在线餐饮网站用户就餐点及口味预测

指导老师	傅晨波
姓　　名	范璐
班　　级	自动化1302
学　　院	信息工程学院

前情提要

一大堆分析：Yelp这个数据集问题真多啊

11个城市里——就在拉斯维加斯玩吧

按照时间先后划分了测试集训练集

~~总算可以开始搞事子~~

推荐引擎测试框架搭建和展示

Highlight

调研餐馆推荐学界历史发展

Yelp数据集的时空属性和口味属性分析，规则抽取

推荐引擎测试流程实现

R算法开发：内存优化的推荐算法实现

GeoC：提出推荐算法改进框架

GeoCUI：基于Shiny的推荐结果展示用户界面（本地/web）

KEYWORDS

一大坨数据报告

无监督学习 LBSNs Yelp 深夜报社美食部

推荐系统

时空数据

局部模型

Restaurant Recommendation

TIMELINE

1995-1999	第一支无线联网便携式手机StarTAC发布（翻盖，不是大哥大） 协同过滤已经广泛应用于电商网站如亚马逊之中 ~~餐馆推荐？不存在的。~~Schulman K等人讨论了餐馆领域知识结合协同过滤进行推荐的实现思路。
1999	DocoMo发布了第一支联网手机并迅速流行开来 Pazzani M J利用44份在线问卷，通过考察他们对58个餐馆界面打分研究了当时流行的推荐算法的应用和结合，包括基于内容，基于人口统计和协同过滤。
2001-2012	基于位置的服务手机应用雏形出现，研发者利用采集到的数据开始研究餐馆推荐，对情景（时间，空间）的利用尤为流行。
2013-	RecSys2013 Yelp数据竞赛首次公开大规模餐馆数据集，LBS/LBSNs持续吸引人们广泛关注，餐馆推荐迎来井喷，Yelp数据集开始频繁的单独或和Twitter，Foursquare等数据集应用在研究中。

总体思路

去哪儿吃>

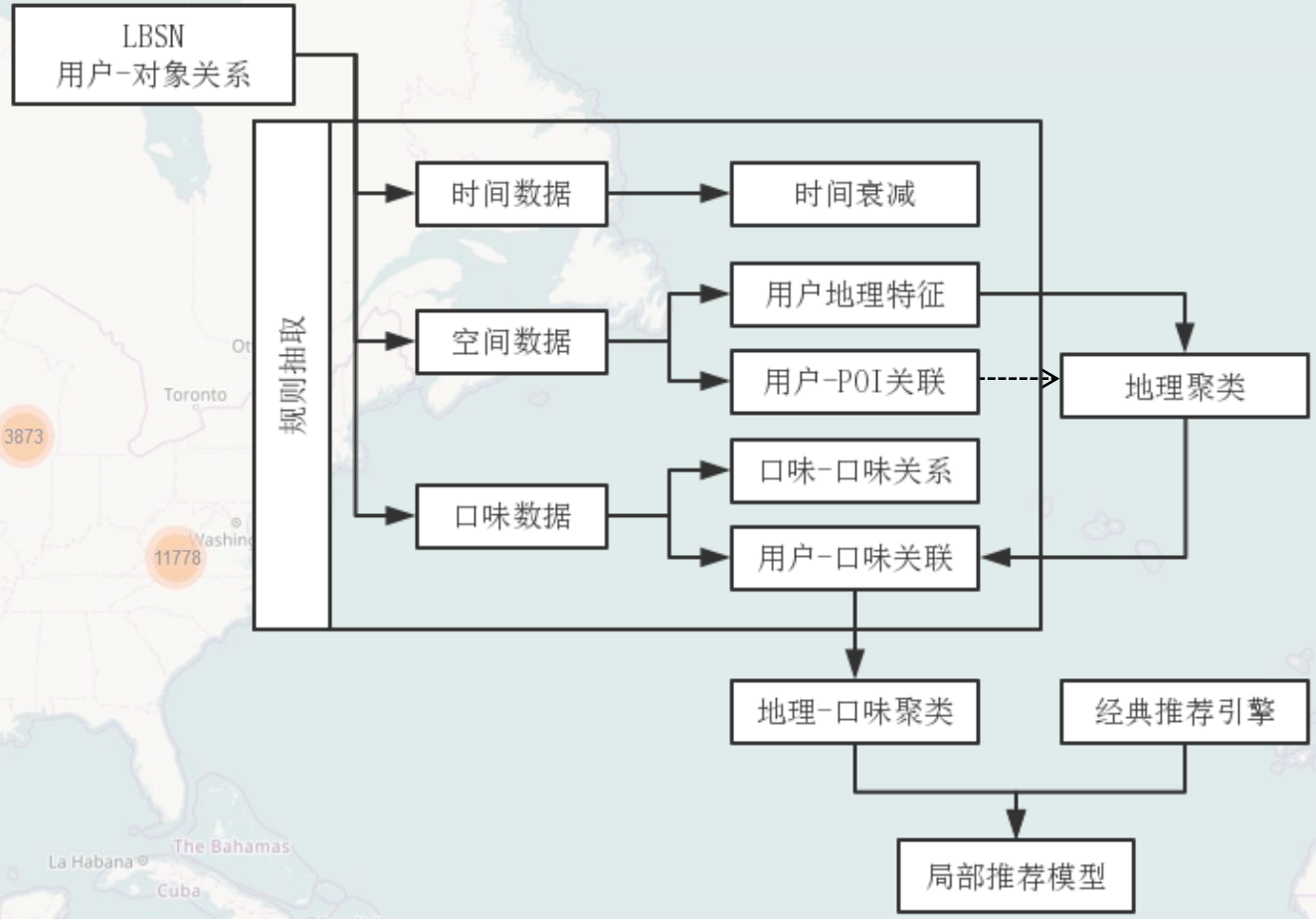
>吃什么?

时间、空间

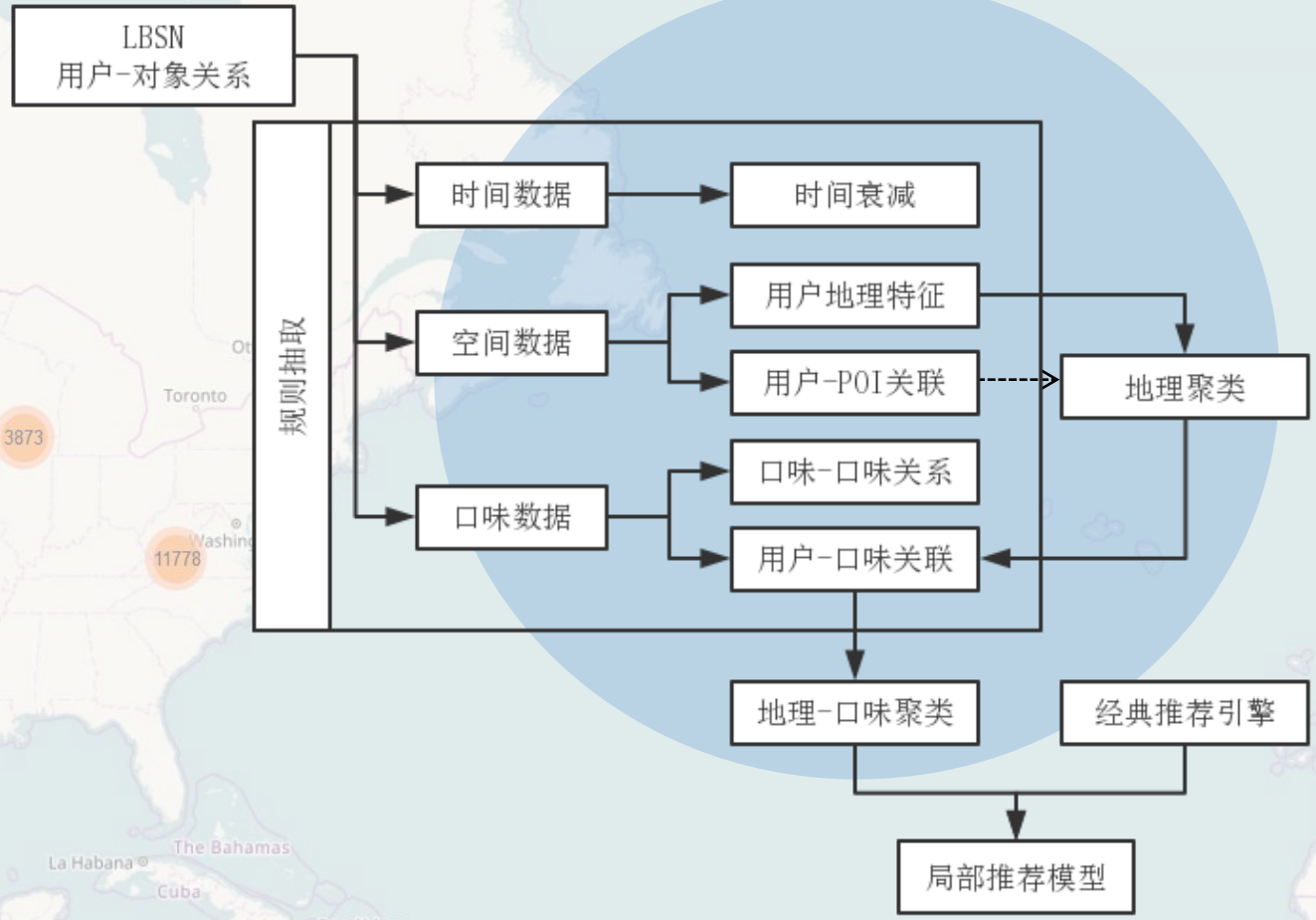
口味

GeoC

段首例行概括

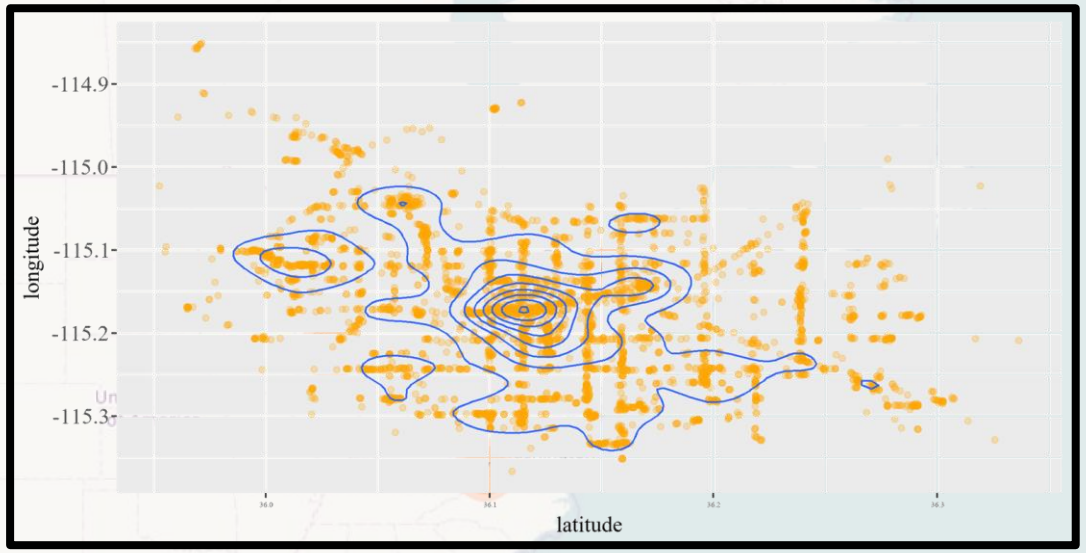

LBSN
用户-对象关系
规则抽取
时间数据
时间衰减
空间数据
用户地理特征
用户-POI关联
口味数据
口味-口味关系
用户-口味关联
地理聚类
地理-口味聚类
经典推荐引擎
局部推荐模型

规则抽取 用户聚类

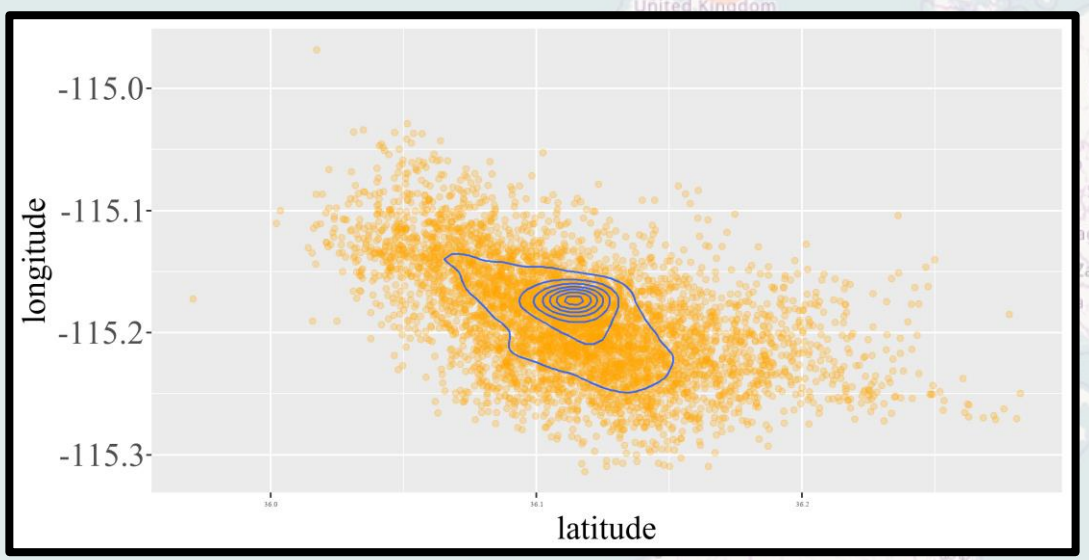

LBSN
用户-对象关系
规则抽取
时间数据
时间衰减
空间数据
用户地理特征
用户-POI关联
地理聚类
口味数据
口味-口味关系
用户-口味关联
地理-口味聚类
经典推荐引擎
局部推荐模型

用户空间特征

餐馆用户都在那儿



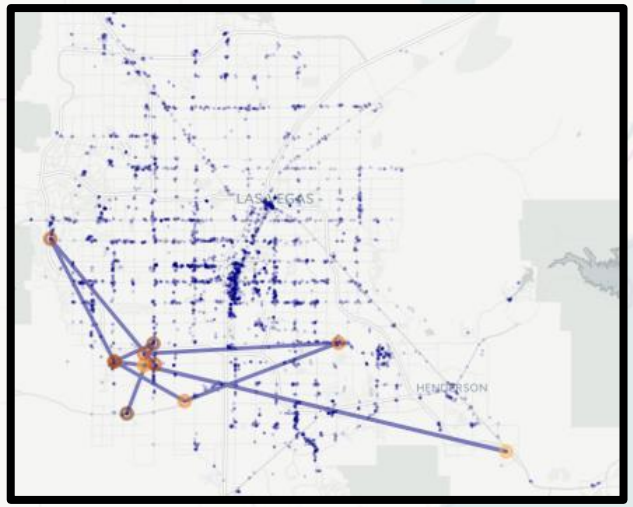
餐馆分布核密度估计



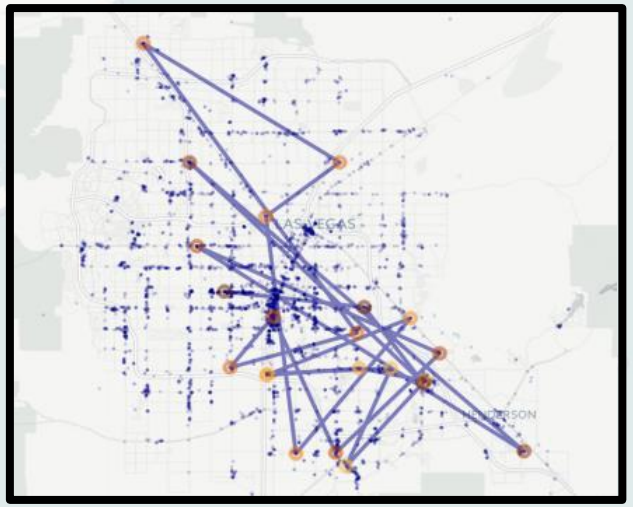
用户活动范围平均坐标分布

用户空间特征

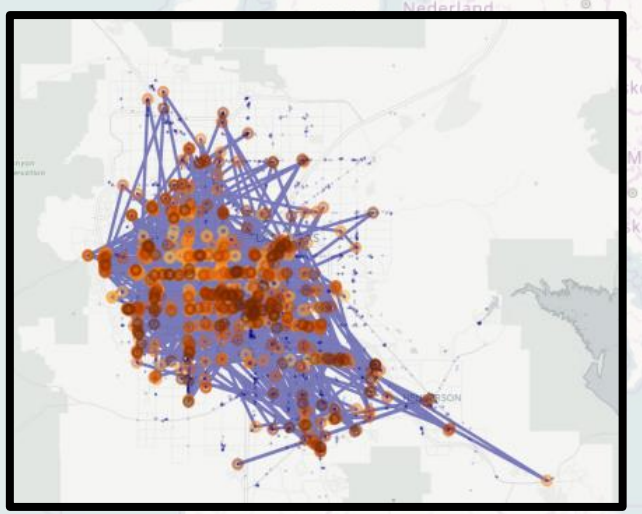
哪些用户爱乱跑



香农熵=3.070



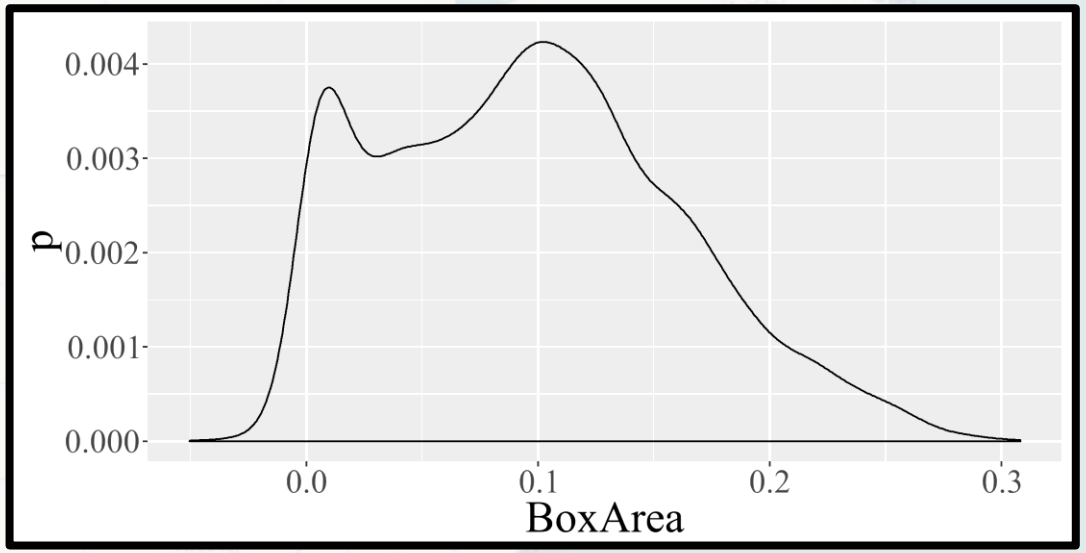
香农熵=3.954



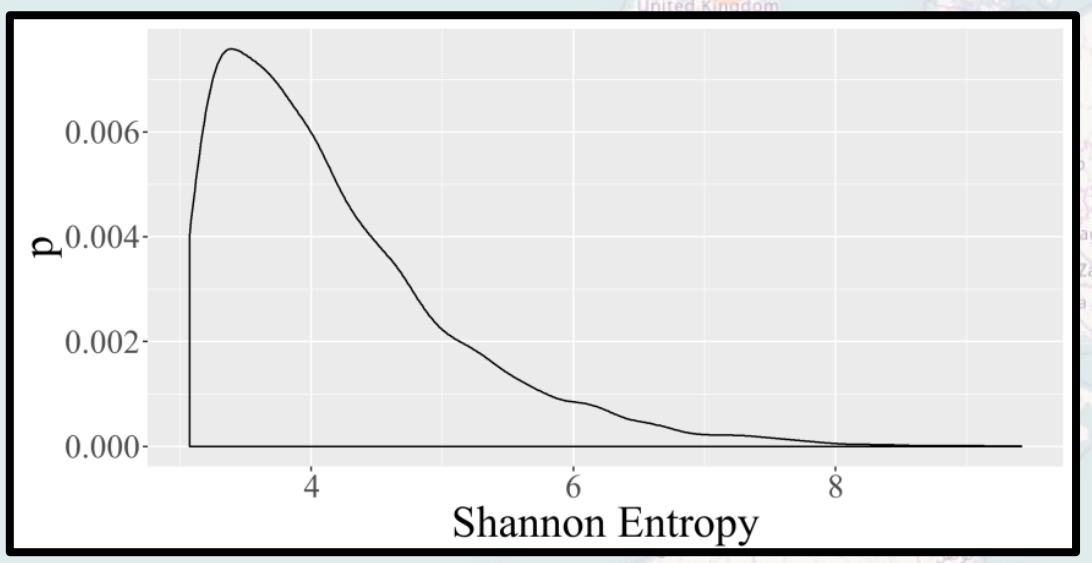
香农熵=9.416

用户空间特征

Overview



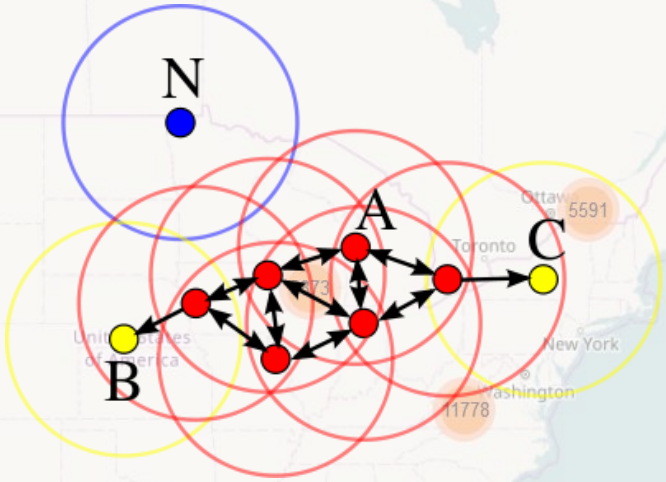
活动范围外接矩形面积/区域2外接矩形面积分布

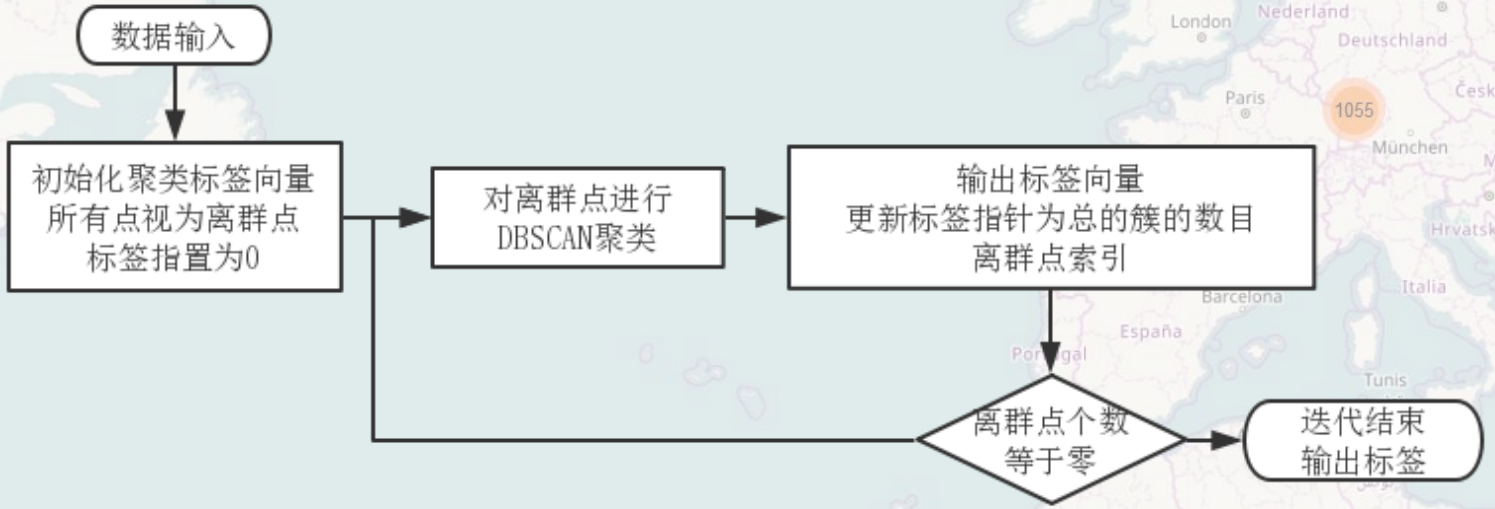


香农熵分布

用户空间特征

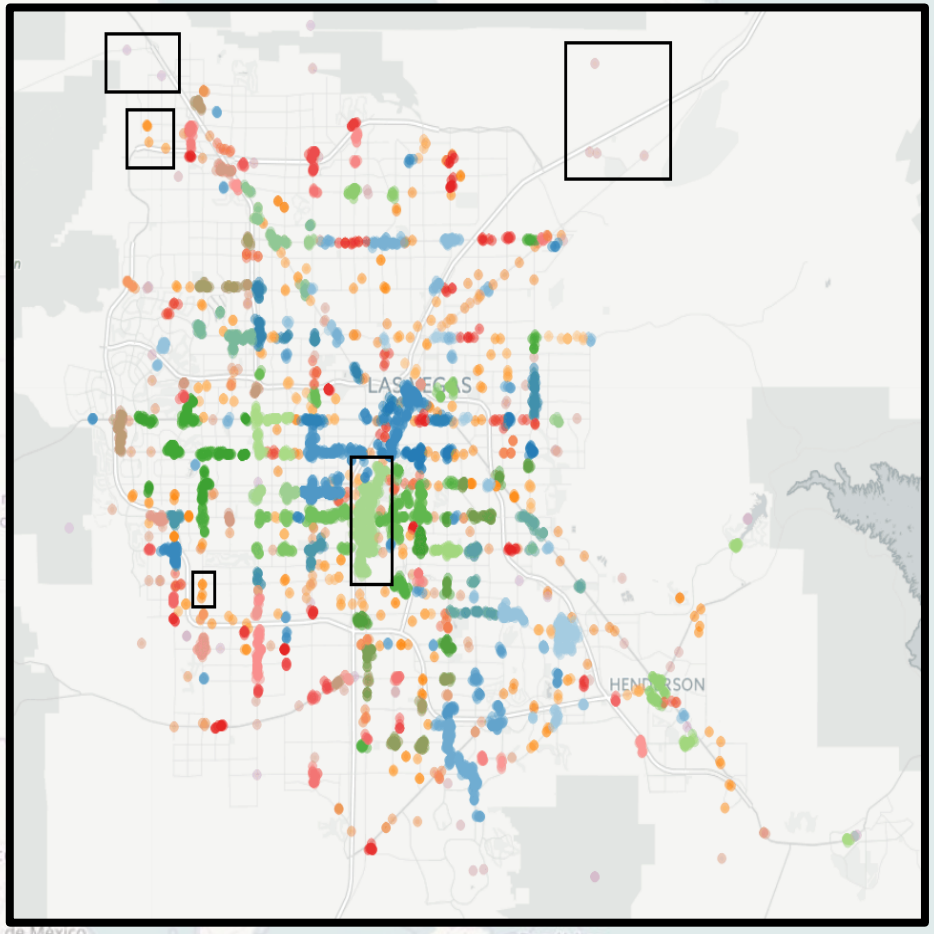
邻域半径扩展的DBSCAN





用户空间特征

邻域半径扩展的DBSCAN



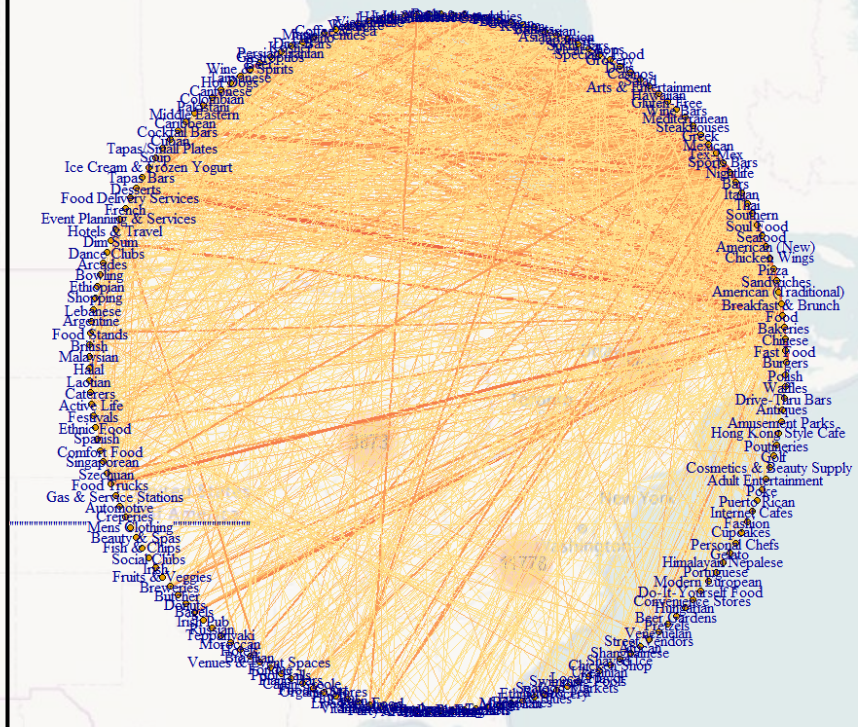
每次迭代，扩展邻域搜索半径，在上一轮聚类结果的离群点内再次聚类

好处：城区内餐馆分布稠密，抑制簇的生长
郊区餐馆分布稀疏，大范围也能聚类为一个簇

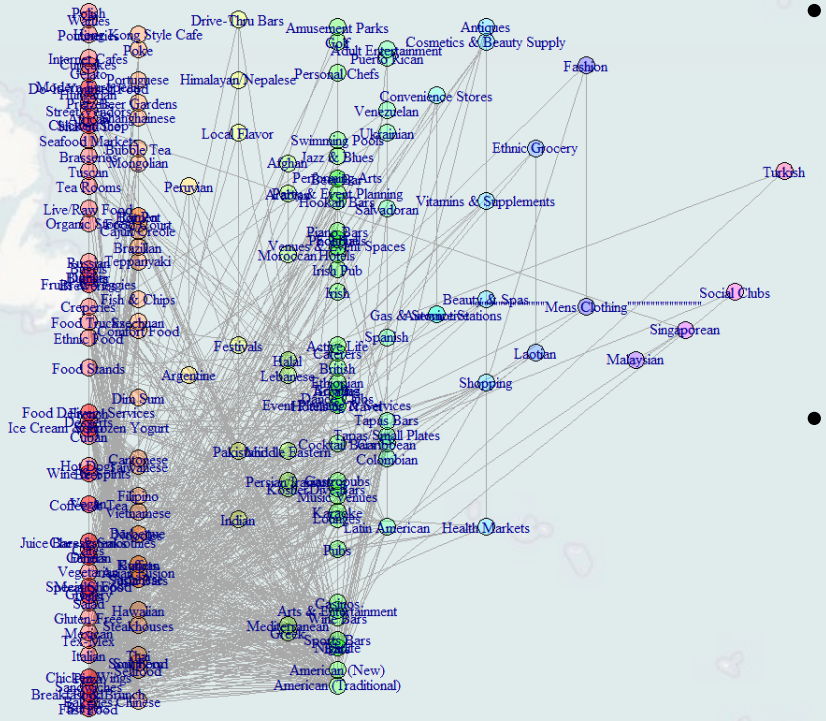
最后总共得到300个POI（商圈）

口味——口味关系

口味平滑策略



口味共现网络
（**Weight**=在餐馆中共同出现次数）



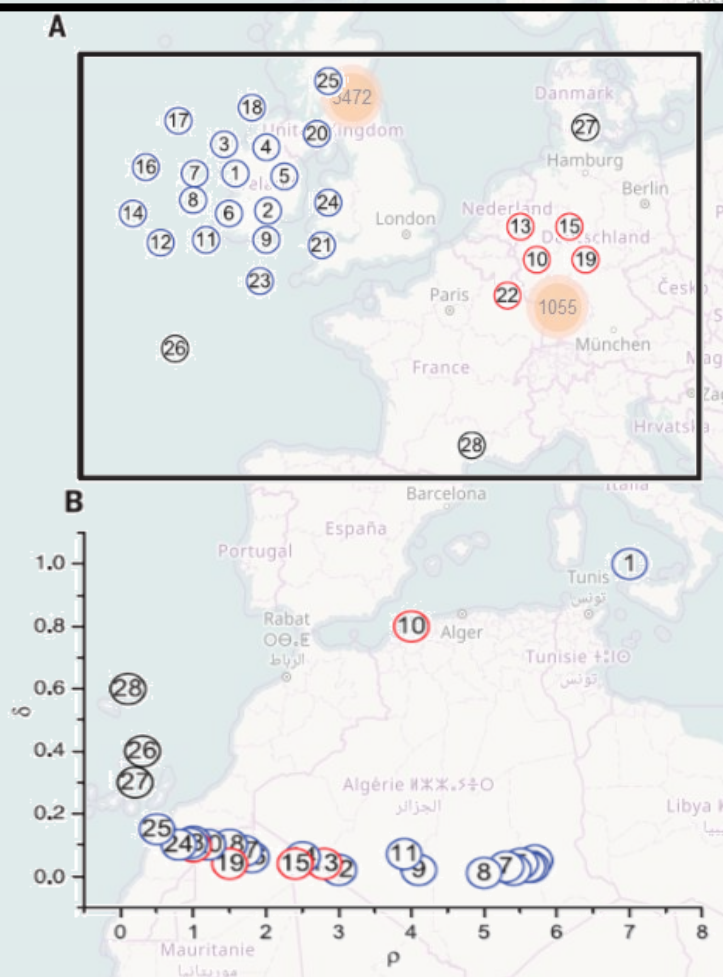
口味聚类结果

- 区域2餐馆中，总共204个口味标签，15类口味标签，可以看到餐馆，服装店，酒吧等标签被区分
- 考虑使用聚类内口味-口味的最短路径或wRA在用户口味向量中平滑，但考虑到用户口味向量和口味共现网络的提取方式存在重合，没有贸然引入

地理和口味聚类

聚类算法

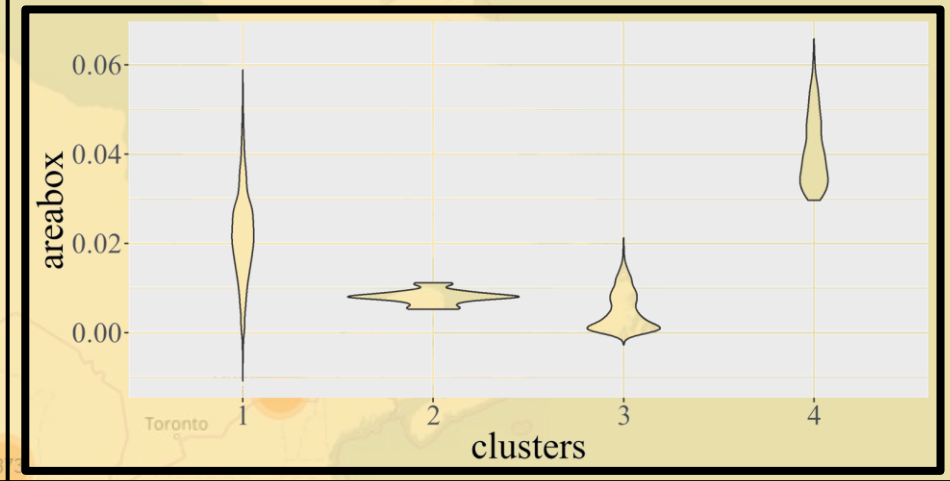
素材	• 地理聚类：用户-POI关联矩阵；用户活动范围外接矩形面积、平均坐标和香农熵特征向量 • 口味聚类：使用用户-口味关联矩阵
算法	• 14年Science上提出的一种新的聚类思路。这篇文章中的创新点在于提出了一个聚类中心刻画的新思路。文中将聚类描述为对聚类中心的寻找，聚类中心的密度应当是邻居中的峰值，而其他密度高的区域应该远离它。 • 本次研究中使用高斯核对每一个数据点求取局部密度，并求取每一个数据点和比它局部密度大的数据点的距离。通过寻找局部密度和“与其他局部密度比它大的点距离都大”的方式，找到聚类中心，进而通过近邻的方法得到簇。

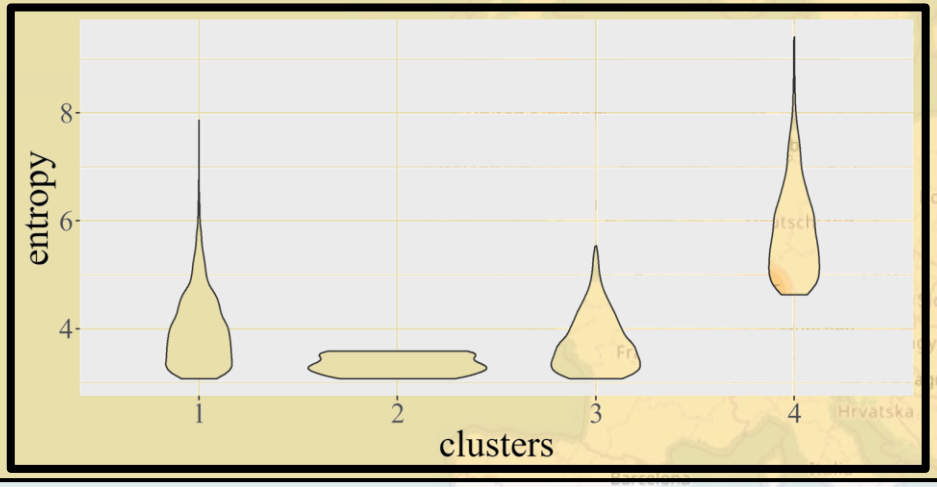


地理聚类结果对比

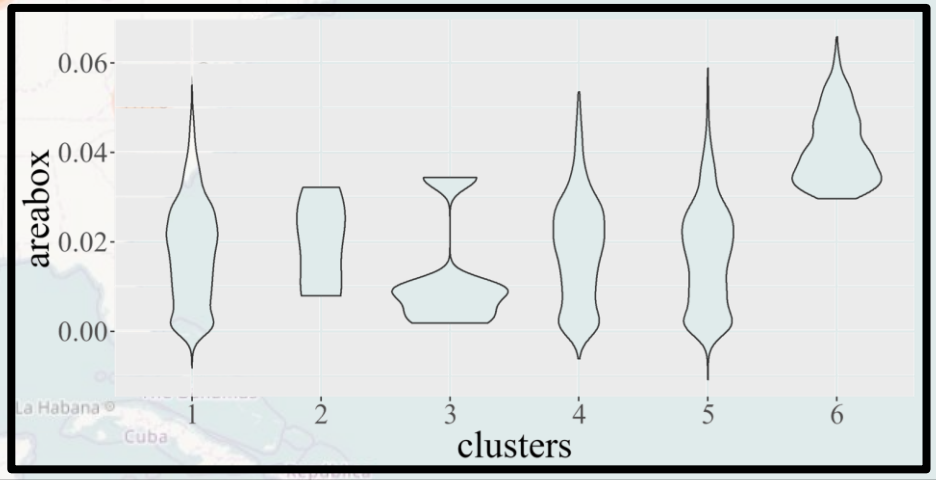
两家聚类谁更强

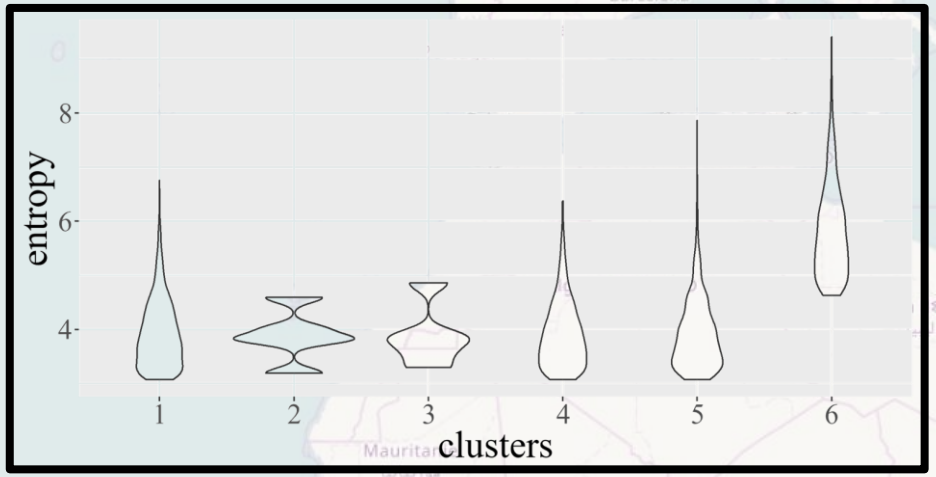
- 抑制活跃用户
- 地理特征
  - 熵
  - 范围大小
  - 平均坐标
- 谱聚类



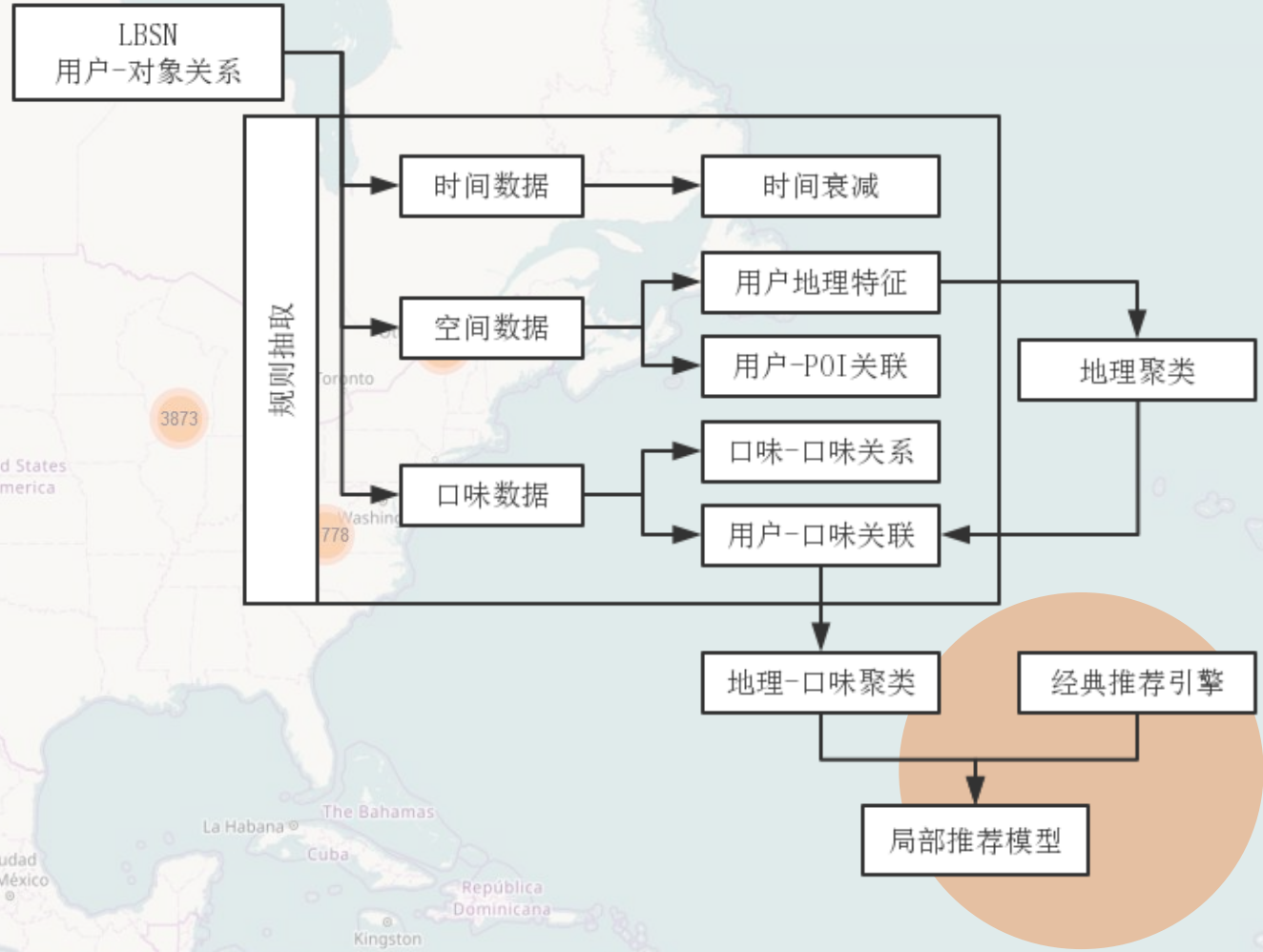


- 抑制活跃用户
- 用户-POI关系
- 谱聚类





推荐模型 局部模型



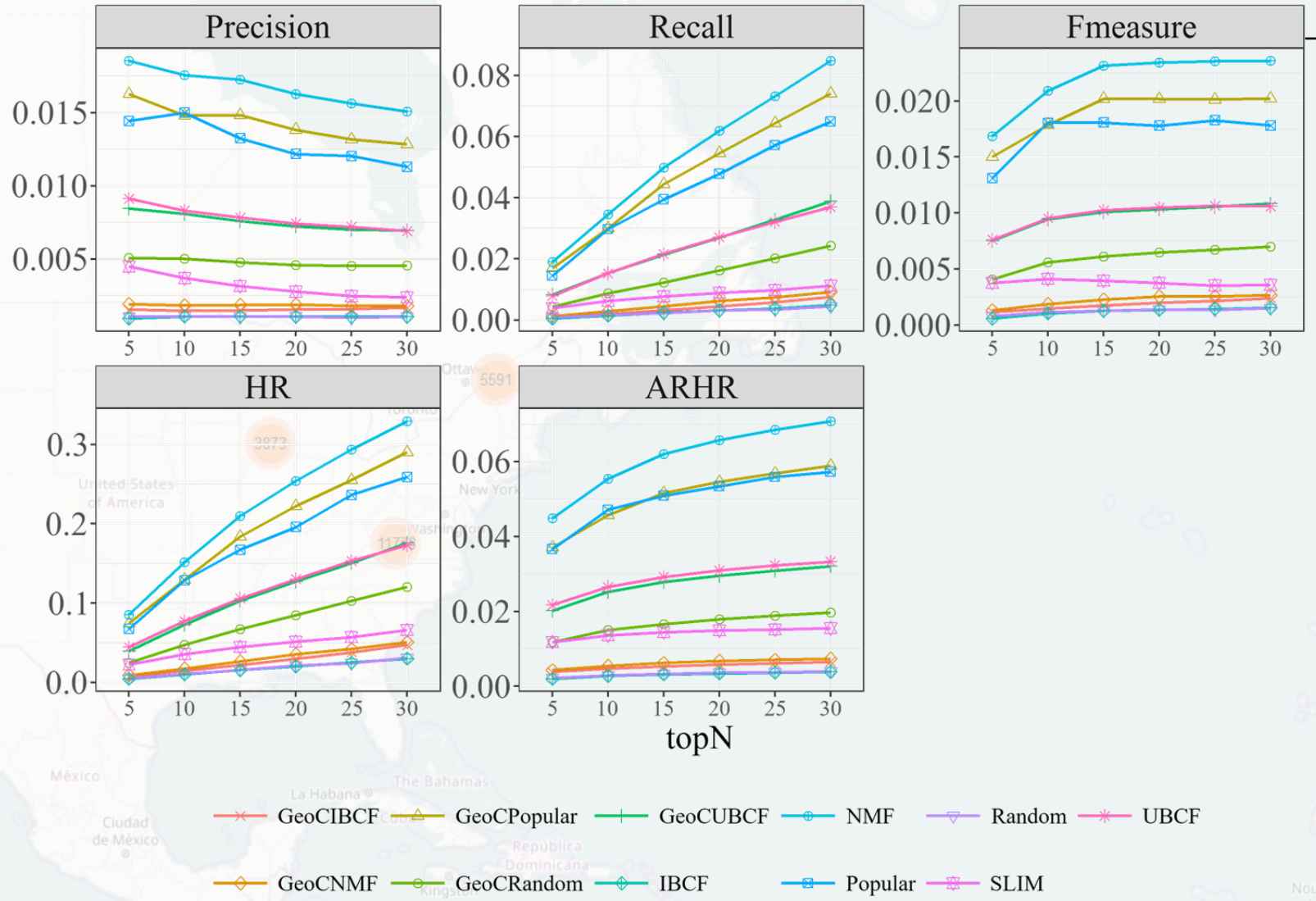
- 最终共得到3个地理簇，3个口味簇共计9个类簇
- 进入推荐环节

推荐引擎		
	Random	随机推荐 在用户-餐馆关联矩阵中，随机抽选餐馆推荐	• 冷启动 • 改善长尾效应 • 防止用户可达范围封闭
	Popular	流行度算法 在用户-餐馆关联矩阵中，推荐最热门的餐馆	• 古老，朴实有效 • 冷启动获取用户偏好 • 负样本的建立
	UBCF	基于用户的协同过滤 给用户推荐与用户相似的用户喜欢的餐馆	• 利用群体智慧 • 协同过滤算法的流行在一定程度上标志着推荐系统领域的诞生
	IBCF	基于物品的协同过滤 给用户推荐其有过行为的餐馆相似的餐馆	• 将机器学习，基于图的方法整合进推荐系统的常用框架

推荐引擎			
	NMF	• 非负矩阵分解 $W_{m\times n} = H_{m\times r} * V_{r\times n}^{T}$ • 将矩阵分解成两个低秩矩阵，在推荐系统的应用中，可视为用户和物品在隐空间的投影，每个用户和每个物品各自得到一个隐空间（**latent factor**）上的向量，则用户**i**对物品**j**的预测评分为两向量的内积 $\hat{r} = u_i \bullet i_j$	• **06**年**Netflix**大奖赛中获得成功以来一直广受关注，近年来最流行的推荐算法之一
	SLIM	• 稀疏线性方法 $\tilde{A}_{m\times n} = A_{m\times n} W_{n\times n}$ • 解决下式，带**l1**范数和**l2**范数规则项的，约束最优化问题求得矩阵**W**用户**i**物品**j**的预测评分为 $\tilde{a}_{ij}$ $\min_{w_i} \frac{1}{2}\|A - AW\|_F^2 + \frac{\beta}{2}\|W\|_F^2 + \lambda\frac{1}{2}\|W\|_1$ subject to $W \geq 0, \mathrm{diag}(W) = 0$	• **11**年提出，在保证精度的情况下推荐速度快（这里是指的离线训练，在线推荐速度快） • 近年来正得到越来越广泛的关注

评估指标	
准确率（Precision）	$\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}}$
召回率（Recall）	$\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}$
F-measure	$\frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$
HR（Hit Rate）	$\frac{\mathrm{tp}}{\mathrm{m}}$
ARHR（Average-Reciprocal Hit Rank）	$\frac{1}{\mathrm{m}}\sum_{i=1}^{tp}\frac{1}{p_i}$

评估结果



FACTS

- 最优模型NMF（对，很遗憾，GeoC*并没有拔得头筹）；
- 最差模型Random；
- GeoCRandom比Random各项指标提升了4-10倍；
- GeoCIBCF比IBCF各项指标提升约50%~100%；
- GeoCPopular和GeoCUBCF比原模型各项指标稍有提升；
- GeoCNMF各项指标反而下降了；
- SLIM算法跑崩了。

展望与发展 不足与局限

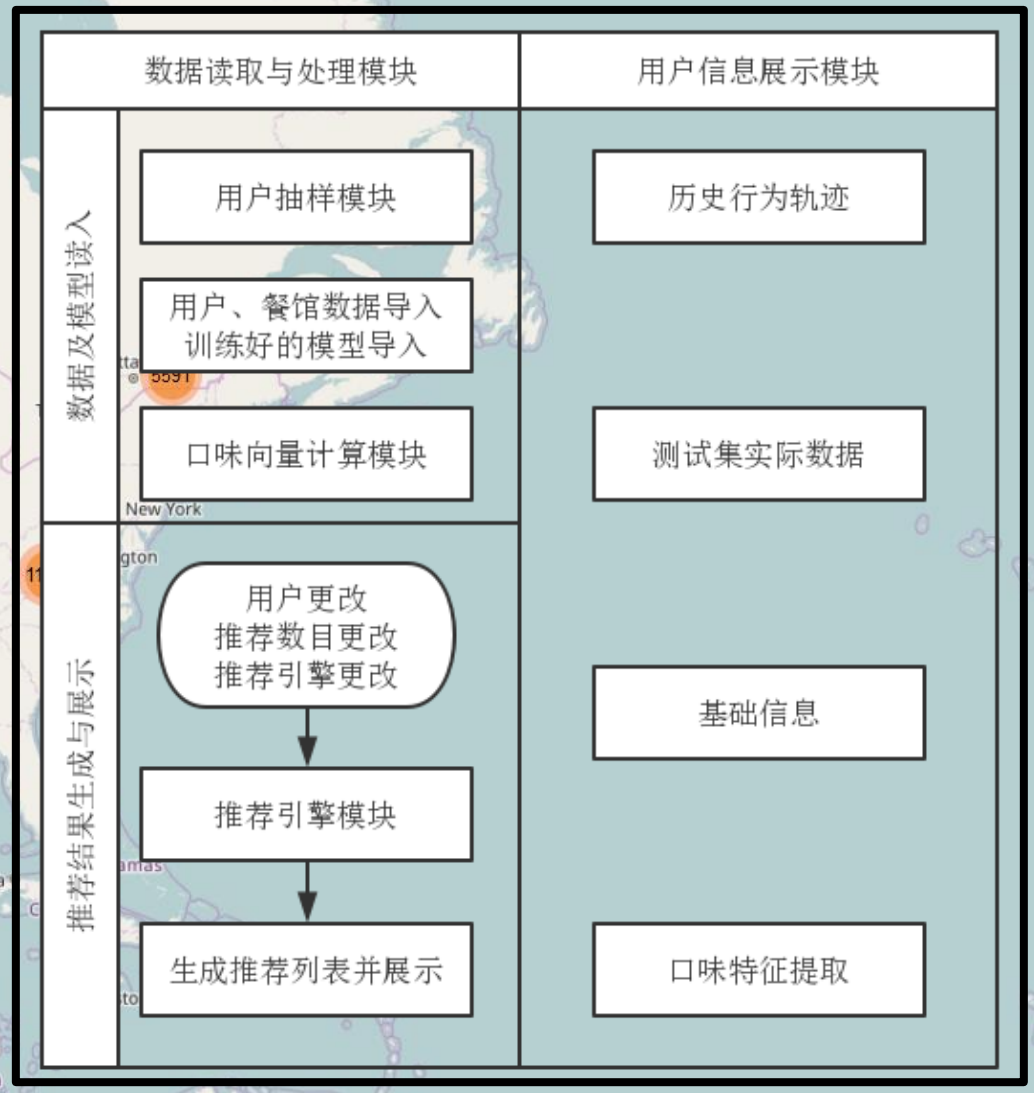
时间不足，基于模型的NMF和SLIM方法训练起来较慢，而且随机初始化方法导致模型最好训练多次，取最优值或平均值，但是时间不足，GeoCNMF中眼瞅着最大的一个簇没收敛，还是拿他做了结果，评估果然不好。另外，SLIM算法也是由于训练时间慢，调宽了收敛域（1%），这大概是SLIM跑崩的直接原因

没有将Yelp数据最具优势和特色的评论文本使用起来，而将评论记录作为评分数据使用，这对于LBSNs这个用户成本比较高的对象来说，缺失就比较大了

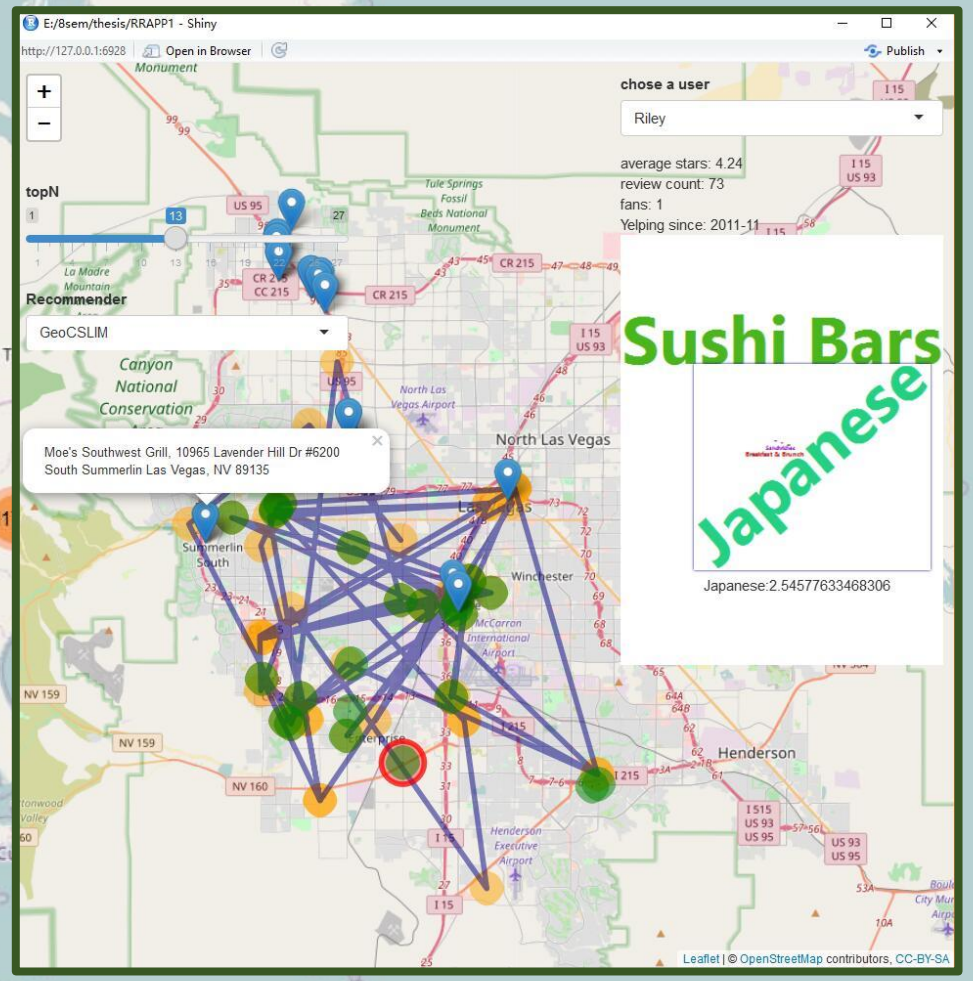
更充分的利用提取出的地理信息和口味信息，考虑POI-POI之间的距离和口味共现网络之间口味的距离。考虑引入不同口味在地理上的概率估计模型

加入情景感知（模拟时间和空间），放在手机应用场景之中

GeoCUI：功能设计

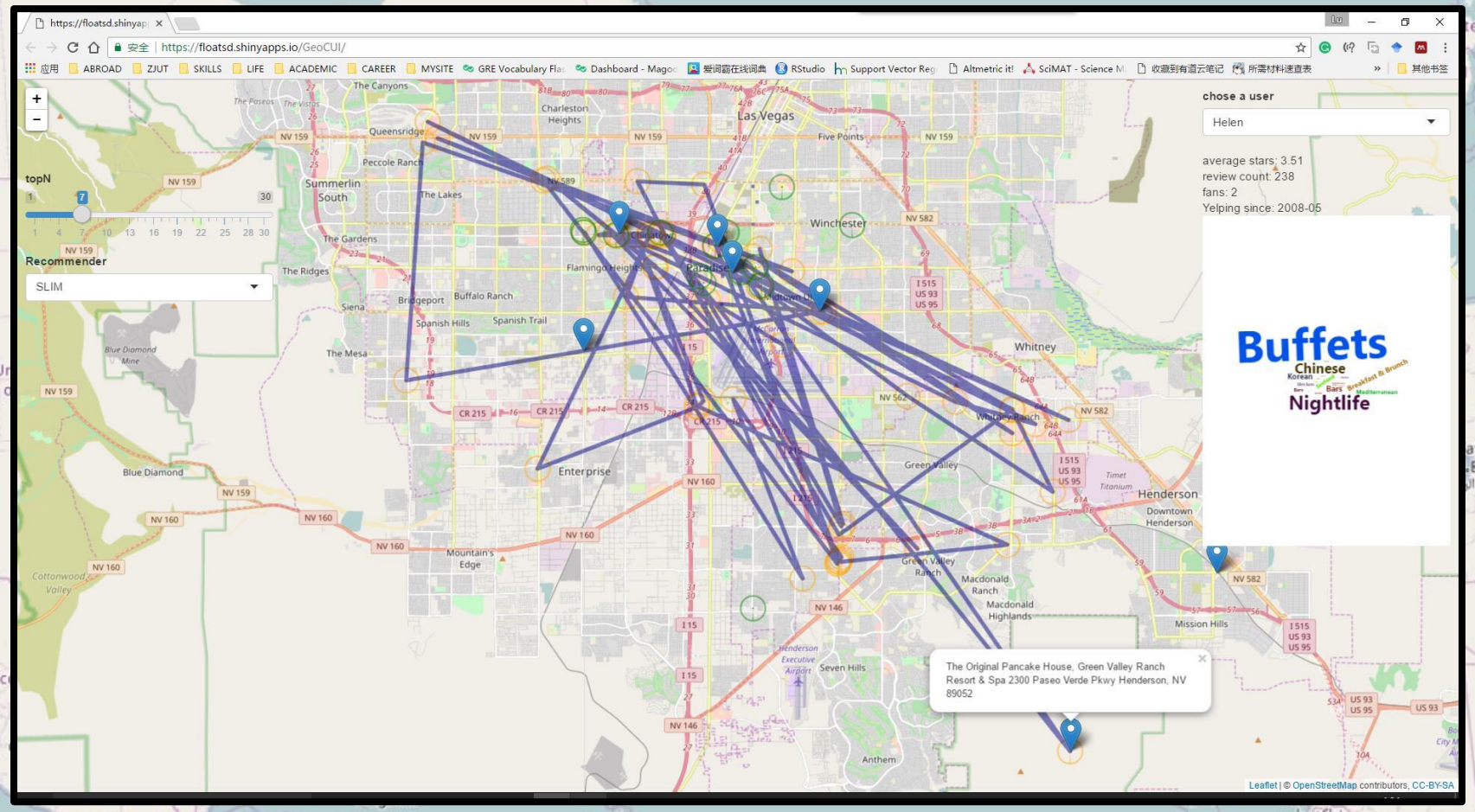

数据读取与处理模块
用户信息展示模块
数据及模型读入
用户抽样模块
用户、餐馆数据导入
训练好的模型导入
口味向量计算模块
历史行为轨迹
测试集实际数据
推荐结果生成与展示
用户更改
推荐数目更改
推荐引擎更改
推荐引擎模块
生成推荐列表并展示
基础信息
口味特征提取

GeoCUI: Local App



GeoCUI：Web App

http://floatsd.shinyapps.io/GeoCUI



致　谢

感谢一年来宣老师、傅老师对我的启迪、引导和提携，感谢明浩师兄、火火师姐、永立师兄、鸣鸣师兄、洋洋师兄、诗迪、心怡、大超、余斌、阿丁和老浩对我毕设及毕设阶段生活学习提供的支持和帮助。

最后，为这个美好的世界献上祝福