Estimating Home Location of Foreigners in Japan Using Photograph Location

Masaharu Hirota
Department of Information Science, Okayama University of Science
1-1 Ridaicho, Kita-ku, Okayama-shi, 700-0005, Japan

Tetsuya Oda
Department of Information and Computer Engineering, Okayama University of Science
1-1 Ridaicho, Kita-ku, Okayama-shi, 700-0005, Japan
E-mail: hirota@mis.ous.ac.jp, oda@ice.ous.ac.jp

Abstract

The attributes of travelers such as home location and age could be useful for many applications such as information recommendation and targeted advertisement. However, this information is not accessible in most Web services because users do not reveal them. We propose a method to estimate a foreign tourist's home location based on each region's tendency, in which tourists from a region tend to visit certain places when traveling abroad. The feature for the estimation uses the frequency of photograph location in a geohash. In this paper, we use foreigners in Japan as a case study. We evaluate the performance of our proposed method by using photographs obtained from their user accounts on Flickr.

Keywords: home location estimation, user trajectory, machine learning, Flickr.

1. Introduction

With the rapid growth of social media websites such as Flickr or Twitter, it has become common for tourists to upload geo-tagged content from their trips. The contents are recommended to the other users based on user information and its friendship network. Therefore, the content strongly influences tourists' decisions, such as their travel destinations and routes, because tourists may refer to this information when they plan their trips.

The user attribute of users on social media websites (e.g., home location, age, and gender) is important for many applications. The user's home location is especially used for targeting advertisement, information recommendation, and user behavior analysis. However, many users on social media websites do not provide information on the home location. It is also possible that the user provides false information.

To estimate the user's home location, we can use the frequency of geo-tags annotated to the user's content. For example, a simple way is to regard the area with the most geo-tags indicating a particular area as the user's home location. However, this method's limitation is that it cannot be applied when the analyst only has data for a specific region. For example, it is impossible to estimate the country of residence outside of Japan if the analyst only has data posted in Japan. Therefore, a method to estimate the home location from a dataset for a particular region only is necessary.

This paper proposes a method to estimate the user's home location based on geo-tags visiting from abroad using contents obtained from a specific area. Our hypothesis is that travelers in different home locations visit different areas. Some places are widely popular tourist spots, like the Eiffel Tower in France and the

Leaning Tower of Pisa in Italy. However, when travelers go abroad, tourists from a particular country may be interested in specific places, not necessarily conventional tourist spots. These are places associated with a specific region, such as a famous movie in their origin region, a restaurant whose owners are from their region, or a place featured on a local TV show. If we can extract the tendency of the user's visitation by home location, we assume that it is possible to estimate the user's home location from geo-tags.

In this paper, we estimate the region of residence of users who come to Japan from abroad. We used geo-tagged photographs obtained from Flickr. We evaluate the performance of our proposed method by estimating the home location of users in Flickr. Also, we discussed the effect of area size of geohash on the estimation performance.

This paper's remainder is organized as follows: Section 2 describes the work related to this topic. Section 3 presents our proposed method for estimating the user's home location. Section 4 presents the experiment conducted to evaluate the performance of our proposed method. Finally, Section 5 concludes our work and discuss results and areas of future work.

2. Related Work

A wide range of home location methods estimating from social media websites content has been proposed in recent years.

Hironaka et al. used data from Twitter to analyze users' home location based on friend relationships[1]. Hu et al. proposed a method to estimate home location from sparse and noisy Twitter data within 100 by 100-meter squares at high accuracy using users' trajectories in their home country[2]. Jurgens et al. evaluated several methods for geo-location prediction using data from Twitter[3].

The above methods focus on estimating users' home location of social media websites by using content posted in their home country. These researches estimate the home location of users within the posted range of the data. Therefore, it is difficult to estimate the home location of users visiting from other regions. Our previous research[4] proposed a new method for home location inference by analyzing foreign tourists' tendencies in a different region from their home country. In this research, we proposed a method for estimating the user's home location and evaluating the method's performance.

3. Proposed Method

Our proposed method consists of two steps as follows.

(i). Extracting features based on occurrence of users' visits from user's contents.

(ii). Estimating the home location of a tourist from their trajectory using a machine learning algorithm.

3.1. *Feature extraction*

We describe the generation of features for machine learning.

In this study, we use geohash generated from a photographic location as a user's feature. A geohash is a string representation of a geographical area, created by encoding latitude and longitude. Geohash converts a point expressed in latitude and longitude into a geographical area, which reduces the accuracy of representing a geographical point. In addition, depending on the string's length of geohash, it is possible to control the degree of accuracy loss. For example, the latitude and longitude of Okayama Castle (latitude, longitude) = (34.66568196, 133.93593695) is converted to 'wypjpy9' in 7-characters geohash.

To create features, we count the number of photographs taken in each geohash for each user. Then, we delete geohashs with fewer than ten users who took the photographs. Finally, we apply L2 normalization to the features.

3.2. *Estimating Home Location*

In this research, we use Random Forest[5] (RF) to estimate the user's home location based on feature vectors obtained from Section 3.1. This algorithm is a multi-class classifier algorithm in supervised learning. In this research, we regard the predicted class from the feature vector obtained from a user's photostream as the user's estimated home location.

4. Experiment

In this paper, we evaluate the performance of our proposed method through an evaluation experiment based on classification. We describe the experiment conditions of the dataset and the evaluation criteria.

As described in Section 3.1, the geohash allows controlling the degree to which latitude and longitude are quantized depending on the string's length to be converted. Therefore, our experiment evaluates the effect of geohash string length on classification performance.

4.1. *Dataset*

In this section, we describe the dataset used for evaluating our proposed method. We used Japan as a particular country for this experiment. We estimate the home location of users who visit from other regions and using the photographs taken in Japan.

The evaluation experiments used photographs obtained from Flickr. The photographs include the metadata: latitude, longitude, and home location. The home location is the text data where the user in Flickr describes their home location.

Because Flickr's home location field is an open text field, the data in it varies widely. Therefore, we mapped all the abbreviations and significant states/cities to the same country name (i.e., "usa", "u.s.a" and "Dallas, Texas" were mapped to "usa").

We excluded photographs where either the latitude or the longitude does not have a value with a precision less than or equal to the third decimal place. We also exclude users who describe more than one home location in their photographs and users who have less than ten photographs.

In this research, we used three home location: Taiwan, America, and United Kingdom. This reason is that the number of users in which we obtained photographs from Flickr is Taiwan, USA, and UK. Those regions are the top three regions in the number of users in our photographs obtained from Flickr. Therefore, our method tackles three-class classification.

Consequently, the number of photographs used in this experiment was 461,454 and the number of users is 1,683. Also, we randomly split into a ratio of 9:1 for train and test data.

4.2. *Evaluation criteria*

We used the following widely used performance measures for classification: Accuracy, Recall, Precision, and F-score. In this research, our classification is multi-class classification. Therefore, we used the macro average of them.

Table 1. Evaluation result.

Geohash	Precision	Recall	F1-score	Accuracy
3	0.499	0.494	0.463	0.552
4	0.543	0.461	0.426	0.554
5	**0.716**	**0.529**	0.455	0.581
6	0.537	0.527	0.479	0.625
7	0.597	0.522	0.491	**0.680**
8	0.570	0.513	**0.515**	0.649

4.3. *Experimental conditions*

This section describes the procedure used for the classifiers to estimate home locations.

This experiment used entropy and Gini impurity for a split of nodes in RF. In addition, the hyperparameters such as max nodes and the number of trees were searched using Optuna[6] with five cross-validations, which is a software for automatically optimizing hyperparameters. We used the parameters with the highest accuracy measured through this experiment. In addition, we used the Python software scikit-learn[7] for the implementation of RF.

4.4. *Evaluation results*

This section describes and discusses the evaluation results of classifying home locations and the effects on string length of geohash.

Table 1 shows the evaluation results of Accuracy and macro average of Recall, Precision, and F-score. The first column represents the string length of geohash. The best value of the length in Precision and Recall is 5. In this case, the UK's Precision value was less than 0.5 when other parameters were used, but it was 1.0 when this parameter was used. The Recall values for the other two regions were also high. Therefore, the result of this parameter may overfit to the UK. Also, in Table 1, the larger the value of the length of geohash, the larger the value of Accuracy. The proposed method may have the best classification performance when the length of the geohash string is 7.

Table 1 shows that the proposed method works well because some performance is obtained for each evaluation criterion.

5. Conclusion

In this paper, we proposed a method to estimate the home location using the user's trajectory. Our approach uses

the tendency of users from the same location to travel to the same places. Also, our approach uses the geohash obtained from the photograph location. Experimental results showed that our classifier could estimate the candidates of home location.

Our future work will include a more detailed experiment. This paper is not enough in the analysis of the home location predicted by our proposed method. Also, the experiment in this paper uses a few photographs. As a result, the experiment result has the limitation of the validity of our proposed method's effectiveness. Therefore, another further study may make the data which includes more number of a dataset.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP19K20418 and Grant for Promotion of OUS Research Project (OUS-RP-20-3).

References

1. Shiori Hironaka, Mitsuo Yoshida, Kyoji Umemura, User's Centrality Analysis for Home Location Estimation. Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, 14–17, 2019.
2. Tianran Hu, Jiebo Luo, Henry Kautz and Adam Sadilek, Home Location Inference from Sparse and Noisy Data: Models and Applications. Proceedings of IEEE 15th International Conference on Data Mining Workshops, Atlantic City, 14-17, 2015.
3. David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, Derek Ruths, Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. Proceedings of the Ninth International AAAI Conference on Web and Social Media, 26-29, 2015.
4. K. Lugasi, M.Hirota, Inferring Home Location of Foreign Tourists Based on Travel Routes Extracted from Social Media Sites, The 2020 International Conference on Artificial ALife and Robotics, 2020.
5. L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.
6. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, 2019.
7. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.