Putting Ubiquitous Devices' Data to Use

@neal_lathia
University of Cambridge
January 23, 2013
what data can we collect?
recommender systems
aim to match users to items that will be of interest to them
recommender systems
aim to match users mobility profiles to items social events that will be of interest to them
use mobility data to recommend social events

(1) infer attendance at events
(2) recommend (test 6 different algorithms)
(3) evaluate recommendation quality
task

(1) get users' data, split temporally
(2) run algorithm that outputs recommendations...
(3) evaluate the quality of the recommendations
algorithms

(1) popular events (in the city)
(2) geographically close
(3) popular events (where you live)
(4) TF-IDF
(5) k-Nearest Locations
(6) k-Nearest Events
what is a good recommendation?
what is a good recommendation?

evaluate by **ranking**: are the events you went to 'near' the top of the recommendation list?

metric: percentile ranking. small value = good. high value = bad.
Average rank for different events

- Shakespeare: 0.13, Popular(1)
- Red Sox: 0.28, Popular in Area(3)
- Cirque: 0.29, TF-IDF(4)
- Friday Nights: 0.37, TF-IDF(4)
- Summer Concerts: 0.40, TF-IDF(4)
- Friday Flicks: 0.41, TF-IDF(4)
future

how would you use other smartphone sensors to improve recommendations?
what tools could we design to help travellers?
sensing mobility:
5%-sample, 2 x 83-days

time-stamped location (entry, exit), modality payments (top-ups, travel cards) card-types (e.g., student)
## Adult

<table>
<thead>
<tr>
<th>Zone</th>
<th>Cash</th>
<th>Oyster pay as you go</th>
<th>Travelcards</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Peak single</td>
<td>Day Anytime</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Off-peak single</td>
<td>Day Off-peak</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peak price cap</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Off-peak price cap</td>
<td></td>
</tr>
</tbody>
</table>

| Zone 1 only only      | £4.00| £1.90                 | £8.00       | £8.00 | £27.60  | £106.00| £1,104 |
| Zones 1-2             | £4.00| £2.50                 | £8.00       | £8.00 | £27.60  | £106.00| £1,104 |
| Euston - Zone 2*      | £4.00| £2.00                 | £8.00       | £8.00 | £27.60  | £106.00| £1,104 |
| Zones 1-3             | £4.00| £2.90                 | £10.00      | £10.00| £32.20  | £123.70| £1,288 |
| Euston - Zone 3*      | £4.00| £2.70                 | £10.00      | £10.00| £32.20  | £123.70| £1,288 |
| Zones 1-4             | £5.00| £3.40                 | £10.00      | £10.00| £39.40  | £151.30| £1,576 |
| Euston - Zone 4*      | £5.00| £3.10                 | £10.00      | £10.00| £39.40  | £151.30| £1,576 |
| Zones 1-5             | £5.00| £4.10                 | £15.00      | £15.00| £47.00  | £180.50| £1,880 |
| Euston               | £5.00| £3.80                 | £15.00      | £15.00| £47.00  | £180.50| £1,880 |
questions

(1) what is the relation between how we travel & how we spend?
(2) do travellers make the correct decisions? (no)
(3) can we help them with recommendations? (yes)
<table>
<thead>
<tr>
<th>(%)</th>
<th>pay as you go purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.8</td>
<td>&lt; 5 GBP</td>
</tr>
<tr>
<td>24.2</td>
<td>5 – 10 GBP</td>
</tr>
<tr>
<td>15.5</td>
<td>10 – 20 GBP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(%)</th>
<th>travel card purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>70.8</td>
<td>7-day travel card</td>
</tr>
<tr>
<td>15.8</td>
<td>1-month travel card</td>
</tr>
<tr>
<td>11.6</td>
<td>7-day bus/tram pass</td>
</tr>
</tbody>
</table>

**Ongoing Trips**

**Purchase Behaviour**

- Travel Cards
- PAYG
Purchase Geography

Mobility Flow

Zone 1
Zone 2
Zone 3
Zone 4
Zone 5
Zone 6

arrive
depart

PAYG
Travel Cards
the data shows that:
(a) there is a high regularity in travel & purchase behaviour
(b) travellers buy in small increments and short-terms
(c) most purchases happen upon refused entry
(2) do travellers make the correct decisions? compare actual purchases to the optimal (per traveller)

how:
(a) clean data
(b) build & search on a tree ~ sequence of choices
how: build a tree with each user's mobility data where a node is a **purchase (expire, cost)** that is expanded when it has expired (reduced) example:
how: build a tree with each user's mobility data where a node is a **purchase** (expire, cost) that is expanded when it has expired (reduced) example:

![Tree diagram with PAYG, £aa.aa, 7-day £bb.bb, 30-day £cc.cc nodes]

we reduce the space-complexity of searching on this tree by implementing expansion rules, pruning heuristics
the cheapest sequence of fares can then be compared to what the user actually spent

PAYG, £aa.aa

PAYG, £aa.aa

7-day £bb.bb

30-day £cc.cc
the cheapest sequence of fares can then be compared to what the user actually spent

in each 83-day dataset, the 5% sample of users where overspending by $\sim \£2.5$ million

An estimate of how much everybody (100%) is overspending during an entire year (365 days) is thus $\£200$ million
overspending comes from
(a) failing to predict one's own mobility needs
   ...but we have observed that mobility is predictable
(b) failing to match mobility with fares (in a complex fare system)
   ...which is an easy problem for a computer

can we help travellers?
recommender systems
aim to match users to items that will be of interest to them
recommender systems aim to match users mobility profiles to items fares that will be of interest the cheapest for them
three steps
1. for a given set of travel histories, compute the cheapest fare (by tree expansion)
2. reduce each travel history into a set of generic features, describing the mobility (next slide)
3. train classifiers to predict the cheapest fare given the set of features
we have a set of \{d, f, b, r, pt, ot, N\} = F

where

d = number of trips
f = average trips per day
b / r = proportion of trips on the bus / rail
pt / ot = proportion of peak & off-peak trips
N = zone O-D matrix
F = cheapest fare (label)
two baselines, three algorithms:
0. baseline – everyone on pay as you go
1. naïve bayes – estimating probabilities
2. k-nearest neighbours – looking at similar profiles
3. decision trees (C4.5) – recursively partitions data to infer rules
4. oracle – perfect knowledge
<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Savings (GBP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dataset 1</td>
<td>Dataset 2</td>
</tr>
<tr>
<td>Baseline</td>
<td>74.99</td>
<td>76.91</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>77.46</td>
<td>80.71</td>
</tr>
<tr>
<td>k-NN (5)</td>
<td>96.74</td>
<td>97.09</td>
</tr>
<tr>
<td>C4.5</td>
<td>98.01</td>
<td>98.29</td>
</tr>
<tr>
<td>Oracle</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

- Baseline: Accuracy: 74.99% for Dataset 1, 76.91% for Dataset 2; Savings: £326,447.95 for Dataset 1, £306,145.85 for Dataset 2.
- Naïve Bayes: Accuracy: 77.46% for Dataset 1, 80.71% for Dataset 2; Savings: £393,585.81 for Dataset 1, £369,232.24 for Dataset 2.
- k-NN (5): Accuracy: 96.74% for Dataset 1, 97.09% for Dataset 2; Savings: £465,822.17 for Dataset 1, £426,375.85 for Dataset 2.
- C4.5: Accuracy: 98.01% for Dataset 1, 98.29% for Dataset 2; Savings: £473,918.38 for Dataset 1, £434,082.81 for Dataset 2.
- Oracle: Accuracy: 100% for both datasets; Savings: £479,583.91 for Dataset 1, £438,923.30 for Dataset 2.
station interest ranking

current system: free travel alerts – manually set up by traveller

future system: predict (and rank) the stations that travellers will visit in their future trips for personalised notifications
station interest ranking

can we automate this?

baseline: rank by visit popularity

proposal: station similarity neighbourhood (visit co-occurrence) and traveller trip history
station interest ranking

Percentile Ranking Results

- Baseline: [0.2467, 0.2561]
- User History: [0.0642, 0.0611]
- Station NN: [0.0591, 0.0565]
accurate ranking
without knowing who travellers are, the network topology, train schedule, disruptions and closures, we designed: no context
today:
(a) mobile location recommendations
(b) fare purchase recommendations
(c) travel alerts
Further reading:


N. Lathia, L. Capra. Mining Mobility Data to Minimise Travellers' Spending on Public Transport. In ACM KDD 2011, San Diego, USA.


Android app to try: www.tubestar.co.uk
Questions?

@neal_lathia

neal.lathia@cl.cam.ac.uk