DOI: http://dx.doi.org/10.14236/ewic/FDIA2015.11

Users Location Prediction in Location-based Social Networks

Jarana Manotumruksa
University of Glasgow
j.manotumruksa.1@research.gla.ac.uk

The wealth of user-generated data in Location-Based Social Networks (LBSNs) has opened new opportunities for researchers to model and understand human mobile behaviour, including predicting where they are most likely to check-in next. In this paper, we propose a model that leverages the use of Global Temporal Preferences and Spatial Correlation, to help make predictions for a previously unseen user - the so-called cold-start problem. The experimental results on a real-world LBSN dataset show that our proposed model outperforms the state-of-the-art approaches on prediction accuracy and can alleviate the cold-start problem.

Keywords: Location-Based Social Networks, User's Location Prediction, Cold-Start Problem

1. INTRODUCTION

Location-Based Social Networks (LBSNs), such as Foursquare and Brightkite, provide enormous user-generated data containing location data and human activity, in the form of *check-ins*, which can be exploited to understand the social and temporal characteristics of users on LBSNs. This includes predicting the user's location at a certain time, which can be useful for the design of future mobile location based services, traffic forecasting or urban planning.

In this paper, we address the issue of predicting the next location of an individual based on his/her historical check-in data. An existing state-of-the-art approach for user's location prediction proposed by Gao *et al.* (2013) used the user's daily and weekly cyclic check-in patterns to model his/her temporal preferences. However, LBSN check-in data is usually very sparse, resulting in difficulties when aiming to effectively model the personal temporal preferences of a user. To overcome this problem, Gao *et al.* (2013) proposed the use of smoothing techniques and also employ social correlation, i.e. using the preferences of the user's friends on the social network to improve suggestions.

Although the aforementioned approaches can effectively tackle the data sparsity problem, these approaches do not tackle the problem of previously unseen users, i.e. the cold-start problem. In this paper, we propose a model that leverage the use of Global Temporal Preferences and Spatial Correlation to alleviate the cold-start problem. Global Temporal Preferences exploits the historical check-ins of other users to model the temporal popularity of locations. Spatial Correlation estimates a distance that a user is willing to visit a location based on his/her current location.

The remainder of this paper is organised as follows. First, we review relevant related work in Section 2. We define the problem and our approach in Section 3. The experiment setup and results are described in Section 4. Finally, conclusions and direction for future work follow in Section 5

2. RELATED WORK

The availability of check-in data in LBSNs has recently attracted the researchers' attention. Gao *et al.* (2013) proposed to use the historical check-ins of users and their friends' in LBSNs to identify daily and weekly cyclic patterns in the users' mobile behaviour. In particular, they proposed to model the temporal preferences of a user with a Gaussian mixture model that estimates the distribution of user check-ins and predicts their location. In comparison with our proposed approach, we consider only the most recent check-in of a user as a user's current location and we use the historical check-in data of other users regardless of their relationship to the user, i.e. friendship, to model the popularity of locations at specific time. The most popular location at a specific time which is nearby the user current location is inferred as the user's next location.

Noulas *et al.* (2012) is the most related to our work where they considered the popularity of a location, i.e. the total number of check-ins at the location, as a feature and used supervised learning technique to predict the user's location. Deveaud *et al.* (2014) showed that the popularity of a location is an effective feature in a Learning-to-Rank technique for Point-of-Interest (POI) recommendation. In contrast to these works, we consider the temporal popularity of the location at a specific time period instead of the overall popularity.

Besides the popularity of a location, there have been some attempts to incorporate spatial influence for POI recommendation. Yuan *et al.* (2013) used a power law distribution to model the willingness of a user to visit a distant POI. In our work, we calculate an average distance between two successive check-ins based on their elapsed time. We then use these average distances as a threshold to filter out any locations that are far away from the most recent check-in location.

3. METHODOLOGY

In this section, we firstly explain the problem of predicting the user's location in LBSN in details (Section 3.1). We then describe our proposed model that consists of 2 components, Global Temporal Preferences (Section 3.2) and Spatial Correlation (Section 3.3).

3.1. Problem Definition

The problem of predicting the user's location in a LBSN can be formally defined as follows. Given a time t, the problem is to predict the location $l \in L$ that user $u \in U$ will visit based on his/her historical visits, $C_{u,t}$, where U and L are the set of users and locations respectively and $C_{u,t}$ is the set of check-ins for the user u before time t. Let C be the global set of all check-ins, with each check-in $c \in C$ is represented as a tuple $(u, l, t) \in C$, indicating a check-in generated by a user $u \in U$ at location $l \in L$ at time t.

It is possible to represent a specific time t (e.g. "2015-02-15 17:45:22") as a *time-slot*, for instance as a specific hour of the day (17:00) or a day of the week (Sunday). Given a time t, $T_m(t)$ is a function that returns a time slot w.r.t the specific time slot granularity m. For example, this function can be chosen to produce a time slot for each hour of the day, i.e. $T_D(t) \in \{0, 1, \ldots, 23\}$.

Next, we describe how the global temporal preferences are modeled using the historical check-in data of all users.

3.2. Global Temporal Preferences

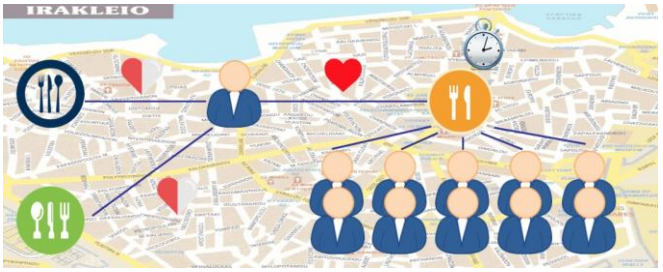

Figure 1: *An influence of mostly visited location to user's preferences*

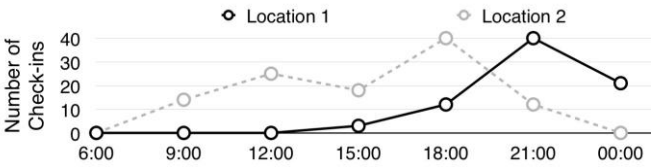

Figure 2: *A distribution of number of check-ins over time periods of two locations*

The popularity of a location, i.e. the total number of check-ins from all users on the location, is an important factor affecting human's check-in behaviour and has been exploited in venue recommendation and user's location prediction in earlier studies, e.g. Noulas *et al.* (2012); Deveaud *et al.* (2014). In this work, we assume that users are influenced by other users regardless of their relationships. Intuitively, as illustrated by Figure 1, if many users have visited a venue at a particular time, this venue may be more attractive to visit than other venues at that time. We can infer the popularity of a location l, as follows:

$$Popular(l) = |\{(u_i, l_j, t_k); (u_i, l_j, t_k) \in C_{u,t}, l_j = l\}| \quad (1)$$

Based on experimental check-in data from the Brightkite LBSN used by Gao *et al.* (2013), we found that the popularity of different locations varies over a time period as shown in Figure 2. Hence, to capture the temporal popularity of a location, we calculate the total number of check-ins of the location l at a specific time t, as follows:

$$Popular_m(l, t) = |\{(u_i, l_j, t_k); (u_i, l_j, t_k) \in C_{u,t}, l_j = l, T_m(t) = T_m(t_k)\}| \quad (2)$$

where m is the specific time slot granularity. To model the global temporal preferences of all users, we propose to compute the probability that the user will visit location l at a given time t, regardless of the user's historical check-in data, as follows:

$$P_m(l, t) = \frac{Popular_m(l, t)}{Popular(t)} \quad (3)$$

3.3. Spatial Correlation

Figure 3: *Distance between successive check-ins*

In the previous section, we described how to predict a probability that a user will check in at location ***l*** by the global temporal preferences using Equation (3). However, capturing the user mobile behaviour in LBSNs solely using Global Temporal Preferences is insufficient. Yuan *et al.* (2013) suggested that users are more willing to check-in at nearby locations to their current location.

In our initial analysis using an experimental Brightkite dataset used by Gao *et al.* (2013), we found that there is a correlation between successive check-ins. Figure 3 shows a distribution of the distance between successive check-ins over time periods. In particular, the distance between two check-ins is correlated with the elapsed time of these two check-ins. Namely, within a short period of time, users are more likely to check-in at a location close to their previous check-in.

We calculate the elapsed time between the testing check-in and the most recent check-in. Then we filter out those locations whose distance to the most recent check-in location is larger than a threshold, namely an average distance with respect to the elapsed time. The qualified locations will be ranked based on their distance to the location of the user's most recent check-in using the following equation:

$$P_d(l) = dist(l, l_{recent}) \quad (4)$$

where *dist* is a function that returns a distance between two locations in kilometers which are calculated using the Haversine Formula Shumaker, B. P., and R. W. Sinnott (1984). We use a linear combination to incorporate Spatial Correlation (Equation (4)) with Global Temporal Preferences (Equation (3)) as follows:

$$P_m(l, t) = \alpha P_d(l) + (1 - \alpha) P_m(l, t) \quad (5)$$

where α is a parameter that controls the relative contributions of Global Temporal Preferences and Spatial Correlation.

4. EXPERIMENTS AND RESULTS

In this section, we report experiments conducted to evaluate the effectiveness of our approach in alleviating the cold-start problem using a real-world check-in data from a LBSN.

Table 1: *Salient statistics of the Brightkite LBSN dataset.*

Duration	04/2008-10/2010
# of Users	26,915
# of Check-ins	4,532,151
# of Unique Locations	751,176
# of Test Check-ins	134,575
Average Check-ins per user	168

Dataset. The publicly available check-in data from Brightkite[1] is used in our experiment. Salient statistics of the dataset are listed in Table 1.

Setup. We consider 26,915 users who have at least 10 check-ins in evaluating the effectiveness of our approach (All Users). We also consider 8,613 users who have less than 20 check-ins in evaluating the extent to which our approach alleviates the cold-start problem (Cold-Start Users). Both experiments are conducted using a 5-fold cross validation on a user level, where for each testing user, we randomly select 5 check-ins as test check-ins. For each test check-in, we consider its check-in time t as given, its check-in location as the ground truth data, and a set of check-ins of a user before time t, $C_{u,t}$, as observed data. We rank all locations extracted from the observed data based on their prediction scores using the Spatial Correlation and Global Temporal Preference model (SGTP) in Equation (5). Then we select the top ranked location as the predicted location where the user is most likely to visit next. To set α, we use 5-fold cross validation, varying α from 0.0 to 1.0 in 0.1 incremental steps, to determine the value that maxmimises Success@1 (see below).

Measure We evaluate the accuracy of the predicted locations using Success@1, which was called *prediction accuracy* by Gao *et al.* (2013), i.e. the ratio of the number of accurately predicted locations to the total number of testing check-ins.

Baselines Two baselines used in our experiments are the state-of-the-art user's location prediction approaches proposed by Gao *et al.* (2013) : (i) Personal Temporal Preference (PTP) where they predict the next location based on the user's daily and weekly historical check-in data and (ii) Temporal Social Correlation (TSC) where they use a collaborative filtering technique to predict the next location based on the temporal preferences of user's friends.

Table 2 shows the prediction accuracy of our approach (SGTP) in comparison with the baselines

[1] http://snap.stanford.edu/data/loc-brightkite.html

Table 2: *Prediction accuracy (success@1) for various models*

	PTP	TSC	SGTP
All Users	0.340	0.334	**0.402**
Cold-Start Users	0.327	0.326	**0.341**

(PTP and TSC). The first row of the table presents the results of the effectiveness evaluation using 26,915 users. The second row of the table compares the effectiveness in alleviating the cold-start problems between our approach and the baselines using 8,613 users. From Table 2, we observe that the linear combination of Spatial Correlation and Global Temporal Preference (SGTP) improves the prediction accuracy by 18% and 20% in comparison with Personal Temporal Preferences (PTP) and Temporal Social Correlation (TSC), respectively. SGTP obtains the optimal results when α is set to 0.9. This clearly demonstrates that users are more likely to check-in at nearby locations than at the most popular ones. In particular, the spatial correlation plays a more important role in predicting the user's location than the global temporal preferences in this dataset. Finally, our approach is promising in alleviating the cold-start problem more effectively than the baselines, by approximately 4% (0.341 vs 0.327 and 0.326, respectively).

5. CONCLUSIONS AND FUTURE DIRECTIONS

The availability of historical check-in data in LBSNs can be exploited to understand the user mobile behaviour. In this paper, we propose a model that leverages Global Temporal Preferences and Spatial Correlation to alleviate the problem of unseen users (cold-start users). The experiment results on real-world check-in data in a LBSN shows that our proposed approach outperforms the stat-of-the-art approaches both in terms of effectiveness and in alleviating the cold-start problem.

The study of user mobile behaviour on LBSNs can be exploited in many applications. A venue recommendation or geo-based advertising application could take this into account in order to improve its effectiveness based on the user's location and the time of the day.

6. ACKNOWLEDGMENTS

I would like to thank my supervisors, M-Dyaa Albakour, Craig Macdonald and Iadh Ounis for their supports. We thank ACM SIGIR for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Gao, Huiji and Tang, Jiliang and Hu, Xia and Liu, Huan (2013) *Modeling temporal effects of human mobile behavior on location-based social networks. Proceedings of CIKM, pp. 1673-1678*.

Noulas, Anastasios and Scellato, Salvatore and Lathia, Neal and Mascolo, Cecilia (2012) *Mining user mobility features for next place prediction in location-based services. Proceedings of ICDM, pp. 1038-1043*

Deveaud, Romain and Albakour, Dyaa and Macdonald, Craig and Ounis, Iadh and others(2014) *On the Importance of Venue-Dependent Features for Learning to Rank Contextual Suggestions. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1827-1830*

Deveaud, Romain and Albakour, Dyaa and Macdonald, Craig and Ounis, Iadh and others(2014) *Time-aware point-of-interest recommendation. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 363-372*

Shumaker, B. P., and R. W. Sinnott (1984) *Astronomical computing: 1. Computing under the open sky. 2. Virtues of the haversine." Sky and telescope 68, pp. 158-159*