轨迹预测暑期项目

阶段性报告：基于 Markov 链的轨迹预测算法

刘前　　liuqian14@mails.tsinghua.edu.cn

2017 年 7 月 6 日

近两周尝试使用 Markov 链对用户的轨迹进行建模与预测。主要工作是：

1. 分析数据，获取**轨迹数据的基本分布**（轨迹长度和时间分布）；

2. 实现基于 Markov 链的轨迹预测算法:

a) 使用**经纬度**数据进行预测，尝试一些 trick(降低数据精度，选择较长的轨迹等)，观察对预测效果的影响；

b) 使用 **POI** 数据进行预测，使用同样的 trick 并观察预测结果。

3. 在 2 的基础上，考虑时间分布，对轨迹点较长的用户轨迹进行**时间分片**，得到尽可能多的有效轨迹，提高预测效果。

1　数据分析

1.1　数据基本情况

The number of users: 59199
The number of trajectories: 500175
The length of the longest trajectory: 738

表 1: 数据基本情况

使用的数据是 “tweets.txt”，包含了用户的 ID、轨迹点的经纬度、时间以及 POI 等信息。表1数据共包含 59199 名用户、500175 个轨迹点。从平均意义上讲，**平均每名用户不到 10 个轨迹点**，轨迹点过少对于轨迹预测必然是不利的。

1.2　轨迹长度分布

为了得到轨迹长度的具体分布情况，可以继续分析得到表2：

轨迹长度最大值	轨迹数量	所占比例
1	22253	0.375902
21	54229	0.916046
41	56846	0.960253
61	57822	0.976739
81	58310	0.984983
101	58581	0.989561
141	58873	0.994493
181	59010	0.996807
201	59057	0.997601
301	59154	0.999240
401	59176	0.999611
501	59186	0.999780
601	59191	0.999865
701	59198	0.999983
741	59199	1.000000

表 2: 轨迹长度分布

发现有 22000 多名用户只有 1 个轨迹点，是无法使用 Markov 链进行预测的。另外，**91% 以上的用户数据点少于 20 个** (而且大部分数据点数少于 10，见图1)。因而可以得到，大部分的用户轨迹数据其实并不完善，能够预见到使用基于 Markov 链的预测方法效果不会特别好。

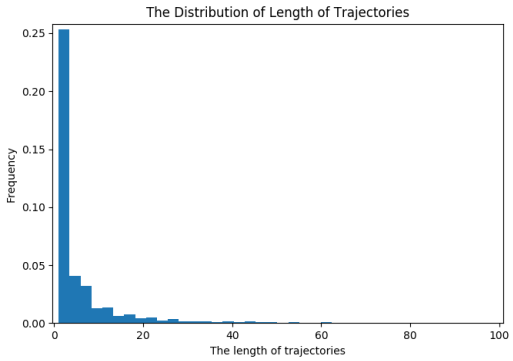

图 1: 轨迹长度的分布情况 (100 以内)

为了得到较好的预测效果，后面会增加一些参数，尝试使用一些 trick，并得到各参数对预测准确率的影响。

2 基于 Markov 链的轨迹预测算法

2.1 问题描述

轨迹预测的主要目的是：对未来用户的位置点进行预测。轨迹数据中的轨迹点对应于 Markov 链的状态。设已有用户的 n 个轨迹点，对应于 k 个位置，现在需要根据前 n 个轨迹点预测得到下一次用户会处于 k 个位置中的哪一个。

2.2 基本思路

基于 Markov 链的轨迹预测算法主要思路是：根据之前的轨迹点，统计得到转移概率矩阵 P，并基于转移概率矩阵，对当前位置点的下一位置进行概率的计算，选择概率最大的位置点作为预测结果。

2.3 问题描述

设对于一个轨迹 T，包含 n 个轨迹点，对应于 k 个地理位置，则转移概率矩阵是一个 k×k 的矩阵，记为 P。设矩阵 P 的第 i 行第 j 列为 p_{ij}，表示位置 i 到位置 j 的概率，可以通过使用历史轨迹数据统计得到。

在已有的真实数据中，统计位置 i 转到位置 j 的次数，用 N_{ij} 表示，则：

$$p_{ij} = \frac{N_{ij}}{\sum_{j=1}^{k} N_{ij}} \quad 1 \le i, j \le k \tag{1}$$

由此就可以得到一步转移概率矩阵 P。

虽然基于得到的一步转移概率矩阵已经可以对未来位置进行预测，但没有充分利用历史轨迹点带来的信息，因而可以保留 h 个历史轨迹点，使用 **h 步转移概率矩阵**进行预测，则：

$$P(h) = P^h \tag{2}$$

假设保留 h 步的历史轨迹，h 步之前的轨迹点认为对未来位置的影响可以忽略不计。因而可以基于 Markov 链，使用加权的方式，时间越久远的数据权重越小，最后从概率向量中选择概率最大的位置作为预测结果，如果概率相同，则预测结果返回一个向量。

此处需要说明，在计算预测正确率时，只要**真实位置存在于预测向量中，则认为预测结果正确**。

3 算法实现及结果

3.1 使用经纬度进行预测

算法实现时使用的 “tweets.txt” 数据中，只有 293559 条数据含有准确的经纬度信息，其余数据可能在经纬度数据采集时出现了问题。因而，如何使用轨迹点的经纬度信息进行预测，只能**使用这 29 万多条数据**。

轨迹长度和经纬度的精度会对预测的效果产生较大的影响，尝试引入了两个参数，分别表示所使用的轨迹最短长度和经纬度的小数位数。表3展示了不同参数下的用户平均预测准确率: 第一列

表示使用轨迹的最小长度，第一行表示经纬度的小数位数 (给的数据小数位数为 4)。

	4	3	2	1
1	15.26%	19.96%	30.22%	44.80%
10	23.34%	29.24%	38.65%	59.73%
20	25.78%	30.80%	38.32%	61.12%
50	28.28%	30.74%	38.32%	63.61%

表 3: 轨迹预测准确率：基于经纬度

3.2 使用 POI 进行预测

“tweets.txt” 数据中包含了 POI 信息，每个位置对应于唯一的 POI。使用 POI 替代经纬度，使用基于 Markov 链的预测算法，得到表4所示的预测结果。

最小轨迹长度	1	10	20	50
准确率	14.4%	25.8%	27.7%	27.0%

表 4: 轨迹预测准确率：基于 POI