Social Media as Sensor in Real World: Geolocate User with Microblog

Xueqin Sui[1], Zhumin Chen[1,⋆], Kai Wu[1], Pengjie Ren[1], Jun Ma[1], and Fengyu Zhou[2]

[1] School of Computer Science and Technology, Shandong University, Jinan, 250101, China
[2] School of Control Science and Engineering, Shandong University, Jinan, 250002, China
chenzhumin@sdu.edu.cn

Abstract. People always exist in the two dimensional space, i.e. time and space, in the real world. How to detect users' locations automatically is significant for many location-based applications such as dietary recommendation and tourism planning. With the rapid development of social media such as Sina Weibo and Twitter, more and more people publish messages at any time which contain their real-time location information. This makes it possible to detect users' locations automatically by social media. In this paper, we propose a method to detect a user's city-level locations only based on his/her published posts in social media. Our approach considers two components: a Chinese location library and a model based on words distribution over locations. The former one is used to match whether there is a location name mentioned in the post. The latter one is utilized to mine the implied location information under the non-location words in the post. Furthermore, for a user's detected location sequence, we consider the transfer speed between two adjacent locations to smooth the sequence in context. Experiments on real dataset from Sina Weibo demonstrate that our approach can outperform baseline methods significantly in terms of *Precision*, *Recall* and $F1$.

Keywords: Location Detection, Social Media, Words Distribution over Locations.

1 Introduction

Location based services, such as dietary recommendation, shopping advertisement and travel routine plan, have increasingly become significant and popular not only for research but also for industry. How to detect users' regular locations automatically is necessary. New social media such as Twitter and Sina Weibo have spawned greatly as human-powered sensing networks. More and more users actively publish short messages about bits and pieces of their lives at any time and any places. This makes it possible to detect a given user's locations from

⋆ Corresponding author.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 229–237, 2014.

his/her published posts in social media. Some people may argue that most posts in social media contain GPS information since most equipments used by users to publish posts, such as mobile phone, have GPS modules. However, there are only about 0.42% posts contain GPS information according to [1], because most users close their GPS modules. Thus, it is not easy to detect a user's locations automatically.

In this paper, we propose a method to detect a user's locations purely based on the content of his/her published Sina Weibo posts in the absence of any other geospatial cues. For a post of a given user, we combine two components, i.e. a Chinese location library and a Bayes model based on words distribution over locations, to judge whether it contains location information and further to detect the location. If the post contains an explicit location name of the location library, the first component is used to match it directly. If the post does not refer any location name, there maybe some words which can imply the location where the user is. Thus, the second component first learns a model of words distribution over locations from Wiki and use it to mine the implied location information under the non-location words in the post. Furthermore, for a user's detected location sequence, we compute the minimum transfer time between two adjacent locations and use it to smooth the sequence in context. Experiments on real dataset from Sina Weibo demonstrate that our approach can outperform baseline methods significantly in terms of *Precision*, *Recall* and $F1$.

The rest of this paper is organized as follows. We introduce related work in Section 2. Section 3 gives our method in detail. In section 4, we discuss the corresponding experiments. We make some conclusions in addition to introducing future works in Section 5.

2 Related Work

There's a large amount of previous work on location prediction. There are three kinds of works: Location prediction based on content, location prediction based on social relationship and application of predicted locations.

Location prediction based on content: In [1], it finds word geographical spatial distribution based on the method of probability. According to the results, the words are divided into the local words of position sensitive (local words) and the location is not sensitive to nonlocal term (non-local words). Then it is based on local words to find the location of the user. [2] studies the location identification problem in blog. First, for each post it uses a named entity identifier which is based on GeoName gazetteers to identify location in the entity. After that, it identifies the place name entity. At last, it uses an ontology to represent hierarchical relationships between the place names which are stored in GeoName gazetteer. [3] uses place name dictionary to identify the main places of a web page. First, it recognizes all mentioned place names in Web page. Then for each case it gives a place and the corresponding confidence level. Finally, the confidence level of the highest place will be as the main place of the whole page. [4] studies the spatial transformation problem in search engine query.

[5] builds a $m * n$ grid based on longitude/latitude coordinate. Each cell represents a grid position. It uses some pictures of location known as the training set in Flickr, according to the label whether in the GeoName place names library to determine whether the label on behalf of the specific geographical location, and training a probability model of the language based on the label marked by the user. For a given image, the method is to predict its geographic location based on this model. [6] puts forward a system combined with the use of text and visual features from 20 million images crawling from Flickr and then mapping to the map. In this paper, limiting 10 markers in a city, this method uses the 10 markers image as a positive example and others as a negative example to train a classifier, then uses the classifier to implement the classification of the image. [7] puts forward several supervised methods, only using text to mark a location for document. First, the earth's surface is divided into multiple cells according to the longitude and latitude. After that, it uses the training set to train term distribution, document distribution and cell distribution. Then [7] uses three supervised methods to choose one of the most possible cell for each document as the geographical location of the document.

Location prediction based on social relationship: [8] explores supervised and unsupervised learning scenarios method to predict location, concluding reconstructing the entire friendship graph with high accuracy even when no edges are given and inferring people's fine-grained location, even when they keep their data private and we can only access the location of their friends. Li et al. [9] proposes a system to integrate both network and content-based prediction via a unified discriminative influence model which combine locations that a Twitter user mentions with the locations of the user's followers. Backstrom et al. predicts the home address of Facebook users based on provided addresses of one's friends [10], Cho et al. focuses on modeling user location in social networks as a dynamic Gaussian mixture, a generative approach postulating that each check-in is induced from the vicinity of either a person's home, work, or is a result of social influence of one's friends [11]. [12] proposes a novel network-based approach for location prediction in social media that integrates evidence of the social tie strength between users for improved location prediction.

Application of predicted locations: [12] studies the application of location prediction in public health. [13] studies the application of location in earthquake and other disasters. From [12,13] you can say that location prediction is very important in our real world.

3 Content-Based Location Detection

We first give a general description of the location detection problem with Sina Weibo.

Definition. Given a post bo_i published by a Sina Weibo user u, Our purpose is to predict the probability $p(l_j|bo_i)$ that user u is at the location l_j. The location with maximum probability is predicted as the user's location $l_{est(i)}$ when he publishes the post.

$$l_{est(i)} = l_j = \arg\max_{l_j \in L} P(l_j|bo_i) \tag{1}$$

where L denotes all the locations in the predefined Chinese location library.

In this paper, $P(l_j|bo_i)$ is defined as:

$$P(l_j|bo_i) = \alpha \cdot P_g(l_j|bo_i) + (1-\alpha) \cdot P_w(l_j|bo_i) \tag{2}$$

The function above consists of two components: $P_g(l_j|bo_i)$ and $P_w(l_j|bo_i)$ which are computed based on the location library and the Bayes model of words distribution over locations respectively. α is a predefined tradeoff parameter and its optimized value is determined by experiment. In this paper, we set $\alpha = 0.5$. Table 1 shows the result of $F1$ when α has different values.

Table 1. The values of $F1$ when α has different values

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$F1$	0.4685	0.4805	0.4734	0.4713	0.5863	0.5086	0.4688	0.4753	0.4780

A post may refer to multi locations. The $P_g(l_j|bo_i)$ is computed as:

$$P_g(l_j|bo_i) = \frac{f(l_j)}{\sum_{l_q \in L} f(l_q)} \tag{3}$$

where $f(l_j)$ represents the frequency of the location l_j mentioned in the post bo_i. $\sum_{l_q \in L} f(l_q)$ is the frequency of all locations in the Chinese location library L mentioned by bo_i.

A post may not contain any direct location names of the library. But there are some words which can imply the location where the user is. For example, when a user publishes a post "I am watching games at GuoAn home." It is well known that 'Beijing' is the home of GuoAn team. Then, we can refer the location is 'Beijing' because the probability that 'GuoAn home' coexists with 'Beijing' is very high. Thus, $P_w(l_j|bo_i)$ can be computed with words distribution over locations:

$$P_w(l_j|bo_i) = \frac{P(bo_i|l_j) \cdot P(l_j)}{P(bo_i)} \tag{4}$$

where $P(bo_i)$ denotes the prior probability of the post, and is the same for any post. So,

$$P_w(l_j|bo_i) \propto P(bo_i|l_j) \cdot P(l_j) \tag{5}$$

where $P(bo_i|l_j)$ denotes the probability that the location l_j generates the post bo_i and $P(l_j)$ is the prior probability of the location l_j mentioned by users.

$P(bo_i|l_j)$ can be computed as:

$$P(bo_i|l_j) = \prod_{w_s \in BO_i} P(w_s|l_j) \tag{6}$$

where bo_i is segmented into word set BO_i with the tool ICTCLAS[1] and w_s is a word of the post bo_i. $P(w_s|l_j)$ denotes the the probability of words distribution over locations.

We observe that there are many words such as local snacks and local scenic spots can reveal where the user is. In order to get the probability of word distribution over locations, we collect data from the Chinese Geography in Wiki[2] to train a Bayes model to compute $P(w_s|l_j)$. The Geography contains all Chinese cities and there is a detailed description for each city such as administrative division, government name, traffic, school, history, hospital, tourist attraction and so on. Finally, we collect 372 city-level locations and corresponding 6355 words (denoted as W) with typical location character. $P(w_s|l_j)$ is calculated as:

$$P(w_s|l_j) = \frac{\log count(w_s)}{\sum \log count(w_h)}; \quad w_s, w_h \in \{W \cap BO_i\} \tag{7}$$

where $count()$ denotes the word frequency in bo_i.

For different locations, values of $P(l_j)$ are different because their city levels, ranges and population are different. We assume the popularity of a location can be evaluated by the number that it is indexed by a search engine. Therefore, $P(l_j)$ is computed as:

$$P(l_j) = \frac{\log count(l_j)}{\sum\limits_{l_q \in L} \log count(l_q)} \tag{8}$$

where $count(l_j)$ represents the result number returned by Google when we submit l_j to Google. Intuitively, the higher the number is, the higher the popularity is.

There is a strong chronological order and context in a user's location sequence. Thus, we further consider the speed of transports to smooth the posts with abnormal locations and the posts without detected locations by using their adjacent locations. We observe that even by an airplane, it is impossible to transfer from one place to another faraway place within a limited time interval. Thus, we compute the threshold time of a person from one location to another as follows:

$$th(l_j, l_k) = \frac{Dis(l_j, l_k)}{v} \tag{9}$$

$th(l_j, l_k)$ denotes the time interval that a user moves from location l_j to location l_k. $Dis(l_j, l_k)$ is the distance in Kilometers between locations l_j and l_k. v is the speed of a kind of transport. Because it is difficult to decide which kind of transport the user takes, in this paper we consider the speed of airplane because it is the fastest transport and set $v = 800$ Kilometers. Thus, $th(l_j, l_k)$ is the minimum time that a user transfers from l_j to l_k.

For a user's two detected adjacent locations l_j and l_k, they correspond to two posts which have the property of publishing time t_j and t_k respectively. We can

[1] http://ictclas.org/

[2] http://en.wikipedia.org/wiki/China#Geography

get the moving time interval $t(l_j, l_k) = t_k - t_j$. If $t(l_j, l_k) < th(l_j, l_k)$, then we set $l_k = l_j$.

4 Experiments

4.1 Data Set

We have to contrast the dataset since there is no standard corpus for evaluating location prediction problem with social media. We collect real data from Sina Weibo. We first select 1000 active users and crawl all their posts published from Aug. 2009 to Apr. 2014. Then if the number of a user's published post is smaller than 10, we remove the user. After that, there remains 772 users. Finally, we annotate all posts with Chinese location library manually. The specific of the data set is shown in Table 2.

Table 2. Dataset from Sina Weibo

item	number
users	772
posts	826,018
locations	372
posts with annotated location	304,384

4.2 Baseline Methods

There are three baseline methods we compare with in this paper. The first method only uses the Chinese location library [2], represented as GL_{lib}. The second method is only based on the probability model [4], denoted as GL_{pm}. The third baseline method is the improved Probability model, represented as GL_{ipm}, which is extended from the second method with the transfer speed smoothing.

4.3 Evaluation Measures

We use standard measures *Precision*, *Recall* and $F1$ to evaluate the user's geolocation results. If the detected location generated by the methods agrees with the manually annotated location, we view it as a correct geolocation.

Precision is the fraction of detected locations that are correct.

Recall is the fraction of correct locations that are detected.

$F1$-*score* is calculated using following function: $F1$ = 2*(*Precision***Recall*) / (*Precision*+*Recall*).

Note that all measures above are macro-average for all users.

Table 3. Experiment results

Methods	*Precision*	*Recall*	$F1$
GL_{lib}	0.8226	0.2029	0.2989
GL_{pm}	0.8394	0.2649	0.3802
GL_{ipm}	0.8335	0.3681	0.4795
Our Method	0.8231	0.4991	0.5863

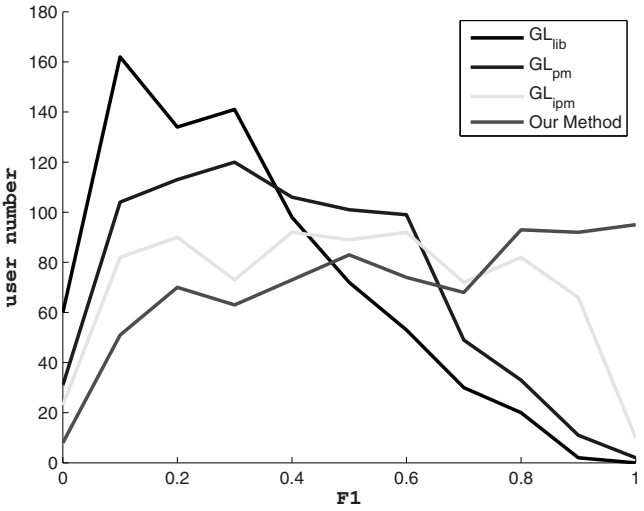

Fig. 1. Distribution of the number of users over $F1$

4.4 Results and Discussion

Table 3 shows the result of our method and the three baseline methods. It is obvious that our method achieves the best performance. Although the *Precision* of our method is slightly less effective than GL_{pm} and GL_{ipm}, the *Recall* and $F1$ of our method are significantly outperform that of baselines. Our approach improves the *Recall* significantly, it performs about 1.5 times than the second method. It is because that our method considers the smoothing technology, including removing abnormal locations and adding some detected locations. However, the values of *Recall* for all methods are not very high. The reason is that we consider all posts of a user that we can get the actual locations, but a lot of posts don't contain any location information. Therefore, we can't predict locations of these posts.

We can also see that the *Recall* of GL_{ipm} is 6% higher than that of GL_{pm}. This further confirms the effectiveness of the smoothing technology. We further have a look at the posts we predict false. They have these features: (i) there are many location names mentioned in the post. (ii) the nuptial problem. For

example, 'Chaoyang' is not only a city in Liaoning province but also a county in Beijing.

To further evaluate micro-performance of the four methods, we analyse the location detection performance of each user. Fig.1 shows the distribution of the number of users over the measure $F1$. It is obvious that when $F1 > 0.7$, the curve of our method is above those of the other methods. This means that our method can achieve good performance for most users, for the reason that many posts contain explicit location names and many contain some information about locations. The GL_{lib} can mine explicit location names and GL_{pm} can mine implicit location names. However, our method can mine not only explicit location names but also implicit location name, therefore, our method can achieve a better performance than others.

5 Conclusion

In this paper, we have proposed an approach to detect users' locations automatically using their published posts in social media. Our method considers both the direct matching with location name and the undirect mining of implied word distribution over locations. The transfer speed between locations is also utilized to smooth the detected location series. We implement both three baseline methods and our new method, and experimentally verify that our method can outperform the baselines especially in terms of the measure of *Recall*.

In the future, we plan to consider the social relationship of the user to further improve the performance. For applications, we can detect the hot and fine-grained locations and recommend them to users.

Acknowledgement. This work is supported by the Natural Science Foundation of China under Grant No. 61272240 and 61103151, the Doctoral Fund of Ministry of Education of China under Grant No. 20110131110028, the Natural Science Foundation of Shandong Province under Grant No. ZR2012FM037, the Excellent Middle-Aged and Youth Scientists of Shandong Province under Grant No. BS2012DX017 and the Fundamental Research Funds of Shandong University.

References

1. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: A content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759–768. ACM (2010)
2. Fink, C., Piatko, C., Mayfield, J., Finin, T., Martineau, J.: Geolocating blogs from their textual content. In: Working Notes of the AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0. AAAI Press (2009)
3. Amitay, E., Har'El, N., Sivan, R., Soffer, A.: Web-a-where: Geotagging web content. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 273–280. ACM (2004)

4. Backstrom, L., Kleinberg, J., Kumar, R., Novak, J.: Spatial variation in search engine queries. In: Proceeding of the 17th International Conference on World Wide Web, pp. 357–366. ACM (2008)
5. 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, pp. 484–491. ACM (2009)
6. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of the 18th International Conference on World Wide Web, pp. 761–770. ACM (2009)
7. Wing, B.P., Baldridge, J.: Simple supervised document geolocation with geodesic grids. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 955–964. Association for Computational Linguistics (2011)
8. Bigham, J.P., Sadilek, A., Kautz, H.: Finding your friends and following them to where you are. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 723–732. ACM (2012)
9. Sun, E., Backstrom, L., Marlow, C.: Find me if you can: Improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th International Conference on World wide Web, pp. 61–70. ACM (2010)
10. Myers, S.A., Eunjoon, C., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1082–1090. ACM (2011)
11. Caverlee, J., Jeffrey, M., Cheng, Z.: Location prediction in social media based on tie strength. In: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, pp. 459–468. ACM (2013)
12. Dredze, M., Paul, M., Bergsma, S., Tran, H.: Carmen: A twitter geolocation system with applications to public health. In: AAAI Workshops (2013)
13. 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, pp. 851–860. ACM (2010)