Next Place Prediction Using GPS traces and Web Search Queries

Ryo Imai
Tokyo Instutitute of Technology
Tokyo, Japan
imai@miubiq.cs.titech.ac.jp

Kota Tsubouchi
Yahoo Japan Corporation
Tokyo, Japan
ktsubouc@yahoo-corp.jp

Masamichi Shimosaka
Tokyo Institute of Technology
Tokyo, Japan
simosaka@miubiq.cs.titech.ac.jp

***Abstract*—Thanks to the popularity of GPS-enabled devices, destination prediction, meaning predicting the future location of users, has been investigated as a core technology for various applications in location-based services. For the last decade, predicting daily activities specific to an individual user, such as one's home and office (familiar destination prediction) is explored, whereas it exhibits difficulties in prediction for unfamiliar destinations such as for shopping on weekends and sightseeing activities. To resolve this limitation, we propose a new framework that exploits web search queries of users as well as GPS. This is inspired by the fact that users tends to perform web searches related to their unfamiliar destinations. To the best of our knowledge, our model is the first attempt that deals with web search queries (connecting natural language processing) and location-oriented research. The experimental results using over 670 users from commercial services show that our proposed method achieves better prediction performance in comparison with the state-of-the-art approaches.**

***Index Terms*—Destination prediction, User location, Next place prediction, familiar destination prediction, unfamiliar destination prediction**

I. Introduction

The widespread use of smartphones and car navigation services has enabled access to numerous GPS logs, making it easier to predict human activity based on GPS data. In a human activity prediction task, predicting the future location of users, i.e., destination prediction, can be applied to various practical applications, such as forecasting gathering events [1], and optimizing taxi dispatch [2].

Destination prediction can be categorized into two types based on the property of destinations: *familiar* destinations, which are high-frequency destinations specific to an individual user, such as home and office location, and *unfamiliar* destinations, which are the places the user may visit such as sightseeing locations or shopping areas over the weekend. Existing studies on destination prediction also led to categorization into two categories corresponding to the type of destinations: one focuses on individual trips [3], [4], whereas the other focuses on group flow [5], [6].

In the former approach based on individual trips, methods are designed for car navigation systems and map applications; the former approach can detect familiar destinations owing to the use of user-specific information such as frequently visited places and routes. Although the prediction accuracy for familiar destinations is relatively high, these studies have observed unfamiliar destination prediction infeasible owing to the difficulty associated with the prediction of low-frequency destinations such as a sightseeing location or a shopping area.

In the latter approach, population-scale people flow is employed to analyze trends of popular places; the population-scale people flow is also designed for advanced geographic information system (GIS) applications, such as optimizing taxi allocation on demand, and congestion analysis based on crowd-scale event detection. In contrast, familiar destination prediction is uncovered in the latter approach due to the lack of user-specific information. Previous studies in literature on destination prediction are suitable either for familiar destinations or for unfamiliar destinations.

In this paper, we focus on destination prediction for both familiar and unfamiliar destinations as the destination prediction for both types of destination evolutionary updates the user experience in car navigation systems, where users do not need proactive determination of one's destination for sightseeing places, and users are also automatically taught to avoid congested areas when returning home. Following from the trend of this field, the primary objective of this paper is to pursue a new destination prediction model that covers the both types of destinations. Instead of concentrating on GPS-based destination prediction, we are motivated to exploit information sources reflecting user intentions for an unfamiliar destination. In this paper, we newly employ web search queries of the user, such as what words are searched, into the destination prediction model for unfamiliar destinations. It is natural to think that users search topics is related to unfamiliar destinations prior to their trips.

To the best of our knowledge, our work is the first attempt to incorporate both web queries and GPS logs to provide accurate destination prediction. The main contributions of this paper can be summarized as follows:

- We propose unified prediction approach for handling both familiar and unfamiliar destinations that incorporates the unfamiliar destination model using web search queries into the familiar destination prediction model using GPS. We introduce a highly flexible context-aware combination prediction model.
- We construct a highly flexible prediction model to predict personalized unfamiliar destination using web search

queries as the information source reflecting user intention to go to the unfamiliar destinations
- The performance of the proposed model is evaluated using GPS and web search queries of more than 670 users obtained from the commercial service.

II. Related Work

As mentioned in the introduction, destination prediction researches can be divided into two types: one based on individual trips for personalized familiar destination prediction and the other based on people flow to understand the POI popularity on an urban scale.

The objective of research focused on individual trips is to develop applications for individuals, such as car navigation systems and map applications on smartphones. Most of these studies use the information obtained from GPS associated with user identifier.Owing to the association between the user identifier and GPS, familiar and user-specific destinations, such as home and office, can be detected. However, unfamiliar-destination prediction still remains a challenge because prediction is based on the frequency of trips to the destination in GPS logs.

On the other hand, studies focusing on group flow target applications where trends of many users' trips are important, such as optimizing taxi flow and detecting gathering events. In addition to studies on GPS-based prediction [1], [5], [6], there are several studies on POI popularity-based prediction.In these studies, it is assumed that users are attracted to places where many people gather. However, predictable destinations are limited in the methods proposed in these studies due to the fact that user's intentions are not reflected and only popular places are predicted.

III. Problem setting of destination prediction

The purpose of this study is to maximize the accuracy and extend predictable destination to both familiar and unfamiliar destinations by focusing on individual trips of users.

A. Data source for prediction for both types of destinations

In this section, we present the setting of the destination prediction problem. First, the information used for prediction is defined as a variable. As described in Section I, we use GPS and web search queries as the information sources. Let $\mathbb{U}$, $\mathbb{G}$, and $\mathbb{Q}$ be the set of users, GPS data, and web search queries (i.e., collection of words) in the dataset. Then, GPS data collected at timestamp $t \in \mathbb{T}$ from $u \in \mathbb{U}$-th user is denoted by $\boldsymbol{g}^{(u)}(t) \in [-90°, 90°] \otimes [-180°, 180°]$, e.g., $(35.5°, 139.8°)$, and web search queries at time t from the same user is also denoted by $\boldsymbol{q}^{(u)}(t)$, e.g., "Tokyo station bullet train". For each user, let $\mathbb{G}_u \subset \mathbb{G}$ be the set of GPS data, and $\mathbb{Q}_u \subset \mathbb{Q}$ be the set of web search queries for user u.

B. Preprocessing step toward problem setting

Next, users, trips, and destination candidates are extracted from the dataset. In this study, destination candidates are divided into *familiar* destinations, *unfamiliar* destinations, and *other* destinations. Familiar destinations are usually specific to each user, such as home and office, and unfamiliar destinations are popular candidates common to all users. Specifically, we define familiar destination as the destination where frequency of trip in one week is equal to or higher than the threshold, this threshold is given in the problem setting, and unfamiliar destination as the destination that is not a familiar destination and is included in the list of unfamiliar destination areas. This list of unfamiliar destinations should be chosen by the developer while considering which predictions are important in the application, e.g., for a transit application, stations should be chosen as unfamiliar destinations. Other destinations are neither familiar nor unfamiliar destinations. For example, when popular stations are chosen as unfamiliar destination candidates, POIs located far from a station or stations not popular will be defined as other destinations. The extraction of users, trips, and destination candidates is performed in eight steps. We describe the process in these eight steps in the following paragraphs.

1) Choosing unfamiliar destination: First, we choose some areas as unfamiliar destination candidates. This step is done manually, e.g., select 10 famous POIs. These areas are preferred for places where many people gather. Let $\mathbb{D}_{\mathrm{pub}}$ be a set of unfamiliar destinations: common for all users.
2) Discretizing GPS data: Next, we use DBSCAN clustering [7] for discretizing GPS data $\mathbb{G}_u$ of each user.
3) Labeling of GPS data as "staying" or "moving," and 4) Extracting destination candidates and trips: In the third and the fourth step, for each $\mathbb{G}_u$, we sort $\boldsymbol{g}^{(u)}(t) \in \mathbb{G}_u$ by its timestamp, and gather GPS data where the same cluster index continues. Using this sequence of GPS data, we label the sequences with duration exceeding the threshold as "staying;" the other sequences are labeled as "moving." For each "staying" sequence, we extract the GPS with the smallest timestamp as the destination and extract before one "moving" sequence and "staying" sequence as trip. Subsequently, we obtain some trips for each user. Let the trip whose starting time is t_s and arrived time is t_a be $\boldsymbol{S}^{(u)}(t_s, t_a)$. We define $\boldsymbol{S}^{(u)}(t_s, t_a)$ as $\boldsymbol{S}^{(u)}(t_s, t_a) = \{\{\boldsymbol{g}^{(u)}(\tau) | t_s \leq \tau < t_a\}, \{\boldsymbol{q}^{(u)}(\tau) | t_s \leq \tau < t_a\}\}$ Here, $\boldsymbol{g}^{(u)}(\tau)$ and $\boldsymbol{q}^{(u)}(\tau)$ are GPS and web search queries for u on timestamp τ.
5) Extracting familiar destination candidates and trips to familiar destination: In the fifth step, we set the threshold percentage of familiar destinations in past destinations to enable the extraction of familiar destinations. Moreover, for each user, we extract familiar destinations using this threshold. Let $\mathbb{D}_{u,\mathrm{freq}}$ be the set of familiar destinations for user u.
6) Extracting trips to unfamiliar destination: In the sixth step, we extract trips, whose destinations are included in the range of one of the unfamiliar destination candidates and are not familiar destinations, as unfamiliar destinations.

7) Extracting other destination candidates and trips to other destination: In the seventh step, for each user, we extract destination candidates which are neither familiar nor unfamiliar destinations, and trips whose destinations are included in these destination candidates. Let $\mathbb{D}_{u,\mathrm{other}}$ be the set of these destinations.

8) Extracting users: We obtain the set of destination candidates for u, $\mathbb{D}_u = \mathbb{D}_{\mathrm{pub}} \cup \mathbb{D}_{u,\mathrm{freq}} \cup \mathbb{D}_{u,\mathrm{other}}$ using the above steps. We set the threshold of data acquisition days, number of trips, and number of trips to unfamiliar destination and then extract users based on these thresholds. We describe specific values of thresholds for prepossessing on Section V.

C. Problem setting

In this study, the destination prediction problem can be formalized as a problem to predict where is the destination of the trip when an imperfect trip is given. As the destinations are classified as clusters, this problem is considered as a multi-class classification problem. Let an imperfect trip be $\boldsymbol{S}^{(u)}(t_s, t) = \{\{\boldsymbol{g}^{(u)}(\tau)|t_s \leq \tau < t\}, \{\boldsymbol{q}^{(u)}(\tau)|t_s \leq \tau < t\}\}$, where t is a current timestamp. Using $\boldsymbol{S}^{(u)}(t_s, t)$, destination prediction can be formulated as follows: $\hat{d} = \mathrm{argmax}_{\tilde{d}\in\mathbb{D}_u} p^{(u)}(\tilde{d}|\boldsymbol{S}^{(u)}(t_s, t))$. Here, $\hat{d}$ is the cluster index of predicted destination, and $\tilde{d}$ is the cluster index of destination candidates. In this paper, the prediction model peculiar to a user is $p^{(u)}$, and the probability model common among users is $p^{(0)}$. When $\hat{d}$ matches ground truth d, the destination prediction of the destination is correct. Our objective is to maximize the prediction accuracy and to extend predictable destination to both familiar and unfamiliar destinations. For simplicity, we represent $\boldsymbol{g}$ and $\boldsymbol{q}$ as $\{\boldsymbol{g}^{(u)}(\tau)|t_s \leq \tau < t\}$ and $\{\boldsymbol{q}^{(u)}(\tau)|t_s \leq \tau < t_s\}$, that is, $\boldsymbol{S}^{(u)}(t_s, t) = \{\boldsymbol{g}, \boldsymbol{q}\}$. Note that prediction based on GPS, that has been extensively studied in recent years can be formulated as follows: $\hat{d} = \mathrm{argmax}_{\tilde{d}\in\mathbb{D}_u} p^{(u)}(\tilde{d}|\boldsymbol{g})$.

IV. Proposed method: Predicting next familiar / unfamiliar places

We introduce the proposed method in this section. For destination prediction, it is important to consider whether the user is likely to go to the familiar or unfamiliar destination, and this depends on user's current context such as the time of the day. To consider that next destination is likely to be familiar or unfamiliar, we propose a combination model showcasing probability that the next destination will be a familiar destination. In this model, the probability of destination d is calculated with coefficient λ, which is the probability that the next destination will be familiar destination. We introduce the following two types of combination method, and will select one with the better result after comparison. One is linear interpolation (LI):

$$p^{(u)}(d|\boldsymbol{g}, \boldsymbol{q}) = \lambda p^{(u)}(d|\boldsymbol{g}) + (1-\lambda)p^{(0)}(d|\boldsymbol{q}). \quad (1)$$

The other is geometric interpolation (GI):

$$p^{(u)}(d|\boldsymbol{g}, \boldsymbol{q}) = \frac{p^{(u)}(d|\boldsymbol{g})^{\lambda} p^{(0)}(d|\boldsymbol{q})^{(1-\lambda)}}{\sum_{d'\in\mathbb{D}_u} p^{(u)}(d'|\boldsymbol{g})^{\lambda} p^{(0)}(d'|\boldsymbol{q})^{(1-\lambda)}}. \quad (2)$$

Whether next destination is likely to be familiar or unfamiliar destination should depend on context. This means that λ in (1) and (2) should be adjusted by current context such as the day, time, starting location, the number of web search queries. In this paper, we employ these information as context to estimate proper λ using bilinear logistic regression.

The GPS-based destination prediction method and the web search query-based destination prediction have to be chosen for our model. We chose multi-class logistic regression [8] for destination prediction based on GPS and extended the prediction model from latent Dirichlet allocation (LDA) [9] to predict the destination for web search query-based destination prediction.

A. GPS based destination prediction using multi-class logistic regression

To calculate the probability $p^{(u)}(d|\boldsymbol{g})$, we apply multi-class logistic regression on user contexts such as the day, time, and starting location, which were extracted from user logs [8]: $p^{(u)}(d|\boldsymbol{g}) = \frac{\exp(\boldsymbol{w}_d^{\mathsf{T}}\boldsymbol{f}(\boldsymbol{g}))}{\sum_{d'\in\mathbb{D}_u}\exp(\boldsymbol{w}_{d'}^{\mathsf{T}}\boldsymbol{f}(\boldsymbol{g}))}$. In this equation, $\boldsymbol{f}(\boldsymbol{g}) = \boldsymbol{f}^{(1)}(\boldsymbol{g}) \otimes \boldsymbol{f}^{(2)}(\boldsymbol{g}) \otimes \boldsymbol{f}^{(3)}(\boldsymbol{g})$, and $\boldsymbol{f}^{(1)}$ is a feature indicating whether that the day is a weekday or a weekend by one-hot encoding, $\boldsymbol{f}^{(2)}$ is a feature indicating the time, $\boldsymbol{f}^{(3)}$ is a feature indicating the starting point label, and $\boldsymbol{w}_d = (\boldsymbol{w}_1 \otimes \boldsymbol{w}_2 \otimes \boldsymbol{w}_3)^{\mathsf{T}}$ is a weight parameter for d.

B. Web search query-based destination prediction based on Latent Dirichlet Allocation (LDA)

As described in Section I, the advantage of web search query stems from the fact that it can categorize the destinations and infer the current context of the user. Therefore, we selected LDA [9]. At first, we define the probability for a familiar destination as $p_{\mathrm{fd}}^{(0)}(\boldsymbol{q}) \propto \prod_{i=1}^{|\boldsymbol{q}|} 1/|\mathbb{Q}|$. The destination probability based on web search query is calculated as

$$\begin{aligned} p^{(0)}(d|\boldsymbol{q}) &\propto p^{(0)}(\boldsymbol{q}|d)p(d) \propto p^{(0)}(\boldsymbol{q}|d) \\ &\propto \begin{cases} p_{\mathrm{ud}}^{(0)}(\boldsymbol{q}|\alpha, \beta, d) \ (d \in \mathbb{D}_{\mathrm{pub}}) \\ p_{\mathrm{fd}}^{(0)}(\boldsymbol{q}) \ (d \notin \mathbb{D}_{\mathrm{pub}}). \end{cases} \end{aligned}$$

In this equation, $p_{\mathrm{ud}}^{(0)}$ is a probability distribution of an unfamiliar destination, and $p_{\mathrm{fd}}^{(0)}$ is a probability distribution of a familiar destination. We use LDA to optimize $p_{\mathrm{ud}}^{(0)}$. We correlate this problem with a topic model problem as follows: destination category as topic, trip as document, and morpheme in web search query as word. We train this LDA for each unfamiliar destination.

V. Experimental results

We also compare the performance of the proposed method against prediction accuracy to evaluate whether the proposed method performs properly in destination prediction for both types of destinations.

A. Dataset

We obtained a GPS dataset and a web search query dataset from a commercial application provided by Yahoo! JAPAN. GPS data includes user identifier, latitude, longitude, GPS accuracy, and timestamp. A search data includes user identifier, query, and timestamp. We introduce specific unfamiliar destination areas and thresholds introduced in III-B. For unfamiliar destination areas, we choose 20 areas around the station in the Kanto region with many passengers as unfamiliar destination candidates using the dataset published by the Ministry of Land, Infrastructure and Transport in Japan[1]. This is because trains are the main means of transport in the Kanto region, and various commercial facilities, shopping malls, and sightseeing spots are situated around the station. For thresholds, we used 20 min as the threshold for labeling "staying," once a week as familiar destination extraction, then 667 users are employed.

B. Evaluation

Our goal is to achieve that high accuracy of both familiar and unfamiliar destination prediction.

1) Evaluation metric: To quantify the performance, we employ the two following metrics as prediction accuracy and prediction likelihood measures. To verify our objective, that is, if the proposed method can predict both familiar and unfamiliar destinations, we evaluate the performance with familiar ($\in \mathbb{D}_{u,\mathrm{freq}}$) and unfamiliar destinations ($\in \mathbb{D}_{\mathrm{pub}}$). We compare the destination prediction accuracy of a user using a prediction accuracy and five-fold cross-validation.

2) Comparison methods: We chose to follow destination methods as comparison methods. We use components of our proposed combination method to verify the impact of combination of two types information. **GPSOnly**: This model is included in the proposed model and serves as a GPS-based predictor. **QueryOnly**: This model is a part of our model and serves as a prediction model without GPS. This model calculates the destination probability by (3).

3) Results: We evaluate five-fold cross-validation result in two parts, trips with and without queries. At first, we show the breakdown of the number of trips with and without queries for familiar and unfamiliar destinations. It should be noted that the prediction method is not given the information whether the true destination is familiar or unfamiliar because it is not known whether the true destination is familiar or unfamiliar until arriving at the destination in actual cases.

Table I and II show the average prediction accuracy for users for the proposed method and for each comparison method. In this table, three types of results are important. That is, familiar and unfamiliar destination prediction results in Table I, and familiar destination prediction results in Table II. Unfamiliar destination prediction results in Table I and familiar destination prediction results in Table II indicate the basic prediction performance using web search queries and the GPS-based model. QueryOnly and GPSOnly should be on the top in these results, and the good performance of the combination model enables it to be close to each "Only" model. Familiar destination prediction results in Table I indicates how well models consider staying and search context correctly. For the combination model, the closer the GPS results are, more correctly this model consider them.

[1]http://nlftp.mlit.go.jp/ksj/gml/cgi-bin/download.php

TABLE I
THE AVERAGE PREDICTION ACCURACY FOR USER WITH QUERIES

Method	familiar destination	unfamiliar destination
GPSOnly	**0.617**	0.135
QueryOnly	0.167	**0.440**
Proposed (LI)	**0.562**	**0.431**
Proposed (GI)	**0.564**	**0.435**

TABLE II
THE AVERAGE PREDICTION ACCURACY FOR USER WITHOUT QUERY

Method	familiar destination	unfamiliar destination
GPSOnly	**0.883**	0.114
QueryOnly	0.000	0.079
Proposed (LI)	**0.883**	0.114
Proposed (GI)	**0.883**	0.114

VI. CONCLUSION

We developed a novel approach on destination prediction to predict both familiar and unfamiliar destinations, that is, to extend predictable destination to both familiar and unfamiliar destinations by exploiting web search queries. To the best of our knowledge, our model is the first attempt to provide accurate destination prediction in both familiar and unfamiliar settings using GPS traces and web query logs. Our proposed approach calculates the destination probability based on GPS traces and web search queries separately in a simple manner. To validate the proposed method, we evaluated the performance of the proposed method using GPS and web search query logs obtained from a commercial service. The experimental results using over 670 users with GPS and query logs demonstrate the strength of our approach.

ACKNOWLEDGEMENT

This work was partly supported by CREST Grant Number JPMJCR1403, Japan.

REFERENCES

[1] A. Vahedian, X. Zhou, L. Tong, Y. Li, and J. Luo, "Forecasting gathering events through continuous destination prediction on big trajectory data," *Proc. of SIGSPATIAL2017*.

[2] M. Xu, D. Wang, and J. Li, "Destpre: a data-driven approach to destination prediction for taxi rides," in *Proc. of Ubicomp 2016*.

[3] J. Krumm and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," in *Proc. of UbiComp 2006*.

[4] C. Manasseh and R. Sengupta, "Predicting driver destination using machine learning techniques," in *Proc. of ITSC 2013*, pp. 142–147.

[5] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu, "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction," *Proc. of ICDE 2013*, pp. 254–265.

[6] A. Y. Xue, J. Qi, X. Xie, R. Zhang, J. Huang, and Y. Li, "Solving the data sparsity problem in destination prediction," *J. of VLDB 2015*.

[7] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "Density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of KDD 1996*.

[8] R. Imai, K. Tsubouchi, T. Konishi, and M. Shimosaka, "Early destination prediction with spatio-temporal user behavior patterns," *PACM IMWUT*, 2017.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. of Machine Learning Research 2003*, vol. 3, no. Jan, pp. 993–1022.