DESTPRE : A Data-Driven Approach to Destination Prediction for Taxi Rides

Mengwen Xu
Joint work with Dong Wang, Jian Li

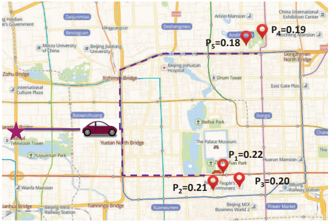
Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University.

Contents

Talk Structure

Introduction

- A growing number of mobile devices users.
- An increasing demand for predicting a driver's destination.
- Destination prediction is to predict the destination of a trip given a partial passed trajectory.



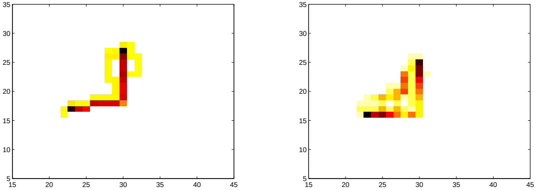
Applications

Several applications can benefit from destination prediction.

- Location-based advertising.
- Route recommendation.
- Navigation system.
- Study the dynamic flow of the traffic.

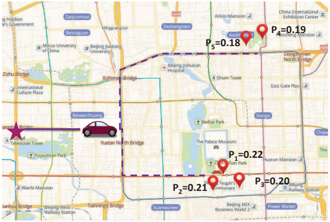
Existing approach-Markov Chain Models

- Hard to describe the behavior of trajectory.
- High order model is hard to train.
- Modified approach[1]: only retain the random walks that are not much longer than the shortest path.



(a) Real life trajectories (b) Modified Markov Chain

Top-k destinations

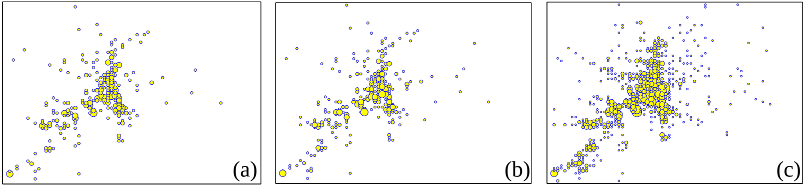


Several previous work[1-5] used probabilistic inference to compute and return the top-k destinations.

- does not consider destinations' geographic locations.
- The returned top-k places may close to each other, and ignore some faraway places with similar probabilities.

Talk Structure

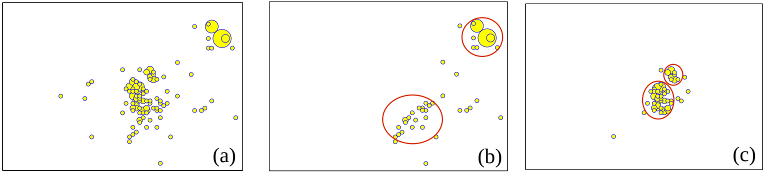
Observation 1



Observation 1

If the prefixes of the trajectories are similar, the distributions of the geographic locations of their destinations also tend to be similar.

Observation 2



Observation 2

The destinations of trajectories with similar prefixes and length are clustered.

Framework

The problem now transforms into finding similar historical trajectories, and predicting destination using them.

- Offline stage: Build the index using BPR quadtree and Minhash index for quickly finding similar trajectories.
- Online stage: Given the partial trajectory, predict the destinations.
    - Search the candidate trajectories from the index.
    - Classify the similar trajectories into different groups
    - In each group, cluster the destinations.
    - The centers of all the clusters are returned.

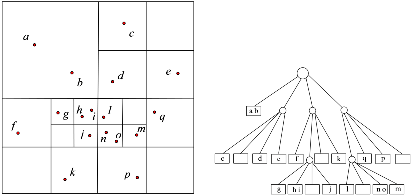
Talk Structure

Map Representation

Divide the map into:

- cells according to the historical destination density(leaves in the BPR Quadtree).
- cells in the uniform grids.

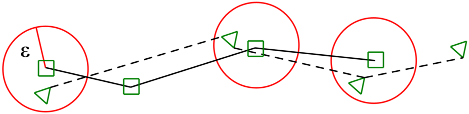
BPR Quadtree: a quadtree, recursively split into 4 blocks, can adapt to various densities of different regions.



Trajectories can be represented by a sequence of cells.

Trajectory Similarity

$$\mathsf{dsim}_\epsilon(T_a, T_b) = 1 - \mathsf{Maxm}_\epsilon(T_a, T_b)/\mathsf{clen}(T_a).$$

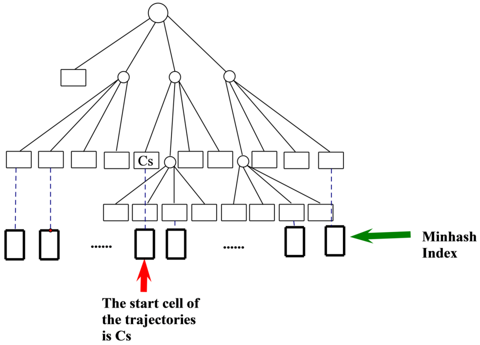


A cell trajectory T_a is similar to another cell trajectory T_b , if $\mathsf{dsim}_\epsilon(T_a, T_b) \leq \theta$.

Index construction

Minwise hashing: a class of Locality-sensitive hashing (LSH), used to quickly identify similar sets(Jaccard similarity).

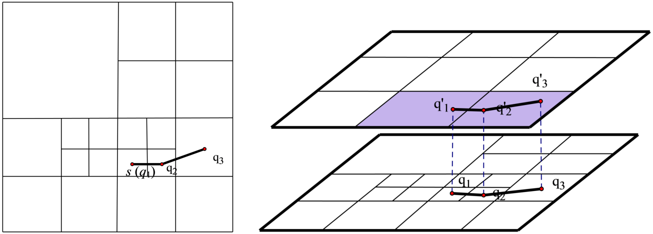
- Traverse the BPR Quadtree to every leaf.
- Build a Minhash index for each leaf.
- Insert trajectories (whose start cell is c) into the index I_c using a set of uniform cells.



Talk Structure

Candidate Trajectory Retrieval

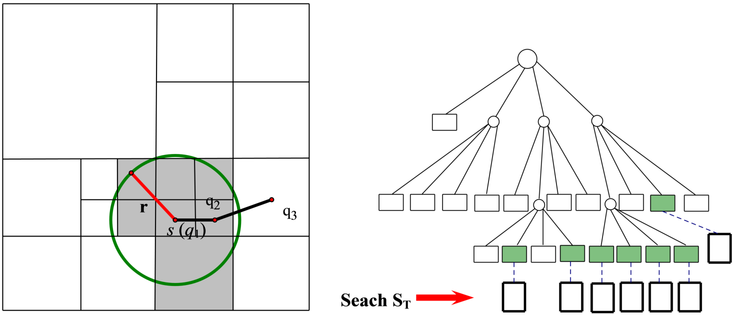
- Input: The partial cell trajectory, $T_q = q_1, q_2, \cdots, q_l$.
- Output: Candidate Trajectories $\mathcal{CT}$.



Transform cells in T_q into a set of grid cells S_T

Candidate Trajectory Retrieval

- Input: The partial cell trajectory, $T_q = q_1, q_2, \cdots, q_l$.
- Output: Candidate Trajectories $\mathcal{CT}$.



Prediction Algorithms

- Calculate the real similarity of the partial trajectory and the candidate trajectories.
- The predicted destinations can then be calculated using Observation 2.
- Instead of Choosing the most frequent destinations (FREQ), choose the cluster centers(CLUSTER).

Talk Structure

Experimental Evaluation

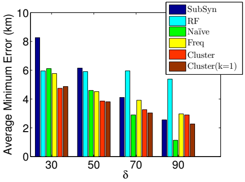
Real dataset:

- GPS points of 12000 taxis, 3 months, in the urban area about $50km \times 50km$ in Beijing.
- Preprocess the data to get the trajectories.
- Test set: randomly pick 1,000 trajectories.
- Cut the test trajectories into different length. The complement of a trajectory is δ.
- The number of returned destinations k.

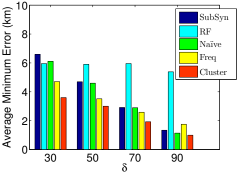
Compared algorithms

- Naïve : Predict the last cell of the partial trajectory as the destination.
- RF : Random Forests.
- SubSyn[1]: The modified Markov Chain approach, use the same information as ours.
- Cluster is the diameter cluster based algorithm, and Freq is the destination frequency based algorithm.
- Cluster ($k = 1$) is an extension prediction algorithm.

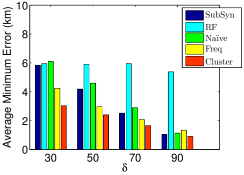
Accuracy Performance



(o) k=1



(p) k=3

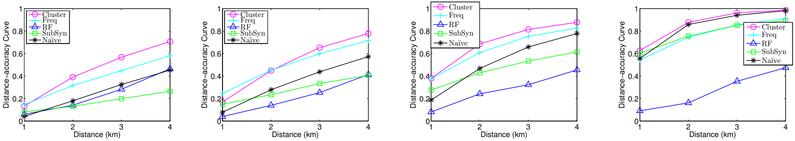


(q) k=5

Figure: Average minimum error. ($\epsilon = 500m$, $\theta = 0.15$, $r = 500m$).

AvgMinErr $= \frac{1}{n}\sum_{j=1}^{n}$ MinErr(q_j), $q_j \in Q$, where
MinErr$(q_j) = \min_{i\in[k]}$ $\mathsf{L}_1(d_{p_{ji}}, d_r)$.

Accuracy Performance



(a) $\delta = 30\%$ (b) $\delta = 50\%$ (c) $\delta = 70\%$ (d) $\delta = 90\%$

Figure: Distance-accuracy curve of $k = 3$ predicted destinations ($\epsilon = 500m$, $\theta = 0.15$, $r = 500m$).

DAV(λ) is the percentage of queries for which the error is no more than λ.

References

1. Andy Yuan Xue, elt. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In ICDE (2003).
2. Juan Antonio Alvarez-Garcia, elt. Trip destination prediction based on past GPS log using a Hidden Markov Model. Expert Systems with Applications 37, 12 (2010).
3. John Krumm and Eric Horvitz. Predestination: Inferring destinations from partial trajectories. In UbiComp (2006).
4. Josh Jia-Ching Ying, elt. Semantic trajectory mining for location prediction. In SIGSPATIAL (2011).
5. Jing Yuan, elt. Driving with knowledge from the physical world. In KDD (2011).

End

Thank you!