Automated and Adaptive Activity-Travel Survey using Online Interaction with Travelers

Toru SEO
Postdoctoral Researcher, Asakura Lab., Tokyo Institute of Technology

Joint work with Dr. Kusakabe and Prof. Asakura

The 7th International BinN Seminar @ UTokyo
20160925
Travel behavior data is essential for activity/travel behavior modeling

- Actual activity/travel data is often collected by conducting a survey
  - realistic compared to stated preference (virtual) data
Conventional survey methods

**Questionary survey**

- Travel data is collected by paper- or web-based questionary

- Limitations:
  - **inaccurate** due to incomplete memory
  - **Long-term, large-scale, and accurate** data collection is difficult (panel attrition, fatigue)
Conventional survey methods

**GPS data collection**
- accurate spatiotemporal data
- Activity type (trip purpose) is missing

*source: Asakura and Hato (2004)*
**Conventional survey methods**

- GPS data collection + questionary survey
  - accurate spatiotemporal data
  - relatively accurate activity data
    - GPS log can help memory recalling
  - burden for participants
    - Extensive manual input is still mandatory

The activity type is asked by questionary survey.
Existing advanced survey methods

- **GPS data collection + imputation based on a priori info.**
  - Automatic imputation based on a priori information (eg: Wolf et al., 2001; Shen and Stopher, 2013; Gong et al., 2013)
    - land use
    - behavior model calibrated by using existing data
  - Collected data may have several limitation
    - Traveler heterogeneity is ignored
    - The data is not suitable for behavior modeling purposes
Existing advanced survey methods

GPS data collection + imputation based on offline info.
(Kim, Ben-Akiva, et al., 2014, 2015)

- Automatic imputation based on preliminary survey
  - Data is collected by preliminary survey for the same participants
  - Traveler heterogeneity can be captured
Proposed survey method

**GPS data collection + imputation based on online info.** (Kusakabe et al., 2015; Seo et al., 2016)
- Automatic imputation based on data collected from previous online interaction
  - online interaction: the survey system will ask a question automatically and dynamically

- Possible features of the proposed method:
  - Reducing frequency of questions
  - Keeping quality of data high
  - Considering traveler heterogeneity
  - Adaptive tracking of behavioral change
  - Automatic process (No need of manual control by survey administrator)
Illustration of methods

Introduction

Methodology

Empirical Validation

Conclusion

Manually collected

Automatically estimated

GPS + a priori info.

Learning process can differ

Learning process can track behavioral change

GPS + online info.

Manually collected

Automatically estimated

GPS + offline info.

Manually collected

Automatically estimated

Questionary survey

GPS survey

time
**Methodology**

- **Interaction** Determine whether ask a question or not
- **Estimation** Calculate estimation confidence
- **Interaction** Ask activity type
- **Interaction** Participant answers the activity type
- **Learning** Update estimation model based on the answer

\[
\hat{c} = \arg\max_c \ P(c|Y)
\]

\[
\hat{c} = \arg\max_c \ \prod_{k} P(x_k|c) P(c) \quad \{(c_1, Y_1), (c_2, Y_2), ... (c_i, Y_i), ...\}
\]

\[
P(\hat{c}|Y)
\]
Concept

- The proposed method estimates activity type
- The system always measures activity situation using standard sensors – date, time, location
- The system detects occurrence of an activity – move-or-stay identification
- The system can ask a question about activity to the survey participants

What are you doing now?

Are you working now?
The system can ask a question, if the system cannot estimate the activity.
The system learns a traveler-specific behavior pattern.
The system runs automatically.

Day 1, 9am, home to office

What are you doing?

Day 2, 9am, home to office

This must be commuting.

Commuting!

Good. It stops bothering me.

Day 3, 9am, home to somewhere else

What are you doing?

Enjoying my vacation!
Overview

- **Activity detection phase**
  - Detects the survey participant staying somewhere to do unknown activity
  - out of scope of this study

- **Estimation phase**
  - Estimates the activity type using an traveler-specific estimation model

- **Interaction phase**
  - Will ask activity type to the survey participant, if the estimation is not confident
  - Will not ask the question, if the estimation is confident

- **Learning phase**
  - Updates the estimation model based on the participant’s answer
Activity type estimation problem:
\[
\hat{c} = \arg\max_c P(c|Y)
\]
where
- activity type: \(c\)
  - eg: work, leisure
- activity situation: \(Y = (x_1, x_2, \ldots)\)
  - eg: time: \(x_1\), location: \(x_2\)

Naive Bayes assumption:
\[
\hat{c} = \arg\max_c \prod_k P(x_k|c)P(c)
\]

\(P(x_k|c)\) and \(P(c)\) are calculated based on historical data
= learning
Learning phase

- **Historical data:**
  \[ H_{\text{traveler, time}} = \{(c_1, Y_1), (c_2, Y_2), \ldots (c_i, Y_i), \ldots\} \]
  - traveler-specific
  - dynamically updated

- \( P(x_k | c) \) and \( P(c) \) can be easily calculated based on the historical data

- How to collect the historical data?
  → online interaction between the system and participant
Interaction phase

- Interaction:
  - If estimation confidence is high enough, the estimation result will be stored as a survey result.
  - Otherwise, the system will ask the actual activity to the survey participant.
    - The answer is stored to historical data for learning, as well as a survey result.

- Estimation confidence:
  \[ P(\hat{c}|Y) \]
  - probability of performing activity \( \hat{c} \) under situation \( Y \)
  - easily calculated based on historical data.

- The system will ask question with certain probability \( p_q \) so that expected error rate will be equal to a given acceptable error rate \( p_a \)
  - \( p_a \) is given by the survey planner
    \( \text{eg: } 5\% \)
Summary

- The survey admin will be happy if *Wrong* is small
- The survey participants will be happy if *Question* is small
  - It makes long-term survey easier
- The acceptable error rate $p_a$ is given by admin
  - quality control
  - trade-off: $Wrong \downarrow \iff Question \uparrow$
- Estimation model keeps being updated during the entire survey period, for each travelers (=online)
  - At the initial stage, the estimation model is dumb
    - *Question* will be frequent
  - As the survey progresses, the estimation model will become accurate
    - *Question* will decrease; *Correct* will increase
  - Long-term behavioral change can be tracked
  - Traveler heterogeneity can be captured
### Empirical Validation

#### Composition Ratio

<table>
<thead>
<tr>
<th>Date</th>
<th>Value</th>
<th>Correlation</th>
<th>Composition Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007/11/7</td>
<td>142</td>
<td>0.000000</td>
<td>0.7816</td>
</tr>
<tr>
<td>2007/11/8</td>
<td>146</td>
<td>0.000000</td>
<td>0.640988</td>
</tr>
<tr>
<td>2007/11/9</td>
<td>131</td>
<td>0.000000</td>
<td>0.625</td>
</tr>
<tr>
<td>2007/11/10</td>
<td>154</td>
<td>0.000000</td>
<td>0.818182</td>
</tr>
<tr>
<td>2007/11/11</td>
<td>97</td>
<td>0.000000</td>
<td>0.75</td>
</tr>
<tr>
<td>2007/12/17</td>
<td>315</td>
<td>0.000000</td>
<td>0.785714</td>
</tr>
<tr>
<td>2007/12/18</td>
<td>361</td>
<td>0.000000</td>
<td>0.833333</td>
</tr>
<tr>
<td>2007/12/19</td>
<td>355</td>
<td>0.000000</td>
<td>0.818182</td>
</tr>
<tr>
<td>2007/12/20</td>
<td>369</td>
<td>0.000000</td>
<td>0.787879</td>
</tr>
<tr>
<td>2007/12/21</td>
<td>368</td>
<td>0.000000</td>
<td>0.75</td>
</tr>
<tr>
<td>2007/12/22</td>
<td>252</td>
<td>0.000000</td>
<td>0.837888</td>
</tr>
<tr>
<td>2007/12/23</td>
<td>258</td>
<td>0.000000</td>
<td>0.818182</td>
</tr>
</tbody>
</table>

#### Chart

- **Question**
- **Correct**
- **Wrong**

The chart above shows the correlation values and composition ratios for different dates. The correlation values range from 0.000000 to 1.000000, while the composition ratios range from 0.000 to 1.000. The chart highlights key events and their corresponding values.
The proposed method is validated by using existing travel survey data
- collected by GPS + questionary survey

Validation procedure
1. Suppose that the survey data is true (ground truth)
2. Emulate the proposed method
3. Compare the estimation result with the ground truth

<table>
<thead>
<tr>
<th>Date</th>
<th>Dec 17-30, 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• duration: 2 weeks</td>
</tr>
<tr>
<td>Location</td>
<td>Matsuyama city, Japan</td>
</tr>
<tr>
<td>Number of participants</td>
<td>92</td>
</tr>
<tr>
<td>Number of trips</td>
<td>4120</td>
</tr>
<tr>
<td></td>
<td>• 3.5 trips/person/day</td>
</tr>
<tr>
<td>Activity type (trip purpose)</td>
<td>commuting, returning home, business, shopping, food/leisure, others</td>
</tr>
<tr>
<td>Activity situation</td>
<td>weekday dummy, arrival time, location</td>
</tr>
</tbody>
</table>
For weekdays, Question decreases and Correct increases as time progresses

- 37% Correct on the 12th day
- Number of Question was almost halved

Wrong is almost constant at 5%

- equal to the given acceptable error rate $p_a$

Commuting and returning home trips were easier to be estimated
Conclusion
Achievements

- Travel behavior data collection method is proposed for long-term and large-scale survey

- Features:
  - reduced frequency of questions
  - guaranteed quality of data
  - considers traveler heterogeneity
  - adaptive tracking of behavioral change
  - automatic process (no need of manual control by survey administrator)

- Methodology:
  - machine learning based on online interaction between the survey system and participant

- The proposed method is validated using existing travel data
  - It almost halved the frequency of questions on the 12th day
    - It may be reduced further if the survey period is longer
  - The error ratio was 5% as intended
References


