一、TRec

TRec: A Taxi Recommender System for Finding Passengers via Deep Neural Networks
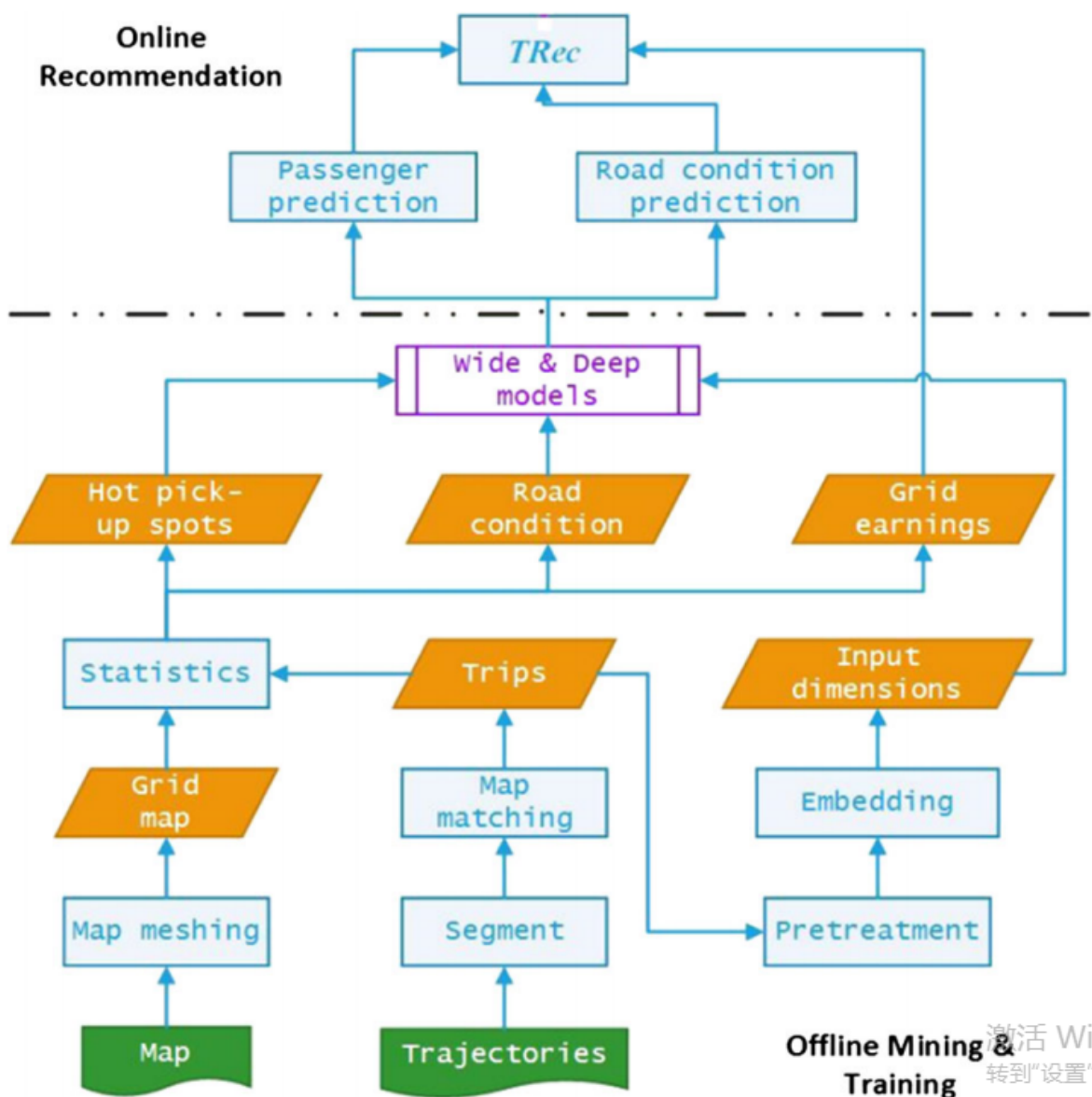
1.方法：

1、使用word2vec对轨迹进行表征学习

2、基于最大利润

3、避免出租车聚集

3、使用上海10000出租车2个月真实数据

问题目标：

目前车在网格O，考虑网格**Di的热度**，O到Di的路况以及**Di的平均收入**，并为每个网格生成最终得分以表示选择度。 最后，将选择**得分最高的网格Di**并将其推荐给出租车司机

2.框架：

(1) Offline mining and training

(2) Online recommendation.

1、地图网格划分

2、 GPS轨迹分割

3、 通过1，2的信息，得到三类信息：

1、接送点

2、道路信息

3、每个网格的收益

4、词嵌入

5、降维

最终做三件事：

(1) The prediction model of passengers,
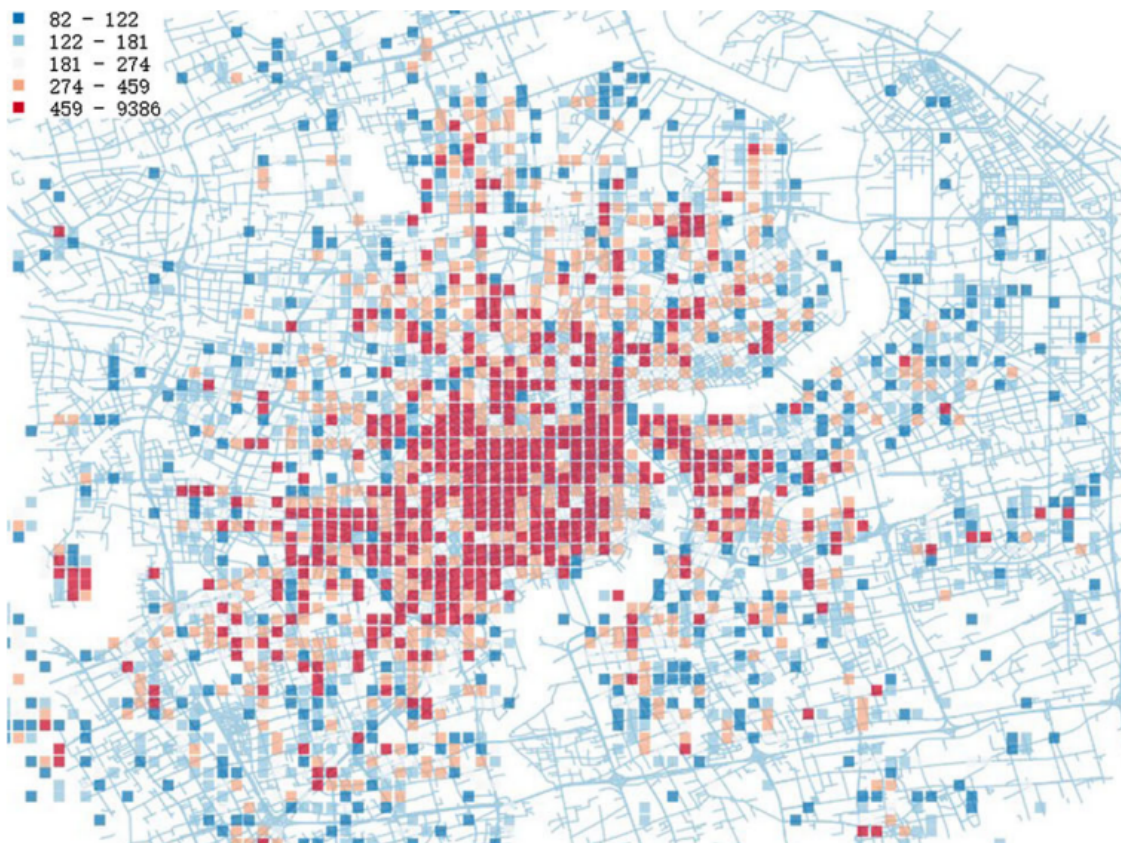
(2) The prediction model of road condition,
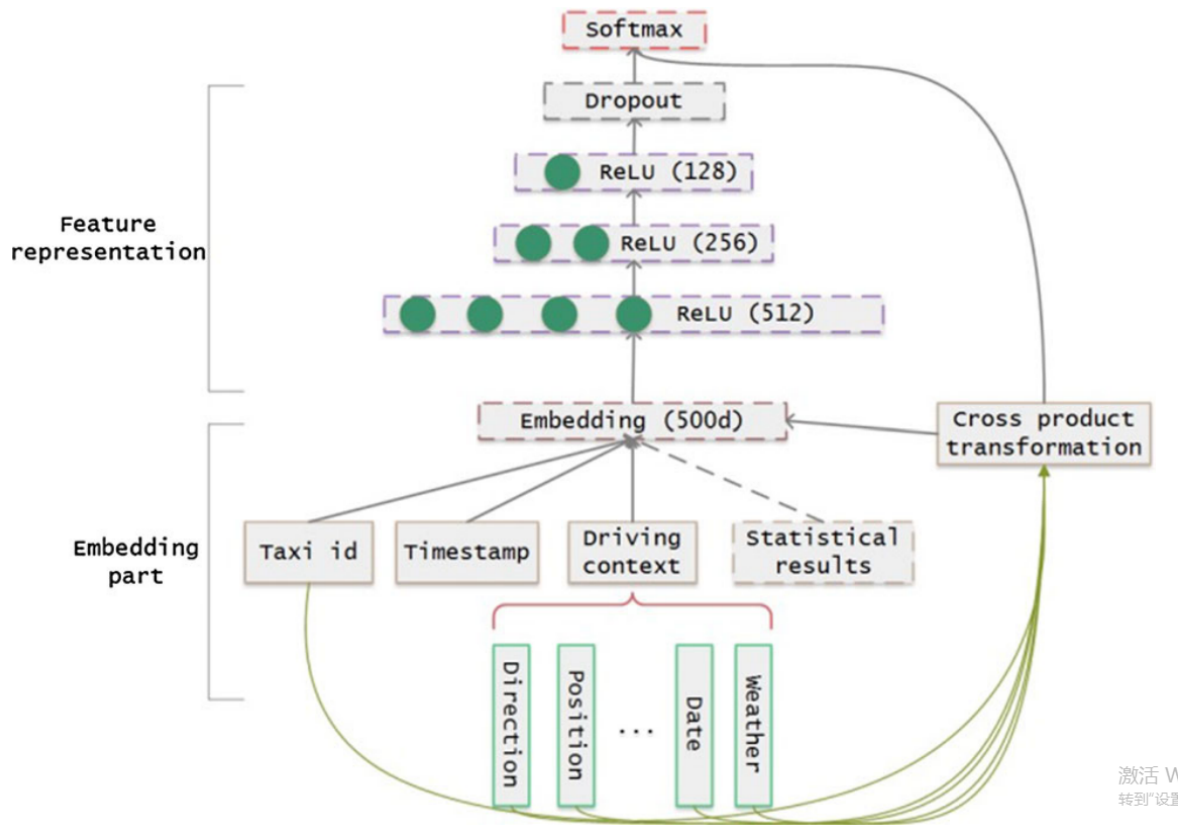
(3) The evaluation model of grid earnings.

3.地图分析

绘制了pick-up，drop-off网格划分图，300*300

4.Embedding

为每个属性值进行onehot编码，然后进行嵌入

5.预测

we select (taxi id, direction), (taxi id, date, weather) and (taxi id, direction, position) as crossed feature.

6.实验：

dataset：

we choose more than ten thousand taxis in the same company and collect their trips for about 2 months (June and July) in 2017.

再将trips分为两类：

(1) idle trips

(2) occupied trips.

Table 1 Embedding on input attributes of passenger prediction

Meta-data	Embedding dimension	Number of classes
Taxi id	10	6012
Original grids	10	19,865
Timestamp	10	240
Directions	5	9
Weather	5	3
Day of week	5	7

we pick the top 70% of drivers as the experienced ones. And the formula used involves miles/driving time and could evaluate the effectiveness of drivers.

7.performance

Table 2 Performance evaluation for passenger prediction

Model	Accuracy (%)
Wide & deep	64.1
MLP	48.2
DNN	61.3
XGBoost	23.1

Table 5 Performace evaluation for road condition prediction

Model	Accuracy (%)
Wide & deep	77.6
DNN	73.2
SVM	71.1
Random forest	72.3

(b)

Fig. 12 Average trip length originates from different girds. **a** Week-day. **b** Weekend

8.Trec的代码实现：

https://github.com/guangxush/Trec

Requirement

Python 2.7/3.5
Keras 2.x
sklearn
pandas
numpy
Tensorflow

Run

代码运行方式如下：

- 训练过程：

```
python cab_embedding.py  train
```

- 测试过程：

```
python cab_embedding.py  test
```

测试结果

模型/方法	Tranin Acc	Dev Acc	备注说明
mlp	0.628	0.000	[512, 256, 128] neurons
XGBoost	0.211	0.000	max_depth: 40, eta: 0.1, silent: 0
RF	0.117	0.000	max_depth: 30, n_estimators: 10,min_samples_split: 2
SVM	---	0.000	C: 1.0, kernel: rbf, degree: 3

文件组织方式

文件名称	文件描述
data文件夹	存放处理好的训练集数据及测试集数据
logs文件夹	存放训练日志
models文件夹	存放训练好的模型
raw_data文件夹	存放原始16年3、4月份的出租车数据
util文件夹	存放数据预处理的代码
README.md	项目描述