User similarity based on trajectory

Eric Bach

February 17, 2014
Similarity between users

... from the information-type point of view

Relationships

- direct connection
- knowing by the same people
- being known by the same people

Places

- friends

Interests

- same friends
- famous people
Similarity between users

... from the information-type point of view

- physically (also passed by)
  - shops, museums, meetings
- virtually
  - online communities, blogs, web-pages
Similarity between users

... from the information-type point of view

- derived from places
- derived from movement
- device usage-profile

partying, studying, coding
biking, jogging
gaming, working
Similarity between users
... from the information-source point of view

Active
• online behavior
• social behavior
• consumer behavior
• usage behavior

Passive
• GPS
• cell-site
• public WLAN’s
• tracker for online behavior
Similarity between users
... from the exploitation point of view

Two particular users

- comparison of people  friend recommendation, expert search
- identification of people “Are two different people/profiles the same person?”

Segmentation of users

- collaborative filtering personalized advertisement
- group-analysis “What kind of people using my service?”

Abstract user pattern

- hidden class model “Is it possible to categorize people?”
Today

using mobile devices to access users’ trajectory and interests

Information-type

- sequence of physical places $\leadsto$ trajectories [5, 8, 6]
- interests $\leadsto$ mobile phone usage-logging [6]

Information-source

- passive $\leadsto$ GPS, cell-site [5, 8, 6]
- active $\leadsto$ mobile phone usage-logging [6]

Exploitation

- comparison of people $\leadsto$ friend recommendation [5, 8]
- general user pattern $\leadsto$ raw similarity [6]
Today

Data

Preprocessing

Similarity calculation

Model approach

Summary
Today

Data

Preprocessing

Similarity calculation

Model approach

Summary
Source of trajectory data I

different accuracy / availability

Cell-site

- **Pros:** availability (focus on mobile phones)
- **Cons:** requires information about cell tower position, accuracy varies (from $m^2$ to $km^2$)

Increasing accuracy related to **range**, **selectivity** and **density** of cell-towers

Omnidirectional antenna  120° sector selective antenna  more than one cell-tower
Source of trajectory data II

different accuracy / availability

GPS

- **Pros**: high accuracy (few \( m \) [4])
- **Cons**: no satellite-connection within buildings, energy consumption (in mobile phones) [1]

WLAN access-points

- **Pros**: high accuracy (few \( m \) [2]), works withing buildings
- **Cons**: energy consumption [1], requires information about access-point position
Data structure

representation of a trajectory

User log (matrix)

\[
\text{user} = \begin{bmatrix}
  t_1 & pos_1 = \langle \text{lat}, \text{long} \rangle & act_1 \\
  t_2 & pos_2 = \langle \text{lat}, \text{long} \rangle & act_2 \\
  \vdots \\
  t_m & pos_m = \langle \text{lat}, \text{long} \rangle & act_m
\end{bmatrix}
\]

\(act_j\) . . . user’s activity

NULL, playing, web-surfing, shopping (maybe related to \(pos_j\))

User log (list)

\[
\text{user} : tr_1 = \{ t_1, pos_1, act_1 \} \to \ldots \to tr_m = \{ t_m, pos_m, action_m \}
\]
Direct comparing trajectories

... it may be meaningless

Uniqueness of trajectories

[... \textbf{four} spatial-temporal points are enough to uniquely identify 95\% if the individuals.” [3]

Sparseness of high dimensional spaces

\begin{figure}
\centering
\includegraphics[width=\textwidth]{graph.png}
\caption{followers (similar user) according to high dimensional data [6]}
\end{figure}
Direct comparing trajectories I

... and the challenging details

problem: How to handle spatial / temporal close locations?

problem: Most probable location is different, even if user visited the same semantic location.
**Direct comparing trajectories II**

... and the challenging details

*problem*: Even if the trajectories / locations are completely different, it is reasonable, that both person are similar, if one considers the semantic locations.
Today

Data

Preprocessing

Similarity calculation

Model approach

Summary
Spatial & temporal clustering per user

How to handle spatial / temporal close locations?

Define stay-points [8, 5]

\[ D_{pos}(tr_j^{(p)}, tr_j^{(q)}) \leq thr_{pos} \quad p, q \ldots \text{two different user trajectories} \]

• drop: close loc. with short stay-time: \[ D_{time}(tr_j^{(p)}, tr_j^{(q)}) \leq thr_{time} \]
  riding a car along a street

• cluster: close loc. with long stay-time: \[ D_{time}(tr_j^{(p)}, tr_j^{(q)}) > thr_{time} \]
  semantic location with multiple locations

\[ D_{pos}(tr_j^{(p)}, tr_j^{(q)}) = 0 \]

• cluster: adjacent identical locations variance in locations

most probable locations

trajectories

semantic locations
Abstraction from physical location

Location is different, even if user visited the same semantic location.

Most probable locations

Semantic locations

Transformed user log (list)

\[
\text{user} : \ tr_j = \{ t_j, \ pos_j, \ act_j \} \sim \ str_j = \{ t_j, \ \text{semanticpos}_j, \ act_j \}
\]

\text{semanticpos} \ldots \text{semantic position} \quad \text{School, Cinema, Park, ...}
Today

Data

Preprocessing

Similarity calculation

Model approach

Summary
Similarity between two trajectories based on the longest common subsequence (LCS)

Subsequence
“[...] a subsequence is a sequence that can be derived from another sequence by deleting some elements without changing the order of the remaining elements.”\(^1\)

Longest common subsequence (LCS)
“[...] longest subsequence to all sequences in a set of sequences [...]”\(^1\)

Similarity value

• LCS based on the semantic positions
• averaged coverage of the particular trajectories by the LCS

\[
D \left( str(p), str(p) \right) = \frac{1}{2} \left( \frac{\text{length}(LCS)}{\text{length}(str(p))} + \frac{\text{length}(LCS)}{\text{length}(str(q))} \right)
\]

\(^1\)wikipedia: Longest common subsequence problem, Subsequence
Similarity between two trajectories

Example from [8]

- users: \( p, q \), sequences: \( str(p), str(q) \), \( sp \): get sem. pos.
  - \( str(p) \supseteq \langle \{School\}, \{Cinema\}, \{Park, Bank\}, \{Restaurant\} \rangle \)
  - \( str(q) \supseteq \langle \{School, Market\}, \{Park\}, \{Restaurant\} \rangle \)
  - \( LCS (str(p), str(q)) \supseteq \langle \{School\}, \{Park\}, \{Restaurant\} \rangle \)

Similarity value

- \( D_{LCS} (str(p), str(q)) = \frac{1}{2} \left( \left\lvert \frac{3}{4} \right\rvert + \left\lvert \frac{3}{3} \right\rvert \right) = \frac{7}{8} \)

Note (see [8])

- unclear semantic location \( \{School, Market\} \)
- different reliability of the trajectories (keyword: “support”)
- more than one trajectory per user: \( D^*_{LCS} (STR(p), STR(q)) \)
Today

Data
  trajectory
  context-logs

Preprocessing
  clustering, abstraction, ...

Similarity calculation
  all data
  general behavior pattern, ...

Model approach
  pair of user
  similarity

Summary

References
Model-driven approach

based on constrained Bayesian Factor Analysis [7, 6]

Bayesian Factor Analysis

- unsupervised method with parameter $k$ as the amount hidden classes
  
  student, early-bird, party-animal, sportsman

- user is linear mixture of hidden classes
  
  $user_i = \text{student} + \text{early-bird} + \text{party-animal}$

- **constrained** limiting the parameter space
  
  positive mixture coefficients $\in [0, 1]$, sum of mixture coef. $= 1$

Mathematical model

Model: $P(X|\Lambda, Z, \sigma^2) = \prod_{i=1}^{m} \prod_{j=1}^{n} [\mathcal{N}(X_{ij} | \Lambda_i Z_j, \sigma^2)]^{l_{ij}}$

$X \in \mathbb{R}^{m \times n}$... $n$ user and $m$ feature, $\Lambda \in \mathbb{R}^{m \times k}$... hidden classes

$Z \in \mathbb{R}^{k \times n}$... user specific mixture, $\sigma^2$... noise variance
Similarity between two trajectories

based on the Likelihood

Distance in the hidden class space

- calculate most probable hidden class (HC) according to Likelihood (prev. slice): \( HC(\cdot) \)
- distance between hidden classes: \( D(HC(str^{(p)}), HC(str^{(q)})) \)
  cosine, euclidean

Distance in the mixture distribution space

- distance between distributions Kullback-Leibner (KL)
- get the mixture distribution (MD): \( MD(str^{(p)}) = Z_p \)
- distance betw. dist.: \( D_{KL}(str^{(p)}, str^{(q)}) = KL(Z_p, Z_q) \)
- more than one trajectory per user: \( D^*_{KL}(STR^{(p)}, STR^{(q)}) \)
Summary

User similarity
- based on different kind users’ properties relationships, places, interests
- different fields of application friend recommendation, personalized advertisement

Data / Preprocessing
- different data sources (availability / accuracy) GPS, cell site, WLAN
- spatial / temporal clustering (averaging)
- abstraction from location to semantic location School, Cinema, Museum, ...

Similarity calculation
- direct comparing of trajectory sequences longest common sequence
- model based approach factor model
Questions?


Conference on Advances in Geographic Information Systems, GIS ’08, pages 34:1–34:10, New York, NY, USA, 2008. ACM.

