Mobility Sequence Extraction and Labeling Using Sparse Cell Phone Data

Yingxiang Yang
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, USA

Peter Widhalm
Austrian Institute of Technology
Giefinggasse 2, Vienna, Austria

Shounak Athavale
Ford Motor Company
Palo Alto, CA

Marta C. González
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, USA

Abstract

Human mobility modeling for either transportation system development or individual location based services has a tangible impact on people's everyday experience. In recent years cell phone data has received a lot of attention as a promising data source because of the wide coverage, long observation period, and low cost. The challenge in utilizing such data is how to robustly extract people's trip sequences from sparse and noisy cell phone data and endow the extracted trips with semantic meaning, *i.e.*, trip purposes. In this study we reconstruct trip sequences from sparse cell phone records. Next we propose a Bayesian trip purpose classification method and compare it to a Markov random field based trip purpose clustering method, representing scenarios with and without labeled training data respectively. This procedure shows how the cell phone data, despite their coarse granularity and sparsity, can be turned into a low cost, long term, and ubiquitous sensor network for mobility related services.

Introduction

Until the end of last century, the data source of transportation modeling had come from manual travel surveys, which are very expensive and have limited sample size and update frequency. But the wide spread of GPS and mobile phone usage has provided new perspectives to acquire such information (Gonzalez, Hidalgo, and Barabasi 2008; Wang et al. 2012). Compared with GPS and survey data, cell phone call detailed records (CDR) can provide much larger sample size and longer observation period. But the spatial measurement accuracy and the average temporal sampling rate for each individual are low. Another common shortcoming of both GPS and mobile phone CDR based trip extraction is the trip sequences are semantically poor. Neither trip purposes, *i.e.* the activities performed at trip destinations, nor the social-demographic information can be directly observed. Moreover, most existing studies on trip purpose labeling focus on trips extracted from GPS traces with labeled training data (Bohte and Maat 2009; Liao, Fox, and Kautz 2005), leaving the question what is the limit of activity labeling accuracy for trip sequences extracted from CDR data and without labeled training data.

Problem Definition

The notations and problems are defined as:

Notation: Mobile phone record r. A mobile phone record contains user id, timestamp, longitude, latitude.

Notation: Activity a. In the current study, activities are classified as home, work, leisure, shopping, other.

Notation: Stay S. A stay means performing an activity at a fixed location. A stay has properties start time $\tilde{t}_{arr}$, duration $\tilde{\delta}$, and location l.

Notation: Stay Sequence. A stay sequence represents the consecutive stays a person made in a day.

Problem 1. Stay sequence extraction. The task is to filter out passing by records (records made when a person is moving) and signal hyperspacing in the records and estimate stay start time $\tilde{t}_{arr}$, duration $\tilde{\delta}$ and location l for each extracted stay s.

Problem 2. Activity classification. Given the extracted stay sequences and labeled travel survey data. Identify the activity type a for each stay.

Problem 3. Activity clustering. When there is no labeled training data, we want to cluster the stays into meaningful categories based on properties such as stay start time, duration, nearby land use types, *etc*.

Methodology

Stay sequence extraction

In the first step we extract actual stay locations from cell phone records and filter out passing by points. We consecutively examines the cell phone records of a mobile device in their temporal order and incrementally creates and appends clusters of phone records with small distances. The arrival times and stay durations of the stays are estimated based on the timestamps of the cell phone records and by assuming constraints on the travel speed. Specifically, we set spatial and temporal thresholds on the stay extraction to filter out mobile signal jumps and records made when a person is moving. The outputs of the algorithm are the expected stay start time $\tilde{t}_{arr}$, expected duration $\tilde{\delta}$, maximum duration δ_{max}, minimum duration δ_{min}, and location l.

Bayesian classification method

The activity categories "home", "work", "leisure", "shopping", and "other" are denoted as number $i = 1, ..., 5$. The

classification problem deals with assigning the activity a of a stay S to one of the numbers in $i = 1,...,5$. We utilize the start time $\tilde{t}_{arr}$, duration $\tilde{\delta}$, location l of each stay and transition probability $P(i_1, i_2)$ between activities i_1 and i_2 $(i_1, i_2 = 1,...,5)$ from the labeled training data to construct the classifier. Among the 5 activity types, "home" and "work" have more distinct start time, duration and location distributions, while the other three activities are harder to distinguish based on these three features. Based on this characteristic of extracted trips, we perform a stepwise classification. The first step classifies the activities into "home", "work", and "the rest" based on activity start time, duration and location. Then in the second step we further distinguish "leisure", "shopping", and "other" while considering the transition probability between activities. For example, in an activity sequence $H - a_1 - a_2 - W - a_3 - H$, the first, fourth, and sixth activities have been identified to be home or work after step one. These activities divide the entire activity sequence into blocks of unknown labels and then we can decide the activity labels in each block simultaneously.

Markov random field clustering method

When there is no labeled training data for classification, we can still reveal the spatial-temporal structure of activities by performing clustering. For this we adopt a technique we first introduced in (Widhalm et al. 2015) which combines different properties of an activity such as activity start time, duration, location, and the number of visits to that location using an undirected graphic model, Markov random field. The inputs are: vectors $\mathbf{P} = (\mathbf{p}_1, ..., \mathbf{p}_n)$ of land use shares in proximity of the estimated locations; arrival times $\mathbf{t} = (t_1, ..., t_n)$; stay durations $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$, and the sequence of stay location indexes $\mathbf{i} = (i_1, ..., i_n)$ which number consecutively the distinct locations visited during a day. We define interrelations between the variables $(\mathbf{P}, \mathbf{t}, \boldsymbol{\delta}, \mathbf{i})$ describing an activity sequence as six cliques $C_1, ..., C_6$ with potential functions $\phi_1, ..., \phi_6$ representing the relationships between:

- C_1: prior probability of activity type A itself;
- C_2: activity type A and land use type L ;
- C_3: activity type A, starting time T and duration D of the activity;
- C_4: the activity type A and a binary indicator R that the activity location is visited more than once during the day;
- C_5: the activity type A and a binary indicator U_p that the activity is performed at only one unique location;
- C_6: an indicator U_a that only one unique activity is performed at location i_j.

Given the unsupervised nature of the problem, we use an EM (expectation-maximization) based learning scheme, to fit the potential functions to the data.

Preliminary Results

To quantitatively show how the clusters from the MRF method can be matched to the five survey defined classes, we use a travel survey data as input to the trained MRF cluster model and observe the confusion matrix in Table 1. The entire table sums up to 100%. Overall speaking, 66.2% of stays in the five clusters could be matched to their corresponding survey classes.While in travel surveys, sometimes activities are divided into three classes, "home", "work", and "other". If we use this simpler classification scheme, 93.9% activities could be matched to their corresponding clusters. While if labeled training data is available, in the Bayesian classification method. 79.4% of all the activities could be correctly classified, as in shown in Table 2. More importantly, the classification accuracy of the confounding classes, "leisure", "shopping", "other" increases dramatically to over 50%. They are 52.3%, 54.1% and 68.1% respectively.

Table 1: Confusion matrix between stay clusters C and survey activity classes

	C 1	C 2	C 3	C 4	C 5
Home	35.4%	0.3%	0.1%	0.0%	0.1%
Work	0.2%	14.3%	0.8%	0.3%	1.2%
Leisure	0.7%	0.8%	2.5%	3.1%	6.0%
Shopping	0.1%	0.0%	1.8%	2.4%	5.2%
Other	0.7%	0.4%	7.9%	3.6%	11.8%

Table 2: Confusion matrix between classified and real activity classes

	Class 1	Class 2	Class 3	Class 4	Class 5
Home	35.7%	0.1%	0.1%	0.0%	0.1%
Work	0.1%	15.1%	1.1%	0.2%	0.4%
Leisure	0.6%	1.2%	6.8%	1.9%	2.5%
Shopping	0.1%	0.1%	1.3%	5.2%	2.8%
Other	0.6%	0.8%	2.9%	3.4%	16.6%

References

[Bohte and Maat 2009] Bohte, W., and Maat, K. 2009. Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: A large-scale application in the netherlands. *Transportation Research Part C: Emerging Technologies* 17(3):285–297.

[Gonzalez, Hidalgo, and Barabasi 2008] Gonzalez, M. C.; Hidalgo, C. A.; and Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature* 453:779–782.

[Liao, Fox, and Kautz 2005] Liao, L.; Fox, D.; and Kautz, H. 2005. Location-based activity recognition using relational markov networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, 773–778. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Wang et al. 2012] Wang, P.; Hunter, T.; Bayen, A. M.; Schechtner, K.; and González, M. C. 2012. Understanding road usage patterns in urban areas. *Scientific reports* 2.

[Widhalm et al. 2015] Widhalm, P.; Yang, Y.; Ulm, M.; Athavale, S.; and González, M. C. 2015. Discovering urban activity patterns in cell phone data. *Transportation* 42(4):597–623.