


SPECIAL ISSUE PAPER

Users' location analysis based on Chinese mobile social media

Zhibo Wang^{1,2}  | Yuechuan Guo² | Senzhe Zheng² | Wei Xu¹ | Lin Liu¹ |
Zixin Liu¹ | Xiaohui Cui²

¹Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, East China University of Technology, Nanchang, China
²International School of Software, Wuhan University, Wuhan, China

Correspondence

Zhibo Wang, Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, East China University of Technology, Nanchang 330013, China; or International School of Software, Wuhan University, Wuhan 430079, China.
Email: rs_wzb@whu.edu.cn

Xiaohui Cui, International School of Software, Wuhan University, Wuhan 430079, China.
Email: xcui@whu.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Number: 61440054 and 61462004; Fundamental Research Funds for the Central Universities of China, Grant/Award Number: 216274213; National Natural Science Foundation of Hubei, China, Grant/Award Number: 2014CFA048; Outstanding Academic Talents Startup Funds of Wuhan University, Grant/Award Number: 216-410100003; Natural Science Foundation of Jiangxi Province, Grant/Award Number: 20151BAB207042; Youth Funds of Science and Technology in Jiangxi Province Department of Education Grant/Award Number: GJJ150572, GJJ160589, GJJ160590 and GJJ170481; Outstanding Foundation of Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology, Grant/Award Number: JELRGBDT201709

Summary

After the rapid development for more than 20 years, Internet has gradually become the main carrier of people's information and behaviors in people's daily life. In addition, the innovation and popularization of smartphone GPS makes user location information much more available and accurate, helping it to create remarkable values by which people are attracted to focus on social media-related data mining and applications. However, because of the sparsity of social media geographical information, direct inferences of locations have plenty of difficulties. Under the background of big data, this research has revised the UGC-LI model in the preprocess of texts and the creation of the local dictionaries in which we take existed local dictionaries from the Internet into consideration, with the purpose of the inferences for users' and texts' locations. At the time of writing, through the crawler, we acquire users' personal information, the blog content, and customer relationships' (follows, fans) information more than 410 331 pieces from Sina Weibo. The experimental results show that the recall rate of the user location inference is 86.0%, whereas the precise rate is 77.4%, and the accuracy of text posted location inference is 66.8%. Compared with some other related algorithms, this revised model has comparatively better results in location inference for users and text publication.

KEYWORDS

big data, location mining, mobile social media, web crawler

1 | INTRODUCTION

With the rapid development and worldwide application of the Internet, an increasing number of people invest a large amount of efforts to the activities online, including entertainments (such as watching films and listening to music) and social contacts (like strolling Post Bar and interaction on Sina Weibo). And what makes the public publish and get texts faster is the advent and popularization of social networking service, such as Facebook, Twitter, Sina Weibo, and WeChat. What's more, just because of the rising of a new medium based on the Internet, its influences on the society are strengthening dramatically. According to the DCCI's report in 2013, the data size has attained 1.2 ZB by 2010, and over 1.8 ZB one year later. Besides, according to "Chinese Sina Weibo Blue Book in 2012", until 2012, there are 3.27 million Weibo users, which means that each user has

1.45 accounts, posts 2.13 pieces of Weibo, and forwards 3.12 pieces of Weibo in average". These data seem too redundant to handle. Actually, they are so comprehensive that they have recorded customers' overall information, which can afford us a more objective description about customers' behaviors and traits.

Among the various data from social media, especially with the popularization of global positioning devices such as smartphones and free Wi-Fi, geographical location information has become a relatively stable resource of data with high quality, which exerts huge influences on practical application, such as monitoring the spread of diseases, advertisement & arrangement, and analysis of urban floating population. Because of the 4V attributes of social media data, constraints of applications, and protection of privacy, however, a large number of users provide inaccurate geographical locations even null geographical locations, which leads to data's sparsity in social media. It has been experimented that, take Sina Weibo as an example, there are only 1% piece of Weibo with location tags among more than 100 million pieces of Weibo posted by 2 million users. Therefore, it may be a little inaccurate to deduce users' location if we only depend on location information provided in social media and ignore its sparsity. What we should take into account indirectly is that how to calculate user's location on the basis of the content of Weibo.

2 | RELATED WORK

Nowadays, there are two main research directions: to locate the users and to locate where the Weibo is posted. In the former fields, based on language consistency, Eisenstein et al¹ apply a multilayer generation model to connect the potential topic of the text and geographic districts and analyze the language differences in different regions, thus locate the users. In the contract, Hecht et al² deal with users' history trajectory data to design users' behavior model and then use multinomial Naïve Bayes model to locate the users. All of them rely both on high efficient and precise geographical database, and different database may lead to errors. Considering of that, researchers prefer to analyze the position from the aspect of users' content. Backstrom et al³ mathematically analyze the diversity of regional space by proposing a probabilistic framework, realizing rough geographical location and data mining in searching for regional interest points. Similarly, Cheng et al⁴ apply this kind of probabilistic framework in social media Microblog, which deduce users' position only by analyzing the content of posted Weibo. They design the model by baseline geographic and defining regional vocabulary and correct the location errors by building smoothing algorithm based on relative attributes like social relationship. However, the work of Cheng et al requires for the amount of manual work in the stage of building classifiers to analyze the words, which is heavy and leads to errors in research results. Ryoo and Moon⁵ revise the method of Cheng et al. They apply the GPS data provided by Tweet instead, which improve the accuracy of estimating position. Based on the content of the text, Backstrom et al realize that social relationships have an influence on position, so they explore the connection. Chandra et al,⁶ Jurgens,⁷ Li et al,⁸ and McGee et al⁹ all have related research in deducing position with the social relationship graph. Li et al propose a multiple profile model, which simulates users to post texts to get multiple position information and interactive position. In this way, the results are accurate and their supports and explanations are provided. Kotzias et al¹⁰ design a primary model to calculate the position with GPS data from Twitter. Depending on a language-position model, this model calculates the probability of each tweet and rank the possible position. Kotzias et al improve the method of Kinsella, explore users who are close, and locate the tweet.

In domestic, Rongjiao and Sansheng¹¹ propose the alternative method of positioning source users. Chi et al¹² analyze and compare the research methods of bulk data of positioning. In addition, Jingnan et al¹³ summarize the methods and developments related to bulk data of positioning. Compared to aboard research studies and application in location data science, the relative fields at home is at primary stage, which are short of valuable research results. Considering the inner relationship between text and location distribution, Wang et al¹⁴ propose a revised probabilistic model, which better reflects the potential relationships between the virtual and the real ones. It has several advantages, such as recognizing location relative words automatically, combining users' relationship spectrum and connection threshold to improve accuracy, and reducing the processing errors by giving weights.¹⁵ On the contrary, having not taken Internet words into account, the local lexicon based on the limited experimental data is not representative enough.

3 | MATERIALS AND METHODS

3.1 | Preprocessing

Comparing with news or other social data written under the strict rules, the content of microblogs could be much more casual and personal. Consequently, there could be some noise or useless information in our data sets, which need to be filtered before our location inference process to prevent negative influence caused by these unrelated factors. Content of microblog in our data set has a following preprocess:

The removal of URL: Many users add some related URLs to the content of microblog when they release it. The websites that such URLs redirecting to have various and complex structure and content. In this project, we decide not to analyze these redirecting websites completely since these websites could hardly generate project-related values and almost equal to useless data. We finally use regular equation to match up these potential URLs and remove or replace it: "((http|ftp|https):/)(([a-zA-Z0-9_-]+ \. [a-zA-Z]{2,6}))(([/0-9]{1,3} \. [0-9]{1,3} \. [0-9]{1,3})|([0-9]{1,4})*(/[a-zA-Z0-9\&%_\.-~]*)?)"

The cut of forwards: According to the Sina Microblog's UI distribution, the original microblog would be located in an area other than the area with forwards' user information. Under this condition, other forwards' information could be seen as a noise, which might interfere the inference of

current user's location that needs to be removed. In this project, we try to use “//@” to match and find the start position of the forwards and remove the content between it and the end of the microblog. The remains are the useable microblog with the values to be used in inference.

The removal of stop words: There are considerable numbers of stop words in the content of our data sets, like function words “的(of)”, “吗(may)” and pronouns “你(you)”, “我(I)”, “他(he)”, “她(she)”. These words are not helpful to extract features of texts and they should be removed. In this project, we set the stop word dictionaries from websites as a standard for matching up and remove stop words in microblog texts.

3.2 | Released Texts Content driven Location Inference model (RCT-LI)

Distinguished by research directions, the mainstream of location inference on social media is various from each other. Researcher define these two problems as the following:

1. Locating social media users' permanent location as detailed as possible.
2. Locating geolocation that specific texts have released by users.

In this paper, we propose the RTC-LI (Released Texts Content driven Location Inference) model, a potential efficient solution for the above problems. Taking existed dictionaries into consideration, this model has analyzed users' information and released texts, inferring comparatively accurate geolocation data. In texts released by users, there are always some characteristic words that contain geographical information of texts, referring to areas in reality. Because of their locality, these words are called local words. Local words have distinctively various types, and typical classes includes cities, streets, towns, buildings, and so on.

3.2.1 | The design and implementation of local dictionaries

The initial job for RCT-LI is to verify if a word is a local word for a specific area, belonging to a local dictionary W_{GL} or not. We need to set an equation to quantify the correlation degree between the word and areas, in some words, to quantify the local value of the word.

We regard α_w as the local value. The smaller the distance between current location and destination is, the more frequent word w is and the bigger α_w is. Besides, we set T_w as the set of texts contain geographical tags, d_{tw} as the distance between current locations and destinations. Consequently, the maximum log likelihood for destinations is

$$f(\mathbb{C}_w, \alpha_w) = \sum_{t \in T_w} \log(\mathbb{C}_w \times d_{tw}^{-\alpha_w}) + \sum_{t \notin T_w} \log(1 - \mathbb{C}_w \times d_{tw}^{-\alpha_w}). \quad (1)$$

In the equation, α_w is the local value of words, representing the correlation degree between the words and the areas. \mathbb{C}_w is a constant and a correction parameter for the equation to acquire the local maximum. Related research infers that there must have and only have a local maximum on the domain. When the equation $f(\mathbb{C}_w, \alpha_w)$ gets its local maximum, the pair of values (\mathbb{C}_w, α_w) is the optimum solution.

After getting the local values of each word w for different destinations, according to the preset threshold, we select n largest words to form local dictionaries. After that, we use part of the test data for evaluation and accordingly change the value of threshold to acquire the optimum of such a process.

However, considering the one-sidedness of released texts and the amount of training data, we import the existed and completed dictionaries from the Internet as a correction for the previous calculation.

In this project, we use the dictionaries from Sogou input method. Sogou is one of largest input method in China, having most users in mobile devices. Up to the third quarter of 2016, China's third party mobile phone input method has reached 605 million users, a quarter increase of 2.3%. Among them, the market share of Sogou input method is as high as 71.2%. It also gets the highest score in many dimensions, such as brand name, product function, and user experience, and becomes the market leader. And Sina Weibo, our data resource has quite a percentage of users from mobile terminal. As a result, the dictionaries of Sogou could somehow represent the textual and geographical features of users.

For word w' in text t , if it belongs to Sogou dictionaries W' while it does not belong to generated dictionary W_{GL} , its local value is

$$\alpha_{w' \in t, w' \in W'} = \frac{\text{count}(w', t)}{\text{length}(t)}. \quad (2)$$

While $\text{count}(w', t)$ represent the frequency of word w in text t .

3.2.2 | User permanent location inference

The process of user permanent location inference is the following:

Parameters:

T_u : the set of texts for user u

S_u : the social relationship for user on Sina Weibo

W_{GL} : local dictionaries

L : the list of destinations

δ : the threshold of α

$$Avg_u(\alpha) = \sum_{w \in W_{GL-u}} \frac{a_w}{|W_{GL-u}|}$$

If $Avg_u(\alpha) > \delta$ then

Coord(u) = GeoCenter(W_{GL-u})

else

for $l_i \in L$

$$S_{likelihood}(l_i | u_{GLocalWords}) \sim 0$$

for $w \in W_{GL-u}$

$$S_{likelihood}(l_i | u_{GLocalWords}) = S_{likelihood}(l_i | u_{GLocalWords}) + S_{likelihood}(l_i | w)$$

$$S_{likelihood}(l_i | u_{socialnetwork}) = \text{Distribution}(S_u, l_i)$$

$$S_{likelihood}(l_i | u_{socialnetwork}) = \text{Mix}(S_{likelihood}(l_i | u_{GLocalWords}), S_{likelihood}(l_i | u_{socialnetwork}))$$

End if

From the process of local dictionaries creation, we could get for w in local dictionary W_{GL} , the possibility for user u in city l_i is

$$S_{likelihood}(l_i | w) = C_w \times d_{l_i w}^{-\alpha_w}. \quad (3)$$

For word w not in dictionary W_{GL} but in Sogou dictionary W' , the possibility for user u in city l_i is

$$S_{likelihood}(l_i | w') = d_{l_i w'}^{-\alpha_{w'}}. \quad (4)$$

For all $w \in T_u$, the possibility for user u in city l_i is

$$S_{likelihood}(l_i | u_{GLocalWords}) = \prod_{w \in T_u} S_{likelihood}(l_i | w) + \prod_{w' \in T_u} S_{likelihood}(l_i | w'). \quad (5)$$

We also consider about the 2-hop social relationship of users. For user u and its related user u' , the possibility for user u in city l_i is

$$S_{likelihood}(l_i | u_{socialnetwork}) = \frac{\text{Count}(U_o, l_i) + \text{Count}(U_A, l_i) + \text{Count}(U'_A, l_i)}{|U_A| + |U_o| + \sum |U'_A|}. \quad (6)$$

While U_A is the number of follows of user u , U_o is user u 's following users, and U'_A is the number of follows of user u' . $\text{Count}(U_o, l_i)$ and $\text{Count}(U_A, l_i)$ represent the numbers of users in U_A , and U_o are in city l_i .

Combining the abovementioned equation, we set a correction parameter to evaluate the hybrid model

$$S_{likelihood}(l_i | u) = \lambda \cdot S_{likelihood}(l_i | u_{GLocalWords}) + (1 - \lambda) (S_{likelihood}(l_i | u_{socialnetwork})). \quad (7)$$

3.2.3 | Texts released location inference

For city l_i and word w belong to dictionary $w_i \in (W_{GL} + W')$, we set $S_{likelihood}(l_i | w_i)$ to represent the weight for word w_i to city l_i . If word w does not belong to dictionary w_i , to avoid the condition of underflow, value it as $\min S_{likelihood}(L | W)$. In this way, the weight for word w_i to city l_i is

$$Q_{w,i} = \begin{cases} S_{likelihood}(L | W), & w_i \in (W_{GL} + W') \\ \min_{w_i \in W_{GL}} \{S_{likelihood}(L | W)\}, & \text{else.} \end{cases} \quad (8)$$

The possibility for specific text released in city l_i is

$$L(t) = \arg \max_{l_i \in L} p(l_i | \theta_L) \prod_{w_m \in t} p(w_m | l_i; \theta_L) \times Q_{m,i}. \quad (9)$$

While θ_L is the location-related parameterized model, $Q_{m,i}$ is the weight for word w_i to city l_i .

4 | EXPERIMENTAL RESULTS AND ANALYSIS

4.1 | Data set

Before the operation of preprocessing data and inferring the location, I firstly analyzed the raw data crawled from the Sina Weibo to explain the meaning of the inference algorithm based on the location of the user-generated contents. If there is a considerable proportion when the user's blog

TABLE 1 Location information to fill in personal information and the distribution of post geotagged

Location Level	Number of Users
Prefecture-level city	720
Province	246
Default	176
Total	1142
Type	Number of Blogs
Blogs geotagged	5211
Total	410 331

with a location label, then the proposed algorithm has a little meaning in this paper. We can totally infer user's location according to Weibo's geotag. However, if Weibo's data present sparsity, then the algorithm in this paper is a viable solution. At the beginning of choosing the user, considering the position inferred training process that requires a certain amount and a personal style blog language as a data set, we should choose the users who have a certain influence on social media and then choose their friend network too. At the same time, take into account the identity of the region to which the user belongs, and in order to avoid discrete, we should avoid choosing user group that has a big mobility, such as actors, singers, etc. Taking into account the blog's theme should be more to life and avoid tending to the user's own occupation, so do not choose opinion leaders and known from media. Finally, I chose a radio host as the beginning user. As of this report, we collected a total of 7994 personal information blocks and 410 331 blogs of 1142 users. Personal location information and blog post with a location tag are shown in Table 1. So, it is not difficult to draw the conclusion: the location data of Sina Weibo are very sparse, which cannot be inferred using the direct position and reference.

4.2 | RTC-LI algorithm results and analysis

4.2.1 | Construction of the local lexicon

A local dictionary is constructed by training 5211 blogs with geotag. Use ICTCLAS data dictionary to handle Chinese word segmentation and match the segmentation results. After getting data, then construct the local dictionary. First for each word, give it a value that contains the location of the blog, which contains the word, in other word, stats the location that each word appears. After the extraction of numerical and match of segmentation leads to a segmentation-geographical coordinates of index. In addition, the geographic location that each word appears is relative to the complement of all blog posts' location and has some relations between the word, which leads to a segmentation-non-geographical coordinates of index. The result is shown in Figures 1 and 2.

We have 33 provincial-level administrative units across the country as the default location of segmentation geographic distribution, and the construction of local dictionary size is 33. Logarithms of maximum likelihood values within a certain range of the image are shown in Figure 3.

Segmentation for 33 of the preset position of the logarithm of the maximum likelihood value part of the optimal solution of alpha values is shown in Figure 4.

Choosing an alpha value for each preset position descending order the top 20 Word constitute the region's local lexicon. Shanghai local dictionary word cloud is shown in Figure 5.

After the statistics, all the alpha values of the word in the dictionary is shown in Table 2.

It can be seen from Table 2 these 10 words for the city name, town name, places of interest, and proper nouns with a landmark geographical information, and their calculations focus in the location that also corresponds to the geographical area. "Pudong" usually refers to the districts of Shanghai, its focus is also displayed within its executive in Shanghai. So tell us, the local words to some extent to display the location, design, and construction of the local dictionary have some accuracy and reference significance.

```

1 Old,121.315552 31.192734,121.315552 31.192734,121.315552 31.192734,116.306777 39.982997,116.306777
2 official,120.4041 36.2064,120.4041 36.2064,2.325601281 48.8665827,2.325601281 48.8665827
3 sometimes,116.374826 40.080349
4 ago,116.34736 40.03062,116.34736 40.03062,116.34736 40.03062,120.0004 31.81315,120.0004 31.81315
5 once 114.504493 36.573295,114.504493 36.573295,114.504493 36.573295,114.504493 36.573295,114.5044
6 Kao 16.332695 39.863029
7 Wujiaochang,121.529831 31.30316,121.529831 31.30316,121.529831 31.30316,121.529831 31.30316,121.
8 firework,117.07901 33.627407
9 Zhe,108.934258 34.218021,108.934258 34.218021,108.934258 34.218021,116.374826 40.080349,116.3748;
10 Occupied,116.32564 40.07979,116.32564 40.07979,116.38996 39.99563,116.38996 39.99563

```

FIGURE 1 Part segmentation-geographical coordinates of index

1 Old, 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734, 121.
2 official, 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734,
3 sometimes 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734,
4 ago , 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734,
5 once, 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734,
6 Kao , 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734, 121.
7 Wujiaochang, 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734,
8 firework , 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734,
9 Zhe, 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734, 121.
10 Occupied, 121.330551 31.19298, -123.113503 49.288673, 121.315552 31.192734, 121.315552 31.192734,

FIGURE 2 Part segmentation-non-geographical coordinates of index

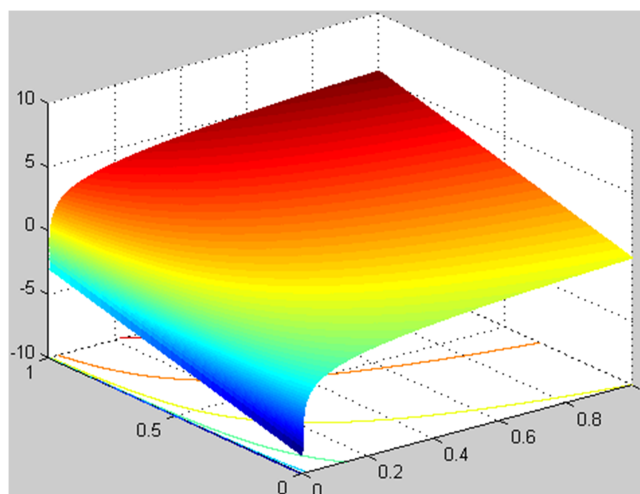


FIGURE 3 Logarithms of maximum likelihood values in the domain of image

```

1 ago,0.275000000000000013,0.995000000000000008,0.54500000000000004,0.83500000000000006,0.71000000000000005,(
2 once,0.2250000000000000012,0.995000000000000008,0.45500000000000003,0.86500000000000007,0.64000000000000005,
3 Wujiaochang,0.250000000000000001,0.995000000000000008,0.53500000000000004,0.82500000000000006,0.7050000000000000
4 exclusive interview,0.255000000000000001,0.995000000000000008,0.54500000000000004,0.84000000000000006,0.71
5 prepare,0.250000000000000001,0.995000000000000008,0.53500000000000004,0.85000000000000006,0.7050000000000000
6 one,0.995000000000000008,0.0,0.995000000000000008,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,
7 DiqingTibetan Autonomous Prefecture,0.260000000000000001,0.995000000000000008,0.54500000000000004,0.83500000000000006,0.71000000000000005,
8 value,0.275000000000000013,0.995000000000000008,0.54500000000000004,0.83500000000000006,0.71000000000000005,
9 Qilu,0.260000000000000001,0.995000000000000008,0.54500000000000004,0.84000000000000006,0.87000000000000007,
10 storage,0.300000000000000016,0.995000000000000008,0.70000000000000005,0.81000000000000006,0.6900000000000000
11 question,0.185000000000000008,0.995000000000000008,0.65000000000000005,0.99500000000000008,0.6800000000000000

```

FIGURE 4 Part of the optimal solution alpha values for preset position

4.2.2 | User location inference

After based on user-generated content and user relationships presumption of mixture model training, experimental results that were accurate and recall rates were shown in the Figure 6.

Figure 6 shows that under the same conditions, based on accurate inferred user location, user-generated content and success rates are relatively high, and the possible reasons are that the amount of data is used to build the local dictionary smaller relative to the test set to 5,211:405,120. Training with a user in the collection characteristics of the word number is still less, which is unable to fully reflect the correlation between a user's language and location. And accurate rates have certainly improved than recall rates and the possible reasons are that the existing derived word, to some extent, better reflects the characteristics of the location where your users reside to do good user inferences. Models that are based on the user's network of friends is better probably because crawling user social relations remain small, which focused mainly on its own area. This model uses the location for all Twitter users' inferred effect may be subject to a certain impact. Mixed model absorbed the advantages of both, to a certain extent make up for their deficiencies, achieved the best results in three of them.

4.2.3 | Text generation location inference

Tolerance for different results, text generation location inference results are shown in Table 3.



FIGURE 5 Shanghai local dictionary word cloud display

TABLE 2 Alpha and the focus position of the highest parts of words

ord	Longitude (°)	Latitude (°)	Province
Pudong	121.48	31.01	Shanghai
Hangzhou	120.20	30.27	Zhejiang
Beijing	116.42	39.92	Beijing
Baiyun	113.38	23.51	Guangdong
Beijing east	116.42	39.82	Beijing
Hutong	116.42	39.92	Beijing
Xihu	120.06	29.71	Zhejiang
Guilin	110.10	24.30	Guangxi
Wuchang	114.45	30.07	Hubei
Binhai	117.20	39.13	Tianjin

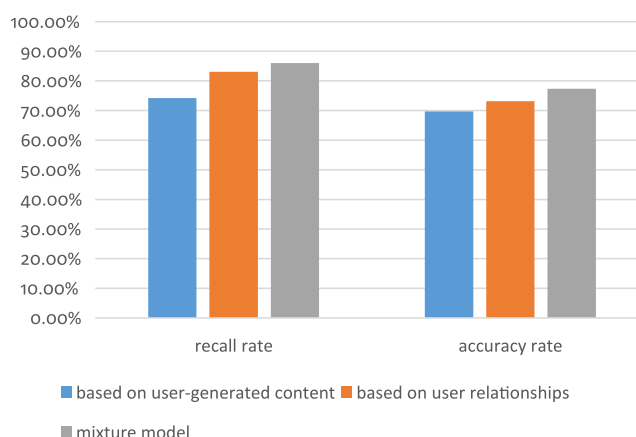


FIGURE 6 The recall rate and the accuracy rate of the user and the results of different models

TABLE 3 Text generation location inference results

Type	Result		
	Right belongs to provinces	Provinces that properly belongs within a distance of 1	Provinces that properly belongs within a distance of 2
Accuracy rate(%)	66.8	83.4	96.8

5 | CONCLUSIONS

This subject is based on user social relationship of network, where Crawler gets quite number of Sina Weibo user information and blogs released Weibo. On Weibo content in the segmentation for localized degree of measure and assigned value and Sogou dictionary of amendment of assigned value, building out different geographical locations of local Word library and taking this through the RTC-LI algorithm for user location inferred and text generated location inferred of process have made the experiment effective. Of course, there still has some problems in this topic, such as training data are small and experimental training set is too small, which cannot meet all the model's measurement of the degree of local needs, leading to

not precise enough results of accuracy, and there may be a certain degree of error. In addition, small led to the one-sidedness of the results of the training set, part of experimental method and process is not guaranteed to all users and post data.

ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China (No 61440054 and No 61462004), Fundamental Research Funds for the Central Universities of China (No 216274213), National Natural Science Foundation of Hubei, China (No 2014CFA048), Outstanding Academic Talents Startup Funds of Wuhan University (No 216-410100003), Natural Science Foundation of Jiangxi Province (No 20151BAB207042), Youth Funds of Science and Technology in Jiangxi Province Department of Education (No GJJ150572, No GJJ160589, No GJJ160590, and No GJJ170481), and Outstanding Foundation of Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology (No JELRGBDT201709).

AUTHOR CONTRIBUTIONS

Zhibo Wang and Xiaohui Cui conceived and designed the experiments; Jinchao Qin performed the experiments; Linlin He and Yuechuan Guo analyzed the data; Yuechuan Guo and Lin Liu contributed materials and analysis tools; Jinchao Qin, Yuechuan Guo, and Zhibo Wang wrote the paper together.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ORCID

Zhibo Wang  <http://orcid.org/0000-0001-5815-6567>

REFERENCES

1. Eisenstein J, O'Connor B, Smith NA, Xing EP. A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; 2010; MIT Stata Center.
2. Hecht B, Hong L, Suh B, Chi EH. 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 (CHI '11); 2011; Vancouver, BC.
3. 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; 2008; Beijing, China.
4. Cheng Z, Caverlee J, Lee K. A content-driven framework for geolocating microblog users. *ACM Trans Intell Syst Technol*. 2013;4(1):2.
5. Ryoo KM, Moon S. Inferring Twitter user locations with 10 km accuracy. In: Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion); 2014; Seoul, South Korea.
6. Chandra S, Khan L, Muhaya FB. Estimating Twitter user location using social interactions—a content based approach. Paper presented at: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing; 2011; Boston, MA.
7. Jurgens D. That's what friends are for: inferring location in online social media platforms based on social relationships. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM); 2013; Cambridge, MA.
8. Li R, Wang S, Chang KC-C. Multiple location profiling for users and relationships from social network and content. *Proc VLDB Endow*. 2012;5(11):1603-1614.
9. McGee J, Caverlee J, Cheng Z. Location prediction in social media based on tie strength. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM '13); 2013; San Francisco, CA.
10. Kotzias D, Lappas T, Gunopulos D. Addressing the sparsity of location information on Twitter. In: Proceedings of the Workshops of the Joint Conference of the 17th International Conference on Extending Database Technology and the 17th International Conference on Database Theory (EDBT/ICDT); 2014; Athens, Greece.
11. Rongjiao Z, Sansheng C. A method of recommendation based on user relationship. Paper presented at: 11th Conference on Internet and Video Broadcasting Development; 2012; Wuhan, China.
12. Chi G, Jingnan L, Yuan F, Meng L, Jingsong C. Value extraction and collaborative mining method of location big data. *J Softw*. 2014;25(4):713-730.
13. Jingnan L, Yuan F, Guo C, Gao K. Research on the analysis and processing of big data. *Inf Sci Ed Wuhan Univ*. 2014;39(4):379-385.
14. Wang K, Yu W, Yang S, Wu M, Hu YH, Li SJ. Location inference method in online social media with big data. *Ruan Jian Xue Bao J Softw*. 2015;26(11):2951-2963.
15. Wu X, Yu K, Ding W, Wang H, Zhu X. Online feature selection with streaming features. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(5):1178-1192.