移动社交网络与用户位置预测

谢 幸
微软亚洲研究院
第29届中国数据库学术会议，合肥

用户位置预测

- 保安的三个哲学终极问题
  - 你是谁？
  - 你从哪来？
  - 你到哪去？
- 换个问法
  - 你是什么样的人？
    - 推断你偏好哪些类型的地方
  - 你去过哪些地方？
  - 你又将会到哪些地方去？

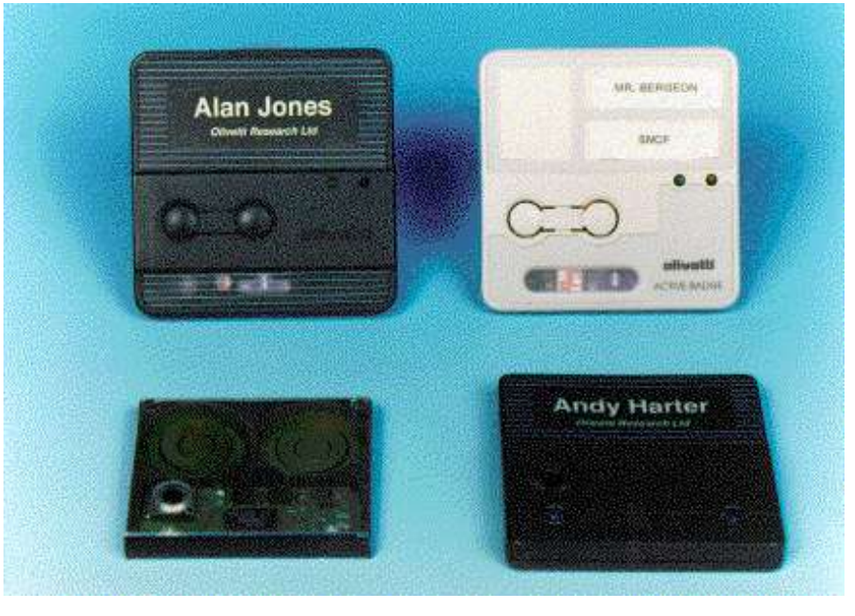
定位技术的对比

定位方法	定位精度	响应时间	使用范围	费用	终端和网络要求
传统GPS	30m	30s-3m	无遮挡	免费	GPS接收机
Cell-ID	>100m	3s	通信网络覆盖区域	收费	手机+通信网络
TOA、TDOA	100m-500m	<10s	通信网络覆盖区域	收费	手机+通信网络
AOA	150m	<10s	通信网络覆盖区域	收费	手机+通信网络
E-OTD	100m-500m	<10s	通信网络覆盖区域	收费	手机+通信网络
混合定位 gpsOne\XPS	3-30m	1s-10s	无限制	收费	CDMA/3G+ GPS + WiFi
RFID	1m-3m	1s	特定区域	免费	RFID装置、射频标签读写器
蓝牙	0.1m	\	10m有效范围 最大增益到100m	免费	蓝牙设备

李清泉, 乐阳. 基于位置服务的分析与展望. 中国计算机学会通讯, 6(6), 2010

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

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

用户位置数据

- 位置数据可能存在于各种类型的数据中
  - 带地理标注的照片、微博、游记
  - 位置搜索日志
  - 地图服务日志

- 缺乏一个很好的机制能够集中管理这些来自不同设备，不同服务和不同用户的位置数据。

移动社交网络

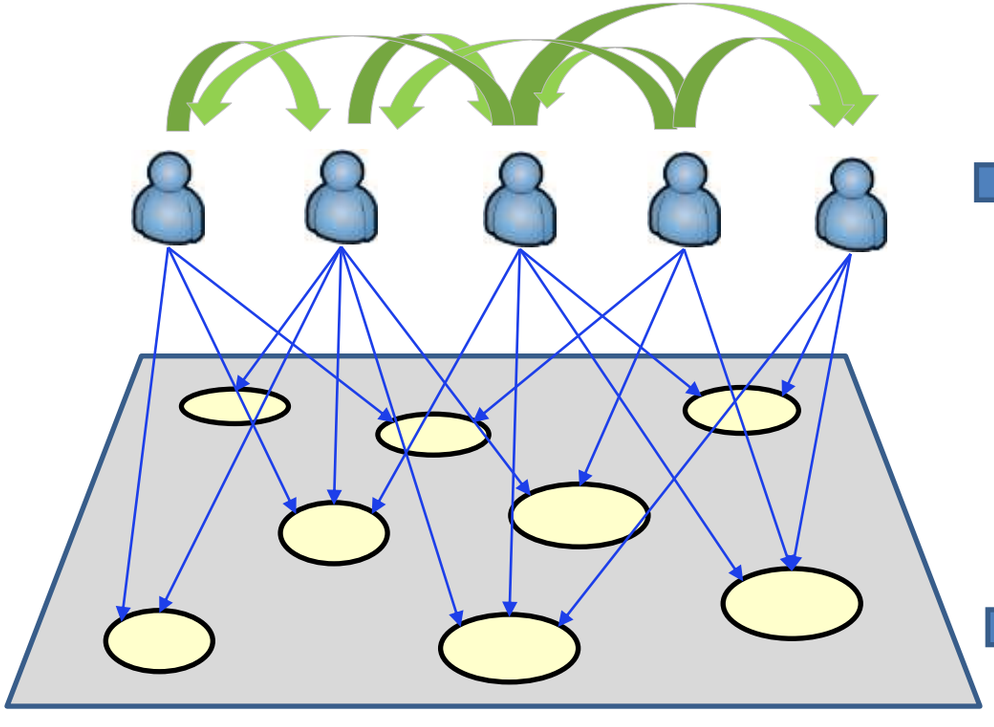
- 在社交网络中，用户主动和他们的朋友们分享心情、爱好、活动和照片等各种信息。这其中的很多信息都显式或隐式的包含了用户的位置。

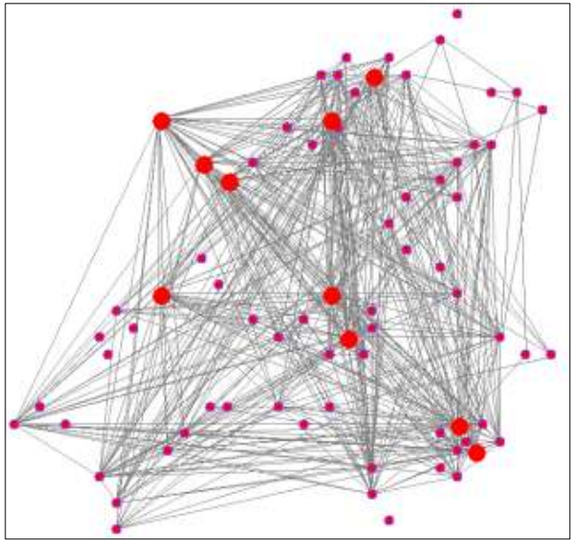
- 基于位置的社交网络
  - Location Based Social Networks，或称为签到服务
  - 共享各自的位置以及与位置相关的信息
  - 产生了非常庞大的用户位置数据集

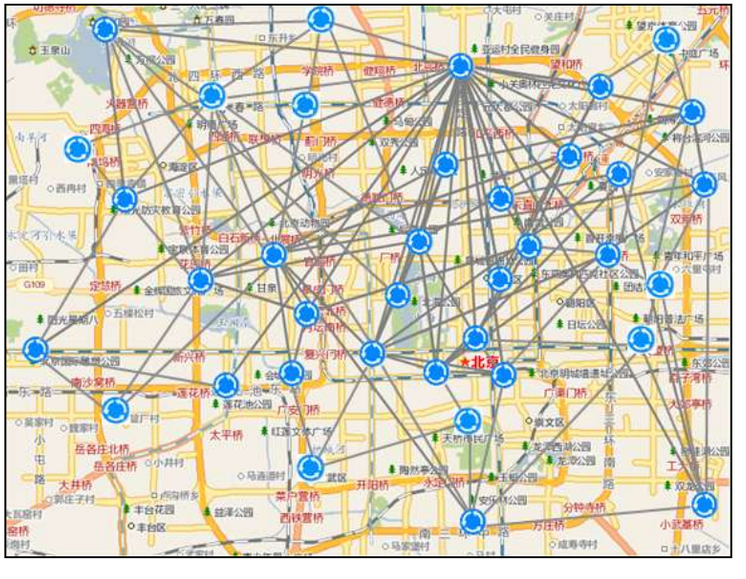
CCCF "移动社交网络"专辑(2012年5月)

- 瞬时社交网络
- 移动社交网络中的感知计算模型、平台与实践
- 移动社交网络中的用户行为预测模型
- 移动社交网络与用户位置
- 移动轨迹数据分析与智慧城市
- 社交媒体中的时空轨迹模式挖掘

GeoLife: Building Social Networks Using Human Location History

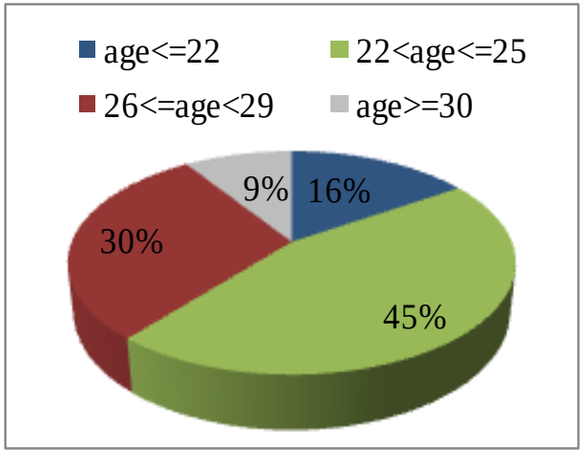


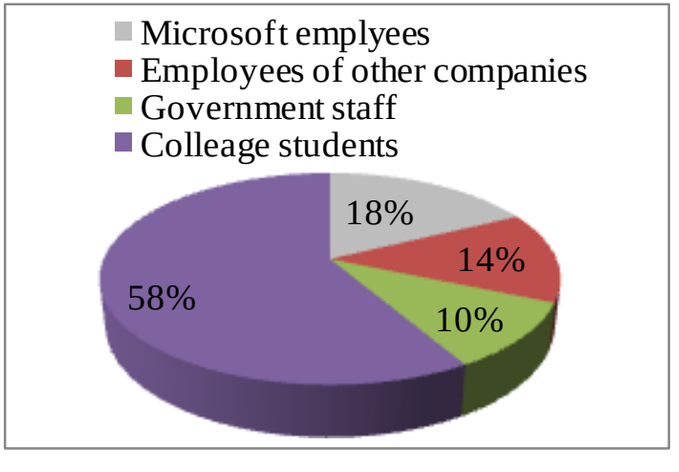




GPS Devices and Users

- 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

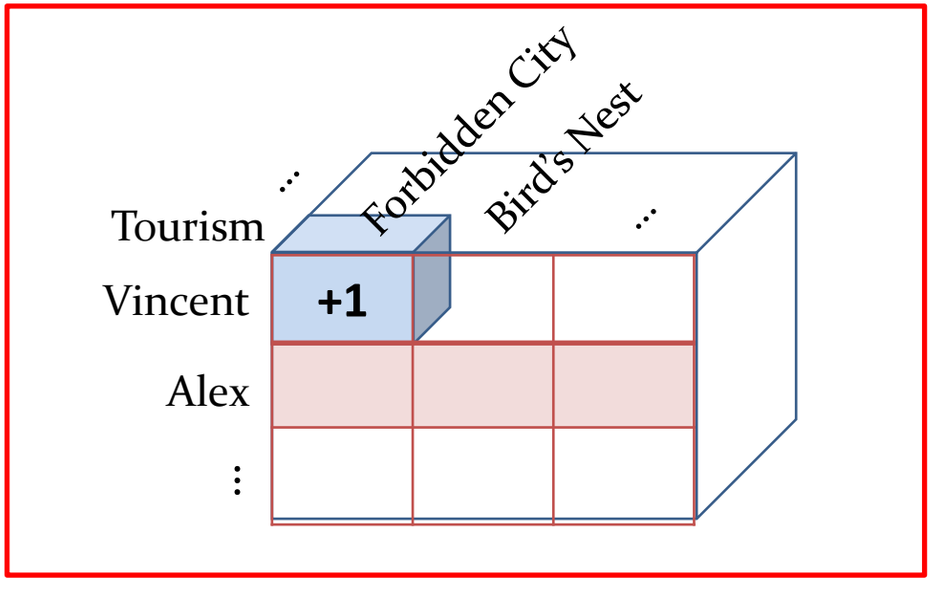
- User <-> Location <-> Activity

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 ..."

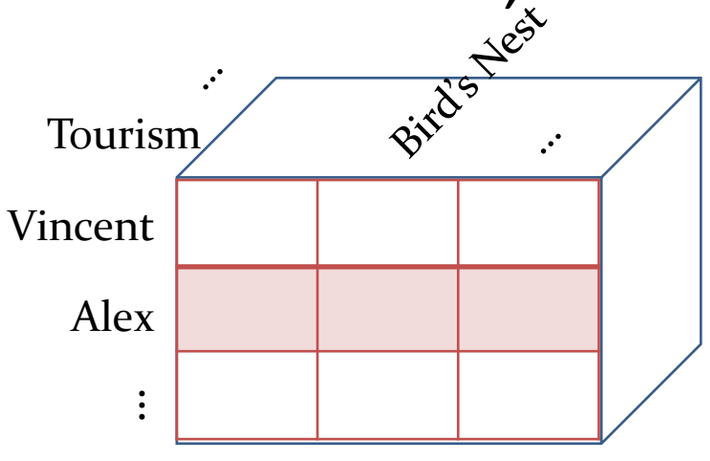
Activity: tourism

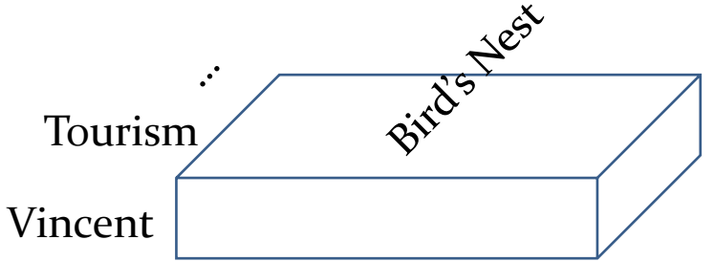


Activities	Descriptions
Food and Drink	Dinning/drinking at restaurants/bars, etc.
Shopping	Supermarkets, department stores, etc.
Movie and Shows	Movie/shows in theaters and exhibition in museums, etc.
Sports and Exercise	Doing exercises at stadiums, parks, etc.
Tourism and Amusement	Tourism, amusement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

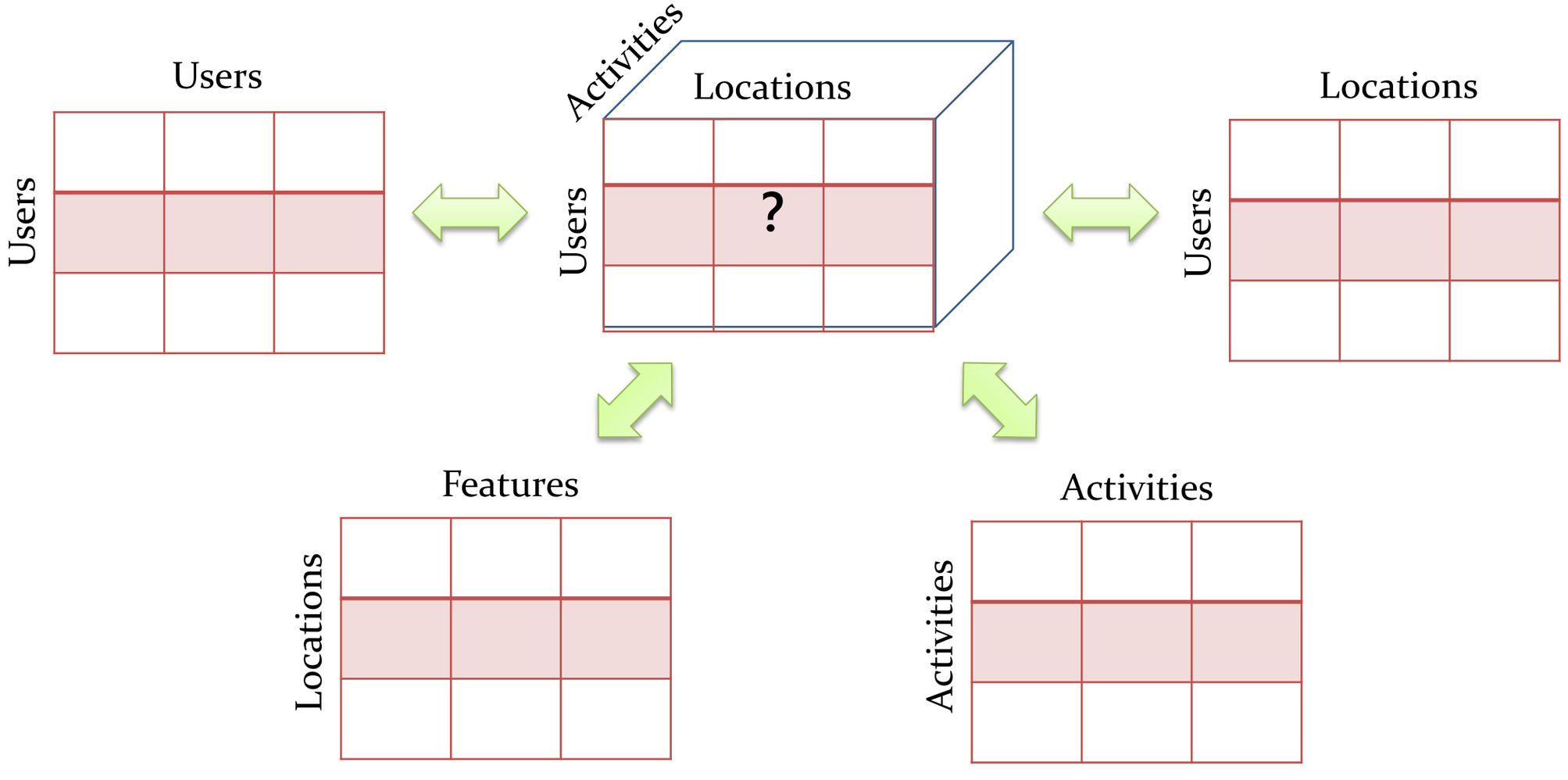
Location recommendation for Vincent
Tourism:
Forbidden City > Bird's Nest > Zhongguancun

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

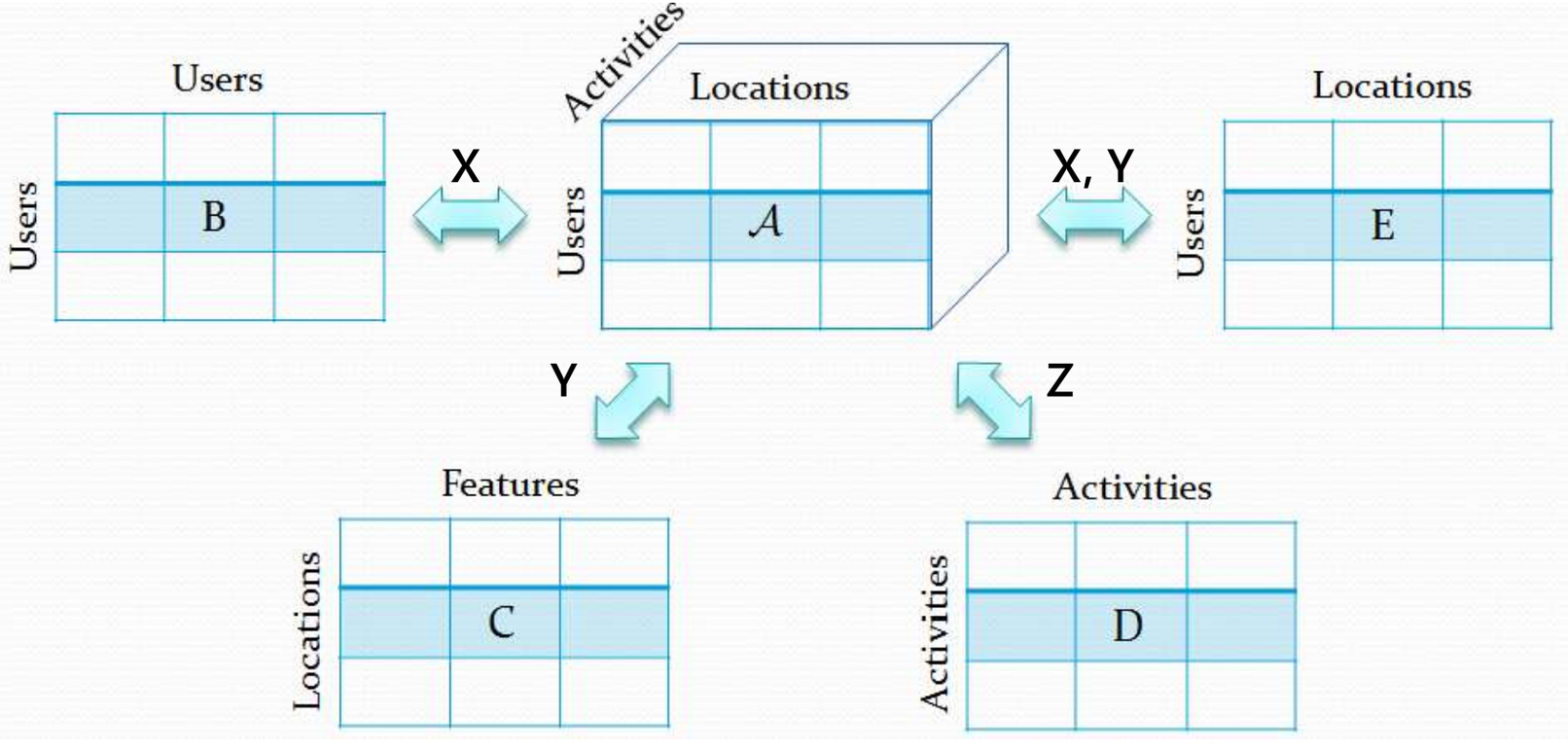
Unfortunately, in practice, the tensor is usually sparse!

Our Solution

- Regularized Tensor and Matrix Decomposition

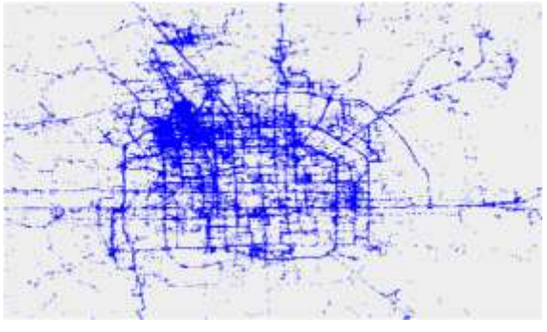


Our Model



$$\mathcal{L}(X,Y,Z,U) = \tfrac{1}{2}\left\|\mathcal{A} - [\![X,Y,Z]\!]\right\|^2 + \tfrac{\lambda_1}{2}\mathrm{tr}(X^T L_B X) + \tfrac{\lambda_2}{2}\left\|C - YU^T\right\|^2 + \tfrac{\lambda_3}{2}\mathrm{tr}(Z^T L_D Z) + \tfrac{\lambda_4}{2}\left\|E - XY^T\right\|^2 + \tfrac{\lambda_5}{2}(\|X\|^2 + \|Y\|^2 + \|Z\|^2 + \|U\|^2)$$

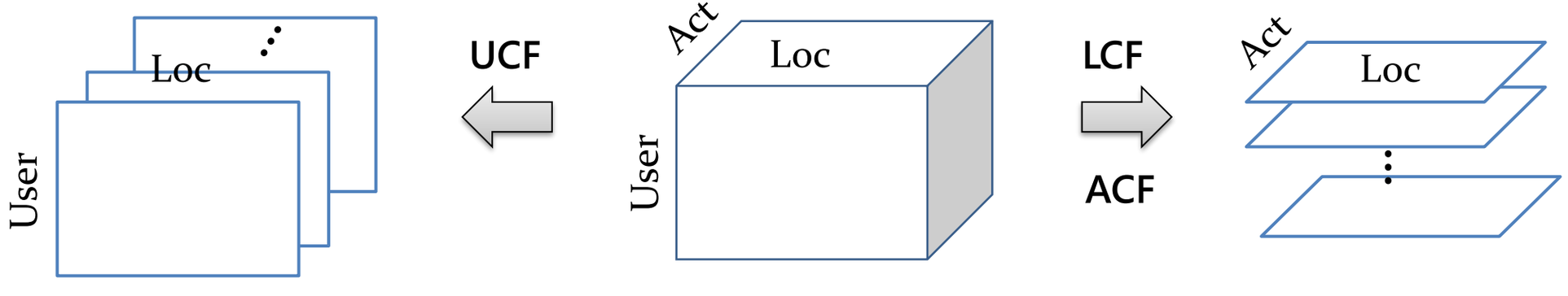
Experiments



- 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*)

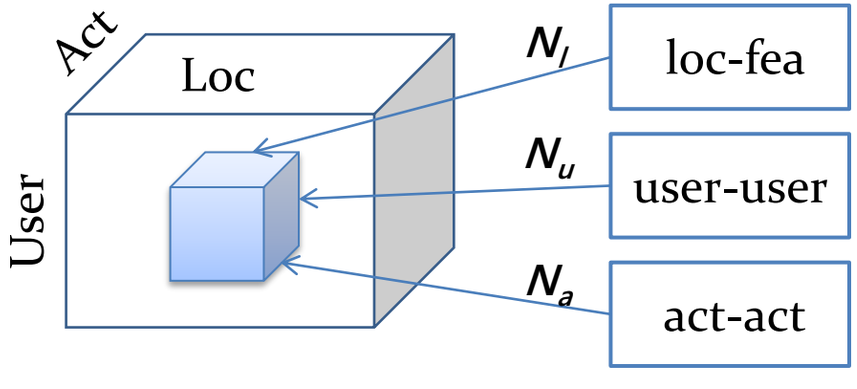
Baselines – Category I

- 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

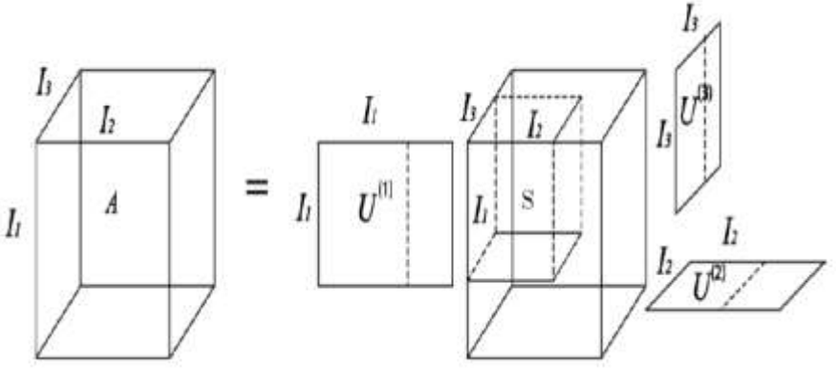


Baselines – Category II

- 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

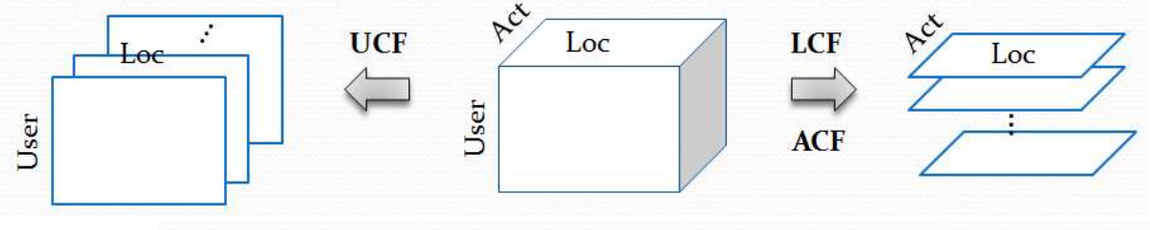


HOSVD

Comparison with Baselines

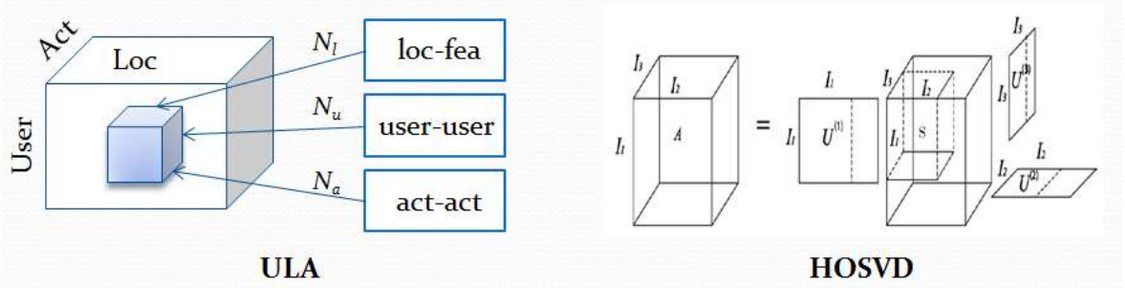
- Reported in “mean ± std”

	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]



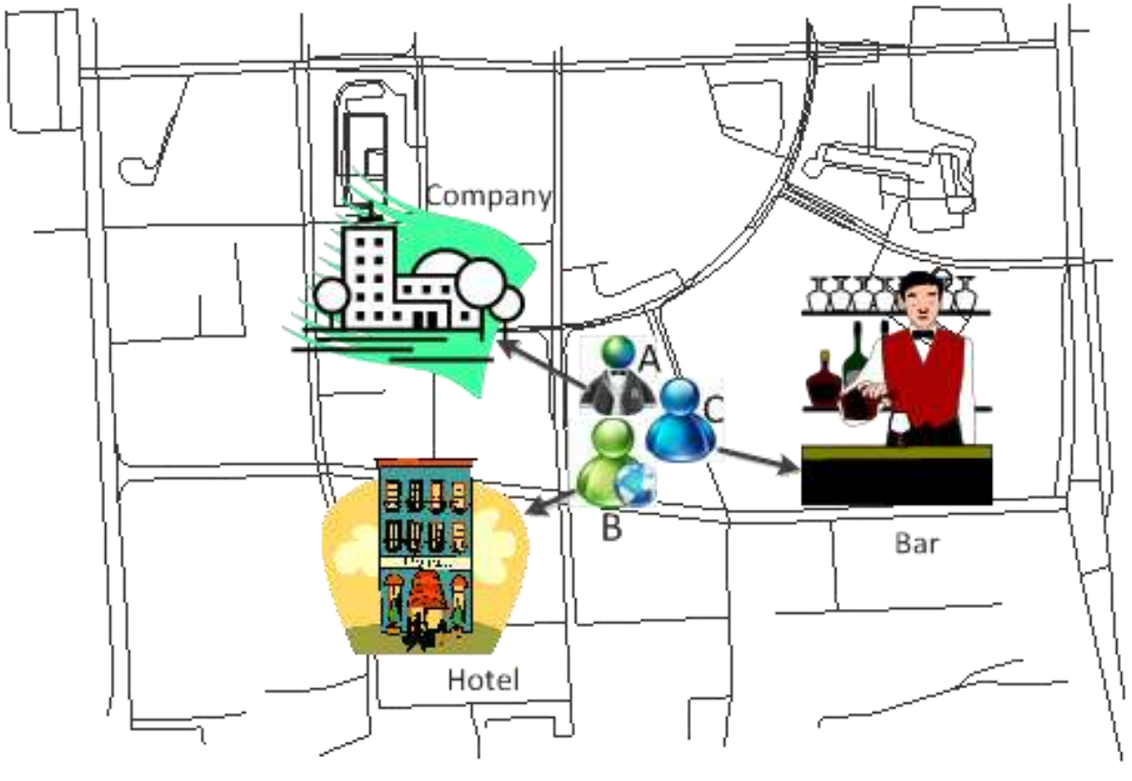
[Symeonidis et al. 2008]

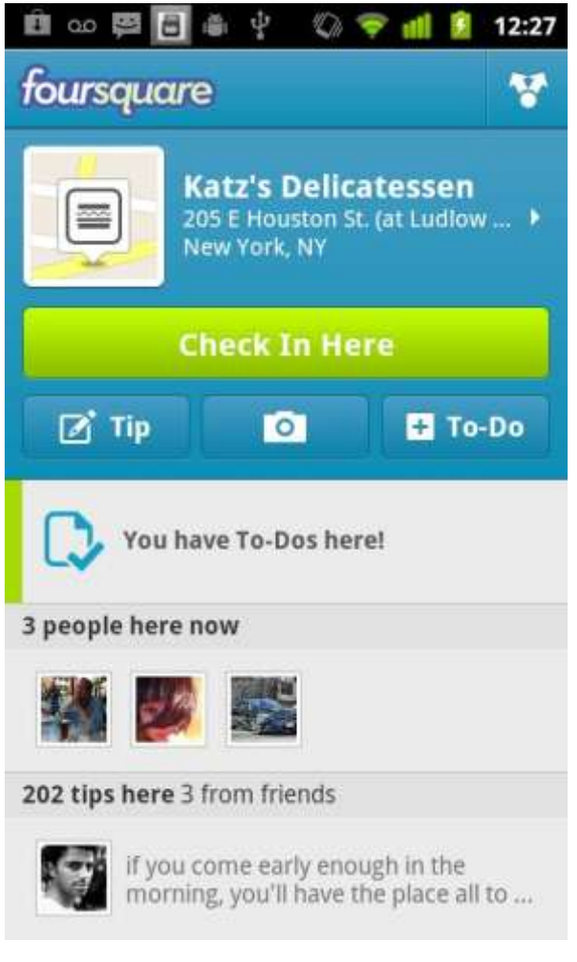
Collaborative Activity and Location Recommendation

- 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





Problem Definition

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

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

Related Work

- 通常先从轨迹中抽取重要地点，然后根据情境信息和历史数据对它们进行标注
  - 使用关系马尔可夫网为地点标注上用户在该地点发生的活动
  - 使用层叠条件随机场来同时抽取并标注重要地点。在标注集合中只有四类名称，即公司、家庭、朋友家和停车场。很难增加大量新的地点名称。
  - 当用户和朋友共享位置信息时，对位置命名方式的偏好。研究了位置的访问频度熵、社交网络、朋友对地点的熟悉程度和分享人对地点的隐私程度等影响因素。但该工作将位置命名偏好只分为三类，即街道地址、语义地址和混合地址，并不试图得出具体的名称。
- 我们希望能标注具体的位置名称，也就是在兴趣点(Point of Interest，缩写为POI)层次上进行标注

Positioning Error & Dense POI

- 坐标的误差
- 兴趣点的高密度、多层次以及大尺度特性

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

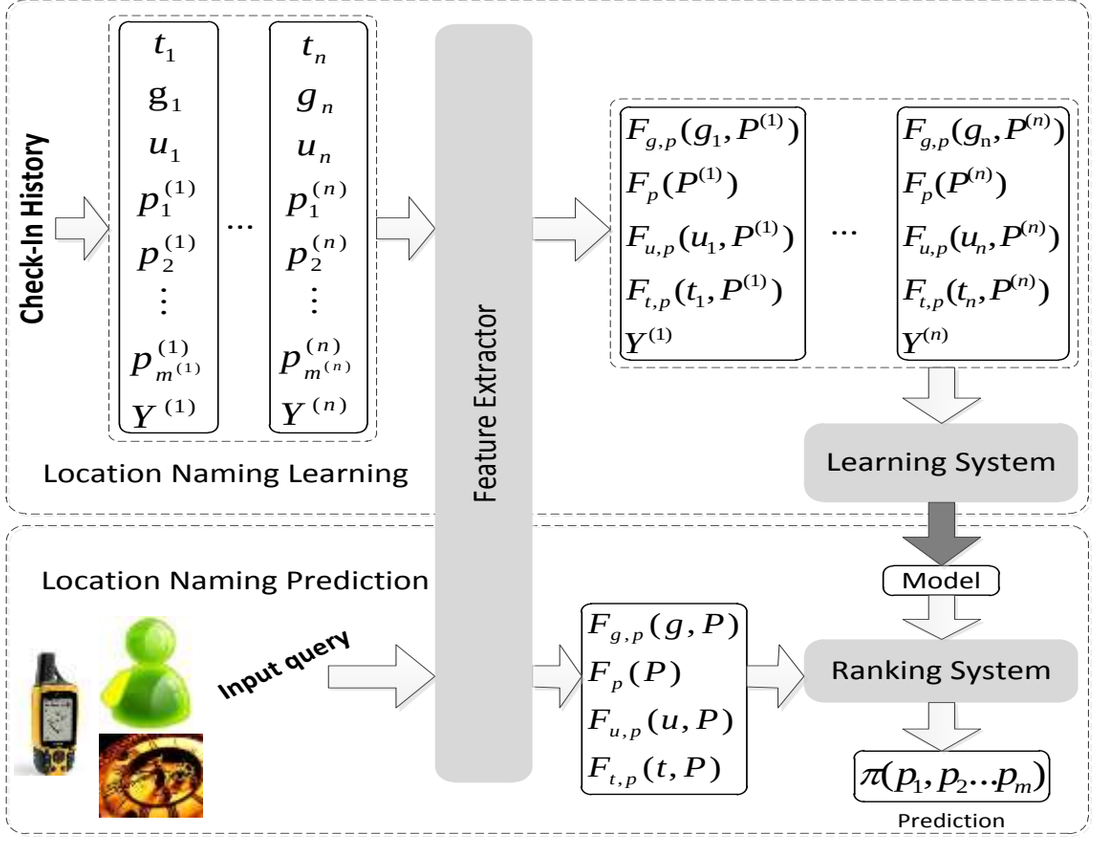
Data Sparsity

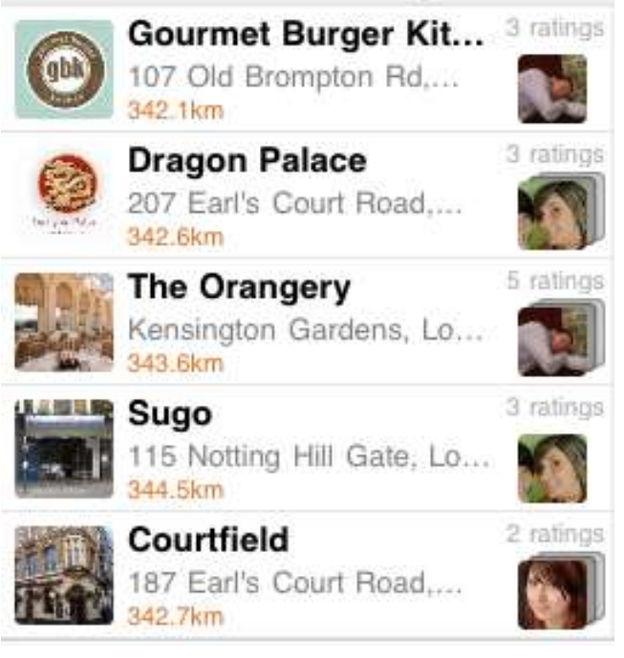
- Dianping
  - Reviews of local businesses
  - Check-in functionality

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





Static Features

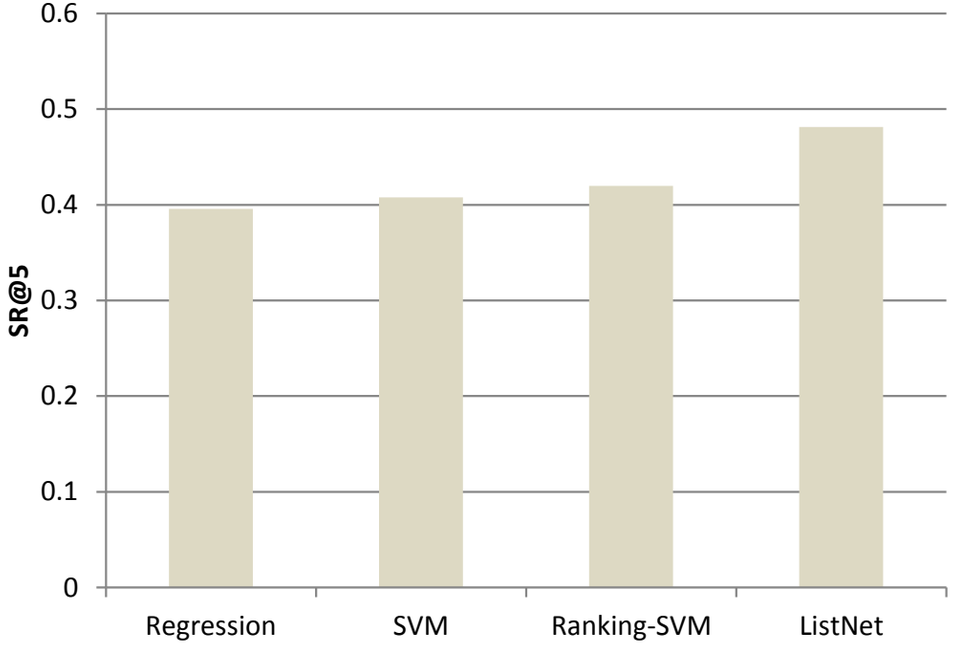
- Number of reviews related to it
- Average score given by social network users
- Number of web pages referring to it
- Number of check-ins
- Number of people checked-in
- Number of photos users have uploaded

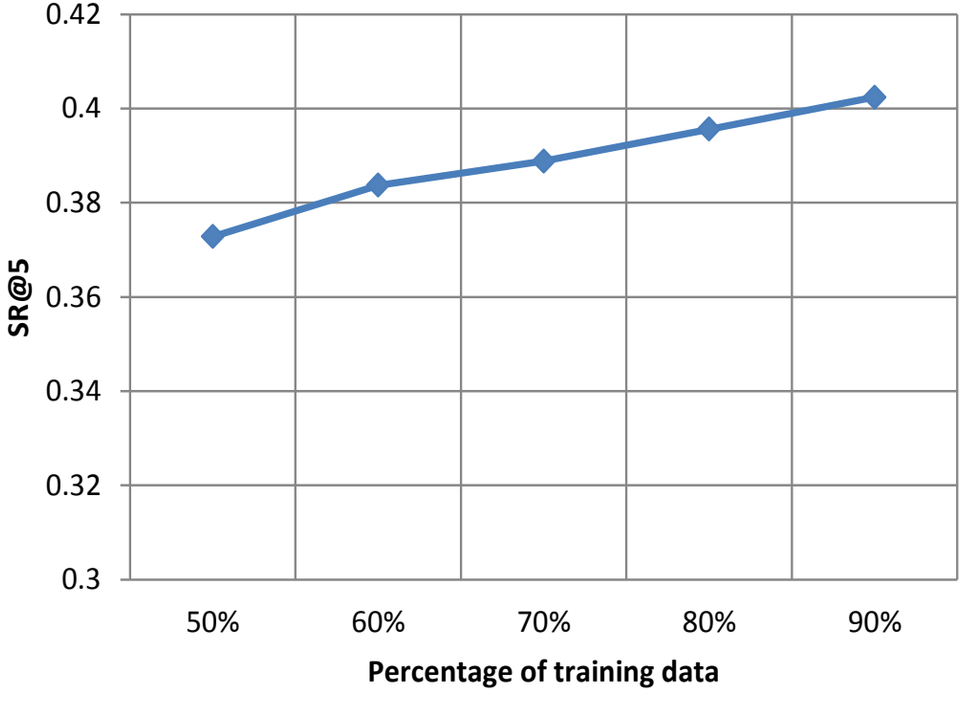
Dynamic Features

- 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

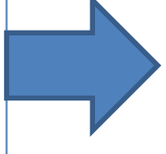
- Evaluation metric
  - Success Rate (SR) at k
  - $SR@k = \frac{|\{query|query\ is\ tested\ as\ accurate\ at\ k\}|}{|\{query\}|}$
- Ranking algorithm selection
- The impact of training data size





Fake Check-In Problem

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

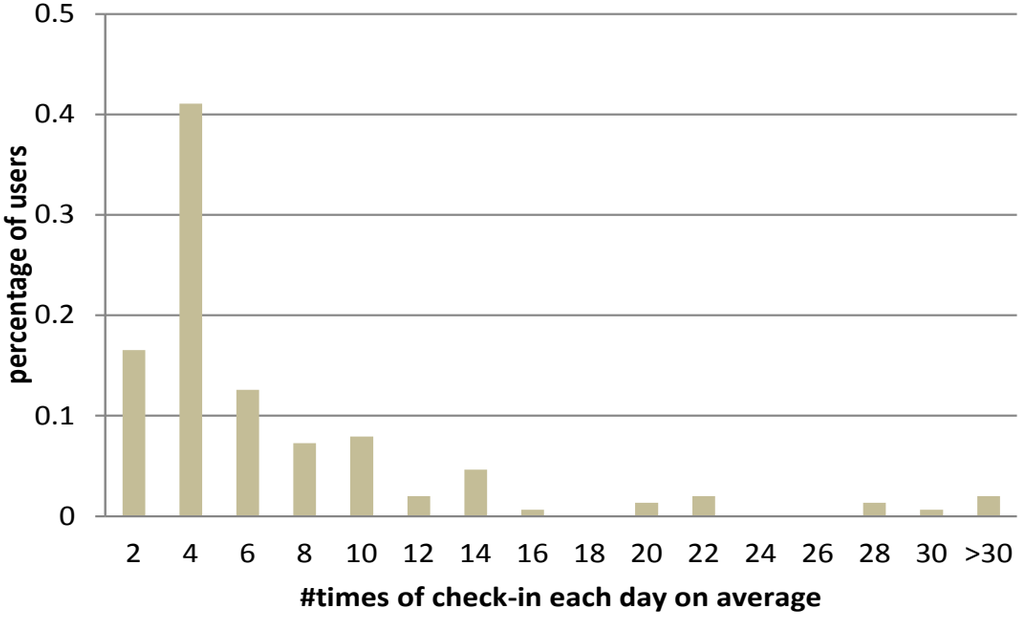


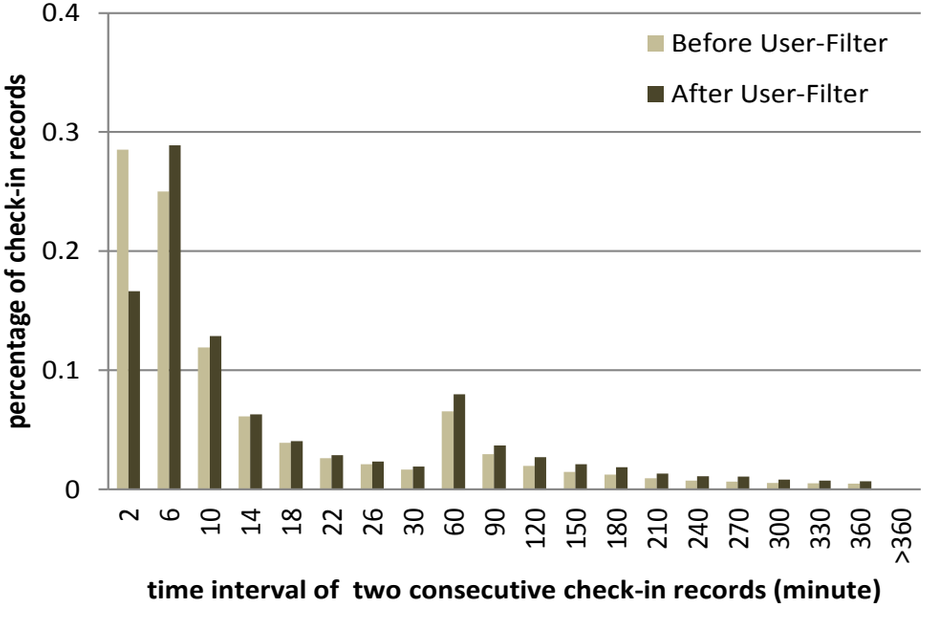
- Frequent check-ins
- Super human speed
- Rapid-fire check-in

Fake Check-In Problem

- Fake users - If a user check-in a lot of locations each day
- Fake check-in record r - if the following condition meets

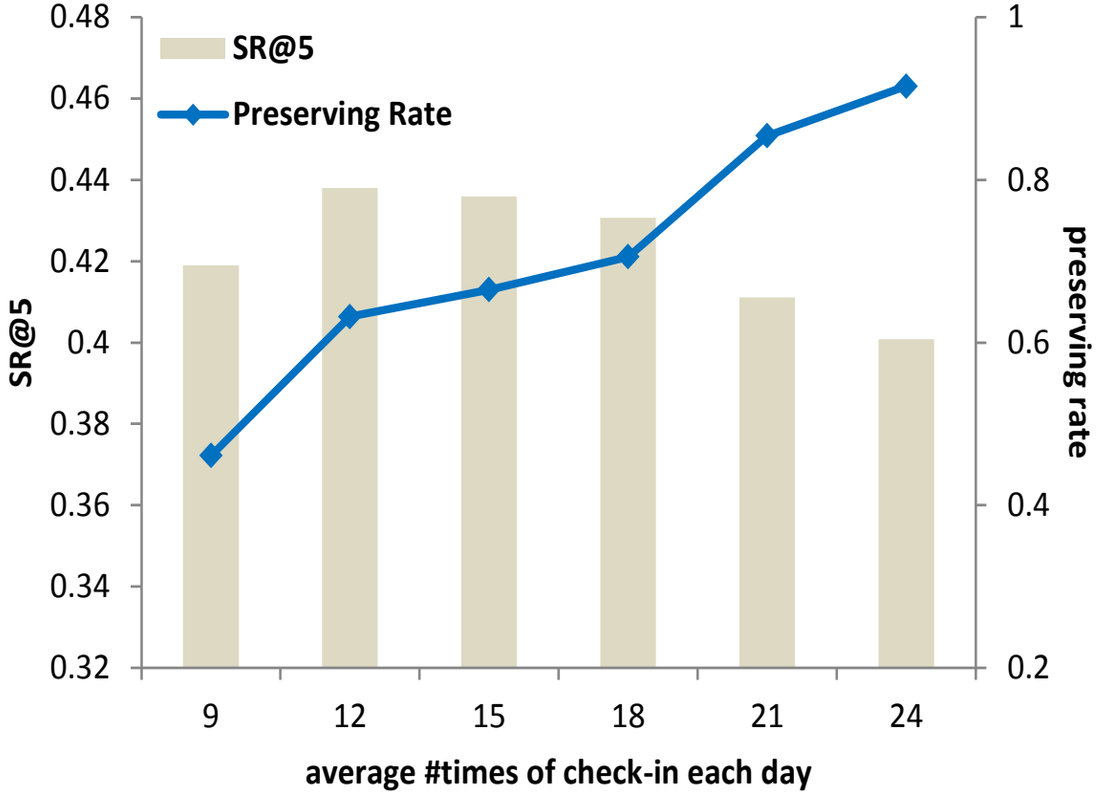
- r has a subsequence record r_{sub}
- $r_{sub}.t - r.t < th_{rf}$.
- $dist(r_{sub}.g, r.g) < 10\ meters$
- $r_{sub}.p! = r.p$

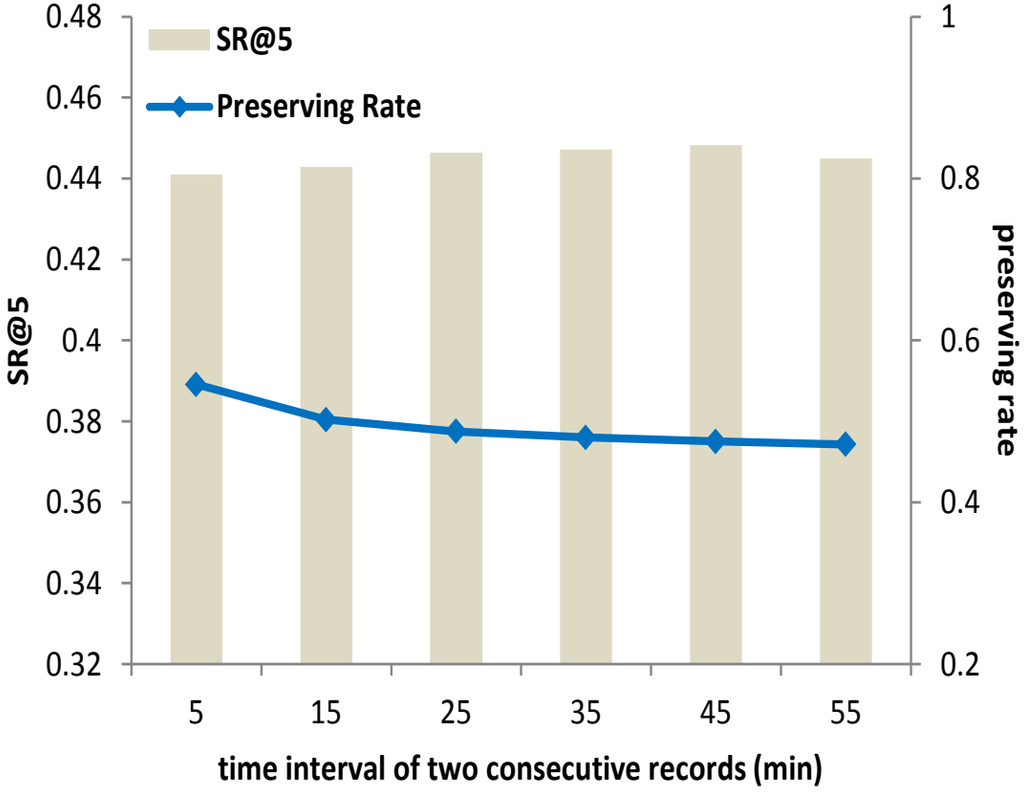




Fake Check-in Filtering

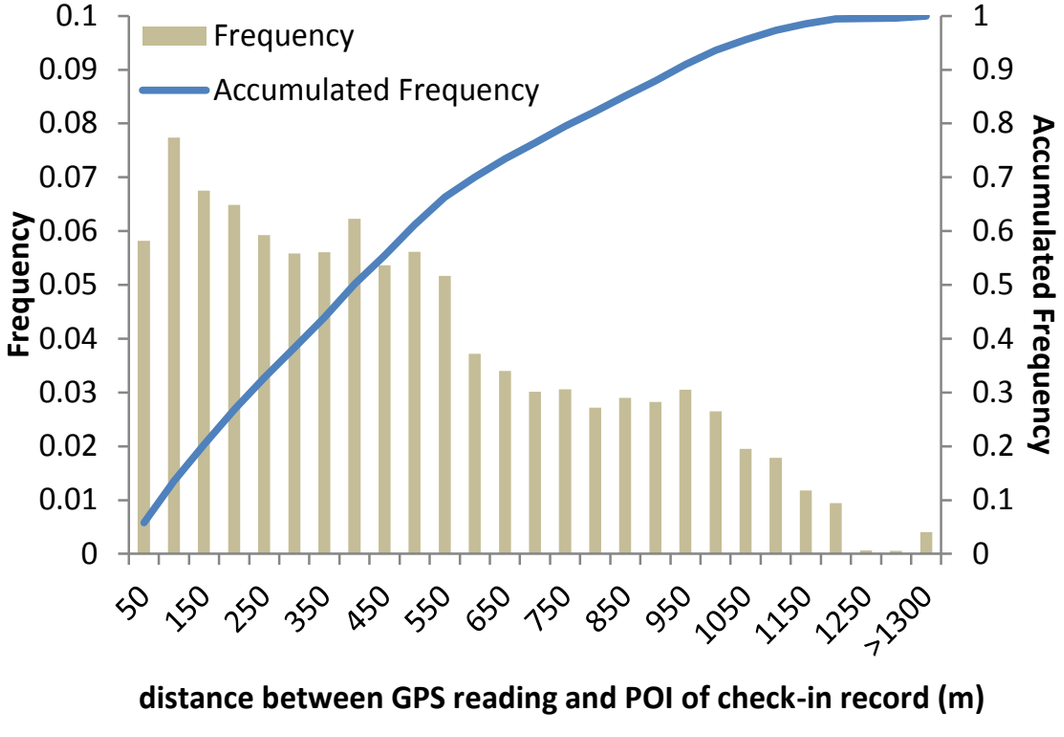
- Large impact of filtering fake users
  - Fake users are so random that it is difficult to predict their check-in
- Little impact of filtering fake check-in records

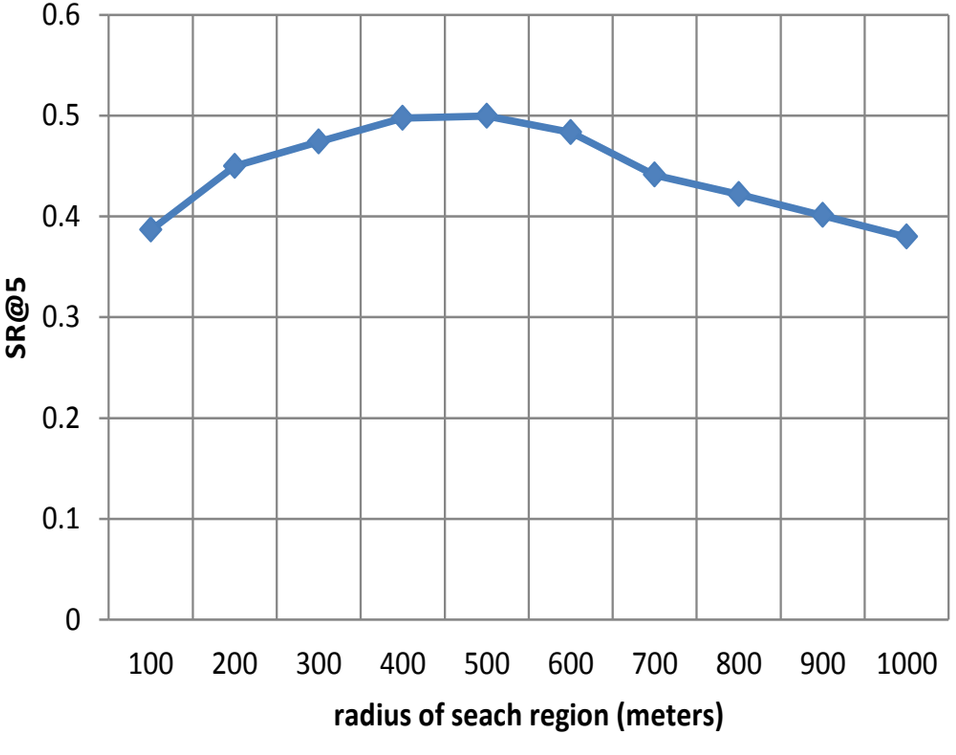




Search Radius

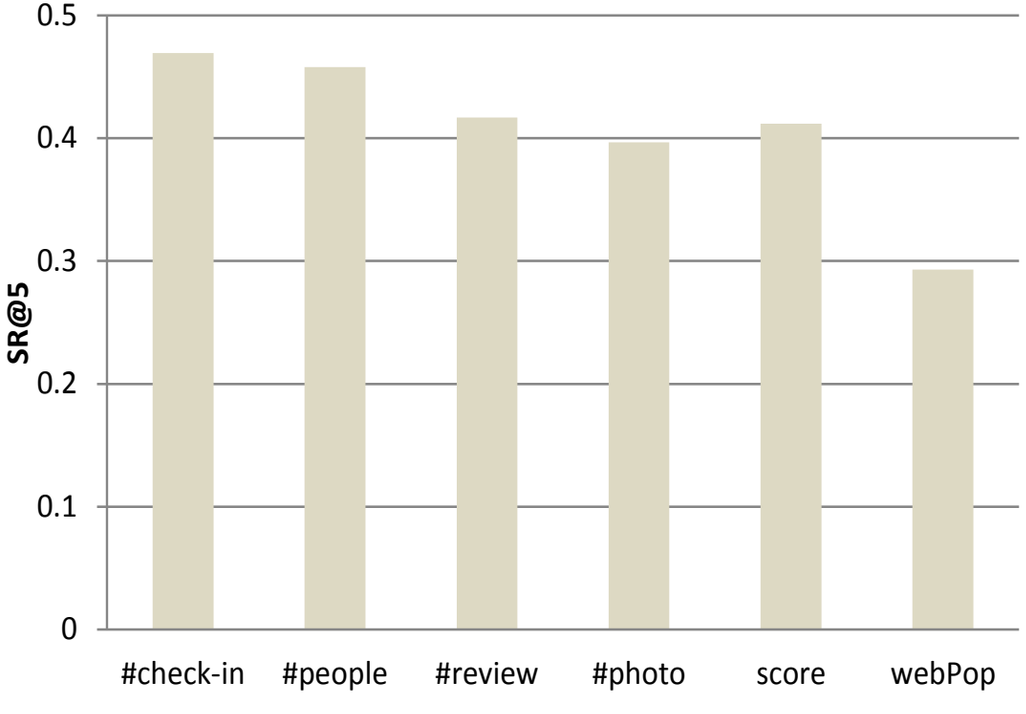
- Most check-ins are at nearby locations
- Distant check-ins are considered as noises
- Significant impact of different search radius

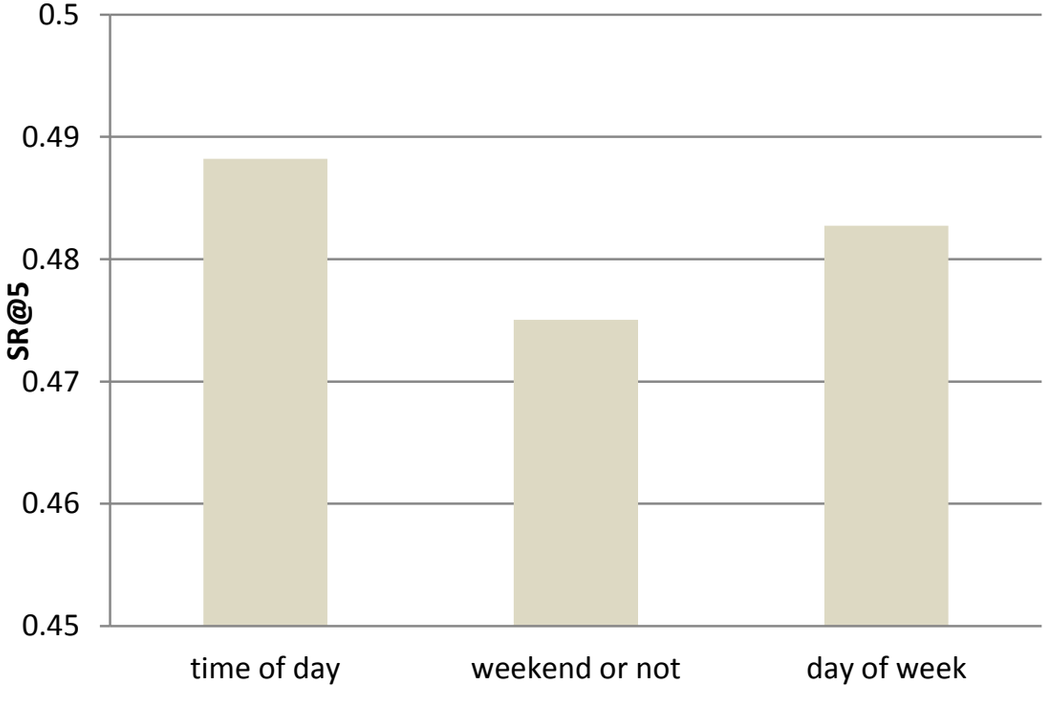




Feature Effectiveness

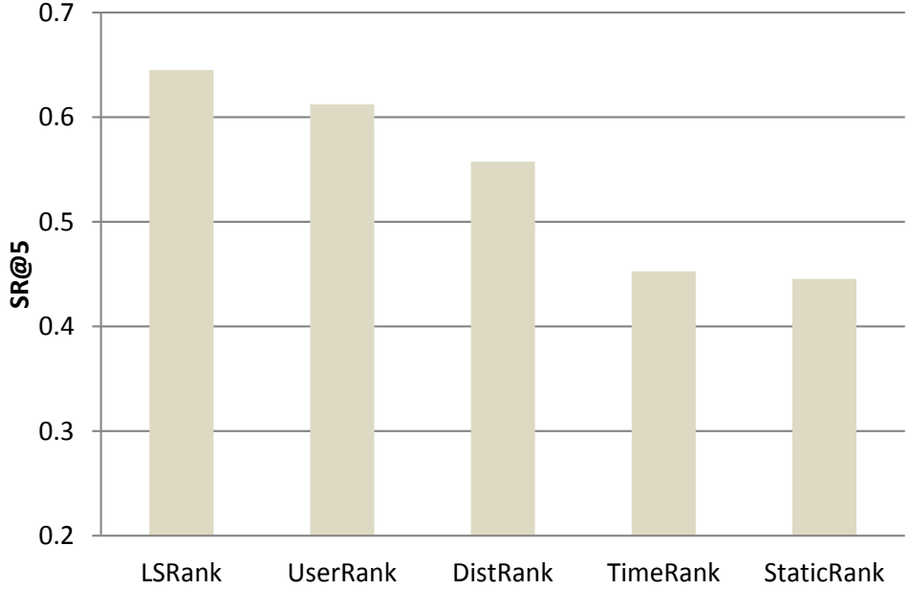
- #check-in is significantly better than webPop
- No big difference of different temporal patterns





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)

用户位置预测

- 签到的顺序性
  - 用户当前的位置对他们下面要去的位置有很大的影响。例如，在用户上午签到了公园以后，他们一般将会去吃午饭，所以很可能会在附近的餐馆签到
- 签到的空间邻近性
  - 大部分相邻签到之间的距离都不超过4公里，距离越长，签到的概率也是急剧衰减的
- 签到的周期性
  - 不管是周末还是工作日用户中午都会倾向在餐馆签到
- 签到兴趣点类型的动态性
  - 周末的下午用户可能会有更多运动和娱乐方面的签到，而在工作日这类签到可能会发生在晚上

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
- User location prediction

Thanks!

Xing Xie

Microsoft Research Asia