Location-based community mining and location prediction
基于地理位置的社区挖掘与地理位置预测

Social network mining

content

Part 1 描述分析

1.1 数据及介绍

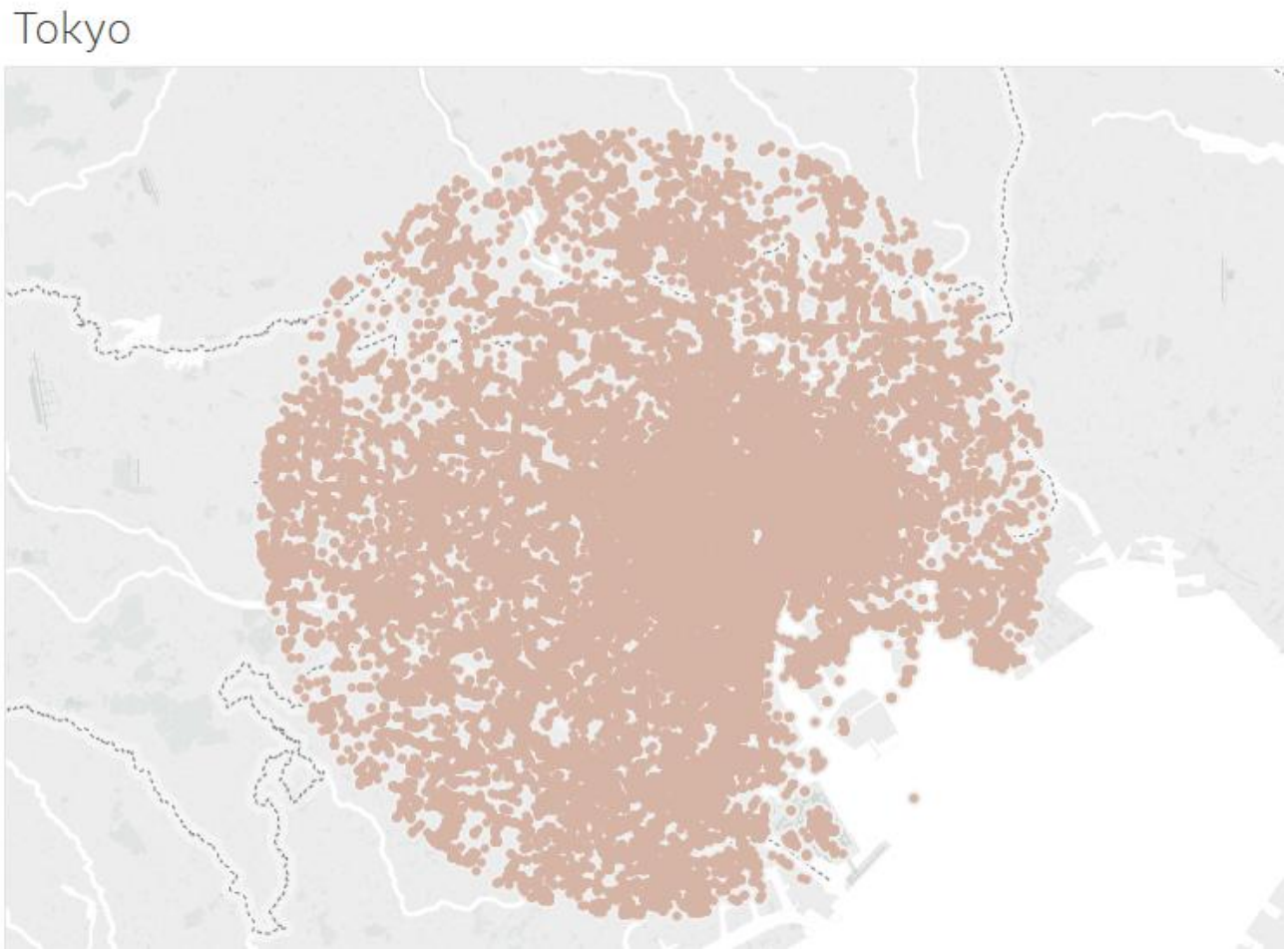
- Foursquare
- https://www.kaggle.com/chetanism/foursquare-nyc-and-tokyo-checkin-dataset

- ——Tokyo
- 2293 用户
- 537303 条签到记录
- 2012/4/4 – 2013/2/16

•变量：
•User ID (anonymized)
•Venue ID (Foursquare)
•Venue category ID (Foursquare)
•Venue category name (Fousquare)
•Latitude
•Longitude
•Time

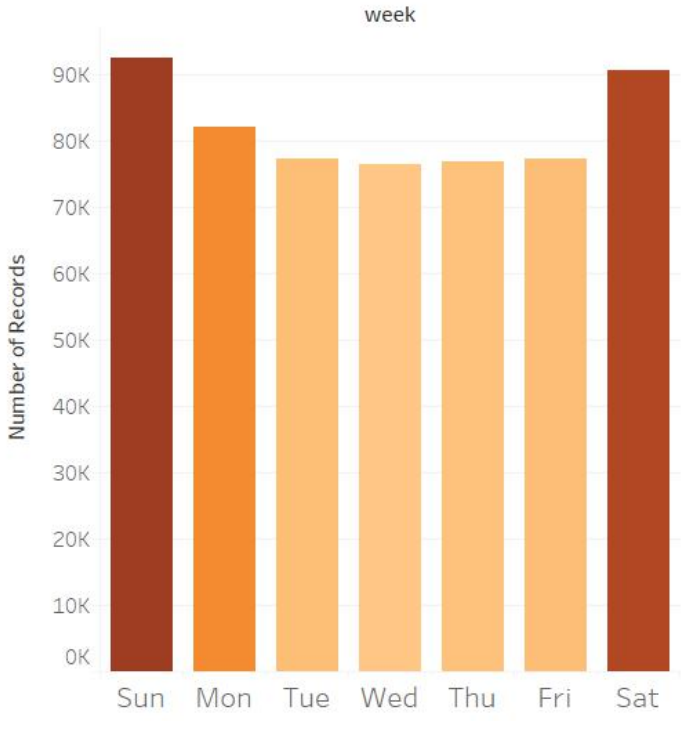
Part 1 描述分析

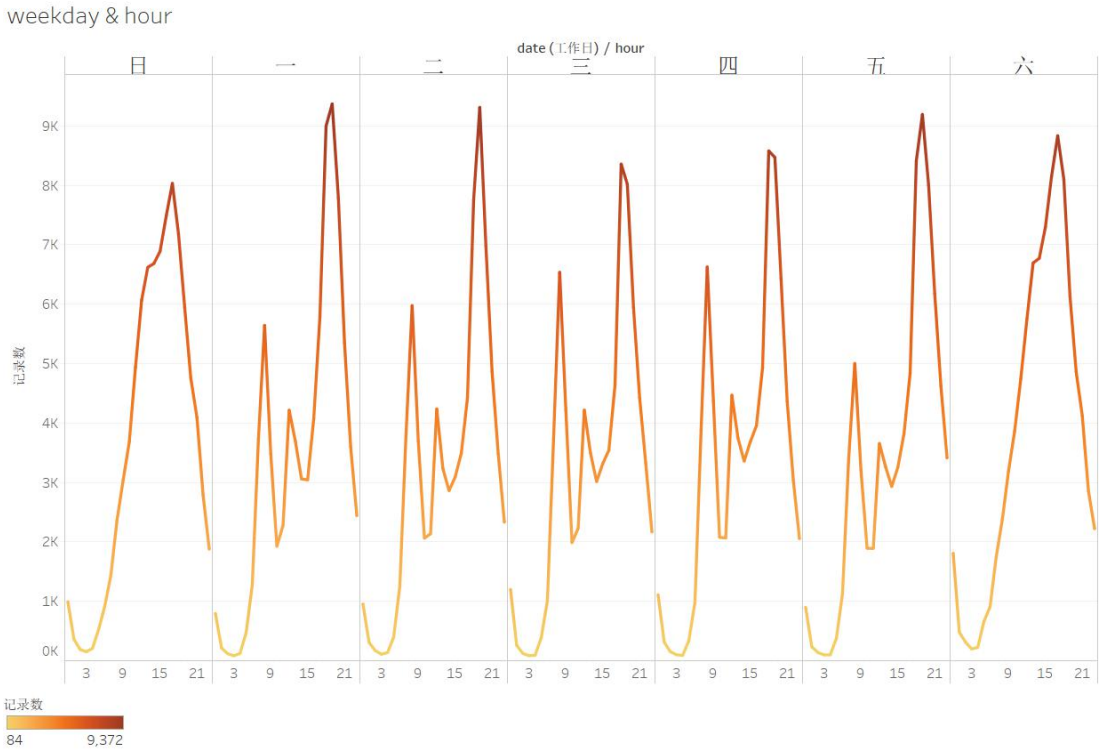
1.2 数据可视化



Part 1 描述分析

1.3 数据特点　　周期性&连续性

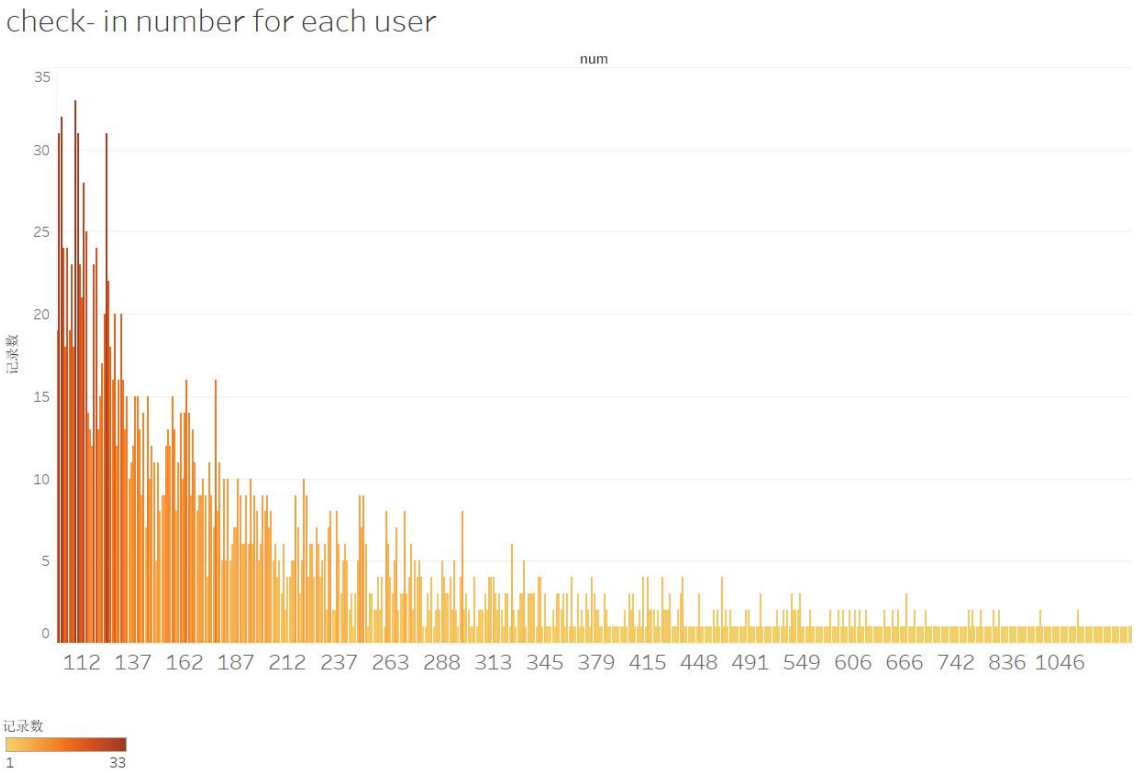




Part 1 描述分析

1.3 数据特点

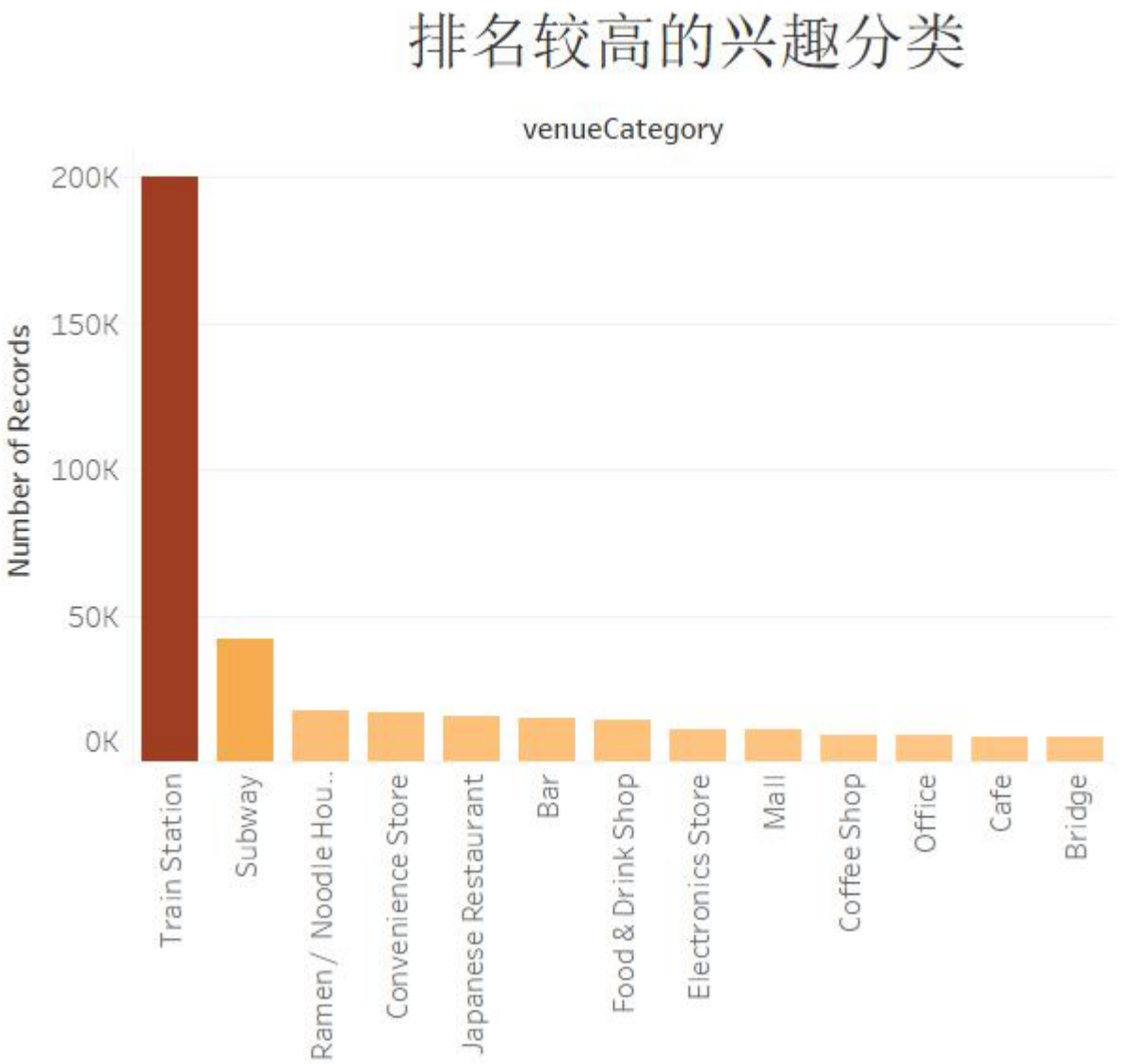
长尾



大多数用户打卡次数在100-200之间

Part 1 描述分析

1.3 数据特点



Part 2 项目情况

2.1 数据预处理

Time：timezone（540）有9小时时差，需要因此改变时间

venueCategory：247
train station（37.3%）

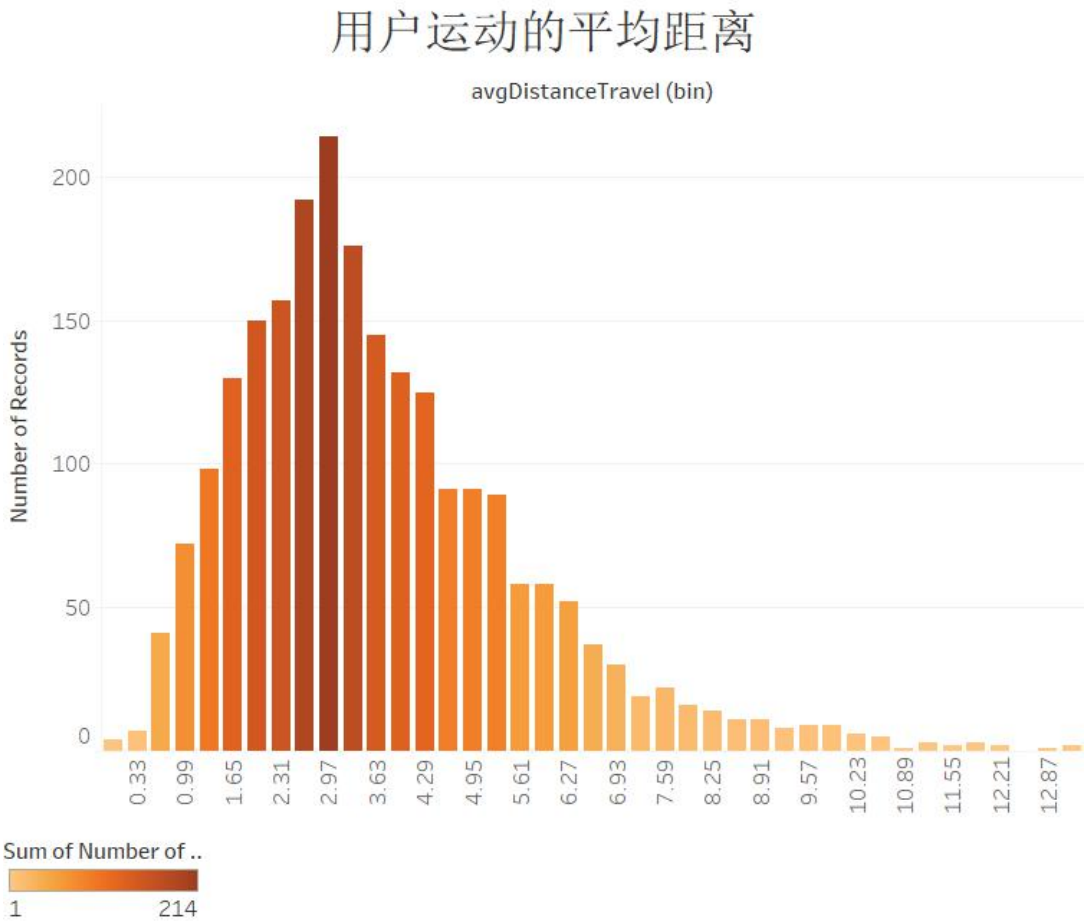
venueId：61858

Tool：
Python —— Plotly, Bash
Gephi

Part 2 项目情况

2.1 数据预处理

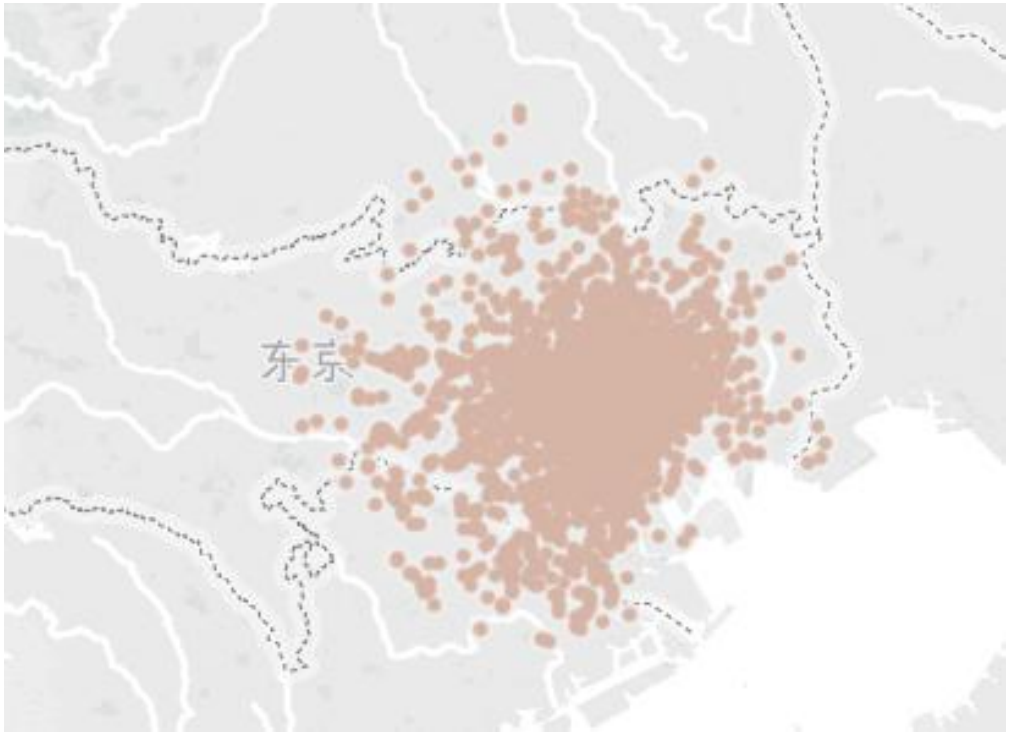
计算每个用户运动的平均距离



Part 2 项目情况

2.1 数据预处理

计算每个用户的中心地点



Part 2 项目情况

2.2 计算相似度

- User influence —— user based **collaborative filtering**
- Social influence —— friends
- Geographical influence —— spatial clustering

我们采取了3种思路：

1. 基于地点类型（ venueCategory ）计算相似度
2. 基于具体地点（ venueId ）计算相似度
3. 基于地点+时间（时空轨迹）计算相似度

Reference：杜翠凤,陈少权. 基于社交网络和地理位置的用户关系预测[J]. 移动通信, 2017,41(8): 21-23.

Part 2 项目情况

2.2 计算相似度

基于地点类型/具体地点计算用户相似度

- user based collaborative filtering
- Pearson 相关系数

$$sim(U_u, U_v) = \frac{\sum_i (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_i (r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_i (r_{v,i} - \bar{r}_v)^2}}$$

r：某个用户在某地的打卡次数

基于具体地点的相似度矩阵过于稀疏：绝大多数用户间相似度为0

Reference：杜翠凤,陈少权. 基于社交网络和地理位置的用户关系预测[J]. 移动通信, 2017,41(8): 21-23.

Part 2 项目情况

2.2 计算相似度

基于地点+时间（时空轨迹）计算用户之间的相似度

$$UserSim\ (LocTra\ 1, LocTra\ 2) = (\sum_{j=1}^{m} SeqSim\ (seq[j])) / m$$

- m=4ε
- 半小时重合（1:00-2:00, 1:30-2:30, 2:00-3:00）

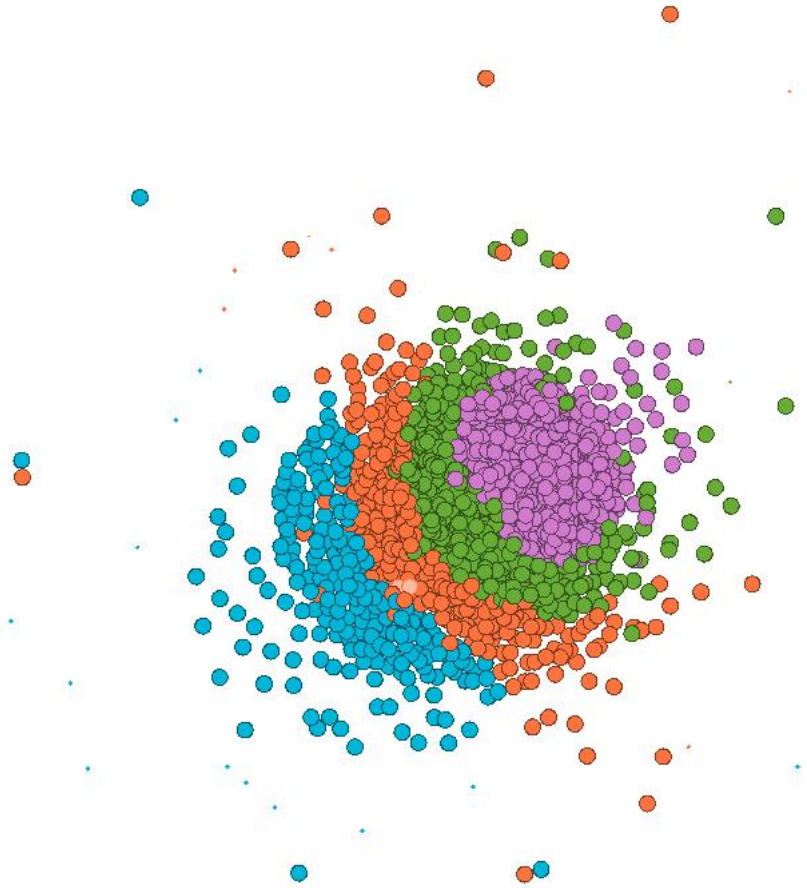
- 其中：时间轨迹的相似度

- Ci 和Ci’分别代表用户1 的位置 $SeqSim(seq) = \frac{1}{1 + dis\tan ce(C_i, C_i^{'})}$ 位置点

Reference：杜翠凤,陈少权. 基于社交网络和地理位置的用户关系预测[J]. 移动通信, 2017,41(8): 21-23.

Part 2 项目情况

2.3 社群划分 ——基于兴趣相似度的社群划分

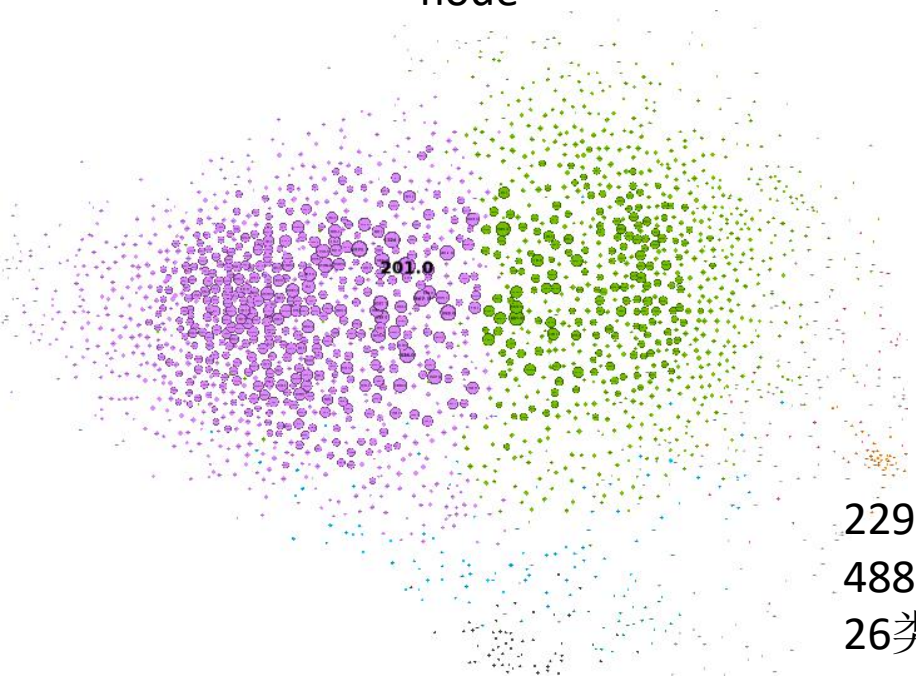


Part 2 项目情况

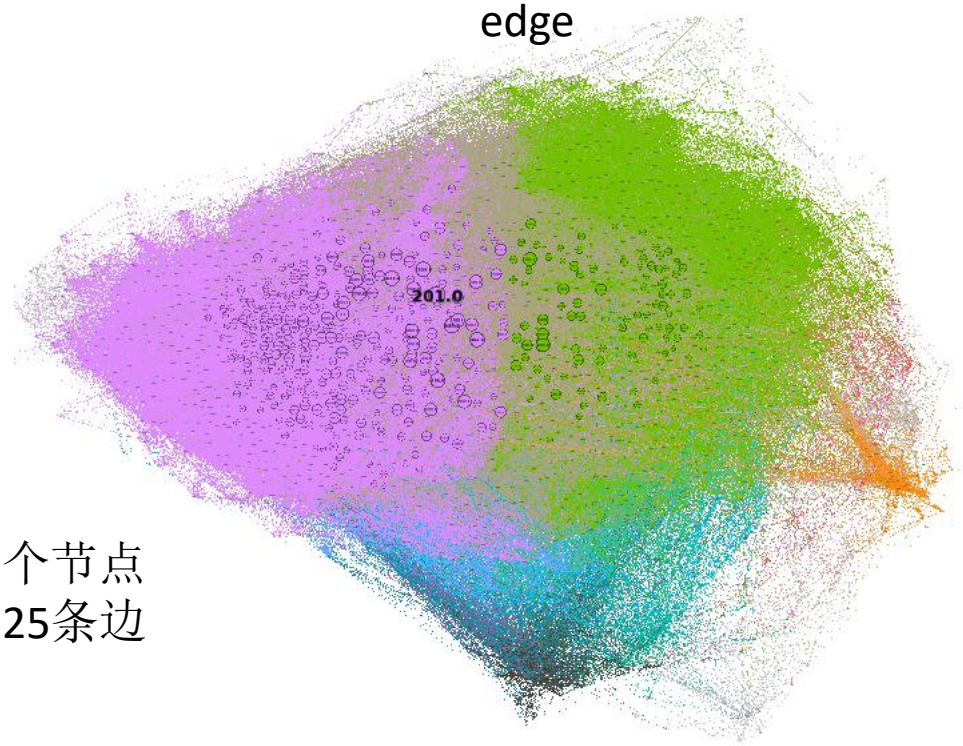
2.3 社群划分 ——基于轨迹相似度的社群划分

上面标有 userId

node



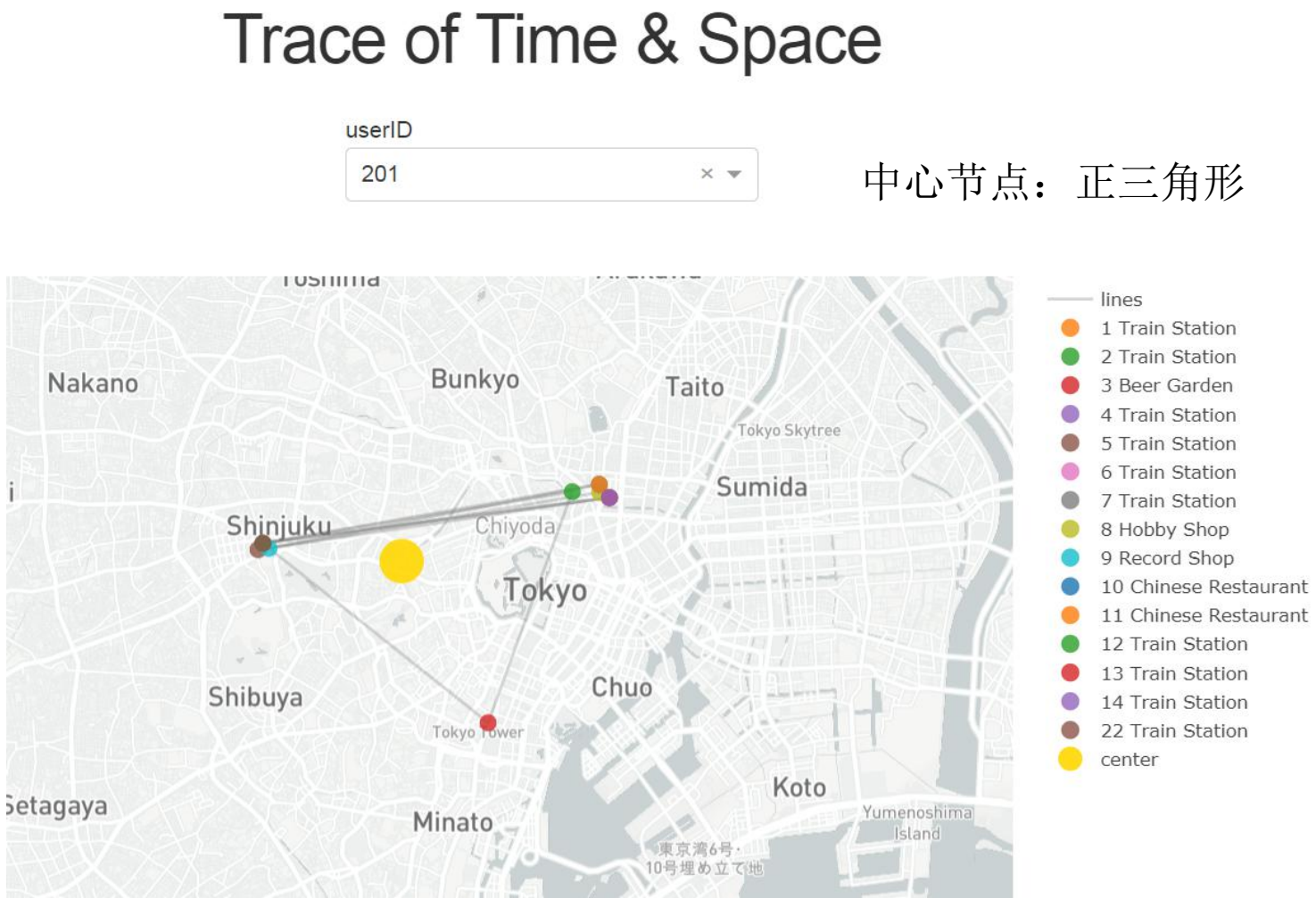
edge



2293个节点
488725条边
26类

Part 2 项目情况

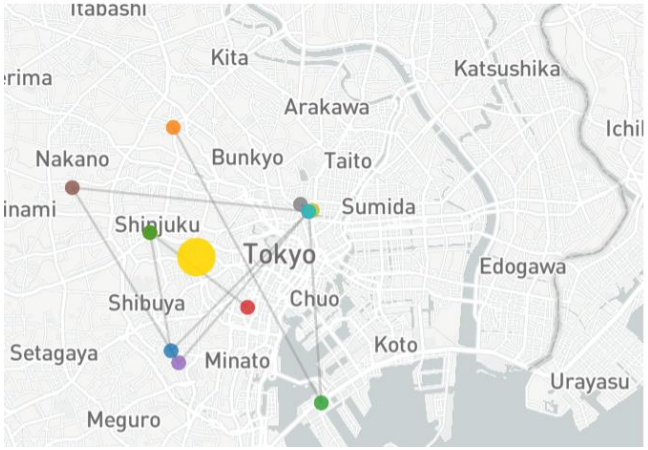
2.4 时空轨迹

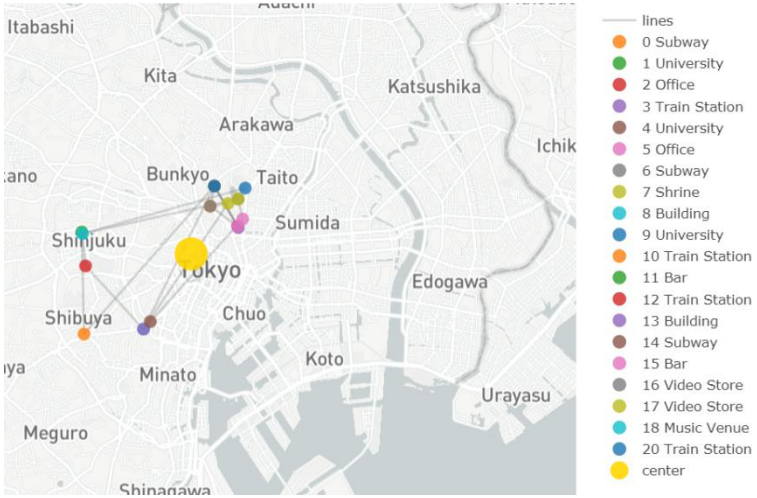


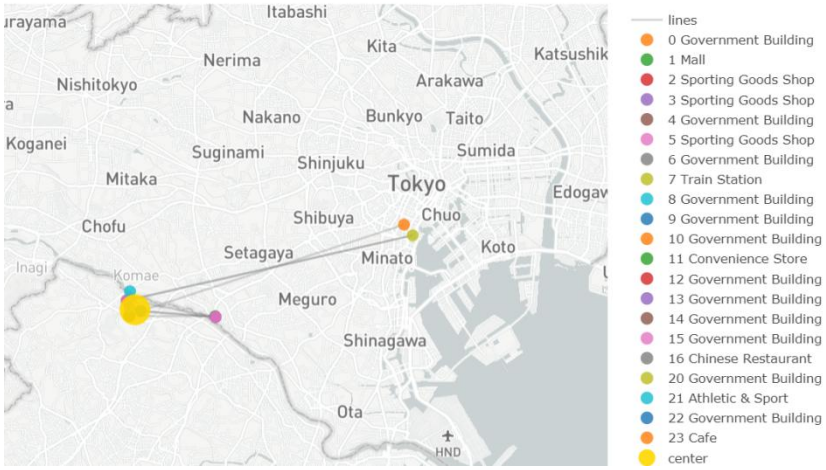
Part 2 项目情况

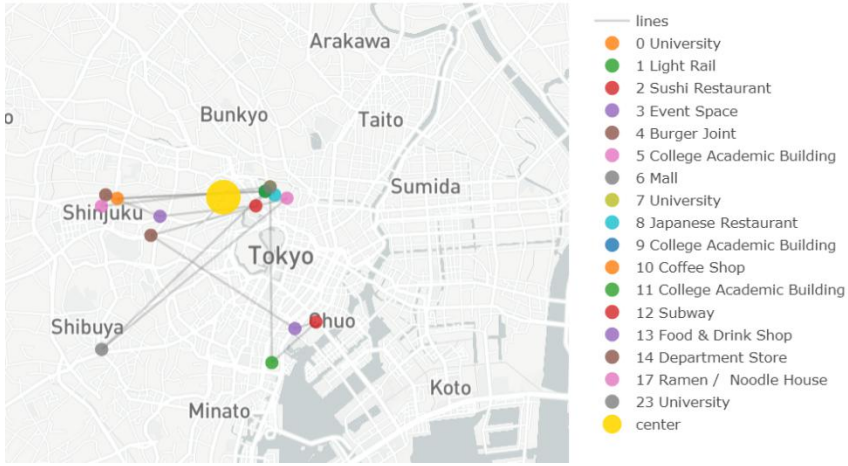
2.4 时空轨迹

第一类：





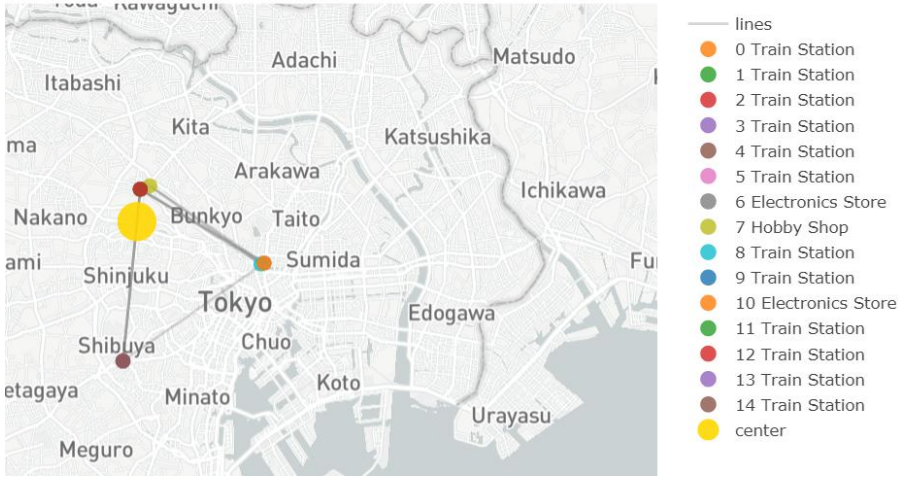


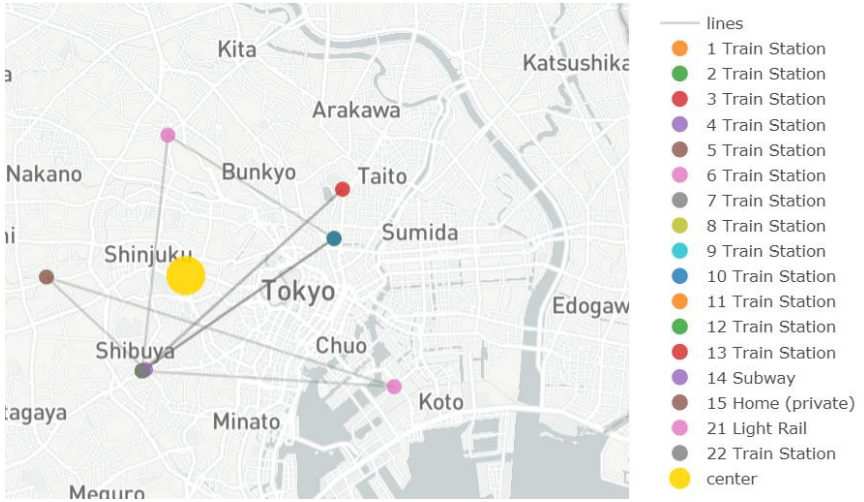


Part 2 项目情况

2.4 时空轨迹

第二类：

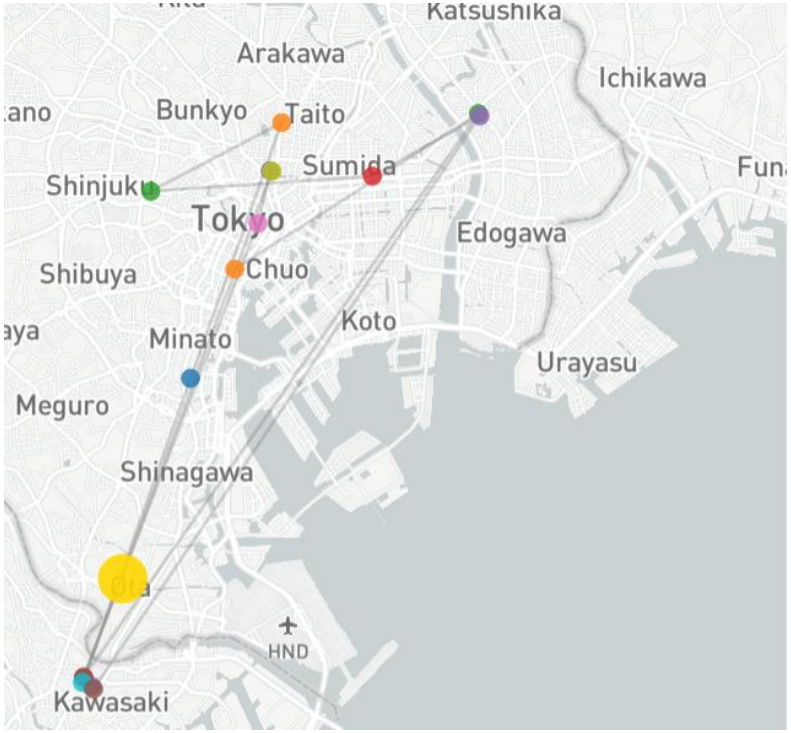


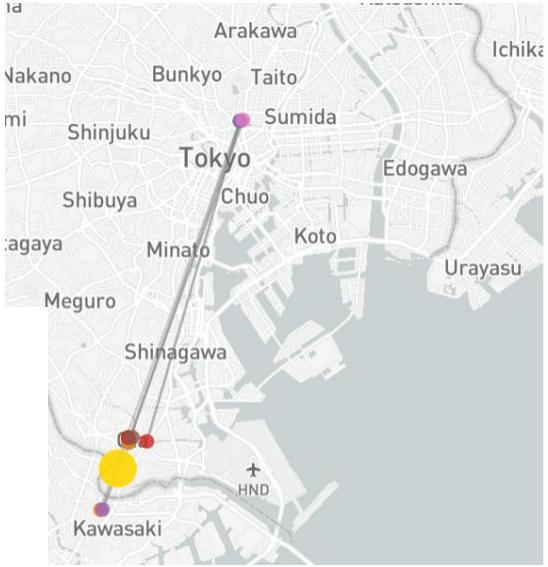


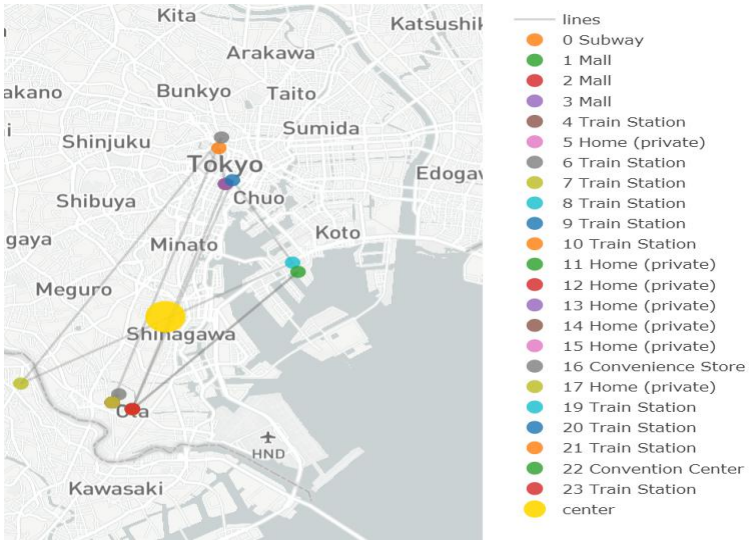
Part 2 项目情况

2.4 时空轨迹

第三类：



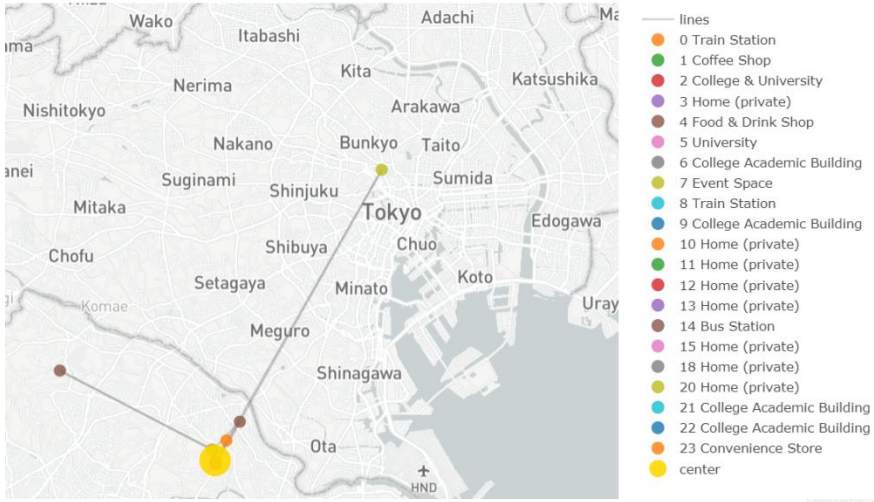




Part 2 项目情况

2.4 时空轨迹

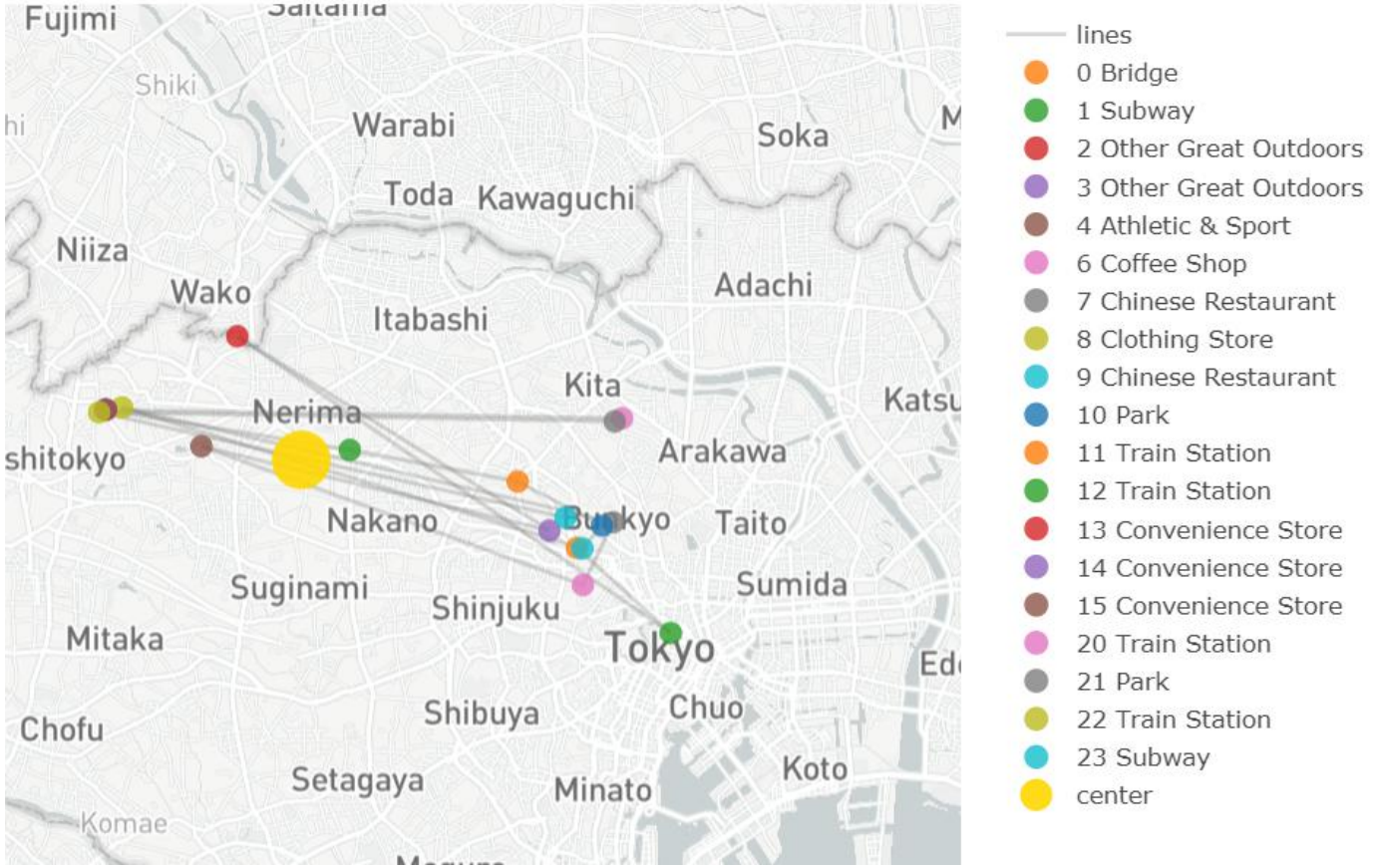
第四类：



Part 2 项目情况

2.4 时空轨迹

单独一类：



Part 2 项目情况

2.5 推荐系统

POI： point of interest

托布勒Topler第一地理定律：
“一切都相互关联，但近处的事物比远处的事物联系更紧密”
(1)人们**倾向于**访问离家或办公室较近的POI;
(2)人们**可能会**有兴趣探索他们喜欢的但是距离较远的POI

W. R. Tobler. A Computer Movie Simulating UrbanGrowth in the Detroit Region. Economic Geography,46:234–240, 1970.

Part 2 项目情况

2.5 推荐系统

推荐算法思路：

- 根据相似度矩阵，筛选与用户i相似度最高的10个用户
- 统计10个用户去过的总地点，计算每个地点的总次数，排序后从中挑选次数最多的地点，并过滤掉用户i已经去过的地点
- 根据用户对地理位置远近的偏好筛选地点

（如果用户的平均移动距离 avgDistanceTravel 较大，则不需筛选；如果用户的平均移动距离较小，
则过滤掉距离远的地点，只为他推荐距离近的地点），

Part 3 总结与期望改进

3.1 一些小不足

- 很多问题来源于数据集的不完善
- 无法找到一个统一的标准来度量社群划分的准确度与合理程度
- 无法判断各种矩阵相似度的准确程度
- 无法通过实践来计算RMSE，即无法判断推荐系统的准确程度

Part 3 总结与期望改进

3.2 总结： 如何做一个好的推荐系统

- 需要更好的综合各个元素：[地理位置、时间、地点类型]

- 需要明确用户的目的
——可以给用户的行为做层次分类：某次行为是基于兴趣爱好，还是只是为了生活需要？

- 需要搜集哪些数据才能做更好的推荐？
——需要用户的反馈（feedback）

Part 3 总结与期望改进

3.3 可以参考的文献算法

用户关系预测方法

- u,v 为用户，F(u),F(v) 为邻居集合
- 分子：共同邻居

$$\text{Similarity}(u,v)=\frac{\mathbf{F}(u)\cap\mathbf{F}(v)}{\mathbf{F}(u)\cup\mathbf{F}(v)}$$

A-A系数

在考虑社交关系的基础上，对上述公式进行改进，在链接预测中，一个兴趣被越少的人拥有，则拥有此兴趣的人越可能成为朋友，而大众兴趣的人之间成为朋友的可能性要低一些，因此该系数给度数较少的节点分配较高的相似度值。

$$\text{Similarity}(u,v)=\sum_{t\in\mathbf{F}(u)\cap\mathbf{F}(v)}\frac{1}{\log(d_t)}$$

Adamic L A, Adar E. Friends and neighbors on the Web[J]. Social Networks, 2003,25(3): 211-230.

Part 3 总结与期望改进

3.3 可以参考的文献算法

地理位置和时间的重合度

设用户u的轨迹可以用时间戳和地点ID的组合来表示，如：<t1, l1>, <t2, l2>, …, <tn, ln>。

对于用户u的区域分布GP(u, r)为：

$$GP(u,r)=\sum_{i=1}^{n}\frac{\delta(r,l_i(u))}{n(u)}$$

当$r= l_i(u)$时，$\delta(r, l_i(u))=1$，否则为0。

结合时间因素，以△T为时间精度（一般设为1个小时），反映所有用户在邻近时间相同地理位置的比例。

同时，考虑工作时间与非工作时间的影响因素，在工作时间段和非工作时间段设置不同的权重θ。

$$CoL(u,v)=\frac{\sum_{i=1}^{n(u)}\sum_{j=1}^{n(v)}\theta(\Delta T-|T_i(u)-T_j(v)|)(\delta(l_i(u),l_i(v))}{\sum_{i=1}^{n(u)}\sum_{j=1}^{n(v)}\theta(\Delta T-|T_i(u)-T_j(v)|)}$$

Adamic L A, Adar E. Friends and neighbors on the Web[J]. Social Networks, 2003,25(3): 211-230.

Part 3 总结与期望改进

3.3 可以参考的文献算法

基于用户社交网络和地理位置用户关系预测模型

$$MR(u,v)=\gamma \sum_{t\in \mathbf{F}(u)\cap \mathbf{F}(v)} \frac{1}{\log(d_t)} + (1-\gamma) \sum_{r\in \mathbf{Loc}} \frac{GP(u,r)\times GP(v,r)}{\| GP(u,r) \| \times \| GP(v,r) \|}$$

用户社交网络

地理位置

通过对用户的“时空切片”求余弦相似性

谢谢