doi:10.11887/j.cn.201403023 http://journal.nudt.edu.cn

依据历史轨迹构建城市出租车移动概率模型*

马武彬,刘明星,黄宏斌,邓 苏
(国防科技大学 信息系统工程重点实验室,湖南 长沙 410073)

摘 要:针对无法在线实时获取移动出租车实时状态信息的条件下,根据对历史轨迹信息的处理分析,提出将隐马尔科夫理论应用到城市出租车移动轨迹模型中,通过实际数据的分析建立出租车运动模型,通过对模型的计算来预测节点的位置分布概率,并在此模型上针对不同的用户需求进行查询处理,为用户提供搭车路线决策支持。通过利用真实数据集的实验证明,模型能够较好的模拟出出租车节点的运动状态,用户也能够从模型中获取较高精度的位置状态信息。

关键词:城市计算;出租车移动模型;隐马尔科夫;移动模型
中图分类号: **文献标志码**: A **文章编号**:1001-2486(2014)03-0129-06

The moving probability model of urban cabs based on history trajectory

MA Wubin, *LIU Mingxing*, *HUANG Hongbin*, *DENG Su*
(Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China)

Abstract: The model of urban cabs moving is one of the key issues for model building. The model needs reflecting moving state information of cabs. More importantly, users can quickly query the moving cabs. Under the condition of real time information of cabs which is hard to obtain, we should model the moving points and forecast the state information according to the moving history. A method which applies Hidden Markov theory to model of the moving trajectory is proposed. Through an analysis of real trajectory data of San Francisco, the caps moving model which is used to query the caps by users was constructed. Experiment with real datasets shows that the method proposed can simulate the moving state of caps. Users can also quickly obtain the useful location information from the model.

Key words: urban computation; cabs' moving model; hidden-markov; moving model

在城市交通日益拥堵的今天,寻找可以搭载的出租车成为人们出行的急切需求。传统的基于位置的服务,如 Google Latitude、Loopt 等,让我们能够获得大量关于出租车的移动时空数据,位置服务系统在线的对出租车实时观测,并将出租车的位置、状态信息反应给用户,能够给用户的出行搭车方案提供重要的参考。然而,在实际情况中,由于网络负载过大或者用户网络环境不好等原因,这些数据有时无法实时获取。同时,目前基于位置的服务系统也没有准确地获取到某辆出租车上是否已经有乘客的信息,这个信息对于出租车所在的位置状态来讲是隐性的。为了给用户能够更加方便的寻找出租车,需要建立有效的出租车运动历史数据的移动模型,通过数据分析为用户提供决策支持。同时,出租车作为一种社会中典型的移动载体,对出租车运动历史数据的分析还有利于社会网络中的人群聚集情况研究以及兴趣图分布情况研究。

例如图1所示,某一用户在景山公园内想要寻找一个可以搭车的地方。红色圆形区域代表用户所在区域,用户有四个方向的选择,可以去东南西北四个搭车点进行搭车。此时,用户希望能够在较短时间内获取一个具有较大概率存在空车的搭车点信息,指导他的下一步行动。本文正是希望能够通过大量的出租历史数据分析,提出一种方法,能够快速地给出用户在某个特定时刻的合理搭车地点信息。

通过对多个出租车数据集进行分析发现,出租车在长期的运行过程中,在某些特定时段,位于某些特定的空间状态存在着一定的规律,本文通过隐马尔科夫模型来对出租车的移动轨迹进行建模,将出租车本身空间状态的转移作为显式状态,

* **收稿日期**:2013-08-30
基金项目:国家高技术研究发展计划(863计划)资助项目(2011AA010106)
作者简介:马武彬(1986—),男,重庆北碚人,博士研究生,E-mail:mawubin417@gmail.com;
邓苏(通信作者),男,教授,博士,博士生导师,E-mail:dengsu@nudt.edu.cn

这些信息可以根据位置服务系统直接获取数据，将出租车是否有乘客作为隐式状态，在无法全部观测的情况下，来估计出租车在位置状态下“空车”或者是“有人”的概率。

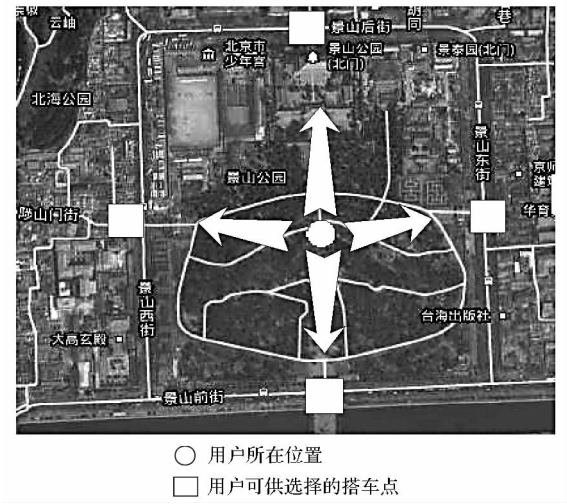


图1 用户搭车点选择示意图

Fig. 1 Sketch map of location selection of taking taxi

本文试图为出租车用户解决以下三类问题：

1)在 t 时刻，对于第 i 个出租车位于位置状态 k 的概率。

2)考虑隐状态，在 t 时刻，第 i 个出租车位于位置状态 k 并且无人的概率。

3)考虑隐状态，在 t 时刻，在位置状态 i 存在任意一个空车对象的概率。

这几个问题也都是用户在城市中寻找出租车比较关注的问题。另外，对于出租车显式状态和隐式状态的模型分析对于出租车的运行方向也有较大的指导意义，出租车可以根据模型的分析结果，选择那些有较大概率出现为“有人”的空间位置状态行走。同时，从社会网络的角度出发，出租车运行的隐马尔科夫模型对于社会节点之间的关系分析也具有重要的意义，模型能够计算出在某些特定的时刻人群经常聚集的区域，通过对模型稳态的求解，结合空间状态的信息，得出社会网络研究中“兴趣区域”的概率分布。

1 相关工作

目前许多学者对于出租车的节点运动分析做了大量的工作。数据分析的目的主要有两个：一是为了让出租车能够减少没有必要的空跑以及乘客的等待时间，达到资源节约和优化配置的目的；二是从社会网路的角度，分析出租车的历史移动轨迹，挖掘人的行为规律特征，为社会网路的研究提供数据基础。

Yuan 等人[1-2]提出了一种基于云计算的出租车实时轨迹规划路线指导推荐系统，该系统假设能够实时的收集出租车的位置信息，通过历史数据以及收集到的信息计算出出租车的最优运动路线，并返回给出租车。由于目前许多城市没有健全出租车的位置信息实时采集系统，所以该系统还不能在实际中取得很好的应用。文献[3]提出了为出租车司机以及乘客提供服务的推荐系统，针对司机，建立下一个可能遇到乘客的概率模型来指导司机的运行方向。同理，该文还建立了针对乘客的出租车可能到来的概率，评估每一个路段的平均等待时间，供用户参考。

Amorim[4]等人针对出租车运行的历史数据进行了数据分析，建立了更加切合实际的出租车移动轨迹模型，将出租车的载人和空车考虑到了轨迹模型中，然而该文却没有给出如何在该模型上进行更深一步的数据挖掘和处理。

Zheng [5]将出租车的历史运行数据的分析结果运用到交通规划中，从北京市出租车的移动轨迹历史信息中发掘出城市公共交通中需要改进的路段，从而指导城市交通道路的规划和建设。同时，Zheng[6]等人还提出了一种基于 landmarks(路标)的出租车移动目标预测方法，路标是指对每一个道路片段进行标记，基于这些道路标志进行预测。Castro[7]等人还提出了一种基于出租车密度流的预测方法。

Ma[8]等人从共享出租车的运行轨迹角度出发，首先定义了共享轨迹问题，并针对大规模出租车运行轨迹数据进行动态查询。在此基础上，将出租车作为一种资源引入到资源的调度问题中进行研究，为出租车提供最优化运行路径的服务。另外，通过移动轨迹数据来显示城市交通情况的文献还有文献[9-10]。

出租车节点作为社会网络节点中的代表之一，也有学者根据出租车的移动历史信息分析其社会性质，Peng[11]等人针对上海出租车的移动历史轨迹进行分析，发现了上海市市民在周末主要利用出租车进行三个方面的社交活动：从工作地到家庭之间活动，在不同的商圈进行活动以及各种其他娱乐活动。Veloso[12]等人对葡萄牙 Lisbon 的出租车移动轨迹进行了数据分析，主要分析在不同时间段内出租车的位置分布，通过出租车在运行过程中的运行时间和空载时间来为出租车的运行方向提供指导。

Yuan[13]等人以发现城市中各种 POI(point of interests)为目标，分析北京市里 2010、2011 年中3

个月内 12000 辆出租车的运行轨迹。利用该数据集验证了文中提出的基于模型的 POI 发现识别方法。Zhang[6] 等人利用出租车移动轨迹数据去分析人们原始出发点之间的语义关系,该文利用杭州的出租车移动数据,首先评估出合理的出发点数据流,然后根据出发点分析出人们的出发社会行为与时间具有较强的关联性。例如购物行为的出发流(从车站到超市)多发生在早上,而迁移行为的出发流(从车站到车站)多发生在白天时间。

本文利用隐马尔科夫理论来对出租车移动轨迹进行建模,主要考虑利用历史数据来推测未来出租车在某个时段的位置信息,以概率的形式展现给用户。另外,在出租车的搭载乘客信息无法精确获取的情况下,本文以出租车的位置信息作为显式状态,将乘客信息以隐状态的形式处理,通过状态转移的思想来预测城市出租车交通状况。

2　问题描述与定义

首先,定义一个时间和空间序列状态,建立出租车对象的时空模型。

定义 1　时间状态序列。本文中的时间状态序列用 $T=\{t_0,t_1,t_2,\cdots,t_n\}$ 表示,其中 $t_i>t_j$, if $i>j$。时间序列对应实际数据集中收集轨迹点的每一个时刻。

定义 2　网格空间状态序列。用 $S=\{s_1,s_2,\cdots,s_m\}$ 来表示。网格状态空间序列对应物理世界中的地理位置,状态空间序列以网格的形式划分,网格状态空间之间不存在交集,每一个网格状态空间有自己的矩形区域,并且这些区域的大小一样。

随着时间序列的推移,出租车在网格状态空间之间互相转换,任意一个时刻的出租车都有一个状态空间与之相对应。

然后,定义出租车对象以及在时空模型中的移动轨迹模型。

定义 3　出租车对象集合。用 $C=\{c_1,c_2,\cdots,c_p\}$ 来表示。在本文中,每一个出租车对应一个对象,在时间序列的约束下,出租车对象在网格状态空间中不停地转换。

显然,存在一个映射关系,在任意时刻 t_i,对象 c_j 所处的状态为 $s_t(c_j)\in S$。

定义 4　出租车轨迹。出租车轨迹 T_{ca} 为一系列带有时间顺序的 GPS 点集组合而成的序列值,每一个点都包括位置信息和时间信息,形式化定义如下,对于具有 n 个点的出租车轨迹 $T_{ca}:P_1\rightarrow P_2\cdots\rightarrow P_n$,其中 P_i 代表出租车的某一个位置点(px,py,t), $\forall k\in[0,n]$, $(P_{k+1}.t-P_k.t)=\Delta t,\Delta t>0$。$\Delta t$ 表示取出租车位置值的间隔时间,假设对于某一个固定的出租车轨迹,间隔时间是一个大于 0 的固定值。

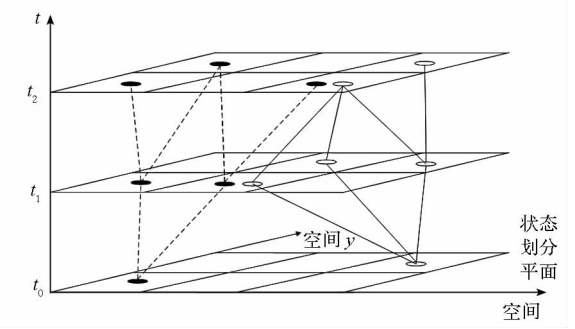


图 2　出租车在时间序列下的状态转移示例

Fig. 2　Transition of cab's location state in the time series

出租车在运行过程中,除了其所处的位置是显示状态意外,还存在一个隐性的状态,即是否有人在车上,对该隐式状态建立模型如下。

定义 5　隐式状态序列。用 $O=\{o_1,o_2,\cdots,o_m\}$ 表示。根据划分区域图,将出租车所处状态用地理位置空间网格来划分,每一个方格区域代表一个状态,状态变量用 S 表示,$S=\{s_1,\cdots,s_N\}$,每一辆出租车在这些状态中不停地转换,每一个状态具有特定的区域,用 $s_i.area$ 表示。

定义 6　轨迹状态函数。对于每一个轨迹序列,轨迹序列 $T_{ca}:P_1\rightarrow P_2\cdots\rightarrow P_n$ 映射到状态序列中的值,定义一个从 T_{ca} 到显式观测变量 Q 的函数,$Q=f(T_{ca})$,其中 $Q=(q_1,\cdots,q_n)$。$q_i=f(p_i)$,当且仅当 $\exists t\in T,p_i.point\in s_j.area$。

3　隐马尔科夫模型

3.1　模型定义

下面将建立出租车移动的隐马尔科夫模型。对于显示观测序列,状态变化对应第 2 节中定义 2 的空间状态变化。出租车在某个空间中的移动轨迹可以看作是出租车的显示状态不停的转移过程,状态转移矩阵定义为

$$A=[a_{ij}](i,j\in\{1,\cdots,n\}) \tag{1}$$

其中　$a_{ij}=P(q_t=s_j/q_{t-1}=s_i)$。

同样存在隐式状态序列,本文做一个假设,假设在一个 Δt 时刻之内,隐式状态是不变的。隐式状态的观测变量为 $V=\{v_1,v_2\}$,分别表示出租车处于有人和无人的观测状态。隐式观测矩阵定义为

$$B=[b_i(k)](i\in\{1,\cdots,n\},k\in\{0,1\}) \tag{2}$$

其中 $b_i(k)=P(x_t=v_k/q_t=s_i)$。

每一辆出租车有一个初始状态,记为 π,$\pi=[\pi_i](i\in\{1,\cdots,N\})$,$\pi_i=P(q_i=s_i)$,对应移动轨迹的初始点。

因此,对于城市出租车的隐马尔科夫模型可以表示为:(A,B,π),下节将讨论如何通过轨迹数据来得出模型中的参数估计值。

3.2 状态转移矩阵估计

利用轨迹数据集 $T_{ca}:P_1\rightarrow P_2\cdots\rightarrow P_n$,计算状态转移矩阵的估计值。状态转移矩阵的似然估计值为

$$\hat{a_{ij}}=P(q_t=s_j/q_{t-1}=s_i)=\frac{\sum_{t=1}^{n-1}\varphi(q_t,s_i)\times\varphi(q_{t+1},s_j)}{\sum_{t=1}^{n-1}\varphi(q_t,s_i)} \tag{3}$$

其中 $\varphi(x,y)$ 是定义的函数:

$$\varphi(x,y)=\begin{cases}1 & \text{if} \quad x=y\\ 0 & \text{if} \quad x\neq y\end{cases}$$

$\sum_{t=1}^{n-1}\varphi(q_t,s_i)\times\varphi(q_{t+1},s_j)$ 代表在 t 时刻位于状态 s_i,而在 $t+1$ 时刻位于状态 s_j 的次数。$\sum_{t=1}^{n-1}\varphi(q_t,s_i)$ 表示在 t 时刻位于状态 s_i 的总次数。

隐式状态的概率为

$$\hat{b_j(k)}=\frac{\sum_{t=1}^{n-1}\varphi(q_t,s_j)\times\varphi(o_t,v_k)}{\sum_{t=1}^{n-1}\varphi(q_t,s_j)},\quad k=0,1 \tag{4}$$

当 $k=0$ 时,$\sum_{t=1}^{n-1}\varphi(q_t,s_j)\times\varphi(o_t,v_0)$ 表示出租车位于状态 s_j 并且没有乘客的次数。相反,当 $k=1$ 时,通过上式可以计算出租车位于状态 s_j 并且有乘客的次数,$\sum_{t=1}^{n-1}\varphi(q_t,s_i)$ 表示在 t 时刻位于状态 s_i 的总次数。

在隐式和显式状态确定以后,当给定出租车的初始状态后,城市出租车的隐马尔科夫模型也就确定了。下面讨论如何利用该模型对上文提出的 3 个问题进行求解。

3.3 概率模型求解

显然,上文提出的问题 1 可以描述为:对于出租车 m 的隐马尔科夫模型 $\lambda=(A_m,B_m,\pi_m)$,通过轨迹数据可以计算 A_m,B_m,在给定初始状态 π_m 的情况下,在 t 时刻,计算出租车在状态 s_i 的概率为

$$P(q_t=s_i)=\pi_m A_m^{t-1}I_i,i=1,\cdots,N \tag{5}$$

其中 I_i 表示在第 i 位取值 1,其他位取值为 0 的 N 维基向量。

针对问题 2,在 t 时刻,出租车 m 在状态 s_i 并且无乘客的概率 $P(o_t[m]=v_i)$ 为

$$P(o_t[m]=v_i)=\pi_m A_m^{t-1}b_i[m],i=1,2 \tag{6}$$

对于问题 3,在城市交通的实际网络中,针对多辆出租车,假设出租车集合为 $C=\{c_1,\cdots,c_M\}$,并且每一辆出租车之间的运行轨迹是独立分布的,可以计算在时刻 t,状态 i 存在出租车并且没有乘客的概率为

$$P(T=t,o_i=v_1)=I-\prod_{m=1}^{M}P(T=t,o_i[m]=v_1)=I-\prod_{m=1}^{M}(I-\pi_m A_m^{t-1}I_i b_0[m]) \tag{7}$$

其中 $o_i[m]$ 表示第 m 辆出租车在 i 时刻的观测值。

4 实验分析

本文采用 San Francisco 的真实出租车数据集[14-15],利用隐马尔科夫模型对该出租车 154540 个移动轨迹点进行数据分析。对于 San Francisco 数据集,当取状态值为 4,即将区域划分为 2×2 的,可以得到该出租车的隐马尔科夫模型中的状态转移矩阵取值如表 1、表 2 所示。

表 1 状态转移矩阵

Tab. 1 State transition matrix

状态值	状态 1	状态 2	状态 3	状态 4
状态 1	0.998	0.001	0	0
状态 2	0.014	0.979	0.002	0.004
状态 3	0.143	0.142	0.714	0
状态 4	0	0.05	0	0.95

表 2 隐状态的概率估计值

Tab. 2 Probability estimates of hidden states

状态值	状态 1	状态 2	状态 3	状态 4
车号 BJ001	0.007	0.385	0.571	0.775
车号 BJ002	0.993	0.615	0.429	0.225

从本文建立的模型可以进行分析,根据式(5)~(7),可以对 3 个问题进行求解。然而,该例中空间的划分粒度太大,查询出来的区域给予用户的参考意义不大,而过细的空间划分粒度会带来巨大的时间复杂度,本实验分析不同状态集划分对计算性能的影响,首先验证对问题 1 和问题 3 的求解,用户查询某一辆出租车在某个时刻

处于某个状态的概率以及对某一个状态是否有空车进行查询,针对不同数据集,第 1 类问题的求解时间随状态集合的划分粒度变化情况如图 3。从计算过程来讲,问题 1 和问题 2 具有相似的特征见本文 3.3 节,因此对于问题 2 的时间性能就不再验证。

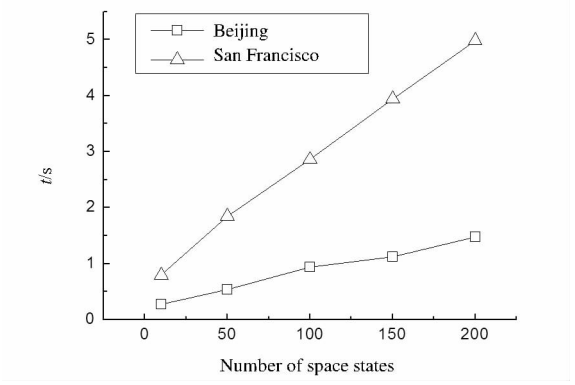


图 3　问题 1 求解时间随状态集合的划分粒度变化

Fig. 3　Computing time for problem 1 with the change of partition granularity

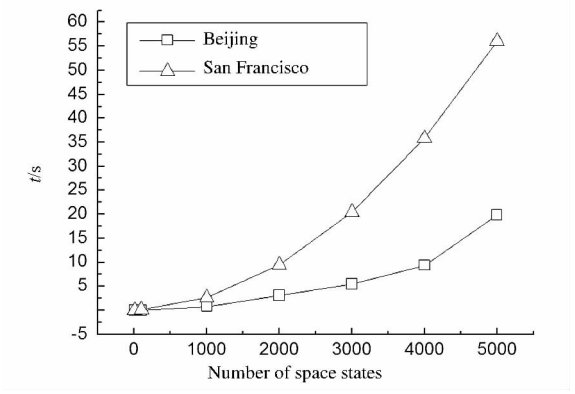


图 4　问题 3 求解时间随状态集合的划分粒度变化

Fig. 4　Computing time for problem 3 with the change of partition granularity

接下来本文针对 San Francisco 数据集,取状态数量为 100 个,验证模型求解方法的有效性。

问题 1、2 的求解可以看成是该辆出租车在整个状态空间上的存在概率分布计算。本实验取在 100 次用户查询结果的概率分布平均值。例如对出租车 2 的问题 1 的查询结果概率分布平均值如图 5 所示。

对于问题 2 的求解,用户需要知道在某个状态下出租车存在且是无人的概率情况。在考虑隐状态的情况下,问题 2 的平均查询结果计算情况,即在各个状态下出租车无人的概率分布如图 7 所示。

本实验还验证了本文方法对于问题 1 的预测误差随历史轨迹点数量的变化,将分布概率中的

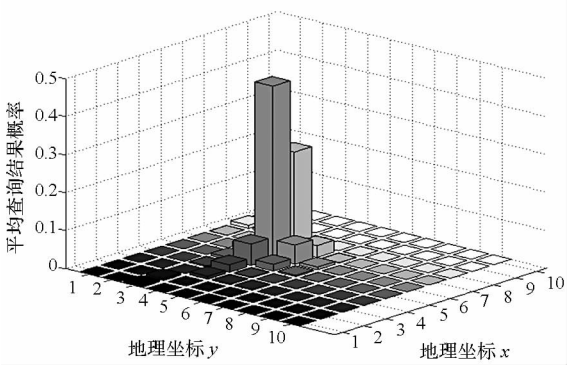


图 5　出租车 2 的问题 1 平均查询结果的概率分布

Fig. 5　Probability distribution of average query results for problem 1 with cabs 2

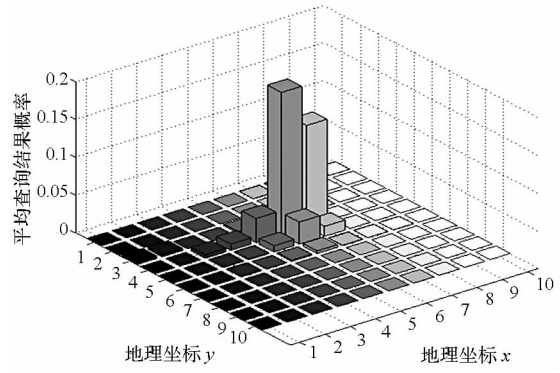


图 6　出租车 3 对于问题 2 平均查询结果的概率分布

Fig. 6　Probability distribution of average query results for problem 2 with cabs 3

最大值所在的空间状态作为出租车的预测运行目标,对比文献[6 -7, 16]所提出对出租车的下一个运行目标的预测方法误差。结果如图 7 所示。

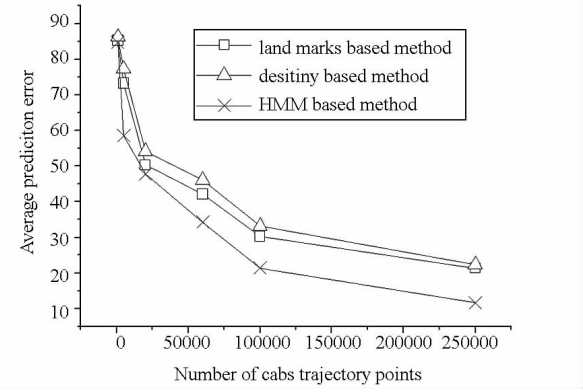


图 7　预测误差随着轨迹点数量的变化比较

Fig. 7　Prediction error with the change of number of trajectory points

从图 7 中可以看出,本文提出的基于 HMM 模型预测方法在轨迹点数量较多时预测误差较小,精度相对于其他两个方法较高。在轨迹点数量为 100 000 个以上时,平均预测误差提高比基于密度流和基于路标的预测算法分别减少

16.3%和12.4%。

对用户和社会网络分析者来讲,更有意义的是对整个城市的所有出租车运行轨迹的运算分析结果。也就是对于问题3的求解。利用式(5),计算San Francisco数据集中的所有出租车问题3的解,得出平均概率分布如图8.

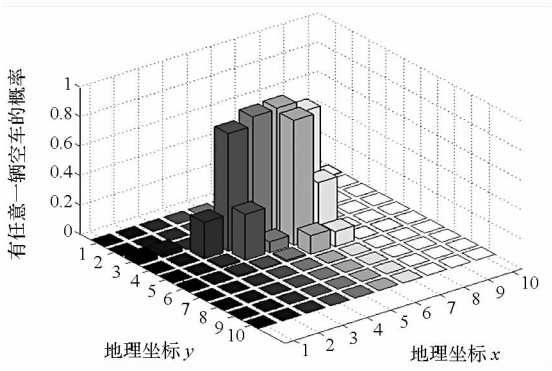


图8 问题3平均查询结果概率分布

Fig. 8 Probability distribution of average query results for problem 3

从图8中可以看出,对于San Francisco而言,整个城市的存在空出租车概率较大的空间网格区域集中在7~8个状态中,用户在这些区域里面能够很容易的寻找空的出租车。同时,从社会网络的角度来看,这些区域往往是人们密集集中的区域。

至此,利用真实数据集,解决前文提出的3个问题的相关概率分布。在实际应用中,例如用户处于状态55,根据图8中的结果,用户搭车路线应该选择位置状态45的方向,而不是向周围的54、56等方向移动。用户可以根据这些信息选择向合适的方向移动,有较大概率能够搭到出租车。

5 总结与展望

本文利用隐马尔科夫理论对出租车移动轨迹数据进行分析,得出在城市中搭载出租车难易程度的概率分布图,利用量化的概率模型来表达出租车在某一特定时刻的分布情况,为乘客出行方向和出租车司机的运行方向提供决策支持。在未来的工作中,由于概率分布的计算时间会随着状态变量的增加而指数级别的增加,而整个城市中的有效状态占据的比例不高,需要进行状态清理。因此,将隐马尔科夫理论模型进行改进,使之适合应用到大量状态的城市出租车数据分析是未来研究的重点。

参考文献(References)

[1] Yuan J, Zheng Y, Xie X, et al. Driving with knowledge from the physical world[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011: 316-324.

[2] Yuan J, Zheng Y, Zhang C Y, et al. An interactive-voting based map matching algorithm [C]//Proceedings of the Eleventh International Conference on Mobile Data Management, 2010: 43-52.

[3] Yuan N J, Zheng Y, Zhang L H, et al. T-Finder: A recommender system for finding passengers and vacant taxis[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(10): 2390-2403.

[4] Amorim A M, Campos J. A conceptual model for representation of taxi trajectories[C]//Proceedings of the 8th Symposium on GEOINFO, Campos do Jordao, 2012: 96-107.

[5] Zheng Y, Liu Y C, Yuan J, et al. Urban computing with taxicabs[C]//Proceedings of the 13th International Conference on Ubiquitous Computing, 2011: 89-98.

[6] Zhang W S, Li S J, Pan G. Mining the semantics of origin-destination flows using taxi traces [C]//Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012: 943-949.

[7] Castro P S, Zhang D Q, Li S J. Urban traffic modeling and prediction using large scale taxi GPS traces [J]. Pervasive Computing, Lecture Notes in Computer Science, 2012,7319: 57-72.

[8] Pergolizzi J V, Gharibo C G, Gudin J A, et al. Development of federally mandated risk evaluation and mitigation strategies (REMS) for transmucosal immediate-release fentanyl products[J]. Pain Practice, 2013, 13(4): 259-263.

[9] Ban X G, Gruteser M. Towards fine-grained urban traffic knowledge extraction using mobile sensing[C]//Proceedings of the ACM SIGKDD International Workshop on Urban Computing, 2012: 111-117.

[10] Castro P S, Zhang D Q, Li S J. Urban traffic modeling and prediction using large scale taxi Gps traces [J]. Pervasive Computing, LNCS, 2012,7319: 57-72.

[11] Peng C B, Jin X G, Wong K C, et al. Collective human mobility pattern from taxi trips in urban area[J]. PLoS ONE, 2012, 7 (8): 34-48.

[12] Veloso M, Phithakkitnukoon S, Bento C. Urban mobility study using taxi traces [C]//Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis, 2011: 23-30.

[13] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs [C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012: 186-194.

[14] Piorkowski M, Sarafijanovoc-Djukic N, Grossglauser M. A parsimonious model of mobile partitioned networks with clustering [C]//Proceedings of the First International Workshops on Communication Systems and Networks, 2009: 1-10.

[15] Zheng Y, Zhang L Z, Xie X, et al. Mining interesting locations and travel sequences from GPS trajectories[C]// Proceedings of the 18th International Conference on World Wide Web, 2009: 791-800.

[16] Xue A Y, Zhang R, Zheng Y, et al. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction[C]//Proceedings of the 2013 IEEE International Conference on Data Engineering, 2013: 254-265.