Fine-Scale Prediction of People's Home Location using Social Media Footprints

Hamdi Kavak[1][0000−0003−4307−2381], Daniele Vernon-Bido[2], and Jose Padilla[2]

[1] George Mason University, Fairfax VA 22030, USA,
hkavak@gmu.edu,
http://www.hamdikavak.com
[2] Virginia Modeling Analysis and Simulation Center,
Suffolk VA 23435, USA

Abstract. In this study, we develop a machine learning classifier that determines Twitter users' home location with 100 meters resolution. Our results suggest up to 0.87 overall accuracy in predicting home location for the City of Chicago. We explore the influence of *time span of data collection* and *location-sharing habits of a user*. The classifier accuracy changes by data collection time but larger than one-month time spans do not significantly increase prediction accuracy. An individual's home location can be ascertained with as few as 0.6 to 1.4 tweets/day or 75 to 225 tweets with an accuracy of over 0.8. Our results shed light on how home location information can be predicted with high accuracy and how long data needs to be collected. On the flip side, our results imply potential privacy issues on publicly available social media data.

Keywords: human mobility, social media, home location inference

1 Introduction

The availability of large-scale behavioral data allows researchers to scratch the surface of human behavior understanding and prediction [2]. The spatial movement of people provides a starting point as it has the potential to reveal the relationship between places, activities, and social interactions that make up one's daily life. In this respect, home is one of the most important hubs for people when transitioning from one daily activity to another [8]. However, predicting 'home' is challenging due to the sparsity of geo-tagged social media footprints.

There is a wide range of methods to infer people's home location from their social media content [5, 7, 6, 4]. In this study, we develop a machine learning classifier that used geo-tagged tweets for home location prediction following and extending Hu et al.[4]'s work. We start building our classifier with five human mobility features identified to be important in [4]. We advance Hu et al.[4]' s work by (1) adding two mobility features (land use patterns and distance from most checked-in location), (2) constructing place visit history through clustering, and (3) exploring the effect of data collection length, tweeting rate, and the number of tweets on classifier accuracy. Our study considers 100 meters resolution and outperforms in applicable scope (100%) and accuracy (up to 0.87) of any previous studies. We also report data collection requirements necessary for home location prediction problem.

In section 2, we describe the dataset used in this study and report the preparation process we follow. In section 3, we present our home location prediction classifier and explain how we train and evaluate its performance. We then report our results in section 4 and investigate several factors that are influential on prediction accuracy. Finally, we conclude the paper in section 5.

2 Dataset Preparation

We perform three steps in preparing social media data for home location prediction. First, we collect data from Twitter and identify a subset suitable for the study. Next, we create a *ground-truth* dataset that contains tweets known to be sent from users' home. Finally, we clean the *ground-truth* dataset and cluster it to identify unique locations at the individual level.

We use Twitter's Streaming API to collect public tweets with exact Global Positioning System (GPS) locations. Tweet collection is performed between May 16, 2014 and April 27, 2015. We choose the city of Chicago, Illinois because it is one of the significant metropolitan areas in the United States. We focus on active users with at least five geo-tagged tweets [4]. The active users' dataset contains ≈7.78 million location footprints from 92,296 Twitter users.

We create a dataset with a portion of active users whose home locations are known with confidence. We follow a process similar to the one in [4] that relies on crowdsourcing to identify whether tweets that contain home-related keywords [3] are sent from home or not. We develop a web application for labeling these home-related tweets. The web application simply displays a tweet from the dataset and asks the user (crowdsource) to choose a label: *from home*, *not from home*, or *unsure*. Each question is displayed up to three times randomly. Precedence is given to the ones that already have an answer. We received 14,076 responses for 4,679 questions. We only focused on tweets with an agreement in all three that the user is sending the message from home. Approximately 38% of tweets (1,797) from 1,268 users satisfy this criterion.

Lastly, pre-processing consists of two steps: *cleaning* and *clustering*. Cleaning prevents biasing the dataset with a significant number of tweets from the same location in a short time interval. We clean tweets that are consecutively shared in less than sixty minutes and within 100 meters distance. At the end of the cleaning, 62.2% of messages were removed. Clustering uses the cleaned tweets and identifies tweets sent from same places. This is important because GPS data usually has some inaccuracy even when shared from the same location. We cluster these points by giving them the same location ID label using the DBSCAN algorithm[3] with 100-meter as the maximum distance parameter and one as the minimum number of points in a cluster. Pre-processing generated

[3] *home* and at least one of the following keywords: *shower*, *sofa*, *TV*, *sleep*, *nap*, *bed*, *alone*, *watch*, *night*, *sweet*, *stay*, *finally*, *tonight*, *arrived*

a dataset containing 462,409 location footprints with the following properties: *anonymized user ID*, *local date-time*, *location (latitude-longitude)*, *cluster label ID*, and *is home*.

3 Home Location Prediction

We consider home location prediction a process that takes a set of location footprint history of a user and predicts the footprints that are shared from home. To this end, we create a method that receives a location footprint set of a user and generates a mobility feature set (X) that contains one record *per unique cluster label ID*. This mobility feature set has following features where the first six are identified in [4] and the last two are proposed by the authors (see online supplemental for detailed feature descriptions and value distributions).

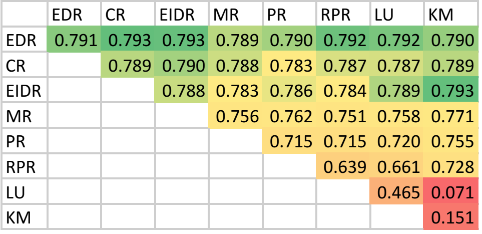
- Check-in Ratio (CR)
- Check-in Ratio during Midnight (MR)
- Check-in Ratio of Last Destination of a Day (EDR)
- Check-in Ratio of Last Destination of a Day with Inactive Midnight (EIDR)
- PageRank (PR)
- Reverse PageRank (RPR)
- Land Use Pattern (LU)
- Kilometer Distance from Most Checked-in Location (KM)

We use Support Vector Machines (SVM) with linear kernel as our classifier because it is a robust approach for binary classification problems and has been successfully implemented in the home location prediction problem [4]. SVM works by creating an optimally placed decision boundary (hyperplane) to separate elements of classes with maximum margin [1]. It is computationally costly to train an SVM classifier due to the involvement of numerical optimizations, but it is computationally efficient to use as a trained classifier. For instance, when a linear SVM classifier is trained, the classification problem turns into a simple calculation of $c = W * X + b$. When c is non-negative, it denotes one class and when it is negative, it denotes the other class given that W is the weight vector for the hyperplane, b is the intercept parameter, and X is the input whose class is investigated.

To train and test the classifier, we apply repeated 5-fold cross-validation. That is, we split the ground-truth users in five equal groups, train the classifier with four groups and test it with the remaining one, and repeat until all groups are used in training four times and in testing one time. We also repeat this procedure five times with randomly shuffling the user list in each time. In total, each evaluation takes 25 runs. In each fold, we predict home location at the user-level by calculating the SVM score for each unique location label ID and pick the one with the highest score. In the end, we capture average accuracy.

4 Results

We first report the accuracy of each mobility feature separately and with their combinations shown in Fig. 1. As a single feature, *End of Day Ratio* (EDR) has the highest accuracy with 0.791 while general *Check-in Ratio* (CR) and *End of Inactive Day Ratio* (EIDR) are marginally lower. *Midnight Ratio* (MR) is slightly lower with 0.756 followed by *PageRank* (PR) and *Reverse PageRank* (RPR) scores 0.715 and 0.639. Finally, *Land Use* (LU) feature has the accuracy score of 0.465 and *Kilometer Distance to Most Visited Location* (KM) feature performs the worst with the accuracy score of 0.151.



	EDR	CR	EIDR	MR	PR	RPR	LU	KM
EDR	0.791	0.793	0.793	0.789	0.790	0.792	0.792	0.790
CR		0.789	0.790	0.788	0.783	0.787	0.787	0.789
EIDR			0.788	0.783	0.786	0.784	0.789	0.793
MR				0.756	0.762	0.751	0.758	0.771
PR					0.715	0.715	0.720	0.755
RPR						0.639	0.661	0.728
LU							0.465	0.071
KM								0.151

Fig. 1. Accuracy scores of single features (diagonal) and all combinations of two features. The color intensity is given based on values.

Pairing the features provides marginal improvements to the best performing single features; however, the lowest performing feature, KM, when combined with PR and RPR improved the accuracy by 8 to 14%. The combined best score, 0.793, outperforms the best-reported score in the literature [4] in that the accuracy is the same but the scope of applicability is greatly improved. Hu et al. [4] have an applicability of 30 - 40% while our classifier covers 100%. To check the robustness of this result, we examine the change of accuracy based on data collection length and number of footprints.

Fig. 2 shows that accuracy scores are not linear with respect to data collection length. The top performing single and combined feature scores reach their maximum accuracy in 14 days (except for PR which peaks at 21 days). For single features, there is a slight difference in the rankings of the top three features although they are still very close. For combined features, their accuracy appears to be a bit better than the single features especially when data collection length is 30 days or lower.

In addition to the data collection length, we investigate the number of tweets per user and the user tweeting rate. The number of tweets per user (G_n) captures the direct relationship between footprint size and classifier accuracy. We define a measurement - *tweeting rate* - to standardize the unit over different tweeting habits of users (G_r). Tweeting rate (Eq. 1) is given as the total number of tweets of a user divided by the number of days between the first and last tweet.

$$G_r = \frac{number\, of\, tweets}{number\, of\, days\, between\, first\, and\, last\, tweet} \tag{1}$$

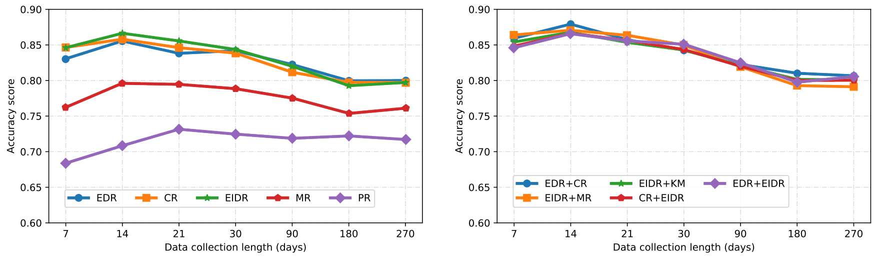


Fig. 2. Accuracy score based on data collection length for single best performing features (left) and combined best performing features (right).

We create four groups of users for each of the two measures keeping a similar number of users in each group (see supplemental). Fig. 3 shows the average accuracy for the top five performing single and combined features. For users with lower tweeting rate, the classifier's accuracy averages 0.6 and 0.7 respectively for single features and combined features. Additionally, PR and RPR perform poorly on this group because of an insufficient number of check-ins.

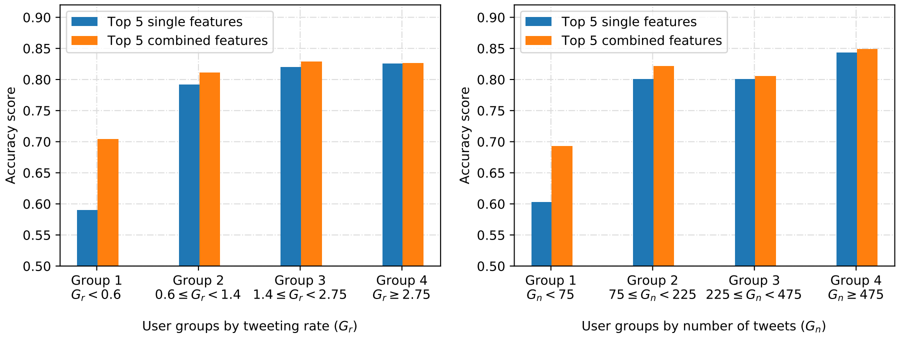


Fig. 3. The change of accuracy by four groups of users gathered based on tweeting rate (left) and number of tweets (right).

Groups 2 through 4 demonstrate no significant difference in accuracy. The higher tweet rates do not provide sufficient benefit to the classifier's accuracy rating to suggest that the increase is relevant. As such, we consider the Group 2 users to have the optimal tweet rate for predicting home location. We conclude that with our classifier, 0.6 to 1.4 daily geo-tagged tweet activity or 75 to 225 total tweets per individual allow for the predication of the user's home location with over 0.8 accuracy.

5 Conclusion

In this paper, we addressed the fine-scale (100-meter) prediction problem of Twitter users' home locations. We developed an SVM classifier with several mobility features including check-in ratios at locations, graph-based features, and distance between locations. We then trained this classifier with geo-tagged Twitter data from the City of Chicago and explored the accuracy of home location prediction under different conditions. The best accuracy for the entire dataset was 0.795. When considering several subsets of the dataset, we gathered empirical insights that were not clearly present in the entire dataset. For instance, we found that a high number of tweets and high tweeting activity did not significantly increase prediction accuracy. In fact, 0.6 to 1.4 daily tweeting rate or 75 to 225 number of tweets is enough to perform over 0.8 accuracy. Lower numbers, on the other hand, reached 0.71 accuracy which is still very high given that our classifier covers all the instances in the applicable scope whereas the previous study by [4] only applies to 71-76% of instances for achieving a similar accuracy.

Notes: Additional information, code, and datasets of this manuscript are freely available at https://github.com/hamdikavak/home-location-prediction. We thank our colleagues at Old Dominion University who helped labeling our dataset.

Acknowledgments and Disclaimer: This material is based on research sponsored by the Office of the Assistant Secretary of Defense for Research and Engineering (OASD(R&E)) under agreement number FAB750-15-2-0120. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the OASD(R&E) or the U.S. Government.

References

1. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20(3), 273–297 (1995)
2. Eagle, N., Pentland, A.S.: Reality mining: Sensing complex social systems. Personal and Ubiquitous Computing 10(4), 255–268 (2006)
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the 2nd KDD. AAAI Press (1996)
4. Hu, T., Luo, J., Kautz, H., Sadilek, A.: Home Location Inference from Sparse and Noisy Data: Models and Applications. Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015 pp. 1382–1387 (2016)
5. Mahmud, J., Nichols, J., Drews, C.: Where Is This Tweet From? Inferring Home Locations of Twitter Users. Icwsm pp. 511–514 (2012)
6. Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., Almeida, V.: Beware of what you share: Inferring home location in social networks. ICDMW 2012 pp. 571–578 (2012)
7. Ryoo, K., Moon, S.: Inferring Twitter user locations with 10 km accuracy. Proceedings of the 23rd international conference on WWW (2014)
8. Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C.: Unravelling daily human mobility motifs. Journal of the Royal Society, Interface 10(84), 20130246 (2013)