密级：　　保密期限：

北京郵電大學

硕士学位论文

题目：移动协作感知下基于真实移动路径的轨迹研究与应用

学号：2012111408

姓名：周萌

专业：计算机科学与技术

导师：刘志勇

学院：网络技术研究院

2015 年　3 月　13 日

独创性（或创新性）声明

本人声明所呈交的论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京邮电大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名：_____ 日期：_____

关于论文使用授权的说明

学位论文作者完全了解北京邮电大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属北京邮电大学。学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

非保密论文注释：本学位论文不属于保密范围，适用本授权书。

本人签名：_____ 日期：_____

导师签名：_____ 日期：_____

移动协作感知下基于真实移动路径的轨迹研究与应用

摘 要

随着移动终端设备定位技术的发展，越来越多的用户愿意记录并共享他们的位置信息，在此基础上也就兴起了诸多基于轨迹的应用。移动协作感知作为一种新型的通过鼓励移动用户参与并共同完成感知任务收集的数据收集方式，一方面为轨迹数据的收集提供了便捷的平台，另一方面源于轨迹数据的价值又能为其提供更多的应用。在轨迹数据的应用领域，轨迹预测是其中最具研究和实用价值的方向之一。

本文首先对移动协作感知系统的技术背景进行介绍，具体介绍了移动协作感知的概念及应用场景，明确了移动协作感知系统的功能结构，并着重强调了轨迹预测模块在系统中的作用。接下来基于现有的轨迹预测方法对轨迹预测部分进行详细地分析，包括轨迹预测的可行性、应用场景、整体流程等，并从不同维度对当前的轨迹预测方法进行分类，不仅帮助了对轨迹预测的理解，也为系统选择合适的预测方法提供了依据。在对现有轨迹预测方法分析的基础上，本文选择两种典型的轨迹预测方法在实验部分基于同一数据集进行了实验验证与结果分析。

本文在移动协作感知的背景下对轨迹预测重点进行了研究。不仅在理论角度对现有轨迹预测方法从不同维度进行了分类比较，同时介绍了轨迹预测模块在系统中的功能及其与其他模块的交互关系，并在实验中采用了状态空间和模式匹配的预测方法分别进行验证和分析。在最后的部分对轨迹预测未来的发展方向进行了讨论。

关键词：移动协作感知　轨迹预测　状态空间　模式匹配

RESEARCH AND APPLICATION OF TRAJECTORY IN PARTICIPATORY SENSING

ABSTRACT

With the widespread function of location in mobile phones, more and more users are encouraged to record and share their location information, which leads to the fast growth of trajectory-based applications. Participatory Sensing is a new way of collecting sensor data, in which people are encouraged to perform the sensing task together. It provides an available way to collect trajectory data easily. On the other hand, the knowledge under trajectory records would be the feedback to the participatory sensing system. Among the applications on trajectory data, trajectory prediction is one of the most valuable research area.

This thesis introduced the technical background of participatory sensing first, including the concept and application scene of participatory sensing system and its function and overall structure, especially the function of trajectory prediction in the system. Based on the understanding of current trajectory prediction methods, we then deeply analyzed trajectory perdition from several aspects such as the feasibility of trajectory prediction, the use-case and the whole workflow. According to different classification standards, we further classified those current trajectory prediction methods. This would not only assist reader understanding trajectory prediction but also provide recommendations when adopting prediction method in specific applications. What's more, we adopted two trajectory prediction methods in our participatory sensing system and described their requirement design and detailed design accordingly. In the experiment, we validated and evaluated the two prediction methods based on the processed dataset.

The highlight of the thesis is trajectory prediction analysis and application. We not only classified and compared current trajectory

prediction methods from different aspects, but also introduced the application of trajectory prediction in the system and its interactive relationship with other modules. In the experiment, we adopted the state space and pattern matching prediction methods for evaluation and analysis. At last, we discussed the future direction of trajectory prediction.

KEY WORDS: participatory sensing, trajectory prediction, state space, pattern matching

目录

第一章 绪论

本章首先对课题的研究背景进行介绍，进而对具体的研究内容及主要工作进行阐述，并对全文的结构进行了说明。

1.1 课题研究背景

近年来，随着通信技术的发展和硬件水平的提升，移动智能终端（如智能手机、平板电脑等）日益普及，这些智能终端拥有强大计算能力，并且集成了大量先进的传感器，这些传感器可以收集光照、声音、轨迹位置等信息。在这样的背景下，为了更好地完成感知任务，研究人员提出了一种通过不同个体共同完成感知任务的感知数据收集方式——移动协作感知（Mobile Participatory Sensing）。与传统的传感网工作方式不同，移动协作感知不需要预先部署特定传感设备，它将普通手持终端用户设备所嵌入的多种传感器作为基本感知单元，保证了信息获取的多元化与及时性、实现了感知任务（来源于任务需求方）与感知动作（来源于任务参与者）的充分互动、感知数据的有效收集和分析，从而形成了众多移动对象共同参与的移动协作感知网络。利用有意识或无意识的用户参与，感知系统具有了覆盖面广、数据丰富、不需要额外硬件投资等优点，能够向基于移动协作感知网络的应用提供全面、便捷、深度的数据，在完成大规模、复杂的社会感知任务方面拥有广阔的应用前景。

移动协作感知是指利用手机、智能车辆用户携带移动终端设备收集并共享感知数据信息进而提供各种新型服务，有着广阔的应用前景和诸多优点。其中用带有时间戳的位置感知数据表示而成的轨迹信息尤其具有研究意义。轨迹在一定程度上反映了用户的生活兴趣和爱好，通过对收集到的历史轨迹数据的分析可以在某种程度上预测用户接下来的运动趋势，从而有针对性地为用户推荐相关的信息或下发激励。因此，对采集到的用户历史轨迹数据进行研究具有现实意义：一方面轨迹预测与移动协作感知中的激励机制有着密切的联系，高正确率的预测结果可以实现有针对性地下发激励，从而优化激励机制的实施；另一方面通过对轨迹数据的分析可以了解用户的行为特征及爱好，这无论是在社交关系的推测还是在提供一系列的前瞻性服务的问题上都具有很大的现实意义。

此外，对于轨迹数据的研究也具有学术意义。目前，在对轨迹预测的研究问题上，诸多学者提出了很多不同的轨迹预测方法，这些方法的适用场景、实验结果、预测算法的选择及预测结果所含的信息都值得进一步研究比较。另一方面，

当前的轨迹预测方法预测的结果准确率都普遍不高，然而有研究指出人类移动性的可预测率可达 93%，因此，在提高轨迹预测结果准确率的领域还有很大的可研究空间。

1.2 课题研究内容

随着移动互联网的发展，以及诸多学者对人类移动性研究的探索，人类的移动轨迹越来越受到广泛的关注，并且其具有很高的现实研究意义和学术研究意义。移动协作感知是利用移动终端设备中嵌入的各式传感器来代替传统的固定位置的传感器，以达到实现覆盖面更广、灵活性更高、实用性更强的感知网络的目的。对移动对象的轨迹位置的研究与把握能更好地保证移动协作感知的部署与实现。准确地预知移动对象的未来位置能掌握传感器位置的分布，在对移动对象轨迹掌握的基础上，实施以激励策略，能主观协调传感器位置的分布，从而更好地收集满足感知任务的感知数据。

因此，本文首先对移动协作感知的概念及系统整体架构进行介绍，接下来对轨迹预测的算法从各个维度进行研究，然后介绍轨迹预测模块在移动协作感知中的设计和实现，并基于真实的数据集进行实验，进而对实验结果进行比较与分析。最后，对内容进行总结，并计划下一步工作。

1.3 本文的主要工作

本文基于移动协作感知算法研究与验证项目，研究移动对象的轨迹数据的应用场景，对轨迹预测的方法在多维度进行研究，并在项目中进行实现，且针对两种不同的轨迹预测算法基于同一数据集进行实验验证与结果分析。具体工作如下：

1. 对移动协作感知系统技术背景的介绍。从移动协作感知的概念出发，说明其典型的应用场景，并对具体的系统架构进行分析介绍，着重介绍了轨迹预测在移动协作感知中的功能与应用。
2. 对轨迹预测方法从多个维度进行研究分析与比较归纳。在基于移动对象的真实轨迹数据的研究中，轨迹预测尤其具有研究价值与现实意义。而目前轨迹预测的准确率相对来说普遍不高，因此具有很广阔的研究进步空间。同时，轨迹预测在移动协作感知中具有重要作用，准确的轨迹预测结果可以为激励策略的实施提供依据与条件，激励潜在的用户去往缺少数据的地方收集感知数据，进而从整体上协调移动设备传感器的分布。这部分主要在研究轨迹预测方法的基础上总结出轨迹预测的一般流程及

应用场景，以及从预测算法、数据集规模、方法输入输出内容等方面进行比较分析。

3. 系统所需数据集的归一化处理。在对现有的真实轨迹数据集调研的基础上，选择合适的含有经纬度及时间信息的轨迹数据集，并将其与包含有温度、湿度等其他传感器信息的数据集结合进行归一化处理，通过坐标偏移的方法构造出适合整个移动协作感知系统所使用的既含有轨迹又含有传感器读数的数据集。
4. 移动协作感知算法研究与验证项目中轨迹预测模块的设计与实现。这部分内容主要从轨迹预测模块的需求分析及详细设计开始，到轨迹预测模块的具体实现进行系统的介绍。
5. 轨迹预测算法的实验结果及分析。此部分在对轨迹预测方法研究分析的基础上，采取两种轨迹预测算法，基于同一数据集进行实验，并对结果进行分析比较。

1.4 论文结构

本文的论文结构如下：

第一章　绪论主要介绍了课题的研究背景、研究内容等，引入了移动协作感知的概念，以及移动对象轨迹数据的应用前景及在移动协作感知中对移动对象轨迹数据的研究所带来的价值。

第二章　移动协作感知系统技术背景更进一步具体介绍了移动协作感知的概念，并举例说明了其典型的应用场景。此外还对本文所采取的具体的移动协作感知系统的整体架构进行了阐述，从功能及流程上分别做出介绍。

第三章　轨迹预测方法的分析与研究着手于轨迹预测的可行性，根据对现有轨迹预测方法的研究，举例说明轨迹预测的典型用例，从整体上总结出轨迹预测的一般流程，并从各个方面对现有的轨迹预测方法进行比较分析，包括预测算法的选择、预测方法的输入输出、预测过程中考虑的信息维度及建模数据集规模的选择确定等。

第四章　系统中轨迹预测模块的设计与实现从移动协作感知项目的角度出发，考虑轨迹预测模块在系统中发挥的作用与预期效果，分别阐述轨迹预测模块的需求分析、详细设计以及具体实现。

第五章　轨迹预测算法的验证与结果分析选择两种轨迹预测算法，基于

同一数据集进行实验验证，并分析其预测效果。此外，还阐述了为生成满足整个系统所需数据集的归一化处理过程。

第六章　结束语对本文中的主要内容进行总结概括，并对下一步进行的工作进行探讨。

第二章　移动协作感知系统技术背景

本章首先介绍移动协作感知的概念，及移动协作感知最常见的应用场景。在此基础上进一步介绍本文所采用的移动协作感知系统的具体架构。

2.1 移动协作感知概述

2.1.1　移动协作感知的概念

移动协作感知的概念在文章[1]中首次被提到。移动协作感知主要是指利用移动终端设备（如手机等）去构建交互式的协作感知网络，从而鼓励普通用户去收集、分析和共享当地情况。根据在个体、群体及整个城市这三种不同规模上的协作，构建不同的协作感知应用场景。

当前越来越多的人都持有移动终端电话，这些逐渐普及的终端设备无论是在交互式还是自发式地捕获、识别信息及传输照片、声音、位置等数据方面的能力越来越强。只要合理地利用这些移动设备，他们就能够代替传统的传感器结点和基于位置感知的数据收集设备。我们对无线传感网并不陌生，也知道分布式感知网能在科技、工业和军事上发挥重大的作用，但我们却对这些每天运转在用户手上的传感设备能在现实中发挥的功能和作用知之甚少。对于传统的传感网来说，传感器均处在一个中心观察者的控制之下。而不同于传统传感网，这些传感器或联网的移动设备是被配备在移动性难以琢磨的对象身上，如人和他们建造的环境。因此，要想有效切实地利用这些传感设备就需要对人类在感知过程中的参与情况进行研究和掌握。

虽然移动协作感知方式比起传统传感网技术来说有较多的优点，然而，前者在相关系统实际部署和运行的过程中仍然充斥着很多挑战和潜在问题。首先，节点（移动终端设备）的移动具有随机性，节点数量众多分布不均匀，节点提供的服务具有动态不确定性；其次，节点的各方面能力和资源相对有限；再次，不同节点拥有的资源存在差异，节点的异构性导致提供感知服务的差异性；最后，数据收集过程受参与者的影响，参与者不同的文化、教育背景等对感知数据收集在精度、可信度、时延等方面有影响，并且与参与热情有关。

2.1.2　移动协作感知的应用场景

基于移动协作感知应用的典型应用场景如图 2-1 所示。

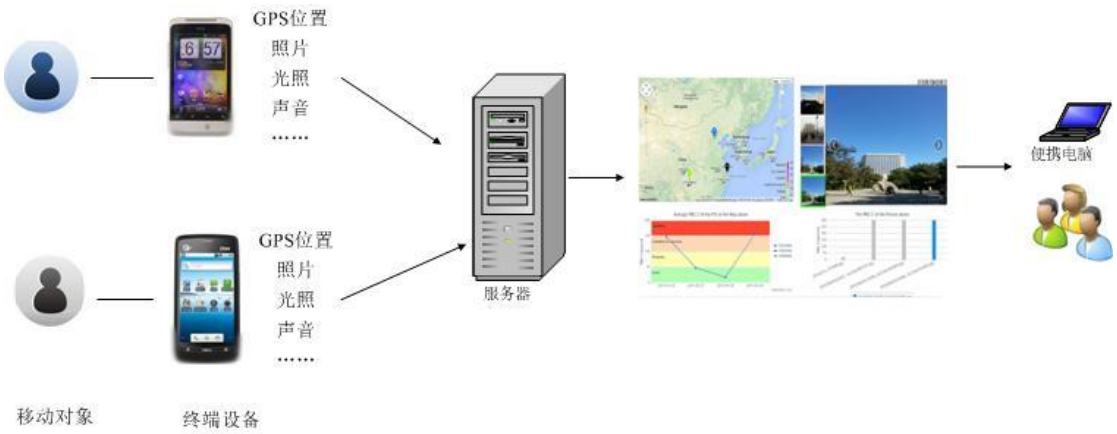


图 2-1 移动协作感知的应用场景

图 2-1 从整体上描述了典型的移动协作感知应用的场景。各参与对象利用移动终端设备采集感知数据（包含位置信息、光照、声音、拍摄的照片等），并采用一定的机制上传至服务器。在服务器端，通过对感知数据的分析、处理，挖掘其内在价值，一方面可以将现有的感知数据以一定的方式展现来体现其规律性，并呈现给用户，如在地图上根据相应的位置展示照片和声音光照等感知数据；另一方面可对感知数据进一步挖掘计算，进而为用户提供新型的应用，如根据用户所贡献的感知数据分析其长期所处的环境，并对其健康状况提供建议等。此外，服务器端可通过向参与者发布感知任务，以便更好地实现感知数据的收集，如在缺少数据的情况下，可向潜在的有可能去往缺少数据地区的用户下发激励，鼓励他们去采集数据。系统将根据参与者贡献的数据的价值下发相应的报酬，以激励用户更加积极地完成感知任务的收集。

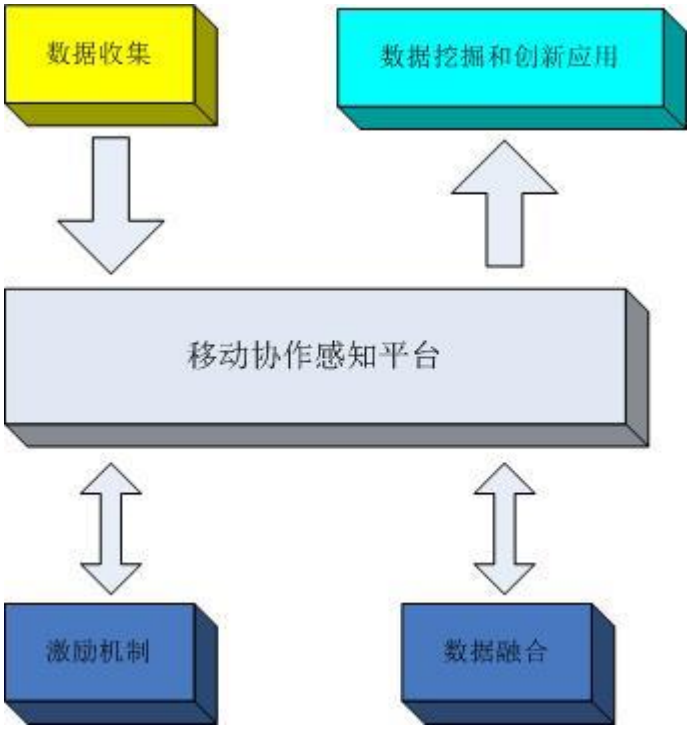


图 2-2 移动协作感知模块图

图 2-2 描述了移动协作感知的应用系统所包含的模块。基于协作感知思想的应用是鼓励大众使用他们的手持设备去收集需要的感知数据，因此数据的有效收集是进行移动协作感知的基础目标。同时，感知数据的收集会对移动对象的隐私带来威胁，如果用户一旦发现隐私被泄漏，这将会严重影响到他们后续参加数据收集的积极性。因此，在大量收集移动对象感知数据的同时，还应考虑移动对象的隐私保护问题。激励机制的作用是寻找合适的策略，鼓励用户积极参与其中去收集数据，以期能以最小的代价获取在覆盖率和准确度上满足要求的有效数据。数据融合是对用户上传的数据进行处理，计算补齐出原本缺失的但却可以通过已有的数据计算得出的数据。基于收集到的数据所进行的数据挖掘和创新应用是移动协作感知应用系统的目标所在，通过对收集的大量数据的分析挖掘，可以给参与者反馈出有用的信息，如根据移动对象每日所处的环境提出改善健康状况的建议，根据移动对象历史的移动轨迹展示其日常生活习惯、为其推测潜在好友等等。

2.2 移动协作感知系统功能结构

移动协作感知系统旨在鼓励用户使用移动终端设备收集并上传感知数据，通过对收集到的感知数据的分析挖掘，返回有用的信息给用户。本系统主要功能模块如图 2-3 所示，总体上分为客户端和服务器两大部分。在客户端有界面展示的用户 UI 模块，负责采集传感器数据的采集模块，负责与服务器端通信的通信模

块以及根据照片分析 PM2.5 的 PM2.5 分析模块等。在服务器端主要有负责与客户端通信的网络接口模块，对收集到的数据进行融合处理的数据融合模块，激励用户参与收集数据感知任务的激励模块，对用户轨迹进行预测的轨迹分析模块，在 web 端展示收集到的感知数据的 web 展示模块及数据库管理模块。

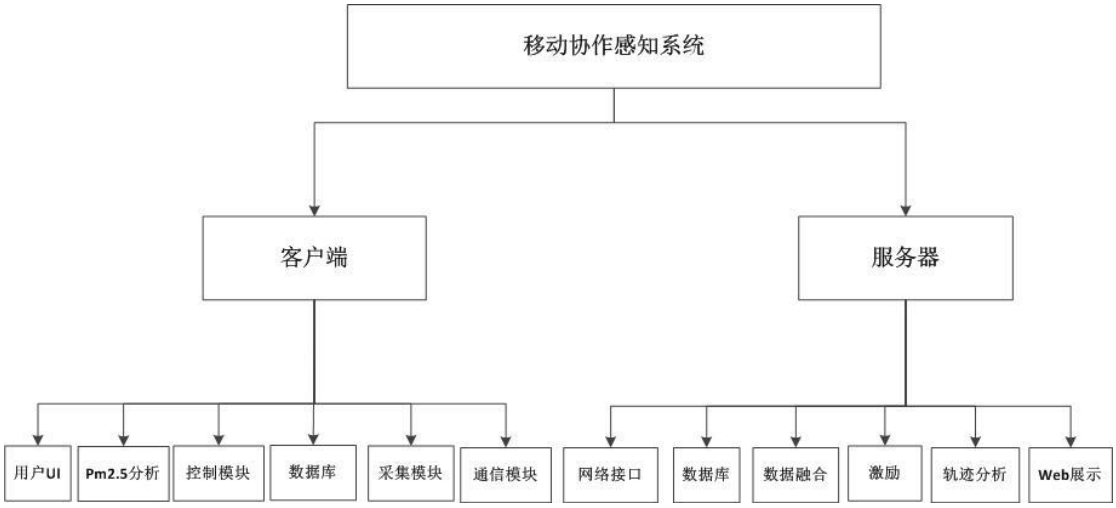


图 2-3 系统功能结构图

本移动协作感知系统搭建的感知平台，收集的感知数据包括位置、光照、声音、照片。相对于移动协作感知应用的各部分来说（图 2-2 所示），本系统数据的收集由用户 UI、采集模块、控制模块组成，经由通信模块及网络接口上传至服务器端的感知平台。基于感知数据的创新应用主要是通过照片分析 PM2.5 的值以及进行轨迹预测。此外系统中还有相应的激励模块协调与鼓励参与者完成感知任务，及数据融合模块对感知数据的融合处理。

轨迹分析模块主要的功能是对用户轨迹进行管理，并在需要进行轨迹预测时给出预测的结果。由于用户分布的不均匀性，导致感知数据的分布也会有不均匀的现象，虽然数据融合模块能够根据已有的数据计算补齐出部分缺少的数据，但在数据融合模块无法计算补齐的时候，就需要用户去缺少数据的地方收集数据。为了更有效地调度人力资源，激励模块应当选取对那些有可能去往该缺少数据地方的用户下发激励，鼓励这些用户前往收集数据。此时便需要轨迹分析模块通过对历史轨迹数据的分析挖掘，预测出可能前往缺少数据地区的用户，进而对这些用户下发激励鼓励其去收集数据。

第三章 轨迹预测的分析与研究

本章从探讨轨迹预测的可行性出发，总结出轨迹预测的应用场景及一般流程，并对现有的轨迹预测方法从各个角度进行分类比较，使得轨迹预测的过程更加直观且易于理解。

3.1 轨迹预测概述与实际可行性研究

3.1.1 轨迹预测概述

在诸多基于轨迹数据的感知应用当中，轨迹预测有着广阔的现实可用前景和较高的研究价值。轨迹预测，即通过对收集到的移动对象在过去一段时间内的移动轨迹进行挖掘分析，在需要对某个指定移动对象将来运动趋势做出预测的时刻，结合其当前的位置或前几个时刻的运动记录，预测出其接下来的运动趋势。

轨迹预测在现实生活各方面中有着较高的实用价值。轨迹预测是在对移动对象运动规律深刻理解的基础上，进而对未来趋势做出预测，因此它有着总结过去决策未来的重要意义。在实际应用中，若能预测某用户某个时刻会到达某区域，则可在提前为其推送该区域相关的优惠促销活动信息，既为用户提供了丰富的信息，又能为商家的活动广而告之。此外，由于人类的活动会给社会环境带来很大影响，因此轨迹预测还能为控制疾病的传播、对城市的合理规划、资源管理等问题提供决策依据。

轨迹预测在学术研究上的价值亦不可忽略。随着定位技术的发展以及带有定位技术的移动终端设备的普及使用，获取轨迹数据的方式变得更加便捷，这也使得对对象移动性的研究变得更加可行。然而尽管对轨迹预测的研究很多，轨迹预测的准确率却普遍不高，这使得在轨迹预测领域还有很大的改进空间和研究价值。

3.1.2 轨迹预测可行性研究

人类行为具有规律性为轨迹预测这一问题提供了理论支持。人类的运动轨迹是否可以预测，可预测程度又是如何，这些问题能从总体上理解轨迹预测的可行性，是进行轨迹预测的理论基础。

Brockmann 等人[10]的实验表明，人类的移动行为可以通过在时间和空间维度上的由两个参数的连续时间的随机游走模型表示，并且能达到较高的准确度。此外实验结果还推断出人类在地理空间维度上的移动是一个矛盾的且有效扩散的过程。Gonzalez 等人[11]的研究表明人类的轨迹在时间和空间上表现出高度的规律

性，对于每个移动对象来说，他/她很有可能会回到那几个过去频繁到访过的地方。据此，人类将来的运动趋势也可以通过对过去的移动性规律的挖掘而预测出来。为了进一步度量出人类移动性的可预测程度，Chaoming Song 等人[12]通过对匿名手机用户的移动模式的研究，在计算每个移动对象轨迹熵值的基础上，表明对象移动性的潜在可预测程度最大能达到 93%。尽管每个用户的移动模式有较大差别，但他们在可预测性程度上并没有太大差别。G. Smith 等人[13]进一步考虑到真实生活中道路拓扑的限制，因此运动的可预测性将受到此限制的影响，并不能达到 93%那么高。

总体说来，之前的研究结果都表明人类移动性的最大预测程度能到达 90%，甚至更高。尽管可能会由于未考虑到现实世界中的道路拓扑而高估了这一结果，但至少能说明人类移动的规律性是可以通过对历史轨迹数据的学习得到的，这也就为根据历史轨迹进行轨迹预测提供了理论依据。另一方面，由于轨迹预测可行性的程度有限，因此这也就可以对轨迹预测的结果进行解释，预测结果的不理想不仅和预测方法的选择有关，也与预测程度的有限性相关。换言之，即使预测的方法再好，预测结果的正确性由于预测程度的局限性也不能够达到 100%。

3.2 轨迹预测的一般过程

3.2.1 轨迹预测的应用场景

轨迹预测常见的应用场景如图 3-1 和图 3-2 所示。对于某一范围内的运动区域，在一段时间内通过移动终端上的特定客户端软件收集轨迹等感知数据，或从诸多社交网站上获得用户的签到位置信息，并将收集得到的位置数据上传至服务器，如图 3-1 所示。利用此方式经过一段时间后收集较多的真实轨迹数据信息，并在服务器端进行一系列的挖掘计算，为实现以后的预测提供依据。图中用不同颜色的曲线代表不同用户的历史轨迹。

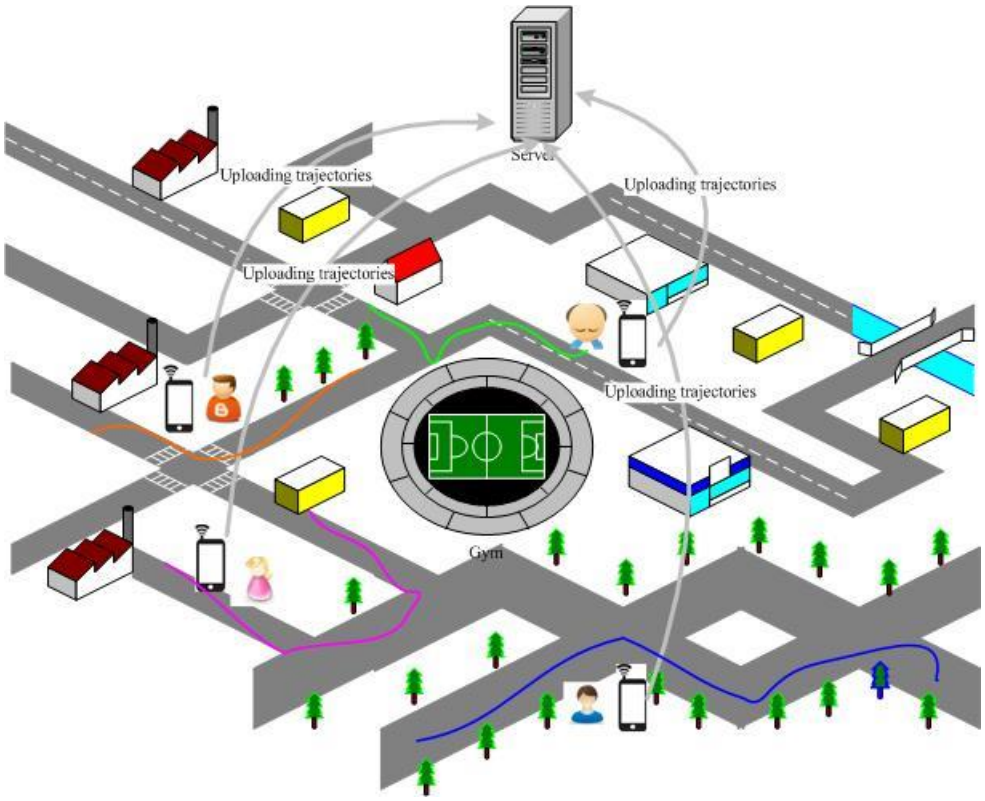


图 3-1 轨迹预测收集数据建立模型的过程

图 3-2 描述的是预测阶段的应用场景。对于处在与之前所收集数据的地点同一区域的某待预测用户，根据该用户当前的运动数据，服务器端将能够预测出其将来的运动趋势。在图 3-2 中用黄色曲线代表用户当前的真实轨迹，红色曲线表示预测出的将来运动趋势。对于待预测用户来说，他/她可以是之前已经贡献过自己的轨迹数据，也可以是之前并没有贡献过其运动数据或是第一次进入该区域。对于后一种情况，可以利用其他用户的历史数据对其进行预测。为了满足类似不同情况下的需求，现有的轨迹预测方法做出了一系列的改进，使其能够适应现实应用中的实际需求，本文将在 3.3 节进行具体分析。

轨迹预测过程中的大量计算均是在服务器端进行。在收集一定的感知数据之后，服务器端通过分析挖掘，建立相应的模型。在预测阶段，服务器接收待预测用户当前的运动情况作为输入，结合之前收集到的历史轨迹信息，返回预测的将来运动趋势作为结果。

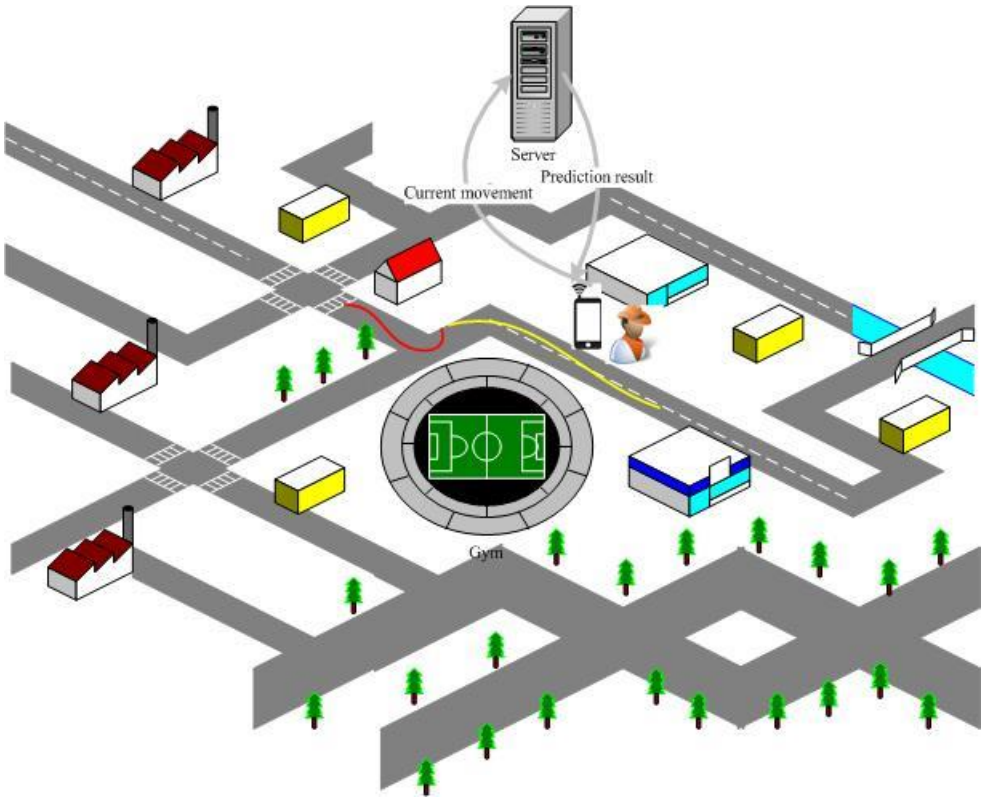


图 3-2 轨迹预测的预测场景

3.2.2 轨迹预测的整体流程

轨迹预测的一般流程包括两个阶段：建模和预测。建模阶段是指通过对历史轨迹数据的计算挖掘，从而训练出一个可用于轨迹预测的模型。预测阶段是指在需要对移动对象进行预测时，根据在建模阶段建立的预测模型和待预测对象当前的运动趋势，预测其接下来的运动趋势。轨迹预测的整体流程如图 3-3 所示。

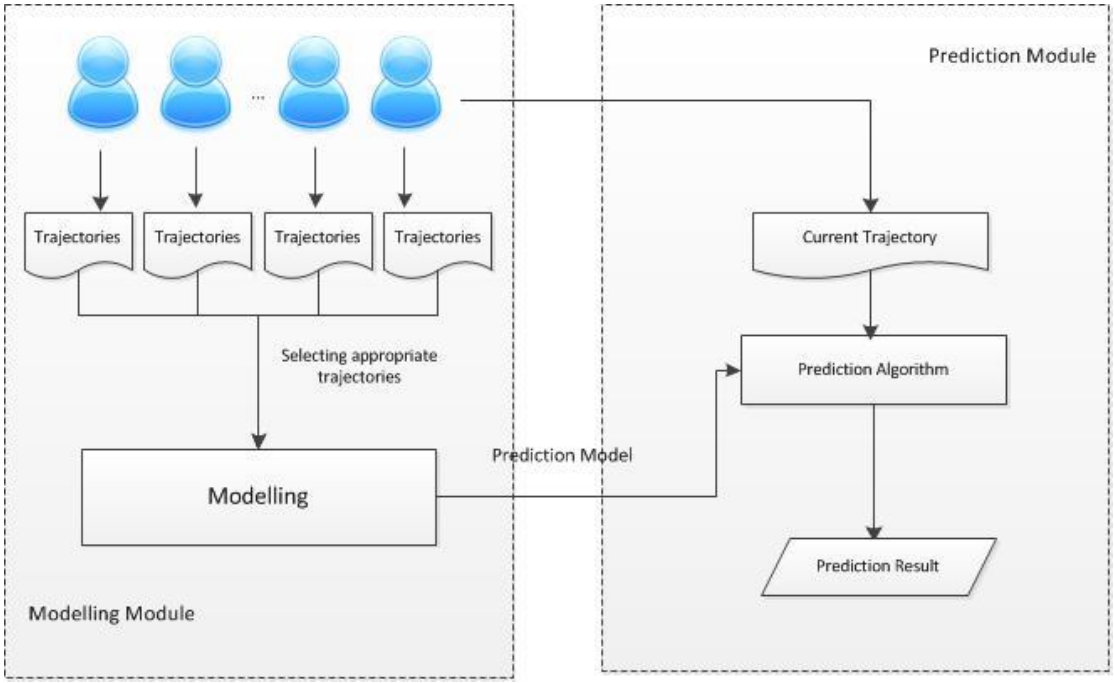


图 3-3 轨迹预测流程图

建模模块的输入是用户在一段历史时间内收集的轨迹数据。通过在收集到的轨迹数据库中选取合适的轨迹数据集建立预测模型是建模阶段的主要任务。建模阶段轨迹数据集的选取在各预测算法中有所不同，若选取的数据集不区分每个移动对象，而是将所有的收集到的轨迹数据作为输入，则建立的是一个通用的预测模型。若选取的数据集与特定对象有关，则是为每个移动对象建立一个特定的预测模型。由于建模阶段需要耗费大量的时间去进行计算，因此一般来说轨迹建模都是在线下进行，保证预测的及时性。建模阶段结束后，会得出一个预测模型，用作预测阶段的输入。

为了满足预测的实时性需求，预测阶段一般来说是在线上进行的，并且只需花费很少的时间。预测模块选取建模阶段建立的预测模型和待预测对象的当前运动数据作为输入，输出则是待预测对象在将来时间的运动趋势。预测的结果可以用作前述各应用的决策依据。此外，在一些预测方法中，预测的结果还可以反过来优化预测模型。

目前大部分的轨迹预测方法都是基于图 3-3 所示的流程进行的，不同的方法在具体实现时对各部分采取的实现方式有所不同。根据预测方法中输入输出的内容、所使用的预测算法、考虑的信息维度及建模时选择的数据集的规模，将在 3.3 节中基于现有的轨迹方法进行具体分类比较。

3.3 对现有轨迹预测方法的分析

轨迹预测的整体流程如图 3-3 所示，但不同的方法在具体实现时有所不同，这也就使得现有的轨迹预测方法具有多样性。针对现有的一些轨迹预测方法，从不同的角度对其进行比较分析，一方面能更好地理解轨迹预测，另一方面可以通过对不同的方式进行比较，分析其各自的特点，从而在实际实现轨迹预测时根据需求选择合适的方法。表 3-1 分别从各个方面对不同的轨迹预测方法进行了整体的比较，后续小节将针对每一点进行具体介绍。

表 3-1 对现有轨迹预测方法的整体比较概览

轨迹预测方法	输入输出	预测模型	考虑的信息维度	数据集规模
Mining Geographic-Temporal-Semantic Patterns in Trajectories for Location Prediction[9]	输入：当前运动趋势 输出：下个位置	模式匹配模型	时间+空间+语义	整个数据集
Distant-time locationprediction inlow-samplingratetrajectories[14]	输入：当前的位置和时刻，需要查询的时刻 输出：需查询时刻所处的位置	状态空间模型	时间+空间	相关数据集
Exploring Spatial-Temporal Trajectory Model forLocation Prediction[15]	输入：当前的位置和时刻，需要查询的时刻 输出：需查询时刻所处的位置	模式匹配模型	时间+空间	整个数据集
Prediction of Moving Object Location Based onFrequent Trajectories[16]	输入：最近的运动趋势 输出：将来可能的路径	模式匹配模型	空间	整个数据集

表 3-1 对现有轨迹预测方法的整体比较概览(续上表)

轨迹预测方法	输入输出	预测模型	考虑的信息维度	数据集规模
Mining Frequent Trajectories of Moving Objects forLocation Prediction[17]	输入：最近的运动趋势 输出：将来可能的路径	模式匹配模型	空间	整个数据集
A “Semi-Lazy” Approach to Probabilistic Path Predictionin Dynamic Environments[18]	输入：当前运动最后 h 步 输出：将来的路径	状态空间模型	时间+空间	相关数据集
Predicting Future Locations with Hidden Markov Models[19]	输入：当前轨迹 输出：下个可能位置	状态空间模型	时间+空间	整个数据集
Next Place Prediction using Mobility Markov Chains[20]	输入：最近的几个位置点 输出：下个可能的位置	状态空间模型	空间	个人数据集
Pedestrian-movement Predictionbased on Mixed Markov-chain Model[21]	输入：当前轨迹 输出：下个可能位置	状态空间模型	时间+空间	整个数据集
A Mixed Autoregressive Hidden-Markov-chain ModelApplied to People’s Movements[22]	输入：当前轨迹 输出：下个可能位置	状态空间模型	时间+空间	整个数据集
Destination Prediction by Sub-Trajectory Synthesisand Privacy Protection Against Such Prediction[23]	输入：当前轨迹 输出：可能的目的地	状态空间模型	空间	整个数据集

表 3-1 对现有轨迹预测方法的整体比较概览(续上表)

轨迹预测方法	输入输出	预测模型	考虑的信息维度	数据集规模
A hybrid prediction model for moving objects[24]	输入：当前的位置和时刻，需要查询的时刻 输出：需查询时刻所处的位置	模式匹配模型	时间+空间	个人数据集
Intelligent Trajectory Classification for ImprovedMovement Prediction[8]	输入：当前轨迹 输出：可能的路径	模式匹配模型	空间	整个数据集
Semantic Trajectory Mining for Location Prediction[25]	输入：当前轨迹 输出：下个可能位置	模式匹配模型	空间+语义	整个数据集
Friendship and Mobility: User Movement InLocation-Based Social Networks[34]	输入：需要查询的时刻 输出：需查询时刻所处的位置	状态空间模型	时间+空间+社交关系	相关数据集
NextCell: Predicting Location Using SocialInterplay from Cell Phone Traces[35]	输入：将来一段时间 输出：将来一段时间内经过的位置列表	状态空间模型	时间+空间+社交关系	相关数据集
Interdependence and predictability of human mobility and social interactions[36]	输入：需要查询的时刻 输出：需查询时刻所处的位置	模式匹配模型	时间+空间+社交关系	相关数据集

3.3.1 预测的输入输出

从表 3-1 中可以看出，根据不同的实际需要，轨迹预测方法的输入输出有所不同。因此，可以根据预测的输入输出，对轨迹预测方法进行分类。

轨迹预测方法的输入一般来说是由两个部分组成，即建模阶段建立的预测模

型和待预测用户当前的运动情况。不同之处在于不同的预测方法对当前用户运动情况所需的信息不同。第一种是需要待预测用户当前所处的位置信息[14,15,24]，第二种是需要用户在当前预测时间之前的运动轨迹，而不是单独的某个位置[8,9,16,17,21,22,23,25]，第三种需要的是用户当前所处的位置及当前位置之前的某几个位置[18,20]，与第二种不同的是，这种方法选择的是之前的若干个位置点，而不是所有的轨迹数据。

预测的输出总体来说有两类，一类是预测某个位置点，一类是预测将来的运动轨迹。而为了满足实际不同应用场景的需要，轨迹预测的输出可以为：

⑴下个位置点[9,19,20,21,22,25]。即当前所处时间点的下个时间点的位置，根据预测粒度的不同，预测的结果可能是一个区域或者是一个代表地点，也可能只是所处的某个状态。

⑵查询某时刻所在的位置[14,15,24,34,35,36]。此时不仅只是考虑空间维度，还涉及到时间维度的信息。能满足获取到某个特定时刻下某移动对象可能所处地点的应用需求。此种情况下在输入方面可能也需要提供当前的时间信息，如在文章[14,15,24]中所提到的方法，输入是待预测对象当前所处的位置和当前的时间，以及需要查询的时间，预测的结果是在该查询的时间点待预测对象所处的位置。而有些方法[34,35,36]只需要以待查询时刻为输入，便可预测出在该时刻可能所处的位置。

⑶目的地位置[23]。同样返回的是位置信息，但这种返回的是移动对象此次运动的目的地。目的地不同于在行进过程中经过的位置点，它能代表着移动对象一定的意图和心理目标，也有理由认为该移动对象在目的地所逗留的时间将长于经过的其他位置点，因此这种类型的输出在实际应用中更会有一定的实用价值。

⑷将来的运动轨迹[8,16,17,18]。前三种预测结果都是离散的位置点，而这种输出是将来的一段运动轨迹，而非某个单独的点。

不同的输入输出方式能满足轨迹预测不同应用场景的需要，现有的轨迹预测方法也是在该分类基础上对输入输出的信息进行组合，以期实现不同的效果。

3.3.2 预测算法的类型

轨迹预测前期的工作就是通过对历史轨迹数据的挖掘，建立轨迹预测的模型。基于对现有轨迹预测方法的研究，可将预测模型分为两大类：状态空间模型和模式匹配模型。

一般来说，用手机等移动设备收集上来的原始轨迹是由多个位置点组成的序列表示而成，令 $T=\prec P_1,P_2\cdots P_n>(n\geq 1)$ 表示一条由 n 个点构成的轨迹序列，其中 $P_i=(x_i,y_i,t_i\cdots)(1\leq i\leq n)$，至少包含了纬度 x_i、经度 y_i 和时间 t_i。由于原始数据中的位置点信息太过精确，粒度太小，不适宜从中发现规律。因此在进行轨迹数据

规律挖掘之前，需将其表示粒度进行放大。常用的方法是将实验区域划分成若干个等大的小区域，在此基础上将具体的位置点映射到小区域中，如此一来，原本用具体位置表示的轨迹就可以用划分出的粗粒度区域表示，规律性也就较易发现。

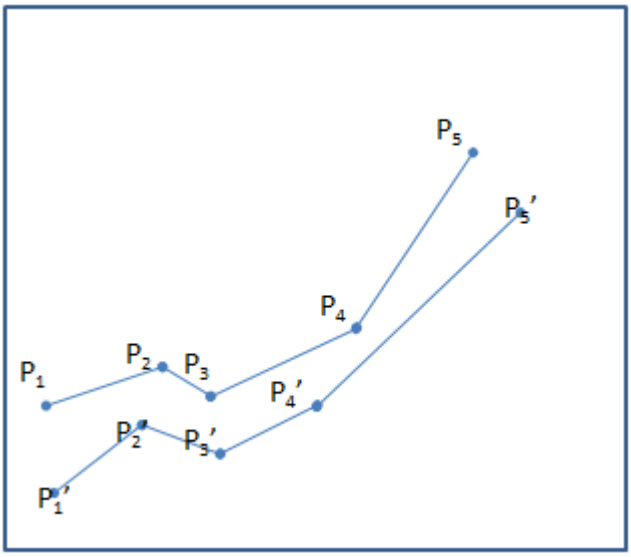


图 3-4 原始轨迹图

图 3-4 中展示的是两条以原始数据表示的轨迹，轨迹$T_1 = < P_1,...,P_5 >$和轨迹$T_1 = < P_1^{'},...,P_5^{'} >$两条轨迹没有任何交汇点，若以原始轨迹数据的角度看，这两条轨迹之间没有任何关系。但其实他们的位置点都比较接近，并且有着相同的走势。在划分子区域之后，可以看到如图 3-5 所示，两条轨迹均可以用$< S_3, S_4, S_2 >$表示。大部分轨迹预测算法均是在对原始数据进行这种处理后的基础上进行的，也有一部分其他方法是采取别的策略将位置数据的粒度进行放大（如聚类成一个区域），但其根本目的都是为了解决原始数据粒度太小不易发现规律的问题。

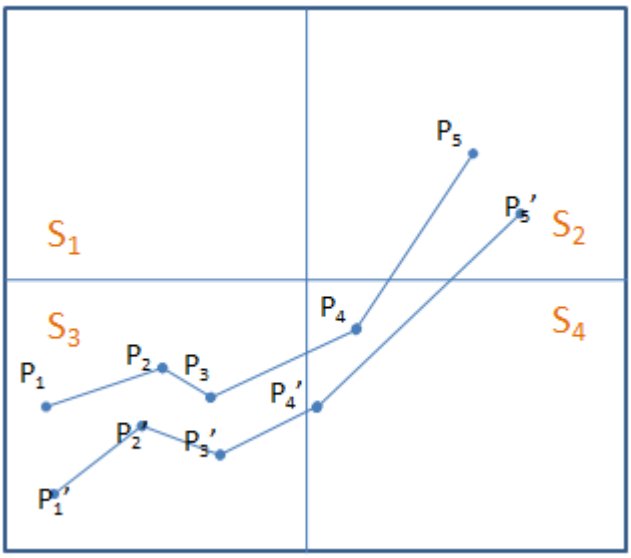


图 3-5 划分区域后的轨迹表示

1. 状态空间模型

在状态空间模型中，每条轨迹都是由多个位置点按时间顺序组成的序列，其中的位置点是为了解决原始数据粒度太小的问题而处理之后的数据，每个位置点被称之为状态。因此，在状态空间模型中，用户在任一时刻都处于其中的一个状态上，该状态反应出相关的位置信息，所有的状态组成了状态空间。一般情况下通过对历史轨迹数据的收集、挖掘分析，可以统计出各个状态之间的转移概率，并存储在转移矩阵中。当需要进行轨迹预测时，根据用户当前所处的状态，查询概率转移矩阵，计算由当前状态转移到其他状态的概率，并认为获得最大转移概率的状态即为下个可能到达的状态。

典型的使用状态空间模型的实现方法是利用马尔科夫特性，基于马尔科夫链计算转移概率。最简单的情况是只考虑空间位置信息。Sebastien Gambs 和 Marc-Olivier Killijian 等人[20]提出了一种基于移动马尔科夫链 Mobility Markov Chain(MMC)的实现方法，其考虑了用户访问的前 n 个位置。由于用户下个位置不仅与当前所处状态有关，很多情况下也与之前的若干个状态有关，因此有理由考虑前 n 个状态下的转移情况。假设 H 代表 Home，W 代表 Work，O 代表 Other，取 n = 2，如表 3-2 记录的为某用户的转移概率矩阵，若当前位置为 W，上一个位置为 H，则下一个转移到位置 H 的概率为 1.00，到其他位置的概率为 0.00。在实现时首先采取了聚类的方法，将原始的位置数据聚类成粒度大的 POI 点，再将轨迹用这些 POI 点表示。通过对用 POI 表示的轨迹的分析，计算出每个状态之间的转移概率，每个状态是由前一个 POI 和现在的 POI 表示的，如表 3-2 所示。图 3-6 描述了状态之间的转移图。文章中介绍到后续的实验结果表明，无论是对于一个用户还是所有用户，预测准确性都是在 n=2 时达到最优，且在 70%到 95%之间。

表 3-2 某用户转移概率矩阵[20]

状态	H	W	L	O
HW	1.00	0.00	0.00	0.00
HL	1.00	0.00	0.00	0.00
HO	0.64	0.34	0.00	0.00
WH	0.00	0.84	0.08	0.08
LH	0.00	0.50	0.00	0.50
OH	0.00	1.00	0.00	0.00
OW	1.00	0.00	0.00	0.00

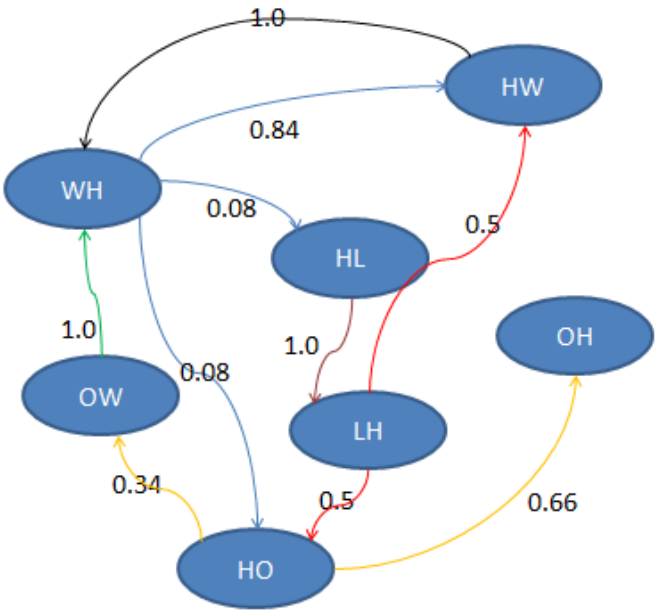


图 3-6 某用户转移概率图[20]

在实际生活中，有些隐含的上下文变量也会部分影响到用户对于下个访问位置所做出的决定，这些隐含的变量可以是用户个人的心理活动和目标驱动等。因此，另外一种用于处理这种隐含因素的典型状态空间模型被提出。Wesley Mathew 等人[19]提出利用隐马尔科夫模型 Hidden Markov Model(HMM)来预测移动对象的将来位置。在隐马尔科夫模型 HMM 中，决定用户选择下个访问位置的状态因素是不可观察的（隐含的），而所能观察到的实际访问位置却依赖于这些不可见的状态。除了在这些不可见的状态之间分布着转移概率之外，每个不可见的状态对于实际访问位置上还分布着概率。图 3-7[19]描述了隐马尔科夫的一般结构，x1, x2, x3 表示的是隐含状态，每个隐含状态之间存在着转移概率。在 t 时刻所处的状态 $x(t)$ 的条件概率保持着马尔科夫特性，其只受前一个状态 $x(t-1)$ 影响。也就是说，在 $t-2$ 及之前时刻所处的状态对 $x(t)$ 无影响。变量 y1, y2, y3, y4 表示的是实际可见的位置状态，每个隐含状态对于这些可见状态都有着概率分布。用户在某个时刻所处的位置只受其当前的隐含状态影响。实验部分使用 GeoLife 数据集，采取不同的配置参数，结果表明最好情况下预测准确率能达到 13.85%。

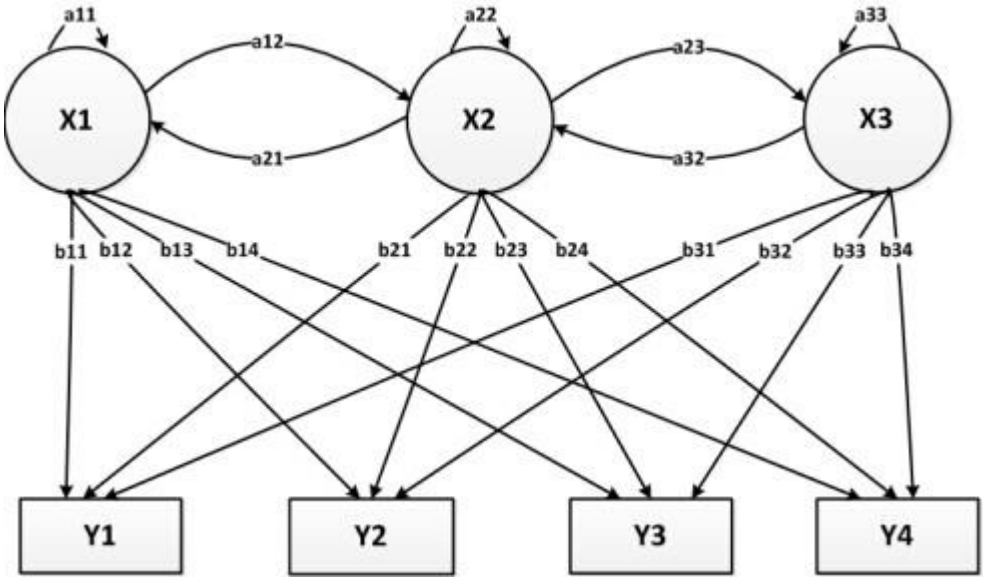


图 3-7 隐马尔科夫模型 HMM[19]

上述的两种模型都只建立了一个通用的模型，这是基于所有的用户都能适用于一个通用模型的假设之上的。然而在现实生活中，人类的活动具有多样性，每个人的个性、行为及心理活动都有所不同，使用一个通用的模型来进行预测是不现实的。因此，Akinori Asahara[21]等人在此基础上提出混合马尔科夫链模型 Mixed Markov-chain Model(MMM)的概念，把用户的个性考虑其中，将其作为一个不可见的决定因素。同时，也把用户之前所处状态的影响考虑其中。图 3-8 所示的是混合马尔科夫链模型，z 表示的是影响可见状态的不可见因素。根据 z 的不同可将用户进行分组，每组的用户行为被认为是相似的，会表现出相似的行为。z 决定了使用哪个模型来生成状态间的转移概率。与隐马尔科夫链不同的是，在隐马尔科夫链模型中，不可见状态是可变的，各状态之间可以相互转移，而在混合马尔科夫链模型中，不可见状态在转移过程中是固定的，代表了用户所属的固定分类。因此，在混合马尔科夫链模型中，相似的用户被认为是一组，具有相同的转移概率分布。在由 z 确定转移概率分布之后，根据移动对象之前所处的状态，预测该移动对象下一个可能的位置。研究[21]中基于在一个步行街收集的真实的数据集之上进行实验验证，证明混合马尔科夫链模型最高的预测准确率能达到 74.4%，并且基于同一个数据集用一般马尔科夫链模型和隐马尔科夫模型进行比较实验，发现其预测准确率分别为 45%和 2%。与一般马尔科夫链模型和隐马尔科夫模型相比，混合马尔科夫链能达到较好的预测效果。混合马尔科夫链模型将相似的用户看作一组，每组用户有相同的转移概率分布，而不是将所有用户都使用一个通用的概率分布模型，也不是单独为每一个用户建立概率分布模型，在效率和准确率上表现的更好。

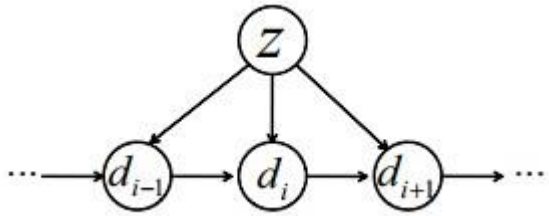


图 3-8 混合马尔科夫链模型 MMM[21]

在隐马尔科夫链模型 HMM 中，下一个位置只受隐含状态的影响，而与之前所处的状态无关。然而在现实中，人们将去的下一个位置不仅仅受当时的隐含因素（如心理状态或目标驱动）所影响，还与其当前所处的位置有关，轨迹上的连续位置点之间具有相关联性。因此，一种自回归隐马尔科夫链模型 AutoregressiveHidde-Markov-chain Model(AR-HMM)被提出来解决这一问题。如图 3-9[22]所示，在自回归隐马尔科夫链模型 AR-HMM 中，下个位置 d_i 不仅受隐含状态 Z_i 影响，还受前一个位置 d_{i-1} 所影响。进一步，Akinori Asahara[22]在此基础上，借鉴混合马尔科夫链模型 MMM 的特征，提出混合自回归隐马尔科夫模型 Mixed Autoregressive Hidden-Markov-chain Model(MAR-HMM)，考虑相似的用户具有类似的行为，受不可见因素的影响。图 3-10[22]所示的是混合自回归隐马尔科夫链模型 MAR-HMM 结构，可见状态之间具有相关性，且有个稳定的、在转移过程中不会改变的不可见因素存在，它反应了一组用户的特征。在实验中，他们在同一个数据集上对四种模型（马尔科夫链模型 MM (Markov-chain Model)，混合马尔科夫链模型 MMM，自回归隐马尔科夫链模型 AR-HMM 及混合自回归隐马尔科夫链模型 MAR-HMM）进行评估，实验结果表明其预测率分别为 35.1%，49.2%，51.5%，56.8%，证明了结合 HMM 和 MMM 特性的混合自回归隐马尔科夫链模型 MAR-HMM 在预测过程中具有较好的表现。

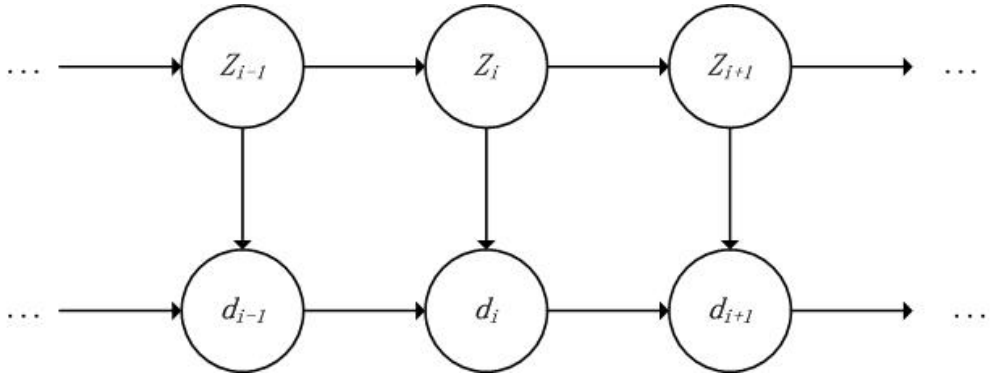


图 3-9 自回归隐马尔科夫链模型 Autoregressive Hidde-Markov-chainModel(AR-HMM)[22]

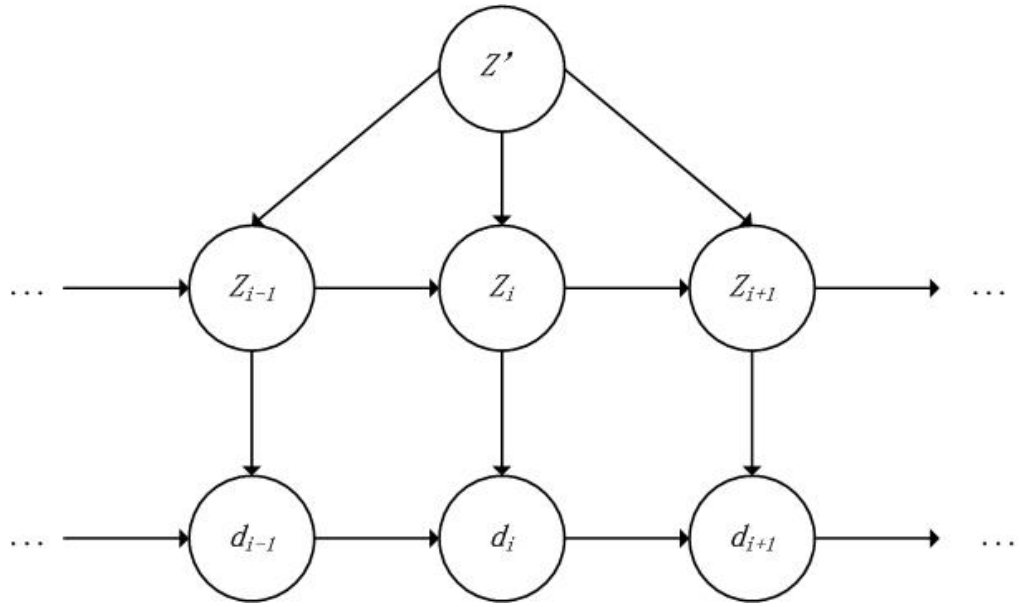


图 3-10 混合自回归隐马尔科夫模型 Mixed Autoregressive Hidden-Markov-chain Model(MAR-HMM)[22]

状态空间模型具体的实现形式主要是马尔科夫链模型 Markov-chain Model(MM)，隐马尔科夫链模型 Hidden-Markov-chain Model(HMM)以及混合马尔科夫链模型 Mixed Markov-chain Model(MMM)。许多方法都是在此基础上进行改进，例如在使用这些模型之前，选择更加合适的历史数据进行建模来优化预测的准确率等。Jingbo Zhou[18]等人提出一种动态选择轨迹来建模的方法，以改进预测方法的效率。在这种方法中，不是将所有的轨迹都选择用来建模，而是将当前轨迹与存储的历史轨迹进行匹配比较，历史轨迹库中与当前轨迹有着相似匹配的轨迹将被选择作为建模时参考的轨迹。然后在此基础上构建本地 HMM 模型，进而预测目标对象的将来移动情况。用这种方法建立的模型与待预测对象具有高度的相关性，在预测时能达到较好的准确率。另一方面，由于学习建模是在小部分具有相关性轨迹的基础上进行的，相比于所有轨迹集来说数量有所减少，因此可以使用较为复杂的学习算法进行建模挖掘。Meng-Fen Chiang 等人[14]利用马尔科夫链模型来处理低采样率情况下的预测问题。通常情况下，签到数据及带有位置信息的照片数据都能反映出用户一定的行为规律，然而其采样间隔频率可能会相隔几小时甚至是几天。在这种情况下，首先将轨迹表示成用粗粒度的区域构成的序列，而这些区域可通过使用现有的聚类算法对位置信息进行聚类得到。继而这些区域被看作是转移状态，并作为转移矩阵的结点，由历史数据计算出各状态之间的转移概率，构造马尔科夫链模型。Andy Yuan Xue[23]等人为了解决数据稀缺的问题，提出用子轨迹合成的方法构建马尔科夫链模型来进行目的地预测。在方

法[34,35]中将用户所处的位置看作是离散的潜在状态，不同于前述的方法计算状态间的转移概率，而是根据历史数据中在各位置上的概率大小总结出其在空间和时间维度上的周期性，并结合用户的社交关系，提高预测的准确率。

2. 模式匹配模型

模式匹配是轨迹预测方法中另一种常见的预测模型，主要包括两个阶段：一是挖掘频繁轨迹和运动模式，二是将当前轨迹与运动模式进行匹配，利用匹配的模式进行预测。频繁轨迹即在历史轨迹中频繁出现的、可以用作预测时作为参考的轨迹。频繁轨迹和运动模式挖掘的方法最初来源于频繁项集的挖掘和关联规则的学习算法 Apriori[26]，其通过引入支持度 *support* 和置信度 *confidence* 的概念来获得频繁项集和关联规则：支持度 *support* 大于最小支持度 *minsup* 的项集为频繁项集，置信度 *confidence* 大于最小置信度 *minconf* 的规则为关联规则。在与运动模式匹配的阶段，诸多学者也提出了多种计算相似度的方法来判断当前轨迹与运动模式是否匹配，根据考虑的信息维度和具体实现方法的不同，可采用不同的方法来进行相似性度量。

Morzy[16]最初在 Apriori 算法的基础上提出类似于 Apriori 且适用于轨迹预测的算法 AprioriTraj，通过计算支持度和置信度，从移动对象轨迹存储库中挖掘频繁轨迹并提取运动规则。头部与当前轨迹相匹配的运动规则被用作预测，并将该运动规则的尾部作为预测结果的候选项。为了解决原始轨迹数据粒度过小的问题，该方法将运动区域划分为固定大小的子区域，并据此表示用户轨迹。在模式匹配阶段，作者提出了四种运动规则匹配的方法。在实验阶段，使用的数据是由基于网络的移动对象生成器[27]生成的伪数据，并通过改变参数最小支持度、子区域大小和移动对象数量的值来观察实验结果。结果表明运动规则的数量随着最小支持度、子区域大小和移动对象数量的增大而均表现出减小的趋势。然而，文章却没有对预测结果的准确率进行度量，只是在结果中分析了运动规则数量与其他各参数之间的变化关系。由于实验数据是合成的，而不是真实数据，因此由四种匹配策略带来的对结果的影响是不可靠的，文章也并没有对这种由这四种匹配策略的表现进行比较。

为了继续并扩展上述研究结果，Morzy[17]进一步提出另外一种实现算法 *Traj-PrefixSpan* 来挖掘频繁轨迹，并修改了 FP-Tree 索引结构来实现快速地查找轨迹工作。此外，在这次实验中，作者提出了一种方法来度量预测算法的准确率，并对四种模式匹配方法进行了比较。结果表明该方法的预测准确率能达到 80%，并且随着最小支持度阈值的减小，预测算法的质量有所提高。

上述的模式匹配方法只考虑了空间地理因素，然而在实际应用中，时间因素

在采集位置信息时也会被同时采集到，而且时间也是影响运动轨迹的一个重要因素，因此，应当充分利用采集到的时间信息进行更加精确而实用的轨迹预测。Hoyoung Jeung 等人[24]在 Apriori 算法的基础上加以修改，考虑进时间因素，使得预测算法能适应时间的要求，而不是像之前的算法仅仅能预测接下来的运动趋势。如图 3-11 所示[24]是修改之后的轨迹模式，类似于 Apriori 算法中的关联规则，不同的是有时间轴表示的时间因素。$R_{t1}^{j_1} \wedge R_{t2}^{j_2} \wedge ... \wedge R_{tm}^{j_m} \xrightarrow{c} R_{tn}^{j_n}$ 是一条轨迹模式，其中 $t_1 < t_2 < ... < t_m < t_n$ 是时间变量，和 Apriori 中的关联规则类似，c 是这条轨迹模式的置信度。在实验阶段，通过生成四个合成的数据集，与其他几种使用运动函数来预测的方法进行比较，并观察当参数变换时预测的准确率变化情况。Po-Ruey Lei 等人[15]将含有空间和时间信息的频繁模式存储在其提出的时空轨迹模型 Spatial-temporal Trajectory Model(STT)中，如图 3-12 所示[15]。时空轨迹模型 STT 表示为一棵概率后缀树，在这之前仍然是将原始的轨迹数据转换成频繁的区域表示，频繁区域表示在该区域内有满足一定数量的轨迹段穿过该区域。在概率后缀树中，每条边表示一个频繁区域，每个树节点是一个频繁区域序列，表示从该节点到根节点的一条路径。每个树节点关联着一张包括空间和时间信息的预测表。其中 next 表示下一区域，$\Pr o_{spatial}$ 表示空间条件概率，$\Pr o_{temporal}$ 表示时间条件概率，时间间隔由(平均间隔，平均间隔偏差)组成，Time 项表示到达该节点的典型时间。在预测阶段，根据提出的计算当前轨迹与树节点序列相似性的方法，定位到与当前轨迹最相似的序列树节点，并利用此节点进行接下来的预测。在实验中使用的是来自于 CarWeb[28]的真实数据集，并与文章[24]中提出的方法在预测准确性上进行比较。这两种方法[15,24]均考虑了时间因素，并在预测结果的表现形式上都是可以返回一个在特定查询时间的位置点。后者[15]在实验中使用真实的数据集与前者[24]所提出的方法进行了比较，并证明了能有较好的预测效果。

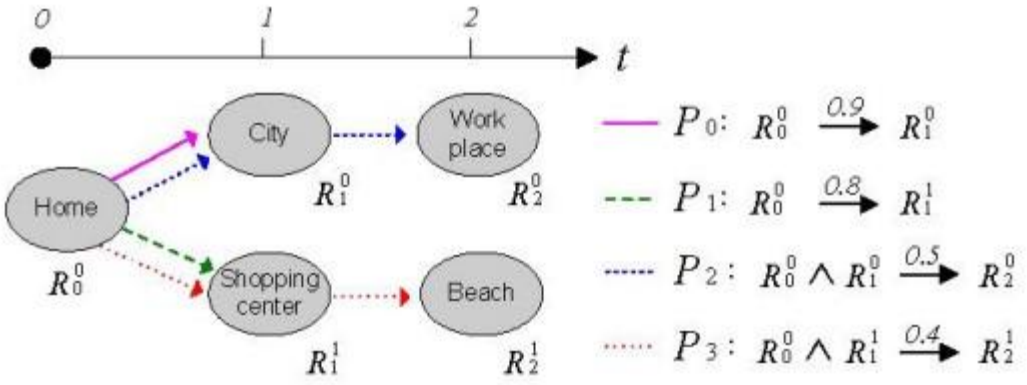


图 3-11 轨迹模式示例[24]

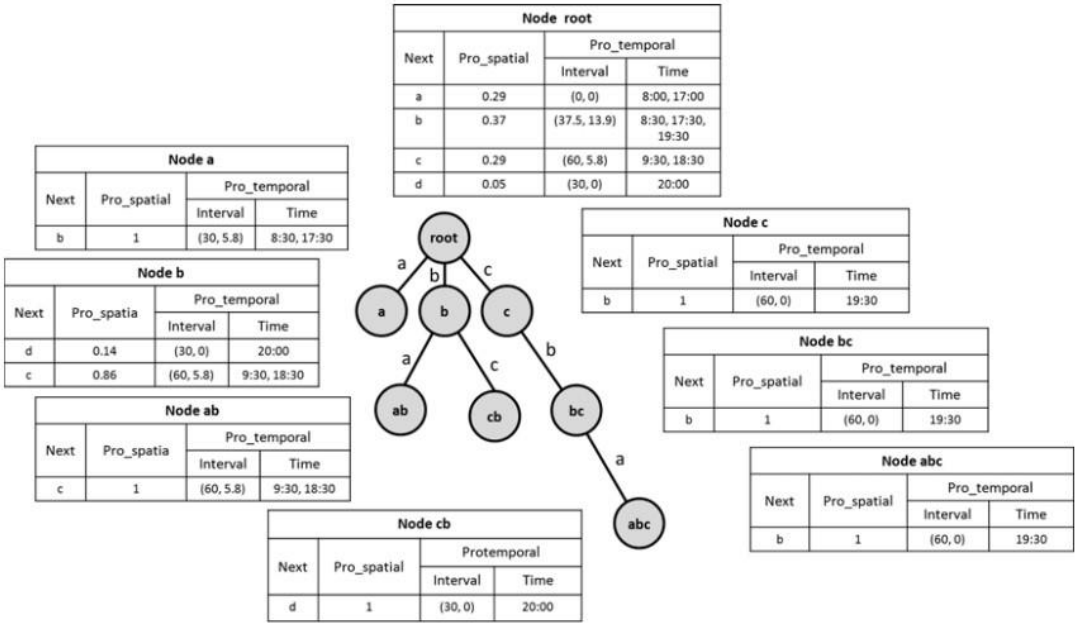


图 3-12 时空轨迹模型 Spatial-temporal Trajectory Model(STT)[15]

原始轨迹数据是由一系列的位置点序列组成，而位置点信息只包含了经度、纬度、海拔、采集时间等。如果在此基础上加上额外的能够描述位置相关属性的信息，便能够更好地理解轨迹。JOSH JIA-CHING YING 等人[9]提出一种新型的轨迹模式，即时空语义 GTS(Geographical-Temporal-Semantic)模式，不仅考虑到空间和时间信息，还加入了语义信息。语义信息的添加使得原本没有意义的数据被赋予了一定的含义，能够帮助理解数据。在这种方法下，首要的工作就是挖掘由这三种信息所驱动产生的频繁模式。之后作者使用了几种匹配策略来计算当前运动与所挖掘的 GTS 模式之间的相似性。实验部分由于需要带有语义信息的轨迹数据，因此选择的是由两个能签到和上传照片的社交网站上提供的用户数据。除了在改变各参数的情况下观察这种位置预测方法的准确性，还将此方法与其他的几种方法进行了比较。结果表明带有时间空间语义信息的轨迹预测方法比目前的方法都要好。

上面介绍的频繁模式匹配的预测方法中，频繁模式的挖掘都是在线下训练得到的，之后在预测阶段被使用。这也是大部分轨迹预测方法所经历的阶段：建模和预测。然而有研究尝试使用在线训练频繁模式的方法进行轨迹预测。Christos Anagnostopoulos 等人[8]提出在线学习机制的方法去更新模式库。初始状态模式库被初始化为只含有一个轨迹模式，在后续的预测过程中，通过不断地更新模式库，使得模式库中保存的模式尽量精简而又有较好的代表性。作者提出一种高级的轨迹模式分类方法，其放松了轨迹分类的严格要求，使得部分相似的模式就被看作

是相似模式。轨迹模式分类的流程如图 3-13[8]所示,假设 Q 是新的待预测轨迹，首先将其与模式库中的模式进行部分匹配，若没有任何模式可以和 Q 匹配，则将 Q 作为新的轨迹模式添加到模式库里。若模式库中的模式 P 与 Q 相匹配，则将 P 作为此次预测的参考轨迹，若预测失败，同样将 Q 作为新的轨迹模式插入到模式库中。若预测成功，为了保证模式库的精简性，需要保留最具代表性的轨迹模式，故而将 P 与 Q 进行比较，若新轨迹 Q 比原有的模式 P 更具有代表性，则将 P 替换为 Q。实验中使用了四个不同采样率的真实轨迹数据集和一个合成的数据集，以验证该分类方法的有效性，此外还和其他的几种预测方法进行了比较。这些方法包括 1）使用在线训练运动规则的模式匹配方法[29][30]；2）采用线下建模的 HMM 模型预测方法[19]。实验不仅将状态空间模型和模式匹配模型进行了比较，还将提出的这种在线预测的方法同其他的在线训练方法进行了比较，以说明这种在线学习轨迹分类方法的有效性。

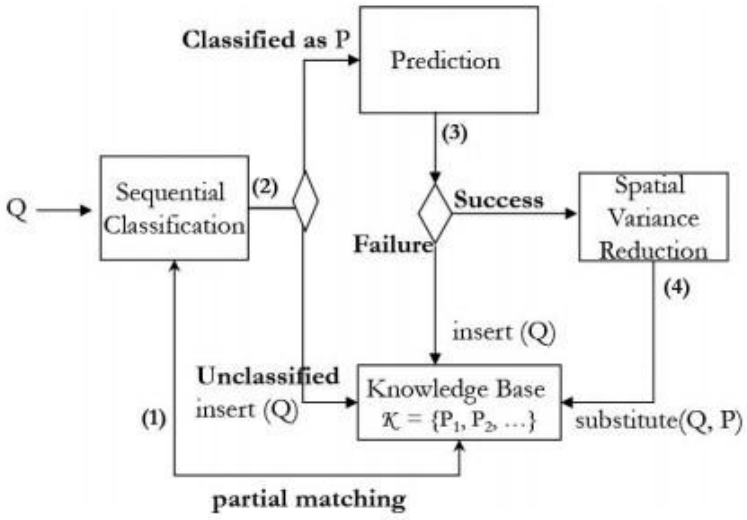


图 3-13 轨迹模式分类的流程[8]

Manlio De Domenico 等人[36]采用的多变量非线性预测的的方法[38]本质上也是采用模式匹配的思想：查找历史轨迹中与当前运动趋势相似的匹配轨迹，并认为若给定当前所处的状态，则下个可能状态与其匹配的历史轨迹中下个状态有极大的可能性是相同的。在此基础上，文章中[36]还考虑了社交关系对预测结果的影响，并提出 *mutual information* 的概念去计算一对用户之间的关联性，在没有社交关系信息的情况下，可通过 *mutual information* 去计算推断两用户之间的关联性，从而提高预测的准确率。

无论是状态空间模型还是模式匹配模型，尽管在各方法的实验中衡量预测结果准确率的度量有所不同，但都表明轨迹预测的准确率不是很高。而人类移动性

的可预测程度能达到90%左右，因此在轨迹预测方面仍有较大的改进空间。在大部分的轨迹预测方法中，大量的时间主要都花费在线下训练预测模型上，而预测阶段为了满足实时的要求只需花费很少量的时间。尽管有些方法[8][29][30]是通过在线学习的机制逐步更新轨迹模式库，但这可能会导致在初始阶段会有较长的一段时间因为没有模式与当前运动相匹配而无法预测的情况。只有在经过一段时间对模式库的更新，才能有足够数量的轨迹模式保证预测的正常进行。

状态空间模型与模式匹配模型各适用于不同的场景。状态空间模型适合于对离散位置点的预测，并有多种方法基于状态空间模型进行了改进，如实现了对下个位置的预测以及与时间有关的预测等。相对而言，模式匹配模型更加适合于对将来连续轨迹的预测及目的地的预测。在使用模式匹配的方法时，若历史轨迹数据不够多可能会导致没有存储的轨迹模式与当前运动相匹配的情况，从而预测会无法进行。而对于状态空间模型来说，缺少历史轨迹数据会使得预测模型不够精确，尽管不能保证结果的准确性，但很少会发生无法预测的情况。

3.3.3 预测过程考虑的信息维度

随着定位技术的发展，记录移动对象的运动轨迹变得更加方便。在轨迹数据中，经度、纬度等空间信息一定会被记录，同时采集位置点的时间也很容易得到。另一方面，还可以结合位置点的特点及其分类属性，可以为其添加标签类别。这种额外添加的能帮助更好地理解位置数据的信息被称之为语义信息。根据预测方法中涉及的不同维度信息，可将其分为如下几种类别：

1. 空间信息

空间地理信息在轨迹预测过程中一定是必需的，主要包含了经度、纬度信息。早期的轨迹预测方法[8,16,17,20,23]都只考虑了空间信息在预测中的影响，尽管时间信息也会同时被采集，但只是将时间作为判别位置顺序的依据，并没有将时间维度的信息真正应用到预测中。同样，预测的结果也只是仅有空间信息的位置，而并不包含时间信息。由 Morzy 等人在合成的轨迹数据集上的实验表明预测准确率能达到 80%[17]。Gambs 等人在真实的轨迹数据集上进行了实验，结果显示出准确率在 70%到 95%之间。然而，只考虑空间信息的预测方法只能进行短时间内的预测，即预测接下来的运动情况，不能满足预测某个待预测时刻时的位置，而且不能估计到达的时间。因此，尽管只考虑空间信息能简化预测过程和减少存储空间，但并不适合于实际应用中的情况。

2. 时间+空间信息

在采集位置数据的时候，时间信息也会一并被采集，然而早期的很多预测方法并没有好好利用而是丢弃了时间信息。加入时间信息不仅能有效地提高预测的

准确率，还能够使得预测方法能够满足更多地应用场景，使预测不仅局限于短时间内的位置预测。例如能够预测在某个待查询时刻的位置，或预测到达某个地点的大概时间等。

与仅考虑空间信息的预测方法相比，时间加空间能够使预测满足长时间间隔预测的应用场景。如文章[14,15,24]中提出的方法均能解决长时间间隔的预测问题，弥补了仅考虑空间信息下短时间间隔的预测问题。而且利用时间信息，由Meng-Fen Chiang 等人提出的方法[14]还能适用于历史轨迹数据稀缺的情况，即数据采集的频率较低，两个相邻位置点之间可能相隔数小时甚至是几天。在实验中，作者基于两个真实的数据集与另一个同样考虑时间信息的轨迹预测方法[24]进行比较，结果表现出更好的预测效果，尤其在数据是采集在低频率的情况下。Po-Ruey Lei 等人[15]在实验中也与方法[24]进行了比较，并证明了其提出的方法预测错误率较低。这两种预测方法[15,24]具有相同的输入输出，输入是移动对象当前的运动轨迹和待查询时间，返回的结果是该移动对象在待查询时间所处的位置。

此外，还有些预测方法[18,19,22]虽然也考虑到了时间因素，但是并没有将其体现在结果中，不能预测某个时间的运动情况。如在 Jingbo Zhou[18]等人提出的方法中，作者结合时间维度信息将与当前轨迹具有相似性的轨迹从历史轨迹库中摘选出来，并在此基础建立与当前待预测对象高度相关的本地模型，从而提高预测的准确率。在隐马尔科夫链模型 HMM 中，可见状态是由不可见状态决定的，而不可见状态是随着时间发生变化的，因此，隐马尔科夫链模型 HMM 也是与时间相关联的。文章[19,22]中所提到的方法均是在使用 HMM 的基础上实现的，因此也都考虑了时间因素。

3. 空间+语义信息

原始的轨迹数据是由一连串的带有时间戳的 GPS 位置点序列构成，其本身没有意义，语义就是额外添加的能够赋予数据意义的信息，它能描述出对象的属性类别等。添加了语义之后的轨迹更加便于理解，也在判断轨迹相似性方面提供了依据。Alvares 等人[31,32]提出语义轨迹的概念，将原始的位置信息与语义信息整合到一起。语义轨迹就是一串贴有语义标签的位置点组成的序列，额外的语义信息能帮助描述移动对象经过的每一个具有代表性的地点。Josh Jia-Ching Ying 等人[25]基于模式匹配模型，在考虑空间地理和语义信息的情况下预测移动对象的下个位置点。如图 3-14 所示为语义轨迹示例，轨迹 2 和轨迹 3 如果只考虑空间信息，这两条轨迹的趋势并不相同，因此也不会被判断为相似轨迹。然而如果加上语义信息，两条轨迹均可表示为(School;Bank;Hospital)，这说明在空间语义维度上两条轨迹是相似的。在实验阶段，数据的语义标签是由用户标注的，同时实验者也提供了一些方法去解决语义无法确定的情况。实验结果比较了添加语义的

预测方法与传统的只考虑空间信息的预测方法，并证明了在预测准确度方面前者比后者要好。

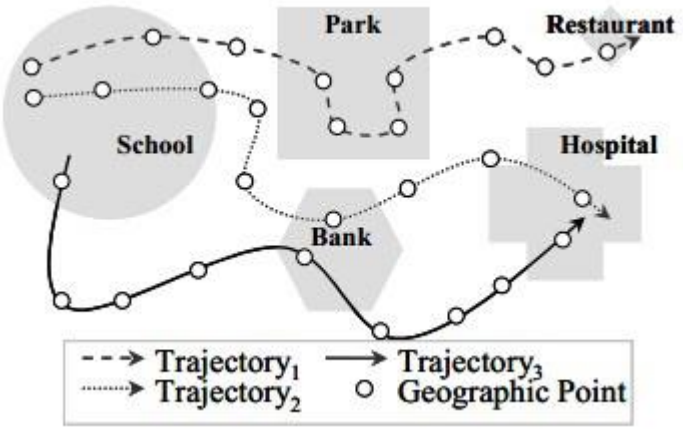


图 3-14 语义轨迹示例[25]

4. 空间+时间+语义信息

综合上述几点，空间信息在轨迹预测中必不可少，时间信息的加入能够使预测结果适合更多的与时间有关的预测应用，并提高预测准确率，语义信息能帮助理解轨迹信息和提高预测准确率，因此把时间空间语义结合起来进行预测将更好地改善预测方法的准确性。JOSH JIA-CHING YING 等人[9]将这三类信息有效结合起来并提出一种由这三种信息同时触发的轨迹模式，即 GTS(Geographical-Temporal-Semantic)模式。在实验中所需的数据集来源于两个能够签到和上传照片的社交网站提供的用户数据，与当前有效的几种方法进行比较之后表现出其准确的预测效果。

5. 空间+时间+社交关系

随着社交网站的快速发展与广泛使用，越来越多的用户愿意在社交网站上分享他们的位置及活动信息，这为轨迹的研究带来了两个好处：1）方便了位置信息的收集，改善过去那种依靠定时采集数据的方式，降低了能耗，提高了效率；2）从社交网络中不仅可以得到离散用户的位置信息，还可以通过其关注或好友列表得到其社交关系。Dashun Wang 等人[39]在研究中表明，人类的移动性、社交网络关系及关系的紧密程度三种之间呈现相互的正相关特性。因此，可以利用其中的已知两点去推断另外一个方面。根据社交关系可以提高人类移动性预测的准确性，反之，根据用户移动的规律性，可以推断出其之间是否存在某种关联性，进而对社交关系进行预测。多种研究将社交关系考虑进了轨迹预测的问题，并通过实验表明考虑了社交关系之后的预测结果要比单独地考虑时空信息取得的结果要准确很多。

Eunjoon Cho 等人[34]首先表明，人类在短距离活动的时候在时空维度上表现出一定的周期性，受社交关系的影响较小，而在远距离活动中受社交关系影响较大。即人们去往远距离的地方极有可能是看望好友，而在短距离内的活动基本上都是受其生活规律所支配。仅有 10%~30%的运动可以用社交关系解释，而50%~70%的活动都表现出一定的周期性和规律性。这表明社交关系只能作为一种辅助因素来作为轨迹预测的影响条件，因此将时空规律性及社交关系相结合才能极大地提高预测的准确性。Eunjoon Cho 等人[34]在此基础上创建了周期性与社交移动性相结合的模型 *Periodic & Social Mobility Model (PSMM)*，并将其与只考虑周期性时空规律的模型 *Periodic Mobility Model (PMM)*进行了比较，模型中参数是使用最大期望算法 EM(Expectation Maximization)来确定。实验中为了得到既含有社交关系又含有位置信息的数据，选取了来自两个社交网站的数据集（Gowalla and Brightkite）及一个两百万移动用户的轨迹数据集合。

除了从社交网站及特定的应用中得到社交关系之外，还可以利用通话记录、短信记录等从通讯网络中获取。Daqiang Zhang 等人[35]选择从蜂窝网络中的数据获取社交关系，利用周期性规律及社交关系的影响相结合，使用一种自适应学习机制对用周期性和用社交关系预测的结果进行整合处理，预测将来一到六个小时内经过的手机发射塔，及最终到达的发射塔。实验数据集来源于 MIT 真实挖掘数据，其包括 32,579 个发射塔位置及 350,000 个小时的用户活动信息。

Manlio De Domenico 等人[36]介绍了 *mutual information* 的概念并将其用来量化两个移动轨迹之间的相关性。在没有提供社交关系的情况下，可通过对 *mutual information* 的计算来推测用户之间的关系，并将朋友及相识的人的轨迹信息作为额外的信息帮助预测。为了将有社交关系和没有社交关系的两种预测情况做比较，实验选择了两个数据集，一个是 Nokia Mobility Data Challenge(NMDC)[40]，包括 GPS 轨迹及通话短信记录等，另一个是 Cabspotting[41]，包含 500 个出租车在旧金山 30 天内的移动轨迹，没有出租车驾驶员之间的关联关系，在这种情况需要计算 *mutual information* 来推断用户之间的关联性。

目前的轨迹预测算法所包含的信息维度主要是上述五类，包含更多的信息维度能带来的好处有：1）满足更多的应用需求，如预测某个时刻移动对象所处的位置；2）提高预测的准确率，包含的信息越多，越能详细地描述移动对象轨迹的特征，从而提供预测准确率。但另一方面，额外的信息也会带来更高的复杂度，并且对存储空间会有更高的要求。当前还并没有研究基于同一数据集将这五种考虑不同信息维度的情况进行比较，尽管 Baumann 等人[33]基于同一数据集比较了 18 种预测下一位置的预测方法的效果，但这十八种预测方法都没有考虑语义维

度的信息及社交关系的影响。随着社交网络的发展以及社交关系对用户移动性的影响，结合社交对用户的轨迹进行建模并预测成为当前轨迹预测的一大趋势。

3.3.4 建模数据集的选择

收集到的轨迹数据集中包含了所有用户的所有轨迹，选择全体数据集还是从中挑选特定有关联的数据集进行建模不仅影响着建模的效率，而且会给预测的结果正确率带来影响。根据建模数据集的选择，当前的轨迹预测方法一般可以分为以下三种：

1. 选择特定的待预测用户的轨迹进行建模

为每个用户选择其自己的历史轨迹建立模型，预测阶段则用属于该用户的预测模型进行预测。这种方法建立的模型具有针对性，但如果某用户自己的历史轨迹数据不够多，则会对其建模造成影响。而且这种建模方法导致了其不能立刻为新加入的用户进行预测，而必须要收集一段时间该用户的历史数据进行建模才可以。此外，为每个用户各自建立预测模型对存储空间也提出了更高要求。在Hoyoung Jeung等人[24]提出的方法中就为每个用户都建立了频繁模式集。

2. 选择所有轨迹进行建模

这种方法指的是选择全部的轨迹集建立一个公共的模型，为所有用户预测时都使用这个模型。因这种方法比较便捷，因此也在众多预测方法建模时被采用。如方法[8,15,16,17,19,22,23]均是利用整体的轨迹数据集来建立一个公共的模型进行预测。

3. 选择具有较高相关性的轨迹进行建模

比较上述两种选择建模数据集的方法，选择每个用户各自的轨迹进行建模尽管建立的模型会与待预测用户高度相关，且每次建模时数据集较少从而可以减少建模的复杂性，但这种方法在缺少某个用户历史数据的情况下会导致无法对该用户进行预测的问题。而选择所有的轨迹建立一个公共的模型时，又会因整体轨迹数据量太大而其中大部分的轨迹又很可能与待预测对象无关而使得效率低下。因此选择与待预测对象轨迹比较相关的轨迹进行建模成为了许多研究者的另一种尝试。

为了处理低采样率的数据稀缺问题，Meng-Fen Chiang等人[14]在时空维度进行扩展，并在扩展后的范围内选择与待预测轨迹相似的轨迹，在此基础上建立一个特定的模型来预测待预测对象在某查询时刻的位置。Jingbo Zhou等人[18]也使用动态选择轨迹数据集的方法来建立预测模型从而进行路径预测。首先通过将待预测对象当前轨迹与历史轨迹进行匹配计算，选出与之相关的小部分轨迹组成参考的轨迹集，再在此基础上建立一个本地模型进行预测。

诸多考虑社交关系因素的预测方法[34,35,36]，根据待预测用户本身的历史数据、好友关系及相关好友的历史数据建立模型。由于用户移动性与社交关系之间存在着很强的相关性，互为好友关系的用户在移动性上也表现出一定的相似性，因此用这种方法选取的建模数据集与待预测用户高度相关。

与选择单个用户的数据集和选择整体数据集相比较，选择特定相关的数据集是前两种方法的折衷处理。然而在相关数据集的选择问题上，第三种方法也会带来一定的时间消耗。因此，在同一实验数据集同一预测方法的基础上，选择不同的建模数据集对整个预测过程带来的影响是个待后续研究的问题。

3.3.5 现有真实轨迹数据集介绍

随着移动设备的普及使用和定位技术的发展，诸多对于对象移动性的研究都是基于真实的移动轨迹数据进行的。相对于通过算法生成的数据来说，真实数据更加可靠，更加能反应出移动对象在现实生活中的移动规律。然而，现有的真实轨迹数据收集数量屈指可数，并且为了满足不同的研究应用需求，轨迹数据收集的方式和包含的信息也有所不同。接下来将介绍一些在轨迹预测中已有的一些真实轨迹数据集。

1. GeoLife 数据集。此数据集是由微软亚洲研究院组织采集得到，记录了一百七十八位用户在历时 4 年的时间内的轨迹数据。该数据集所收集的轨迹是由一系列含有时间信息的位置记录组成，每个位置点包含的信息有经度、纬度和海拔。整个轨迹数据集包含 17,621 条轨迹，总长度达 1,251,654km，总时间有 48,203 个小时。这些轨迹是由各种不同的 GPS 记录设备或 GPS 手机记录的，采样的频率也都各有不同，大部分采样的间隔都比较短，大约是 1~5 秒，或者是每 5~10 米一个点。此数据集记录了大量用户户外的运动轨迹，不仅包括回家、去工作这种生活中的路径，还包括一些娱乐和运动活动，如购物、爬山、骑行等，能满足大部分研究对轨迹的需求，如运动模式挖掘、用户活动推荐、感兴趣地点推荐等等。
2. T-Drive 数据集。这是一个由 33,000 辆出租车在三个多月的时间内轨迹所组成的数据集。总距离达 400 百万千米，GPS 点数量有 790 百万个。平均采样的时间间隔为每 3.1 分钟一个点，没两个连续点之间的平均距离为 600 米。出租车轨迹不同于人们日常出行的轨迹，不能反映出某个特定用户的偏好和生活规律，但可以根据其进行路线推荐或从整体角度来进行预测。

3. ST 数据集[6]。此数据集是新加坡一个多星期时间内 13,200 辆出租车轨迹组成，每辆出租车每 20~80 秒持续地报告其位置，总共含有 268 百万个位置点。
4. HT 数据集[7]。此数据集是从一个火车站 30 分钟的监控视频中获取的行人轨迹，该监控视频分辨率为 480*720，每秒帧数为 24。由于在实际实验中不需要太高的采样频率，因此可以根据需要对轨迹数据进行处理，降低每两个连续位置点间的时间间隔。
5. Trace T0[8]。此数据集是在旧金山区域内出租车的移动轨迹，包含 30 多天内大约 500 辆出租车的轨迹，每两个相邻位置点间的时间差距少于 10 秒。
6. Trace T1[8]。此数据集是在半径为 10km 的区域内由 44 个参与者携带 GPS 设备步行收集的轨迹组成，这些区域包括两个大学校园，一个首都城市（纽约），一个博览会区域和一个主题公园。大部分的点都是步行记录的位置点，当然也有一些是乘交通工具记录的位置点。五个区域的数据加起来总共历时 1000h，采样间隔为 10s。
7. Trace T2[8]。从监控系统 LifeMap 中提取出的人类移动轨迹，包含了用手机在两个月时间内收集的细粒度移动数据，地点分布在首尔和朝鲜，采集间隔是 2 分钟。
8. Trace T3[8]。此数据集来源于用 Place Lab 软件在西雅图不同的三个社区收集的 GPS 数据，总共持续的时间有 2h，大约 55000 个读数记录。
9. EveryTrail Dataset[9]。EveryTrail (http://www.everytrail.com/)是个分享旅游经历的社交网络网站，用户可以上传、分享自己的旅行路线，还可以给旅行经历贴上活动标签。通过 EveryTrail 网站开放的接口可以取得相关的位置数据。
10. Nokia Mobility Data Challenge(NMDC)[40]。包含 GPS 轨迹以及用户通话记录、短信记录、电话号码、蓝牙及 WLAN 历史信息，通过对用户的通讯号码、短息记录的分析可以得出该用户的社交关系，并结合 GPS 在时间和空间维度上的数据记录，可对加入社交因素建立的预测模型进行评估。

为满足不同的应用或实验需要，可选择满足需求的数据集进行实验。采样间隔、移动对象的类型、数据集规模、收集的数据所含的信息种类以及数据获取渠道的可靠性，都是选择数据集时所考虑的因素。

第四章 系统中轨迹预测模块的设计与实现

本章在对轨迹预测的分析和研究的基础上，设计并实现轨迹预测模块在移动协作感知系统中的应用。

4.1 轨迹预测模块需求设计

图 4-1 所示的是移动协作感知系统的整体框架。客户端通过采集模块采集感知数据（声音、光照、轨迹），用户通过 UI 界面拍摄照片，采集到的数据一方面存在本地数据库，同时另一方面通过通信模块传至服务器端。在服务器端的网络接口模块接收由客户端通信模块传来的数据，并暂时存在缓存队列中。图片经由文件管理系统处理，其他感知数据存在分布式数据库中。数据融合模块对数据库中存储的数据进行处理，根据已有的感知数据通过计算补齐出缺少的部分数据并存至数据库中，而那些缺少而又无法补齐的数据则通知激励模块去鼓励用户收集。同时 web 展示模块取出分布式数据库中的数据通过图表的形式展示在 web 端。当缺少数据的时候，数据融合模块会将缺少数据的区域的信息通知给激励模块，激励模块会通知轨迹预测模块。轨迹预测模块返回的预测结果是去往该缺少数据地区的用户列表。

轨迹分析模块分为轨迹建模和轨迹预测两个部分。轨迹建模的输入是数据库中的轨迹数据，通过计算挖掘建立预测模型，并将预测模型的信息存储在数据库中。轨迹建模部分是在线下进行的，占据大部分时间，而轨迹预测模块为了满足实时的要求，是在线上进行。轨迹预测模块的输入是激励模块通知的缺少数据区域的信息和用户当前的运动轨迹信息，输出是可能前往缺少数据区域的用户信息列表。

在整个移动协作感知系统中，轨迹模块主要与数据库模块和激励模块进行交互。轨迹模块从数据库中获取存储的历史轨迹数据，并将建模计算的结果存入数据库。在预测阶段，从数据库中查询待预测用户的当前运动情况，进而对其做出预测。当需要下发激励时，激励模块将缺少数据地区的信息通知轨迹模块，轨迹模块通过计算后将预测的用户列表返回给激励模块。

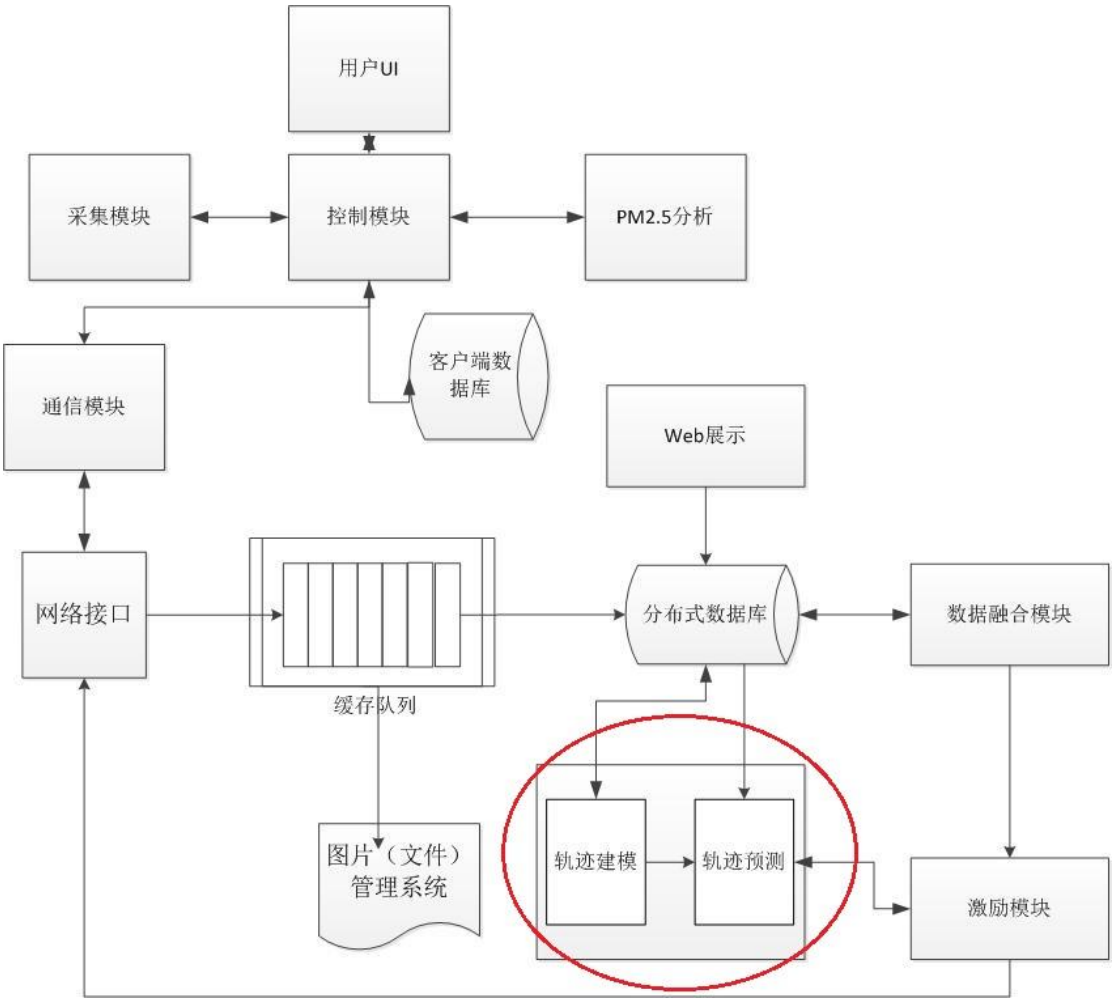


图 4-1 移动协作感知系统整体框架

根据第三章中对轨迹预测的分析研究，本系统中所采取的两种轨迹预测方法，分别属于两种不同类型的预测算法，输入为待预测用户当前的运动轨迹，输出一种是下个位置点，另一种是将来的轨迹。在信息维度只考虑空间地理位置信息，并将整体数据集作为建模数据集。

4.2 轨迹预测模块详细设计

4.2.1 数据库表设计

本文实现两种轨迹预测的方法，一种是属于状态空间模型的二阶马尔科夫链方法，一种是属于模式匹配模型的基于频繁轨迹的预测方法。数据库表设计包括轨迹数据的存储及两种预测方法涉及的表。

轨迹数据表 tbl_traj 如表 4-1 所示，userid 唯一标识一个用户，轨迹 id 作为一条轨迹的标识，相同轨迹 id 的位置点按照时间顺序构成的序列为一条轨迹。time 为该条位置点记录被采集的时间。

表 4-1 轨迹数据表

字段	类型	说明
id	int	自增 id
userid	int	用户 id
tid	int	轨迹 id
latitude	double	纬度，小数点后至少 10 位
longitude	double	经度，小数点后至少 10 位
time	varchar	收集时间

用区域表示的轨迹表 tbl_regiontraj 如表 4-2 所示，是在对原始轨迹数据处理的基础上，将原始位置点转换为用区域表示之后的轨迹。轨迹 id 不变，sequence 是转换后用区域表示的轨迹。

表 4-2 用区域表示的轨迹表

名称	类型	说明
id	int	自增 id
tid	int	轨迹 id
sequence	text	将轨迹点表示成区域的字符串

一阶马尔科夫链表 tbl_firstmarkov 如表 4-3 所示，probability 表示从区域 regionx 一步转移到区域 regiony 的概率。

表 4-3 一阶马尔科夫链转移概率表

名称	类型	说明
id	int	自增 id
regionx	int	区域起点
regiony	int	区域终点
probability	double	转移概率

在一阶马尔科夫链表基础上生成二阶马尔科夫链表 tbl_secondmarkov，如表 4-4 所示，结构类似于一阶马尔科夫链表，不同的是 probability 表示的是由区域 regionx 两步转移到 regiony 的概率。

表 4-4 二阶马尔科夫链转移概率表

名称	类型	说明
id	int	自增 id
regionx	int	区域起点
regiony	int	区域终点
probability	double	转移概率

频繁轨迹表 tbl_freqtraj 是基于频繁轨迹预测情况下存放频繁轨迹的表。如表 4-5 所示，regiontraj 表示由区域组成的序列片段，tidtraj 是包含相应区域序列的轨迹组成的轨迹序列，suf 是频繁估计对应的支持度。

表 4-5 频繁轨迹表

名称	类型	说明
id	int	自增 id
regiontraj	text	区域序列
tidtraj	text	轨迹序列
suf	double	支持度，小数点保留 6 位

运动规则表 tbl_moverule 如表 4-6 所示，head 存储规则头部序列，tail 存储规则尾部序列，confidence 存储规则对应的置信度。

表 4-6 运动规则表

名称	类型	说明
id	int	自增 id
head	text	规则头部
tail	text	规则尾部
confidence	double	置信度，小数点保留 10 位

4.2.2 与其他模块交互的接口设计

轨迹模块主要与数据库和激励模块进行了交互。轨迹建模需要与数据库进行交互并存储预测模型的相关信息，轨迹预测阶段需要激励模块提供待预测区域的信息，并反馈去往该区域的用户列表给激励模块。预测阶段也要与数据库交互，查询用户当前的轨迹信息。主要的交互接口如表 4-7 所示。

表 4-7 与其他模块交互的接口

接口	getAllLocatedRecord()
接口说明	获取所有用户的所有位置数据，没有则返回 NULL
接口	setRegion(int tid,java.util.Date timestamp,java.lang.String region)
接口说明	将原始轨迹点转换成用区域表示的序列，为轨迹号为 tid 的轨迹添加区域 region
参数说明	tid 为轨迹号，timestamp 为轨迹的起始时间，region 为新增的区域
返回值说明	成功返回 true，失败返回 false
接口	getRegionSequence(int tid)
接口说明	查询轨迹号为 tid 的轨迹用区域表示的序列
参数说明	tid 为待查询轨迹号
返回值说明	成功返回区域的序列，失败返回 NULL
接口	setProbability(int startRegion,int endRegion, TransferProbabilityTable.TransferProbabilityType type, double pro)
接口说明	设置转移矩阵的值
参数说明	startRegion 表示起始状态点区域，endRegion 表示终止状态点区域，type 表示识别是一阶还是二阶转移矩阵，pro 为转移概率值
返回值说明	成功返回 true，失败返回 false
接口	getProbability(int startRegion, int endRegion, TransferProbabilityTable.TransferProbabilityType type)
接口说明	查询转移概率矩阵的值
参数说明	startRegion 表示查询的起始状态点，endRegion 表示终止状态点区域，type 表示识别是一阶还是二阶转移矩阵

表 4-7 与其他模块交互的接口(续上表)

返回值说明	返回 double 型的转移概率，查询失败返回-1
接口	setFreTraj(String regionseq, String tidseq, double suf)
接口说明	插入频繁轨迹
参数说明	regionseq 表示频繁区域段序列，tidseq 表示包含频繁区域段的轨迹 tid 列表，suf 表示对应的支持度
返回值说明	成功返回 true，失败返回 false
接口	getSufOfFre(String regionseq)
接口说明	查询对应区域序列的支持度
参数说明	regionseq 表示待查询区域序列段
返回值说明	double 型的支持度，查询失败返回-1
接口	getMovRules()
接口说明	查询所有运动规则信息
返回值说明	返回所有运动规则的头部，尾部，置信度列表
接口	trajPrediction?lat= & lon=
接口说明	提供待预测地点的经纬度，返回预测结果
参数说明	lat 传递纬度，lon 传递经度
返回值说明	返回可能到达该地区的用户列表，没有返回 NULL

4.3 轨迹预测模块的实现

在第三章对轨迹预测方法的比较分析基础上，本小节实现基于状态空间模型的二阶马尔科夫链预测和基于模式匹配的频繁轨迹预测方法。两种方法均使用的是整体数据集建立的一个公用的预测模型，建模阶段在线下进行。

4.3.1 数据预处理

由于包含经纬度信息的原始位置数据太过精确，经纬度能精确到小数点后六位，因此不便于发现其中的规律，在进行轨迹预测之前要对这些数据进行预处理，将轨迹转换为粗粒度的子区域来表示参见 3.3.2 节。

第 i 个位置点可表示为 $P_i=(x_i,y_i,t_i)$，其中 x_i 表示纬度，y_i 表示经度，t_i 表示采集位置点的时间。轨迹由若干个位置点组成的序列表示 $T=(P_1,P_2,...,P_n)$。为了方便划分正方形子区域，将实验区域整体划为矩形。由于原始数据太过精确，而两个位置点之间的差异可能要在小数点后几位才显现出来，因此为了方便数据处理，可先对其放大一定的倍数。

假设以下数据都已经过放大相同的倍数处理，实验区域的坐标起点为 (x_0,y_0)，区域大小为 A*B，A 为纬度坐标差，B 为经度坐标差。若将实验区域划分为 $N_x\times N_y$ 个子区域，则每个子区域的边长 $side=\frac{A}{N_x}=\frac{B}{N_y}$。假设子区域编号由 0 开始，则任一位置点 (x,y) 对应的子区域为 $\left(\frac{y-y_0}{side}-1\right)\times N_x+\left(\frac{x-x_0}{side}-1\right)$。数据处理的类定义如下：

```
public class DataOperation {
    private int partitionNum;//每边划分的份数
    private double length;//区域的边长
    private double length_of_each_region;//每等分的长度
    private double start_x;//区域起点
    private double start_y;//区域起点y坐标
    private int factor;//坐标处理时用到的乘积因子

    //将x坐标放大factor倍数后，取小数点后两位
    public double processX(double x)
    {
        double xx = (x-start_x)*factor;
        DecimalFormat df=new DecimalFormat("#.00");
        xx = Double.parseDouble(df.format(xx));
        return xx;
    }
```

```
//将y坐标放大factor倍数后，取小数点后两位
public double processY(double y)
{
	double yy = (y-start_y)*factor;
	DecimalFormat df=new DecimalFormat("#.00");
	yy = Double.parseDouble(df.format(yy));
	return yy;
}

//根据x y坐标计算区域数
public int xyToRegion(double x, double y)//传入的坐标是已经经过处理的坐标
{
	int i, j;
	for(j=1; j<=partitionNum; j++)
	{
		if(x <= (float)j*length_of_each_region)
			break;
	}
	for(i=1; i<=partitionNum; i++)
	{
		if(y <= (float)i*length_of_each_region)
			break;
	}
	return (i-1)*partitionNum+(j-1);
}
}
```

经由上述方法处理之后，将用经纬度表示的位置点转换成用子区域表示，原始的轨迹序列进而转换为由子区域连接表示的序列。接下来在此基础之上进行基于马尔科夫链预测方法和基于频繁轨迹预测方法的建模处理。

4.3.2 相关定义

下面对基于马尔科夫链和基于频繁轨迹两种预测方法中使用到的概念进行定义，基于马尔科夫链的预测方法相关定义如下：

定义 4.1（马尔科夫链）设有随机过程$\{X_t, t=0,1,2..\}$，若对于任意的时刻$t \in T$和状态序列$i_0, i_1, ..., i_{t+1} \in I$，满足条件概率

$$P\{X_{t+1}=i_{t+1} \mid X_0=i_0, X_1=i_1, ..., X_t=i_t\}=P\{X_{t+1}=i_{t+1} \mid X_t=i_t\},$$

则称$\{X_t, t=0,1,2..\}$为马尔科夫链，即下一时刻所处的状态只与现时刻所处的状态有关，而与之前的状态无关。

定义 4.2（一步转移概率）条件概率$p_{ij}=P\{X_{t+1}=j \mid X_t=i\}$即在 t 时刻由状态 i 经一步转移到 j 的概率，称为一步转移概率。

定义 4.3（一步转移矩阵）设 P 表示由一步转移概率组成的概率转移矩阵，则有$P=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & & & \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}$被称之为一步转移概率。

定义 4.4（n 步转移概率）将条件概率$p_{ij}^{(n)}=P\{X_{t+n}=j \mid X_t=i\}$称之为在 t 时刻由状态 i 经 n 步转移到状态 j 的 n 步转移概率。

定义 4.5（n 步转移矩阵）称矩阵$P^{(n)}=\left(p_{ij}^{(n)}\right)$为 n 步转移矩阵，并有$P^{(n)}=P^n$，其中$p_{ij}^{(n)} \geq 0, \sum_{j \in I} p_{ij}^{(n)}=1$。

基于频繁轨迹预测的相关定义如下：

定义 4.6（轨迹的支持度计数）轨迹段的支持度计数计算公式为：

$$\sigma(t)=\left|\{T_i \mid t \subseteq T_i, T_i \subseteq D\}\right| \quad (4\text{-}3),$$

其中 D 为整体轨迹数据库，$T_i \subseteq D$为一条轨迹。轨迹段的支持度计数表示的是在轨迹数据库中包含某特定轨迹段的轨迹数量。

定义 4.7（轨迹的支持度）在轨迹支持度计数基础上计算轨迹的支持度为：

$$\sup(t)=\frac{\sigma(t)}{|D|} \quad (4\text{-}4),$$

其中|D|为轨迹数据库中所有轨迹的数量。

定义 4.8（频繁轨迹)若轨迹的支持度计数大于给定的最小支持度 min_sup，则称该轨迹为频繁轨迹。频繁轨迹组成的集合称为频繁轨迹集。

定义 4.9 （运动规则的置信度）将频繁轨迹 t 分为前半段轨迹头部 t_{head} 和后半段轨迹尾部 t_{tail}，若满足

$$confidence(t_{head} \rightarrow t_{tail}) = \frac{\sup(t)}{\sup(t_{head})} \geq \min\_conf \quad (4\text{-}5)$$

则称 $t_{head} \rightarrow t_{rail}$ 为一条运动规则，其中 min_conf 为设定的支持度阈值。

定义 4.10 （轨迹的长度）轨迹的长度为轨迹段中包含子区域的个数。单位轨迹为只包含一个子区域的轨迹。

定义 4.11 （轨迹的连接）若轨迹 $t_1=(A_1, A_2,\ldots,A_m)$ 和轨迹 $t_2=(A_2, A_3,\ldots,A_m, A_{m+1})$ 满足 t_1 的后 m-1 项与 t_2 的前 m-1 项相同，则 t_1 与 t_2 可连接为 $(A_1, A_2,\ldots,A_{m+1})$，此过程称之为轨迹的连接。若 t_1 与 t_2 为单位轨迹，则在 t_1 与 t_2 所含的区域为相邻的情况下可连接。

4.3.3 基于马尔科夫链的轨迹预测

本文在基于马尔科夫链的轨迹预测[34]的基础上，实现使用二阶马尔科夫链的轨迹预测。在数据预处理后，将每个子区域看作是马尔科夫链中的离散状态。根据马尔科夫链的性质，每个状态之间的转移是按照一定的转移概率进行的，下个时刻的状态只取决于当前所处的状态和转移概率。

基于 4.3.2 中相关定义，利用马尔科夫链实现轨迹预测。由于在现实生活中，下一刻所处的状态往往不仅与当前的状态有关，也有受之前状态的影响。因此，在本文中选择用二阶马尔科夫链实现轨迹预测，即用户下刻的状态受当前状态及当前状态之前的一个状态所影响。然而另一方面，当前状态和当前状态之前的一个状态对下个时刻所处状态的影响程度又不一样，因此使用权值来表现其不同的影响程度。在此基础上利用马尔科夫链进行轨迹预测的步骤如下：

1. 根据预处理后用子区域表示的轨迹数据，使用统计的方法计算一步转移概率及生成一步转移矩阵。计算一步转移概率的公式为：

$$p_{ij} = \frac{N_{ij}}{\sum_{j \in I} N_{ij}} \quad (4\text{-}1)，$$

其中 N_{ij} 表示在历史数据中由 i 一步转移到状态 j 的次数，$i, j \in I$ 均属于离散状态。在一步转移概率的基础上生成一阶转移矩阵

$P=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & & & \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}$，其中 $n \in I$ 。

2. 由一阶转移矩阵得到二阶转移矩阵 $P^{(2)}=P^2=\begin{pmatrix} p_{11}^{(2)} & p_{12}^{(2)} & \cdots & p_{1n}^{(2)} \\ p_{21}^{(2)} & p_{22}^{(2)} & \cdots & p_{2n}^{(2)} \\ \cdots & & & \\ p_{n1}^{(2)} & p_{n2}^{(2)} & \cdots & p_{nn}^{(2)} \end{pmatrix}$，其中 $n \in I$ 。

3. 根据加权公式预测下个时刻状态，公式为：

$$X_{t+1}=w_t S(t)P+w_{t-1}S(t-1)P^2 \quad (4\text{-}2),$$

其中 t 表示当前时刻，t+1 表示下一时刻，t-1 表示前一时刻，w_t 表示 t 时刻附加的权值，由于当前时刻对于下个位置的影响要大于前一时刻对其的影响，因此权值 w_t 应稍大于 w_{t-1}，X_{t+1} 表示下一时刻可能的状态，是一个一行 n 列的矩阵，S(t)表示 t 时刻所处的状态，也是一个一行 n 列的矩阵，所处状态对应的值为 1，其余的值为 0。预测的结果为选取 X_{t+1} 中概率最大值对应的状态作为下个时刻可能的状态。

在以上基础上，基于马尔科夫链预测方法的建模和预测过程伪代码分别如图 4-2 和图 4-3 所示。

```
//基于马尔科夫链模型建模
Algorithm  MarkovModelling(dataset, matrix1, matrix2, areaSet)
输入：建模数据集dataset，一阶转移矩阵matrix1，二阶转移矩阵matrix2,
状态集areaSet
输出：预测模型matrix1，matrix2

1  For all trajectory∈dataset
2    sequence=preDeal(trajectory) //预处理
3    扫描sequence上每个区域
4    Matrix1[area][nextarea]++;

5 //计算一阶转移概率
6 For all area∈areaSet
7     Total = 0
8      For all nextarea∈areaSet
9          Total += matrix1[area][nextarea]
10    For all nextarea∈areaSet
11     matrix1[area][nextarea] =matrix1[area][nextarea]/total

12  //计算二阶转移概率
13    matrix2 = matrix1*matrix1 //矩阵相乘
```

图 4-2 基于马尔科夫链预测方法的建模阶段伪代码

```
// 基于马尔科夫链的预测
Algorithm MarkovPrediction（trajectory,matrix1,matrix2,areaSet,w1,w2）
输入：待预测轨迹trajectory，一阶转移矩阵matrix1，二阶转移矩阵matrix2，
状态集areaSet,权值w1，权值w2
输出：预测的下个状态

1 sequence[] = preDeal(trajectory)//预处理
2 max = 0
3 for all area∈areaSet
4    Possibility=w1*matrix1[sequence[trajectory.length-
     1]][area]+w2*matrix2[sequence[trajectory.length-2]][area]
5    If possibilty>max
6       max=possibilty
7       state=area
8 return state
```

图 4-3 基于马尔科夫链预测方法的预测阶段伪代码

4.3.4 基于频繁轨迹的轨迹预测

在经典的频繁集挖掘和关联规则生成算法 Apriori 基础上进行改进，可通过挖掘频繁轨迹及运动规则进行轨迹预测[16]。基于此轨迹预测的建模阶段主要包括两个部分：频繁轨迹挖掘和关联规则的生成。具体步骤如下：

频繁轨迹挖掘阶段：

1. 根据预处理后的轨迹数据库，计算出长度为 1 的各轨迹段支持度。
2. 留下支持度大于阈值 min_sup 的频繁轨迹组成频繁轨迹集，剪枝删除非频繁轨迹。
3. 在现有的频繁轨迹集基础上进行轨迹连接，得到在原长度加一基础上的轨迹。
4. 重复 2，3 步至频繁轨迹长度不再增加，得到最大的频繁轨迹集。

频繁轨迹挖掘伪代码如图 4-4 所示。

```
// 基于频繁轨迹的挖掘
Algorithm FrequentMining(dataset,minsup)
输入：预处理后数据集dataset，最小支持度阈值minsup
输出：频繁轨迹集FreSet

For 所有长度为1的轨迹段
  if sup(长度为1的轨迹段t1)>minsup
      f1 = f1 ∪ t1 //f1存储长度为1的频繁轨迹集

For all trajectories ti,tj ∈ f1
  if adjacent(ti, tj) == true && sup(ti ∧ tj)>minsup//若长度为1的轨迹段可连接且连
接之后的轨迹段支持度大于minsup，则加入长度为2的频繁轨迹集合
      f2 = f2 ∪ (ti ∧ tj)

For(k=3;fk-1不为空;k++)
   for all trajectories ti ∈ fk-1
    for all trajectories tj ∈ fk-1
      if concatenate(ti,tj) && sup(ti ∧ tj)>minsup
              fk = fk ∪ (ti ∧ tj)
Return ∪ fk
```

图 4-4 频繁轨迹挖掘阶段伪代码

运动规则生成阶段：

1. 对于最大频繁轨迹集里的每一条频繁轨迹，若长度为 k，则可分裂为 k-1 个序列对 $\langle t_i, t_{k-i} \rangle$，其中 $1 \le t \le k-1$。
2. 对于每一个轨迹序列对，若满足 $confidence(t_i \to t_{k-i}) \ge \min\_conf$，则将其作为运动规则加入运动规则库中。$t_i$ 被称为运动规则的头部，t_{k-i} 被称为运动规则的尾部。

运动规则生成阶段伪代码如图 4-5 所示。

```
//运动规则的挖掘
Algorithm RuleMining（FreSet, minconf）
输入：频繁轨迹集FreSet，最小置信度阈值minconf
输出：运动规则集合MoveSet

For all trajectories t∈FreSet
  for(i=1;i<t.length;i++)
    if conf(t_i,t_length-i)>minconf
      MoveSet = MoveSet ∪ (t_i⇒t_length-i)
Return MoveSet
```

图 4-5 运动规则挖掘阶段伪代码

预测阶段：将待预测用户当前的运动趋势经过预处理阶段表示成区域序列的形式 t，然后将当前的运动序列同运动规则的头部进行匹配，若匹配成功，则按匹配的程度将匹配的运动规则尾部作为预测的结果候选项。在研究[17]中提出三种运动规则匹配的方法：

1. 运动规则 $t_{head} \rightarrow t_{tail}$ 的头部 t_{head} 是当前运动趋势 t 的子轨迹，且 t_{head} 序列的最后一项与当前运动趋势 t 的最后一项相同。
2. 当前运动趋势 t 是运动规则 $t_{head} \rightarrow t_{tail}$ 头部 t_{head} 的子轨迹，且 t_{head} 序列的最后一项与当前运动趋势 t 的最后一项相同。
3. 当前运动趋势 t 与运动规则 $t_{head} \rightarrow t_{tail}$ 头部 t_{head} 完全相同匹配。

使用基于频繁轨迹预测方法在预测阶段的伪代码如图 4-6 所示。

```
//通过运动规则匹配进行预测
Algorithm FrePredict（MoveSet， trajectory）
输入：运动规则集MoveSet,待预测轨迹trajectory
输出：预测的将来轨迹

For all rules(t_i⇒t_j)∈MoveSet
    if ismatched(t_i, trajectory) && conf(t_i⇒t_j)最大
        return t_j
```

图 4-6 基于频繁轨迹预测方法预测阶段伪代码

所有匹配的运动规则尾部都可作为预测结果的候选项，按照每条运动规则对应的置信度大小排列作为预测的结果集，置信度大的运动规则尾部作为预测结果的可能性大。若无匹配项，则预测失败。

第五章 轨迹预测算法的验证与结果分析

本章旨在阐述适合整个系统所用数据集的生成方法，并在第四章介绍的两种轨迹预测方法基础上采用同一数据集进行实验验证，比较两种方法的实验效果。

5.1 系统所需数据集的归一化处理

在前章节 2.3 的介绍中可知，现有的真实轨迹数据集资源比较缺乏，而在仅有的可用的数据集中，收集的数据也是只含有位置和时间信息，不含有其他传感器数据。然而对于整个移动协作感知系统来说，需要既含有传感器读数的数据，又要有采集这些传感器读数相对应的位置信息，以便数据融合模块可以根据位置关系补齐计算出缺少数据位置处的传感器读数，进而使 web 展示模块可以根据位置展示相应的传感器数据。因此为了得到满足整个系统使用要求的数据集，本文通过将只含有位置信息的GeoLife数据集与在固定位置使用传感器去收集感知数据（湿度、温度、光照）的数据集进行归一化处理，生成满足系统要求的数据集。下面首先对这两个数据集依次进行概要介绍。

GeoLife 数据集是由微软亚洲研究院安排采集，记录了一百七十八个用户在历时 4 年的时间内的轨迹数据。该数据集所包含的轨迹是由一系列含有时间信息的位置记录序列组成，每个位置点包含的信息有经度、纬度和海拔。该数据集按用户分文件夹存储了各用户的轨迹，文件夹下每个以轨迹起始时间戳命名的文件里存储的是该用户在该天的轨迹记录。为了得到分布区域大小适中的数据集，本文对原始数据集进行处理，选取数据分布比较集中的区域，提取出轨迹有关的信息，具体格式如图 5-1 所示，从左到右每列的含义分别为：用户 id，轨迹 id，纬度，经度，时间。

001	1	39.991132	116.318231	2008-10-23 04:25:07
001	1	39.991169	116.31848	2008-10-23 04:25:12
001	1	39.991183	116.318691	2008-10-23 04:25:17
001	1	39.991204	116.318904	2008-10-23 04:25:22
001	1	39.991201	116.319098	2008-10-23 04:25:27
001	1	39.991194	116.319315	2008-10-23 04:25:32
001	1	39.991187	116.31949	2008-10-23 04:25:37
001	1	39.991206	116.319683	2008-10-23 04:25:42
001	1	39.991239	116.319866	2008-10-23 04:25:47
001	25	39.99999	116.312615	2008-11-19 13:29:31
001	25	39.999896	116.312589	2008-11-19 13:29:36
001	25	39.99975	116.312576	2008-11-19 13:29:41
001	25	39.999556	116.312593	2008-11-19 13:29:46
001	25	39.999413	116.312626	2008-11-19 13:29:51
001	25	39.99925	116.312677	2008-11-19 13:29:56
001	25	39.999068	116.312754	2008-11-19 13:30:01
001	25	39.9989	116.312792	2008-11-19 13:30:06
001	25	39.998734	116.312833	2008-11-19 13:30:11

图 5-1 处理后的 GeoLife 数据集格式

另一数据集是来自无锡地区某传感网络，由固定位置传感器节点定时收集的传感器读数组成，包括湿度、温度和光照信息，具体格式如图 5-2 和图 5-3 所示。图 5-2 描述的是固定传感器节点的位置信息，第一列表示传感器 id，第三列和第四列分别表示传感器的经度和纬度信息。图 5-3 描述了该传感器网络收集的数据记录，第一列表示时间信息，第二列表示数据包的类型，第三列为源传感器 id，第四列表示收集到该数据记录的传感器 id，第五、六、七列依次表示湿度、温度、光照数据信息。数据集中的所有数据都是在三天内收到的数据记录。

```
1001,NULL,120.372923333333,31.4834383333333,46.1,NULL,2
1003,NULL,120.37285,31.48324,42.6,NULL,2
1004,NULL,120.372768333333,31.4843266666667,8.9,NULL,2
1006,NULL,120.371171666667,31.4829766666667,35.9,NULL,2
1007,NULL,120.37308,31.483715,8.8,NULL,2
1008,NULL,120.37268,31.4837366666667,14.9,NULL,2
1010,NULL,120.372648333333,31.48331,61.8,NULL,2
1011,NULL,120.372356666667,31.4831666666667,50,NULL,2
1013,NULL,120.372236666667,31.4839616666667,1.5,NULL,2
1014,NULL,120.37312,31.4843283333333,12.5,NULL,2
1015,NULL,120.373135,31.4832283333333,39.5,NULL,2
1016,NULL,120.373018333333,31.483345,43.9,NULL,2
1018,NULL,120.373158333333,31.4841833333333,3.5,NULL,2
1019,NULL,120.37227,31.48387,13.6,NULL,2
1020,NULL,120.372921666667,31.483325,49.2,NULL,2
1021,NULL,120.372253333333,31.483475,13.7,NULL,2
1022,NULL,120.372083333333,31.4830833333333,74.8,NULL,2
1023,NULL,120.372266666667,31.48351,16.1,NULL,2
1024,NULL,120.372245,31.484135,5.9,NULL,2
1026,NULL,120.372171666667,31.48327,15.4,NULL,2
1030,NULL,120.372883333333,31.484315,8.7,NULL,2
```

图 5-2 无锡数据集传感器位置信息分布

```
2011:08:03:00:00:05:687: C1  1273 60001  1277  2947  6625     3
2011:08:03:00:00:06:312: C1  1146 60001  1191  3083  6602     2
2011:08:03:00:00:07:500: C1  1078 60001  1095  2927  6602    35
2011:08:03:00:00:08:625: C1  1163 60001  1154  3106  6625     2
2011:08:03:00:00:11:875: C1  1281 60001  1212  3000  6588     3
2011:08:03:00:00:13:562: C1  1220 60001  1293  3123  6603     3
2011:08:03:00:00:16:375: C1  1187 60001  1120  2955  6609     0
2011:08:03:00:00:19:187: C1  1075 60001  1065  3206  6612     2
2011:08:03:00:00:23:125: C1  1063 60001  1105  3061  6592     4
2011:08:03:00:00:23:125: C1  1251 60001  1260  3051  6608     3
2011:08:03:00:00:26:375: C1  1023 60001  1372  3093  6622     4
2011:08:03:00:00:30:187: C1  1218 60001  1298  3115  6588     3
2011:08:03:00:00:30:250: C1  1221 60001  1220  3023  6636     1
```

图 5-3 无锡数据集传感器数据记录

本文首先选取 GeoLife 中数据分布比较集中的子区域与整个的无锡数据集，使用 matlab 仿真后观察两个数据集的位置分布，结果显示前者区域要大于后者，并且后者位置点分布相对来说比较稀疏且数据量少。因此，将无锡数据集位置分布区域扩大至与 GeoLife 子数据集分布区域相同的大小，图 5-4 和图 5-5 所示的分别是处理后区域大小相同的两实验数据集位置分布。如图 5-4 所示，来自无锡数据集的传感器分布区域为纬度范围是 31.48-31.485，经度范围是 12.368-12.373。如图 5-5 所示，处理后 GeoLife 数据集的区域分布为纬度范围 39.995-40，经度范围 116.315-116.32。在此基础上，进一步将无锡数据集的分布区域按照坐标偏移的关系映射到 GeoLife 数据集的坐标系中，保留 GeoLife 数据集的位置信息、用户信息和时间信息，相应地填充来自无锡传感网络的传感器数据信息。由于

GeoLife 数据集中的位置相比传感器网络中固定节点的位置要多的多，因此映射后的数据集中会有大量的位置点没有相应的传感器读数。

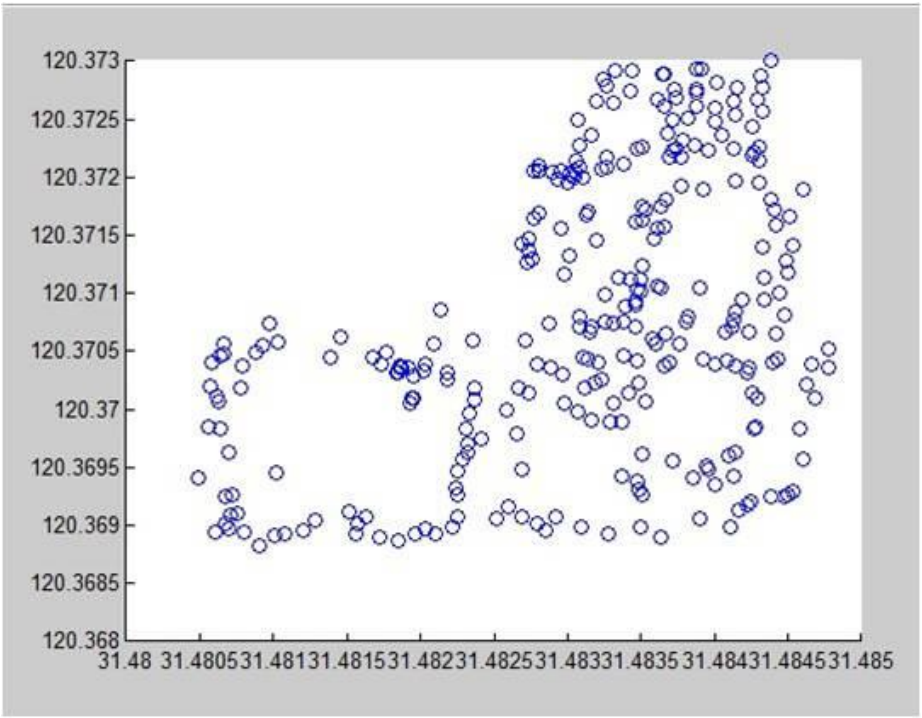


图 5-4 无锡数据集传感器位置分布

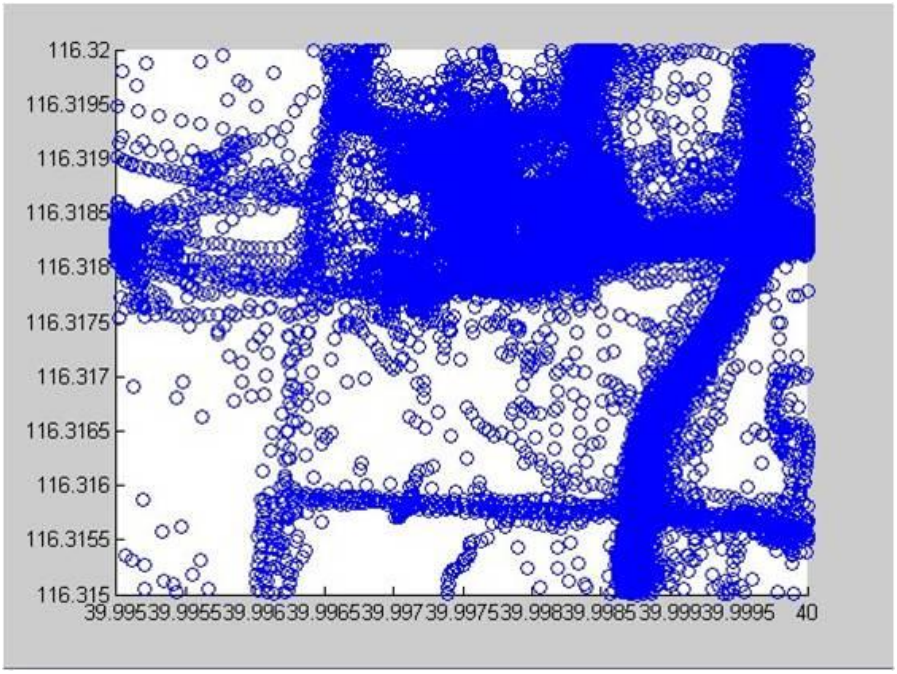


图 5-5 GeoLife 子数据集位置分布

归一化处理数据集过程的伪代码如图 5-6 所示。

```
Algorithm dataNormalize(dataset1, dataset2, ε, gap)
输入：dataset1无锡数据集，dataset2为GeoLife数据集，ε为阈值参数，gap为坐标差
输出：归一化后含有位置及传感器读数的数据集
1 //将dataset1映射到dataset2坐标系
2 For all point∈dataset1
3    point.x += gap
4    point.y += gap
5 //将dataset1中传感器读数填充到dataset2中
6 For all point1∈dataset1
7    For all point2∈dataset2
8     if distance(point1,point2)<= ε
9        point2.readings = point1.readings
10  Return dataset2
```

图 5-6 归一化处理过程伪代码

通过归一化处理之后的数据集包含 28987 条数据记录，包含信息有用户 id、轨迹 id、经度、纬度、温度、湿度、光照和时间，数据覆盖的区域为纬度 39.995-40，经度 116.315-116.32，数据记录如图 5-7 所示。

id	uid	tid	x	y	temperature	light	humidity	time
1	0	37	39.99629100000000	116.31720400000000	2908	27	6551	2008-12-10 02:35:03
2	0	37	39.99635400000000	116.31730600000000	3178	5	6618	2008-12-10 02:35:08
3	0	37	39.99638700000000	116.31740800000000	6665	3	2889	2008-12-10 02:35:13
4	0	37	39.99641300000000	116.31755000000000	6603	3	3123	2008-12-10 02:35:18
5	0	37	39.99648300000000	116.31764800000001	6603	3	3123	2008-12-10 02:35:23
6	0	37	39.99652300000000	116.31771500000001	6603	3	3123	2008-12-10 02:35:28
7	0	37	39.99657700000000	116.31773800000001	2927	11	6551	2008-12-10 02:35:33
8	0	37	39.99664800000000	116.31779600000000	3171	14	6648	2008-12-10 02:35:38
9	0	37	39.99670200000000	116.31787000000000	3033	21	6619	2008-12-10 02:35:43
10	0	37	39.99676300000000	116.31793000000000	3039	7	6605	2008-12-10 02:35:48
11	0	37	39.99683200000000	116.31797299999999	2940	19	6501	2008-12-10 02:35:53
12	0	37	39.99688700000000	116.31795700000001	3199	14	6603	2008-12-10 02:35:58
13	0	37	39.99692800000000	116.31791800000001	2933	28	6555	2008-12-10 02:36:03
14	0	37	39.99705200000000	116.31806400000001	3189	4	6634	2008-12-10 02:36:08
15	0	37	39.99714900000000	116.31824600000000	3182	4	6588	2008-12-10 02:36:13
16	0	37	39.99717700000000	116.31842000000000	3011	8	6683	2008-12-10 02:36:18

图 5-7 映射后实验数据

5.2 实验结果评估参数定义

由于基于马尔科夫链的预测方法输出可以是下个状态（区域），也可以进而预测后续若干个区域，而基于频繁轨迹的预测方法输出是接下来的运动区域序列，因此下面为两种方法分别定义预测准确率来衡量预测效果。

定义 5.1（位置点预测结果准确性）设 q_k 为预测的位置点，q_k^* 为实际真实的

位置点，则用变量

$$\xi_k = \begin{cases} 1, & if \left\| q_k - q_k^* \right\| \le \varepsilon \\ 0, & otherwise \end{cases} \quad (5\text{-}1)$$

表示预测位置点结果的准确性，其中 $\left\| q_k - q_k^* \right\|$ 表示两点之间的欧几里德距离，若两位置点之间的距离小于阈值 ε，则认为预测成功，结果为 1，反之预测失败，结果为 0。

定义 5.2 （基于马尔科夫链预测下一区域的准确率）设 N 为总预测次数，则预测准确率

$$precision = \frac{\sum_{k \in N} \xi_k}{N} \quad (5\text{-}2)$$

定义 5.3 （轨迹上两点距离）设轨迹 $P=(p_1 p_2 \cdots p_n)$ 和轨迹 $Q=(q_1 q_2 \cdots q_n)$，其中 n 为正整数，将 p_i 和 q_i 两个点之间的距离表示为 $d(P,Q,i)$：

$$d(P,Q,i) = \begin{cases} 1, & if \left\| p_i - q_i \right\| \le \varepsilon \\ 0, & otherwise \end{cases} \quad (5\text{-}3)$$

其中 $\left\| p_i - q_i \right\|$ 是 p_i 和 q_i 两点之间的欧几里德距离。

定义 5.4 （轨迹之间的距离）设轨迹 $P=(p_1 p_2 \cdots p_n)$ 和轨迹 $Q=(q_1 q_2 \cdots q_n)$，其中 n 为正整数，则轨迹 P、Q 之间的距离为

$$d(P,Q) = 1 - \frac{\sum_{i \in n} d(P,Q,i)}{n} \quad (5\text{-}4)$$

定义 5.5 （连续轨迹预测结果的准确性）设轨迹 $P=(p_1 p_2 \cdots p_n)$ 和轨迹 $Q=(q_1 q_2 \cdots q_n)$，其中 n 为正整数，则预测的结果为

$$\tau_k = \begin{cases} 1, & if \quad d(P,Q) \le \eta \\ 0, & otherwise \end{cases} \quad (5\text{-}5)$$

k 表示是第 k 次预测，若预测的轨迹与真实轨迹距离小于阈值 η，则预测成功，结果为 1，反之预测失败，结果为 0.

定义 5.6 （连续轨迹预测的准确率）设 N 为总预测次数，则预测准确率为

$$tra\_precision = \frac{\sum_{k \in N} \tau_k}{N} \quad (5\text{-}6)$$

5.3 实验方法及各参数选择

本文将基于马尔科夫链的预测方法与基于频繁轨迹的预测方法在一个统一的数据集上进行实验。经过 5.1 节描述的过程处理后得到的整个系统所需的数据集中数据记录包含位置和传感器读数信息，而在轨迹预测部分并不需要湿度、温度、光照这些传感器读数，因此在实验的时候只考虑位置、时间和用户部分的信息，具体有 302 条轨迹，28987 个位置点。

在诸多的轨迹预测研究中，实验部分都是在改变各参数的情况下观察预测方法的效果。本文在前人研究结论的基础上，改变建模数据集的大小，采取两种类别的方法在同一个数据集上实验并观察预测方法的效果。基于马尔科夫链的预测方法属于状态空间模型，而基于频繁轨迹的预测方法属于模式匹配模型。根据第三章对轨迹预测方法的理解，本文采用的两种方法都是使用整体的数据集建立一个公用的预测模型，预测过程只考虑空间地理信息。为了度量预测方法的效果，实验时在数据集中不区分用户选取部分轨迹进行建模，另一部分轨迹用来验证预测结果的准确性。在进行轨迹预测的时候，将轨迹分为两部分，前半部分作为已知去预测轨迹的后半部分，然后观察预测的结果与实际真实的后半段轨迹之间的差距。具体过程如图 5-8 所示，分别使用两种方法建立的模型进行预测并统计出用两种预测方法预测成功的次数，除以预测进行的总次数，便得到相应的预测准确率。

```
//预测实验
Algorithm predict(dataset,predictSet,n,ε)
输入：建模数据集dataset，待预测轨迹集predictSet，选取前n个状态作为轨迹的已知部分，
距离阈值参数ε， η
输出：使用两种预测方法预测的准确次数

//用两种方法分别建模
MarkovModel = MarkovModelling(dataset)
FrequentModel = FrequentModelling(dataset)
Markovsuccess = 0
Frequentsuccess = 0

For all trajectories t ∈ predictSet
   sequence = preDeal(t)// 预处理
   input[] = getSequence(sequence, 0, n)//选取前n个状态为轨迹的已知状态
   output[] = getSequence(sequence, n, sequence.length)
   pre_state = MarkovePredict(MarkovModel,input)
   pre_traj = FrequentPredict(FrequentModel,input)
   if distancePoint(output[0],pre_state)<ε
       Markovsuccess++
   if distanceTraj(output, pre_traj)< η
       Frequentsuccess++
Resutl.Markov=Markovsuccess; Result.Frequent=Frequentsuccess; return result;
```

图 5-8 实验过程伪代码

在数据处理部分各参数如表 5-1 所示，实验区域为正方形区域，将其划分为 20*20 个小区域，坐标乘积因子是为了方便对位置数据的计算而放大的倍数，整体实验的轨迹条数为 302 条。

表 5-1 数据处理部分各参数取值

实验区域大小	区域划分	坐标乘积因子	总轨迹条数
经度：116.315-116.32 纬度：39.995-40	20*20	10000	302

在基于频繁轨迹挖掘的轨迹预测方法中，支持度阈值与置信度阈值会影响频繁轨迹和关联规则的数量，若支持度阈值过大会导致频繁轨迹的数量减少，从而使运动规则的数量减少，在预测阶段就会因为缺少匹配的运动规则而出现无法预测现象。同样，置信度阈值过大也会导致因缺少匹配的运动规则而无法预测的问题。然而，若两者的值太小又会因挖掘的频繁轨迹过多而给空间存储能力造成压力。本实验在前述研究的基础上，选择合适的实验参数如图 5-9 所示。

支持度阈值	置信度阈值
0.01	0.3

图 5-9 基于频繁轨迹预测方法的各参数

在实验评估阶段，各参数如表 5-2 所示。

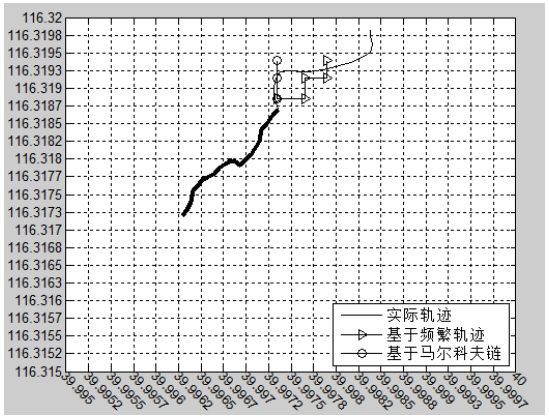
表 5-2 实验结果评估参数取值

ε	η
2.6	0.6

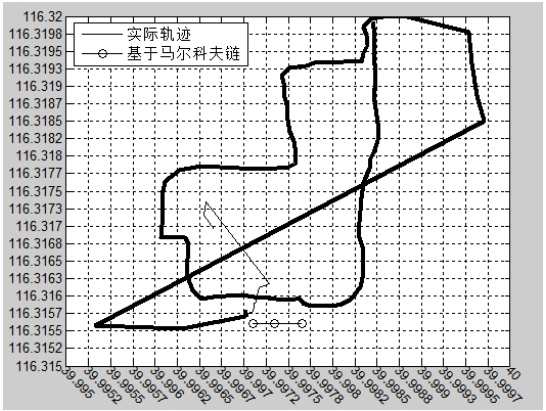
由于轨迹建模阶段是在线下进行，而且建模数据的数量决定了所建预测模型的质量，若建模数据量太少会导致预测模型不准确甚至是在预测阶段无法预测的问题，因此，本实验通过改变建模数据集的大小，用两种方法分别实现轨迹预测，并观察其准确率的变化情况。由表 5-1 可知总体轨迹集合含有 302 条轨迹，每次实验选取的建模轨迹数量分别为 100，150，200，250，290，然后用建立的模型去预测相应的余下轨迹。

5.4 实验结果分析

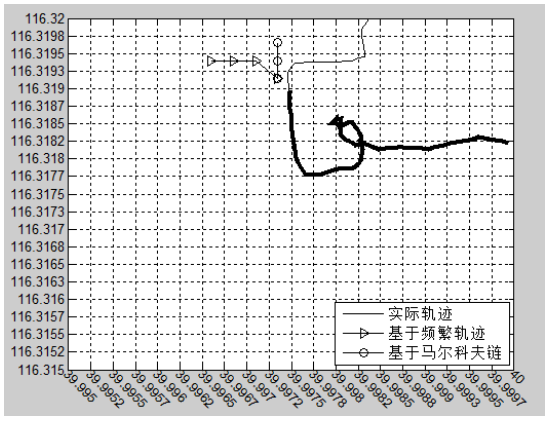
本实验采用前述的来自于 GeoLife 的子数据集作为实验数据集，使用基于马尔科夫链的预测方法和基于频繁轨迹的预测方法分别进行预测，实验效果采取 5.2 节介绍的预测准确率方法进行度量。



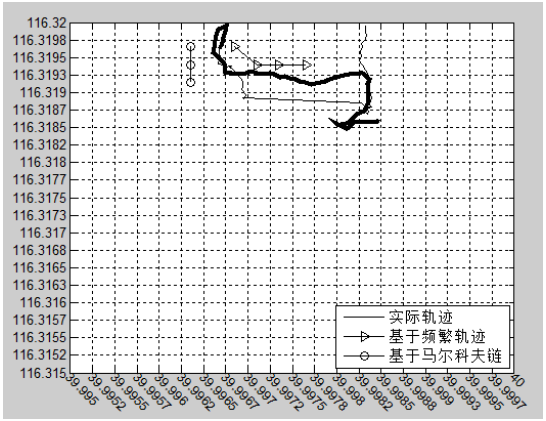
(1) 轨迹 id=37 的预测效果



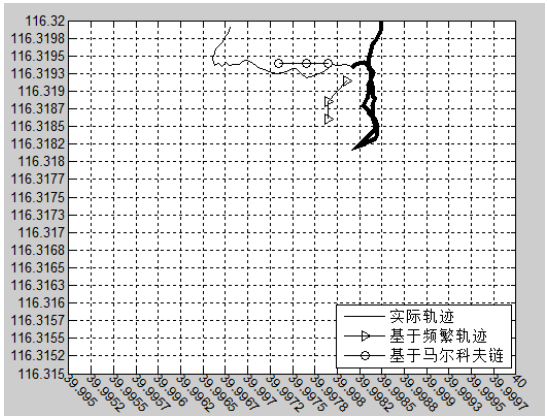
(2) 轨迹 id=40 的预测效果



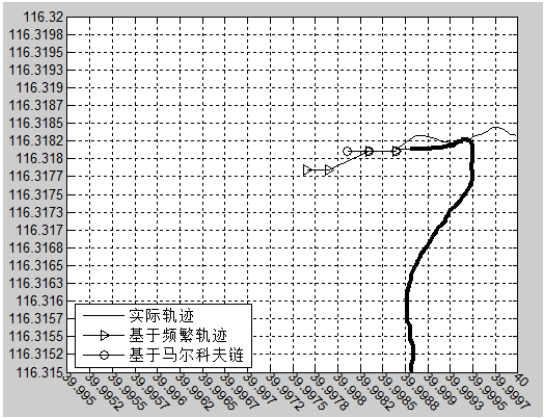
(3) 轨迹 id=6018 的预测效果



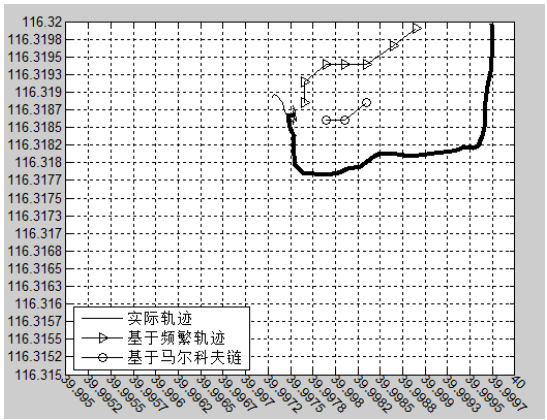
(4) 轨迹 id=770 的预测效果



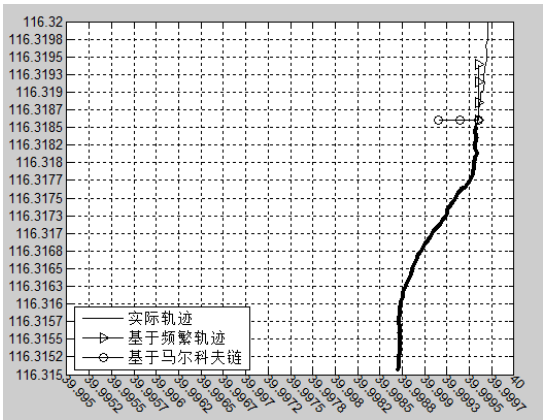
(5) 轨迹 id=1078 的预测效果



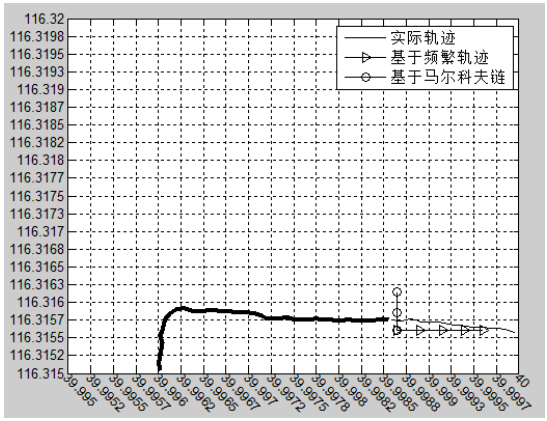
(6) 轨迹 id=18654 的预测效果



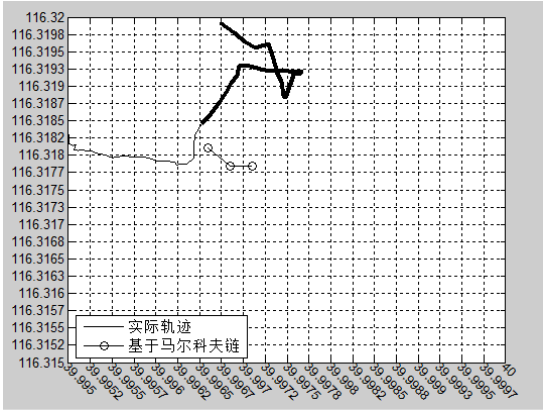
(7) 轨迹 id=1755 的预测效果



(8) 轨迹 id=18591 的预测效果



(9) 轨迹 id=769 的预测效果



(10) 轨迹 id=642 的预测效果

图 5-10 轨迹预测效果

图 5-10 所示的是选取的一组实验数据，使用 290 条轨迹作为整体建模数据集建立的两个模型进行预测的效果展示。将整体实验区域划分成 20*20 大小的子区域，粗实线表示已知轨迹并作为输入，细实线表示真实的后续轨迹，而另两种图标表示用基于频繁轨迹和基于马尔科夫链预测的结果。以两种方法预测的结果均是以子区域为单位表示。按 5.2 节中对实验结果评估参数的定义对该组实验数据结果做出评估，相关阈值参数取值如表 5-2 所示，结果如表 5-3 所示。Result=S 表示预测成功，Result=F 表示预测失败。

表 5-3 示例实验数据预测结果

编号	轨迹 id	基于马尔科夫链预测(下一区域)		基于马尔科夫链预测(连续三个区域)		基于频繁轨迹预测	
		$\|q_k - q_k^*\|$	Result	$d(P,Q)$	Result	$d(P,Q)$	Result
1	37	0	S	0.33	S	0.2	S
2	40	2.5	S	1	F	NULL	F
3	6018	0	S	0.33	S	0.75	F
4	770	2.77	F	1	F	0.75	F
5	1078	2.5	S	0	S	1	F
6	18645	0	S	0	S	1	F
7	1755	5	F	1	F	0.875	F
8	18591	0	S	0.75	F	0	S
9	769	0	S	0.75	F	0	S
10	642	2.77	F	1	F	NULL	F

从图 5-10 及表 5-3 可以看出，选取的这组实验数据中，在采用基于频繁轨迹方法进行预测时会因没有运动规则与当前的运动趋势相匹配而出现无法预测

的情况（轨迹 id 为 40 和 642）。若预测结果与真实结果之间的距离超过一定的阈值则认为预测失败。在用基于马尔科夫链的方法预测连续若干个区域时，将预测的下一区域当作已知，去预测下下个区域，以此类推预测出连续三个区域。使用基于马尔科夫链的预测方法去预测连续区域时，若预测的当前下一区域与真实结果相差较多，则会导致后续的预测结果越来越不准确。

表 5-4 实验结果预测准确率

建模集大小	100	150	200	250	290
基于马尔科夫链(下个区域)	0.25	0.38	0.44	0.5	0.53
基于马尔科夫链(连续区域)	0.22	0.34	0.4	0.41	0.43
基于频繁轨迹	0.2	0.31	0.39	0.43	0.46

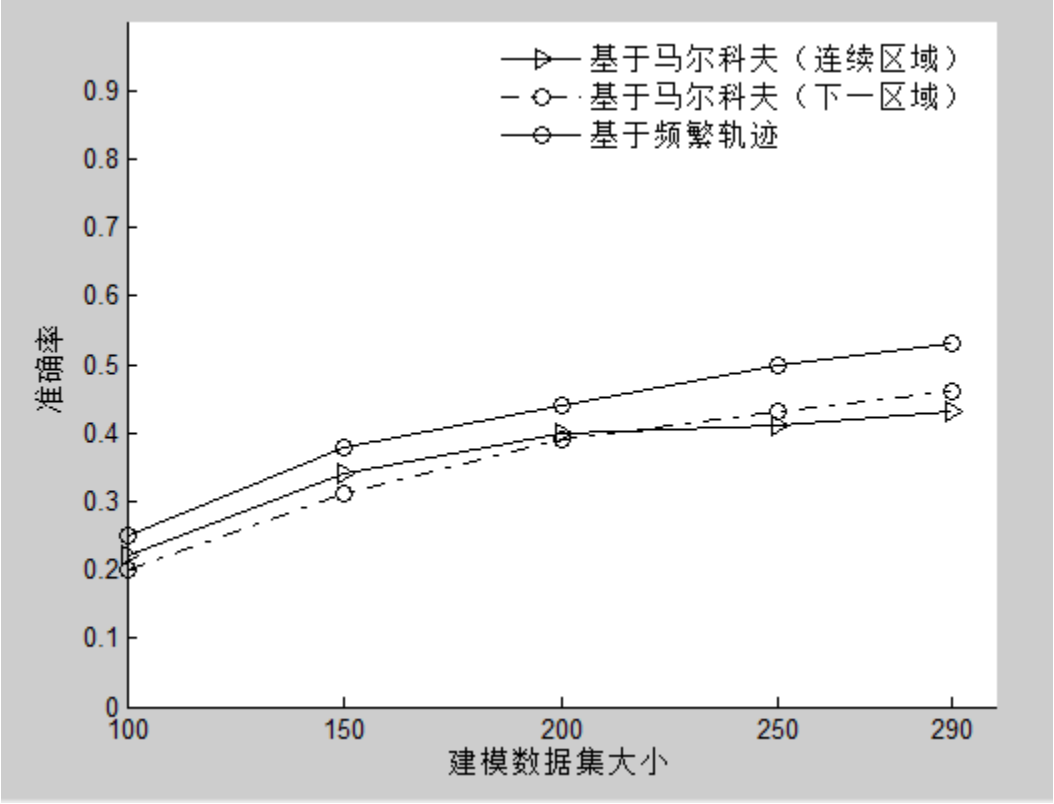


图 5-11 预测准确率

在变化建模数据集大小的过程中，实验结果预测准确率如表 5-4 所示，观察预测准确率变化情况如图 5-11 所示。随着建模数据集的增大，建立的预测模型

更加准确，频繁轨迹和运动规则的数量也相应增多，从而使得预测准确率有所上升。基于马尔科夫链的方法在预测下一区域时有较好的效果，在预测连续区域时准确率会受到影响，产生这一结果的原因在于不确定性结果的累积会导致后续结果更加不准确。在初期建模数据集较小的情况下，基于频繁轨迹的预测会因预测阶段没有可匹配的运动规则而出现预测失败的情况，从而影响预测的准确率。此外，各方法中其他一些参数的选择是否达到最优也会对预测方法的效果造成影响，如区域划分的大小、支持度阈值、置信度阈值的选择等。由于基于马尔科夫链的方法更适合于对下一状态的预测，而基于频繁轨迹的方法在预测连续区域比较适合，因此为了满足不同的应用场景可选择合适的预测方法。

第六章　结束语

6.1 论文工作总结

本文选取了在研究领域和现实应用领域都具有重要意义的轨迹预测课题进行研究，不仅在科学研究的角度分析目前轨迹预测的研究方法，并且在实际应用中分析轨迹预测模块在移动协作感知系统所发挥的作用和体现的价值，并采取真实的数据集对两种不同思想的轨迹预测方法进行实验验证。

首先本文对移动协作感知这个新兴领域进行了介绍，描述了移动协作感知典型的应用场景，并对具体采用的移动协作感知系统架构进行了说明。在第三章中，本文首先对轨迹预测可行性及具体用例进行介绍，然后站在学术研究的角度对现有的轨迹预测方法进行更深层次的理解，总结出轨迹预测一般流程，并根据不同的标准对现有的轨迹预测方法进行分类，在此基础上可根据预测的需求选择合适的预测方法。在第四章本文首先对轨迹预测模块在移动协作感知系统中需求设计进行了说明，进而重点阐述了轨迹预测部分在系统中详细的数据库表设计和其他模块的交互设计，然后对选择实现的两种预测方法进行了详细介绍。第五章介绍了整个系统所需数据的归一化处理过程，并对两种预测方法的实验结果进行了展示和分析。

6.2 下一步工作展望

本文中采取的轨迹预测方法虽能实现轨迹预测的功能，但在很多方面还有所不足，而且本文只是对现有的方法进行分析阐述，未来的轨迹预测还有很大的改善和进步空间，主要有：

1. 现有的轨迹预测方法大多是对个体的运动趋势进行预测，而事实上对群体的预测也有着很广泛的应用前景，比如对敏感活动人群的预测，控制疾病的传播等，并且在科研角度看，对群体预测的方面还有很大的研究发展空间。
2. 在对用户位置采集的同时也会带来暴露用户个人隐私的问题，这就在一定程度上降低了用户参与活动的热情，也为用户的正常生活带来了隐患。因此，如何在保护用户隐私的同时提高预测的准确率将成为将来轨迹预测领域的一个研究问题。
3. 随着社交网络的普遍使用，将来的轨迹预测工作将充分利用社交关系来

提高预测的准确率及效果。在现有预测方法的基础上，综合时间、空间、语义及社交关系，利用多个维度信息对轨迹预测进行改进。

4. 由于各种预测方法所提出的对其预测效果的评估参数计算方法不同，因此在将来的工作中有必要提出一种统一的衡量标准。对于所有的轨迹预测方法，均可以通过这一统一的基准对预测效果进行评估。

参考文献

[1] Burke J A, Estrin D, Hansen M, et al. Participatory sensing[J]. Center for Embedded Network Sensing, 2006.

[2] Ye Y, Zheng Y, Chen Y, et al. Mining individual life pattern based on location history[A]. // MDM'09. Tenth International Conference onMobile Data Management: Systems, Services and Middleware[C], Taipei: IEEE, 2009: 1-10.

[3] Zheng Y, Zhang L, Ma Z, et al. Recommending friends and locations based on individual location history[J]. ACM Transactions on the Web (TWEB), 2011, 5(1): 5.

[4] Tang L A, Zheng Y, Yuan J, et al. On discovery of traveling companions from streaming trajectories[A]. // IEEE 28th International Conference onData Engineering (ICDE) [C],Washington: IEEE, 2012: 186-197.

[5] Yoon H, Zheng Y, Xie X, et al. Social itinerary recommendation from user-generated digital trails[J]. Personal and Ubiquitous Computing, 2012, 16(5): 469-484.

[6] Wu W, Ng W S, Krishnaswamy S, et al. To taxi or not to taxi?-enabling personalised and real-time transportation decisions for mobile users[A]. // IEEE 13th International Conference onMobile Data Management (MDM) [C],Bengaluru: IEEE, 2012: 320-323.

[7] Zhou B, Wang X, Tang X. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents[A]. // IEEE Conference onComputer Vision and Pattern Recognition (CVPR) [C],Providence: IEEE, 2012: 2871-2878.

[8] Anagnostopoulos C, Hadjiefthymiades S. Intelligent Trajectory Classification for Improved Movement Prediction[J]. 2014.

[9] Ying J J C, Lee W C, Tseng V S. Mining geographic-temporal-semantic patterns in trajectories for location prediction[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2013, 5(1): 2.

[10] Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel[J]. Nature, 2006, 439(7075): 462-465.

[11] Gonzalez M C, Hidalgo C A, Barabasi A L. Understanding individual human mobility patterns[J]. Nature, 2008, 453(7196): 779-782.

[12] Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility[J]. Science, 2010, 327(5968): 1018-1021.

[13] Smith G, Wieser R, Goulding J, et al. A refined limit on the predictability of human mobility[A]. // IEEE International Conference onPervasive Computing and Communications (PerCom) [C],Budapest:IEEE, 2014: 88-94.

[14] Chiang M F, Zhu W Y, Peng W C, et al. Distant-time location prediction in low-sampling-rate trajectories[A]. // IEEE 14th International Conference onMobile Data Management (MDM) [C],Milan:IEEE, 2013, 1: 117-126.

[15] Lei P R, Shen T J, Peng W C, et al. Exploring spatial-temporal trajectory model for location prediction[A]. // 12th IEEE International Conference on Mobile Data Management (MDM) [C],Lulea:IEEE, 2011, 1: 58-67.

[16] Morzy M. Prediction of moving object location based on frequent trajectories[M]//Computer and Information Sciences–ISCIS 2006. Springer Berlin Heidelberg, 2006: 583-592.

[17] Morzy M. Mining frequent trajectories of moving objects for location prediction[M]//Machine Learning and Data Mining in Pattern Recognition. Springer Berlin Heidelberg, 2007: 667-680.

[18] Zhou J, Tung A K H, Wu W, et al. A “semi-lazy” approach to probabilistic path prediction in dynamic environments[A]. // Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining [C],New York:ACM, 2013: 748-756.

[19] Mathew W, Raposo R, Martins B. Predicting future locations with hidden Markov models[A]. // Proceedings of the 2012 ACM Conference on Ubiquitous Computing [C],New York:ACM, 2012: 911-918.

[20] Gambs S, Killijian M O, del Prado Cortez M N. Next place prediction using mobility markov chains[A]. // Proceedings of the First Workshop on Measurement, Privacy, and Mobility [C],New York:ACM, 2012: 3.

[21] Asahara A, Maruyama K, Sato A, et al. Pedestrian-movement prediction based on mixed Markov-chain model[A]. // Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems [C],New York:ACM, 2011: 25-33.

[22] Asahara A, Maruyama K, Shibasaki R. A mixed autoregressive hidden-markov-chain model applied to people's movements[A]. // Proceedings of the 20th International Conference on Advances in Geographic Information Systems [C],New York:ACM, 2012: 414-417.

[23] Xue A Y, Zhang R, Zheng Y, et al. Destination prediction by sub-trajectory

synthesis and privacy protection against such prediction[A]. // IEEE 29th International Conference onData Engineering (ICDE) [C],Brisbane:IEEE, 2013: 254-265.

[24] Jeung H, Liu Q, Shen H T, et al. A hybrid prediction model for moving objects[A]. // IEEE 24th International Conference onData Engineering [C],Cancun:IEEE, 2008: 70-79.

[25] Ying J J C, Lee W C, Weng T C, et al. Semantic trajectory mining for location prediction[A]. // Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems [C],New York: ACM, 2011: 34-43.

[26] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Proc. 20th int. conf. very large data bases, VLDB. 1994, 1215: 487-499.

[27] Brinkhoff T. A framework for generating network-based moving objects[J]. GeoInformatica, 2002, 6(2): 153-180.

[28] Lo C H, Peng W C, Chen C W, et al. Carweb: A traffic data collection platform[A]. // MDM'08. 9th International Conference onMobile Data Management [C],Beijing:IEEE, 2008: 221-222.

[29] Anagnostopoulos T, Anagnostopoulos C, Hadjiefthymiades S. An adaptive machine learning algorithm for location prediction[J]. International Journal of Wireless Information Networks, 2011, 18(2): 88-99.

[30] Han S J, Cho S B. Predicting user's movement with a combination of self-organizing map and markov model[M]//Artificial Neural Networks–ICANN 2006. Springer Berlin Heidelberg, 2006: 884-893.

[31] Alvares L O, Bogorny V, Kuijpers B, et al. Towards semantic trajectory knowledge discovery[J]. Data Mining and Knowledge Discovery, 2007.

[32] Bogorny V, Kuijpers B, Alvares L O. ST‐DMQL: a semantic trajectory data mining query language[J]. International Journal of Geographical Information Science, 2009, 23(10): 1245-1276.

[33] Baumann P, Kleiminger W, Santini S. The influence of temporal and spatial features on the performance of next-place prediction algorithms[A]. // Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing [C],New York: ACM, 2013: 449-458.

[34] 彭曲, 丁治明, 郭黎敏. 基于马尔可夫链的轨迹预测[J]. 计算机科学, 2010, 37(8): 189-193.

[35] Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in

location-based social networks[A]. // Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining [C],New York: ACM, 2011: 1082-1090.

[36] Zhang D, Xiong H, Yang L, et al. NextCell: predicting location using social interplay from cell phone traces[J]. 2013.

[37] De Domenico M, Lima A, Musolesi M. Interdependence and predictability of human mobility and social interactions[J]. Pervasive and Mobile Computing, 2013, 9(6): 798-807.

[38] Kantz H, Schreiber T. Nonlinear time series analysis[M]. Cambridge university press, 2004.

[39] Wang D, Pedreschi D, Song C, et al. Human mobility, social ties, and link prediction[A]. // Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining [C],New York: ACM, 2011: 1100-1108.

[40] Laurila J K, Gatica-Perez D, Aad I, et al. The mobile data challenge: Big data for mobile computing research[C]//Pervasive Computing. 2012 (EPFL-CONF-192489).

[41] M. Piorkowski, N. Sarafijanovic-Djukic, M. Grossglauser, Crawdad trace set epfl/mobility/cab (v.2009-02-24) (2009).

致谢

时光荏苒，岁月如梭。转眼间，两年半的研究生生涯即将结束。在北邮读研的这两年半里，我不仅在生活上有所感悟，更加在学术研究上收获颇多。在这里我将对那些曾经帮助鼓励过我的人们表达最诚挚的感谢，本文的撰写当然也离不开这些人的帮助和支持。

首先我要感谢我的导师刘志勇老师和不断给我指导的王文东老师以及实验室的龚向阳老师和阙喜戎老师。感谢刘老师的信任与栽培，能够让我进入北邮的网络与交换技术国家重点实验室进行学习和学术研究。感谢王文东老师在项目研究过程中对我的悉心指导，为我指点迷津，带领大家不断深入研究和探讨项目中所遇到的问题。王老师认真负责的态度和严谨治学的精神及其专业独到的见解，为我的科研道路带来了极其深远的影响。感谢龚向阳老师和阙喜戎老师在学术研究过程中给予我的帮助与支持。

同时，我要感谢项目组中各位师兄弟、师妹及同级同学的帮助与合作。感谢宋峥师兄和田野师兄在学术上对我的指导和鼓励，感谢寇秦荔师妹对于我工作的帮助与支持，感谢项目组的高慧、张波师兄及周雪松、赵露名、刘肖阳、朱致远、李莹、徐登佳、薛潇剑、袁龙运等同学的合作与理解。感谢寝室室友在生活和学习上给我的支持与鼓励。

此外，我还要感谢养育我的父母，感谢他们辛劳伟大的付出，感谢他们对我的抚育和照顾以及对我默默的支持。

最后，再次向在我人生道路上给予过我帮助和鼓励的老师、朋友、家人表示最真诚的感谢和祝福！

攻读学位期间发表的学术论文

[1] 周萌 基于在线学习机制的轨迹预测 科技论文在线 2014.12