软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(11):3043−3057 [doi: 10.13328/j.cnki.jos.005340]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于前缀投影技术的大规模轨迹预测模型*

乔少杰[1], 韩 楠[2], 李天瑞[3], 李荣华[4], 李斌勇[1], 王晓腾[3], Louis Alberto GUTIERREZ[5]

[1](成都信息工程大学 网络空间安全学院,四川 成都 610225)

[2](成都信息工程大学 管理学院,四川 成都 610103)

[3](西南交通大学 信息科学与技术学院,四川 成都 611756)

[4](深圳大学 计算机与软件学院,广东 深圳 518060)

[5](Department of Computer Science, Rensselaer Polytechnic Institute, New York, USA)

通讯作者: 韩楠, E-mail: hannan@cuit.edu.cn

摘 要: 智能手机、车载 GPS 终端、可穿戴设备产生了海量的轨迹数据,这些数据不仅描述了移动对象的历史轨迹,而且精确地反映出移动对象的运动特点.已有轨迹预测方法的不足在于:不能同时兼具预测的准确性和时效性,有效的轨迹预测受限于路网等局部空间范围,无法处理复杂、大规模位置数据.为了解决上述问题,针对海量移动对象轨迹数据,结合频繁序列模式发现的思想,提出了基于前缀投影技术的轨迹预测模型 PPTP(prefix projection based trajectory prediction model),包含两个关键步骤:(1) 挖掘频繁轨迹模式,构造投影数据库并递归挖掘频繁前序轨迹模式;(2) 轨迹匹配,以不同频繁序列模式作为前缀增量式扩展生成频繁后序轨迹,将大于最小支持度阈值的最长连续轨迹作为结果输出.算法的优势在于:可以通过较短的频繁序列模式,增量式生成长轨迹模式;不会产生无用的候选轨迹,弥补频繁模式挖掘计算代价较高的不足.利用真实大规模轨迹数据进行多角度实验,表明 PPTP 轨迹预测算法具有较高的预测准确性,相对于 1 阶马尔可夫链预测算法,其平均预测准确率可以提升 39.8%.基于所提出的轨迹预测模型,开发了一个通用的轨迹预测系统,能够可视化输出完整的轨迹路线,为用户路径规划提供辅助决策支持.

关键词: 轨迹预测;前缀投影;频繁序列模式;轨迹匹配;马尔可夫链

中图法分类号: TP311

中文引用格式: 乔少杰,韩楠,李天瑞,李荣华,李斌勇,王晓腾,Gutierrez LA.基于前缀投影技术的大规模轨迹预测模型.软件学报,2017,28(11):3043−3057. http://www.jos.org.cn/1000-9825/5340.htm

英文引用格式: Qiao SJ, Han N, Li TR, Li RH, Li BY, Wang XT. Large-Scale trajectory prediction model based on prefix projection technique. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):3043−3057 (in Chinese). http://www.jos.org.cn/1000-9825/5340.htm

Large-Scale Trajectory Prediction Model Based on Prefix Projection Technique

QIAO Shao-Jie[1], HAN Nan[2], LI Tian-Rui[3], LI Rong-Hua[4], LI Bin-Yong[1], WANG Xiao-Teng[3], Louis Alberto GUTIERREZ[5]

[1](School of Cybersecurity, Chengdu University of Information Technology, Chengdu 610225, China)

[2](School of Management, Chengdu University of Information Technology, Chengdu 610103, China)

* 基金项目: 国家自然科学基金(61772091, 61100045, 61363037); 教育部人文社会科学研究规划基金(15YJAZH058); 教育部人文社会科学研究青年基金(14YJCZH046); 成都市软科学项目(2015-RK00-00059-ZF); 四川省教育厅资助科研项目(14ZB0458)

Foundation item: National Natural Science Foundation of China (61772091, 61100045, 61363037); Planning Foundation for Humanities and Social Sciences of the Ministry of Education of China (15YJAZH058); Youth Foundation for Humanities and Social Sciences of the Ministry of Education of China (14YJCZH046); Soft Science Foundation of Chengdu (2015-RK00-00059-ZF); Foundation of Educational Commission of Sichuan Province (14ZB0458)

本文由复杂环境下的机器学习研究专刊特约编辑张长水教授推荐.

收稿时间: 2017-04-07; 修改时间: 2017-06-16; 采用时间: 2017-08-23

[3](School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)
[4](College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China)
[5](Department of Computer Science, Rensselaer Polytechnic Institute, New York, USA)

Abstract: Smart phones, GPS equipped vehicles and wearable devices can generate a large number of trajectory data. These data can not only describe the historical trajectory of moving objects, but also accurately reflect the characteristics of moving objects. The existing trajectory prediction approaches have the following drawbacks: both prediction accuracy and efficiency cannot be guaranteed together, effective trajectory prediction is limited to road-network constrained local spatial areas, and complex and large-scale location data are difficult to process. Aiming to cope with the aforementioned problems, a prefix projection based trajectory prediction model targeting massive trajectory data of moving objects is proposed by employing the basic idea of frequent sequential patterns discovery. The new model, called PPTP (prefix projection based trajectory prediction model), includes two essential steps: (1) Discovering frequent trajectory patterns by creating projected databases and iteratively mining frequent prefix trajectory patterns from projected databases; (2) Trajectory matching by incrementally extending the postfix trajectory based on each frequent sequential pattern and outputting the longest continuous trajectory that is greater than the threshold of minimum support count. The advantages of the proposed algorithm are that it can generate long-term trajectory patterns via short frequent sequential patterns in an incremental manner, and it will not generate useless candidate trajectory sequences in order to overcome the drawback of time-intensive in discovering frequent sequential patterns. Extensive experiments are conducted on real large-scale trajectory data from multiple aspects, and the results show that PPTP algorithm has very high trajectory prediction accuracy when comparing to 1st-order Markov chain prediction algorithm and the average improvement of accuracy can reach to 39.8%. A generic trajectory prediction system is developed based on the proposed trajectory prediction model, and the complete prediction trajectories are visualized in order to provide assistance for users in path planning.

Key words: trajectory prediction; prefix projection; frequent sequential patterns; trajectory matching; Markov chain

随着位置大数据分析处理技术[1]的快速发展,离散的时空位置点中蕴藏的移动行为规律被人们挖掘和利用.不同于简单的位置签到数据,轨迹数据是连续和完整的具有时效性的运动个体移动规律和社交信息的展示,因此,理解并利用好时空轨迹数据具有实际意义.车载 GPS 等设备采集的轨迹数据同样具有较高价值,其中最直观的应用包括:

(1) 行驶路线规划.与以往单纯根据路网信息进行推荐的方法不同,当拥有海量基于路网的轨迹数据后,可以根据行驶经验预测出同一时刻不同路段的拥堵情况,并以此为依据,为驾驶员提供最便捷、最省时的行驶路线[2].

(2) 交通流预测.将单个个体的位置信息进行综合,可以得到群体的移动规律并加以利用,可以帮助交警部门及时掌握路况信息,合理进行警力安排,保证交通顺畅.

(3) 出租车调度.分析出租车的轨迹数据有助于了解其营运状态分布、空车常见区域等,并以此为基础,挖掘打车成功率较高的地点,推荐给有出行需求的用户.同样地,对于出租车驾驶员而言,通过对用户的轨迹进行预测分析,可以帮助其更加方便、准确地找到有乘车需求的客户[3].

“大规模移动对象轨迹预测模型研究”是结合智慧交通真实应用提出的新课题,是一项非常困难和富有挑战意义的课题.

- 首先,如何准确和高效地预测移动对象的连续运动轨迹,成为全新的课题.针对海量轨迹数据挖掘移动对象频繁轨迹模式,已有的算法需要多次扫描数据库,代价极高,需要设计新型频繁模式挖掘算法,提高挖掘的效率和准确性.而且,现有的轨迹预测方法很少考虑移动对象运动的复杂场景信息,如主观场景(移动速度和方向的影响)和客观场景(如交通状况、天气情况、出行高峰、季节等).
- 其次,大规模轨迹数据的预处理会极大地影响轨迹预测模型的性能.一条轨迹往往由成千上万个时空点构成,车载 GPS 设备每天可以采集上万条轨迹数据,其体量是大数据级别.轨迹数据具有稀疏性,需要通过数据降维、轨迹压缩、索引等技术对轨迹数据进行预处理,这样才能保证轨迹预测的时效性.
- 再者,基于模拟技术的预测方法依赖于大量输入参数,参数的设置会极大地影响模型的准确性.对具有诸多不确定性的轨迹数据进行预测,需要考虑诸多主客观因素和领域专家知识.而且通过定位系统获得的数据流信息量大,具有不确定性,需要更加稳定和具有可伸缩性的预测方法.

- 最后,位置预测的实时性是一个重要指标[4],需要尽可能快地对运动轨迹进行评价,延时或者滞后将产生无意义的预测结果.如果算法设计不合理,那么随着移动对象数目的增加,模型的计算代价可能呈指数级增长.

大规模移动对象轨迹预测的典型应用场景是智能交通,具体可以解决如下问题:每天哪一时段,具体在哪一区域是机动车出行的高峰期;采用何种手段可以使路网上的交通流量处于最佳状态,改善交通拥挤和阻塞状况,最大限度地提高交通的通行能力,提高整个公路运输系统的机动性、安全性和运输效率.

为了弥补现有轨迹预测方法的不足并提供针对大规模时空轨迹的高效预测模型,本文的主要贡献包括:

(1) 通过分析现有轨迹预测技术的不足,对比基于马尔可夫链的轨迹预测方法,指出其与基于频繁模式挖掘的轨迹预测方法的不同点.提出了前序、后序轨迹和投影数据库的概念,并给出了基于前缀投影技术的轨迹序列模式挖掘的相关性质和推论.

(2) 借鉴频繁序列模式挖掘算法基本思想,结合轨迹数据时空特性,提出一种新型基于前缀投影技术的增量式轨迹预测轨模型 PPTP(prefix projection based trajectory prediction model),通过详细示例介绍算法内部工作原理.

(3) 在真实大规模 GPS 轨迹数据集上进行大量实验,验证了算法的准确率、时间效率等指标.引入对比算法,全面而客观地比较不同轨迹算法的性能优劣.此外,设计并实现了基于 PPTP 算法的轨迹预测系统,对系统功能进行描述.

1 相关工作

轨迹数据存在于空间内,需采用定位等方式采集.常见的轨迹数据采集方式包括卫星定位、传统无线通信网络定位、互联网接入定位、WiFi 接入定位、RFID 定位等技术.常见的轨迹数据,如出租车行驶数据、公共自行车轨迹数据等都是应用 GPS 技术进行采集的.互联网接入定位主要依托网络服务提供商的数据,提供粗略的位置估计.相对于 GPS 定位方式,其定位精度大幅度下降.WiFi 接入定位适用于室内环境定位[5].为了弥补卫星定位和无线通信网络定位在室内定位的不足,WiFi 定位得到了广泛的关注[6].

受定位精度的影响,利用上述方式采集到的轨迹数据往往包含误差,因此需要对轨迹数据进行特征提取,包括地图匹配、校准、分段、化简、去噪等一系列操作[7].常用的轨迹特征提取方法包括贝叶斯滤波、卡尔曼滤波和粒子滤波等[8].针对位置大数据价值提取和挖掘问题,郭迟等人[1]从 3 个粒度层面综述了位置大数据的分析处理方法.Zheng[5]对轨迹数据挖掘技术系统、全面地阐述,介绍了轨迹数据预处理、轨迹数据挖掘及各种智能分析技术,指出针对轨迹大数据挖掘的难点,如数据预处理和优化问题.Dai 等人[9]利用情景感知信息减少移动对象运动的不确定性,对轨迹不确定性进行排序,提高了轨迹数据质量.

针对移动对象的连续位置预测包括:过去轨迹的重现、当前和未来轨迹的预测,主要研究集中于轨迹频繁模式挖掘,通过挖掘频繁模式找出典型运动路径.传统的轨迹预测算法通常假设对象运动是线性变化的,针对这一不足,Tao 等人[10]提出了一种监控和索引移动对象的架构,设计了频繁轨迹模式挖掘算法 STP-tree,算法可以有效地查询移动对象频繁运动轨迹.该方法的不足在于:设计的挖掘方法不具有普适性,对于真实的频繁轨迹查询架构的评价方法尚需进一步验证.Morzy[11]通过构建移动对象位置的概率模型,从移动规则中匹配对象动态运动轨迹.算法的优势在于充分利用了移动对象数据库的历史信息,具有较高的预测准确性.其不足体现在:当移动对象数据量较大时,建立索引的代价较高,计算量较大.Qiao 等人[2]提出了一种基于 FTP-tree(frequent trajectory pattern-tree)并集成双层轨迹索引和轨迹热点区域挖掘技术的高效和准确的轨迹预测算法,但是算法仍然无法解决频繁模式挖掘算法固有的计算代价较高的问题.

近些年来,利用数据挖掘和机器学习技术感知和预测移动对象行为的研究得到了学者的广泛关注.Song 等人[12]在《Science》上发表了一篇介绍如何预测人类移动行为的文章,通过测量个体轨迹的信息熵,定量地计算出了人类动态运动行为具有 93%的可预测性.MIT 的 Reality Mining 项目[13]通过分析不同国家、不同地域手机用户的位置数据,理解不同文化背景人群的社会经济状态、生活节奏、移动性、对社会突发事件的反应.Qiao

等人[4]借助隐马尔可夫模型设计实现了一种可以自适应调整轨迹预测参数的动态预测算法,根据不同类型轨迹数据预测最佳路线,但是这一模型没有考虑大规模轨迹时空数据下的运行时效性问题.Ding 等人[14]提出了一种路网匹配的基于轨迹数据库的交通流分析方法,用于预测移动对象的位置信息.乔少杰等人[15]针对移动对象的复杂多模式运动行为,利用高斯混合回归方法建模,计算不同运动模式的概率分布,利用高斯过程回归预测运动轨迹.Dai 等人[16]利用高斯混合模型描述行驶偏好矢量中随机变量的概率分布,并结合最短路径算法推荐个性化运动轨迹.Yuan 等人[17]提出了一种移动对象多粒度周期性活动发现模型,用于预测个体行为规律.文献[18]针对轨迹稀疏问题,通过计算轨迹熵值预测移动行为的整体趋势.Chaulwar 等人[19]提出了一种混合机器学习方法,用于解决复杂交通场景下的轨迹规划问题.

通过上述工作分析可知,现有研究工作中存在的突出问题是:(1) 在个体行为识别上一般是采用有监督的学习方法,现有的方法通常需要样本训练,计算开销较大,不适合大规模、海量的轨迹预测问题;(2) 现有的轨迹预测方法主要针对单一的简单运动模式,当轨迹模式复杂多变时,无法有效应对诸如轨迹预测算法的准确性、稳定性和可伸缩性的要求.而上述指标正是本文所要考虑的核心问题.

2 问题描述

大规模移动对象轨迹预测包含的主要步骤包括:首先,获取某个对象的大量历史轨迹数据,通过对轨迹数据的简化、抽象和聚集操作,建立轨迹预测模型,提取出移动对象的轨迹运动模式;然后,运用抽取的轨迹模式,借助各种轨迹预测技术和方法挖掘对象的运动趋势.

对于轨迹预测问题,常见的解决方法有基于马尔可夫模型的预测方法等,这一方法具有简单、高效等优势,分析基于马尔可夫链的方法对于认识本文提出的基于前缀投影技术的轨迹预测算法具有一定价值.首先对这一预测模型进行定量描述,如下所示.

定义 **1**(马尔可夫轨迹链). 设有轨迹随机过程$\{X_n,n\in T\}$,若对于任意整数 $n\in T$ 和任意位置点 $i_0,i_1,\ldots,i_{n+1}\in I$,条件概率满足 $P\{X_{n+1}=i_{n+1}|X_0=i_0,X_1=i_1,\ldots,X_n=i_n\}=P\{X_{n+1}=i_{n+1}|X_n=i_n\}$,则称$\{X_n,n\in T\}$为马尔可夫轨迹链.马尔可夫过程具有马尔可夫性质,强调对象的下一位置点仅与之前有限长度的位置点有关.当下一状态仅与当前位置有关时,称为一阶马尔可夫过程.

定义 1 中给出的马尔可夫轨迹链实际上是一阶马尔可夫链.利用一阶马尔可夫链进行轨迹预测时,只需考虑当前点对下一步位置的影响,并没有充分利用历史轨迹点的信息.使用当前点前 n 个轨迹点到预测点的 n 步转移概率进行加权求和,可以得到不同预测点的预测概率.称条件概率$=p_{ij}(n)=P\{X_{n+1}=j|X_n=i\}$为马尔可夫轨迹链$\{X_n,n\in T\}$在时刻 n 的一步转移概率,其中,$i,j\in I$,简称为转移概率.称条件概率 $p_{ij}^{(n)}=P\{X_{m+n}=j|X_m=i\}$,$(i,j\in I,m\geq 0,n\geq 1)$为马尔可夫轨迹链$\{X_n,n\in T\}$的 n 步转移概率.同时,称 $P^{(n)}=(p_{ij}^{(n)})$为马尔可夫链的 n 步转移矩阵.

使用 n 步转移概率进行预测的方法认为:到当前位置点距离不同的历史轨迹点,其对于预测点的贡献不同,因而使用加权的方法进行组合.由 n 步转移概率的定义可知:当计算一个点的 n 步转移概率时,实际上只考虑了相邻两个状态转移过程之间的关系,而并不是一个连续过程.在本文关注的轨迹预测问题中,训练数据的规模较大,因为海量数据可以提供更为准确的特征,所以考虑在马尔可夫轨迹链的基础上提出新的预测模型.

一阶马尔可夫模型的建立过程时间复杂度低、计算速度快,但是准确率低.如果采用高阶马尔可夫模型,其计算复杂度随阶数显著增加,并不适用于海量数据的轨迹预测问题.给定轨迹 $T=\{p_1,p_2,p_3,\ldots,p_n\}$,高阶马尔可夫模型的主要思想类似于利用以前 n 个点作为前缀的轨迹在训练集中出现的次数为概率进行预测,该思想本质上与频繁序列挖掘的思想类似.其基本思路为:首先,在训练数据中挖掘出长度不同的频繁序列作为知识库,在预测过程中,使用该知识库对待预测轨迹进行匹配,待匹配轨迹的长度逐渐缩减,直至匹配完成,输出匹配的频繁序列.本质上讲,使用该匹配方法等价于将可变阶的马尔可夫链应用于轨迹预测算法,同时不会产生状态空间爆炸问题,具有较高的空间利用率.

根据以上描述可以发现:无论是固定阶马尔可夫链预测还是可变阶马尔可夫链预测,均与序列模式挖掘具有一定的相似性.通过分析发现:将频繁序列挖掘算法与轨迹预测问题相结合,可以在马尔可夫链预测模型基础

上简化模型构建,提高预测准确率.因此,针对轨迹数据预测问题,本文提出了基于前缀投影技术的增量式轨迹预测算法.本节首先给出算法中使用的主要概念,并给出相关性质和定理[20].

定义 **2**(轨迹). 给定 GPS 点序列 $T=\{p_1,p_2,p_3,\ldots,p_n\}$,$p_i=(lng_i,lat_i,t_i)$,其中,$lng$ 表示纬度坐标,lat 表示经度坐标,t 表示时间戳.对于任意 $i<j$,有 $p_i.t<p_j.t,p_i\in P,P$ 为 GPS 点的集合,则称 T 为轨迹.

定义 **3**(轨迹序列). 给定轨迹 $T=\{p_1,p_2,p_3,\ldots,p_n\}$,经轨迹特征点提取,转换得到序列$\alpha=\langle\alpha_1,\alpha_2,\ldots,\alpha_i,\ldots,\alpha_n\rangle$,则称序列$\alpha$为轨迹序列.

为了提取序列α上的轨迹角度变化特征点α_i,需要完成两个 GPS 点间距离的计算和轨迹角度变化计算,具体方法如下:

已知点 A,B,φ_1,φ_2 分别表示 A,B 的纬度,λ_1,λ_2 分别表示 A,B 的经度,γ为地球半径,则 A,B 两点距离为

$$d(A,B)=2\gamma\arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \tag{1}$$

已知点 A,B,C,则角$\angle ACB$ 为

$$\cos\angle ACB=\frac{d(B,C)^2+d(A,C)^2-d(A,B)^2}{2d(B,C)d(A,C)} \tag{2}$$

特征点提取算法以提取轨迹角度变化点为基础,首先遍历轨迹数据,利用公式(1)、公式(2)计算轨迹中轨迹点与前序后序轨迹点之间的夹角:若角度值超过给定的角度阈值,则认为该点为角度变化点,将其加入到轨迹序列中,就完成了原始轨迹数据的转换.

定义 **4**(子轨迹). 给定序列$\alpha=\langle\alpha_1,\alpha_2,\ldots,\alpha_p\rangle$和$\beta=\langle\beta_1,\beta_2,\ldots,\beta_q\rangle$,$p\le q$,若有递增下标序列$\langle i_1,i_2,\ldots,i_p\rangle$($i_1<i_2<\ldots<i_p$),使得$\alpha_1=\beta_{i1},\alpha_2=\beta_{i2},\ldots,\alpha_p=\beta_{ip}$,则称序列模式$\alpha$是$\beta$的子轨迹,或$\beta$包含$\alpha$.

定义 **5**(前序轨迹). 给定序列$\alpha=\langle\alpha_1,\alpha_2,\ldots,\alpha_p\rangle$和$\beta=\langle\beta_1,\beta_2,\ldots,\beta_q\rangle$,$p\le q$,$\alpha$为$\beta$的前序轨迹当且仅当

$$\alpha_1=\beta_1,\alpha_2=\beta_2,\ldots,\alpha_p=\beta_p.$$

定义 **6**(后序轨迹). 给定序列$\alpha'=\langle\alpha_1,\alpha_2,\ldots,\alpha_n\rangle$为$\alpha$对应于$\beta=\langle\beta_1,\beta_2,\ldots,\beta_m\rangle$的投影,则序列$\langle\alpha_{m+1},\alpha_{m+2},\ldots,\alpha_n\rangle$为$\alpha$对应于序列$\beta$的后序轨迹.

定义 **7**(投影). 给定轨迹序列α和β,若β是α的子轨迹,则α关于β的投影α'必须满足:β是α'的前序轨迹,α'是α的满足上述条件的最大子轨迹.

定义 **8**(投影数据库). 设α是轨迹序列集 S 中的一个频繁轨迹,那么 S 中所有α的后序轨迹组成的集合即为α在 S 中的投影数据库,记为 $S|\alpha$.

定义 **9**(投影数据库的支持度). 设α是序列集 S 中的一个频繁轨迹,序列β以α为前序轨迹,那么β在 $S|\alpha$中的支持度为 $S|\alpha$中以β为前序轨迹的轨迹序列的数量,记为 $Support_{S|\alpha}(\beta)$.

频繁轨迹的求取过程可以视为递归求解的过程,针对这一特点,可以得到以下引理和推论[20].

引理 **1**. 令α表示长度为 l 的轨迹序列,$\langle\beta_1,\beta_2,\ldots,\beta_m\rangle$为以$\alpha$为前序轨迹的长度为 $l+1$ 的轨迹序列的集合,该集合可以被分为 m 个相互独立的子集,第 j 个子集($1\le j\le m$)中包含所有以β_j为前序轨迹的序列.

注意:默认的轨迹序列是空集合.

由引理 1 可知:轨迹模式挖掘过程通过不断将投影集进行分割,减小对数据的访问次数,同时选取可能的项作为频繁序列的增长项,这一过程是一个可递归的过程.

引理 **2**. 给定两个轨迹序列α和β,$\alpha\in S,\beta\in S$,其中,α是β的前序轨迹,则有:

1. $S|\beta=(S|\alpha)|\beta$;
2. 对于任何以α为前序轨迹的轨迹序列γ,$Support_S(\gamma)=Support_{S|\alpha}(\gamma)$;
3. α的投影集的大小不超过 S 的大小.

引理 2 给出了在不断进行递归挖掘时,频繁轨迹α和以α为基础进行增长的序列β之间的关系.因为 $S|\beta=(S|\alpha)|\beta$,所以使用α的投影集进行关于β的挖掘可以得到所有与β相关的序列.同时,$Support_S(\gamma)=Support_{S|\alpha}(\gamma)$可以保证以$\alpha$为基础进行增长的序列$\beta$同样也是频繁轨迹.

推论 **1**. 轨迹α称为频繁轨迹,当且仅当轨迹满足引理 1 和引理 2 给出的递归性质.

结合上述定义和性质,轨迹预测问题定义如下.

定义 **10**(轨迹预测). 轨迹预测问题定义为求以待预测轨迹序列后序轨迹作为前序轨迹的频繁轨迹序列.

由于引理 2 给出了频繁轨迹挖掘递归过程的原理,因此对于轨迹序列数据,其频繁轨迹挖掘过程可以根据以下步骤进行.

- 第 1 步:首先考虑其每一项在训练数据集中出现的次数,对数据集进行一步扫描,得到所有出现次数满足最小支持度的项,记为 F.
- 第 2 步:划分搜索空间.在第 1 步中得到了长度为 1 的频繁项集合 F,使用 F,可以将轨迹序列集 S 划分成$|F|$个相互独立的子集.其中,每个子集中的所有元素都对应一个频繁 1 项,且该子集中的所有元素都以该频繁 1 项为前序轨迹.
- 第 3 步:挖掘子集中序列模式.子集中序列模式可以通过构造相应的投影集来进行递归挖掘.若有投影集$S|\alpha$,那么α的增长项必然出现在投影集$S|\alpha$中,每次增长长度为 1.遍历投影集$S|\alpha$,计算每一序列中第一项的支持度,取满足最小支持度的项作为α的增长项,同时将该项对应的序列进行再次划分.
- 重复第 3 步.

例 1(频繁轨迹挖掘示例):假设在轨迹数据库中存有如表 1 所示的轨迹序列集合,设最小支持度为 2.

在轨迹序列集中存在{*a*,*b*,*c*,*d*,*e*,*f*}项,各项在数据集中的支持度见表 2,其中,{*f*}项的支持度为 1,不满足最小支持度的要求,因此被舍弃;余下的{*a*,*b*,*c*,*d*,*e*}项满足最小支持度要求,因此可以作为频繁轨迹的起始项进行下一步挖掘,对每一项构建投影集.

Table 1 Example of trajectory sequences

表 **1** 轨迹序列举例

轨迹 ID	轨迹序列
1	⟨*a b d e*⟩
2	⟨*a b d c*⟩
3	⟨*a c k e*⟩
4	⟨*b c a d e f*⟩
5	⟨*d a b c e*⟩
6	⟨*e b a d e*⟩

Table 2 Support counts of trajectory items

表 **2** 轨迹项支持度计数值

项	支持度
a	6
b	5
c	4
d	5
e	5
f	1

以⟨*a*⟩的投影集及其挖掘过程中频繁轨迹序列的增长过程为例,见表 3.频繁轨迹挖掘过程递归进行,第 1 次挖掘首先以⟨*a*⟩为频繁轨迹作为输入,计算得到其投影集为{⟨*b d e*⟩,⟨*b d c*⟩,⟨*b c e*⟩,⟨*c k e*⟩,⟨*d e f*⟩,⟨*d e*⟩},然后计算投影集中各序列第 1 项的支持度,取满足最小支持度的项作为频繁序列增长项,同时将该项对应的投影集中的项带入下一次迭代,作为其投影集.具体来说,在第 1 次迭代中,⟨*b*⟩和⟨*d*⟩满足最小支持度要求,因此分别作为频繁轨迹增长项,进行第 2 次迭代.在表 3 中,第 2 次迭代被分为以⟨*a b*⟩和⟨*a d*⟩为频繁轨迹分别进行挖掘的过程,其中,⟨*a d*⟩的投影集由第 1 次迭代中以⟨*b*⟩为前序轨迹的轨迹项组成.以此类推,最终当挖掘过程进行到第 3 次迭代时,投影集中已无满足最小支持度的轨迹项,因此挖掘过程结束.

Table 3 Example of mining frequent trajectory sequences

表 **3** 频繁轨迹序列挖掘举例

<table>
<tr><td>次数</td><td colspan="2">1</td><td colspan="2">2</td><td colspan="2">3</td></tr>
<tr><td rowspan="3">过程</td><td>频繁轨迹</td><td>投影集</td><td>频繁轨迹</td><td>投影集</td><td>频繁轨迹</td><td>投影集</td></tr>
<tr><td rowspan="2">⟨a⟩</td><td rowspan="2">⟨b d e⟩,⟨b d c⟩
⟨b c e⟩,⟨c k e⟩
⟨d e f⟩,⟨d e⟩
支持度 b:3,d:2,c:1</td><td>⟨a b⟩</td><td>⟨d e⟩,⟨d c⟩,⟨c e⟩
支持度 d:2,c:1</td><td>⟨a b d⟩</td><td>⟨e⟩,⟨c⟩
支持度 e:1,c:1</td></tr>
<tr><td>⟨a d⟩</td><td>⟨e f⟩,⟨e⟩
支持度 e:2</td><td>⟨a d e⟩</td><td>⟨f⟩
支持度 f:1</td></tr>
</table>

例 1 中展示的频繁轨迹挖掘过程最终产生了 5 条频繁轨迹,分别是⟨*a*⟩,⟨*a b*⟩,⟨*a d*⟩,⟨*a b d*⟩,⟨*a d e*⟩.

完成频繁轨迹挖掘后,轨迹预测过程即可简化为对频繁轨迹进行匹配查找.以例 1 中表 1 数据为例,例 2 给出了轨迹预测的具体过程.

例 2(轨迹预测示例):设待预测轨迹序列为⟨*e c a b*⟩,首先在频繁轨迹集合中查找以⟨*e c a b*⟩为前序轨迹的频繁轨迹,在集合中未找到,因此缩短待预测序列,取⟨*e c a b*⟩的后序轨迹序列⟨*c a b*⟩为待预测轨迹序列重复上一步骤,直至取⟨*a b*⟩作为待预测序列,通过匹配查询,找到频繁轨迹序列⟨*a b d*⟩.频繁轨迹⟨*a b d*⟩以待预测轨迹序列⟨*e c a b*⟩后序轨迹⟨*a b*⟩为前序轨迹,满足定义给出的描述,因此取⟨*d*⟩作为轨迹预测结果(如图 1 所示).

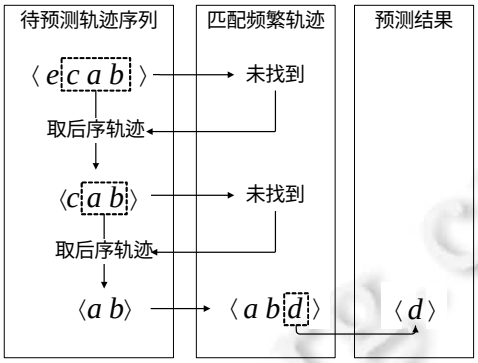


Fig.1 Example of trajectory prediction

图 1 轨迹预测示例

在进行轨迹预测时,可以采用两种轨迹匹配方法:(1) 精确匹配,从频繁轨迹集合中找到包含待预测轨迹后序轨迹的频繁轨迹;(2) 模糊匹配,从挖掘的频繁轨迹集合中找到所有将待预测轨迹后序轨迹作为子序列的所有频繁轨迹.本文所提出的预测算法采用精确匹配,针对待预测轨迹给出一条最可能的轨迹路线.

3 轨迹预测框架

轨迹预测功能的实现需要一系列相应辅助模块共同协作,本文所提出的 PPTP 轨迹预测算法也不例外.为了更好地进行轨迹预测,本文提出一种新的轨迹预测框架,其内部工作原理如图 2 所示.

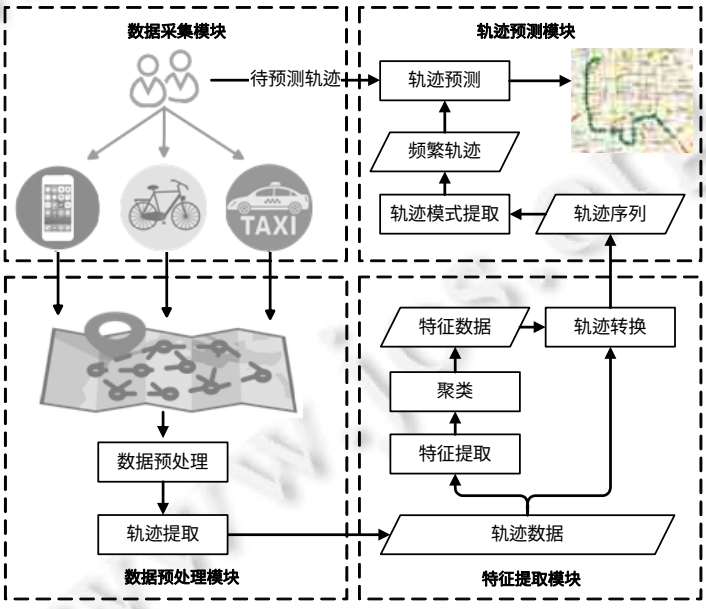


Fig.2 Architecture of trajectory prediction system

图 2 轨迹预测系统框架图

数据采集模块首先完成对用户数据的采集工作,然后将原始数据交由数据预处理模块;数据预处理模块将原始轨迹数据中的噪声去除,同时应用轨迹提取算法从中提取出轨迹数据,并按照时间、距离等条件进行轨迹分割,最终形成格式统一的轨迹数据;在特征提取模块中,首先对轨迹数据按角度变化特性进行特征点提取、聚类等操作,然后使用提取的特征数据对轨迹进行转换,从而得到轨迹序列数据;在最终的预测模块,算法利用轨迹序列数据进行轨迹模式提取,得到频繁轨迹作为预测所需的候选数据,然后根据用户的实时轨迹数据,使用PPTP 轨迹预测算法完成最终的预测.

4 轨迹预测算法及性能分析

4.1 算法描述

前面通过举例说明了基于前缀投影技术的大规模轨迹预测算法的思想及工作原理,下面给出算法的形式化描述,见算法 1.表 4 给出算法中使用参数的说明.

算法 **1**. 基于前缀投影技术的增量式轨迹预测算法——PPTP.

输入:待预测序列 t,预测步数 n.

输出:预测结果.

```
1.  F←getPrefix(t[0]);
2.  IF F=∅
3.     THEN RETURN PPTP(t.suffix(1),n);
4.  END IF
5.  F′←∅;
6.  FOR EACH f IN F
7.     IF f.len<t.len+n
8.        THEN CONTINUE;
9.     END IF
10.    i←0;
11.    WHILE i<t.len
12.       IF f[i]!=t[i]
13.          THEN BREAK;
14.       END IF
15.          i←i+1;
16.    END WHILE
17.    IF i=t.len
18.       THEN F′.add(f);
19.    END IF
20. END FOR
21. F′.sort();
22. IF F′=∅
23.    THEN RETURN PPTP(t.suffix(1),n);
24. END IF
25. RETURN F′[0].get(t.len+n–1);
```

Table 4 Parameter introduction of the proposed algorithm

表 **4** 算法参数说明

参数	说明
t	待预测轨迹序列
n	预测步数
F	频繁轨迹集合
F'	满足条件的频繁轨迹集合

算法第 1 行~第 4 行获取待预测序列 t 的第 1 项,调用 getPrefix 函数获取以该项为前缀的频繁模式,若获取失败,则以 t 的第 2 项开始的子串作为输入,递归调用 PPTP 方法;第 6 行~第 20 行遍历之前获取到的 F 集合,其中,第 7 行~第 9 行判断 f 的长度是否大于待预测序列 t 的长度与预测步数之和,如果不满足,则说明 f 无法提供足够长度的预测步数,因此跳过 f;第 11 行~第 16 行逐项匹配 f 和 t,若失败,则跳出循环;第 17 行~第 19 行判断 f 的长度,若满足预测条件,则将 f 加入结果集;第 22 行~第 24 行判断 F'是否为空,若为空,则以 t 的子串作为输入,递归调用 PPTP 方法;第 25 行返回 P'中第 1 项的相应步作为预测结果.

4.2 算法性能分析

通过分析算法 1,可以得出其时间复杂度为 $O(m^n)$,其中,m 表示频繁轨迹数量;n 表示预测步数,即算法迭代次数.通过第 5.3 节图 5 给出的不同预测步数下算法轨迹预测准确性对比实验得到的结论可知:预测步数通常介于 2 步~4 步之间可以保证较高的预测准确性,而频繁轨迹的数量远远小于原始轨迹数据,因此算法 1 的整体复杂性不高.算法的空间复杂度为 $O(k)$,其中,k 表示所有轨迹点的数量.

算法 1 的正确性和预测结果完整性可以通过引理 1 和引理 2 得到证明,分析算法时间性能得到如下结论.

(1) 算法不会生成多余的候选序列模式.PPTP 算法仅通过较短的频繁序列模式以增量式方式生成长轨迹模式,不会产生投影数据库中不存在的候选轨迹序列模式.

(2) 投影数据库的规模不断缩小.通过表 3 可以发现,经过 3 次迭代操作,投影数据库中项的规模不断缩小,因为其中的候选项是从频繁轨迹序列扩展生成的后序轨迹序列.

(3) 算法的主要时间开销是投影数据库的构建.最坏的情况是,PPTP 算法对每一个轨迹序列模式构建一个投影数据库,如果候选频繁轨迹序列模式规模不是很大,则时间代价不会很高.

5 实验及算法性能分析

5.1 实验环境及数据集描述

本实验所使用的轨迹数据集来源于微软亚洲研究院郑宇研究员所领导的 T-Driver 项目[21],数据采集自北京市真实路网中的出租车 GPS 设备,包含 10 357 辆出租车超过 1 周的行驶轨迹数据.该轨迹数据集中的轨迹点总量超过 15 000 000,总行驶长度超过 9 000 000km,具体描述见表 5.

Table 5 Description of experimental datasets

表 5 实验数据集描述

参数	值
轨迹时间跨度	2008/02/02~2012/02/08
车辆数	10 357
轨迹数量	>25000
轨迹点数量	>15000000
总长度	>9000000km

本文中所提到的算法均采用 Java 程序设计语言实现,使用 Eclipse Juno 作为开发环境.实验硬件平台为:Intel(R) Core(TM)2 Duo P8700 2.53GHz CPU,3GB 内存,操作系统平台为 Windows 7.实验通过对比实验评价所提出方法的性能优劣,对比算法包括:采用 1 阶马尔可夫链轨迹预测算法[4](当前位置由前一个位置点决定)、2 阶马尔可夫链轨迹预测算法[22](代表高阶马尔可夫链,当前位置由前面两个位置点确定,本文以 2 阶马尔可夫链为例,更高阶马尔可夫链情况类似)和本文提出的基于前缀投影技术的 PPTP 轨迹预测算法.实验中针对不同数据集随机选取 90%作为训练数据,其余 10%的轨迹数据用作测试数据.

5.2 性能评价指标

本文提出的轨迹预测算法工作过程包括:首先,在数据集上使用轨迹特征提取算法,提取轨迹特征点;然后,通过轨迹转换技术将 GPS 轨迹数据转化为由轨迹特征点表示的特征序列;最后,使用基于特征序列的轨迹预测

算法进行预测.为了方便量化表示算法的性能优势,本文采用如下性能评价指标[2].

定义 **11**(预测命中). 已知轨迹序列 $T=\{T_1,T_2,...,T_k\}$,预测轨迹序列 $P=\{P_1,P_2,...,P_n\}$,$k<n$,$dist(m,n)$表示时空轨迹点 m 和 n 间的欧氏距离,δ表示距离阈值,则 $dist(T_i,P_i)<\delta$时,表示预测命中,定义为

$$H(T_i,P_i)=\begin{cases}1, & dist(T_i,P_i)<\delta\\ 0, & dist(T_i,P_i)>\delta\end{cases} \tag{3}$$

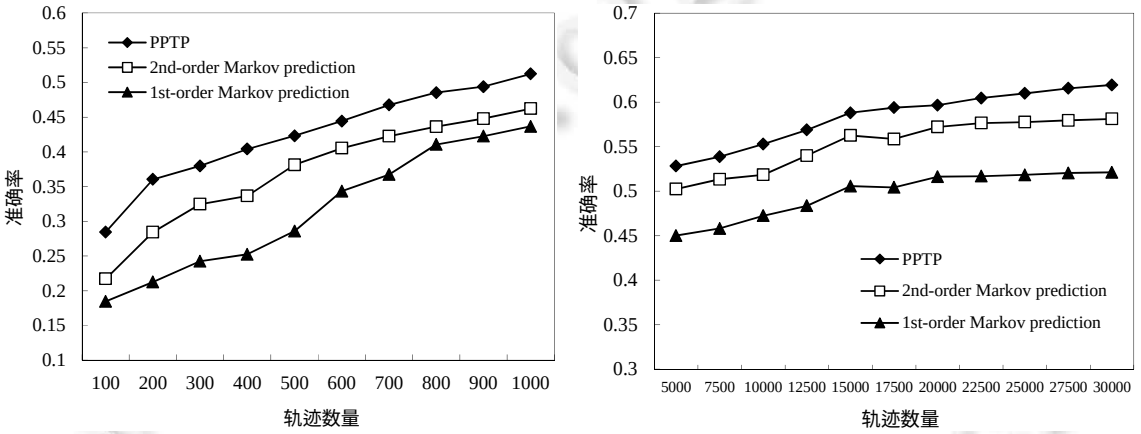
定义 **12**(预测准确率). 已知轨迹序列 T,预测轨迹序列 P,则预测准确率定义为

$$Accuracy=\frac{\sum_{i=1}^{n}H(T_i,P_i)}{|P|} \tag{4}$$

其中,$|P|$表示预测轨迹序列的长度.

5.3 轨迹预测准确性对比

本节实验首先观察轨迹预测算法在不同规模数据集上的预测准确性,图 3(a)、图 3(b)分别展示了在小规模和大规模轨迹数据集上进行的对比实验结果.其中,横轴表示训练轨迹的数量,纵轴表示预测准确率 Accuracy.



(a) 小规模数据集 (b) 大规模数据集

Fig.3 Prediction accuracy comparison with different number of trajectories

图 3 不同数量训练轨迹下预测准确性对比

实验所用训练数据规模依次递增,通过观察发现:

(1) 本文提出的 PPTP 轨迹预测算法在预测准确率上明显优于其他两种算法.在小规模数据集图 3(a)上实验,相对于 2 阶马尔可夫链和 1 阶马尔可夫链轨迹预测算法平均提高了 15.8%和 39.8%;在大规模数据集图 3(b)上实验,相对于 2 阶马尔可夫链和 1 阶马尔可夫链轨迹预测算法平均提高了 5.5%和 17.4%.主要原因在于:1 阶马尔可夫链预测算法和 2 阶马尔可夫链预测算法仅考虑了轨迹序列中较短项之间的相互影响,对训练数据的利用不够全面,因此当训练数据量不断增加时,其预测准确率上升空间有限.反观本文所提出的 PPTP 轨迹预测算法,充分利用了训练数据集,将不同长度的训练数据都作为依据,对训练数据的使用更为全面,因此具有更好的预测准确性.

(2) 随着训练数据集规模的逐渐增大,3 种轨迹预测算法的预测准确率均不断提高,PPTP 算法比较稳定,尤其是在大规模训练数据集下,准确性保持在一个较高的水平,且增长趋势比较平缓;而其他两种基于马尔可夫链的预测算法均会出现波动,进一步说明本文提出的基于前缀投影技术的增量式轨迹预测算法具有较好的稳定性.

为了更加全面地展示 PPTP 轨迹预测算法对不同数据集的适应性,使用 8 种不同规模的数据集进行如下实

验.图 4 展示了 3 种轨迹预测算法在不同数据集上的预测准确率,通过对比可以发现,PPTP 轨迹预测算法的预测准确率均优于其他两种算法,相对于 2 阶马尔可夫链和 1 阶马尔可夫链轨迹预测算法,平均提高了 7.1%和19.9%.进一步证明了本文提出的算法不依赖于训练数据集,具有很好的普适性.且 PPTP 算法的预测准确性均高于其他两种算法,原因与上面的描述相同.

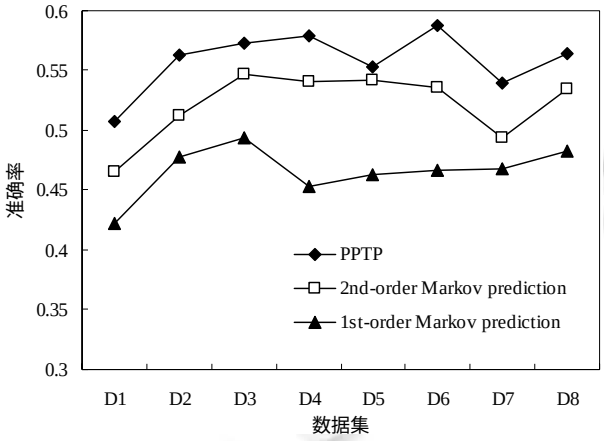


Fig.4 Prediction accuracy comparison with different datasets

图 4 不同数据集下算法预测准确率对比

在轨迹预测问题中,未来 n 步位置预测也是重点考虑的问题.n 步预测指的是以当前位置为起始,预测 n 步后移动对象的位置,可以验证算法的长轨迹预测能力.针对这一问题,使用上述 3 种算法进行对比验证,实验结果如图 5 所示.

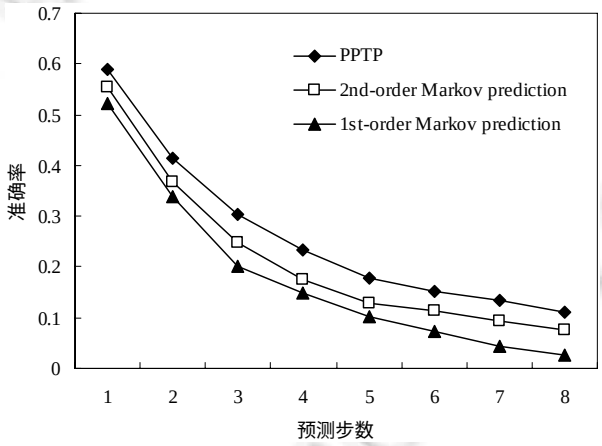


Fig.5 Relationship between prediction steps and accuracy

图 5 预测步数与准确率的关系

通过观察可以发现,随着预测步数的增加,3 种算法的预测准确率均有所下降,但 PPTP 算法在各个阶段均优于其他两种算法.原因在于,在进行 n 步预测对比实验时,基于马尔可夫链的对比算法采用 n 步转移概率,即条件概率 $p_{ij}^{(n)}=P\{X_{m+n}=j|X_m=i\}(i,j\in I,m\geq 0,n\geq 1)$来计算其 n 步转移矩阵 $P^{(n)}=(p_{ij}^{(n)})$.相对于 PPTP 算法使用可变长模式序列进行预测的方式,对比算法仅考虑了固定长度的训练轨迹预测的作用,对于数据的适应性较差,因此在实验中,PPTP 轨迹预测算法的预测准确率相对较高.

在 PPTP 轨迹预测算法中,支持度表示训练数据中某一序列模式出现的次数,该参数的选择对预测准确率

有一定影响,实验针对支持度的选择问题,在不同数据集上进行对比实验,实验结果如图 6 所示.通过观察实验结果可以发现,预测准确率随支持度的不同而产生波动,在所有实验数据集上都表现出了先上升后下降的趋势.当支持度位于 2~4 区间时,算法在所有实验数据集上都表现出了较高的预测准确率,因此可以认为在实验所用数据集上进行实验,支持度在 2~4 区间内选择可以获得最佳的预测准确率.

由于支持度这一概念本身与训练数据集的规模具有较大关联,因此实验本身并不能给出十分精确的参数选取范围.结合实验结果和支持度的定义可以总结出结论:当训练数据集的规模有显著提升时,支持度的选取范围也应该相应提高.

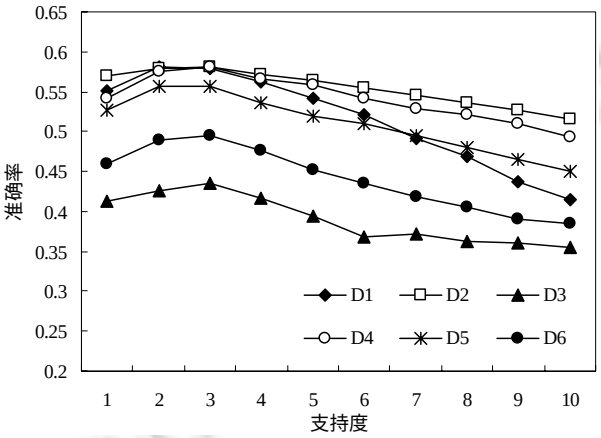


Fig.6 Effect of support counts on prediction accuracy

图 6 支持度对准确率的影响

5.4 轨迹预测时间性能对比

为了验证 PPTP 算法在时间上相对于其他算法的优劣,进行了时间性能对比实验.图 7 给出了 3 种算法在轨迹数量不断递增情况下的时间消耗情况.

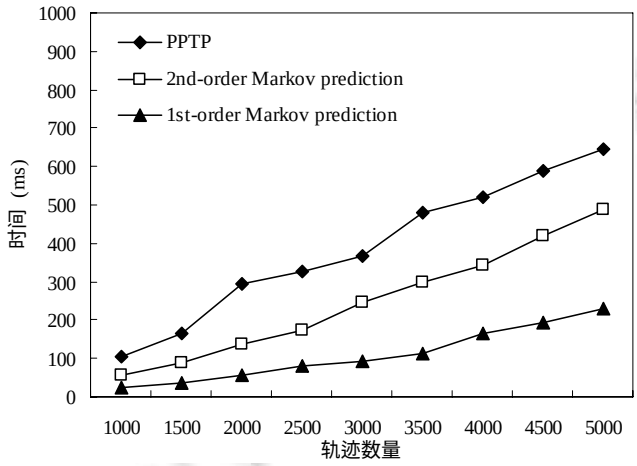


Fig.7 Prediction time comparison of algorithms with different number of trajectorys

图 7 不同轨迹数量下算法预测时间比较

通过观察可以发现,PPTP 轨迹预测算法相对于其他两种算法,在模型训练过程中要花费相对多一些的时间,因为需要构建投影数据库.PPTP 算法相对于 2 阶和 1 阶马尔可夫链轨迹预测算法平均高出 0.138s 和 0.277s,

总体时间开销差异维持在 ms 级别,即使在实时性要求较高的轨迹预测系统中,这一差异也是可以接受的.

3 种算法在不同数据集上的时间代价如图 8 所示,PPTP 轨迹预测算法的时间代价略高于其他两种对比算法,但是总体上维持在 ms 级的差异.主要原因在于,构建前缀投影数据库的时间开销略大于马尔可夫链预测算法中计算条件概率转移矩阵的时间代价.

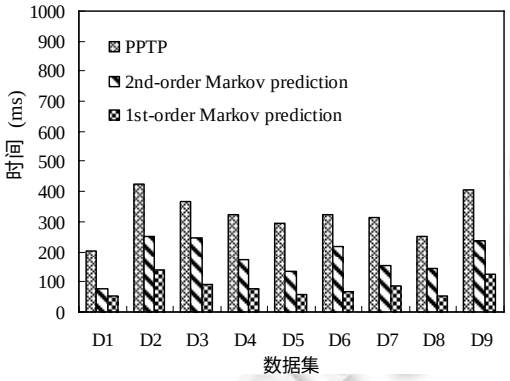


Fig.8 Prediction time comparison of algorithms on different datasets

图 8 不同数据集上算法预测时间比较

5.5 轨迹预测系统展示

在本文提出的 PPTP 算法的基础上设计实现了轨迹预测可视化系统,该系统包括 3 个主要功能模块:轨迹数据加载、数据可视化模块和轨迹预测模块.该系统是普适的轨迹预测系统,可以按照数据格式加载不同的地图信息和轨迹数据集.这里以轨迹预测模块为例,介绍系统主要功能.

轨迹预测模块以前面所提到的增量式前缀投影技术为理论基础,提供轨迹实时预测功能.模块分为训练部分和预测部分,训练部分首先对轨迹数据进行特征提取、聚类等操作,训练过程所得结果存储于移动数据库中.预测算法使用训练结果作为依据,结合当前输入轨迹进行预测.轨迹预测模块的功能界面如图 9 所示.

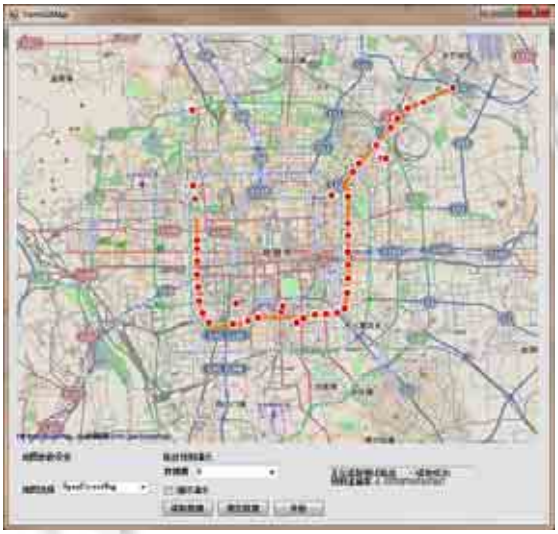


Fig.9 Graphical user interface of trajectory prediction system and visualization results

图 9 轨迹预测系统图形化用户界面及可视化结果

可视化界面中,主要包括地图选择、数据加载、清除、轨迹预测等功能,在加载数据并开始执行预测后,预测详细信息由右侧窗口打印输出至界面,直观显示给客户.

如图 9 所示,用户可以通过左下方“地图选择”功能加载需要的地图,通过点击“读取数据”按钮添加不同的轨迹数据集,点击“开始”按钮完成训练和预测的过程,轨迹预测准确率结果展示在右下方的文本框区域内.并在地图上可视化输出一条供用户参考的最佳路线.

6 结束语

针对海量移动对象轨迹数据,结合频繁序列模式挖掘算法,本文提出了一种面向大规模位置数据的前缀投影轨迹预测模型,提出前序、后序轨迹和投影数据库的概念,并给出前缀投影轨迹序列模式挖掘的相关性质和定理.利用真实轨迹数据进行多角度实验,对算法性能进行全面检验.本文提出的轨迹预测算法在实验中表现出较高的预测准确率,相对于 1 阶和高阶马尔可夫链轨迹预测算法,其平均预测准确率可以得到提升.以本文提出的预测模型为理论依据,开发了一个普适的轨迹预测系统,提供轨迹可视化、轨迹预测等功能.

未来的研究工作包括:(1) 结合实时交通情况,如交通拥堵、天气变化等因素,改进轨迹预测算法,提供更为准确的轨迹预测算法;(2) 引入社交数据,分析个体用户的出行模式,对用户进行聚类分析,提取群体的出行特征,以此为辅助改进轨迹预测算法;(3) 优化轨迹预测算法的时间效率,可对部分数据采用预计算的方式合理安排计算策略,从而降低实时计算量,提高预测效率.

References:

[1] Guo C, Liu JN, Fang Y, Luo M, Cui JS. Value extraction and collaborative mining methods for location big data. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):713−730 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4570.htm [doi: 10.13328/j.cnki.jos.004570]

[2] Qiao SJ, Han N, Zhu W, Gutierrez LA. TraPlan: An effective three-in-one trajectory prediction model in transportation networks. IEEE Trans. on Intelligent Transportation Systems, 2015,16(3):1188−1198. [doi: 10.1109/TITS.2014.2353302]

[3] Yuan J, Zheng Y, Xie X, Sun G. T-Drive: Enhancing driving directions with taxi drivers' intelligence. IEEE Trans. on Knowledge and Data Engineering, 2013,25(1):220−232. [doi: 10.1109/TKDE.2011.200]

[4] Qiao SJ, Shen DY, Wang XT, Han N, Zhu W. A self-adaptive parameter selection trajectory prediction approach via hidden markov models. IEEE Trans. on Intelligent Transportation Systems, 2015,16(1):284−296. [doi: 10.1109/TITS.2014.2331758]

[5] Zheng Y. Trajectory data mining: an overview. ACM Trans. on Intelligent Systems and Technology, 2015,6(3):Article 29. [doi: 10.1145/2743025]

[6] Bao J, Zheng Y, Wilkie D, Mokbel M. Recommendations in location-based social networks: A survey. Geoinformatica, 2015,19(3):525−565. [doi: 10.1007/s10707-014-0220-8]

[7] Zheng K, Zheng Y, Yuan J, Shang S, Zhou XF. Online discovery of gathering patterns over trajectories. IEEE Trans. on Knowledge and Data Engineering, 2014,26(8):1974−1988. [doi: 10.1109/TKDE.2013.160]

[8] Prentow T, Thom A, Blunck H, Vahrenhold J. Making sense of trajectory data in indoor spaces. In: Proc. of the 16th IEEE Int'l Conf. on Mobile Data Management. Washington: IEEE, 2015. 116−121. [doi: 10.1109/MDM.2015.44]

[9] Dai J, Ding Z, Xu J. Context-Based moving object trajectory uncertainty reduction and ranking in road network. Journal of Computer Science and Technology, 2016,31(1):167−184. [doi: 10.1007/s11390-016-1619-5]

[10] Tao Y, Faloutsos C, Papadias D, Liu B. Prediction and indexing of moving objects with unknown motion patterns. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2004. 611−622. [doi: 10.1145/1007568.1007637]

[11] Morzy M. Mining frequent trajectories of moving objects for location prediction. In: Proc. of 5th Int'l Conf. on Machine Learning and Data Mining in Pattern Recognition. Berlin: Springer-Verlag, 2007. 667−680. [doi: 10.1007/978-3-540-73499-4_50]

[12] Song C, Qu Z, Blumm N, Barabási AL. Limits of predictability in human mobility. Science, 2010,327(5968):1018−1021. [doi: 10.1126/science.1177170]

[13] Pentland A. Society's nervous system: Building effective government, energy, and public health systems. IEEE Computer, 2012,45(1):31−38. [doi: 10.1109/MC.2011.299]

[14] Ding Z, Yang B, Güting RH, Li Y. Network-Matched trajectory-based moving-object database: Models and applications. IEEE Trans. on Intelligent Transportation Systems, 2015,16(4):1918−1928. [doi: 10.1109/TITS.2014.2383494]

[15] Qiao SJ, Jin K, Han N, Tang CJ, Gesangduoji, Gutierrez LA. Trajectory prediction algorithm based on Gaussian mixture model. Ruan Jian Xue Bao/Journal of Software, 2015,26(5):1048−1063 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4796.htm [doi: 10.13328/j.cnki.jos.004796]

[16] Dai J, Yang B, Guo C, Ding Z. Personalized route recommendation using big trajectory data. In: Proc. of the 31st IEEE Int'l Conf. on Data Engineering. Washington: IEEE, 2015. 543−554. [doi: 10.1109/ICDE.2015.7113313]

[17] Yuan G, Zhao J, Xia S, Zhang Y, Li W. Multi-Granularity periodic activity discovery for moving objects. Int'l Journal of Geographical Information Science, 2017,31(3):435−462. [doi: 10.1080/13658816.2016.1205194]

[18] Zhang L, Liu L, Xia Z, Li W, Fan Q. Sparse trajectory prediction based on multiple entropy measures. Entropy, 2016,18(9):No.327. [doi: 10.3390/e18090327]

[19] Chaulwar A, Botsch M, Utschick W. A hybrid machine learning approach for planning safe trajectories in complex traffic-scenarios. In: Proc. of 15th IEEE Int'l Conf. on Machine Learning and Applications. Washington: IEEE, 2016. 540−546. [doi: 10.1109/ICMLA.2016.0095]

[20] Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, Dayal U, Hsu M. Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Trans. on Knowledge and Data Engineering, 2004,16(11):1424−1440. [doi: 10.1109/TKDE.2004.77]

[21] Yuan J, Zheng Y, Zhang C, Xie W, Xie X, Sun G, Huang Y. T-Drive: Driving directions based on taxi trajectories. In: Proc. of the 18th SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. New York: ACM Press, 2010. 99−108. [doi: 10.1145/1869790.1869807]

[22] Gambs S, Killijian M, Cortez DP, Miguel N. Next place prediction using mobility Markov chains. In: Proc. of the 1st Workshop Measurement, Privacy, and Mobility. New York: ACM Press, 2012. [doi: 10.1145/2181196.2181199]

附中文参考文献:

[1] 郭迟,刘经南,方媛,罗梦,崔竞松.位置大数据的价值提取与协同挖掘方法.软件学报,2014,25(4):713−730. http://www.jos.org.cn/1000-9825/25/713.htm [doi: 10.13328/j.cnki.jos.004570]

[15] 乔少杰,金琨,韩楠,唐常杰,格桑多吉,Gutierrez LA. 一种基于高斯混合模型的轨迹预测算法.软件学报,2015,26(5):1048−1063. http://www.jos.org.cn/1000-9825/4796.htm [doi: 10.13328/j.cnki.jos.004796]

乔少杰(1981－),男,山东招远人,博士,教授,CCF 高级会员,主要研究领域为轨迹数据挖掘,机器学习.

李斌勇(1982－),男,博士,讲师,CCF 专业会员,主要研究领域为大数据,云服务.

韩楠(1984－),女,博士,讲师,主要研究领域为移动对象数据库.

王晓腾(1991－),男,硕士,主要研究领域为移动对象数据库,轨迹预测.

李天瑞(1969－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为智能信息处理,数据挖掘.

Louis Alberto GUTIERREZ(1980－),男,博士,Researcher,主要研究领域为数据挖掘.

李荣华(1985－),男,博士,助理教授,CCF 专业会员,主要研究领域为图数据挖掘,机器学习.