

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 数据集简介
- 1.2 数据可视化
- 1.3 数据特点
- 1.4 可分析因素

Part 2 项目进展

- 2.1 数据与处理
- 2.2 计算相似矩阵
- 2.3 社群划分
- 2.4 用户时空轨迹
- 2.5 基于地理位置的推荐系统

Part 3 总结和期望改进

- 3.1 改进
- 3.2 总结

Part 1 描述分析

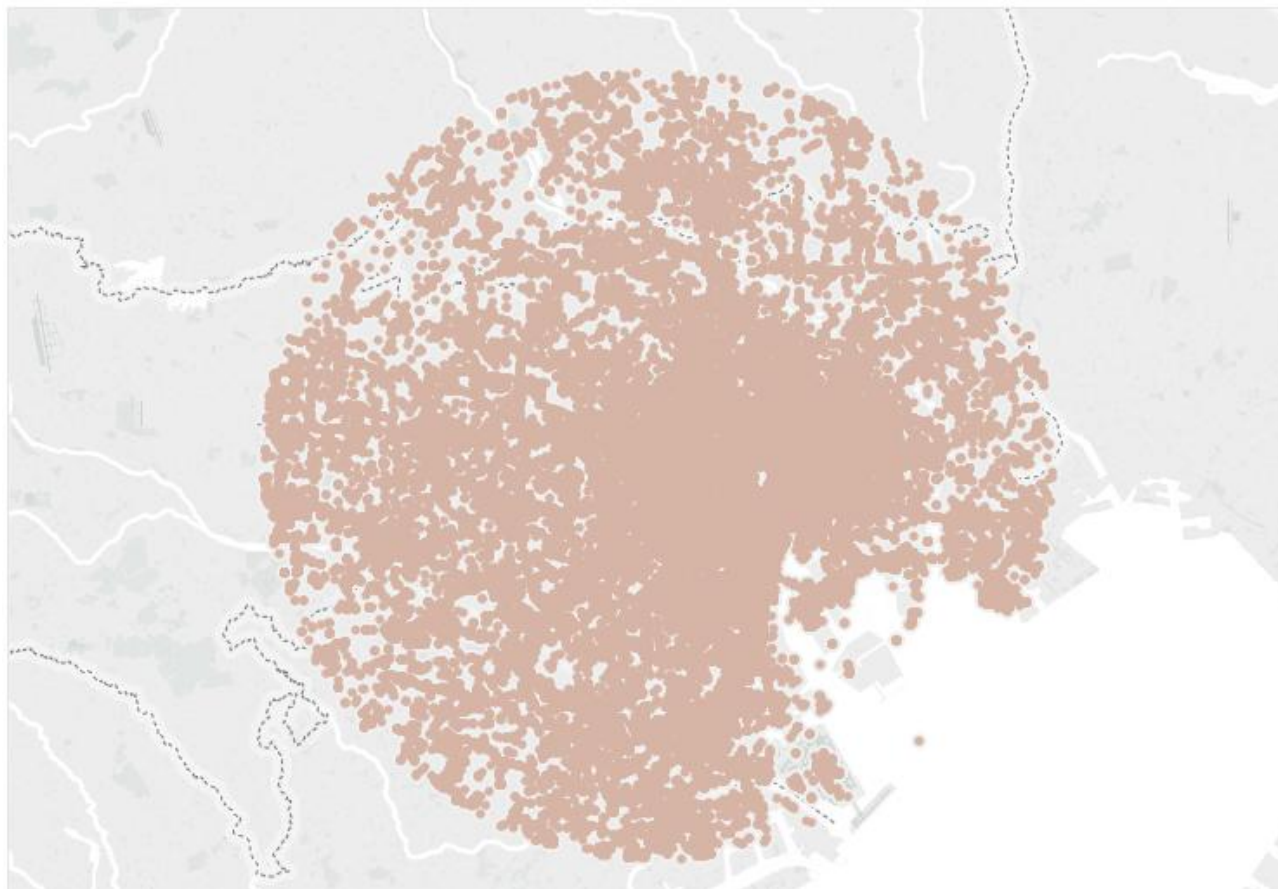
1.1 数据及介绍

- 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 (Foursquare)
 - Latitude
 - Longitude
 - Time

Part 1 描述分析

1.2 数据可视化

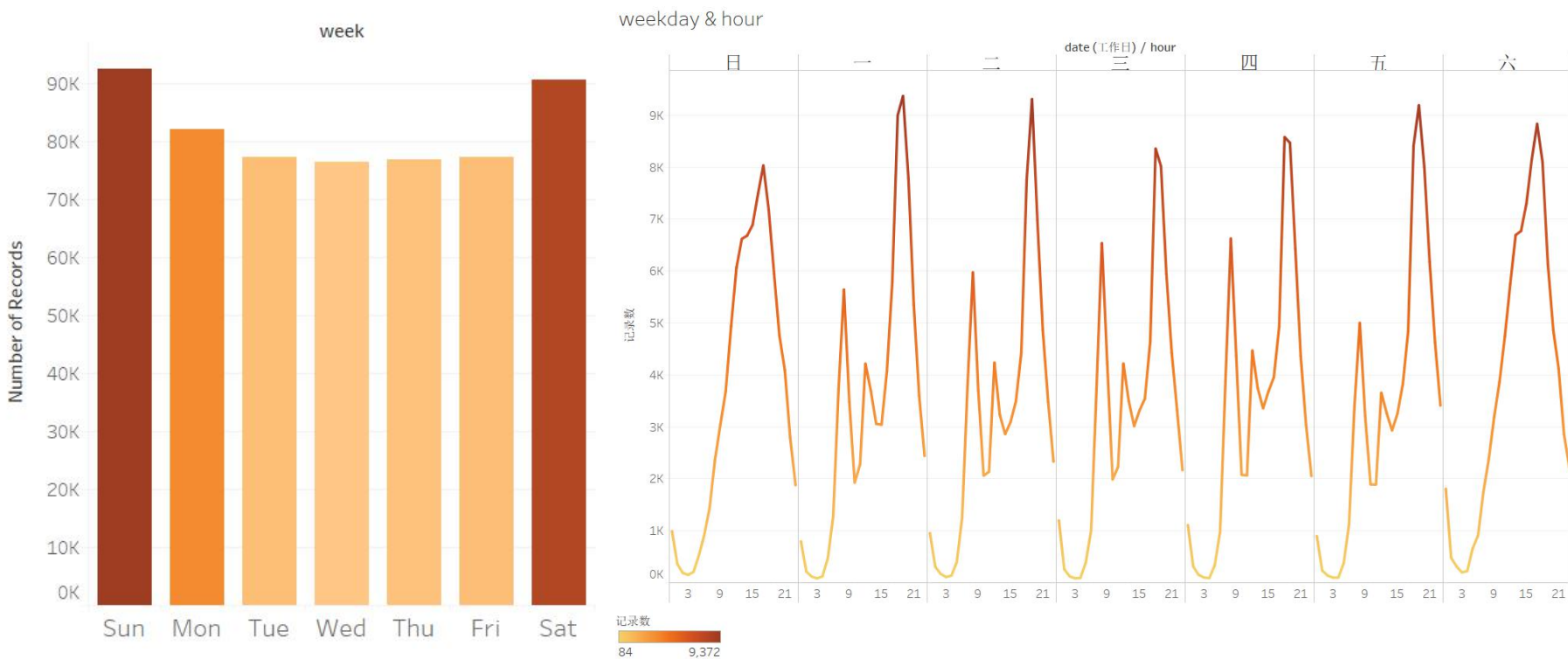
Tokyo



Part 1 描述分析

1.3 数据特点

周期性&连续性

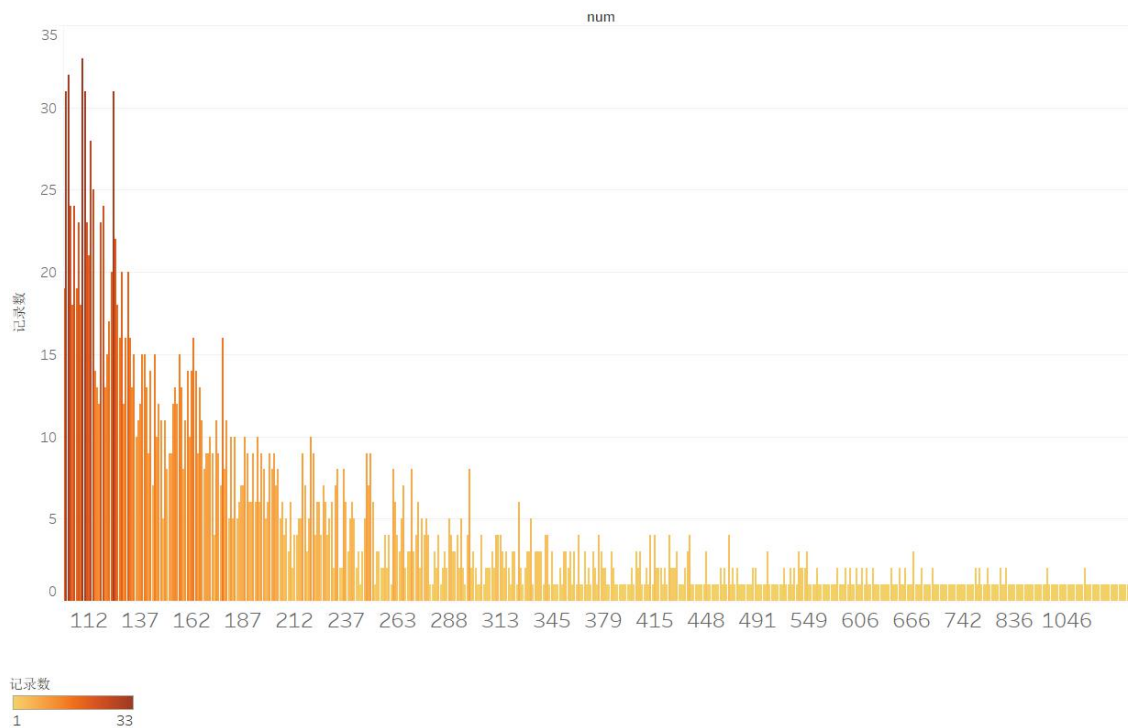


Part 1 描述分析

1.3 数据特点

长尾

check-in number for each user

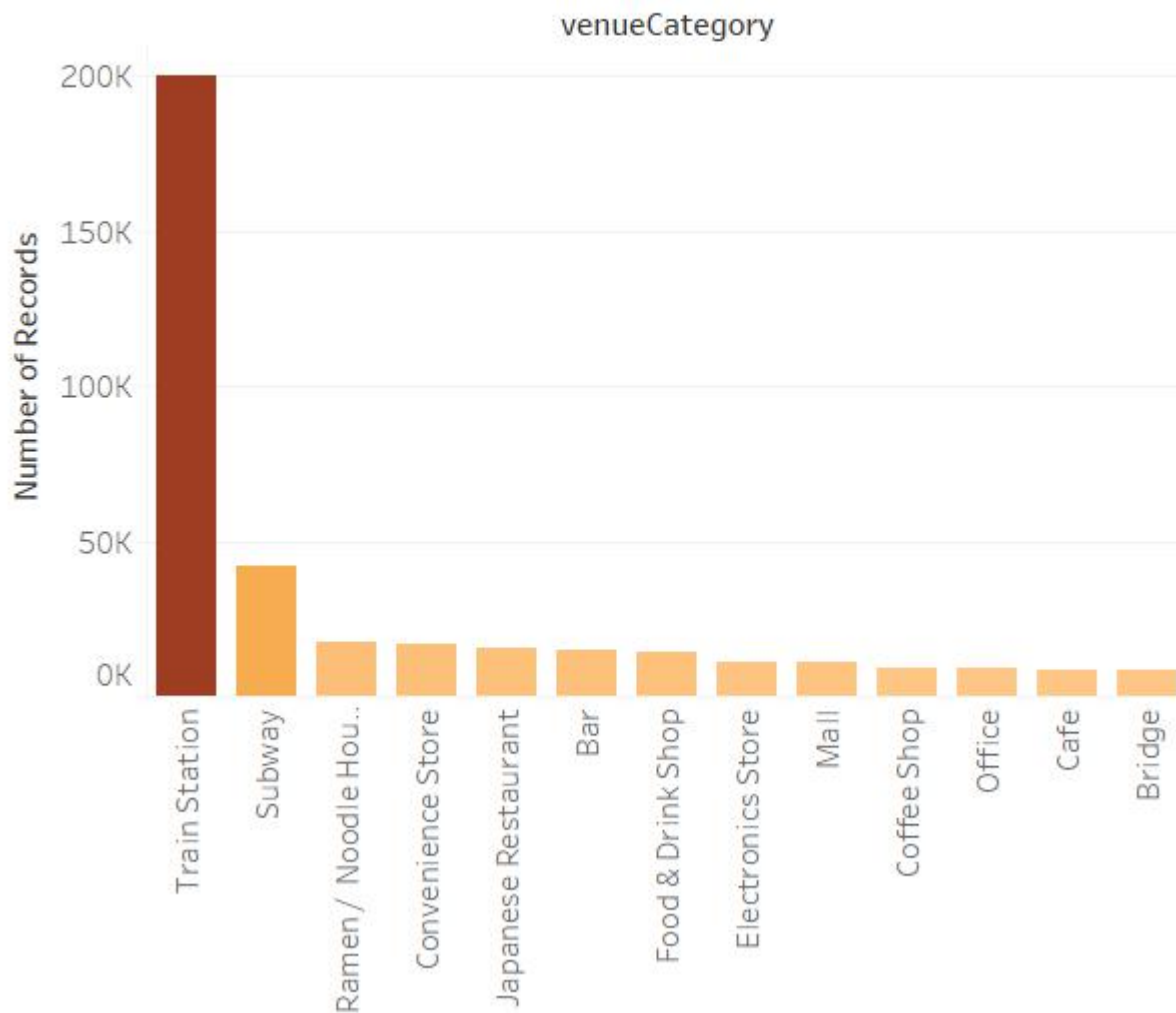


大多数用户打卡次数在100-200之间

Part 1 描述分析

1.3 数据特点

排名较高的兴趣分类



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:

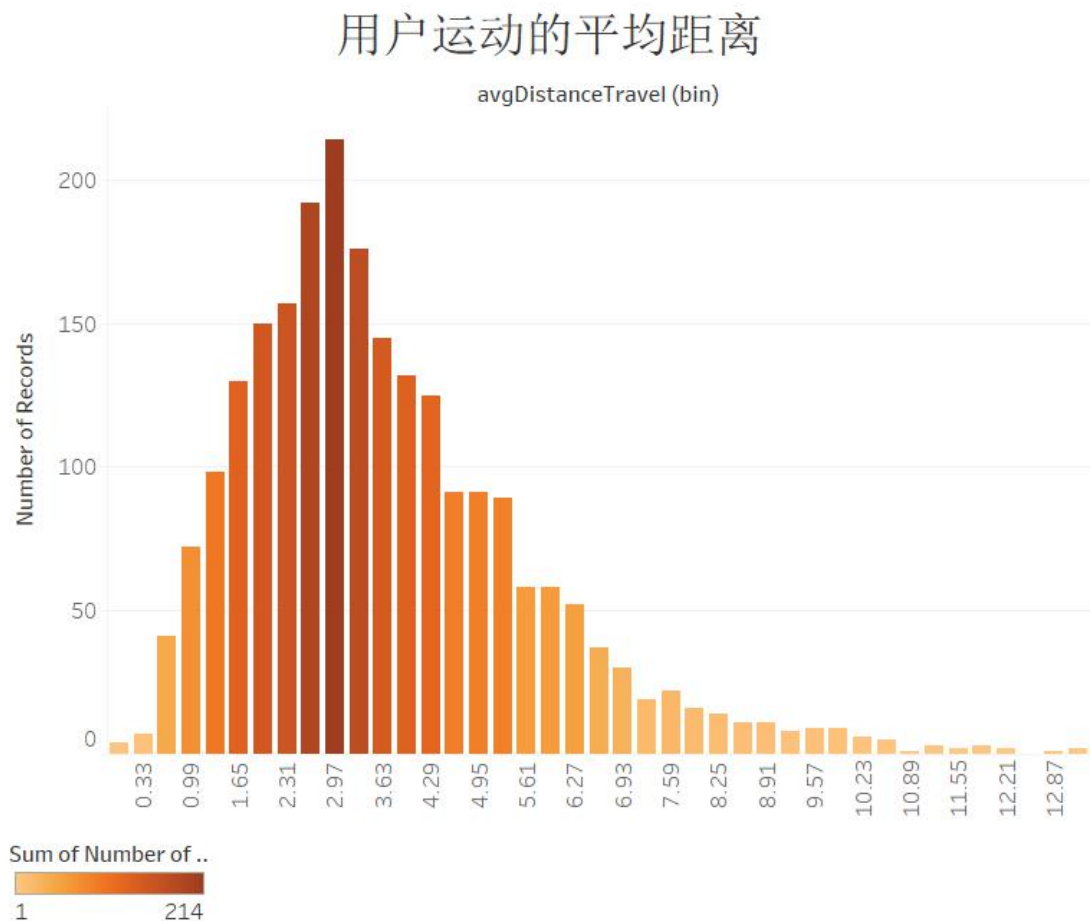
Python — Plotly, Bash

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

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

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=48 小时轨迹序列长度

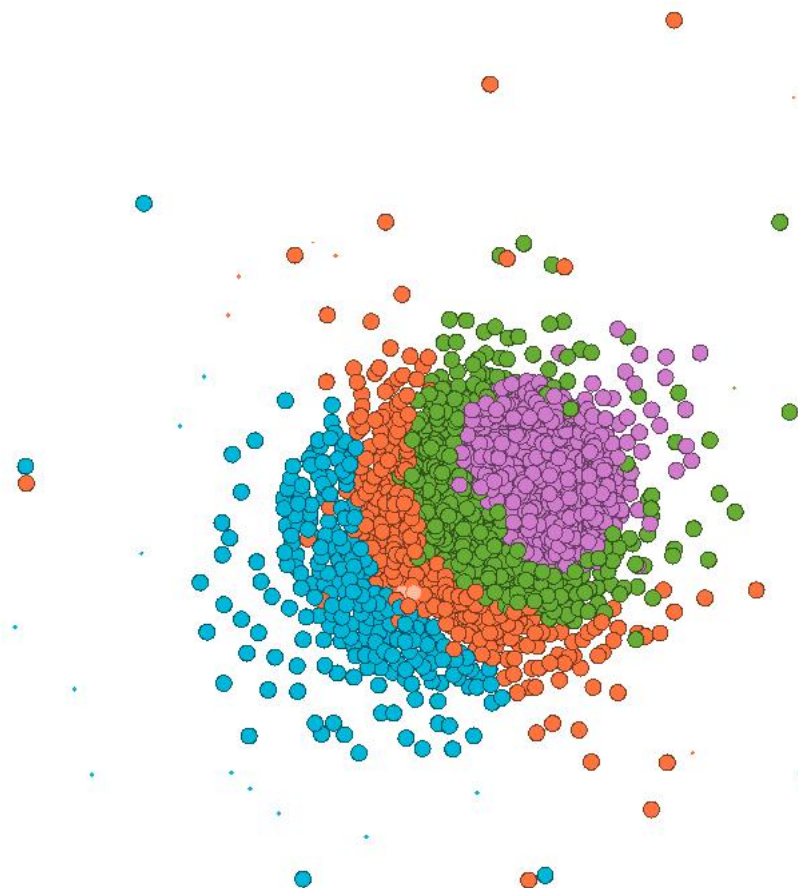
- 半小时重合（1:00-2:00, 1:30-2:30, 2:00-3:00）

- 其中：时间轨迹的相似度

- C_i 和 C_i' 分别代表用户1 的位置 $SeqSim(seq) = \frac{1}{1 + distance(C_i, C_i')}$ 位置点

Part 2 项目情况

2.3 社群划分 —— 基于兴趣相似度的社群划分



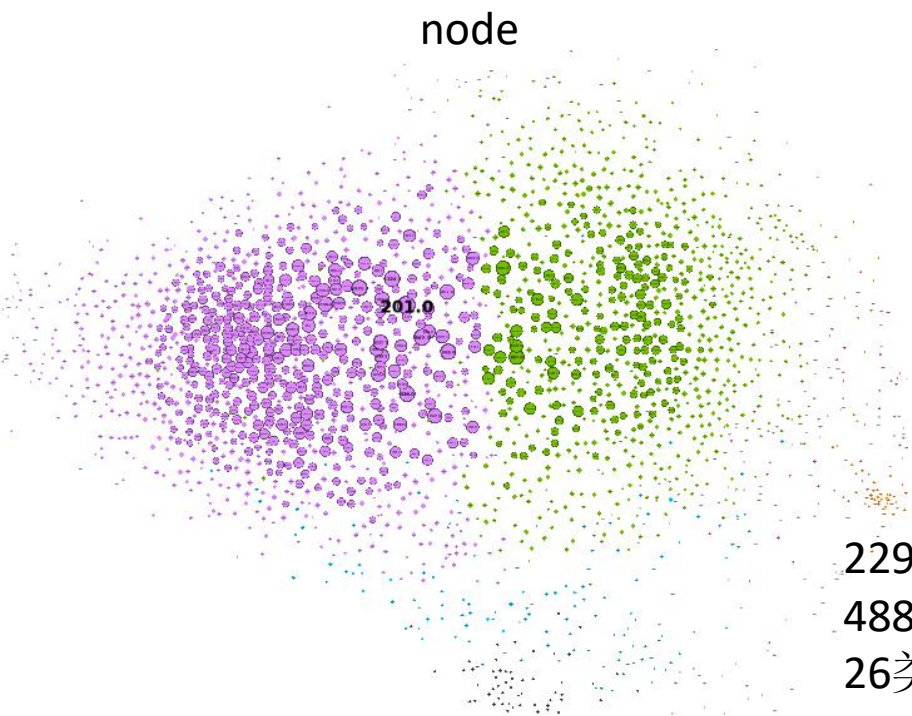
Part 2 项目情况

2.3 社群划分 —— 基于轨迹相似度的社群划分

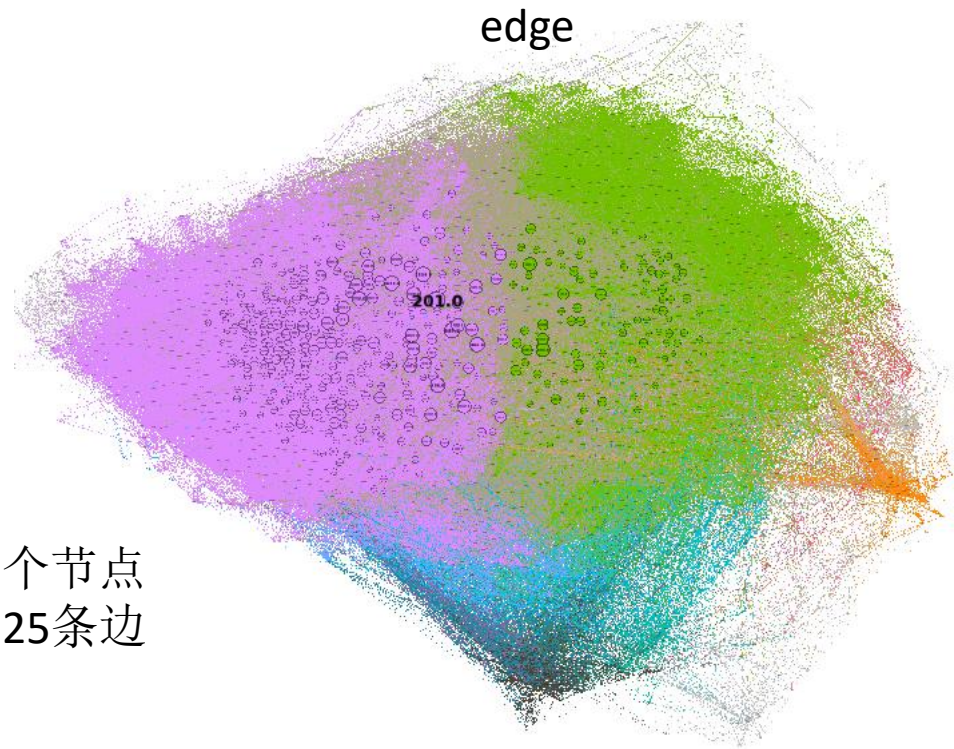
上面标有 userId

node

edge



2293个节点
488725条边
26类



Part 2 项目情况

2.4 时空轨迹

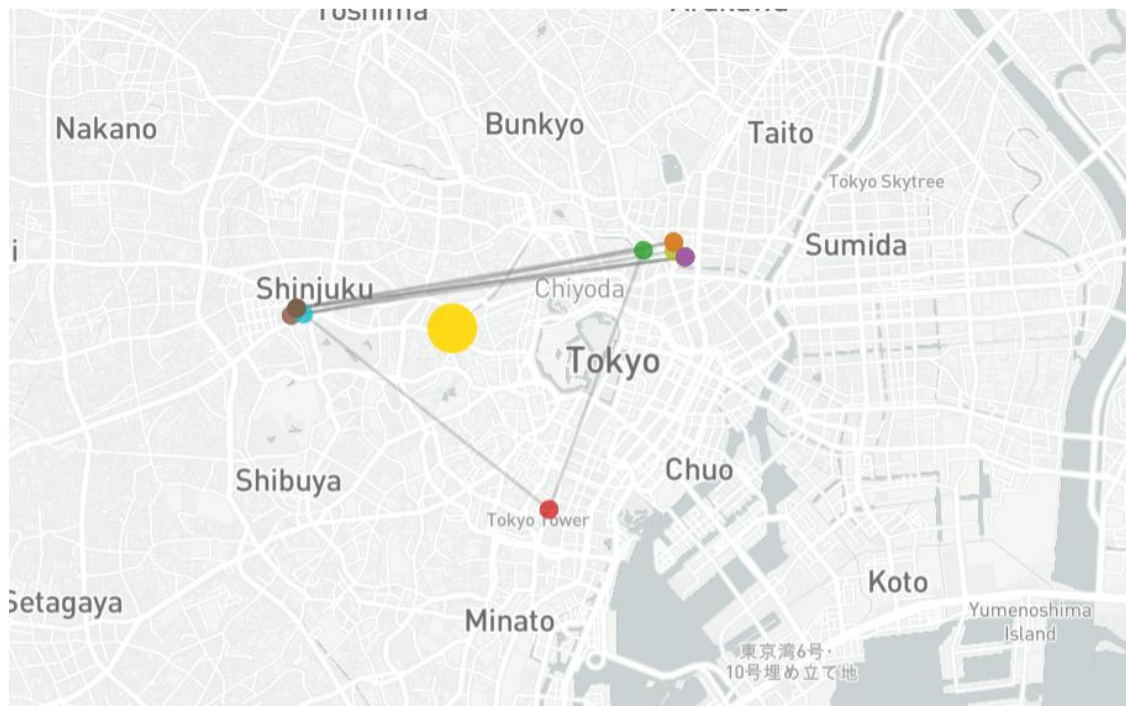
Trace of Time & Space

userID

201

中心节点：正三角形

第一类：

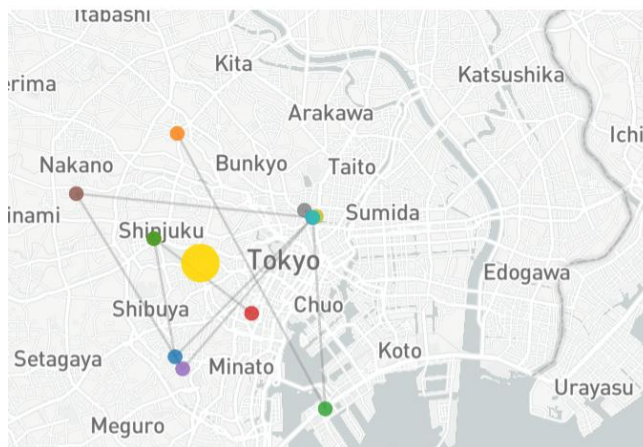


- lines
- 1 Train Station
- 2 Train Station
- 3 Beer Garden
- 4 Train Station
- 5 Train Station
- 6 Train Station
- 7 Train Station
- 8 Hobby Shop
- 9 Record Shop
- 10 Chinese Restaurant
- 11 Chinese Restaurant
- 12 Train Station
- 13 Train Station
- 14 Train Station
- 22 Train Station
- center

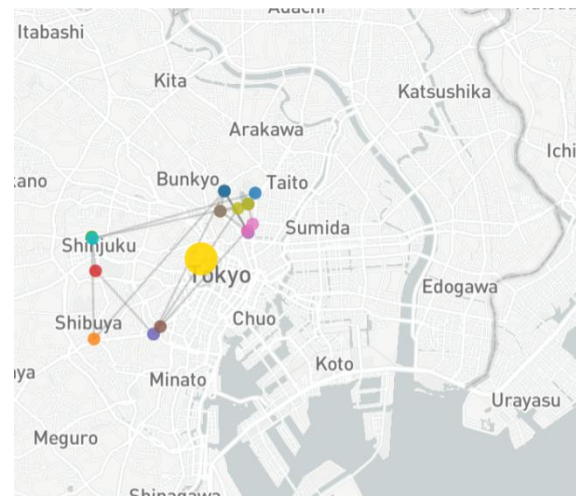
Part 2 项目情况

2.4 时空轨迹

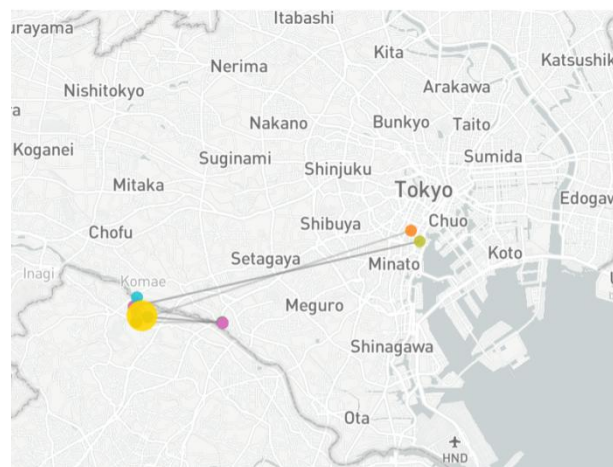
第一类:



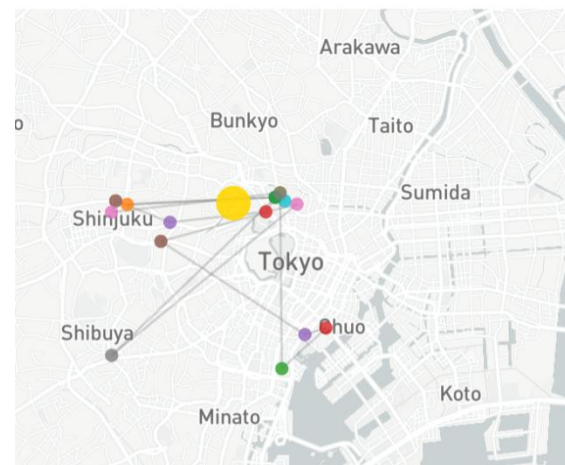
- lines
- 0 Train Station
 - 1 Train Station
 - 3 Train Station
 - 4 Restaurant
 - 5 Convention Center
 - 6 Train Station
 - 7 Ramen / Noodle House
 - 8 Electronics Store
 - 9 Train Station
 - 10 Train Station
 - 11 Train Station
 - 12 Train Station
 - 13 Subway
 - center



- lines
- 0 Subway
 - 1 University
 - 2 Office
 - 3 Train Station
 - 4 University
 - 5 Office
 - 6 Subway
 - 7 Shrine
 - 8 Building
 - 9 University
 - 10 Train Station
 - 11 Bar
 - 12 Train Station
 - 13 Building
 - 14 Subway
 - 15 Bar
 - 16 Video Store
 - 17 Video Store
 - 18 Music Venue
 - 20 Train Station
 - center



- lines
- 0 Government Building
 - 1 Mall
 - 2 Sporting Goods Shop
 - 3 Sporting Goods Shop
 - 4 Government Building
 - 5 Sporting Goods Shop
 - 6 Government Building
 - 7 Train Station
 - 8 Government Building
 - 9 Government Building
 - 10 Government Building
 - 11 Convenience Store
 - 12 Government Building
 - 13 Government Building
 - 14 Government Building
 - 15 Government Building
 - 16 Chinese Restaurant
 - 20 Government Building
 - 21 Athletic & Sport
 - 22 Government Building
 - 23 Cafe
 - center

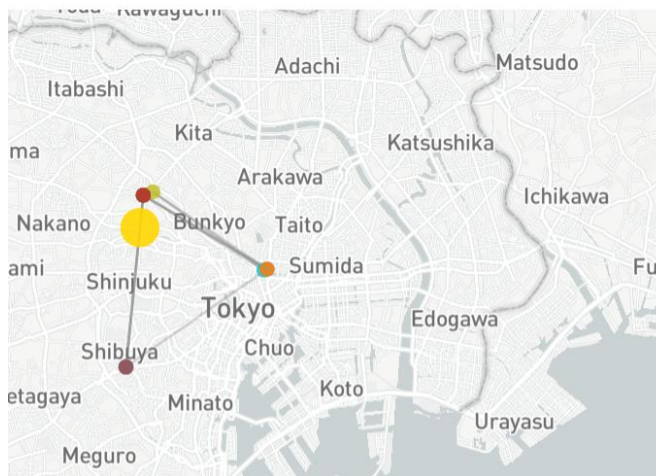


- lines
- 0 University
 - 1 Light Rail
 - 2 Sushi Restaurant
 - 3 Event Space
 - 4 Burger Joint
 - 5 College Academic Building
 - 6 Mall
 - 7 University
 - 8 Japanese Restaurant
 - 9 College Academic Building
 - 10 Coffee Shop
 - 11 College Academic Building
 - 12 Subway
 - 13 Food & Drink Shop
 - 14 Department Store
 - 17 Ramen / Noodle House
 - 23 University
 - center

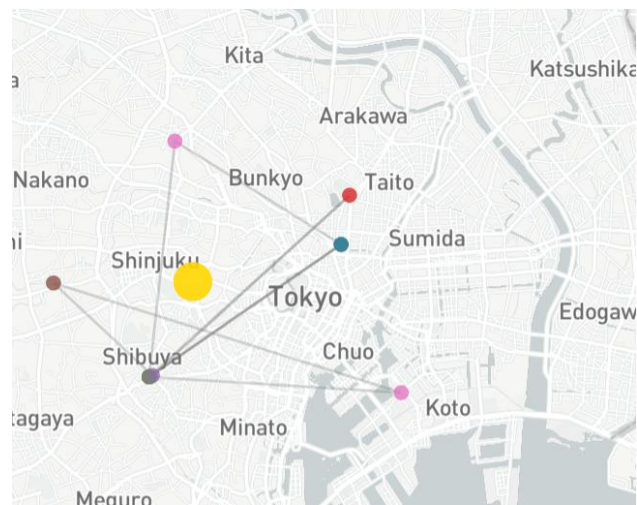
Part 2 项目情况

2.4 时空轨迹

第二类:



- lines
- 0 Train Station
- 1 Train Station
- 2 Train Station
- 3 Train Station
- 4 Train Station
- 5 Train Station
- 6 Electronics Store
- 7 Hobby Shop
- 8 Train Station
- 9 Train Station
- 10 Electronics Store
- 11 Train Station
- 12 Train Station
- 13 Train Station
- 14 Train Station
- center

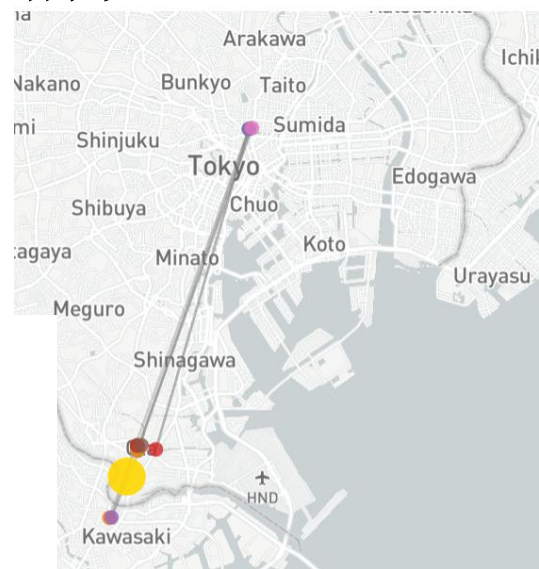
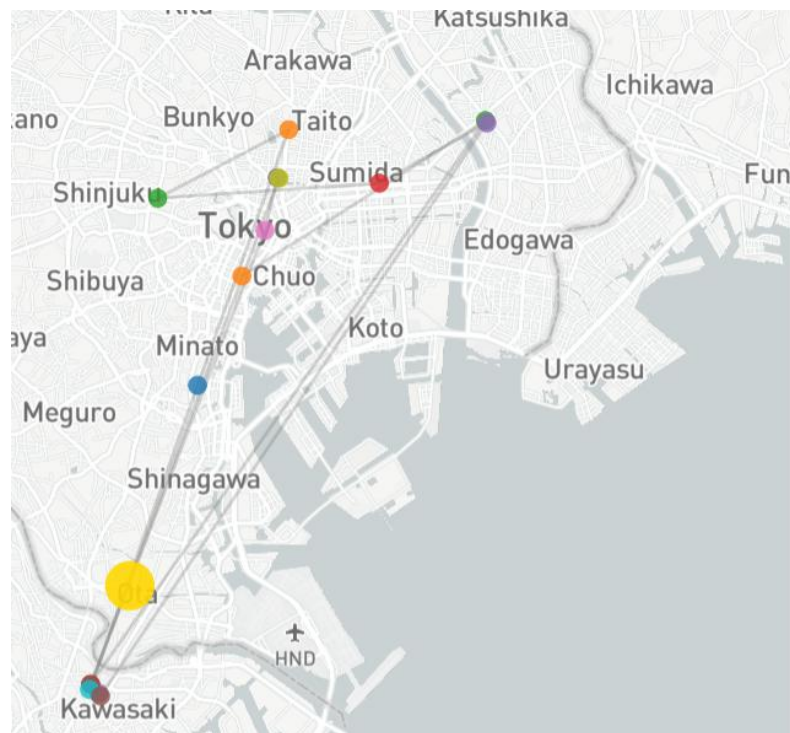


- lines
- 1 Train Station
- 2 Train Station
- 3 Train Station
- 4 Train Station
- 5 Train Station
- 6 Train Station
- 7 Train Station
- 8 Train Station
- 9 Train Station
- 10 Train Station
- 11 Train Station
- 12 Train Station
- 13 Train Station
- 14 Subway
- 15 Home (private)
- 21 Light Rail
- 22 Train Station
- center

Part 2 项目情况

2.4 时空轨迹

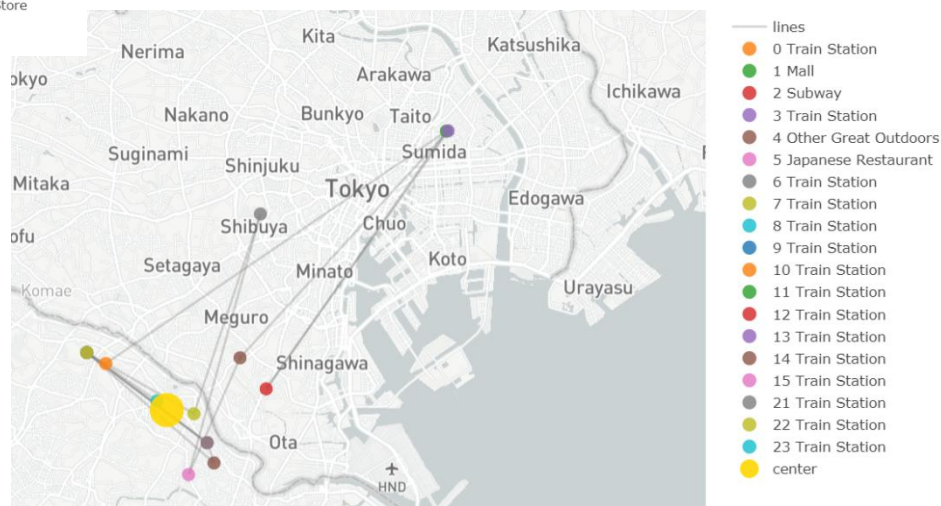
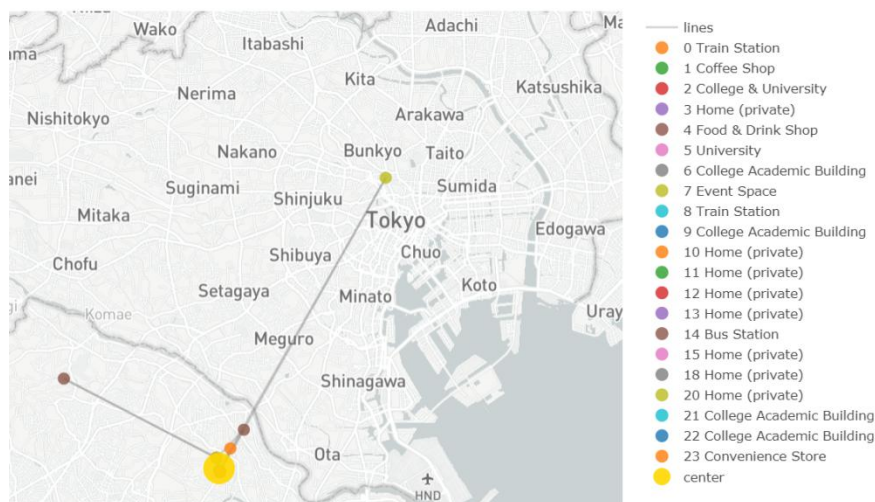
第三类:



Part 2 项目情况

2.4 时空轨迹

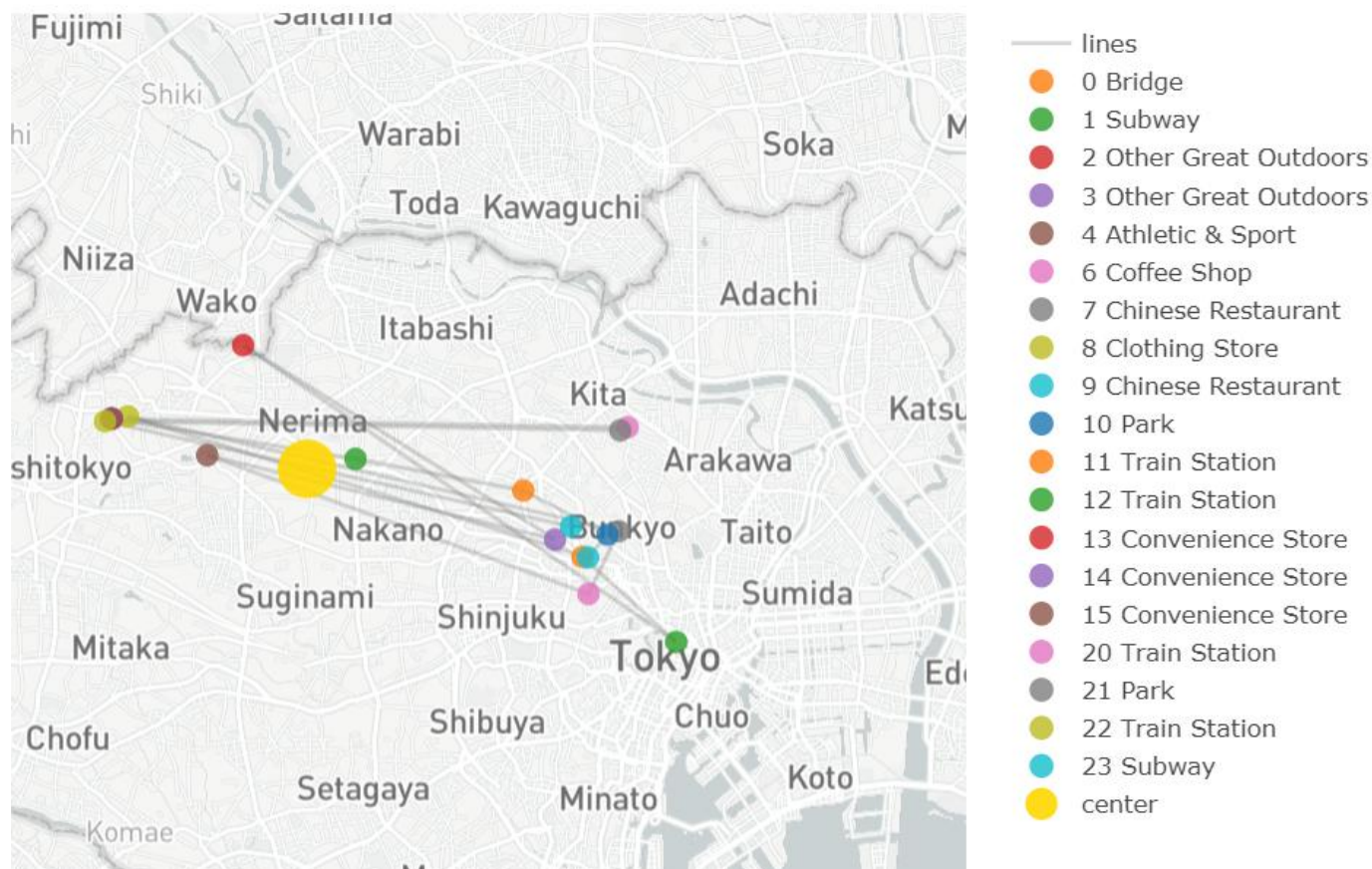
第四类:



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Tobler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

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

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

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。

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

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： $\langle t_1, l_1 \rangle, \langle t_2, l_2 \rangle, \dots, \langle t_n, l_n \rangle$ 。

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 ΔT 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 。

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

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

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

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

