基于用户时空轨迹的社区挖掘、可视化与兴趣点预测

《社交网络挖掘》课程项目汇报

2024年1月5日

冯嘉旭 吴嘉骜 周志凌

FUDAN UNIVERSITY 1905 復旦大學 FUDAN UNIVERSITY

假如你是一个吃货...

热衷于在大众点评上探店打卡

但是...

找不到饭搭子！！！

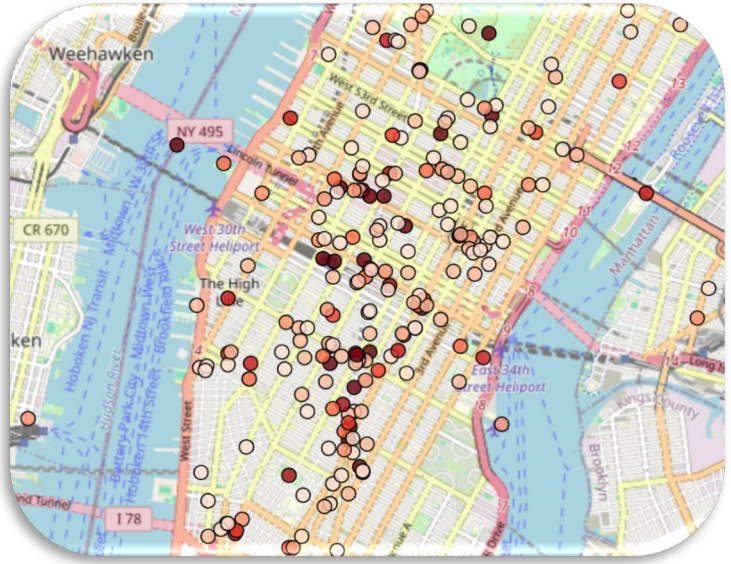
没关系！我们可以帮你

1. 找到吃货组织！

看看谁和你有相同的恰饭喜好！

2. 看看大家去哪吃！

在地图上直接看到大家在吃什么！



3. 猜你喜欢！

还有哪些宝藏饭店没被发掘！

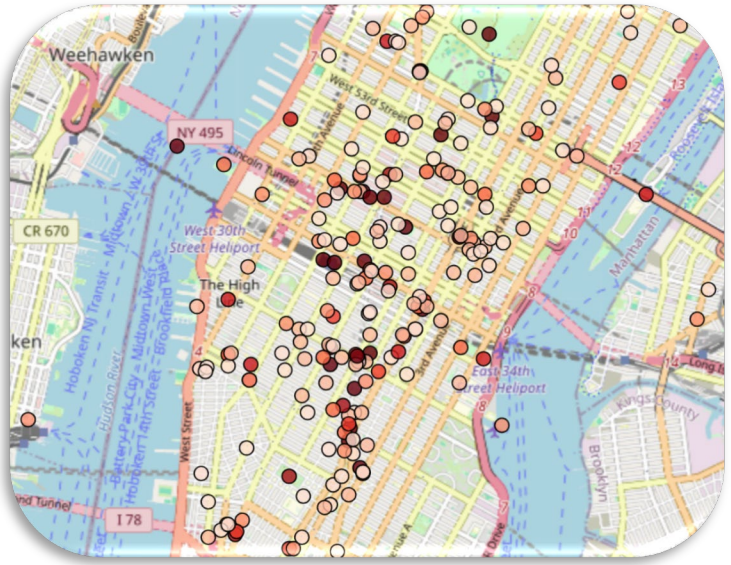
项目介绍

1.社交网络构建与社区挖掘

- 根据用户的兴趣点(POI)建立社交网络，并对用户进行社团划分

2.社区与用户轨迹可视化

- 可视化社团划分结果
- 可视化用户时空轨迹



3.兴趣点预测与推荐

- 根据用户历史上的POI，设计模型预测与推荐用户新的POI

数据集

- **介绍**
  - 227,428条纽约市的打卡记录（来自Kaggle）
  - 12 April 2012 至 16 February 2013
- **内容**
  - 用户ID（1083个用户）
  - 地点ID（34639个地点）
  - 地点类别ID（384个类别）
  - 地点类别名称（中餐馆、健身房、办公室等）
  - 经度、纬度
  - 打卡时间与UTC时间之差（夏令时、冬令时不同）
  - UTC时间

目录

1.1 社交网络构建算法：思想与框架

- **算法思想**
  - 构建一个带权图，每个用户都是图中的一个节点
  - 两个用户之间边的权重通过共同访问地点计算得出
    - 如果两个用户去的共同地点越多，权重越大
    - 如果两个用户去同一地点的时间越近，权重越大
- **算法框架**
  - 对于每一个用户对，搜索其共同去过的地点
    - 对于每个共同去过的地点，计算子权重
  - 对所有子权重求和得到用户对的边权重

1.1 社交网络构建算法：权重计算

- 对于每个用户对共同去过的地点，计算子权重
  - 计算用户A和用户B去该地点的次数：AFreq和BFreq
  - 计算两个用户在1小时内访问该地点的次数visitWithinHour，以及在1天内（1小时外）访问该地点的次数visitWithinDay
  - AFreq ← AFreq + 2 * visitWithinHour + visitWithinDay
  - BFreq ← BFreq + 2 * visitWithinHour + visitWithinDay
  - **子权重← (AFreq + BFreq) / abs(AFreq - BFreq) / 该地点在数据集中被访问的人次**
- 对所有子权重求和得到边权重
- 如果权重大于1，则在两节点之间连带权重的边

1.1 社交网络构建算法：子权重解释

子权重← (AFreq + BFreq) / abs(AFreq - BFreq) / 该地点在数据集中被访问的人次

案例编号	AFreq	BFreq	地点访问人次	子权重
1	2	3	10	0.5
2	9	10	10	1.9
3	2	10	10	0.15
4	9	10	1000	0.019

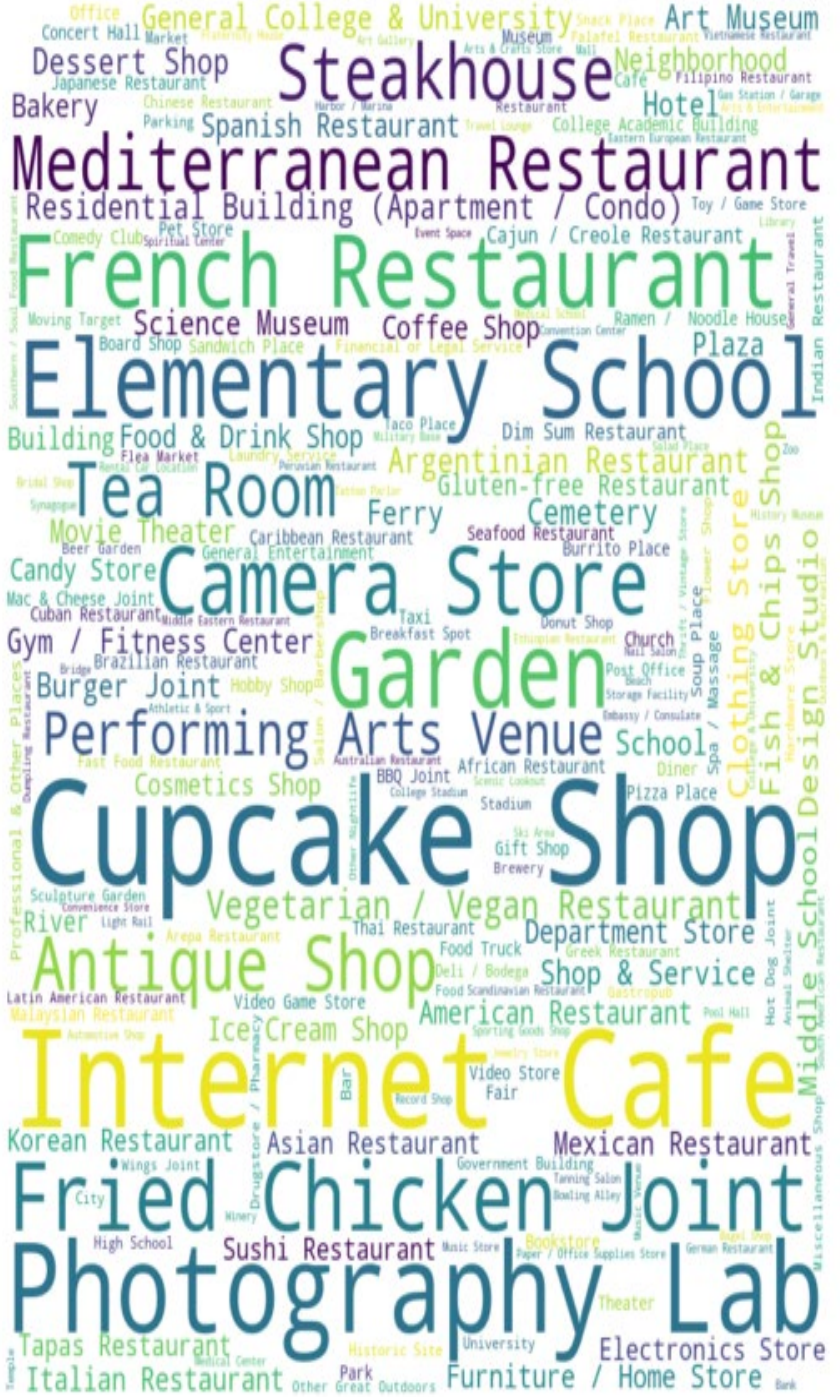
- **案例1与案例2**：两用户去某地点次数越频繁，子权重越大，解释为何将AFreq和BFreq相加
- **案例1与案例3**：一个用户经常去，另一个不经常去，两个用户不相似，解释为何要除以abs(AFreq - BFreq)
- **案例2与案例4**：两个用户都经常去，但是所有用户都经常去（比如地铁、机场），无法代表两个用户之间的特征

1.2 社区挖掘算法：Louvain算法

- **模块度** $Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$
  - 对于带权图，其中节点i和节点j之间的边的权重表示为A_{ij}， A 为邻接矩阵。节点i的度k_i是其所有邻接边的权重之和，m是网络中所有边权重之和
- **算法步骤**
  1. **初始化**：将每个节点视为一个独立的社区
  2. **迭代优化**：对每个节点，遍历其邻居节点，计算将该节点移动到邻居节点所在社区时的模块度增益。选择模块度增益最大的移动，并将节点移动到对应的邻居节点所在社区
  3. **合并社区**：将所有节点移动完毕后，将每个社区合并为一个超级节点，构建超级节点之间的新图。重复步骤2，直到没有进一步的模块度增益为止

1.3 社区特征分析：词云图算法

- 根据社区内用户的打卡信息可以提取最能概括该社团的语义标签
- 词云图中各个标签（地点类别）的权重计算可以分为两步
  - 标签特异度 = $\frac{\text{社区中该地点类别被访问的人次}}{\text{数据集中该地点类别被访问的人次}}$
  - 标签代表性 = $(\frac{\text{社区中访问过该地点类别的人数}}{\text{社区总人数}})^{\alpha}$，$\alpha$为常数
  - 标签权重 = 标签特异度 × 标签代表性
  - **从TF-IDF获得启发**



吃货一族


Arepa Restaurant
Art Museum
Japanese Restaurant
Deli / Bodega
Middle Eastern Restaurant
French Restaurant
Fried Chicken Joint
Gym / Fitness Center
Malaysian Restaurant
Tapas Restaurant
Burger Joint
Clothing Store
Vietnamese Restaurant
Turkish Restaurant
Salad Place
Indian Restaurant
Office
Breakfast Spot
Recycling Facility
Ferry
Spa / Massage
Korean Restaurant
Coffee Shop
Italian Restaurant
Café
Pizza Place
American Restaurant
Cosmetics Shop
Swiss Restaurant
Asian Restaurant
Falafel Restaurant
Dessert Shop
Seafood Restaurant
Greek Restaurant
Gastropub
Beer Garden
Thai Restaurant
Sandwich Place
Eastern European Restaurant
Scandinavian Restaurant
Dim Sum Restaurant
Ramen / Noodle House
Australian Restaurant
Bar
Mediterranean Restaurant
Steakhouse
Ice Cream Shop
Snack Place
Mac & Cheese Joint
Nail Salon
BBQ Joint
Cuban Restaurant
Dumpling Restaurant
Castle
Food Truck
Mexican Restaurant
Sushi Restaurant
Bakery
Hobby Shop
Tea Room

剧迷戏粉


Theater
College Theater
Performing Arts Venue
General Entertainment
Movie Theater
Concert Hall
Music Venue
Museum
Afghan Restaurant
African Restaurant
Bookstore
Bar
Café
Eastern European Restaurant
Other Great Outdoors
Vegetarian / Vegan Restaurant
Medical School
Moroccan Restaurant
Other Nightlife
Sandwich Place
Burger Joint
Dumpling Restaurant
Thai Restaurant
Greek Restaurant
Office
Department Store
Clothing Store
Embassy / Consulate
Government Building
Scenic Lookout
Breakfast Spot
Cupcake Shop
Pool
Food Truck
Indian Restaurant
General College & University
Southern / Soul Food Restaurant
Fish & Chips Shop
Bowling Alley
Pizza Place
Miscellaneous Shop
Pet Store
Flower Shop
Furniture / Home Store
Ice Cream Shop
Mexican Restaurant
College Academic Building
Gym / Fitness Center
Paper / Office Supplies Store
Thrift / Vintage Store
Gift Shop
Seafood Restaurant
South American Restaurant
Rental Car Location
Synagogue
University
Diner
Salon / Barbershop
Coffee Shop
Food & Drink Shop
Cuban Restaurant
Burrito Place
American Restaurant
Restaurant
Sculpture Garden
Bakery
Sushi Restaurant
Art Museum
Mall
Plaza
Historic Site
Fried Chicken Joint
Italian Restaurant
French Restaurant
Music Store
Post Office
Park
Medical Center
Camera Store
Laundry Service
Housing Development
BBQ Joint
Middle Eastern Restaurant
High School
Drugstore / Pharmacy
Video Store
Market
Cosmetics Shop
Spiritual Center
Flea Market
Harbor / Marina
Falafel Restaurant
Outdoors & Recreation
Beach
Department Store
Steakhouse
Winery
Ramen / Noodle House
Public Art
Parking
Playground
Ferry
Neighborhood
Latin American Restaurant
Dessert Shop
Tapas Restaurant
Church
Bridge
Japanese Restaurant
Comedy Club
Peruvian Restaurant
Stadium
Record Shop
Deli / Bodega
Garden
Cemetery
Korean Restaurant
Building
Spa / Massage
Hotel
Arcade
Beer Garden
Tea Room
Chinese Restaurant
Financial or Legal Service
Nail Salon
Taxi
Athletic & Sport
Professional & Other Places
Fraternity House
Moving Target
Library
Toy / Game Store
Brewery
Design Studio
Arepa Restaurant
Bagel Shop
Fast Food Restaurant
Bank
Pool Hall
Bridal Shop
Taco Place
Art Gallery
Convention Center
Filipino Restaurant
General Travel
History Museum
Soup Place
Hardware Store
Argentinian Restaurant
Event Space
Tattoo Parlor
Vietnamese Restaurant
Shop
Hobby
School
Middle School
Spanish Restaurant
Salad Place
Caribbean Restaurant
Casino
Factory
Arts & Crafts Store
Candy Store
Gastropub
Hot Dog Joint
Sporting Goods Shop
Australian Restaurant
Lounge
Auto Dealership
Automotive Shop
Arts & Entertainment
Law School
Student Center

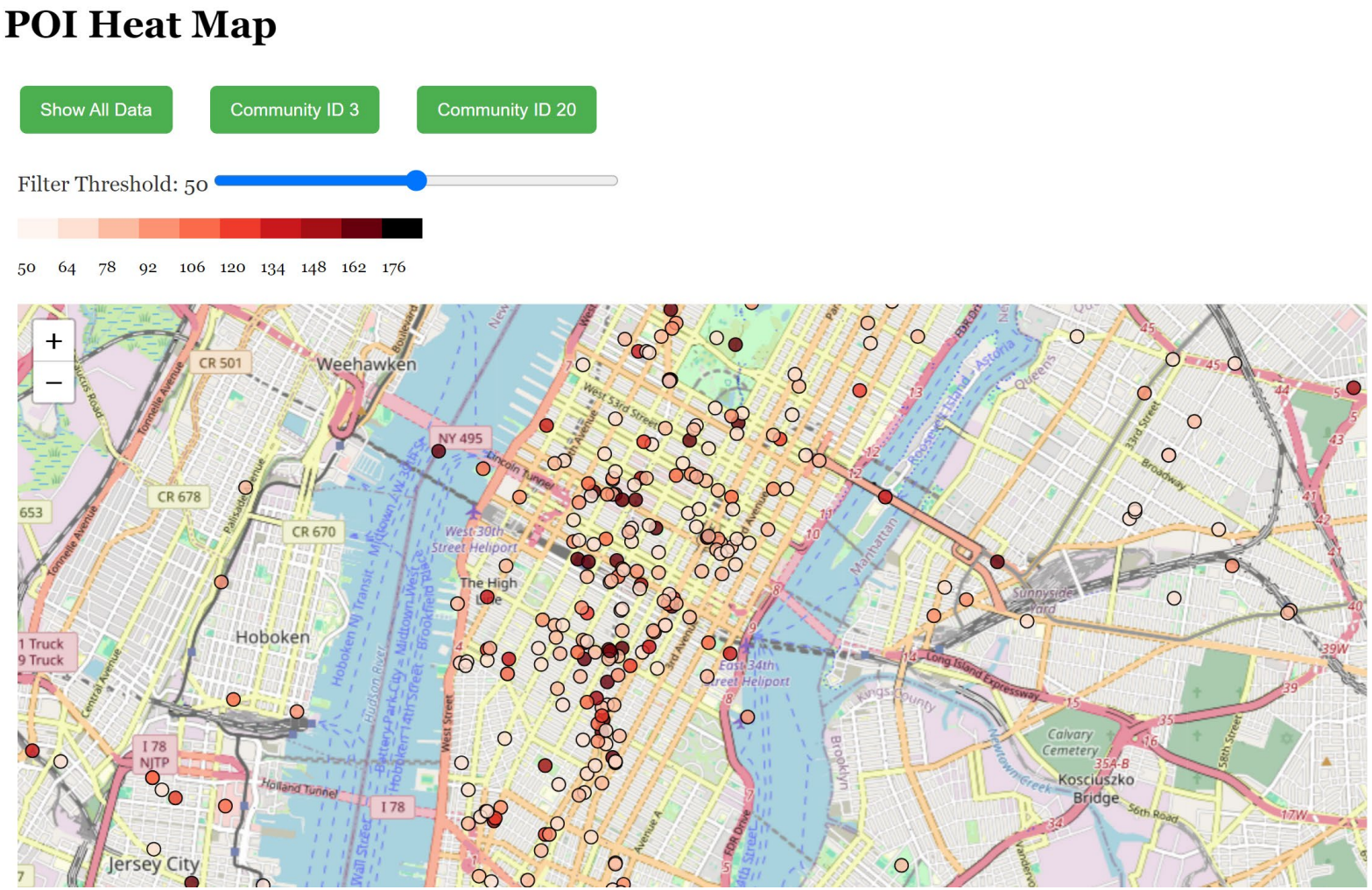
学生群体


Laundry Service
Salad Place
Jewelry Store
Event Space
African Restaurant
BBQ Joint
Clothing Store
College Academic Building
College Theater
Art Gallery
Embassy / Consulate
Burrito Place
Italian Restaurant
Concert Hall
Music School
Other Great Outdoors
Sculpture Garden
Thai Restaurant
Storage Facility
Performing Arts Venue
Chinese Restaurant
Donut Shop
Caribbean Restaurant
Park
Gym / Fitness Center
General Travel
Outdoors & Recreation
Church
Bar
Mediterranean Restaurant
Law School
Bakery
Beer Garden
Financial or Legal Service
Bank
Burger Joint
Latin American Restaurant
Taxi
Coffee Shop
Public Art
Hot Dog Joint
Plaza
Office
Post Office
American Restaurant
Design Studio
Medical Center
Hardware Store
Theater
Fast Food Restaurant
Museum
Aquarium
Sushi Restaurant
Movie Theater
Bowling Alley
Deli / Bodega
Scenic Lookout
Southern / Soul Food Restaurant
Dumpling Restaurant
Seafood Restaurant
Building
Harbor / Marina
Pool Hall
Restaurant
Neighborhood
Department Store

目录

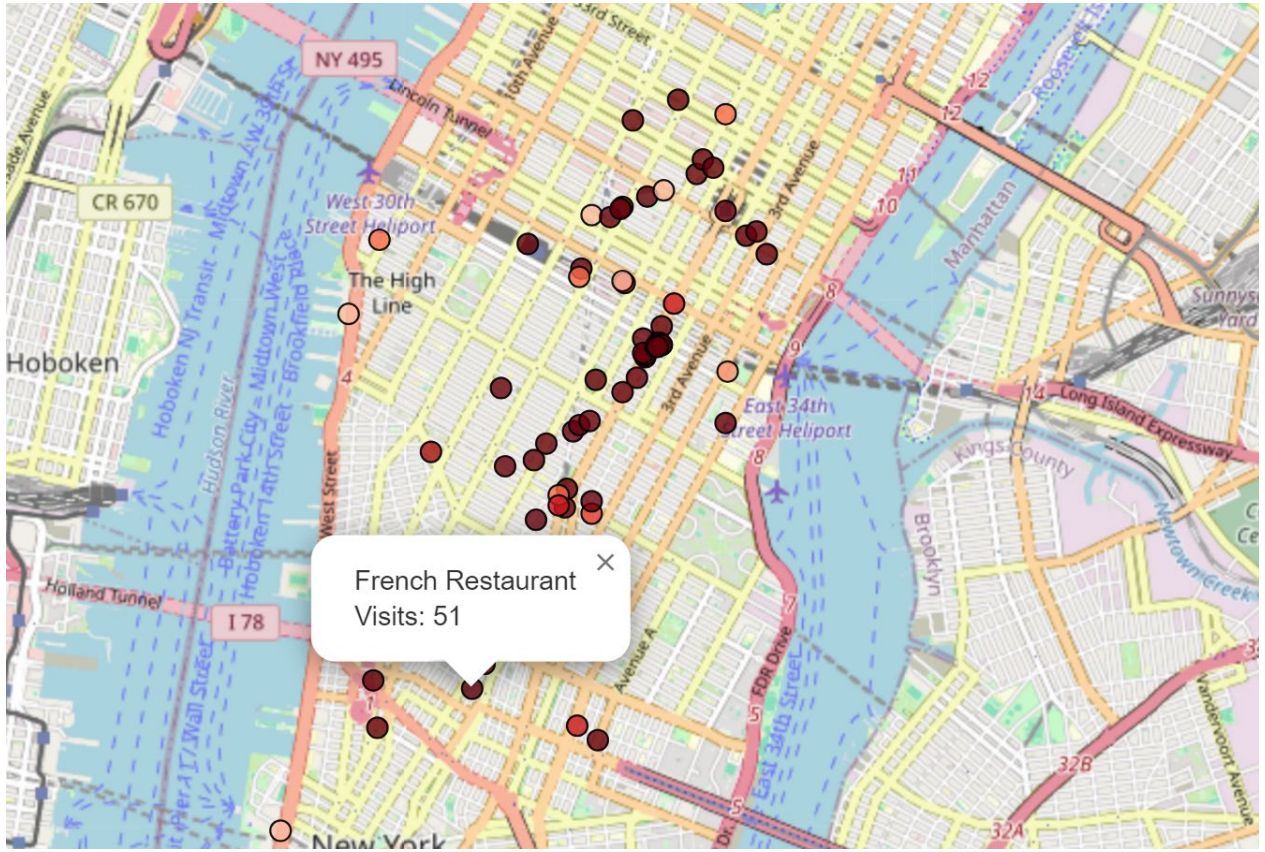
2.1 地图可视化：访问人次热力图

- 考虑所有用户的访问地点，其中深色的地点访问人次更多，分析为地铁、公交等公共设施
- 为了防止过于密集，可以滑动调整地点频数阈值



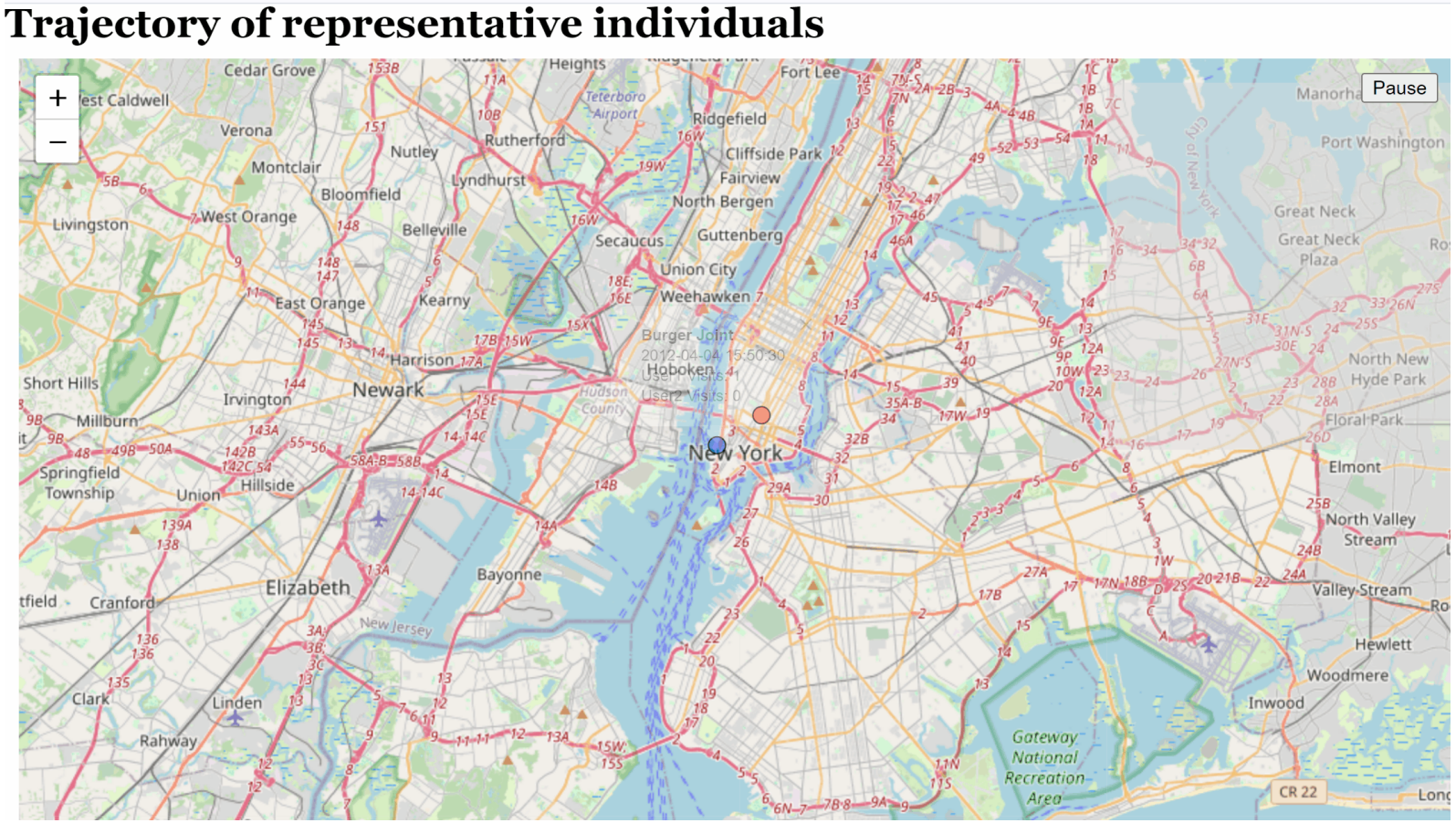
2.2 社区可视化：社区内用户的访问人次热力图

- 考虑吃货社区内用户的访问地点，访问人次最多的地点为各种饭店，与前面词云图分析相符



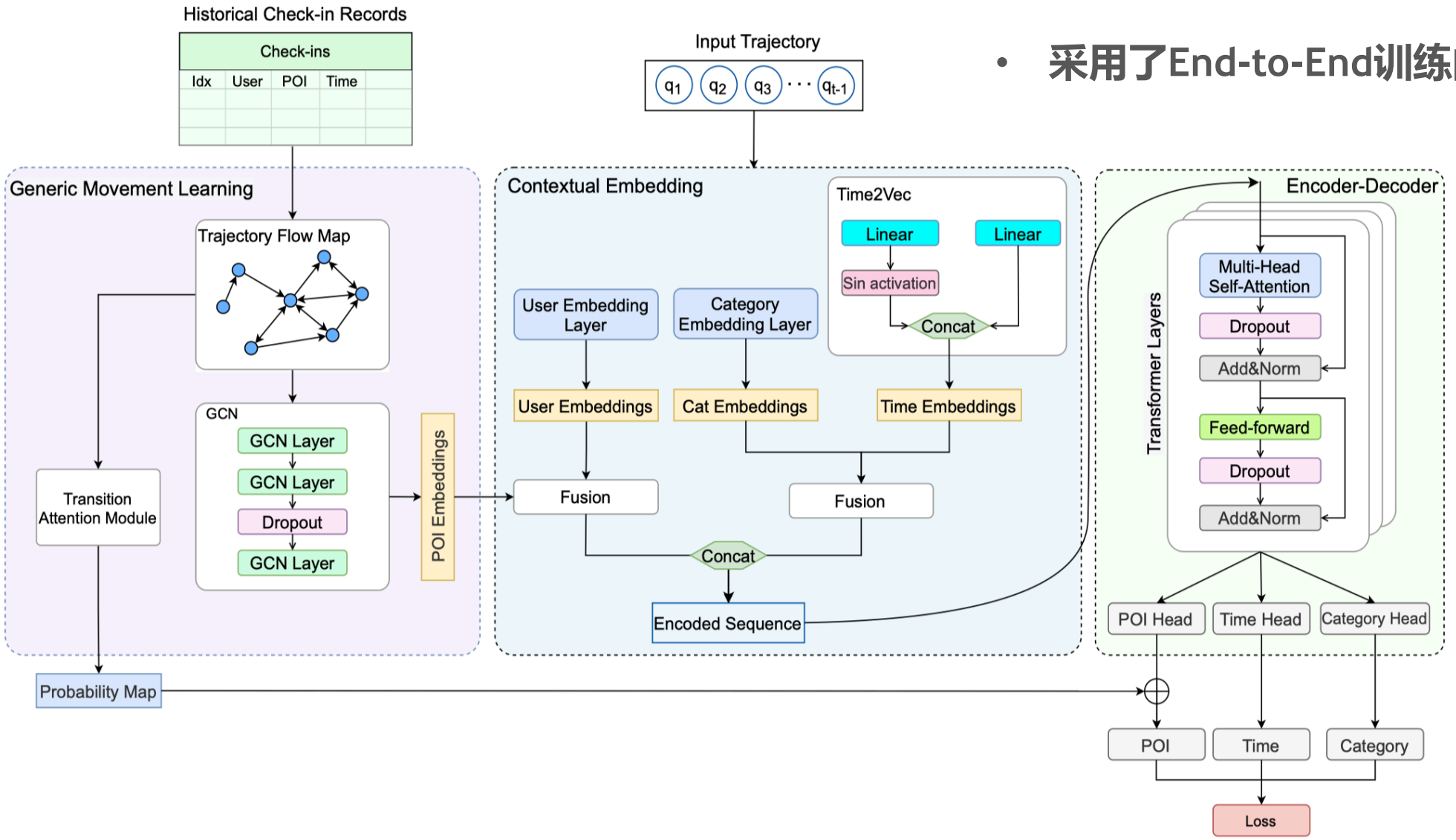
2.3 用户轨迹可视化：见证相识相知相爱

- 两个用户开始轨迹没有什么重叠，某一时刻（相识）之后就轨迹经常重叠（相知），通过可视化发现了动人的友谊（浪漫）故事



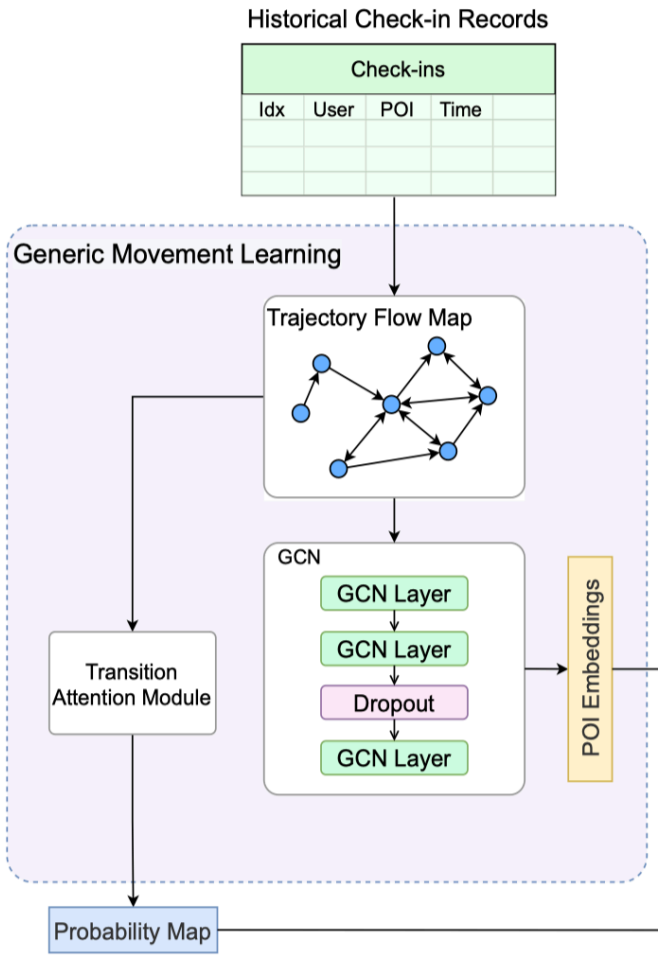
目录

3.1 模型基础：Vanilla GETNext Model



- 采用了End-to-End训练的方式

3.1 模型基础：Vanilla GETNext Model



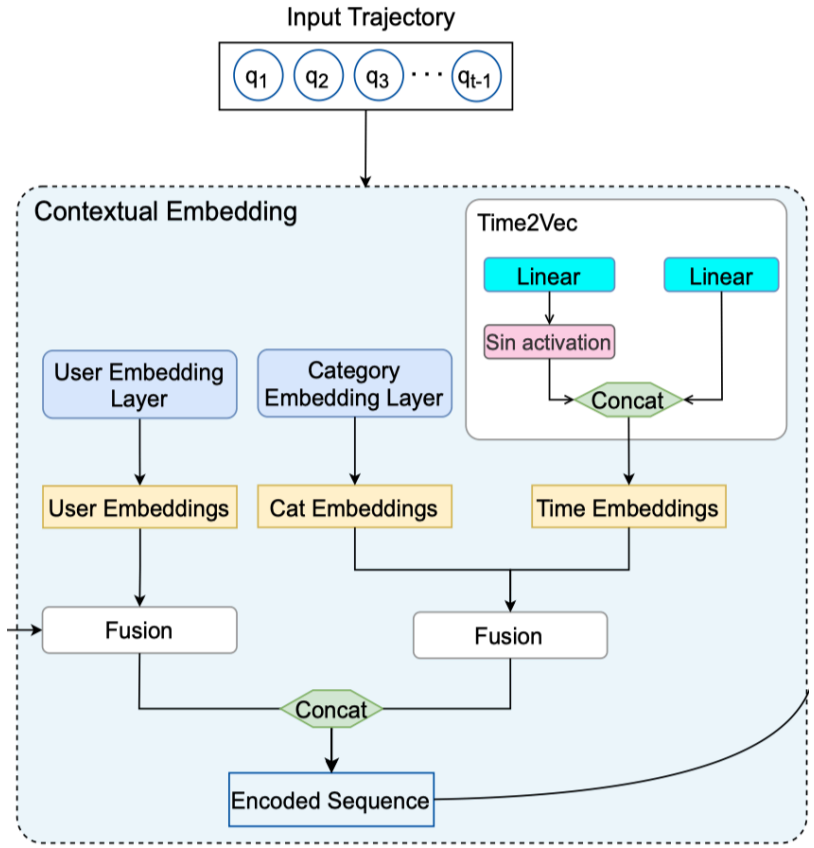
轨迹图与地点表征学习

- 轨迹图中每个节点表示一个地点，特征包含了访问频次、地理位置和类别
- 每条有向边的权重是这条路径的频次
- 使用图卷积神经网络进行地点的表征学习
- 使用图注意力机制处理获得转移概率图

3.1 模型基础：Vanilla GETNext Model

序列表征学习与数据融合

- 融合地点信息和用户信息来刻画用户偏好
- 融合时间信息和类别信息以描绘访问时间与地点类别之间的关系

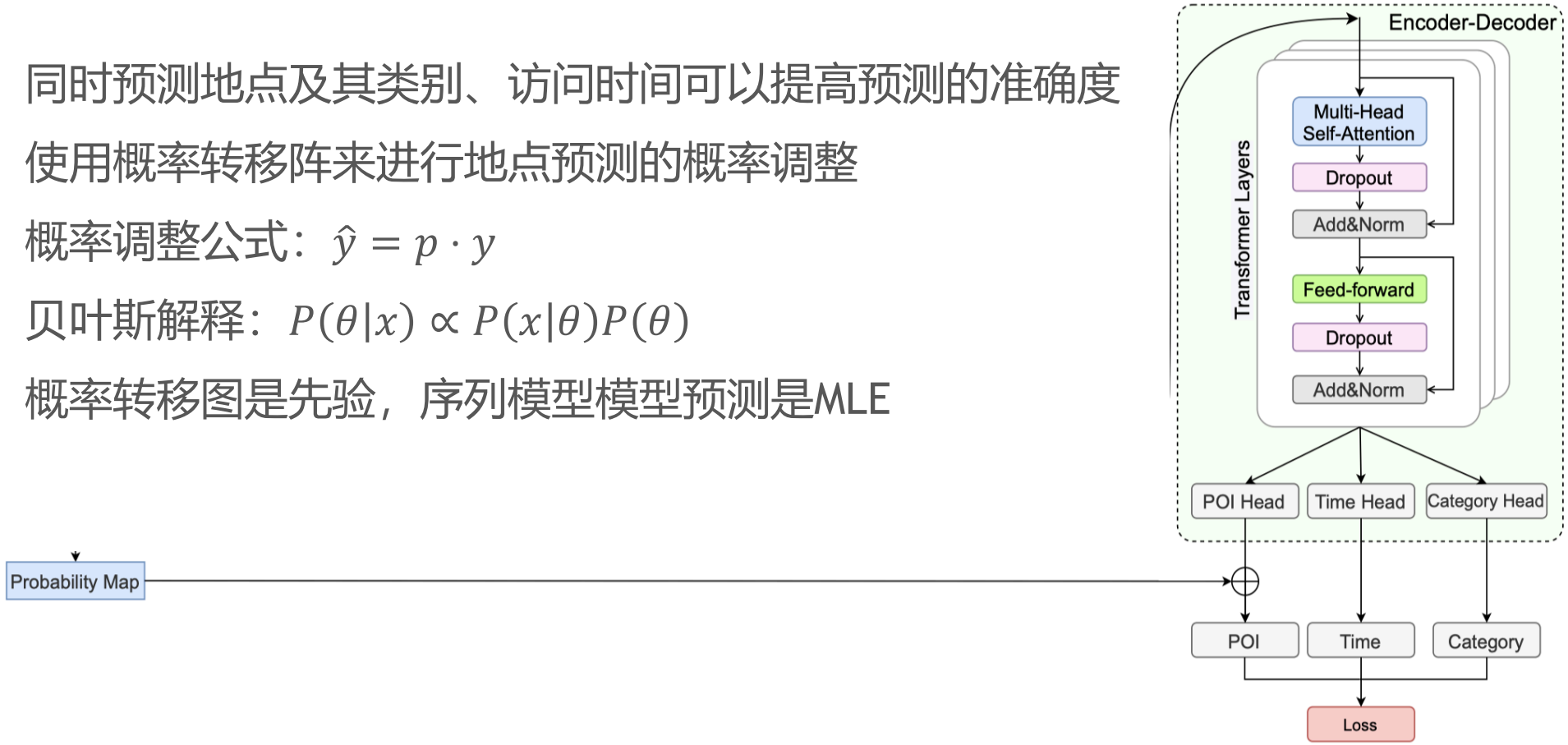


- 例如，一般不会在白天去酒吧；大多数人晚上会回家

3.1 模型基础：Vanilla GETNext Model

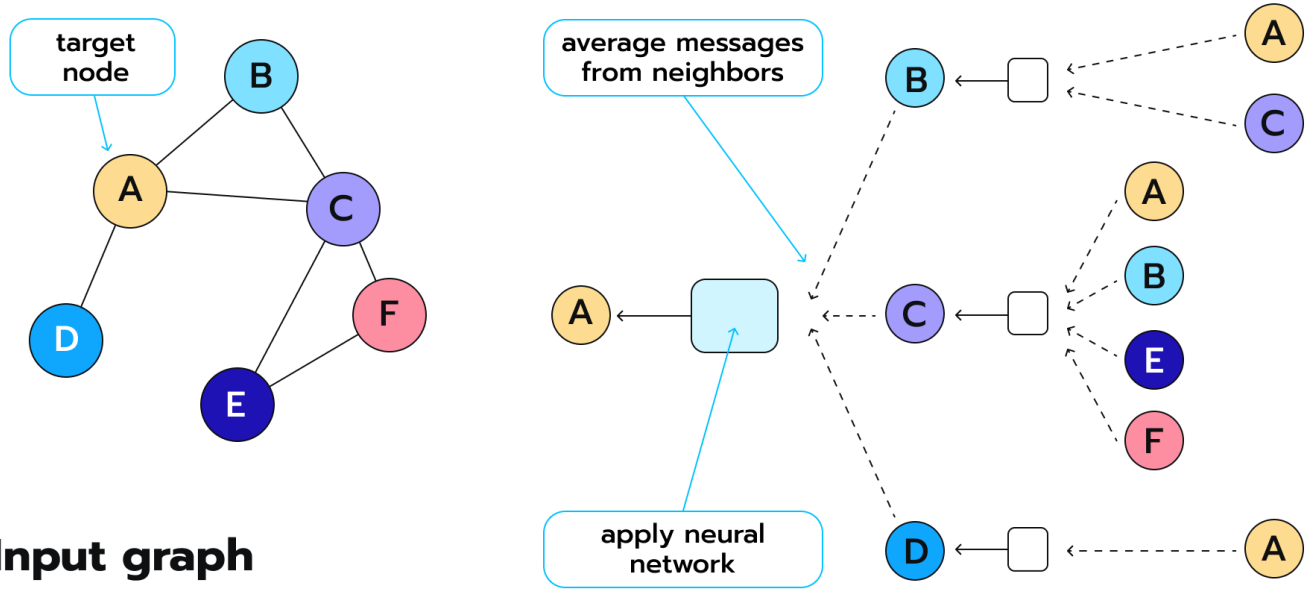
序列预测模型与概率调整

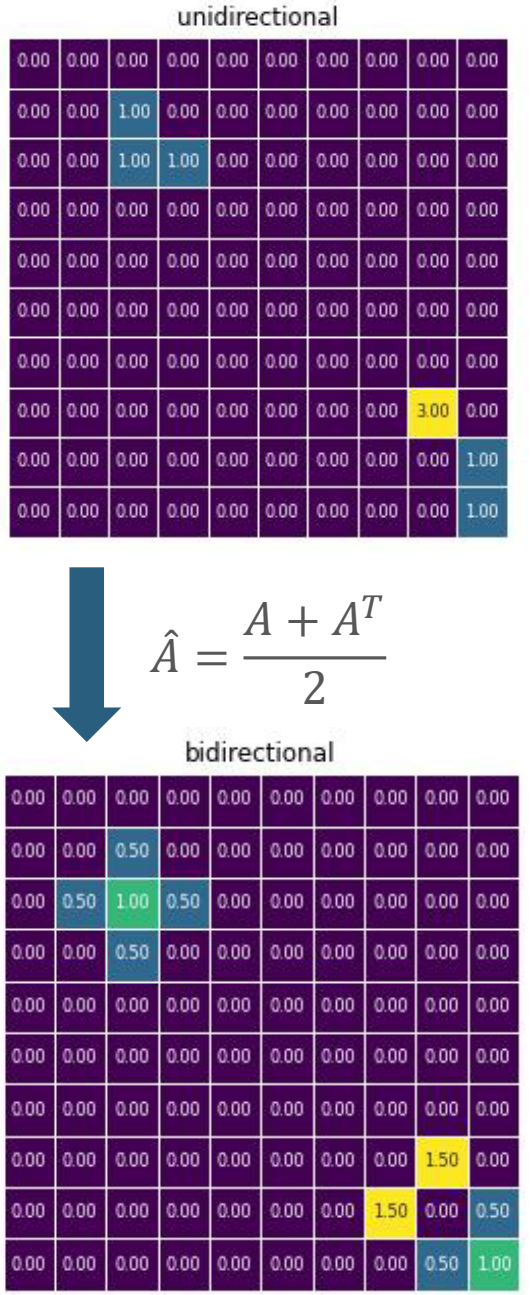
- 同时预测地点及其类别、访问时间可以提高预测的准确度
- 使用概率转移阵来进行地点预测的概率调整
- 概率调整公式：$\hat{y} = p \cdot y$
- 贝叶斯解释：$P(\theta|x) \propto P(x|\theta)P(\theta)$
- 概率转移图是先验，序列模型模型预测是MLE



3.2 模型改进：双向轨迹图

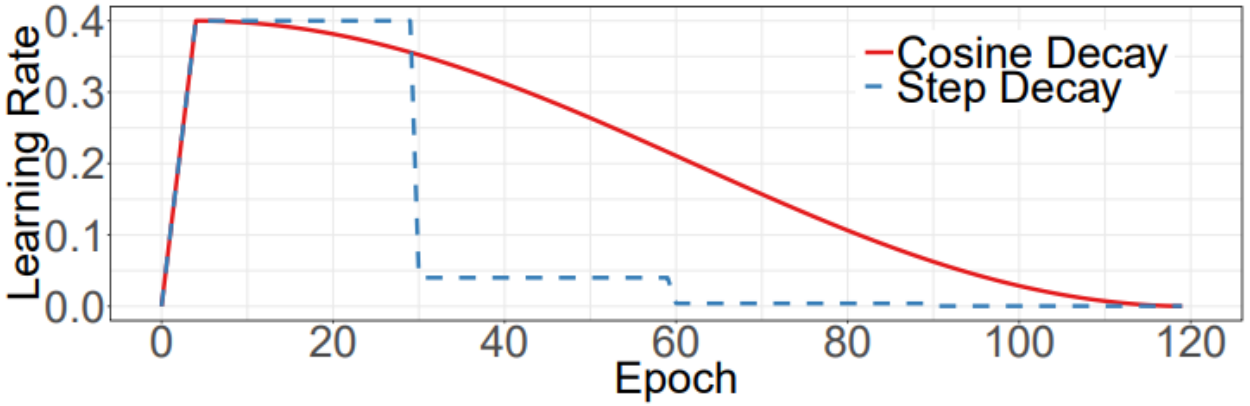
- GCN中每一层汇聚入边节点的信息
- 单向轨迹图会使得节点信息的汇聚不平衡，双向轨迹图可以让更好地GCN更好地处理地点之间信息
- 转移概率图仍然使用单向轨迹图以更好地刻画先后关系

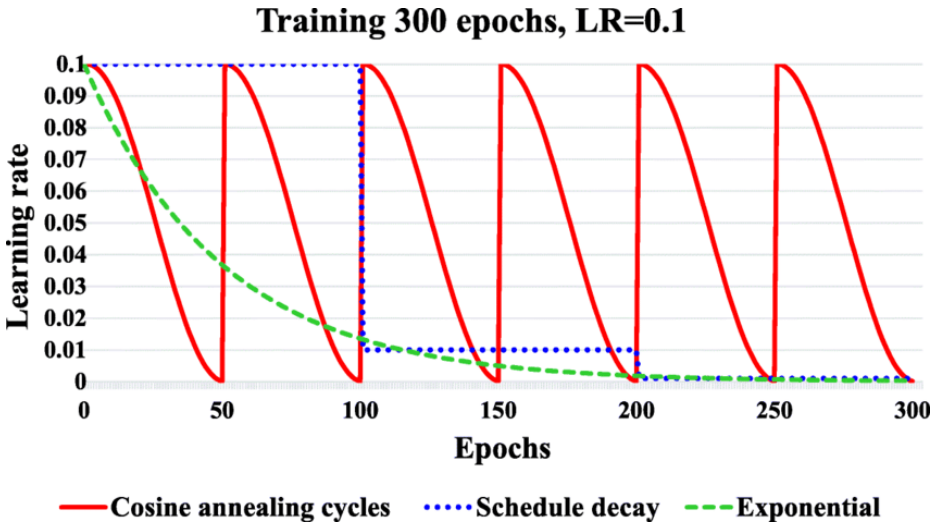




3.2 模型改进：Warm-up, Cosine Scheduler and AdamW

- 学习率预热(warm-up)是训练transformer的常见技巧，可以避免初期的梯度爆炸，使得模型初期的训练更加稳定，帮助模型更平滑地收敛
- 余弦退火学习率调整可以帮助模型逃离局部最小值
- AdamW优化器可以通过weight decay与正则化解耦，帮助改善泛化性能





3.3 实验结果

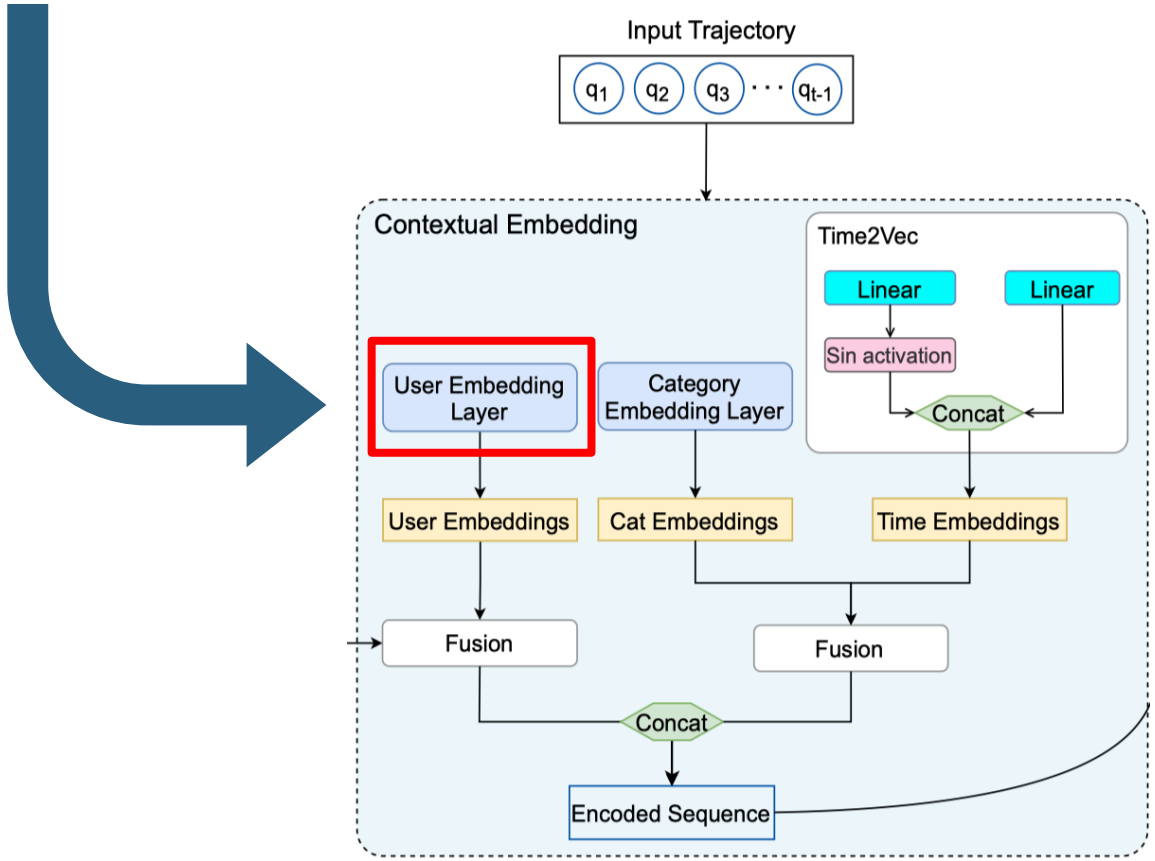
- 盲目的增大参数量、更改模型或激活函数无法改进模型
- 概率转移图和GCN输入均使用双向轨迹图可以有一定的提升
- 概率转移图使用单向轨迹图，GCN输入使用双向轨迹图能够在Top1正确率有1%的提升
- 进一步使用warm-up学习率调整可以显著在大多数指标上有1-2%的提升

Table 1: The accuracy of point of interest prediction

Model	Top1	Top5	Top10	Top20	MRR
LSTM	0.1305	0.2719	0.3283	0.3568	0.1857
ST-RNN	0.1483	0.2923	0.3622	0.4502	0.2198
STGCN	0.1799	0.3425	0.4279	0.5370	0.2806
STAN	0.2231	0.4582	0.5734	0.6328	0.3253
Original	0.2435	0.5089	0.6143	0.6880	0.3621
Reproduction	0.2370	0.5085	0.6148	0.6819	0.3627
Larger Embedding	0.2387	0.5001	0.6047	0.6770	0.3576
GIN	0.2462	0.5003	0.6137	**0.6932**	0.3654
ELU	0.2359	0.5029	0.6015	0.6868	0.3604
ProbMap-Bi + GCN-Bi	0.2465	0.5169	0.6162	0.6882	0.3678
ProbMap-Uni + GCN-Bi	0.2532	0.5014	0.6121	0.6783	0.3646
Warm-up + ProbMap-Uni + GCN-Bi	**0.2585**	**0.5201**	**0.6224**	0.6863	**0.3759**

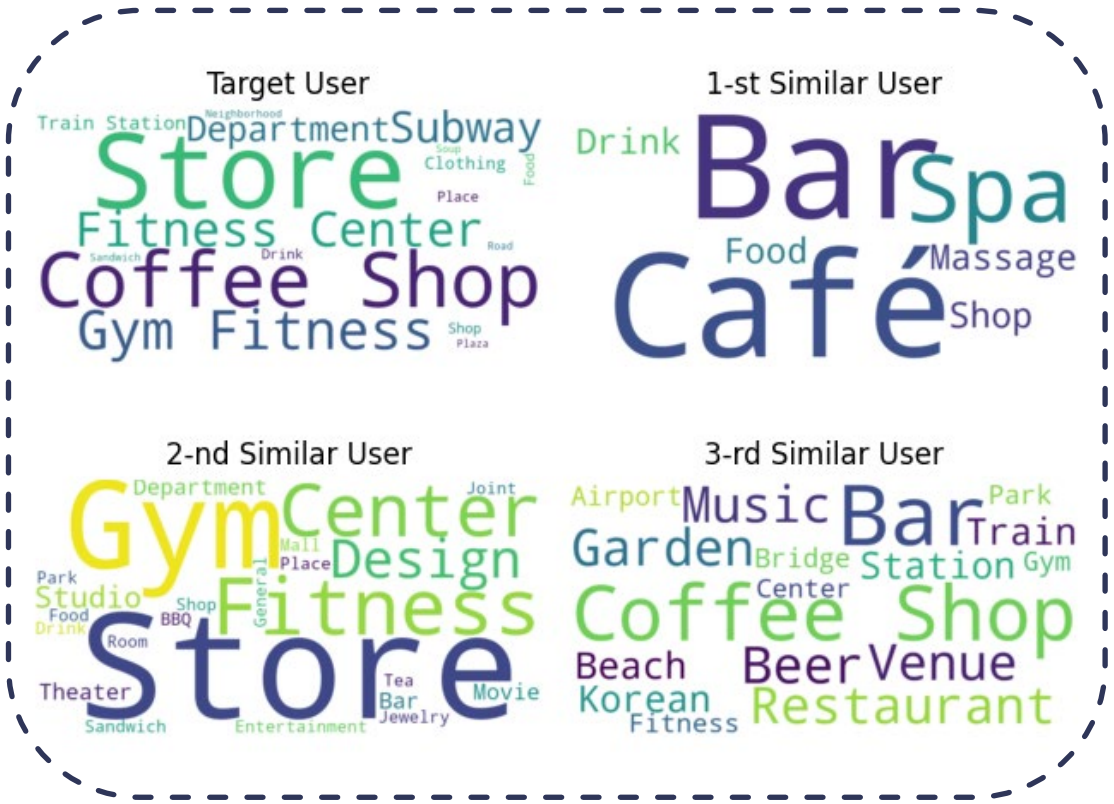
3.4 模型解释

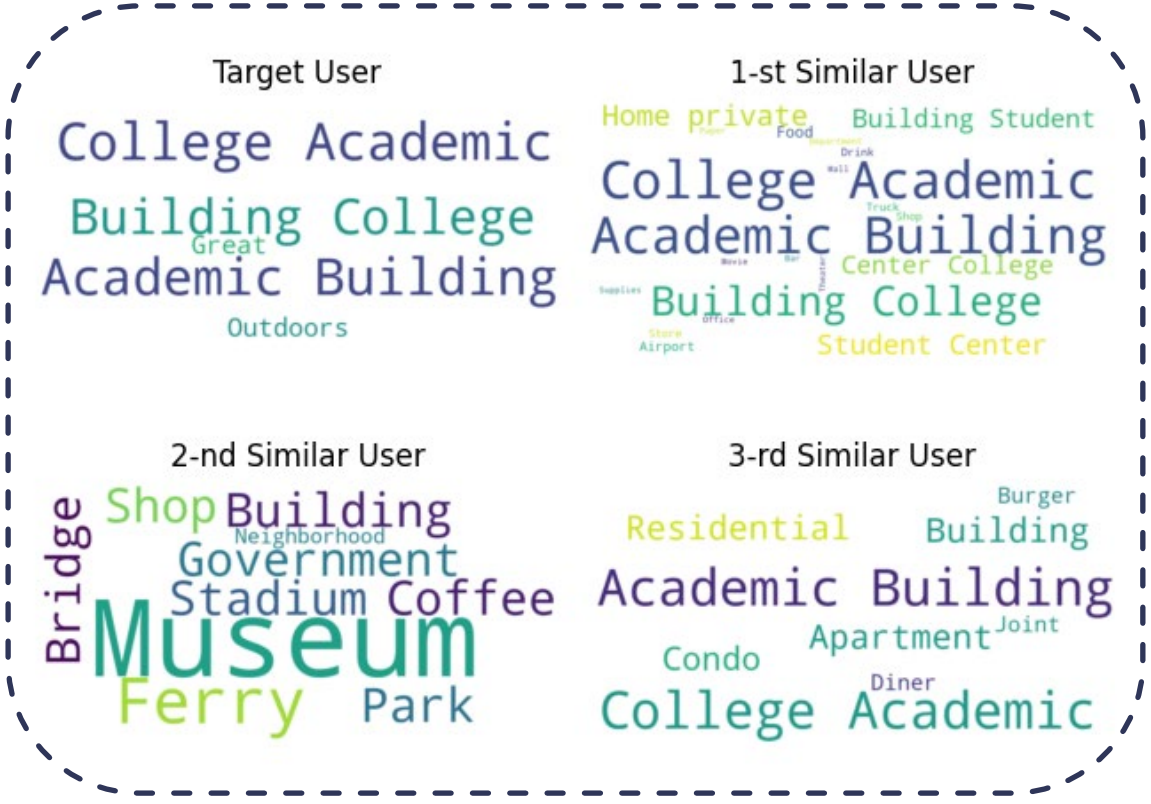
- 目标：解释End-to-End范式训练出来的黑盒模型
- 从训练出来的用户表征入手，探索表征对用户画像的刻画



3.4 模型解释

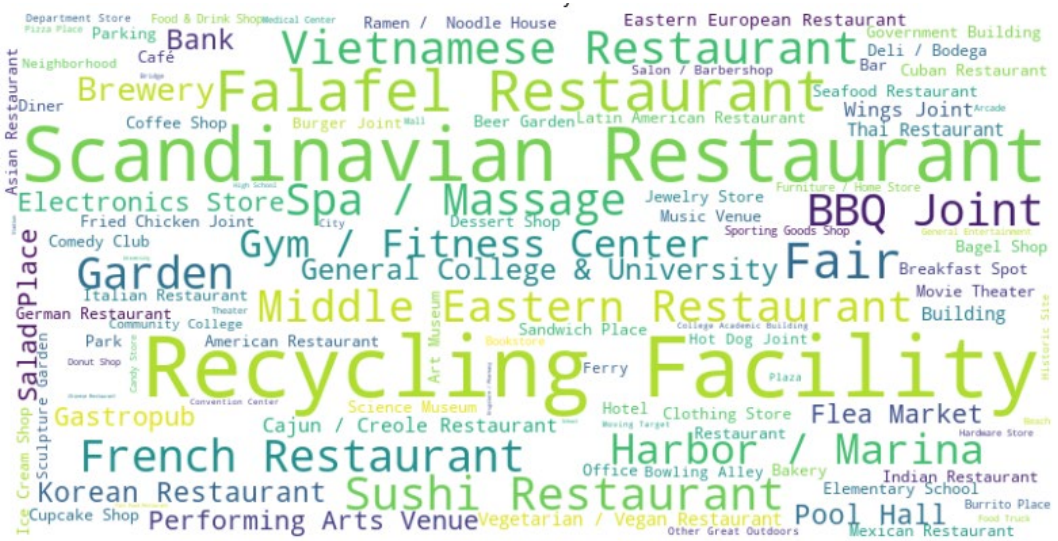
- 如何解构表征与用户画像？→ 利用表征寻找相似的用户，他们是否有相同的偏好
- 相似度计算：$d(x,y) = \frac{x^T y}{\|x\|\|y\|}$





3.4 模型解释

- 进一步将用户与其最相似的三个用户相连，进行社区挖掘
- 基于深度表征学习的结果与基于传统特征工程的结果有相似之处，也能发现一些不一样的特征关联
- 可以通过相似度和社区挖掘进行进一步的兴趣点推荐



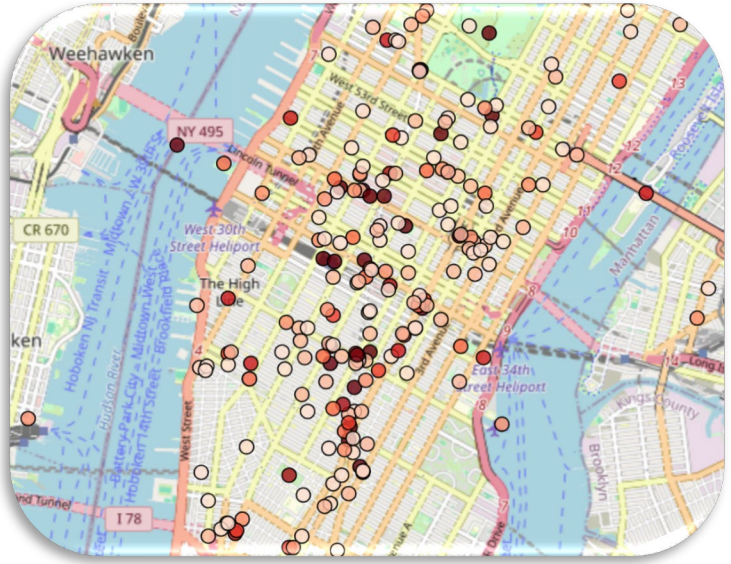
总结：再也不用担心找不到饭搭子和宝藏饭店！

1. 找到了吃货组织！

快来把你拉进群！

2. 知道大家去哪吃！

加入to-do list！



3. 正合我意！

你怎么知道我想去这里吃！

冯嘉旭 吴嘉骜 周志凌

FUDAN UNIVERSITY 1905
復旦大學
FUDAN UNIVERSITY