Semantic Labeling of Places based on Phone Usage Features using Supervised Learning

A. Rivero-Rodriguez, H. Leppäkoski, R. Piché

Tampere University of Technology
Tampere, Finland
www.tut.fi/posgroup

November 21, 2014
Corpus Christi, Texas, USA
UPIN-LBS
Context inference and awareness
This talk describes the design of the algorithms for a smartphone to learn your significant places.

- **Training data**
- **Features**
- **Classifiers**
MDC dataset

Training Data

Idiap and NRC-Lausanne
Lausanne Data Collection Campaign (2009-2011)
Records of 200 users over 18 months
Captures all types of information
Users provide extra information (labels!)
Anonymisation
46 GB of data!
Features Available

Training Data

**Active Phone Usage**
- calls, messages
- calendar, contacts
- application usage

**Passive Phone Usage**
- network information
- system information
- location & movement
The places were identified by clustering, then labeled by the user.
We selected 14 features that could be used by a place-labelling application:

**System**
- duration
- startHour
- endHour
- nightStay
- batteryAvg
- chargingTimeRatio
- sysActiveRatio
- sysActStartsPerHour

**Call logs**
- callsTimeRatio
- callsPerHour

**Accelerometer**
- idleStillRatio
- walkRatio
- vehicleRatio
- sportRatio

Features
We considered two different data representations

<table>
<thead>
<tr>
<th>1 - visits</th>
<th>2 - places</th>
</tr>
</thead>
<tbody>
<tr>
<td>User #1, Home, data ...</td>
<td>User #1, Home, data ...</td>
</tr>
<tr>
<td>User #1, Home, data ...</td>
<td>Work, data ...</td>
</tr>
<tr>
<td>User #2, Work, data ...</td>
<td>Other, data ...</td>
</tr>
<tr>
<td>User #3, Home, data ...</td>
<td>User #2, Home, data ...</td>
</tr>
<tr>
<td>User #2, Other, data ...</td>
<td>Work, data ...</td>
</tr>
<tr>
<td>User #1, Other, data ...</td>
<td>Other, data ...</td>
</tr>
<tr>
<td>User #3, Work, data ...</td>
<td>User #3, Home, data ...</td>
</tr>
<tr>
<td>User #1, Work, data ...</td>
<td>Work, data ...</td>
</tr>
<tr>
<td>User #3, Home, data ...</td>
<td>Other, data ...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

55,932 instances

295 instances
We preprocessed the data to obtain the features for both approaches.

**Features**

- Visits_20min.csv
- Places.csv

**Definitions for DB queries**

- Start times, end times, used ids, place labels

**Make queries**

- System
- Call logs
- Accel activity

**For each visit**

- Compute times, counts, averages

**For each user & place**

- Accumulate times & counts, weight averages

**Compute ratios**

- Feature vectors for visits
- Feature vectors for places
We applied five popular classification methods to the data

\[
P(X \mid A, B) = \frac{P(A \mid X) P(B \mid X) P(X)}{P(A) P(B)}
\]

**Naïve Bayes (NB)**

**Decision Tree (DT)**

**Bagged Tree (DT)**

**K-nearest neighbors (K-NN)**

**Neural Networks (NN)**
Results - Visits approach

Classifiers

<table>
<thead>
<tr>
<th></th>
<th>Well Classified</th>
<th>Misclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>70%</td>
<td>28%</td>
</tr>
<tr>
<td>W</td>
<td>82%</td>
<td>80%</td>
</tr>
<tr>
<td>O</td>
<td>60%</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>70%</td>
<td>28%</td>
</tr>
</tbody>
</table>

NB: 53%
DT: 75%
BT: 77%
NN: 61%
KNN: 58%

H: Home
W: Work
O: Others
Results - Places approach

Classifiers

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Well Classified</th>
<th>Misclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>81%</td>
<td></td>
</tr>
<tr>
<td>BT</td>
<td>85%</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td>71%</td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td>71%</td>
<td></td>
</tr>
</tbody>
</table>

H: Home
W: Work
O: Others
Naive Bayes and Bagged Decision Tree with Places data-representation are best
NN and K-NN underperform and are computationally demanding
Most relevant features are: night stay, stay duration, start time, battery status, idle time

Other classifiers (logistic regression, support vector machine)
Combine Places and Visits data-representations