Improving Location Prediction using a Social Historical Model with Strict Recency Context

Kwan Hui Lim[*†], Jeffrey Chan[*], Christopher Leckie[*†], and Shanika Karunasekera[*]
[*]Department of Computing and Information Systems, The University of Melbourne, Australia
[†]Victoria Research Laboratory, National ICT Australia, Australia
{limk2@student., jeffrey.chan@, caleckie@, karus@}unimelb.edu.au

ABSTRACT

Location-based Social Networks (LBSN) are a popular form of social media where users are able to check-in to locations they have visited and share these check-ins with their friends. An important problem in LBSNs is the prediction of a user's next check-in location, for purposes such as showing location-specific recommendations or advertisements. One state-of-the-art algorithm is the Social Historical Model (SHM) which utilizes a user's historical check-ins and social links to predict his/her next check-in location. Observing various LBSN user characteristics, we improve the SHM by exploiting the recency effect (where users are more likely to revisit places from their recent past than the more distant past) and place-links (where two friends share a common temporal check-in). Using two Foursquare datasets, we then demonstrate how these modifications improve the overall location prediction accuracy of the SHM.

Categories and Subject Descriptors: J.4 [Computer Applications]: Social and behavioral sciences

General Terms: Experimentation, Measurement

Keywords: Location Prediction, Next Check-in Prediction, Location-based Social Networks, Foursquare

1. INTRODUCTION

Location-based Social Networks (LBSN) such as Foursquare and Facebook Places have gained immense popularity in recent years, fueled by the prevalence of smart phones with GPS technology. LBSNs enable users to check-in to any place they visit and share these check-ins with their friends. This ubiquitous technology provides an opportunity for businesses to more effectively market products and services to LBSN users based on their fine-grained location data. More importantly, knowing the next place a user intends to visit will allow these businesses to further optimize their marketing strategy by displaying products and services relevant to these predicted locations.

Related Work. These potential business applications make location prediction an important problem, which has been an active focus of the research community. Most earlier works focus on general location prediction (i.e., predicting the next check-in location for typical users), while others focus on specific domains such as predicting check-ins to novel (new and un-visited) locations [7] or predicting check-ins for new users who have no prior check-in history [5]. Our work focuses on the former and we shall discuss some related work in the general area of location prediction.

In [2], Chang and Sun studied various user features using a logistic regression model and found that a user's past check-ins are the most significant predictor, resulting in the Most Frequent Check-in (MFC) model. Recognizing that users' check-ins exhibit certain temporal patterns (e.g., go to school/work in the morning and home in the evening), the Temporal-based MFC model extends upon the MFC model by considering this temporal nature of check-ins [3]. Observing that a user's check-in is often part of a sequence of check-ins (e.g., go school, canteen, then home), Song et al. [10] proposed the Order-k Markov model, which considers the frequent patterns that exist in users' check-in sequence. More recently, Gao et al. [4] proposed a Social Historical Model (SHM) for location prediction using a language model that considers both a user and his/her friends' past check-ins. By utilizing both a user's historical check-ins and his/her social links, the SHM has been shown to out-perform the earlier discussed models in terms of location prediction accuracy [4, 7]. Using two LBSN user characteristics (check-in recency and place-links), we show how we can further enhance the SHM's performance.

Contributions. Our main contributions are: (i) studying the recency effect of check-ins (i.e., how users tend to revisit more recently visited places), and demonstrating how it can improve location prediction using historical data; (ii) introducing place-links where the users are linked by both an explicit friendship and common recent check-in, and showing how place-links improve location prediction using social links; and (iii) modifying the SHM to include both the recency effect and place-links, resulting in an overall improvement in prediction accuracy.

2. ORIGINAL SOCIAL HISTORICAL MODEL

The SHM comprises both the components of a Historical Model (HM) and Social Model (SM). As restated from [4], we first describe the HM and SM components before elaborating on how they constitute the SHM.

The HM predicts a user's next check-in location using

CaRR '15, March 29, 2015, Vienna, Austria

their past check-in history. A user's check-in history is modeled as a Hierarchical Pitman-Yor (HPY) process [11], which is a language model that extends the traditional Pitman-Yor process and generates a probability distribution of check-in locations, with a discount parameter to capture the power-law effect. In addition, the HPY process also uses an n-gram model to capture the short-term effect where the latest check-in has more importance than an earlier check-in.[1] The HM is formally defined as:

$$P^i_{HM}(c_{n+1} = l) = P^i_{HPY}(c_{n+1} = l) \tag{1}$$

where $P^i_{HPY}(c_{n+1} = l)$ is the probability of user i's next check-in c_{n+1} at location l, calculated using the HPY process with user i's previous check-ins, $c_1, c_2, ..., c_n$.

The SM first computes the similarity between a user i and his/her friend j where the user-friend similarity $sim(i, j)$ is based on their frequency of common check-ins over their total number of check-ins. The computation of this similarity is then repeated for the entire set of user i's friends, denoted F_i. Thereafter, the SM attempts to predict the next check-in location of user i based on his/her friends. Formally, the SM is defined as:

$$P^i_{SM}(c_{n+1} = l) = \sum\nolimits_{j \in F_i} sim(i, j) P^j_{HPY}(c_{n+1} = l) \tag{2}$$

where $P^j_{HPY}(c_{n+1} = l)$ is the probability that a user i's next check-in c_{n+1} is at location l, calculated using the HPY process with friend j's previous check-ins (unlike Eqn. 1 that uses user i's previous check-ins, Eqn. 2 uses the previous check-ins of his/her friend, user j). In short, the SM predicts user i's next check-in location based on the predicted check-ins of user i's friends and their similarity to user i.

The SHM then uses both the HM and SM to predict next check-in locations and is defined as:

$$P^i_{SHM}(c_{n+1} = l) = \eta P^i_{HM}(c_{n+1} = l) + (1 - \eta) P^i_{SM}(c_{n+1} = l) \tag{3}$$

where η is the weighting given to the HM and SM. The original authors experimented with various values of η and found that a value of 0.7 works the best. For more details on the SHM, refer to [4].

3. PROPOSED MODIFICATIONS TO SHM

Our modifications to SHM include adopting more stringent definitions of check-ins' recency and social links for the HM and SM respectively.

Specifically for the HM, we adopt a stringent *recency criterion* for check-ins rather than use all of a user's past check-ins regardless of their time. Let T_n be the time of the latest check-in c_n of a user i. We modify the HM (Eqn. 1) such that we only train the HPY process with user i's previous check-ins, $c_m, c_{m+1}, ..., c_n$, where $T_n - T_m = X\ month$, i.e., we only use check-ins by a user that is within the most recent X months, as training data. In contrast, the original HM use all of a user's past check-ins. As the HPY process emphasizes on both the power-law distribution (frequency) and short-term effect (recency) of check-ins [4], a location that is frequently visited in the past could be incorrectly given a high probability. Our proposed modifications further constrain the HPY process to a much smaller set of recent check-ins, ensuring that the emphasis on check-in frequency is only on recent data. We denote this modified HM (using strict recency) as HM-SR.

Instead of using social links for the SM, we only consider *place-links*, which are defined as social links where the connected users have checked-in to the same venue on the same day. Referring to Eqn. 2, we consider and calculate $sim(i, j)$ only for users i and j who share a common check-in that is performed on the same day. While the SM utilizes an effective user-friend similarity based on their common check-ins, the temporal aspects of such common check-ins have not been considered (e.g., two friends with common check-ins that are months apart). Place-links introduce this temporal criterion and ensure that we only consider friends who are similar in both the temporal and spatial aspects of check-ins. We denote this modified SM (using place-links) as SM-PL.

Similar to original SHM (Eqn. 3), our modified SHM (denoted SHM-PLSR) then combines the results from both HM-SR and SM-PL with a η value of 0.7. We next present some experiments and data analysis that show the effectiveness of these proposed modifications.

4. EXPERIMENTS AND RESULTS

Datasets. We use two large LBSN datasets from Foursquare that comprise 2.29M and 2.07M check-ins, coupled with 47k and 115k friendship links among the 11k and 18k users. All check-ins are time-stamped and tagged to a specific location while user friendships are bi-directional links. The two datasets differ in terms of their time range, one is from Jan 2011 to Dec 2011 while the other is from Mar 2010 to Jan 2011. Both datasets are publicly available at [5] and [4].

Evaluation. Using the two datasets mentioned previously, we divide each dataset into 10 equal time bins and consider users with >1 check-in at each time bin for our experiment. At each time bin, we hide the last check-in location of each user and attempt to predict it based on the preceding data. E.g., for an evaluation using time bin 5, we will try to predict the last check-in location for each user in time bin 5 based on the preceding time bins 1 to 5 (minus the last check-in). Thereafter, we evaluate the various models using the average prediction accuracy, which is based on the total number of correct predictions (for all users over all time bins) out of the total number of predictions made.

4.1 Experiments on Historical Model

Temporal Re-visit Trends. As our proposed HM-SR is built on the premise that users tend to visit (check-in to) more new places than old ones (previously visited places), we first investigate the temporal trends of how users re-visit such old places. Using the set of unique places to which a user has performed a check-in in the first time period, we compute the user's re-visit ratio based on how many of these unique places the same user has re-visited in the subsequent time periods. A re-visit ratio of 1 thus indicates that a user re-visits all of his/her formerly visited places, while a value of 0 indicates otherwise.

Fig. 1 shows the average re-visit ratio for all users over the entire timespan of our two Foursquare datasets based on a time period of one to four months.[2] The results show a gen-

[1] Due to space limit, we are unable to discuss the HPY process in detail and refer readers to [11] for more information.

[2] As dataset 2 is of a shorter duration than dataset 1, the last time period of dataset 2 is not plotted in Fig. 1.

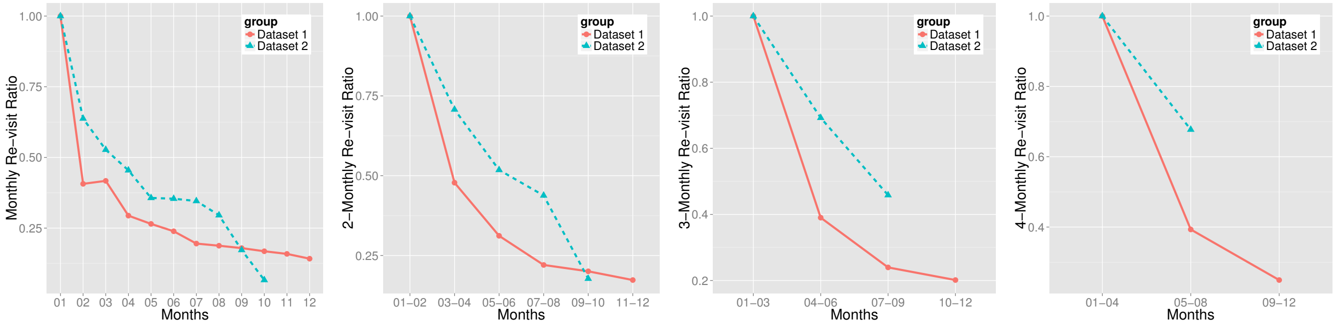

Figure 1: Re-visit Ratio over Time Periods of 1 to 4 Months (left to right)

eral downward trend of the re-visit ratio for both datasets at all time periods, indicating that users tend not to re-visit old places over time. Based on monthly re-visits, users only re-visit approximately half of these old places after three months, and re-visit less than one-fifth of these places after nine months. This observation indicates that LBSN users are less likely to re-visit a place which they have visited long ago (i.e., a short-term check-in trend). Our proposed HM-SR further exploits this short-term trend by implementing a strict recency constraint where we only use check-ins within one month to train our model. This short-term trend has also been exploited in other work such as [6] that uses check-in recency for identifying location-specific domain experts, and [9] that uses song recency for personalized music recommendation.

Building on the results in Fig. 1, we have experimented with various values of check-in recency and found that a value of one month works the best for the HM-SR. Using a higher value of recency reverts the HM-SR to the original HM, while a lower value results in insufficient training data to provide accurate predictions. Thus, for the rest of the paper, we employ a check-in recency of one month for the HM-SR and SHM-PLSR.

Evaluation of Historical Model. We now compare our proposed HM-SR to the original HM in terms of average prediction accuracy. Table 1 shows that our proposed HM-SR provides an improvement of 2.6% and 8% over the original HM for datasets 1 and 2 respectively.

Table 1: Prediction Accuracy for Historical Models

	Dataset 1		Dataset 2	
Model	**Accy.**	**Impv.**	**Accy.**	**Impv.**
HM	0.2741	-	0.2397	-
HM-SR	***0.2812***	2.6%	***0.2589***	8.0%

While our HM-SR offers a modest improvement over the HM, the lack of a bigger improvement is because the original HM is using the HPY process, which gives a heavier emphasis to check-ins that are performed more frequently and recently. However, a place that has been frequently visited in the past may be incorrectly emphasized, especially if the user no longer visits that place (e.g., due to a change of work, school or home). Our HM-SR achieves a further improvement over HM by further constraining the HPY process to a much smaller set of recent check-ins. This strict recency constraint proves to be effective as users only re-visit as few as 40% of the places they have visited in the previous month, as shown in Fig. 1.

4.2 Experiments on Social Model

Comparison of Place-links and Social Links. As our SM-PL uses place-links instead of social links, we now investigate the effectiveness of place-links over social links for location prediction in terms of the check-in similarity of users. For each user i and his/her set of friends $j \in F$ (based on link type L), we define their check-in similarity as:

$$S_L = \frac{1}{|F|} \sum_{j \in F} sim(i,j) \tag{4}$$

Thus, S_P and S_S refers to the check-in similarity of users based on place-links and social links respectively (similarly calculated based on Eqn. 4). We then compute S_P and S_S using the two Foursquare datasets described earlier in §4, and present the results in Fig. 2.

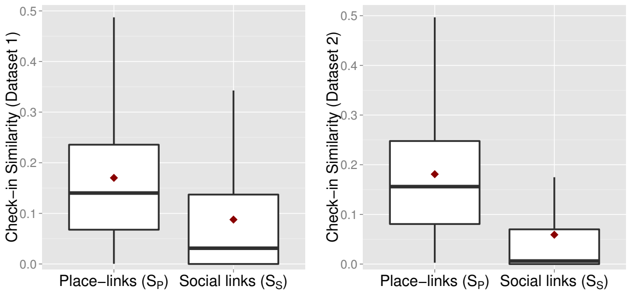

Figure 2: Check-in similarity for users based on place-links (S_P) and social links (S_S)

Next, we conduct a two sample t-test with null hypothesis H_0: $S_P \leq S_S$ and alternative hypothesis H_1: $S_P > S_S$. We obtained p-values of < 2.2e-16 for both datasets 1 and 2, and reject the null hypothesis. This result shows that users connected by place-links share more common check-ins than users connected by social links, thus motivating the use of place-links in our proposed SM-PL, which we evaluate next.

Evaluation of Social Model. Comparing our proposed SM-PL to the original SM, we observe that our proposed SM-PL improves the average prediction accuracy by 30.2% and 20.4% over the original SM for datasets 1 and 2 respectively, as shown in Table 2.

While the original SM utilizes a similarity weighting based on the common check-ins between a user and his/her friends, the SM does not consider the temporal aspects of this check-in, e.g., a common check-in between a user and his/her friend can be days or months apart but are still given the same weighting. Our introduction of place-links enforces a temporal criterion on these social links, ensuring that we only

Table 2: Prediction Accuracy for Social Models

	Dataset 1		Dataset 2	
Model	Accy.	Impv.	Accy.	Impv.
SM	0.1242	-	0.0663	-
SM-PL	***0.1617***	30.2%	***0.0798***	20.4%

consider friends who are similar in terms of a common check-in that is performed within a common time (within a day).

Prior work has also shown that spatial-social links (based on friends with a common check-in regardless of the check-in time) result in place-bound communities comprising users who frequently visit the same venues [1]. Our work extends upon the concept of spatial-social links in [1] and uses place-links that are spatial-social links with a temporal constraint (i.e., friends who share a common check-in performed within a certain time) for location prediction. Similarly, place-links have also been used for other applications such as in the detection of location-centric communities [8].

4.3 Overall Evaluation of Prediction Models

The evaluation thus far shows that HM-SR and SM-PL out-performs their original counterparts, the HM and SM respectively. Moving on, we now evaluate the performance of SHM-PLSR against original SHM, in order to determine the overall improvement of our proposed modifications. In addition, we also compare our SHM-PLSR to various baseline prediction models that were used in [7, 4], namely:

- Most Frequent Check-in (MFC): Predicts next check-in as the most frequently visited location of the user, based on his/her previous check-ins.
- Most Frequent Check-in, Temporal-based (MFC-T): Same as MFC, but also considers the time (hour) of the day when the check-in is performed.
- Most Frequent Check-in, All Users (MFC-A): Same as MFC, but uses the most frequently visited location of all users, instead of a single user.
- Random Check-in Selection (RAND): Randomly select a location that a user has previously visited as his/her predicted next check-in location.

Table 3: Prediction Accuracy for Various Location Prediction Models

	Prediction Accuracy	
Model	Dataset 1	Dataset 2
MFC	0.2693	0.2039
MFC-T	0.1449	0.1357
MFC-A	0.0383	0.0064
RAND	0.0299	0.0472
SHM	0.2767	0.2395
SHM-PLSR	***0.2855***	***0.2619***

Table 3 shows that our proposed SHM-PLSR out-performs all baselines (MFC, MFC-T, MFC-A, RAND), with improvements in prediction accuracy for all cases. In addition, our proposed SHM-PLSR offers an improvement of 3.2% and 9.4% over original SHM for datasets 1 and 2 respectively.

As noted in [4], historical check-ins play a bigger role in location prediction than social links, thus the HM component heavily influences the results produced by the SHM. Similarly, this trend is also reflected in our SHM-PLSR with an overall improvement of 3.2% and 9.4% (for datasets 1 and 2), despite a greater improvement in its SM-PL component of up to 30.2%. While the improvements to the original SHM is modest, our SHM-PLSR has been shown to out-perform various other baselines by large margins. More importantly, we believe that our findings offer some insight into user behavior on LBSN and provide future opportunities for new location prediction algorithms using strict recency and place-links.

5. CONCLUSION

We first examined the recency effect where users exhibit a short-term check-in trend and are less likely to re-visit the same place over longer periods of time. Using this observation, we then applied a strict recency criterion (using only the most recent one month of check-ins) to the Historical Model, which improved its prediction accuracy by up to 8.0%. Thereafter, we introduced place-links, which are essentially social links embedded with spatial and temporal aspects (i.e., a link connecting two friends who check-in to the same place on the same day). Next, we modified the Social Model to use place-links (instead of social links) and succeeded in improving its prediction accuracy by up to 30.2%. Finally, we show how adding the concepts of strict recency and place-links to the Social Historical Model improves its prediction accuracy by up to 9.4%. Future directions include adopting a weighted version of place-links using a time-based decay function of the common daily check-ins.

Acknowledgments. National ICT Australia (NICTA) is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

6. REFERENCES

[1] C. Brown, V. Nicosia, and et. al. The importance of being placefriends: discovering location-focused online communities. In *Proc. of WOSN*, pages 31–36, 2012.

[2] J. Chang and E. Sun. Location 3: How users share and respond to location-based data on social networking sites. In *Proc. of ICWSM*, pages 74–80, 2011.

[3] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. of KDD*, pages 1082–1090, 2011.

[4] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *Proc. of ICWSM*, pages 114–121, 2012.

[5] H. Gao, J. Tang, and H. Liu. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks. In *Proc. of CIKM*, pages 1582–1586, 2012.

[6] W. Li, C. Eickhoff, and A. P. de Vries. Geo-spatial domain expertise in microblogs. In *Proc. of ECIR*, 2014.

[7] D. Lian, V. W. Zheng, and X. Xie. Collaborative filtering meets next check-in location prediction. In *Proc. of WWW Companion*, pages 231–232, 2013.

[8] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera. Detecting location-centric communities using social-spatial links with temporal constraints. In *Proc. of ECIR*, 2015.

[9] M. Schedl, D. Hauger, and D. Schnitzer. A model for serendipitous music retrieval. In *Proc. of CaRR*, 2012.

[10] L. Song, D. Kotz, R. Jain, and X. He. Evaluating location predictors with extensive wi-fi mobility data. In *Proc. of INFOCOM*, pages 1414–1424, 2004.

[11] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. of ACL*, 2006.