Inferring Home Location of Foreign Tourists Based on Travel Routes Extracted from Social Media Sites

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

Masaharu Hirota
Department of Information Science, Okayama University of Science
1-1 Ridaicho, Kita-ku, Okayama-shi, 700-0005, Japan
E-mail: i16i090lc@ous.jp, hirota@mis.ous.ac.jp

Abstract

Tourists from certain regions tend to visit certain places when traveling abroad. The availability of large amounts of data from social media sites allows researchers to profile their tendencies, which could be useful for many applications such as route recommendations or advertisements. We propose a method to infer the home location of a tourist, based on their travel route, using the extracted tendency for each country. Our approach uses supervised machine learning methods to learn the quantized travel route of each user. In this paper, we use foreign travelers in Japan as a case study. We evaluate the performance of our proposed method by using photographs collected from their user accounts on Flickr.

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

1. Introduction

With the rapid growth of social media services, it has become common among tourists to upload geo-tagged contents from their trips to online services, which led to an overflow of available information about tourist attractions. Users of those services are being constantly exposed to the same tourist attractions as their friends, which affects their decisions when planning their next trip, whether it is domestic or international.

Some places are widely popular tourist spots, like the Eiffel Tower in France and the Leaning Tower of Pisa in Italy. However, when traveling abroad, tourists from a particular country may be interested in specific places, which are not necessarily conventional tourist spots. For instance, places where a famous movie in their origin country was filmed, a restaurant whose owners are from their country, or a place that was featured on a local TV show.

By extracting only the differences in travel route between travelers from various origin countries, we assume that it is possible to create a tendency profile that can be used to predict the home location of a tourist based on their travel route, and vice versa.

Knowing a tourist's home location has many applications. For instance, it can be used to improve and personalize travel route recommendation and targeted advertisement for tourists. However, this information is not accessible for the most part. Therefore, there is a need to infer it based on other information. We propose a method to do that by finding the difference in traveling routes between tourists from different countries, based on their publicly-shared contents.

The remainder of this paper is organized as follows: Section 2 describes the work related to this topic. Section 3 presents the methodology used. Section 4 presents the experiment conducted in order to verify our methodology. Finally, Section 5 concludes this paper and proposes future research directions.

2. Related Work

A wide range of methods for inferring home location from social media content has been suggested in recent years.

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

While the above methods focus on inferring home location of social media users by using contents posted in their home country, in this study, we propose a new method for home location inference by analyzing the tendency profile of foreign tourists using travel routes in a country that is different from their home country.

3. Methodology

Our methodology relies on three consecutive steps; extraction of tourists contents (from now on will be referred to as "photo streams"), and dividing into individual visits (first step), clusters creation which will be used for quantizing individual users' trajectory (second step), and inferring the home location of a tourist from their trajectory using machine learning algorithm (third step).

3.1. *Extracting Tourists' photo streams*

To retrieve photographs from users who are not residents of the target country, we extract photo streams that are confirmed to be of a certain length of stay and show mobility patterns that indicate visits to different places in a short period of time.

Because a user may visit the same country several times in a given period, we, therefore, define one visit as being at least three days long and no longer than the standard for tourists' stay permit in that country. The interval between two visits is defined to be at least four weeks because it is unreasonable for a tourist who usually takes and uploads photographs to not take photographs for more than that period.

For each visit, it is required that it contains at least two photographs that have different GPS data (taken in different places).

3.2. *Creating Clusters*

In order to extract a user's trajectory, it is necessary to define areas that can be referred to as a single place between which the tourists can travel.

We create two types of clusters based on the obtained data; The first cluster combines photographs taken in the same region, generated from the photograph's longitude and latitude compared against a predefined regions list.

The second cluster is obtained by using grid clustering. The country is divided into an MxN grid with cells of about 6x6 kilometers width and height. Adjacent cells that meet a certain threshold are merged together up to a width of about 60x60 kilometers. The threshold definition is illustrated in Eq. (1).

$$threshold = \frac{\text{sum(unique users in a cell)}}{\text{total cells with data}} \qquad (1)$$

3.3. *Inferring Home Location*

In this study, we use two types of classifiers; Long Short-Term Memory (LSTM) [4] and Support Vector Machine (SVM) [5]. We perform classification using all-versus-all method in LSTM and one-versus-all method in SVM. The feature vectors for those models uses quantized user trajectories as described in Section 3.2.

4. Experiment

4.1. *Dataset Preparation*

Flickr API[a] was used to collect metadata from publicly posted photographs with Global Positioning System (GPS) location data in the target country.

For the case study, data from Japan between April 1, 2009 and April 1, 2019 was used. We filtered the results to contain only users whose home location (owner

[a] https://www.flickr.com/services/api/

Table 1. Users, Visits and Photographs for each country used for the experiment.

Country	Users	Visits	Photographs
United States	1,490	2,222	168,680
Taiwan	966	2,044	351,333
United Kingdom	674	910	75,252
Australia	460	660	54,045
France	424	574	34,293
Canada	369	518	42,361
Italy	298	374	18,766
China	290	492	22,958
Spain	284	380	25,759
Germany	253	351	21,179
Singapore	226	381	32,298
Hong Kong	195	440	42,427
Netherlands	155	210	14,189

location, in Flickr API terms) is in a country other than the target country.

In order to decrease the negative influence of inaccurate data on our results, we only used photographs that contained GPS data accurate to the millionth decimal place, and photographs with the highest timestamps granularity.

Because "owner location" field in Flickr is an open text field, the data in it varies widely. To cope with that, we mapped all the abbreviations and major states/cities to the same country name. For instance, "usa", "u.s.a" and "Dallas, Texas" were mapped to "United States".

The final result yielded 1,084,645 photographs generated by 8,286 users from 102 countries. However, only countries with more than 200 visits were included, as indicated in Table 1.

4.2. Preliminary Experiment

In order to confirm the existence of variation between users from different locations, we first conducted a preliminary experiment.

The target country's whole area was divided into a 50x50 grid. Photographs of users from specific home locations were mapped into the grid. As a result, each cell in the grid contained the number of unique user visits. The results were normalized and represented as a 50x50 matrix. Then, we performed a matrix subtraction with other origin countries.

We observed a significant difference between different origin countries, a difference that remained about the same regardless of the grid size, indicating its significance.

These results were then compared to the average difference between each origin country's users to confirm that they were indeed larger.

4.3. Evaluation Method

For LSTM evaluation F1-score was used. For SVM a one-versus-all model was created for each country, and therefore we evaluated each model while accounting for the collective result.

We used each of the models' prediction probability to rank them in descending order. As a result, the model with the best probability for a prediction candidate was positioned first in the list. Because tourist data is extremely noisy, we allowed up to 3% of measurement error and therefore checked the next candidates in the range of 3% probability difference.

4.4. Results

Multi-class classification using LSTM showed poor results. Despite numerous optimization attempts varying from changing the number of feature vectors, to adjusting the clustering method and number of clusters, F1-score never crossed 0.3, which is considered low.

Results from SVM one-versus-all model are shown in Table 2. Out of 1,106 users in the test data, 168 users were classified with the best score by their home country's model, and 764 users were correctly classified with a measurement error of up to 3% compared to other country models. The rest 174 users were incorrectly classified with probabilities ranging from 10% to 100% error.

Table 2. Experiment Results Using SVM Classifier.

		Top Candidate		
		Yes	No	*Total*
Model Prediction	Correct	168	764	***932***
	Incorrect	14	160	*174*
	Total	*182*	*924*	*1106*

5. Conclusions and Future Work

In this paper, we proposed a method to infer the home location using a tourist's trajectory. Our approach used the tendency of tourists from the same country to travel to the same places. Experimental results showed that our classifier could estimate the candidates of tourist's home location.

In future work, we plan to expand the features to include not only the origin country but also other parameters such as gender, traveling season, and traveling form.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 16K00157, 16K16158, and Tokyo Metropolitan University Grant-in-Aid for Research on Priority Areas Research on social big data.

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, Thessaloniki, Greece, Oct. 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, NJ, USA, Nov.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, Oxford, England, May. 26-29, 2015.
4. Hochreiter, Sepp & Schmidhuber, Jürgen, *Long Short-term Memory*. Neural computation. 9. 1735-80. 10.1162/neco, 1997.
5. C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning, vol. 20, no. 3, 1995, pp. 273–297.