数据处理：

1、出租车数据：

1、gps限制：

对原始出租车数据集进行gps限制，只保留puck-up点和drop-off点同时处于manhatan岛的出租车行驶记录，得到jan.csv, feb.csv,..., jun.csv六个月的数据集。

2、时间划分

对gps划分后的数据集进行时间片分割，目前的划分规则为，可能要调整：

```
timeSlot0 = []  # 22-7       凌晨
timeSlot1 = []  # 7-9        早高峰
timeSlot2 = []  # 9-12       早班时间
timeSlot3 = []  # 12-14      午高峰
timeSlot4 = []  # 14-17      午班时间
timeSlot5 = []  # 17-19      晚高峰
timeSlot6 = []  # 19-22
```

划分依据：

1、An Effective Taxi Recommender System Based on a Spatiotemporal Factor Analysis Model

2、An Energy-Efficient Mobile Recommender System

3、FUSING GEOGRAPHIC INFORMATION INTO LATENT FACTOR MODEL FOR PICK-UP REGION RECOMMENDATION

4、MPE: a mobility pattern embedding model for predicting next locations

5、Profitable Taxi Travel Route Recommendation Based on Big Taxi Trajectory Data

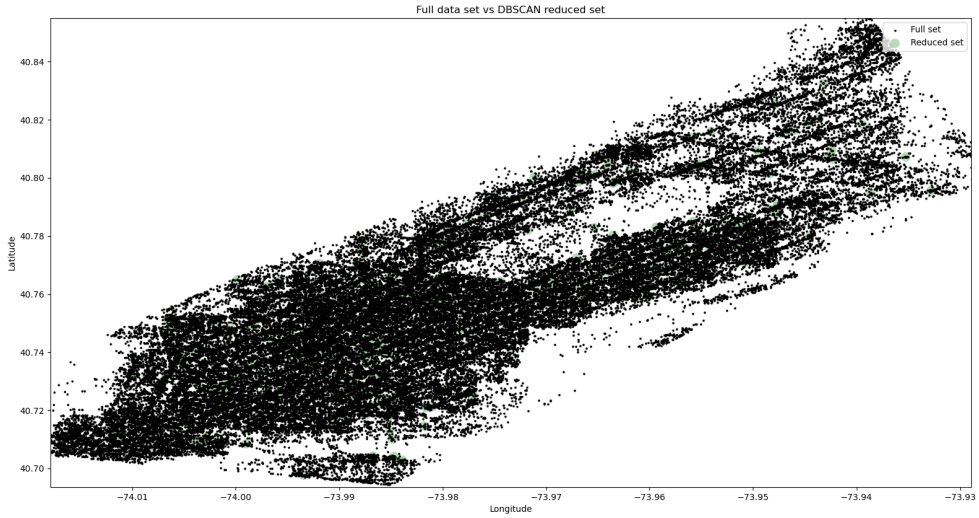
2、POI数据划分

1、gps限制

只保留gps坐标在曼哈顿岛上的POI点

2、防止过于密集，进行网格划分

即使经过DBScan过滤后分布依然密集：



于是取消DBScan的划分，按照网格进行划分，选取每个网格中最活跃的点，POI点缩减到4000，构图很快

实验一：测试评估指标

要得到六个评估指标，实验步骤如下：

- 分割出租车数据D（D1：D2，4：1）（train--test）
- 使用出租车、POI数据构图（train--test）
- 嵌入得到表征结果（R1：R2，train--test）
- 如何对比ground truth和得到的表征结果?

选取ground truth思路：

目前的做法是：

1、出租车数据D本身可以当作Ground Truth，使用D1，D2产生的表征R1，R2产生的**推荐**与D本身的**记录**对比，得到评估参数，那划分数据D的意义在哪里?

2、选取出租车D的一部分（50%）当作Ground Truth，使用剩余的50%数据学习表征R1，R2。用其产生的**推荐**与D本身的**记录**对比，得到评估参数

目前流程使用的出租车数据D就是某个月一个时间片所有数据，比如：2月份所有接载记录在7点-9点之间的行驶记录。

遇到的问题：

1、OpenNE工具包费了好长时间，实验室电脑没有GPU，但代替方案已解决

2、因为出租车数据很多，分为不同时间片，构图费了不少时间

3、选取ground truth思路不清晰

4、代码管理习惯不好，很多之前写好的代码，现在又要重新理解当时写的是什么意思

5、操作很麻烦，打算多跑一些数据集，在不同权重，不同思路下看一下结果。

实验二：自定义评估指标

1、划分出租车群体

实现了三个划分（按照driver）：

1、按照接载客数量划分Top/Bottom15%driver

2、按照实际收入划分Top/Bottom15%driver

3、划分网格收入高低

2、问题：如何反向对比?

思路一（未完成）：

Ground Truth为Top/Bottom15%driver dataset D

再把D直接送入模型，用输出的推荐（A--->[H,K,E]），（C--->[R,Y,E]）评估收入，但如何对比?

比如：

对比方案1：

1、从2月份所有接载记录在7点-9点之间的行驶记录选出TOP15%数据为D1，Bottom15%的数据为D2.

2、将数据D1输入模型，得到基于不同的起点推荐n次（A--->[H,K,E]），（C--->[R,Y,E]），假设都选择去推荐列表的首选地点，分别计算收入并求和S1.

3、从D2中挑选以推荐为起点的真实记录，计算所有对应收入并求和S2。

4、对比S1，S2，说明本模型带来收入提升

问题：

要对多个时间片，多个群体进行处理，操作有点麻烦

补充1：嵌入参数设置：

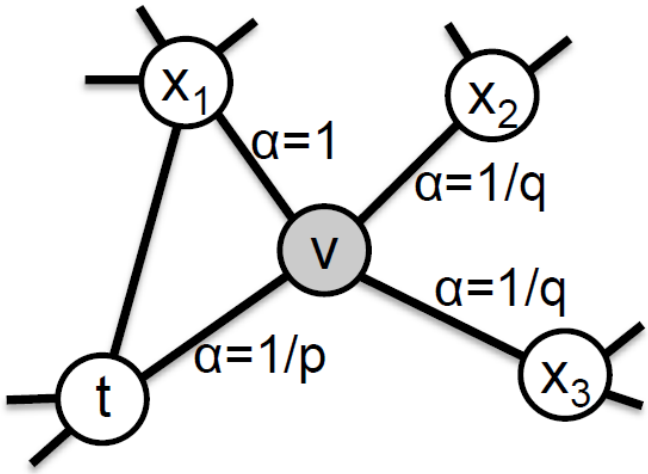


Figure 2: Illustration of the random walk procedure in *node2vec*. The walk just transitioned from t to v and is now evaluating its next step out of node v. Edge labels indicate search biases α.

直观地看，参数p和q控制行走探索和离开起始节点u附近的速度。特别是，参数允许我们的搜索过程(近似地)在BFS和DFS之间插入。

返回参数(return parameter)，p。参数p控制了在遍历中重新访问一个节点的可能性。将它设置为一个高值(> max(q, 1)确保我们不太可能在以下两个步骤中采样一个已经访问过的节点(除非遍历中的下一个节点没有其他邻居)。这种策略鼓励适度的探索，避免了采样中的两跳冗余。另一方面，如果p低(< min(q，1))，它将导致walk回溯一个步骤(图2)，这将使walk接近起始节点u。

向内向外参数(In-out parameter)，q。参数q允许搜索区分"向内"和"向外"节点。回到图2，如果 q > 1，则随机游走偏向于靠近节点t的节点(偏向BFS)。这样的遍历获得了底层图相对于遍历中的起始节点的局部视图，以及近似的BFS行为，因为我们的样本包含了一个小区域内的节点。而当q < 1时，walk更倾向于访问距离t较远的节点(偏向DFS)。因此，采样节点与给定源节点u的距离不是严格递增的，但反过来，我们受益于可处理的预处理和随机游走的优越采样效率。请注意，通过将πv,xπv,x设为t中前边节点的函数，随机游走是2阶马尔可夫过程。

补充2、具有参考意义的论文中使用的自定义指标：

论文一：A Balanced Assignment Mechanism for Online

来源：2017 IEEE 18th International Conference on Mobile Data Managemen

可引点：详细讨论乘客等待时间

论文二：A Cost-Effective Recommender System for Taxi Drivers

来源：kdd14

可引点：讨论grid profit，可做实验

论文三：An Effective Taxi Recommender System Based on a Spatiotemporal Factor Analysis Model

来源：2014 International Conference on Computing, Networking and Communications, Mobile Computing & Vehicle Communications Symposium

可引点1：每个小时划分为一个时间片更好

> In this paper, we divide a day into 24 time slots of equal length (i.e., the length of a time slot is an hour). The number of divisions affects the performance (i.e., total revenue). In our experiments, we conclude that a higher number of divisions (for example, hourly time slots rather than morning, noon, or afternoon sections) will achieve better performance.

可引点2：对比收入的方法：对比训练/测试集在不同时间的收入曲线

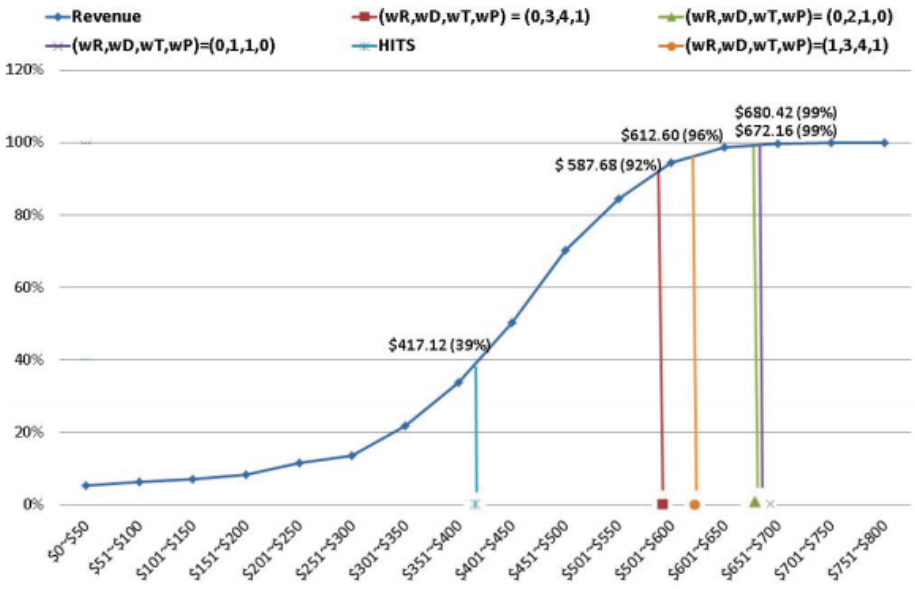


Fig. 1. The cumulative distribution function of revenue using the training data set on weekdays

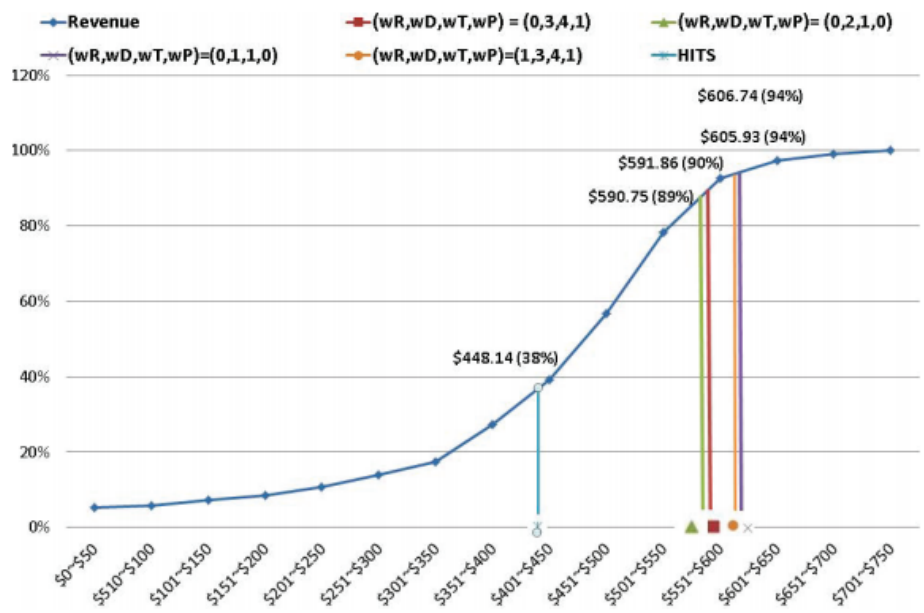


Fig. 2. The cumulative distribution function of revenue using the testing data on weekdays

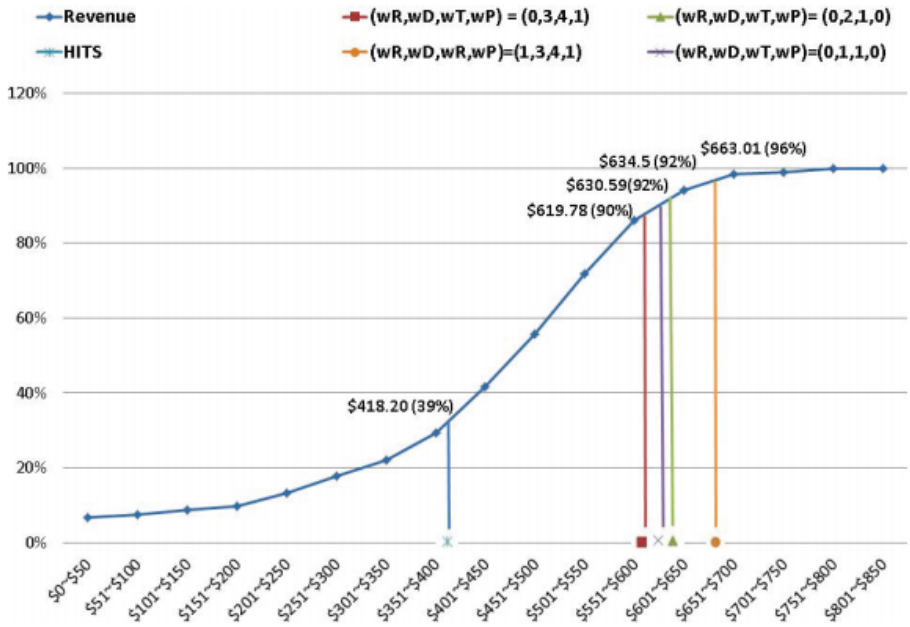


Fig. 3. The cumulative distribution function of revenue using the training data set on weekends

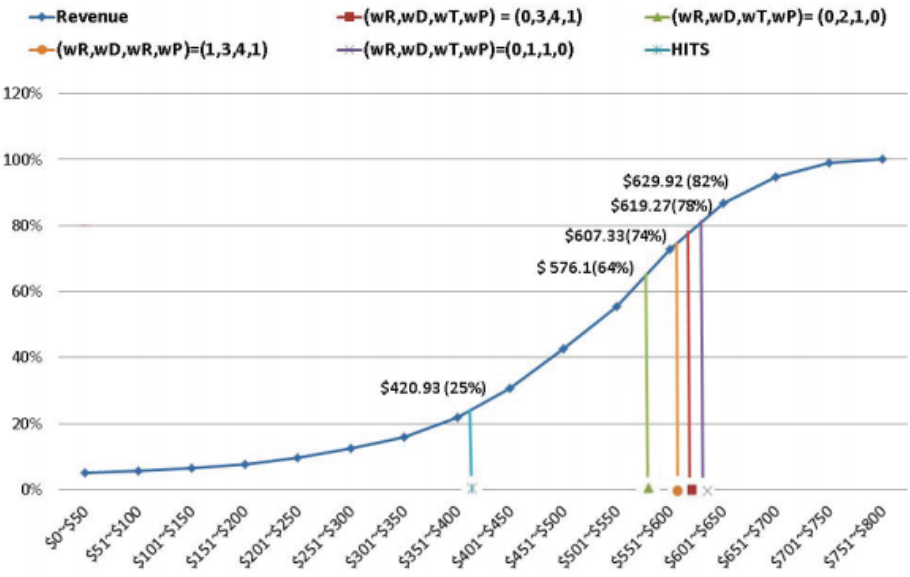


Fig. 4. The cumulative distribution function of revenue using the testing data set on weekends

论文四：An Energy-Efficient Mobile Recommender System

来源：KDD'10

可引点： we focus on two time periods: 2PM-3PM and 6PM-7PM.

论文五：FUSING GEOGRAPHIC INFORMATION INTO LATENT FACTOR MODEL FOR PICK-UP REGION RECOMMENDATION

来源：2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)

可引点： we divide the data into three datasets according to the time attribute. The data before 7, 12 and21 o'clock are used as the training data, and data between 7 to 9, 12 to 14, 21 to 23 o'clock are used as the test data.

论文六：MPE: a mobility pattern embedding model for predicting next locations

来源：World Wide Web 2017

Special Issue on Social Computing and Big Data Applications

可引点：时间片划分很细

> Before applying MPE to our data, we need to map the time-stamp of each record to the time slot it belongs to. We set the size of slot at 1, 5, 10, 15, 30, 60 and 120 minutes respectively and evaluate the performances.

论文七：Profitable Taxi Travel Route Recommendation Based on Big Taxi Trajectory Data

来源：IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS

可引点： one day is divided into 48 time intervals of half an hour