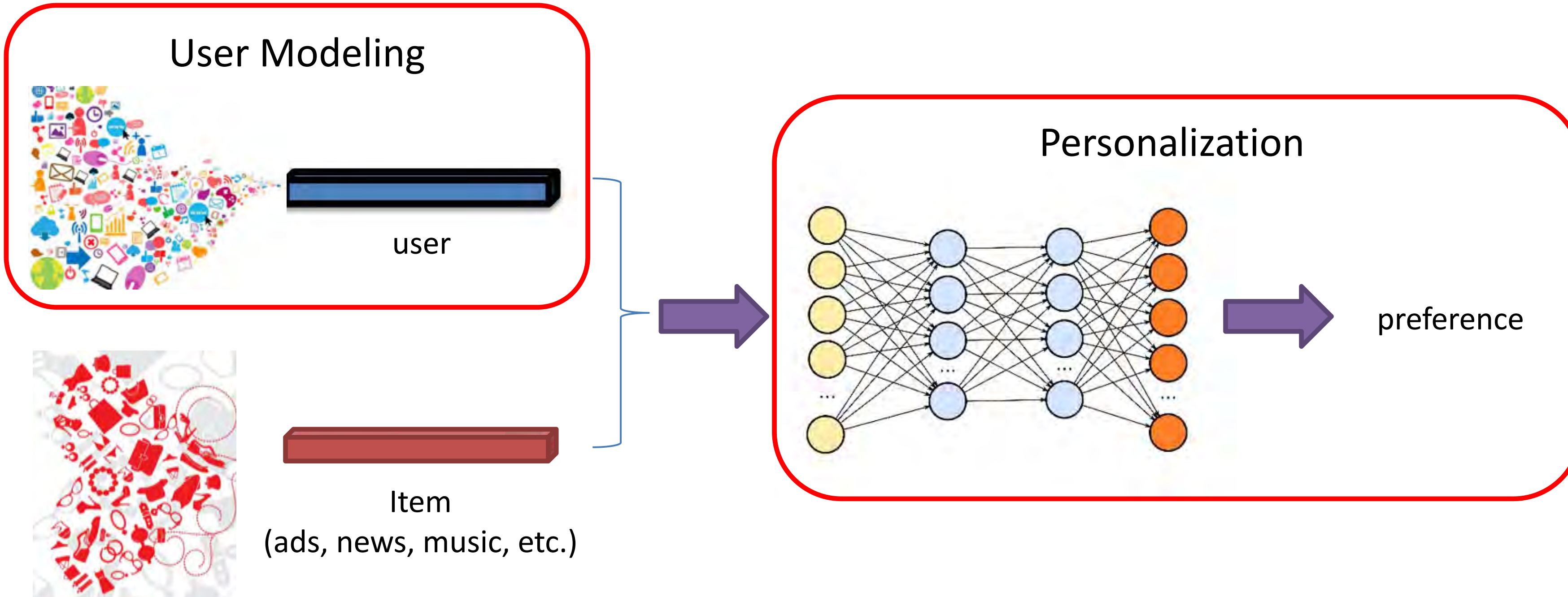
结合跨平台异构数据的推荐系统

谢 幸
微软亚洲研究院

用户画像与推荐系统


User Modeling
user
Item
(ads, news, music, etc.)
Personalization
preference

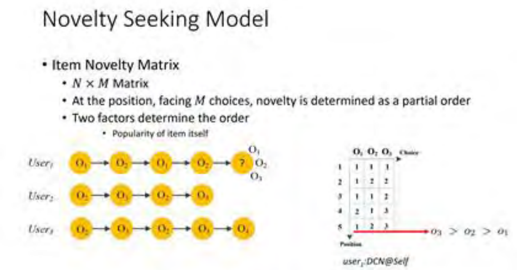
相关研究工作

Big Five Personality

WSDM 2017

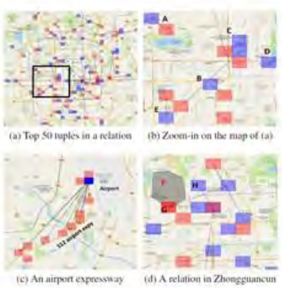
Consumer Impulsivity

UbiComp 2015



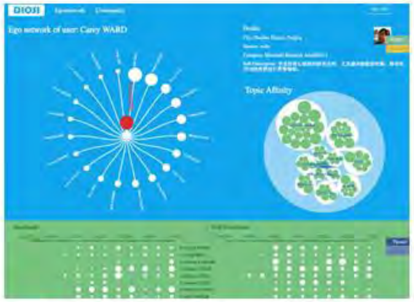
Novelty Seeking Trait

WWW 2015/WWW 2014



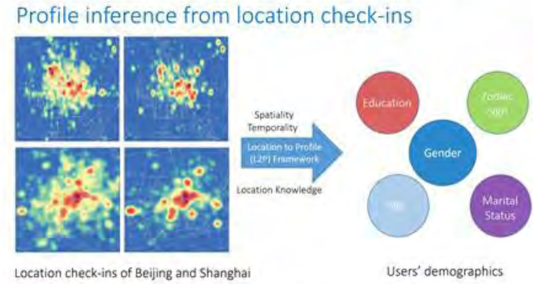
Location Interests

IJCAI 2017



Dynamics of Online Intimacy

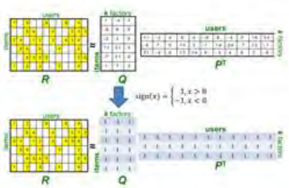
WSDM 2016



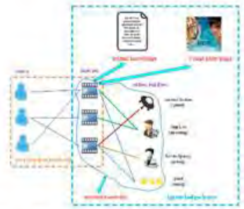
Location to Profile

WSDM 2015

相关研究工作

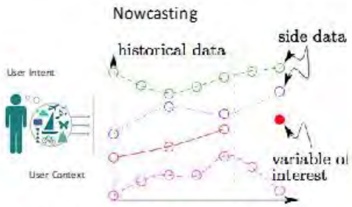


KDD 2017



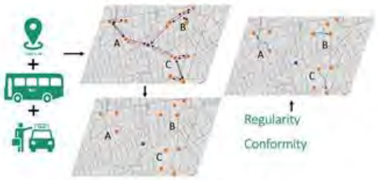
Knowledge Enhanced Recommendation

KDD 2016



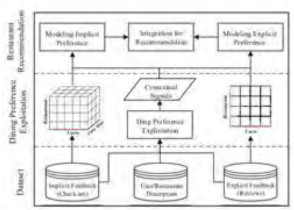
Contextual Intent Tracking

KDD 2016/best student paper

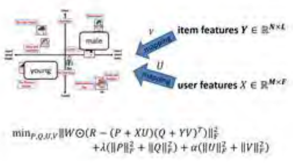


Regularity and Conformity

KDD 2015

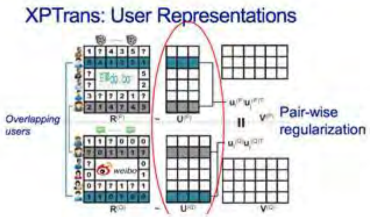


WWW 2016



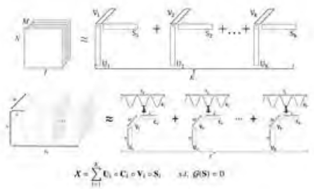
Bayesian Content-aware CF

IJCAI 2016



Cross-Platform Behavior Prediction

AAAI 2016



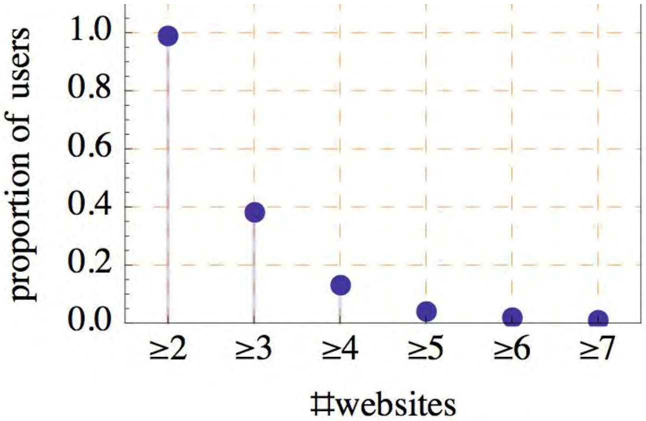
App Usage Forecasting

UbiComp 2016

LifeSpec跨平台用户行为数据集

LifeSpec跨平台用户行为数据集

- 4 (major) networks: Jiepang, Weibo, Douban, Dianping
- 1.4M+ unique (deterministically identified) users accounts
- Heterogeneous footprints: tweets, photos, check-ins, movies, books, music, offline events, online purchase history, etc.
- Rich user profiles integrated from different sites (publicly available)



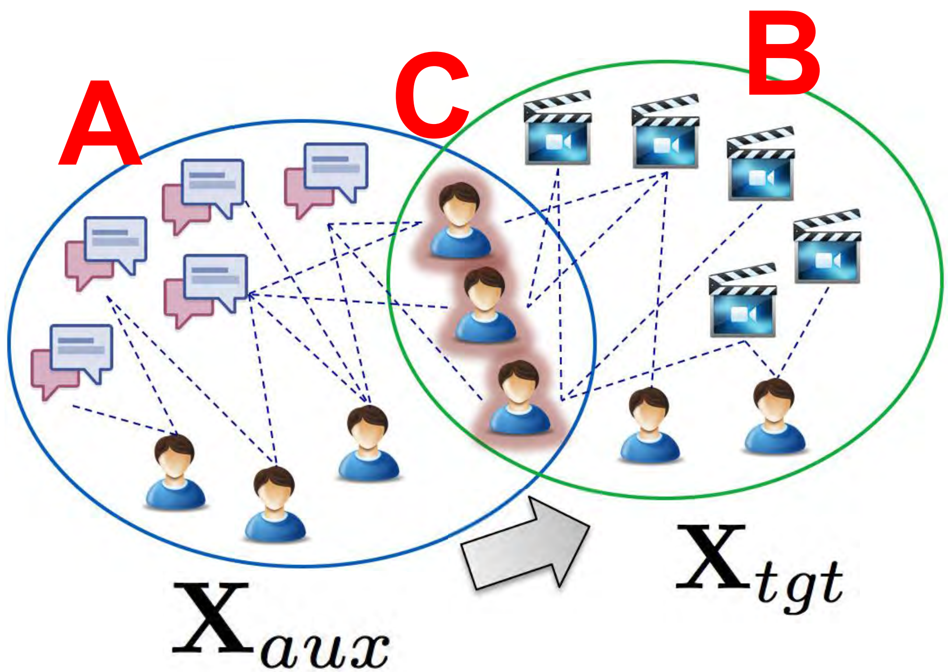
Age	Gender	Residence	Relationship	Occupation	College	High School	Self description	...

LifeSpec跨平台用户行为数据集

- 53 million footprints (check-in, movie, music, events, book, etc.)
- 3 million social links
- 39 million check-ins

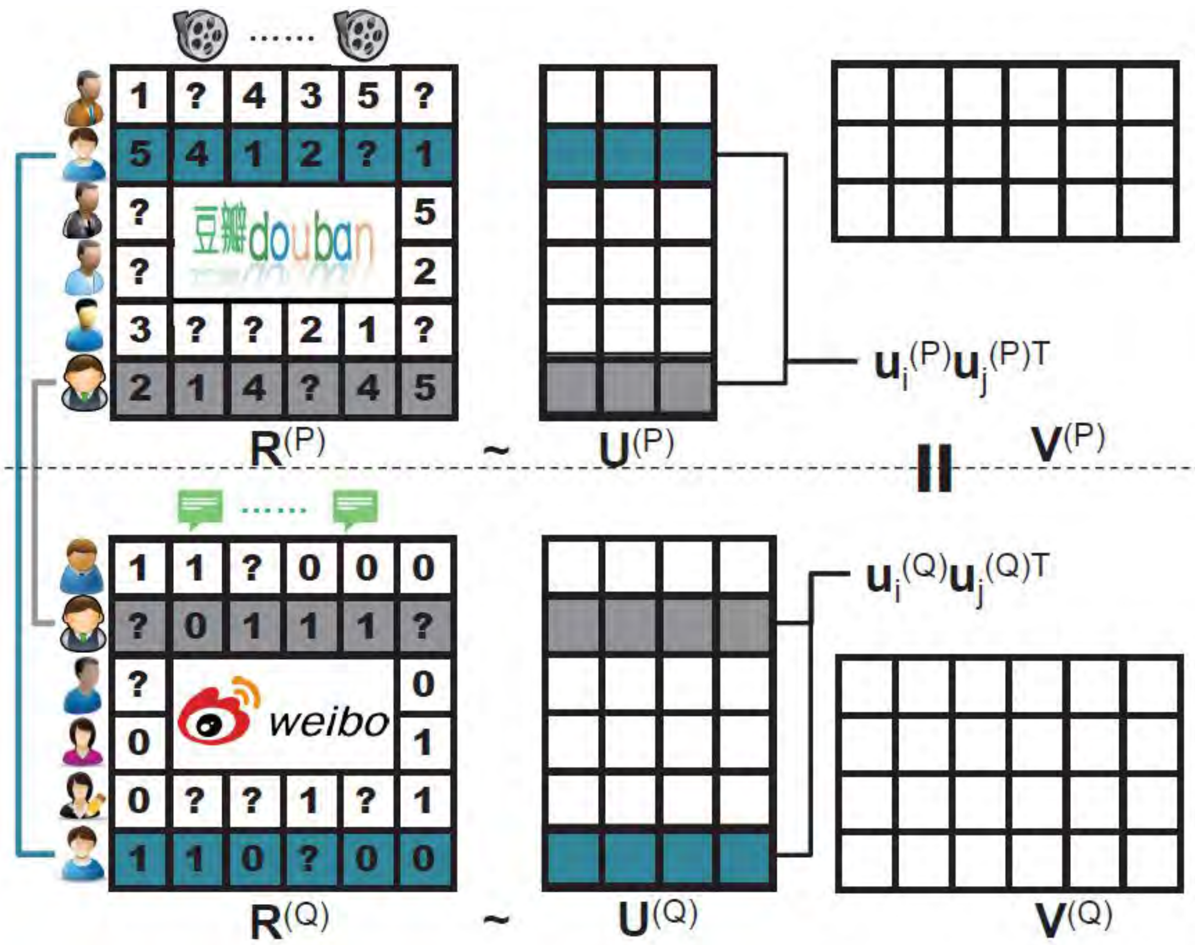
	city	Shanghai	Beijing	Guangzhou	Tianjin	Hangzhou	Hongkong	Xiamen	Suzhou	Nanjing	Chengdu	Wuhan	Xian
	users	417,681	162,764	53,089	15,490	34,322	12,599	10,123	19,673	21,558	23,372	20,975	15,261
Footprints	check-in	25,178,189	5,898,447	1,092,138	392,943	619,219	424,650	369,231	560,274	414,202	327,634	321,646	229,678
	movie	1,661,214	1,466,479	171,789	118,775	238,721	57,003	70,172	89,706	174,664	191,042	166,337	123,223
	music	766,165	737,254	85,953	60,658	103,936	30,313	29,716	39,701	82,513	88,426	76,316	62,876
	book	402,318	387,138	51,913	28,188	57,835	18,117	18,516	19,521	44,345	42,241	44,804	28,435
	event	609,076	803,158	101,246	52,133	78,587	18,277	20,889	27,400	46,788	66,640	44,764	72,902
	total	28,616,962	9,292,476	1,503,039	652,697	1,098,298	548,360	508,524	736,602	762,512	715,983	653,867	517,114

Partially Overlapped Users



XPTrans: User Representations

Overlapping users



Pair-wise regularization

实验结果

NO Transfer

User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A	Auxiliary platform data!	
C	0.779	0.805
B	**1.439**	**0.640**

User set	Douban book to Weibo social tag	
	RMSE	MAP
A	**0.429**	**0.464**
C	0.267	0.666
B	Auxiliary platform data!	

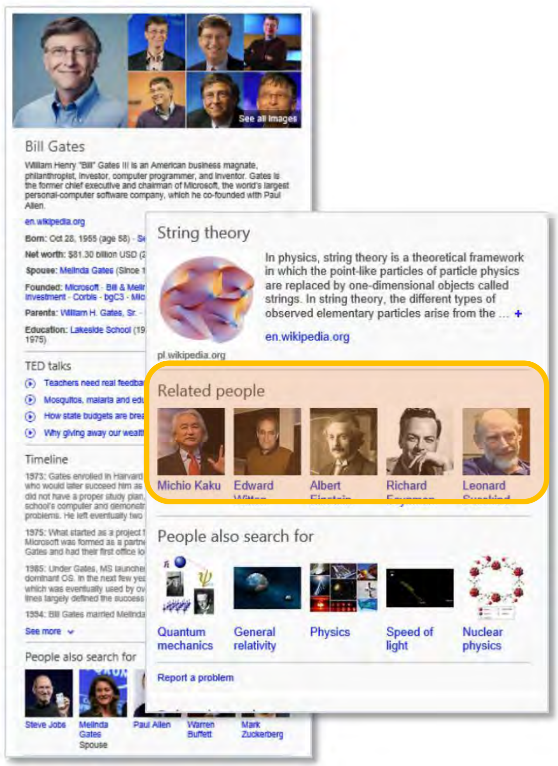
Transfer via Different Latent Spaces

User set	Weibo tweet entity to Douban movie	
	RMSE	MAP
A		
C	0.715	0.821
B	**0.722**	**0.820**

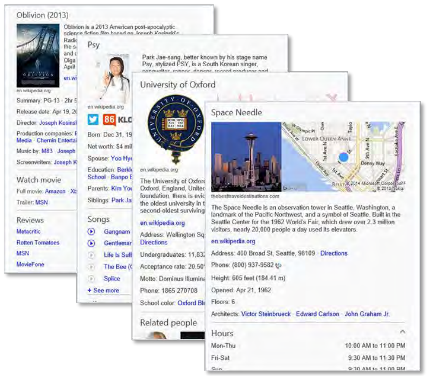
User set	Douban book to Weibo social tag	
	RMSE	MAP
A	**0.374**	**0.533**
C	0.236	0.705
B		

知识图谱

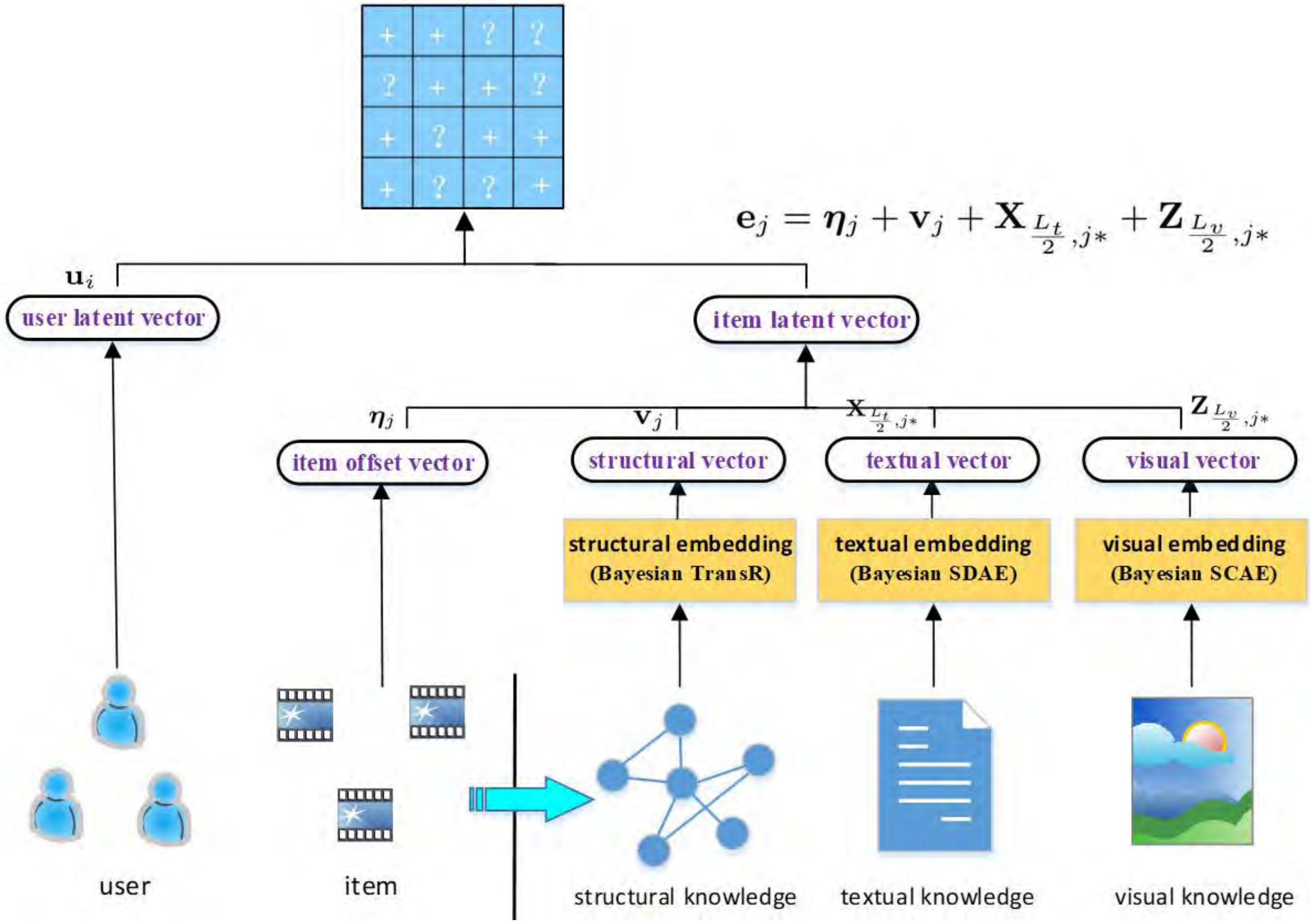
>17B Facets& Relationships



Dozens of domains



结合异构知识的推荐


$\mathbf{e}_j = \boldsymbol{\eta}_j + \mathbf{v}_j + \mathbf{X}_{\frac{L_t}{2},j*} + \mathbf{Z}_{\frac{L_v}{2},j*}$
$\mathbf{u}_i$
user latent vector
item latent vector
$\boldsymbol{\eta}_j$
$\mathbf{v}_j$
$\mathbf{X}_{\frac{L_t}{2},j*}$
$\mathbf{Z}_{\frac{L_v}{2},j*}$
item offset vector
structural vector
textual vector
visual vector
structural embedding
(Bayesian TransR)
textual embedding
(Bayesian SDAE)
visual embedding
(Bayesian SCAE)
user
item
structural knowledge
textual knowledge
visual knowledge

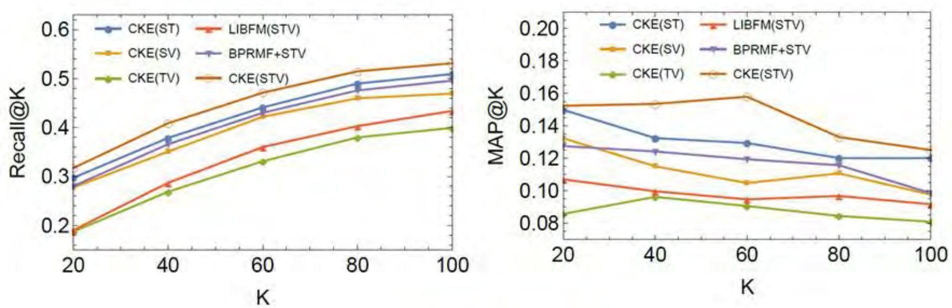
数据

- MovieLens-1M
  - 1-step subgraph includes category, director, writer, actors, language, country, production date, rating, nominated awards, and received awards
- IntentBooks
  - 9-month Bing query logs, apply entity linking to find out book entity
  - 1-step subgraph includes category, author, publish date, belonged series, language, and rating

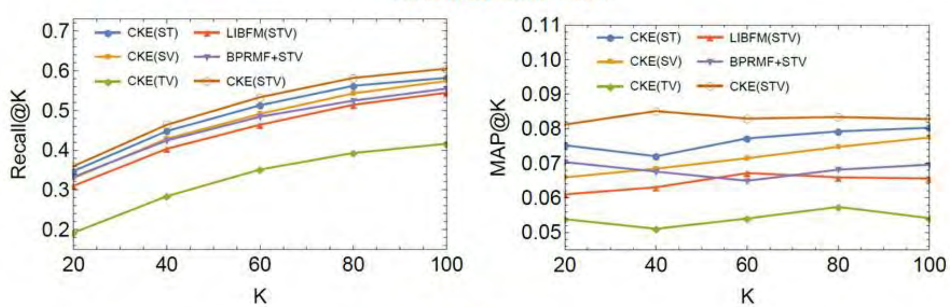
	MovieLens-1M	IntentBooks
#user	5,883	92,564
#item	3,230	18,475
#interactions	226,101	897,871
#sk nodes	84,011	26,337
#sk edges	169,368	57,408
#sk edge types	10	6
#tk items	2,752	17,331
#vk items	2,958	16,719

实验结果

- CKE(ST), CKE(SV), CKE(TV): only two types of knowledge
- LIBFM(STV): all knowledge as raw features
- BPRMF+STV: not joint-learning

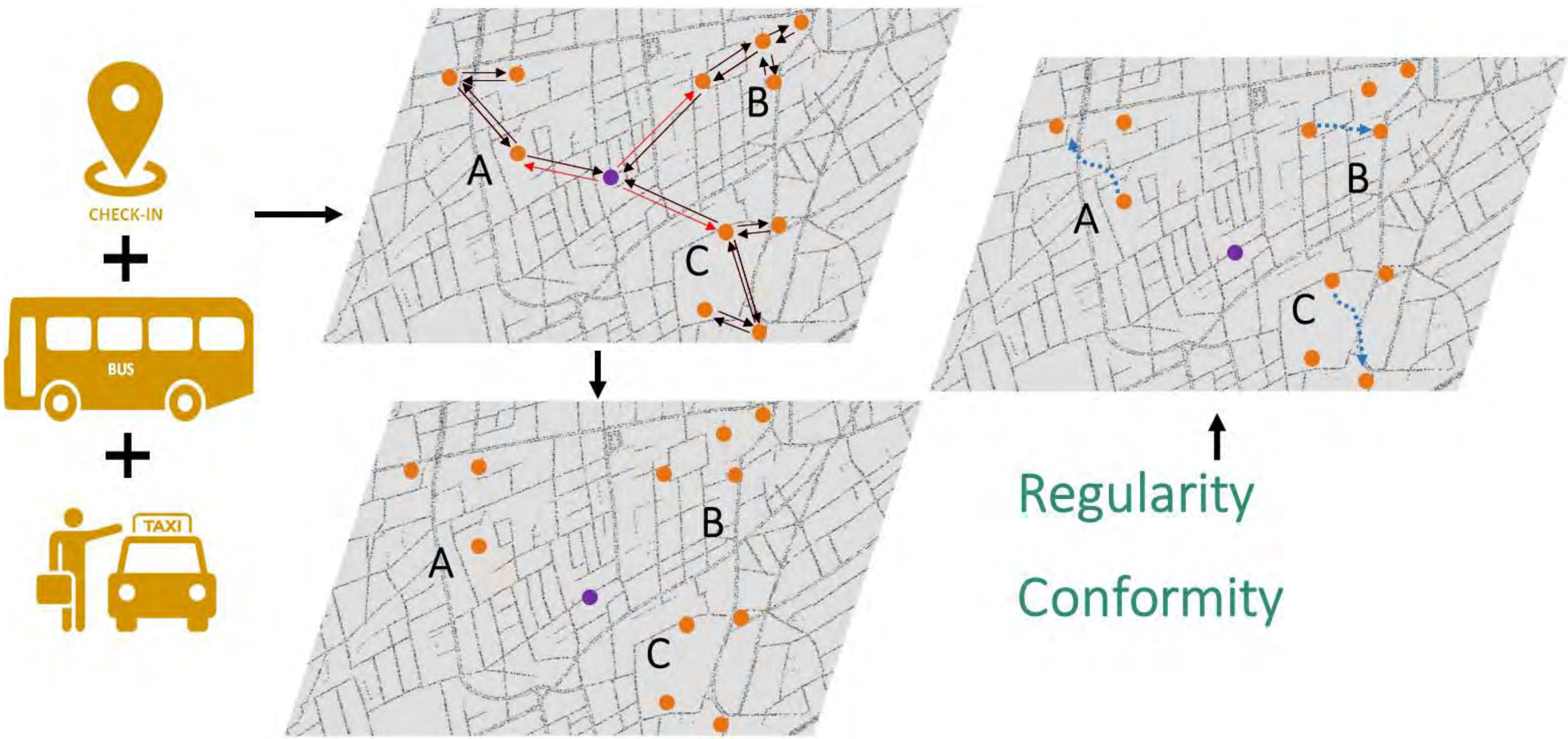


MovieLens-1M

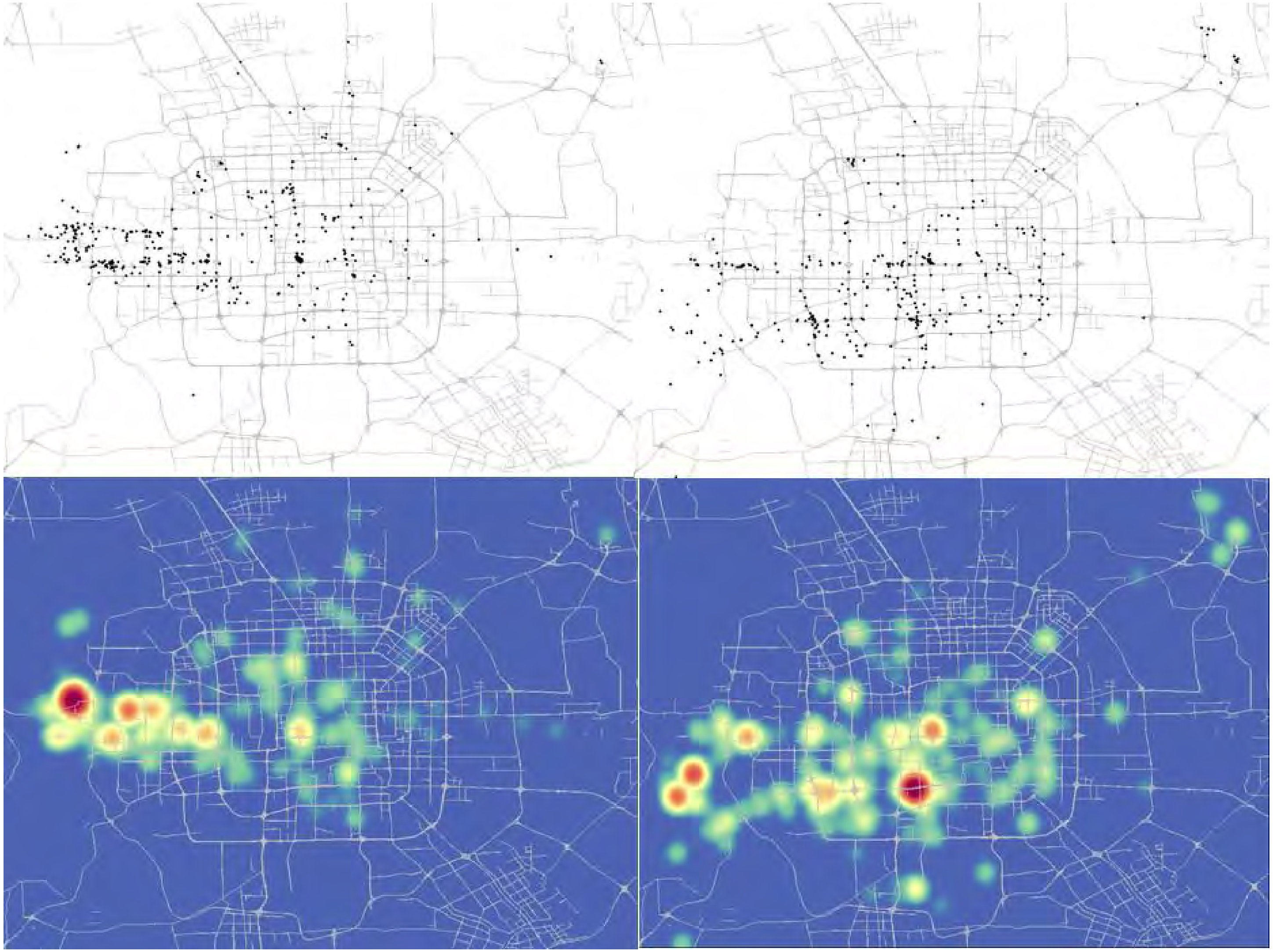


IntentBooks

基于跨平台位置数据的行为预测



Regularity

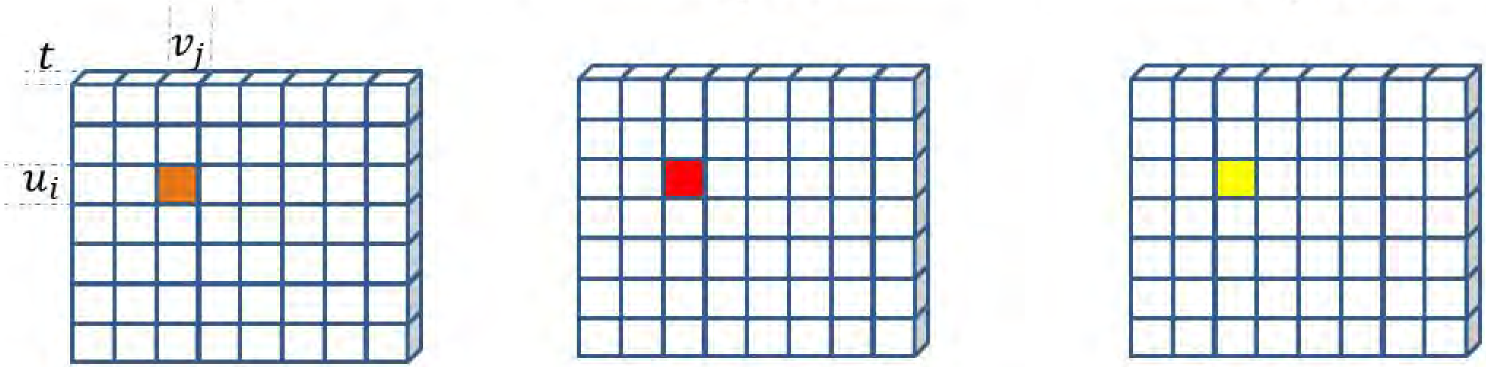


Conformity

Main Idea

- Split days into T time slots $\mathcal{T} = \{t_1, t_2, \dots, t_T\}$
- *M* users and *N* venues
  - $\mathcal{U} = \{u_1, u_2, \dots, u_M\}$
  - $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$
- Preference matrix of $\mathcal{U}$ to $\mathcal{V}$ at time ***t*** : $\mathbf{R}(t) \in \mathbb{R}^{M \times N}$

$$R_{ij}(t) = R_{ij}^{(r)}(t) + R_{ij}^{(c)}(t)$$



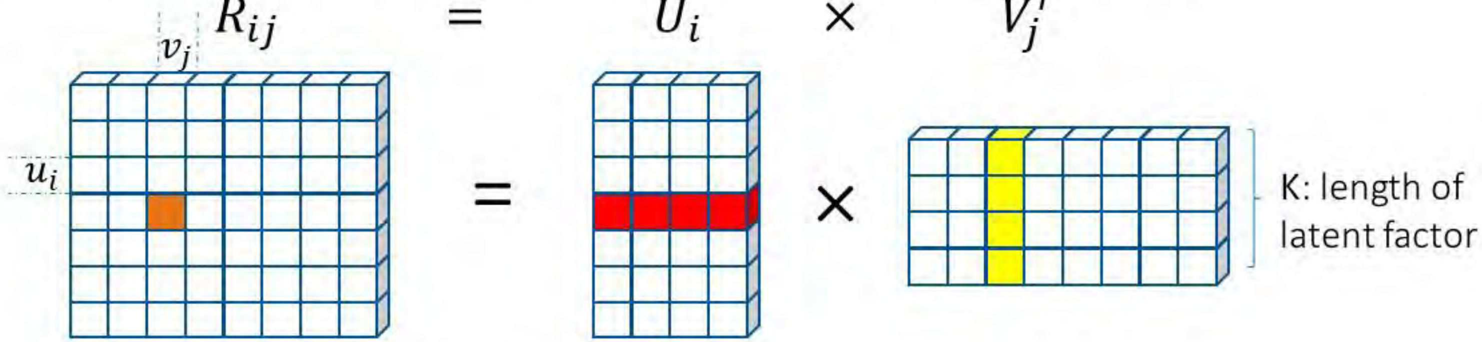
Regularity term

Conformity term

Conformity Term (Check-in Data)

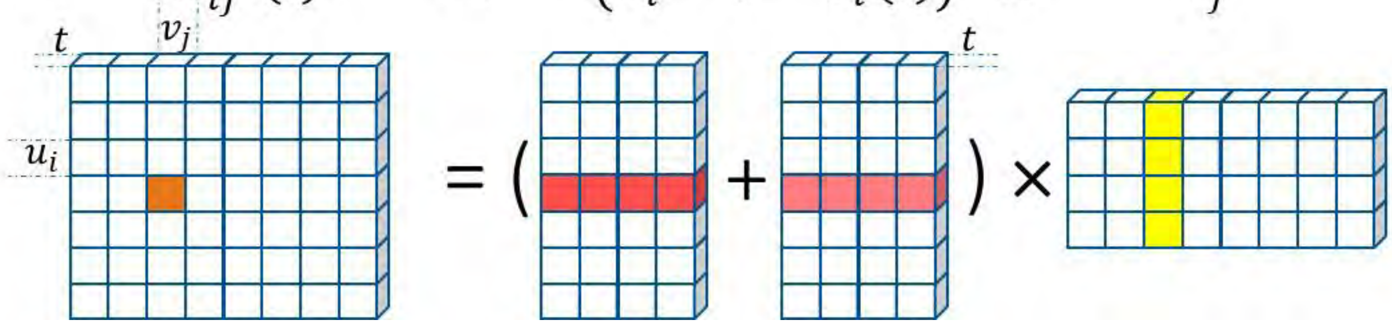
- Traditional collaborative model: Matrix Factorization

$$R_{ij} = U_i \times V_j^T$$



- Time-aware Matrix Factorization

$$R_{ij}^{(c)}(t) = (U_i + U_i(t)) \times V_j^T$$

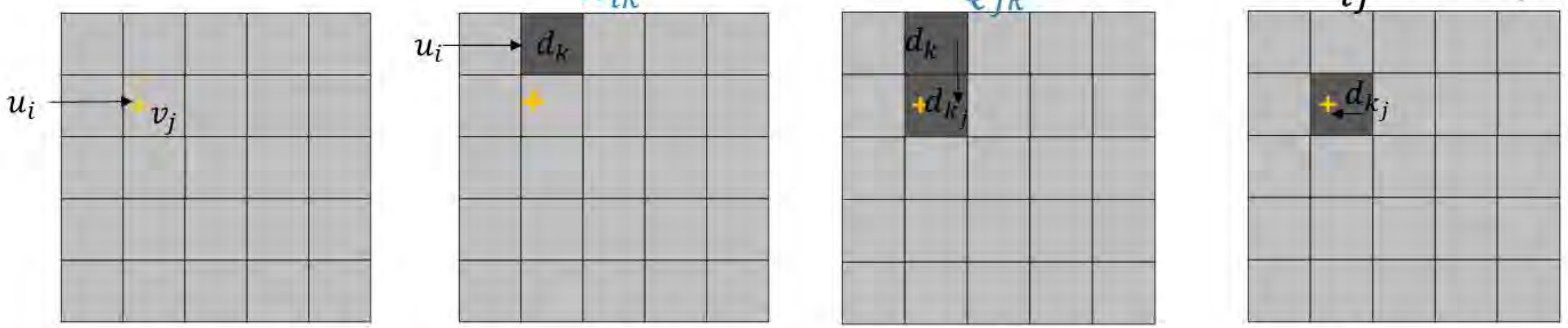


Regularity Term (Heterogeneous Data)

- Split the city into I grid cells: $C = \{d_1, d_2, \dots, d_I\}$
- v_j belongs to a grid d_{kj}
- u_i travels from a grid d_k to v_j

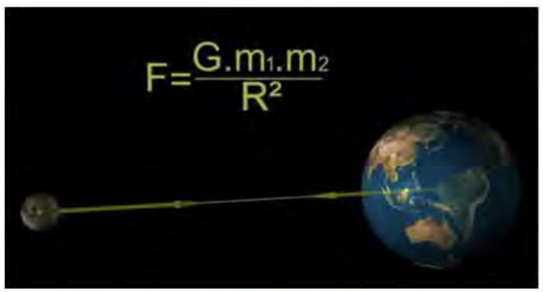
$$\Pr(v_j|u_i) \propto \textstyle\sum_{k=1}^{I} \Pr(d_k|u_i) \cdot \Pr(v_j|d_k)$$
$$= \textstyle\sum_{k=1}^{I} \Pr(d_k|u_i) \cdot \Pr(d_{kj}|d_k) \cdot \Pr(v_j|d_{kj})$$

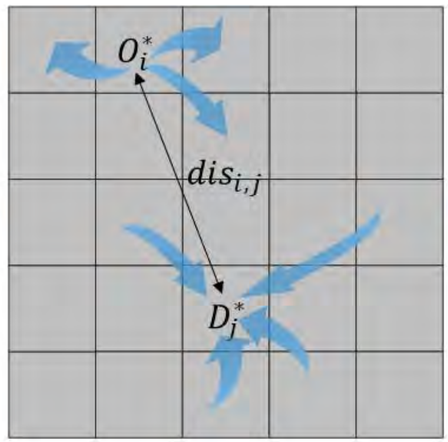
H_{ik} Q_{jk} $\longrightarrow R_{ij}^{(r)} = H_i \cdot Q_j^T$



$\Pr(v_j|u_i)$ $\Pr(d_k|u_i)$ $\Pr(d_{kj}|d_k)$ $\Pr(v_j|d_{kj})$

Gravity Model





$$T_{i,j}^* = c\,\frac{(O_i^*)^a \cdot (D_j^*)^b}{\exp(r \cdot dis_{i,j})},$$

$$* \in \{B, A, C\}$$

$m_1 \rightarrow (O_i^*)^a$, O_i^*: number of individuals leaving grid d_i in data*

$m_2 \rightarrow (D_j^*)^b$, D_j^*: number of people going toward d_j in data*

$R^2 \rightarrow \exp(r \cdot dis_{i,j})$, $dis_{i,j}$: distance between d_i and d_j

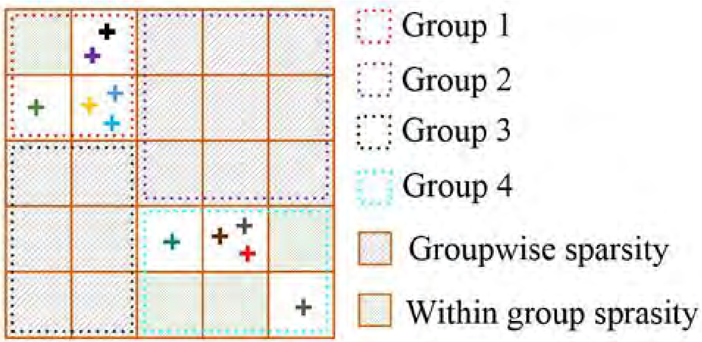
B: bus data

A: taxi data

C: check-in data

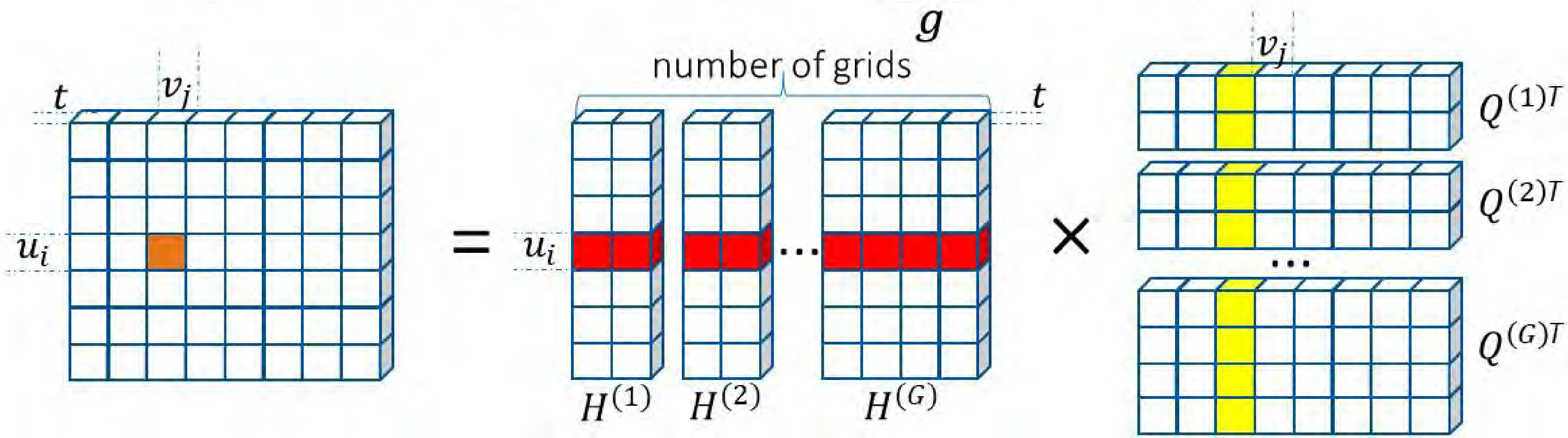
c,a,b,r: constants

Two-level Sparsity



Cluster grids into G group

$$R_{ij}^{(r)} = H_i \cdot Q_j^T \longrightarrow R_{ij}^{(r)} = \sum_{g} H_i^{(g)} \cdot Q_j^{(g)T}$$



RCH Model

- Sparse group lasso
- Objective function:

$$
\begin{aligned}
&\boldsymbol{P}(\boldsymbol{U}, \boldsymbol{U}(t), \boldsymbol{V}, \boldsymbol{H}(t), \theta^{B}, \theta^{A}) \\
&= \sum_{t\in\mathcal{T}} \left\|\boldsymbol{R}(t) - \sum_{g\in\mathcal{G}} \boldsymbol{H}^{(g)}(t) \sum_{*\in\mathcal{P}} \theta^{*} \left(\boldsymbol{Q}^{*(g)}\right)^{T} - (\boldsymbol{U} + \boldsymbol{U}(T))\boldsymbol{V}^{T}\right\|_F^2 \\
&\quad + \sum_{t\in\mathcal{T}} \Big(\underbrace{\boxed{(1-\alpha)\sigma \sum_{j=1}^{M} \sum_{g\in\mathcal{G}} \left\|\boldsymbol{H}_j^{(g)}(t)\right\|_2}}_{\text{Offers group-wise sparsity}} + \underbrace{\boxed{\alpha\sigma \sum_{j=1}^{M} \left\|\boldsymbol{H}_j(t)\right\|_1}}_{\text{Offers within-group sparsity}}\Big) \\
&\quad + \gamma\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right) + \beta \sum_{t\in T} \|\boldsymbol{U}(t)\|_F^2,
\end{aligned}
$$

where $* \in \mathcal{P} = \{A, B, C\}$ and $\theta^{C} = 1$.

- Optimization: alternative minimization

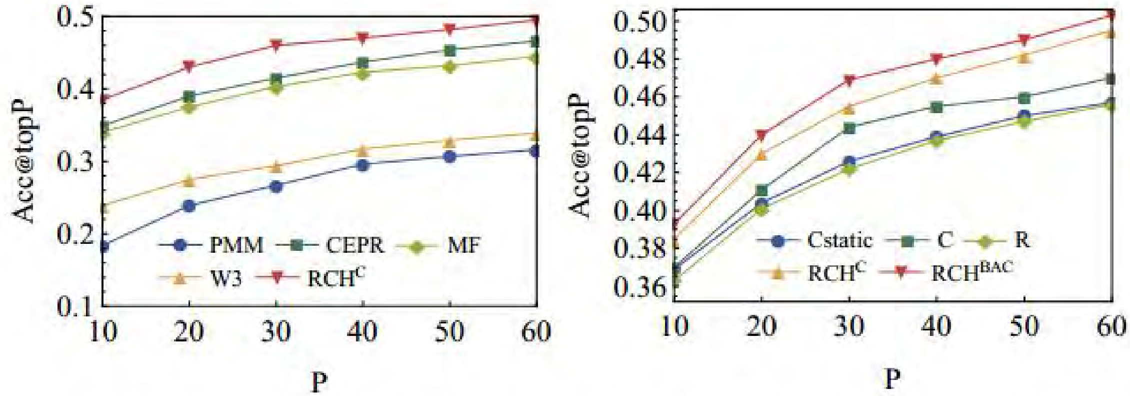
数据

Data Set	Check-in(Sina Weibo)	Bus Data	Taxi Data
City	Beijing	Beijing	Beijing
Scale of Data	12,133,504 check-ins	3,000,000 bus-trips	19,400,000 taxi transitions
Period	Mar. 2011 to Sep. 2013	Aug. 2012 to May 2013	Mar. 2011 to Aug. 2011
Content	user ID, check-in time, venue Id, venue's geo-coordinates	card Id, alighting time, boarding and alighting stops	times, geo-coordinates of boarding and alighting

实验结果

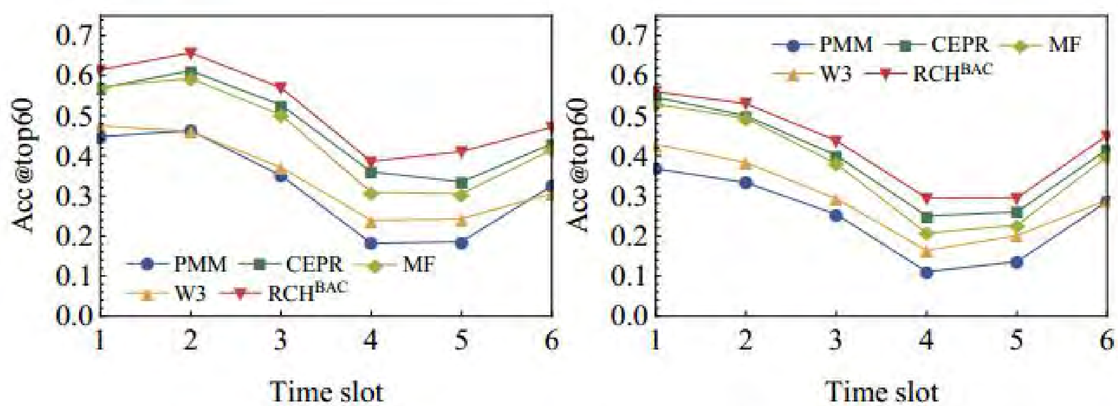
- Baselines
  - MF(Most Frequent Model)
    - Calculate the frequencies of users’ check-ins
  - PMM (Periodic Mobility Model)
    - 2-dimentional (home, work)
    - Time-independent spatial Gaussian Mixture
  - W^3(Who, When, Where)
    - Probabilistic model
  - CEPR
    - Human mobility: regular and novel ones

实验结果



(a) different models (b) different variations of RCH

Acc@top*P*



(a) workdays (b) holidays

Acc@top*P* for different type of days and time slots

未来展望

- 数据
  - 跨平台用户数据链接
  - 用户数据与隐私保护的平衡
- 方法
  - 深度学习与知识图谱的应用
  - 可解释推荐系统
  - 与心理学、社会学、脑科学等领域的结合

结合跨平台异构数据的推荐系统

谢谢！