User similarity based on trajectory

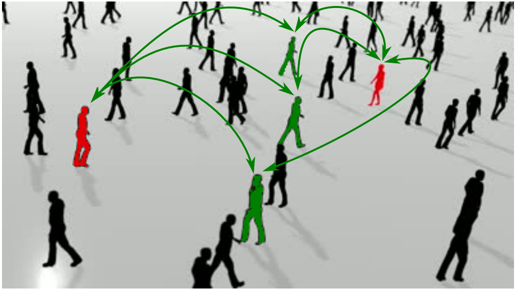
Eric Bach

February 17, 2014

Similarity between users

... from the information-type point of view

Relationships Places Interests



- direct connection — friends
- knowing by the same people — same friends
- being known by the same people — famous people

Similarity between users

... from the information-type point of view

Relationships Places Interests

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

Similarity between users

... from the information-type point of view

Relationships

Places

Interests

- derived from places — partying, studying, coding
- derived from movement — biking, jogging
- device usage-profile — gaming, working

Similarity between users

... from the information-source point of view

Active

- online behavior — blogs, web-pages
- social behavior — facebook
- consumer behavior — amazon
- usage behavior — xfire, spec. logging-software

Passive

- GPS — phones, GPS-receiver
- cell-site — phones
- public WLAN's — computer, phone
- tracker for online behavior — cookies

Similarity between users

... from the exploitation point of view

Two particular user

- 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 $\rightsquigarrow$ trajectories [5, 8, 6]
- interests $\rightsquigarrow$ mobile phone usage-logging [6]

Information-source

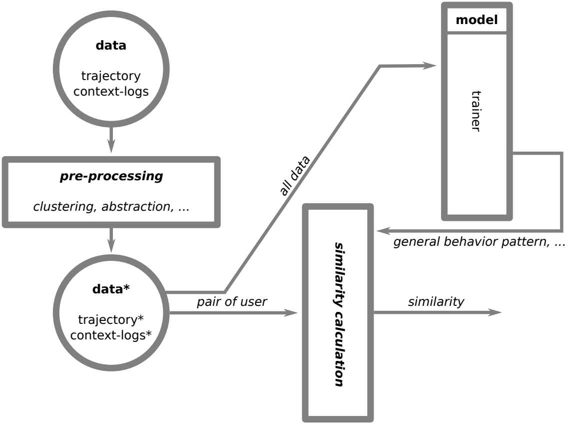
- passive $\rightsquigarrow$ GPS, cell-site [5, 8, 6]
- active $\rightsquigarrow$ mobile phone usage-logging [6]

Exploitation

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

Today

- Data
- Preprocessing
- Similarity calculation
- Model approach
- Summary



Today

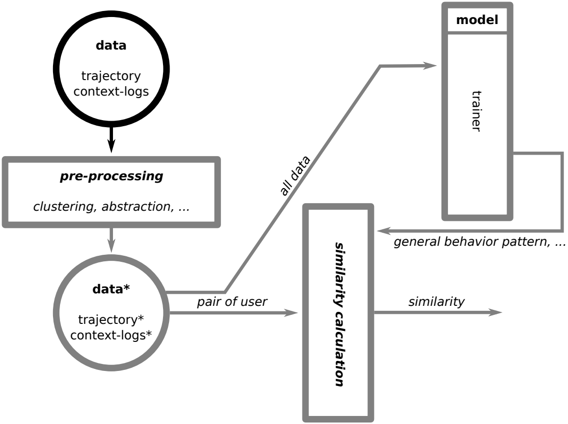
Data

Preprocessing

Similarity calculation

Model approach

Summary

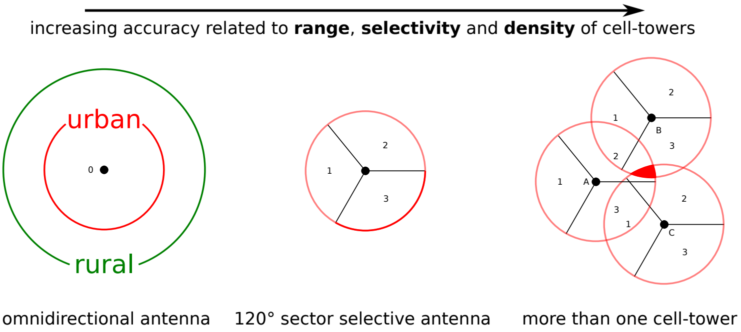


Source of trajectory data I

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)



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

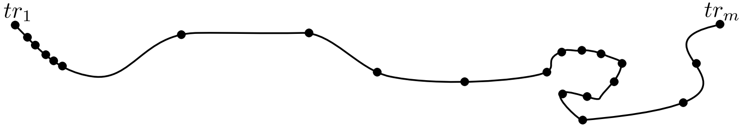
User log (matrix)

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

$act_j \ldots$ user's activity
NULL, playing, web-surfing, shopping (maybe related to pos_j)

User log (list)

$$user : tr_1 = \{t_1, pos_1, act_1\} \to \ldots \to tr_m = \{t_m, pos_m, action_m\}$$



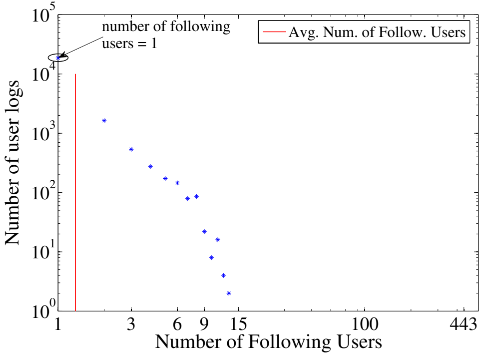
Direct comparing trajectories

... it may be meaningless

Uniqueness of trajectories

[...] **four** spatial-temporal points are enough to uniquely identify 95% if the individuals." [3]

Sparseness of high dimensional spaces

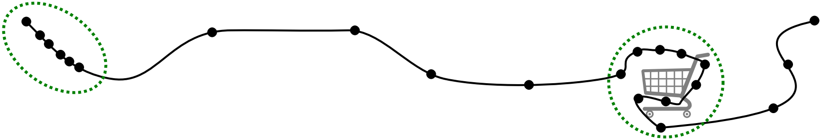


Graph: followers (similar user) according to high dimensional data [6]

Direct comparing trajectories I

... and the challenging details

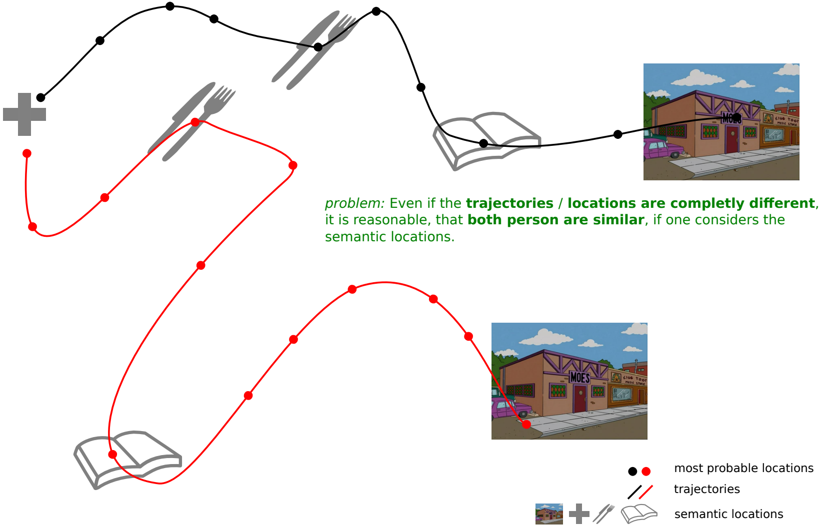
problem: How to handle **spatial / temporal close locations**?



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 completly different**, it is reasonable, that **both person are similar**, if one considers the semantic locations.

Today

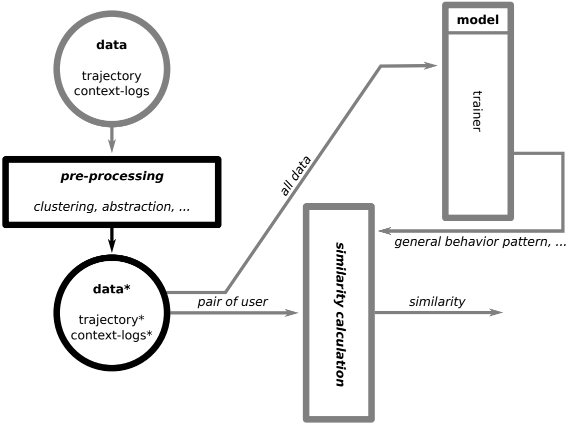
Data

Preprocessing

Similarity calculation

Model approach

Summary



Spatial & temporal clustering per user

How to handle spatial / temporal close locations?

Define stay-points [8, 5]

$D_{pos}\left(tr_j^{(p)}, tr_j^{(q)}\right) \leqq thr_{pos}$ $p, q \ldots$ two different user trajectories

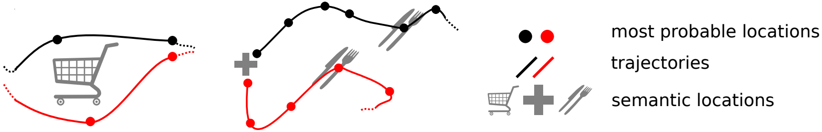
- drop: close loc. with short stay-time: $D_{time}\left(tr_j^{(p)}, tr_j^{(q)}\right) \leqq thr_{time}$
  riding a car along a street
- cluster: close loc. with long stay-time: $D_{time}\left(tr_j^{(p)}, tr_j^{(q)}\right) > thr_{time}$
  semantic location with multiple locations

$D_{pos}\left(tr_j^{(p)}, tr_j^{(q)}\right) = 0$

- cluster: adjacent identical locations variance in locations

Abstraction from physical location

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



Define social-locations [6, 8, 5]

- using external sources `http://maps.google.com`
- using heuristics (frequency / time / heterogeneous)
  at home in the night , at work at the day, private vs. public

Transformed user log (list)

$$user : tr_j = \{t_j, pos_j, act_j\} \curvearrowright str_j = \{t_j, semanticpos_j, act_j\}$$

semanticpos . . . semantic position School, Cinema, Park, ...

Today

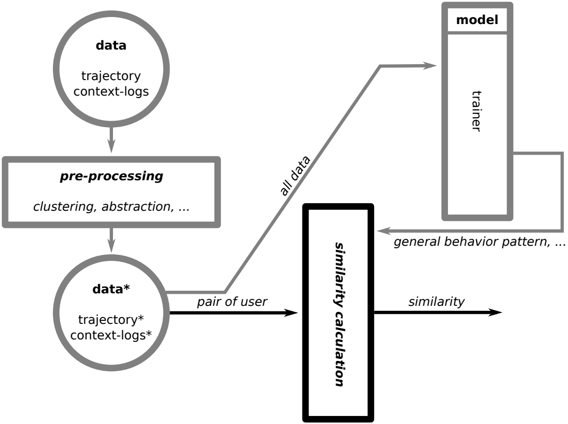
Data

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
- $D\left(str^{(p)}, str^{(p)}\right) = \frac{1}{2}\left(\frac{length\left(LCS\right)}{length\left(str^{(p)}\right)} + \frac{length\left(LCS\right)}{length\left(str^{(q)}\right)}\right)$

[1] wikipedia: Longest common subsequence problem, Subsequence

Similarity between two trajectories

example

Example from [8]

- users: p, q, sequences: $str^{(p)}, str^{(q)}$, $\stackrel{sp}{=}$: get sem. pos.
- $str^{(p)} \stackrel{sp}{=} \langle\{School\}, \{Cinema\}, \{Park, Bank\}, \{Restaurant\}\rangle$
- $str^{(q)} \stackrel{sp}{=} \langle\{School, Market\}, \{Park\}, \{Restaurant\}\rangle$
- $LCS\left(str^{(p)}, str^{(q)}\right) \stackrel{sp}{=} \langle\{School\}, \{Park\}, \{Restaurant\}\rangle$

Similarity value

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

Note (see [8])

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

Today

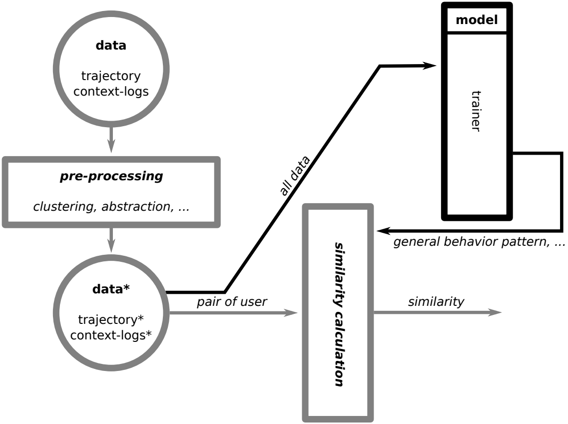
Data

Preprocessing

Similarity calculation

Model approach

Summary



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
  $user_i =$ student + early-bird + party-animal
- *constrained* limiting the parameter space
  positive mixture coefficients $\in [0, 1]$, sum of mixture coef. $= 1$

Mathematical model

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

$X \in \mathbb{R}^{m \times n}$ ... n user and m feature, $\Lambda \in \mathbb{R}^{m \times k}$ ... hidden classes
$Z \in \mathbb{R}^{k \times n}$ ... user specific mixture, σ^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: $D\left(HC\left(str^{(p)}\right), HC\left(str^{(q)}\right)\right)$ cosine, euclidean

Distance in the mixture distribution space

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

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

Questioins?

[1] Aaron Carroll and Gernot Heiser. An analysis of power consumption in a smartphone. In *Proceedings of the 2010 USENIX Conference on USENIX Annual Technical Conference*, USENIXATC'10, pages 21–21, Berkeley, CA, USA, 2010. USENIX Association.

[2] Yongguang Chen and Hisashi Kobayashi. Signal strength based indoor geolocation. In *Communications, 2002. ICC 2002. IEEE International Conference on*, volume 1, pages 436–439, 2002.

[3] Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, March 2013.

[4] Chen Fränti and Tabarcea. Four aspects of relevance: content, time, location and network. Technical report, Department of Computer Science University of Eastern Finland, 2014.

[5] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International*

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

[6] Haiping Ma, Huanhuan Cao, Qiang Yang, Enhong Chen, and Jilei Tian. A habit mining approach for discovering similar mobile users. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 231–240, New York, NY, USA, 2012. ACM.

[7] Mikkel N. Schmidt. Linearly constrained bayesian matrix factorization for blind source separation. pages 1624–1632, Dec 2009.

[8] Josh Jia-Ching Ying, Eric Hsueh-Chan Lu, Wang-Chien Lee, Tz-Chiao Weng, and Vincent S. Tseng. Mining user similarity from semantic trajectories. In *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, LBSN '10, pages 19–26, New York, NY, USA, 2010. ACM.