF_w is the inverse location frequency for word w, N is the number of locations in training set. And n_w is the number of locations in which word w occurs. We set a threshold for filtering the original words which appear in many locations. After the application of the ILF filter, 19,424 local words were left from 46,369 words.

We used the Remote Words filter, which we called the RW filter, to filter out these remote words. In RW, we calculated the average distance of a location with all other locations for a local word, and eliminated the maximum value which exceeded our threshold in a specific iteration. When the average distance of all locations is less than the threshold, the iteration is over. In our experiment, we set the threshold as 200 miles. Before we operated the RW filter, we removed all words which occurred two times in every location. The remedy for these words is dealt with in the next step of NER. The formula of average distance is calculated as follows:

$$RW = \frac{1}{n-1} \sum_{\substack{i=1 \\ j \neq i}}^{n} | loc_j - loc_i | \tag{2}$$

Named Entity Recognition. We processed each tweet, applying NER and location-entity disambiguation to identify the related locations for the focus location. For each tweet, we exploited the named entity recognizer [21] to extract a location entity mentioned in it. Each entity was matched against the Yahoo! Placemaker[3] and got the latitude and longitude; it was then calculated for the distance of location of the entity to disambiguate the extracted location entity. We left these entities which are within 40 miles of the focus location. There were 10229 entities being recognized, and after disambiguation the total locations were 5702. Because the locations which users reveal in their profiles are place names, we put the words in location entities.

Textual Model Estimator. We used the Maximum Likelihood Estimation to geo-locate the users where they are settling down. Given the set of words U_w and location entities U_e extracted from a user's tweets U_T, we proposed the probability of the user being located in city l_i as:

[3] `http://developer.yahoo.com/geo/placemaker/`

$$p(l_i \mid U_w) = \lambda \sum_{w \in U_w} \alpha * p(l_i \mid w) * p(w) + \mu \sum_{e \in U_e} \beta * p(l_i \mid e) p(e) \tag{3}$$

where the $p(w)$is a *priori* probability which means the probability of the word *w*in the whole dataset and the a *priori* probability $p(e)$is equal to 1. The parameters α and β are used to denote the significance of the local word w for estimation location l_i, the other group parameters, λ and μ, are the weight for which the portion is more important for estimation. In our experiment, we set λ=1 and μ=5. The *priori* probability is calculated: $\mathrm{p(w)} = \frac{count(w)}{N \times local(w)}$, $p(l_i|w) = \frac{count_i(w)}{N \times count(w)}$ and $p(l_i|e) = \frac{count(e)}{N}$. Where *count*($w$) is the number of occurrences of the word w in the whole dataset, N is the total number of the word after filtering via the training set, *local*(w)stands for the location number where the word w occurs, *count*$_i$(w) donates the count of word w in location l_i, and *count*(e)stands for the total number of entity of recognized location.

Circular-Based Neighborhood Smoothing. There is a problem in that some locations have a few tweets in our training set because of having a small population or users unaccustomed to update status in the locations. The word distribution is sparse in these locations. How to overcome the location sparsity of words in tweets? We used the approach of circular-based neighborhood smoothing to improve the quality of user location estimation. Circular-based neighborhood smoothing considers all geographic neighbors from which the distance is 40 miles to the centre of a location. The circular probability of a word w can be formalized as:

$$p(r_i \mid w) = \sum_{l_j \in S} p(l_j \mid w) \tag{4}$$

where the r_i is the radius of the circle which the center is estimation location l_i, and S is the collection of locations in the round, at the same time including the estimation location. Then, the probability of the word w to be located in location l_i, $p(l_i|w)$ can be replaced with $p(r_i|w)$.

3.2 Social Network Model

The social network model is a probabilistic estimator based on the user's followers and followings (we called them *friends* consistently and whose locations get from training data) to estimate the location where he/she settles down. We can write down the likelihood of a particular location l_i as:

$$p(l_i \mid U_{SN}) = \frac{\sum \{u_i \mid u_i \in (U_{Fa} \mid U_{Fo}) \wedge u_i \in l_i\}}{N_{Fa} + N_{Fo}} \tag{5}$$

where l_i is the estimation of location, U_{SN} is the social network of user u including all followers U_{Fa} and followings U_{Fo}, N_{Fa} is the total number of followers for user u, and N_{Fo} stands for the number of followings of user u. The equation means how many users in location l_i of user u's followers and followings.

Generally, users do not add friendship connections at random with all other users, but, instead, they tend to prefer other users who are "close" to them in social network. For instance, many links do appear between individuals at closer social distance from each other, with the 2-hop neighborhood of single nodes being the largest source of new ties [7][22]. In Twitter, this phenomenon may be weaker, but it still can be considered. In our experiment, we only considered the number of a user's following's followers. We take into account this portion based on an assumption that the fewer of followers of user u_j (which u_j is a following of user u), the more intimate the relationship between them, the greater the contribution to the prediction. The formula of the likelihood is represented as follows:

$$p(l_i \mid U_{SN}) = \frac{\sum\{u_i \mid u_i \in (U_{Fa} \mid U_{Fo}) \wedge u_i \in l_i\} + \sum\{u_i^{'} \mid u_i^{'} \in U^{'}_{Fa} \wedge u_i^{'} \in l_i\}}{N_{Fa} + N_{Fo} + \sum N_{Fa}^{'}} \tag{6}$$

where $U_{Fa}^{'}$ is the followers of user u_j, $N_{Fa}^{'}$ is the number of followers for user u_j. We also considered the follower's social networks and the following's followings, but the result is lower.

3.3 Hybrid Model

We proposed a hybrid probabilistic model of combining a textual model with a social networks model. Since the dimensions of the two models are not the same, we first normalized the resulting values of the two models into values in the range [0, 1].

$$p(l_i \mid U) = \frac{p(l_i \mid U) - \min\{p(l_i \mid U)\}}{\max\{p(l_i \mid U)\} - \min\{p(l_i \mid U)\}} \tag{7}$$

where max{$p(l_i \mid U)$} and min{$p(l_i \mid U)$} are the maximum and minimum value of $p(l_i \mid U)$ respectively for all the estimated locations. Then the hybrid probabilistic is

$$p(l_i \mid U) = \omega * p(l_i \mid U_w) + (1-\omega) * p(l_i \mid U_{SN}) \tag{8}$$

where ω is the balancing coefficient in the range [0, 1].

4 Experimental Results

4.1 Data Collection

Twitter offers an open API that is easy to crawl and collect data. However, we used two existed corpora: [1] provides tweets and the location of user (we call this data

CHENG) and [2] offers the social network of the user (we call this data KWAK) in Twitter. The commonality of the CHENG and KWAK is that they have the same user ID in Twitter. The data of CHENG provided both the training set and test set. The training set contains 115,886 Twitter users and 3,844,612 updates from the users. All the locations of the users are self-labeled in United States. The test set contains 5,136 Twitter users with over 1000 tweets each of user and the total updates are 5,156,047 from these users. All the locations of users are uploaded from their smart phones with the form of "UT: Latitude, Longitude"[1]. The data of KWAK contain social graphs, mapping tables from numeric IDs to screen names, and restricted user profiles (> 10,000 followers) which are collected from July 6th 2009 to July 31th 2009 in Twitter. In our experiments, we only used the social graph portion which includes 1.47 billion social relations (still calling social graphs KWAK).

4.2 Metrics of Evaluation

We used the metrics which were defined in paper [1] and compared the estimated location of a user versus the actual location based on his/her latitude and longitude coordinates. The first metric is Average Error Distance (*AvgErrDist*) of all test users. The other metric is Accuracy which considers the percentage of users with their error distance categorized in the range of 0-x miles. We regard estimation as city level when x is equal to 100 as mentioned in [1] and [10]; when x is equal to 25, it is town level [3]. *Accuracy@K* means the accuracy metric in the top-k with the least error distance to the actual location which the *ErrDist(u)* is lower x.

4.3 Estimation Methods

1. ILF filter (ILF).We estimated the location of users using the ILF filter to filter the local words in the training set.
2. RW combines with ILF (RW+ILF). An approach that combined the remote word filter with the ILF to select local words in the training set to predict user location.
3. Named Entity Recognition (NER). This is a traditional method to identify a user's location in social networks using the content.
4. Named Entity Recognition augments the two filters (NER+Fs). In this method, we identified the locations from the training set as local words and merged the two local word filters aforementioned.

   All of mentioned approaches combined the circular based neighborhood smoothing to estimate user location in Twitter before the local words filter.
5. Social network predict (SN). For each user, locations were ranked according to the probability that most of their friends are settling down.
6. 2-hop social network (2-hop). We estimated user location using his/her followings' followers and his/her friends.
7. Hybrid estimation (Hybrid_SN, Hybrid_2-hop). Predicting the user's locations used hybrid models combining the two dimensions of textual and social networks.

4.4 Geo-locating on the City Level

For estimating a Twitter user's city-level location, we take into account Cheng et al. [1], whose system, solely based on tweet content, could geo-locate 51% of the 5119 users in the test set within 100 miles of their actual locations and that the *AvgErrDist* across all users was 535 miles as our baseline location estimator. Meanwhile, the traditional location prediction method of NER also is treated as baseline.

First, when we did not use the local word filters, our system only estimated about 8% of 5119 users in their actual locations corresponding to 10% of Cheng's baseline. We observed that the strong positive impacts of the local word filter. With the local word filter ILF alone, we reached an Accuracy of 0.437 which is more than five times as high as the Accuracy without using the local word filter. The filter removed noise resulting from non-local words in the tweet content and significantly affected the quality of user-location estimation. As we continued to use the RW filter after using the ILF to filter out the remote words, the Accuracy increased from 0.437 to 0.501, because the local words, filtered through ILF, were removed from the local focus. The result means that 50% of the test users can be placed in their actual location. We noted that the *AvgErrDist* was reduced significantly, from 652 miles to 516 miles, which is lower than the 535-mile baseline. The result also outperforms the baseline of NER method. When we took the location entities into consideration, the Accuracy reached 0.5098 which is almost identical with the baseline of 0.510. Meanwhile, the Average Error Distance was reduced from 535 miles to 473 miles; the overall estimated error was significantly lower, and the ACC@2, which at most having one location was correct in the first two locations, is 0.635, exceeding the baseline of 0.624.

Table 1. Results for user-location prediction at city level

Method	ACC	*AvgErrDist*(Miles)	ACC@2
Baseline	0.510	535.564	0.624
NER	0.419	499.10	0.591
ILF	0.437	652.449	0.563
RW+ILF	0.501	516.039	0.585
NER_Fs	0.5098	473.617	0.635
SN	0.593	503.588	0.700
2-hop	0.522	573.71	0.643
Hybrid_SN	**0.566**	**442.321**	**0.683**
Hybrid_2-hop	0.560	446.267	0.675

Continuing the observation in Table 1, we found that only using the user friends, the Accuracy could achieve 0.593 and, when using the 2-hop nodes, 52.2% of test users can be geo-located within 100 miles of their actual locations (we do not count the users whose first two or more predicted locations of the probability are equal). However, we did not compare the SN model and 2-hop model with other models, for in SN, we only considered the 704 users who have more than ten followers and followings (whose followers is less than 300) simultaneously and, in the 2-hop model, the user whose followers and followings are more than five are 1421. We still found that the power of using one's social network to estimate the actual locations is strong; to some extent, its ability has exceeded the standard model, purely based on content to predict the user actual location.

Now, let us look at the predictive power of the hybrid model in Table 1. In hybrid models, we calculated the probability of users whose followers or followings are not zero (the Accuracy is about 42% of removing the users whose at least first two estimated locations' score are the same) and set ω=0.501 based on measures when combined with the model of NER_Fs. We also observe the positive impact of combining the two dimensions of textual and social networks. We can see that 2-hop and SN merged with a textual model result in better user-location estimations than only using one dimension. By observing Table 1, we found that the best Accuracy achieved 0.566, which means placing 56.6% of users within 100 miles of their actual location, with an *AvgErrDist* of all users of 442 miles. If the first two locations are considered, 68.3% of 5119 users can predict actual locations.

4.5 Geo-locating on the Town Level

Kinsella et al. [5] considered the users' self-reported location which is extracted from their profiles and tweets using the Yahoo! Placemaker estimated user location on the town level. We treated it, which the Accuracy is 0.362, as our baseline on town-level location estimations. From Table 2, we found that all of the models' (except ILF) predicting accuracy outperform the baseline and NER method. The combining of two local word filters purely based on tweet content provided the overall results: which 40.5% of users placed within 25miles of their actual location. We can see that the location entities have a negative impact on town-level location predictions. The result of estimating user actual location, in which the Accuracy reaches 50.1% of 760 users using the user's followers and following of more than five respectively, is encouraging. It also shown that the hybrid model provided attractive results that 45.2% of 5119 users could be estimated town level actual location.

Table 2. Results of user-location prediction on the town level

Method	ACC	ACC@2
Baseline	0.362	--
NER	0.223	0.377
ILF	0.358	0.458
RW+ILF	0.405	0.477
NER+Fs	0.381	0.489
SN	0.501	0.595
2-hop	0.443	0.537
Hybrid_SN	**0.452**	**0.548**
Hybrid_2-hop	0.445	0.540

4.6 Impact of the Number of User Followers and Followings for Location Estimation

In order to understand the impact of followers and followings on estimating the user's actual location, we investigated the Accuracy while the number of followers and followings is in continuous growth. We have the following four groups of experiments to observe the influence of friends on the prediction Accuracy: 1) the number of followers is growing but ignoring the followings (Fer); 2) the number of followings increases without followers (Fing); 3) the followings and the followers change at the same time (Friends); and 4) the number of friends is in synchronous growth and, meanwhile, the following's follower number is less than 300 (Friends_-300) (we carried out this experiment based on the fact that, if the follower number of a user is on a large scale in Twitter, he/she is likely to be a celebrity, thus bringing noise).

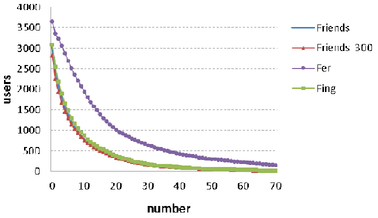

Fig. 1. The number of test users in four conditions

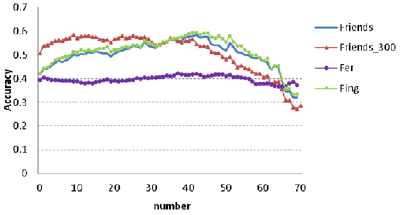

Fig. 2. Accuracy of the four conditions

From Fig.1, we can see that the test users with other three conditions are almost the same variation in different friend number except Fer. However, as the Fig. 2 shows, the estimation of Accuracy varies, and barely around about 40% when the number of followers is growing but ignoring the followings. In conditions 2) and 3), the estimation effect is the same and achieves the maximum when the number is equal, about 42. For the most users, when predicting their actual location using their friends, it is wise to not consider their followings whose own followers' number is more than 300.

5 Conclusion

In this paper, we investigated the problem of geo-locating users, which aimed to estimate user's actual locations via their tweet content and friends in microblogging service of Twitter. Based on the update statuses of a user, we proposed two local word filter methods to filter out noises of tweets content: Inverse Location Frequency (ILF) and Remote Words (RW). ILF filters the words which have a very wide geographic distribution and non-local features. RW is used to eliminate some local words, which are occasionally mentioned in locations far from their local focus. We also separately considered the place name mentioned in tweets to improve the estimation accuracy.

Simultaneously, we attempted to estimate the user's location via where the friends of the user settled most, and also considered the 2-hop friends to refine prediction accuracy. Finally, we combined the two dimensions of textual and social networks to get a better result. The results demonstrated the suitability of our approach and showed that our hybrid estimator can place 56.6% of 5119 users in Twitter within 100 miles of their actual location, 45.2% users geo-located within 25 miles of their actual location.

Next, we plan to investigate users' interactions with each other to refine the accuracy of the estimator, and we are also interested in mining local information from user's self-label tags in futures.

Acknowledgment. This work is partially supported by grant from the Natural Science Foundation of China (No.60673039, 60973068, 61277370), the National High Tech Research and Development Plan of China (No.2006AA01Z151), Natural Science Foundation of Liaoning Province, China (No.201202031), State Education Ministry and The Research Fund for the Doctoral Program of Higher Education (No.20090041110002).

References

1. Cheng, Z., Caverlee, J., Lee, K.: You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In: 19th ACM Conference on Information and Knowledge Management, pp. 759–768. ACM, New York (2010)
2. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media? In: 19th International Cnference on World Wide Web, pp. 591–600. ACM, New York (2010)
3. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: 19th International Conference on World Wide Web, pp. 61–70. ACM, New York (2010)
4. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-Where: Geotagging Web Content. In: 27th annual International ACM SIGIR Conferenceon Research and Development in Information Retrieval, pp. 273–280. ACM, New York (2004)
5. Kinsella, S., Murdock, V., O'Hare, N.: I'm Eating a Sandwich in Glasgow: Modeling Locations with Tweets. In: 3rd International Workshop on Search and Mining User-Generated Contents, pp. 61–68. ACM, New York (2011)
6. Hecht, B., Hong, L., Suh, B., Chi, B.E.: Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles. In: 2011 Annual Conference on Human Factors in Computing Systems, pp. 237–246. ACM, New York (2011)
7. Scellato, S., Noulas, A., Mascolo, C.: Exploiting Place Features in Link Prediction on Location-based Social Networks. In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1046–1054. ACM, New York (2011)
8. Ye, M., Shou, D., Lee, W., Yin, P., Janowicz, K.: On the Semantic Annotation of Places in Location-Based Social Networks. In: 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 520–528. ACM, New York (2011)
9. Lin, J., Xiang, G., Hong, J.I., Sadeh, N.: Modeling People's Place Naming Preferences in Location Sharing. In: 12th ACM International Conference on Ubiquitous Computing, pp. 75–84. ACM, New York (2010)

10. Fink, C., Piatko, C., Mayfield, J., Chou, D., Finin, T., Martineau, J.: The Geolocation of WebLogs from Textual Clues. In: IEEE International Conference on Computational Science and Engineering, pp. 1088–1092. IEEE Press (2009)
11. Li, W., Serdyukov, P., Vries, A.P., Eickhoff, C., Larson, M.: The Where in the Tweet. In: 20th ACM International Conference on Information and Knowledge Management, pp. 2473–2476. ACM, New York (2011)
12. Yardi, S., Boyd, D.: Tweeting from the Town Square: Measuring Geographic Local Networks. In: 4th International AAAI Conference on Weblogs and Social Media, pp. 194–201. AAAI, California (2010)
13. Serdyukov, P., Murdock, V., Zwol, R.: Placing Flickr Photos on a Map. In: 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 484–491. ACM, New York (2009)
14. Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the World's Photos. In: 18th International Conference on World Wide Web, pp. 761–770. ACM, New York (2009)
15. Hays, J., Efros, A.: IM2GPS: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE Press (2008)
16. Popescu, A., Grefenstette, G.: Mining User Home Location and Gender from Flickr Tags. In: 4th International Conference on Weblogs and Social Media, pp. 307–310. AAAI, California (2010)
17. Gallagher, A., Joshi, D., Yu, J., Luo, J.: Geo-location Inference from Image Content and User Tags. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 55–62. IEEE Press (2009)
18. Lee, R., Sumiya, K.: Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp. 1–10. ACM, New York (2010)
19. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what twitter contribute to situational awareness. In: 28th International Conference on Human Factors in Computing Systems, may 2010, pp. 1079–1088. ACM, New York (2010)
20. Anastasios, N., Salvatore, S., Cecilia, M., Massimiliano, P.: An Empirical Study of Geographic User Activity Patterns in Foursquare. In: ICWSM 2011: 4th International Conference on Weblogs and Social Media, pp. 570–573. AAAI, California ((2011)
21. Finkel, J.R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: 43nd Annual Meeting of the Association for Computational Linguistics, pp. 363–370. ACL, New York (2005)
22. Mok, D., Wellman, B., Basu, R.: Did distance matter before the Internet? Interpersonal contact and support in the 1970s. Social Networks 29(3), 430–461 (2007)