Microsoft® Research

Microsoft Research Asia Faculty Summit 2012

Understanding User Location in Mobile Social Networks

Xing Xie

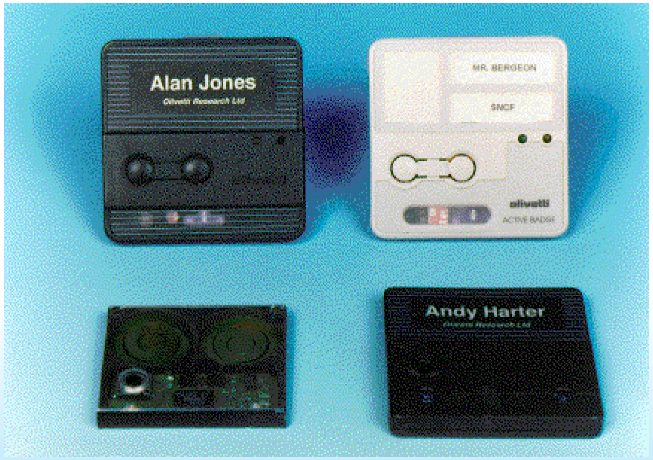
Microsoft Research Asia
Faculty Summit 2012

User Location Prediction

- Three basic questions of philosophy from concierges
  - Who are you?
  - Where do you come from?
  - Where do you go to?
- They can be great research problems too

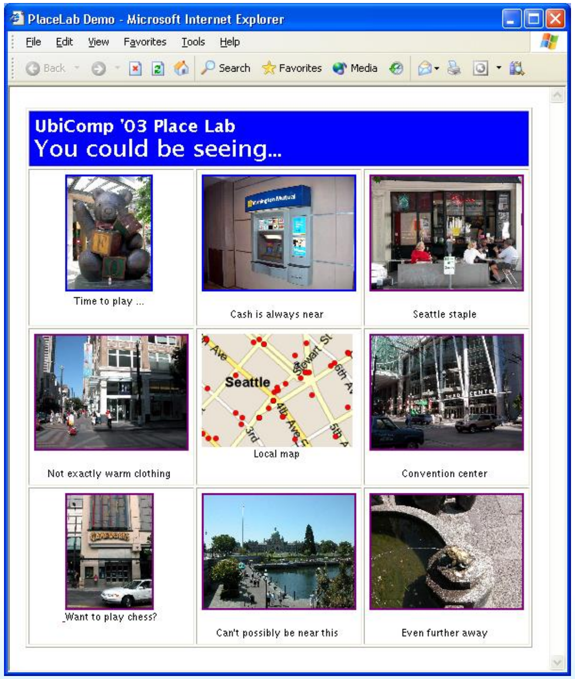
Active Badges (Olivertti Research, 1989)

- First automated indoor location system
- The small device worn by personnel transmits a unique infra-red signal every 10 seconds.
- Each office within a building is equipped with one or more networked sensors which detect these transmissions.



Ubicomp Research Projects

- RADAR (Microsoft, 2000)
  - Wi-Fi signal-strength based indoor positioning system
- Place Lab (Intel, 2003)
  - Low-cost, easy-to-use device positioning for location-enhanced computing applications
  - GSM tower, Bluetooth, 802.11 access points



UbiComp '03 Place Lab
You could be seeing...

Sensors Are Becoming Ubiquitous

- 85% of mobile devices will ship with GPS by 2013
- By 2013, 50% of mobile devices will ship with accelerometers and ~50% with gyroscopes
- Shipments of mobile motion sensors (accelerometers, compasses, gyroscopes, and pressure sensors) will reach 2.2B units in 2014, up from 435.9M in 2009.
- Contextual Computing will be a $160B market by 2015

User Location Data

- User location exist in various type of data
  - Geo-tagged photos, tweets and travelogues
  - Location based search logs
  - Map service logs

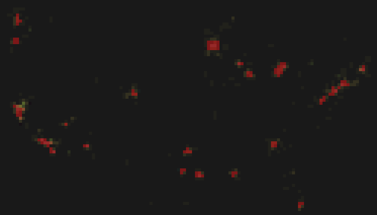
- There is no unified mechanism for managing these location data from different devices, different services and different users

Mobile Social Networks

- In social networks, people proactively share their feelings, interests, activities and photos with their friends. Many of them explicitly or implicitly contain user location information

- Location based social networks
  - Or called check-in services
  - Share location or location related information with each other
  - Generate huge user location data set

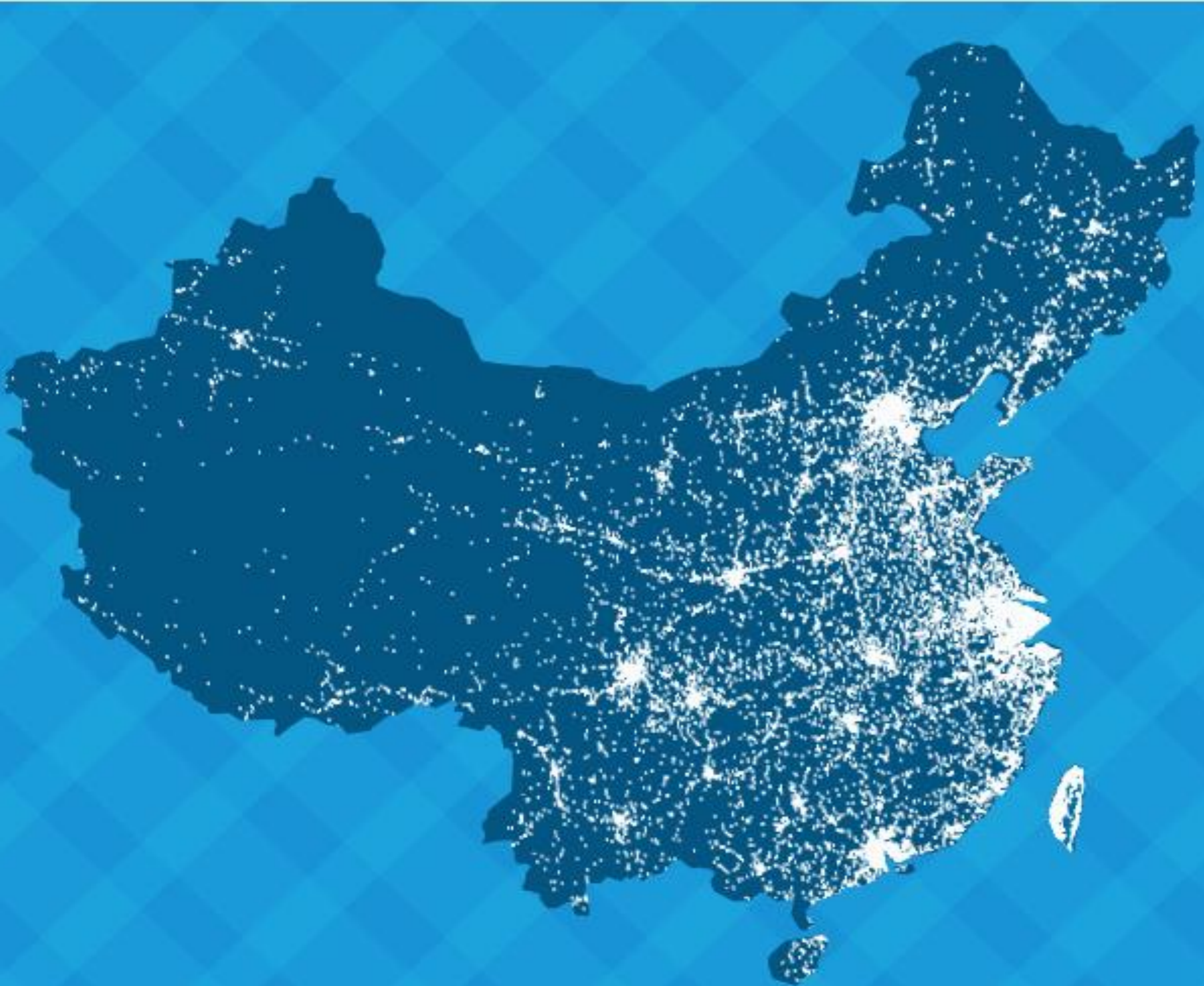
11/2008



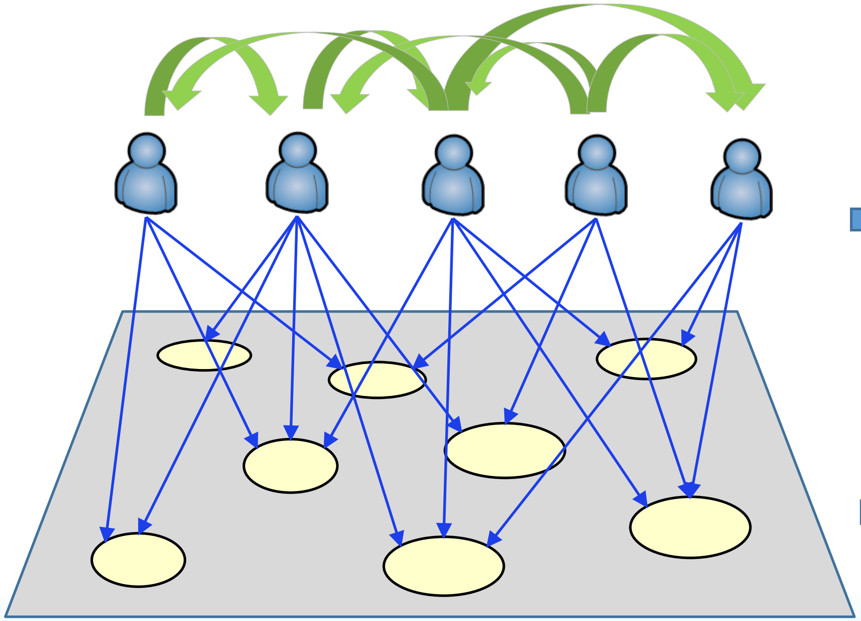
2 billion check-ins from Foursquare users

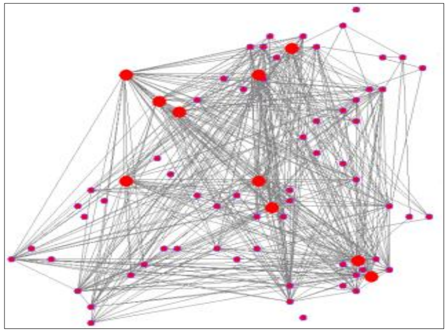
街旁
Jiepang.com

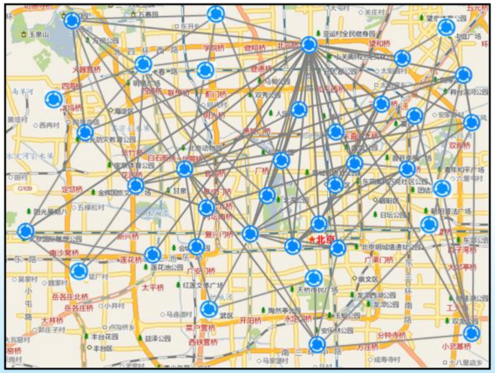
签到点亮中国



GeoLife: Building Social Networks Using Human Location History

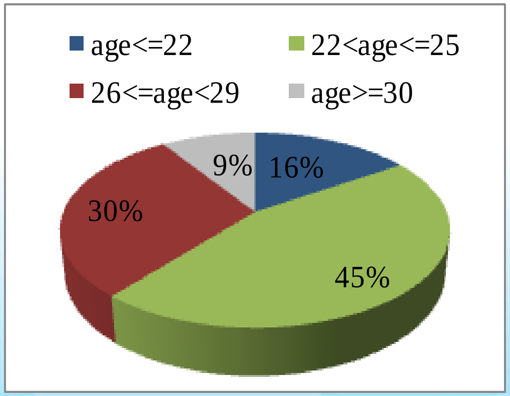


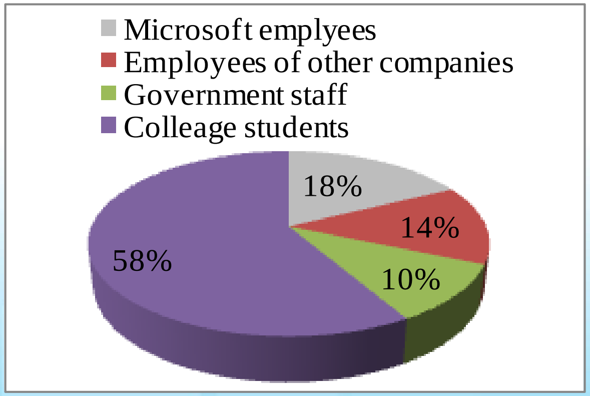




GPS Devices and Users

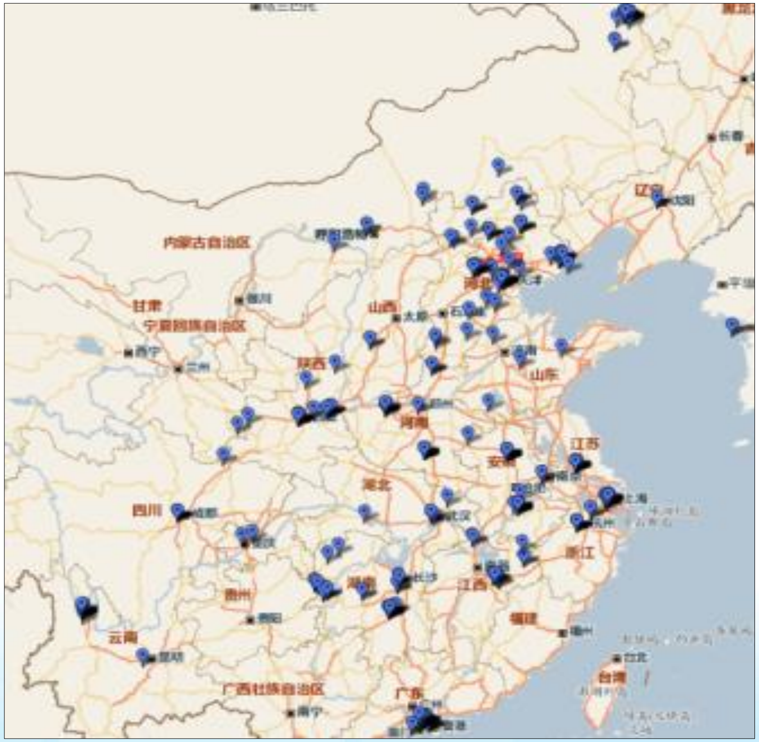
- 178 users, Apr. 2007 ~ Oct. 2011

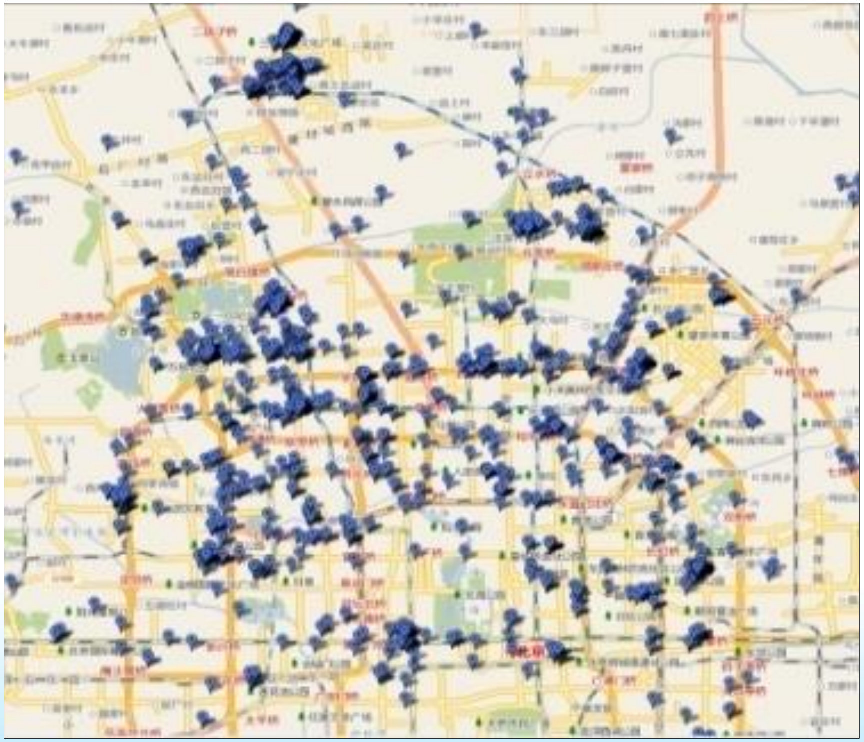




A Free Large-Scale GPS Dataset

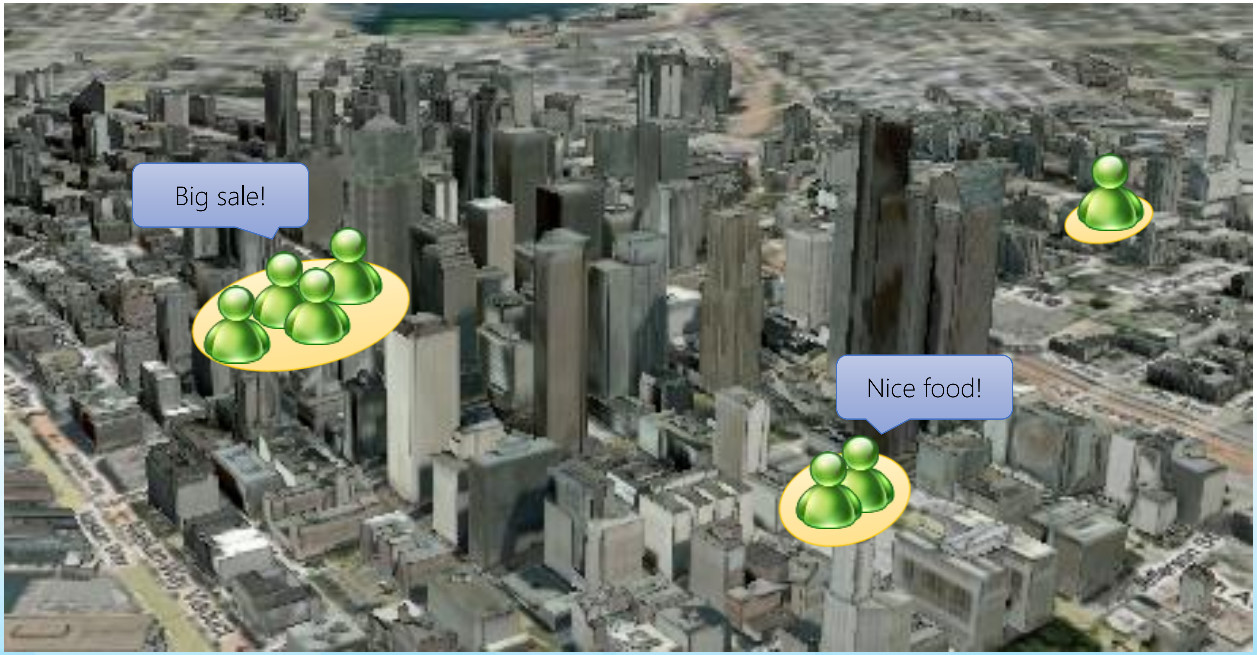
- 17621 trajectories, 1.2 million kilometers, 48000+ hours





Collaborative Activity and Location Recommendation

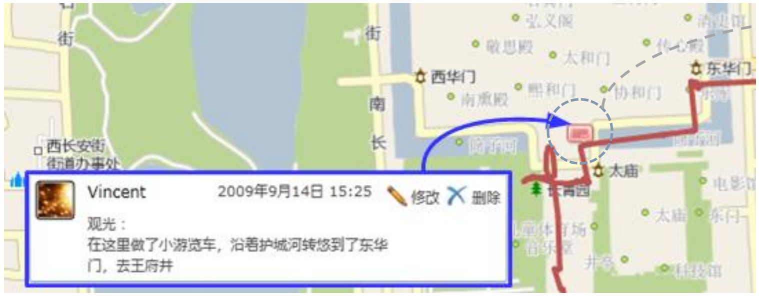
- Location Recommendation
  - Question: *I want to find nice food, where should I go?*
- Activity Recommendation
  - Question: *I will visit the downtown, what can I do there?*



AI Journal, AAAI 2010, WWW 2010

Data Modeling

- User <-> Location <-> Activity



Vincent 2009年9月14日 15:25

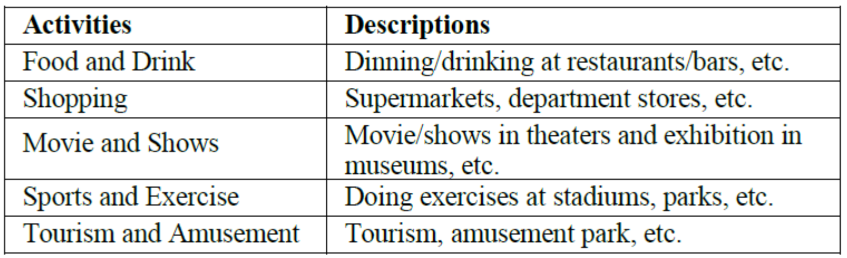
观光：
在这里做了小游览车，沿着护城河转悠到了东华门，去王府井

GPS: “39.903, 116.391, 14/9/2009 15:25”

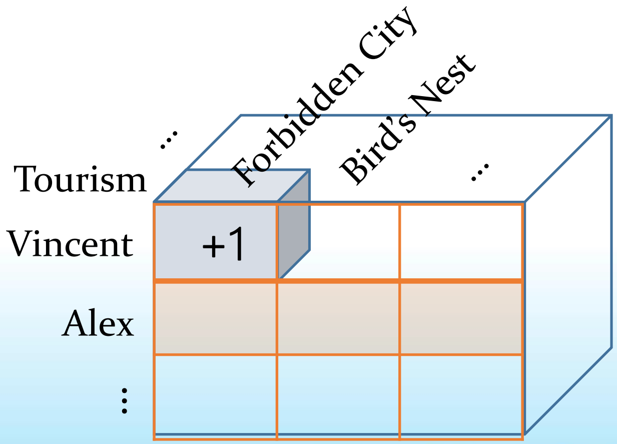
Stay Region: “39.910, 116.400 (Forbidden City)”

“User Vincent: We took a tour bus to see around along the forbidden city moat …”

Activity: tourism

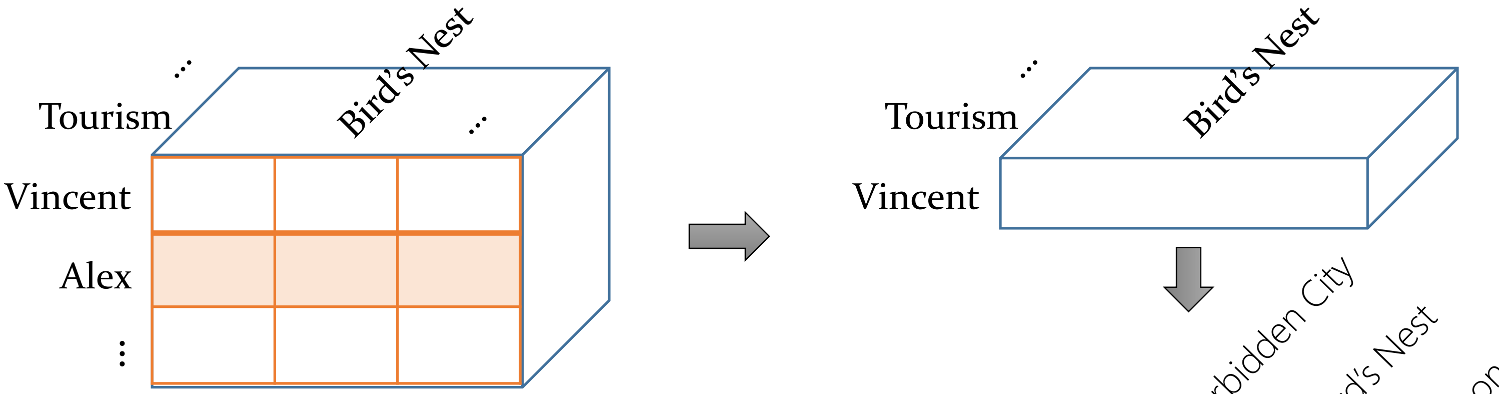


Activities	Descriptions
Food and Drink	Dinning/drinking at restaurants/bars, etc.
Shopping	Supermarkets, department stores, etc.
Movie and Shows	Movie/shows in theaters and exhibition in museums, etc.
Sports and Exercise	Doing exercises at stadiums, parks, etc.
Tourism and Amusement	Tourism, amusement park, etc.



How to Do Recommendation?

- If the tensor is full, then for each user:



	Forbidden City	Bird's Nest	Zhongguancun
Shopping	2	1	6
Exhibition	4	3	2
Tourism	5	4	1

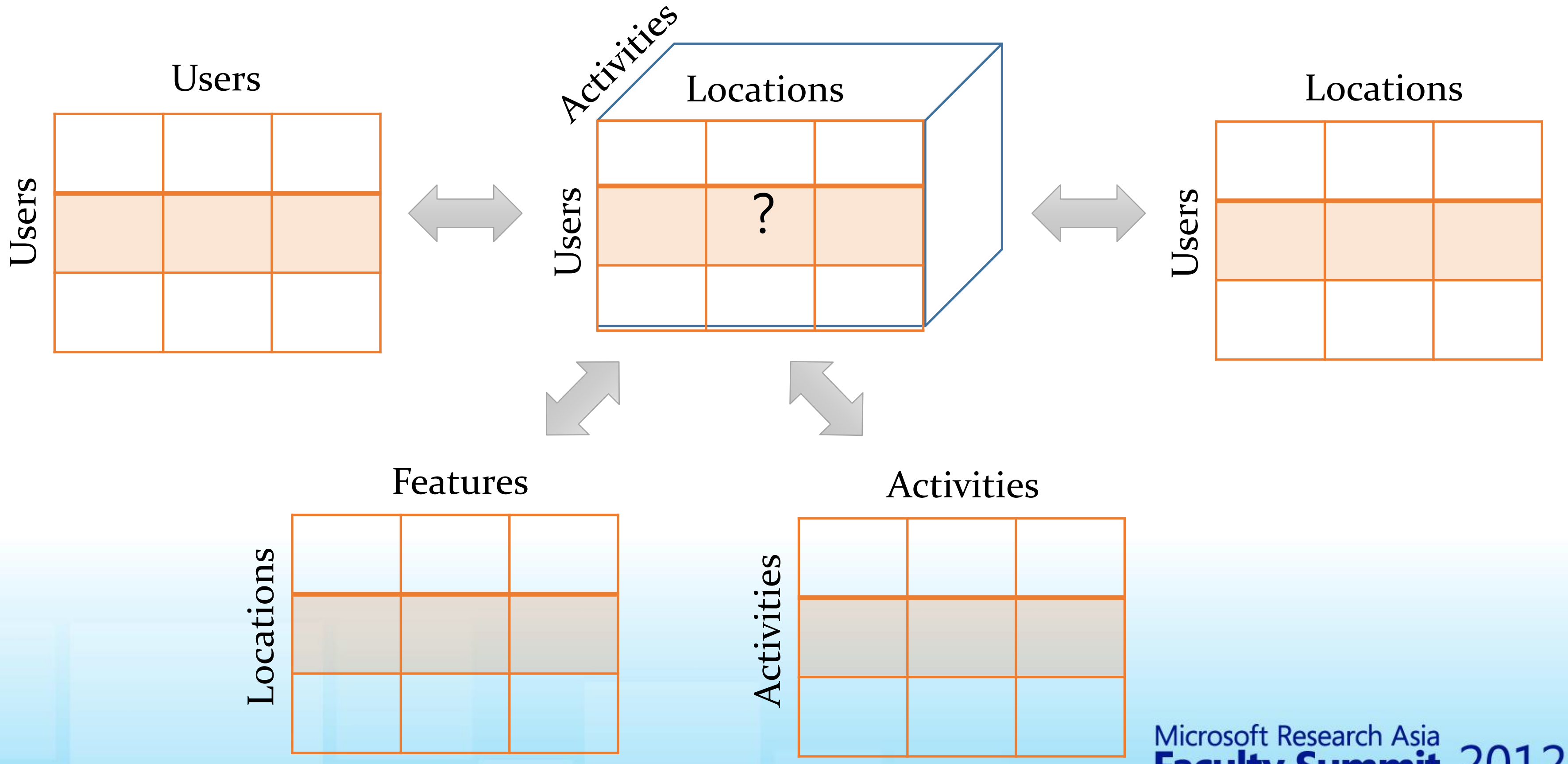
Location recommendation for Vincent
Tourism:
Forbidden City > Bird's Nest > Zhongguancun

Activity recommendation for Vincent
Forbidden City:
Tourism > Exhibition > Shopping

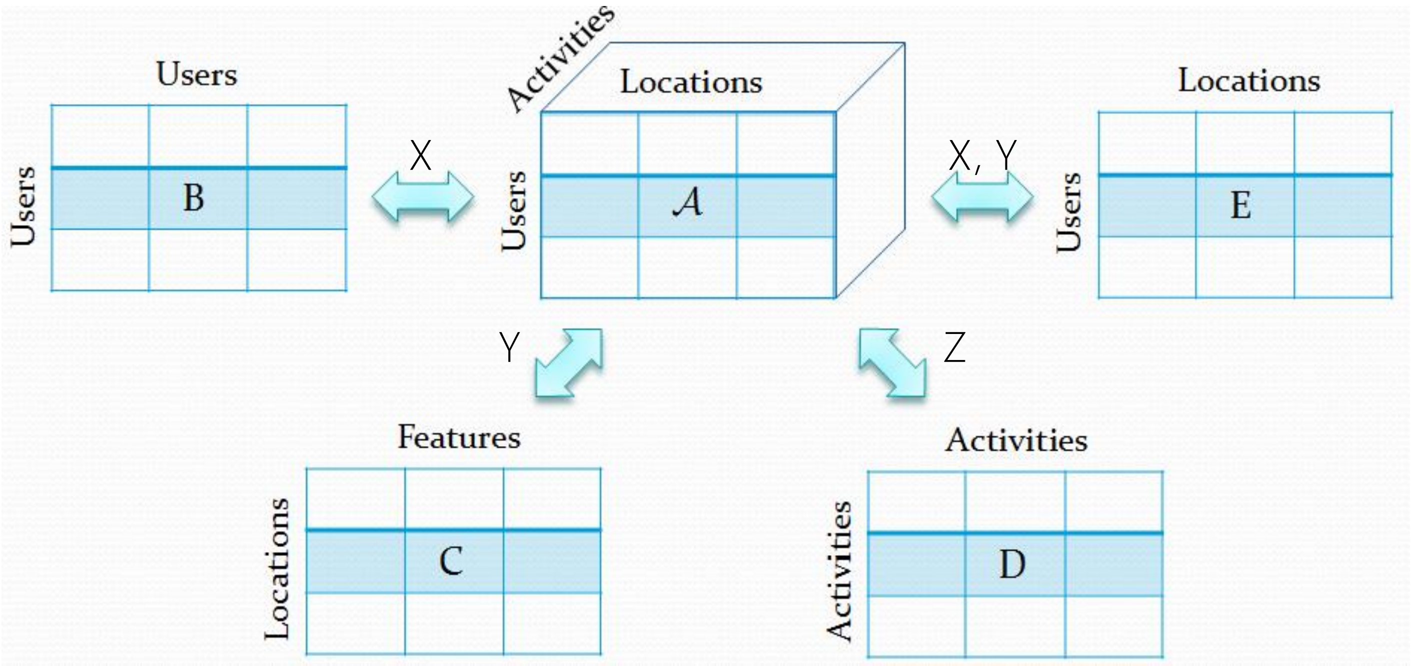
Unfortunately, in practice, the tensor is usually sparse!

Our Solution

- Regularized Tensor and Matrix Decomposition



Our Model

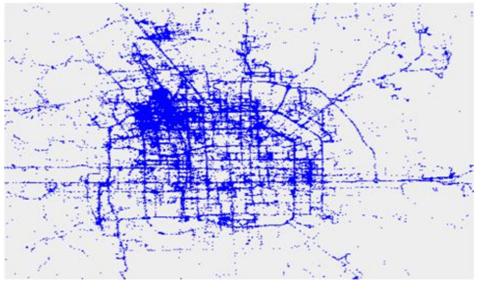


$$\mathcal{L}(X,Y,Z,U) = \frac{1}{2}\left\|\mathcal{A} - [\![X,Y,Z]\!]\right\|^2$$

$$+ \frac{\lambda_1}{2}\mathrm{tr}(X^T L_B X) + \frac{\lambda_2}{2}\left\|C - YU^T\right\|^2 + \frac{\lambda_3}{2}\mathrm{tr}(Z^T L_D Z) + \frac{\lambda_4}{2}\left\|E - XY^T\right\|^2$$

$$+ \frac{\lambda_5}{2}(\|X\|^2 + \|Y\|^2 + \|Z\|^2 + \|U\|^2)$$

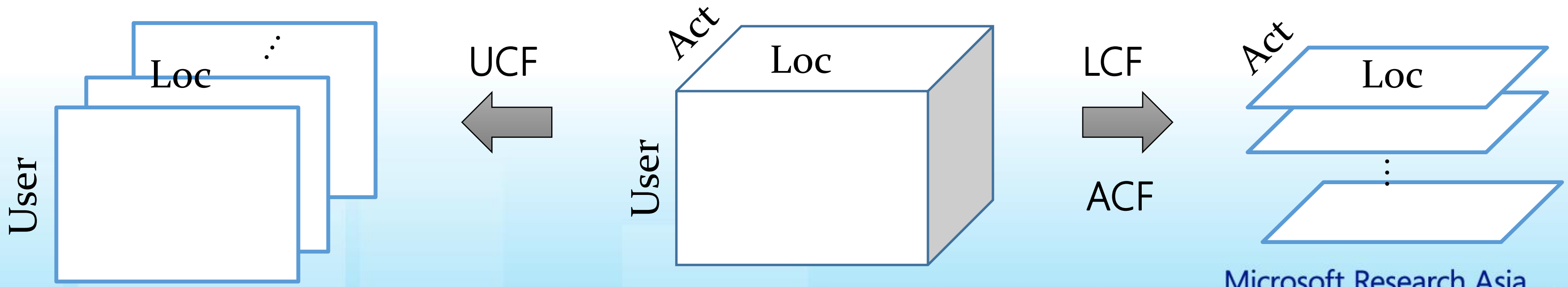
Experiments



- Data
  - GeoLife data set
  - 13K GPS trajectories, 140K km long
  - 530 comments
  - After clustering, #(loc) = 168; #(user) = 164, #(act) = 5, #(loc_fea) = 14
  - The user-loc-act tensor has 1.04% of the entries with values
- Evaluation
  - Ranking over the hold-out test dataset
  - Metrics:
    - Root Mean Square Error (RMSE)
    - Normalized discounted cumulative gain (*nDCG*)

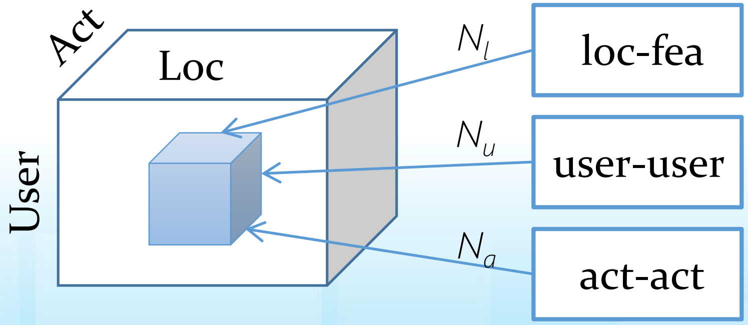
Baselines – Category I

- Tensor -> Independent matrices [Herlocker et al. 1999]
  - Baseline 1: UCF (user-based CF)
    - CF on each user-loc matrix + Top *N* similar users for weighted average
  - Baseline 2: LCF (location-based CF)
    - CF on each loc-act matrix + Top *N* similar locations for weighted average
  - Baseline 3: ACF (activity-based CF)
    - CF on each loc-act matrix + Top *N* similar activities for weighted average

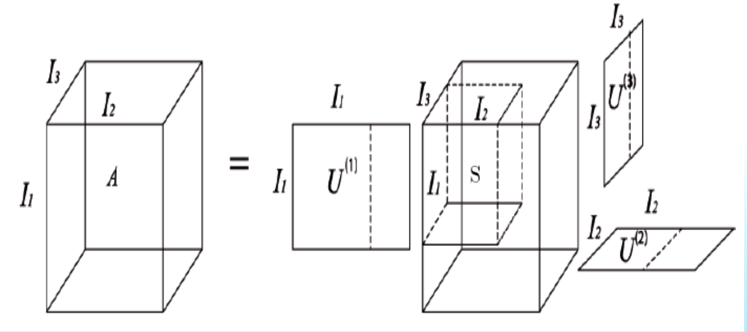


Baselines – Category II

- Tensor-based CF
  - Baseline 4: ULA (unifying user-loc-act CF) [Wang et al. 2006]
    - Top N_u similar users, top N_l similar loc's, top N_a similar act's
    - Similarities from additional matrices + Small cube for weight average
  - Baseline 5: HOSVD (high order SVD) [Symeonidis et al. 2008]
    - Singular value decomposition with matrix unfolding



ULA

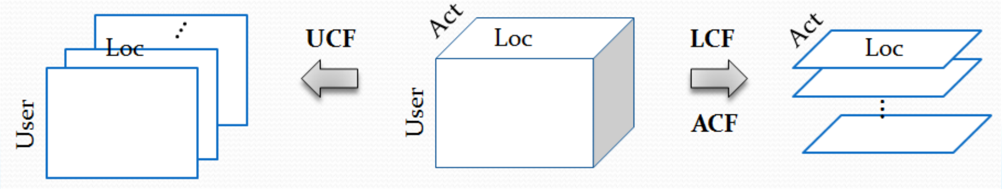


HOSVD

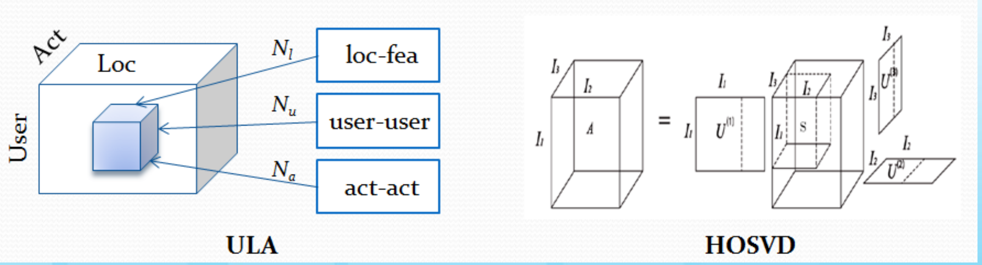
Comparison with Baselines

- Reported in “mean ± std”

	RMSE	nDCG$_{loc}$	nDCG$_{act}$
UCF	0.027 ± 0.006	0.297 ± 0.024	0.807 ± 0.007
LCF	0.009 ± 0.000	0.532 ± 0.021	0.614 ± 0.019
ACF	0.022 ± 0.005	0.408 ± 0.012	0.785 ± 0.006
ULA	0.015 ± 0.003	0.291 ± 0.022	0.799 ± 0.012
HOSVD	0.006 ± 0.001	0.390 ± 0.021	0.913 ± 0.004
UCLAF	$\mathbf{0.006 \pm 0.001}$	$\mathbf{0.599 \pm 0.036}$	$\mathbf{0.959 \pm 0.009}$



[Herlocker et al. 1999]



[Wang et al. 2006]

[Symeonidis et al. 2008]

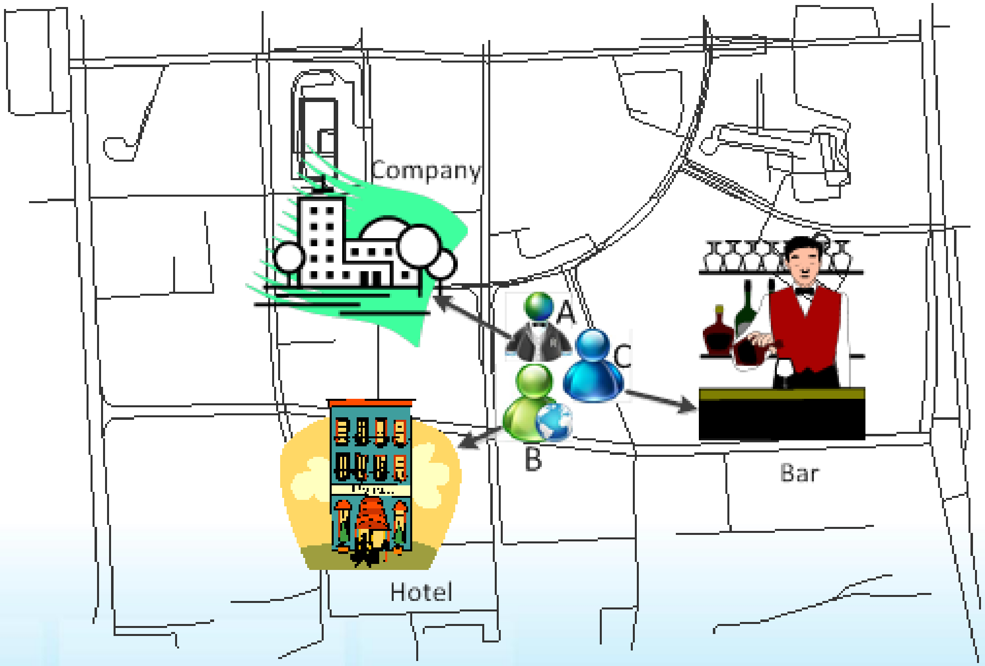
Collaborative Activity and Location Recommendation

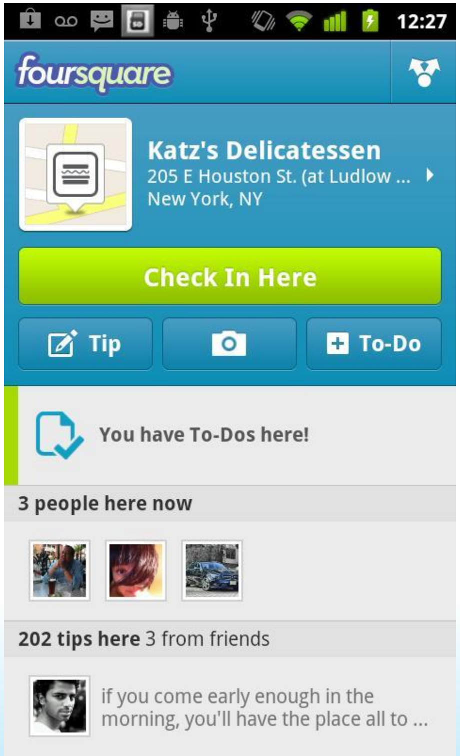
- We showed how to mine knowledge from GPS data to answer
  - If I want to do something, where should I go?
  - If I will visit some place, what can I do there?
- We evaluated our system on a large GPS dataset
  - 19% improvement on location recommendation
  - 22% improvement on activity recommendation

  over the simple memory-based CF baseline (i.e. UCF, LCF, ACF)

User Location Naming

- Mapping from GPS to location name





Katz's Delicatessen
205 E Houston St. (at Ludlow ...
New York, NY

You have To-Dos here!

3 people here now

202 tips here 3 from friends

if you come early enough in the morning, you'll have the place all to ...

Problem Definition

- Given
  - POI database $\boldsymbol{P}$
  - Check-in history $\boldsymbol{C_{ts}^{te}}$, *where* $\boldsymbol{ts, te}$ *is the start and end time*
  - User $\boldsymbol{u}$
  - Time $\boldsymbol{t}$
  - GPS reading $\boldsymbol{g}$

- Rank a subset $\boldsymbol{P'}$ from a POI database $\boldsymbol{P}$
  - $R_{g,u,t,C_{ts}^{te}} = \pi_{g,u,t,C_{ts}^{te}}(P'), P' \subseteq P$

Positioning Error & Dense POI

- GPS errors
- High density, hierarchical and large scale properties of POIs

Size(m^2)	200x200	100x100	50x50
avg #poi	10.6	6.0	3.7
stdvar #poi	21.8	11.2	6.9
max #poi	490	286	237

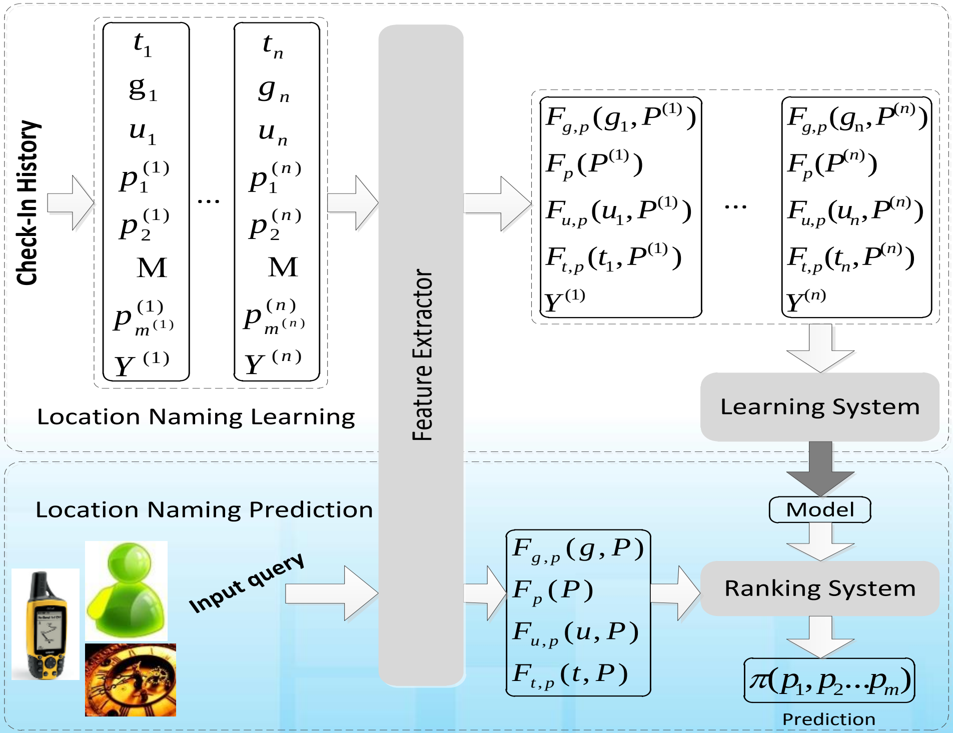
Data Sparsity

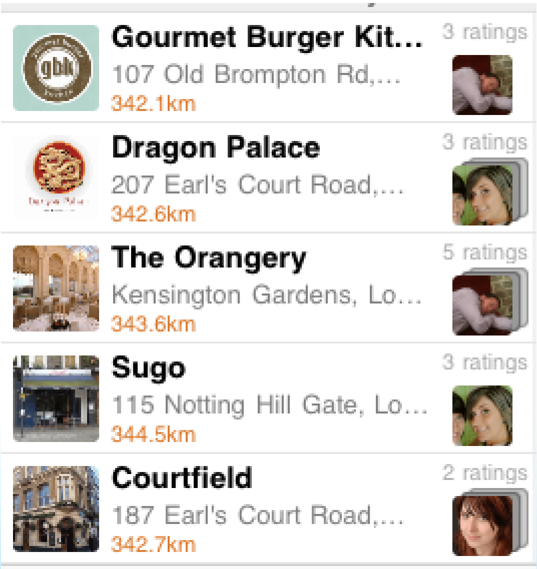
- Dianping
  - Reviews of local businesses
  - Check-in functionality

Dataset—Dianping—Beijing	**2011.1.7—2011.6.11**
#POIs	15664
#Users	545
#Check-in	31811
#Days	152
average #Check-in per POI	2.6
average #Users per POI	1.4
average #Check-in per User	58
average #POIs per User	32

An Analogy to Local Search

- One-to-One mapping is difficult
- Try to provide a better rank of POIs





Gourmet Burger Kit… 3 ratings
107 Old Brompton Rd,…
342.1km

Dragon Palace 3 ratings
207 Earl's Court Road,…
342.6km

The Orangery 5 ratings
Kensington Gardens, Lo…
343.6km

Sugo 3 ratings
115 Notting Hill Gate, Lo…
344.5km

Courtfield 2 ratings
187 Earl's Court Road,…
342.7km

Static Features

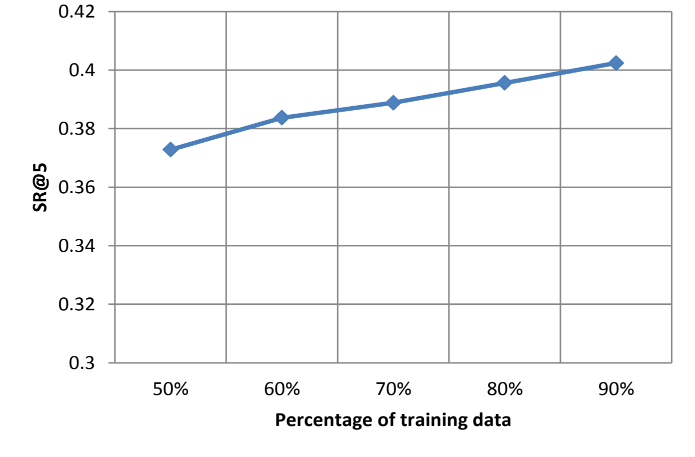
- Number of reviews related to it
- Average score given by social network users
- Number of web pages referring to it
- Number of check-ins
- Number of people checked-in
- Number of photos users have uploaded

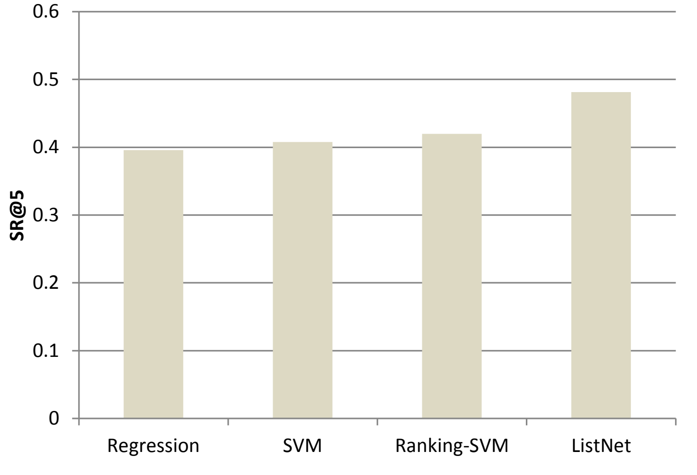
Dynamic Features

- Features for an individual user
  - Distance between the GPS reading and the POI location
  - Preference of user u on POI p
    - Measured by the number of check-ins by user u at POI p
- Features for a group of users
  - Temporal pattern between time t and POI p
    - Measured by the number of check-ins at time t and at POI p

Experiments

- Evaluation metric
  - Success Rate (SR) at $\boldsymbol{k}$
  - $$SR@k = \frac{|\{query|query\ is\ tested\ as\ accurate\ at\ k\}|}{|\{query\}|}$$
- Ranking algorithm selection
- The impact of training data size





Fake Check-In Problem

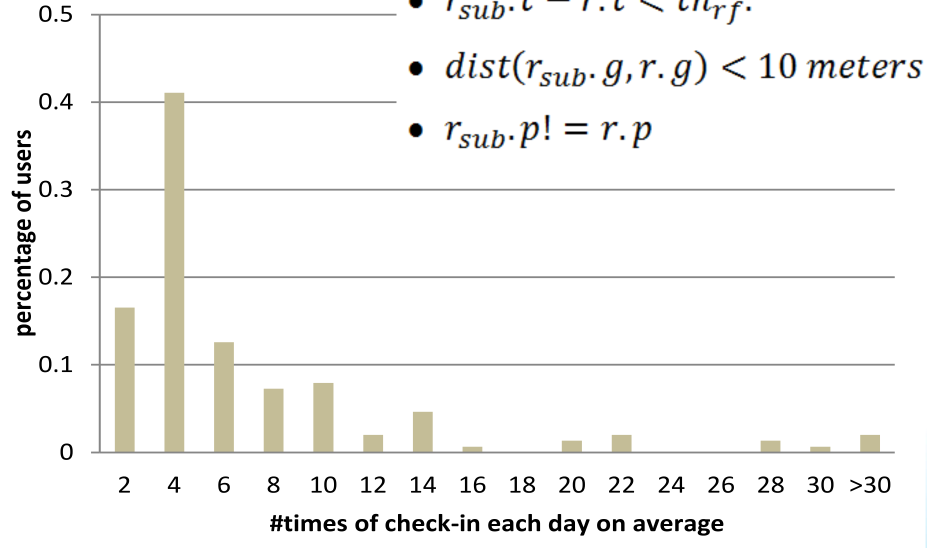
- Benefit driven
  - Getting the coupon
  - Getting the discount
  - Getting the badge
- Killing time, e.g, at the airport
- Interest driven

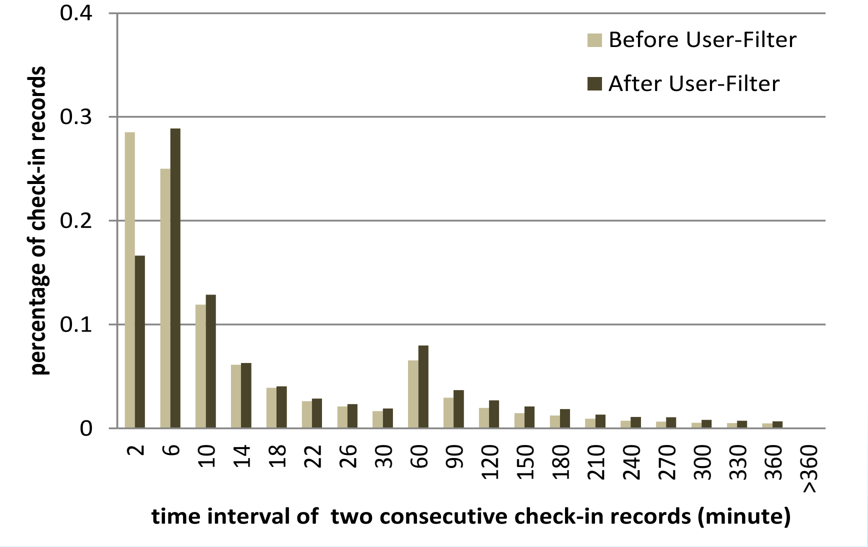
→

- Frequent check-ins
- Super human speed
- Rapid-fire check-in

Fake Check-In Problem

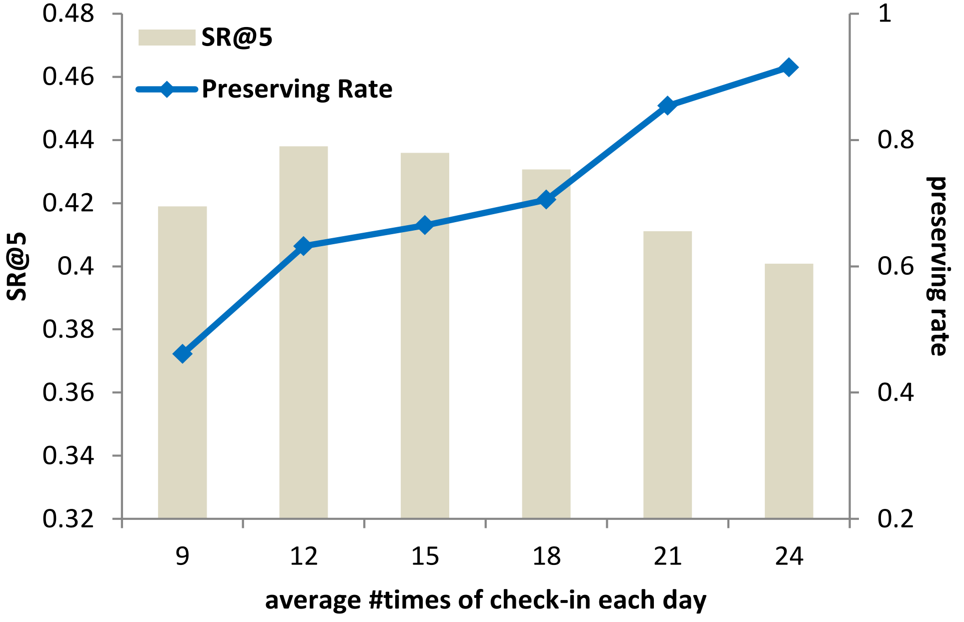
- Fake users - If a user check-in a lot of locations each day
- Fake check-in record $\boldsymbol{r}$ - if the following condition meets
  - r has a subsequence record r_{sub}
  - $r_{sub}.t - r.t < th_{rf}$.
  - $dist(r_{sub}.g, r.g) < 10\ meters$
  - $r_{sub}.p! = r.p$

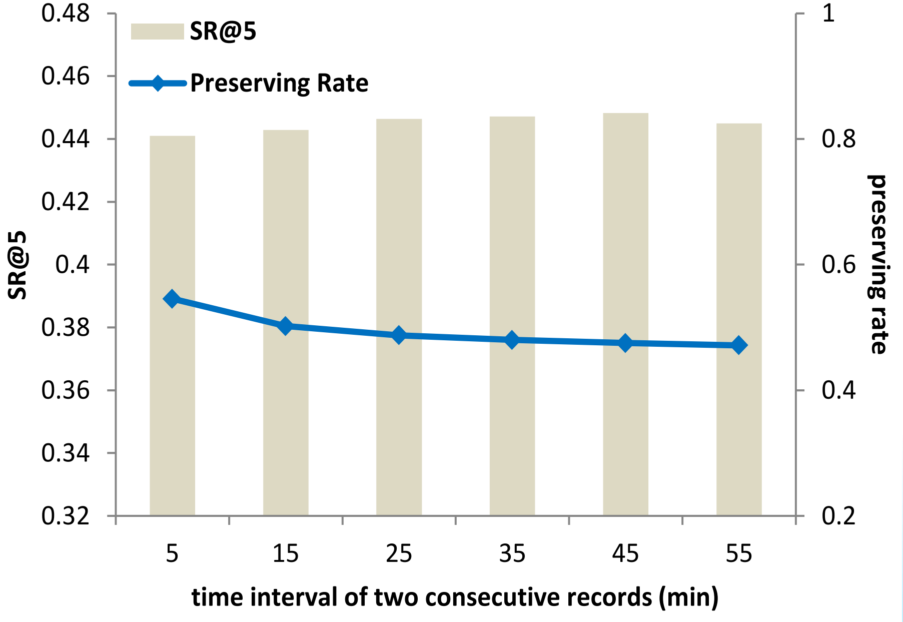




Fake Check-in Filtering

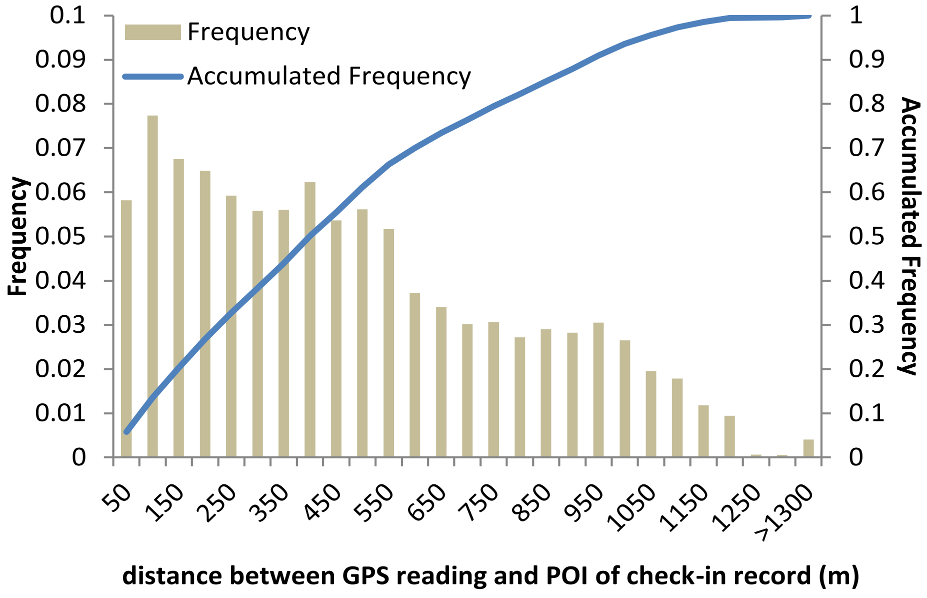
- Large impact of filtering fake users
  - Fake users are so random that it is difficult to predict their check-in
- Little impact of filtering fake check-in records

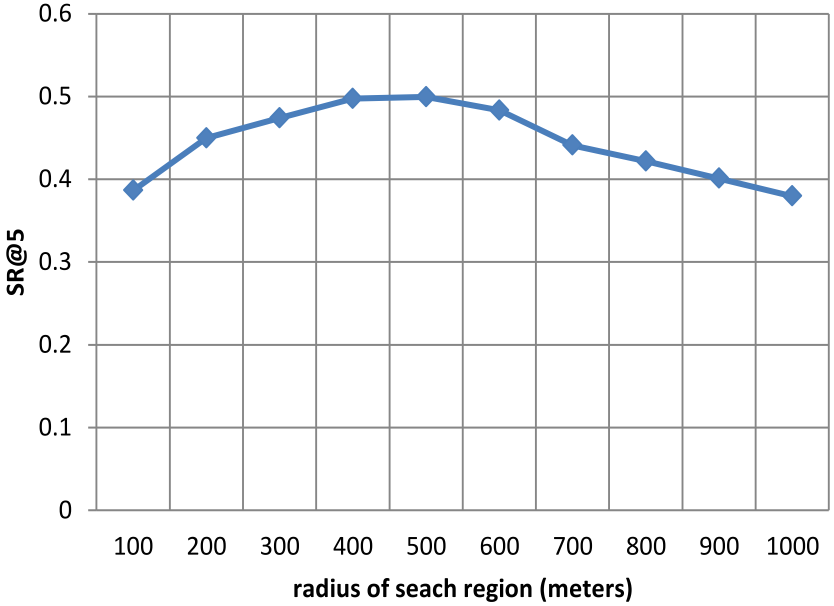




Search Radius

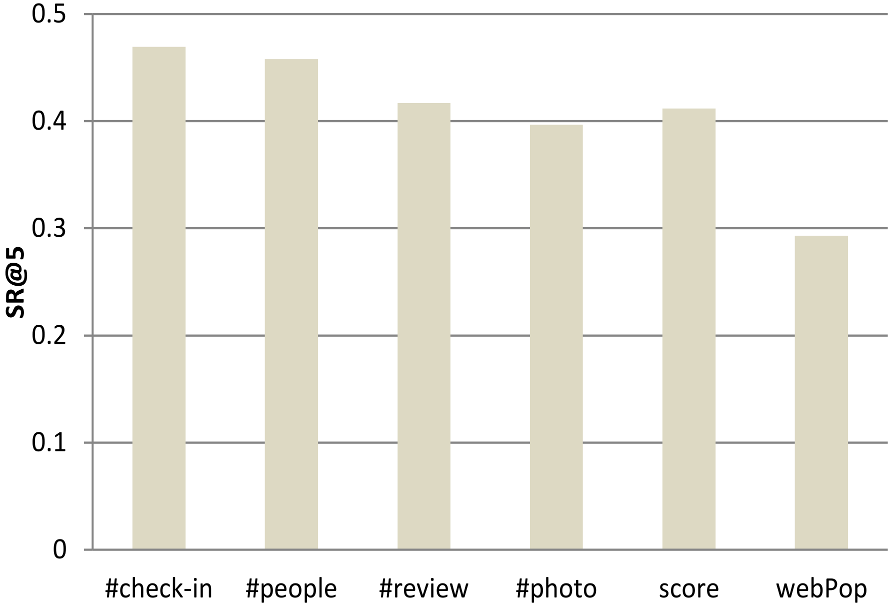
- Most check-ins are at nearby locations
- Distant check-ins are considered as noises
- Significant impact of different search radius

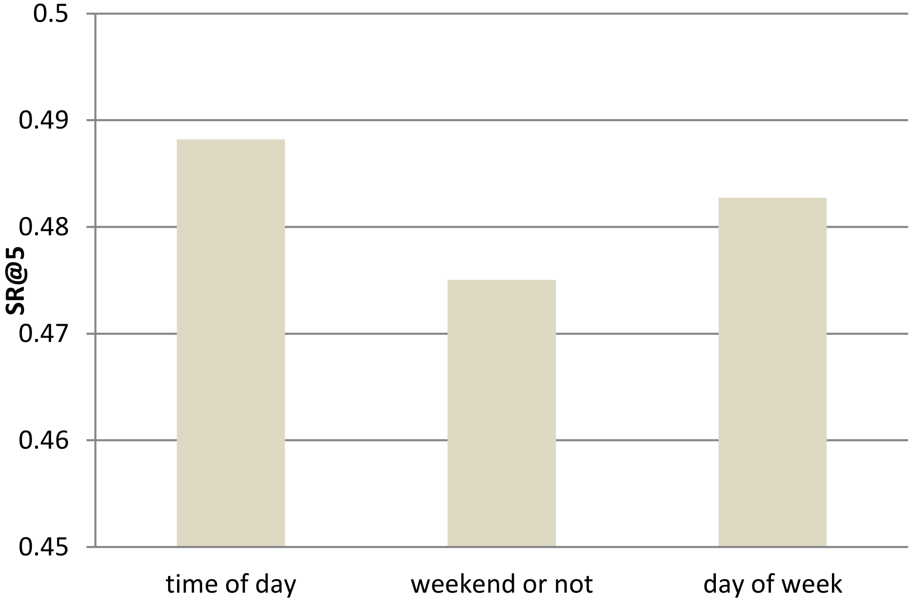




Feature Effectiveness

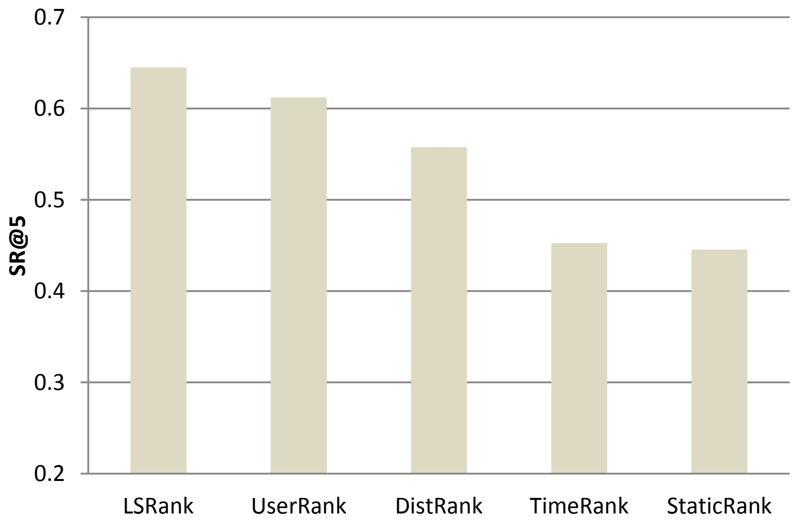
- #check-in is significantly better than webPop
- No big difference of different temporal patterns





Overall Results

- Our proposed LSRank performs the best, but not significantly better than UserRank.
- Distance and interaction between user and POI is important
- Static features can not be ignored.



User Location Naming

- A novel location naming approach which provides concrete and meaningful names to users based on time, GPS reading and check-in histories.
- Most important features
  - User history
  - Distance
  - #review
  - Web popularity
- 64.5% of test queries can return intended POIs within top 5 results

Human Mobility

- Mobility based on Levy Flight and variants (Brockmann et al Nature'06, Gonzalez et al Nature'08, Song et al Nature Physics'10, Rhee et al Infocom'08)
  - Data from Bank notes, CDR, GPS
  - Jump step size analysis
  - Collective and individual behavior
  - Gyration distribution
- Mobility extracted from real traces (Isaacman et al MobiSys'12, Kim et al, Infocom'06, Cho et al KDD'11, Sadilek et al WSDM'11, Krumm et al Ubicomp'06, Yoon et al MobiSys'06, Jing et al KDD'12)
  - Data from GPS, CDR and WLAN, Check-in and Geo-twitters
  - Collective and individual significant places (home/workplace) detection
  - Markov process between hot spots modeling
  - Duration estimation at a location
  - Socially controlling mobility (Geo-twitters and check-ins)
    - Move near friends' home
    - Move similar to friends

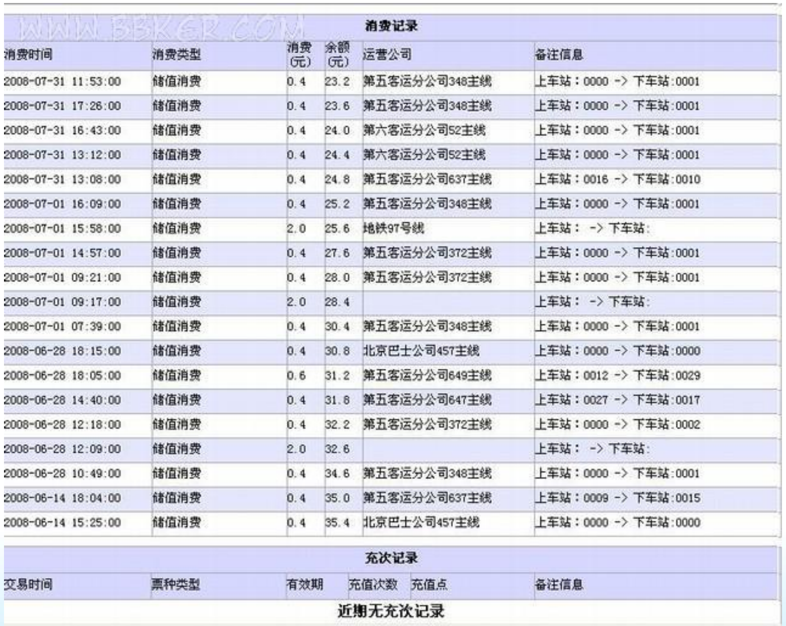
Mobility Prediction

- Predictability (Song et al Science'10 , Jensen et al MLSP, Lin et al Ubicomp'12)
  - Low resolution GSM/WLAN/blue tooth/acceleration with entropy measurement
  - High resolution GPS data with redundancy measurement
- Prediction
  - Spatial (Song et al TMC'06, Eagle et al Pers Ubiquit Comput'06, Scellato et al. Pervasive'11)
  - Temporal (Chon et al PerCom'12, Scellato et al. Pervasive'11)
  - Activity recognition (Eagle et al Behav Ecol Sociobiol' 09)

A Real Story

- Sequential pattern
  - **石佛**营西里-350，406-**朝阳公园**桥-657-**望京**
  - **石佛**营西里-729-**木樨园**-627-**望京**

- Home location:**石佛**营西里
- Work location:**望京**
- Important location:**木樨园**
- Job category: **服装批**发(**旺角市**场)



消费记录

消费时间	消费类型	消费(元)	余额(元)	运营公司	备注信息
2008-07-31 11:53:00	储值消费	0.4	23.2	第五客运分公司348主线	上车站：0000 -> 下车站:0001
2008-07-31 17:26:00	储值消费	0.4	23.6	第五客运分公司348主线	上车站：0000 -> 下车站:0001
2008-07-31 16:43:00	储值消费	0.4	24.0	第六客运分公司52主线	上车站：0000 -> 下车站:0001
2008-07-31 13:12:00	储值消费	0.4	24.4	第六客运分公司52主线	上车站：0000 -> 下车站:0001
2008-07-31 13:08:00	储值消费	0.4	24.8	第五客运分公司637主线	上车站：0016 -> 下车站:0010
2008-07-01 16:09:00	储值消费	0.4	25.2	第五客运分公司348主线	上车站：0000 -> 下车站:0001
2008-07-01 15:58:00	储值消费	2.0	25.6	地铁97号线	上车站： -> 下车站:
2008-07-01 14:57:00	储值消费	0.4	27.6	第五客运分公司372主线	上车站：0000 -> 下车站:0001
2008-07-01 09:21:00	储值消费	0.4	28.0	第五客运分公司372主线	上车站：0000 -> 下车站:0001
2008-07-01 09:17:00	储值消费	2.0	28.4		上车站： -> 下车站:
2008-07-01 07:39:00	储值消费	0.4	30.4	第五客运分公司348主线	上车站：0000 -> 下车站:0001
2008-06-28 18:15:00	储值消费	0.4	30.8	北京巴士公司457主线	上车站：0000 -> 下车站:0000
2008-06-28 18:05:00	储值消费	0.6	31.2	第五客运分公司649主线	上车站：0012 -> 下车站:0029
2008-06-28 14:40:00	储值消费	0.4	31.8	第五客运分公司647主线	上车站：0027 -> 下车站:0017
2008-06-28 12:18:00	储值消费	0.4	32.2	第五客运分公司372主线	上车站：0000 -> 下车站:0002
2008-06-28 12:09:00	储值消费	2.0	32.6		上车站： -> 下车站:
2008-06-28 10:49:00	储值消费	0.4	34.6	第五客运分公司348主线	上车站：0000 -> 下车站:0001
2008-06-14 18:04:00	储值消费	0.4	35.0	第五客运分公司637主线	上车站：0009 -> 下车站:0015
2008-06-14 15:25:00	储值消费	0.4	35.4	北京巴士公司457主线	上车站：0000 -> 下车站:0000

充次记录

交易时间	票种类型	有效期	充值次数	充值点	备注信息
近期无充次记录					

Summary

- Understanding location and people through mobile social networks
- GeoLife: Building Social Networks Using Human Location History
- Learning Location Naming from User Check-In Histories

Microsoft Research

Thank you!

Microsoft Research Asia Faculty Summit 2012