**读 书 报 告**

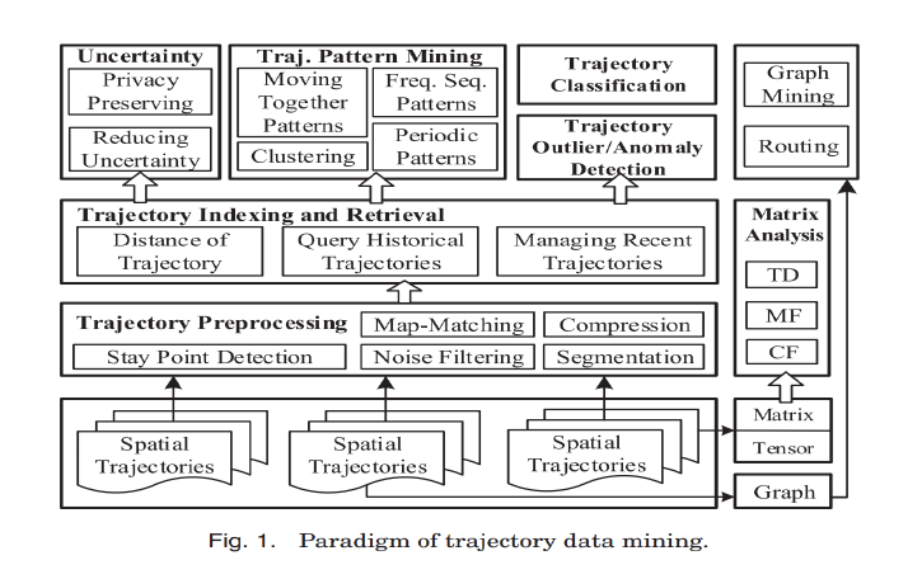
《Trajectory Data Mining:An Overview》

------轨迹数据挖掘综述

杨文杰 2020年09月25日

一、概述

这篇文章对轨迹数据挖掘的主要内容进行了系统性的介绍，提供了该领域的全景及其研究课题的范围。根据轨迹数据的来源、轨迹数据的预处理、轨迹数据的管理以及各种挖掘任务（如轨迹模式挖掘、异常值检测和轨迹分类）的路线图，研究探讨了这些技术之间的相关性以及差异。本文还介绍了将轨迹转换为其他数据格式（如图，矩阵和张量）的方法，可以应用更多的数据挖掘和机器学习技术。最后，提出了一些公共轨迹数据集，有助于读者快速入门轨迹数据挖掘，快速了解这一领域的情况和相关技术。

首先，本文将生成轨迹的数据源分为4类，并列出了每个类别中轨迹数据可以使用的一些应用场景。在使用轨迹数据前，为了保证数据能被正常使用，需要对数据进行预处理，主要包括噪声滤波、停留点检测、轨迹压缩、轨迹分割和地图匹配等。在进行数据挖掘时，我们需要访问轨迹的不同样本或轨迹的不同部分多次，而挖掘大量轨迹是非常耗时的。这需要有效的数据管理技术，可以快速检索所需的轨迹，文章主要介绍了K-最近邻（KNN）查询和范围查询这两种方法。根据前面几个步骤，我们可以进行数据挖掘任务，如轨迹模式挖掘，轨迹不确定性，异常值检测和数据分类等。如Fig.1所示。

二、主要内容

1.轨迹数据

在本文中，我们将生成轨迹的数据源分为4个主要类别，简要介绍每个类别的主要应用场景。

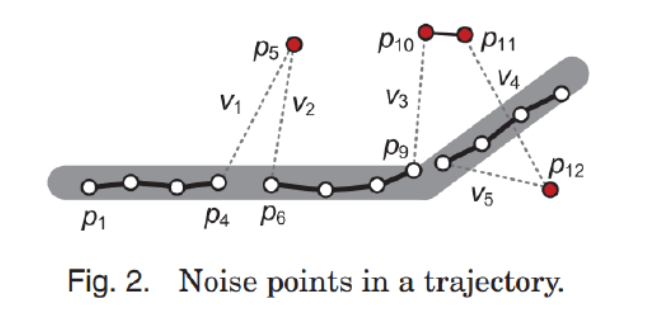
1. 人员流动：长期以来，人们以空间轨迹的形式被动地，积极地记录着现实世界的运动。其数据来源包括各种GPS接收机的数据，社交网络，移动电话和信用卡交易记录等。
2. 运输车辆的流动性：数据主要来源于汽车配备的GPS接收机设备以一定时间频率报告位置。
3. 动物流动：主要研究动物的迁徙痕迹，行为和生活情况。
4. 自然现象的流动：主要是一些自然现象的轨迹，如飓风、龙卷风和洋流等。

2.轨迹数据预处理

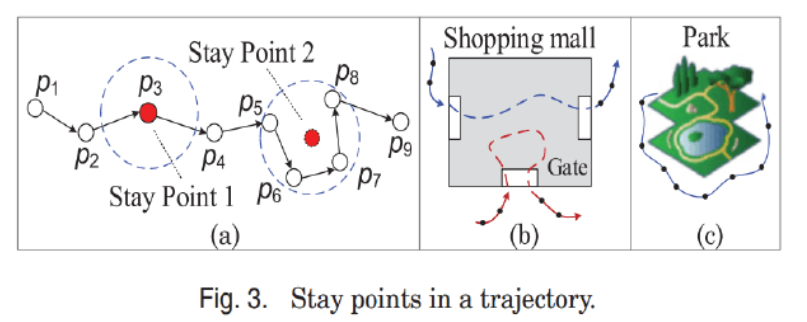
本文介绍了在开始任务挖掘之前处理轨迹所需的四项基本技术，包括噪声滤波，停留点检测，轨迹压缩和轨迹分割。

2.1噪声滤波

由于传感器噪声和其他因素，如在城市峡谷中收到较差的定位信号，空间轨迹永远不会完全准确。因此，在开始数据挖掘任务前，需要去除噪点，目前主流的方法分为三大类。

1. 均值（或中值）滤波器，对于测量点***z***i，（未知）真实值的估计是***z***i及其n-1个前驱在时间上的平均值（或中值）。中值滤波比均值滤波鲁棒性强，当连续处理多个噪声点时，采用均值（中值）滤波误差较大，当轨迹采样率低时也不适合采用均值（中值）滤波。
2. Kalman和粒子滤波器，通过线性模型进行最优估计，去除噪声，其效果取决于初始位置的精度，如果轨迹中的第一点嘈杂，则两个滤镜的有效性会显着下降。
3. 基于启发式的异常检测，主要是通过设置合适的阈值进行滤波。

2.2停留点检测

空间点在轨迹上并不是等重要的。有些地方表示人们停留了一段时间的地方，如购物中心和旅游景点，或加油车辆的加油站，我们称这种点为“停留点”。如图3（a）所示，轨迹中出现两种停留点。一个是单点位置，例如，Stay Point 1，用户保持静止一段时间。这种情况是非常罕见的，因为用户的定位设备通常在相同的位置产生不同的读数。第二种类型，如图3（a）所示的“Stay Point 2”，更为普遍地观察到轨迹，表示人们移动的地方（例如，如图3（b）和3（c）所示）或保持静止但定位读数会转移。

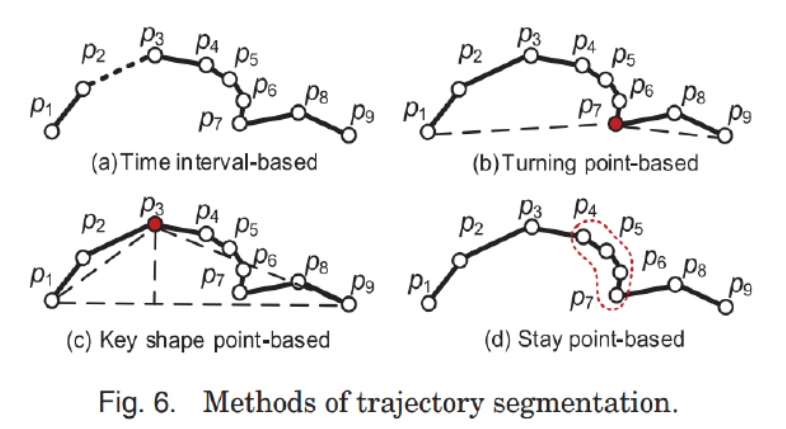
2.3轨迹压缩

基本上，我们可以每秒记录移动物体的时间戳地理坐标。但是，这需要大量的电池电量和通信，计算和数据存储的开销。此外，许多应用程序并不真正需要这样的位置精度。为了解决这个问题，提出了两类轨迹压缩策略（基于轨迹的形状），旨在减少轨迹的大小，同时不会损害其新数据表示的精确度。一种是线下压缩（即批处理模式），它可以在轨迹完全生成后减小轨迹的大小。另一种是在线压缩，当对象行进时，立即压缩轨迹。

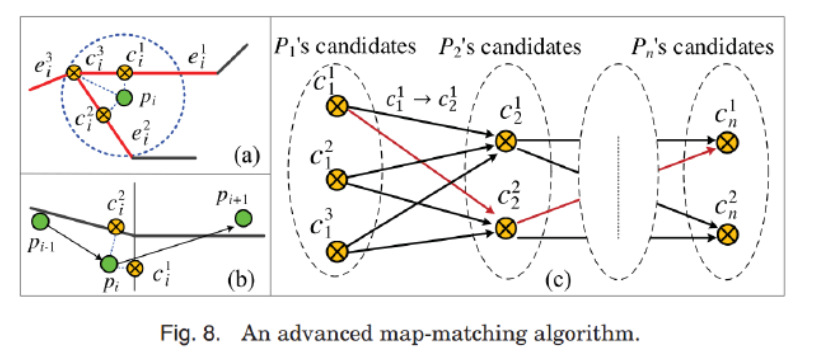
2.4轨迹分割

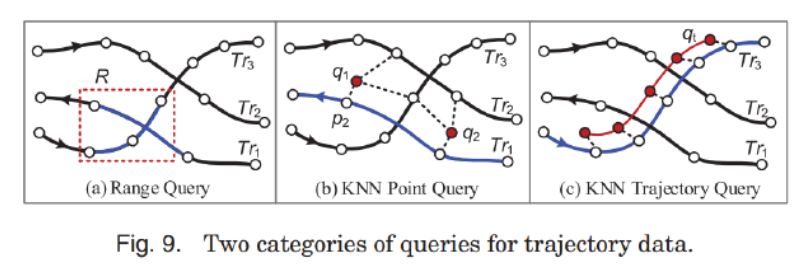
在许多情况下，例如轨迹聚类和分类，我们需要进一步将一个轨迹进行分割。分割不仅减少了计算复杂度，而且使我们能够挖掘更丰富的知识，如子轨迹模式，从而超出了我们从整个轨迹中学到的知识。一般来说，有三种类型的分割方法。即基于时间分割、基于轨迹的形状分割和基于轨迹中点的语义含义进行分分割。

2.5地图匹配

将原始纬度/经度坐标序列转换为路段序列的过程。了解车辆所在的道路对于评估交通流量，引导车辆的导航，预测车辆的行驶路线以及检测起点与目的地之间最常见的行进路径等极为重要。地图匹配算法可以分为四组：几何，拓扑，概率和其他先进技术。几何地图匹配算法考虑道路网络中各个链路的形状，例如将GPS点与最近的道路相匹配。拓扑算法注意道路网络的连通性。代表性算法是使用Fréchet距离来测量GPS序列和候选路线序列之间的拟合的算法。为了处理嘈杂和低采样率的轨迹，概率算法明确规定了GPS噪声，并考虑通过道路网络的多个可能路径找到最佳路线。最近出现了更先进的地图匹配算法，其包括路网的拓扑和轨迹数据中的噪声，例如Lou等。这些算法找到了一系列道路段，它们同时靠近嘈杂的轨迹数据，形成了通过道路网络的合理路线。

根据考虑的采样点的范围，地图匹配算法可以分为两类：局部/增量和全局方法。局部/增量算法遵循从已经匹配的部分顺序扩展解决方案的贪心策略。 这些方法尝试基于距离和方位相似度找到局部最优点。局部/增量方法运行非常有效，通常在在线应用程序中采用。然而，当轨迹的采样率低时，匹配精度降低。相反，全局算法旨在将整个轨迹与道路网络相匹配，例如，考虑到前提和后继者的一点。全局算法比局部方法更准确，但效率更低，通常应用于已经生成完整轨迹的离线任务（例如挖掘频繁轨迹模式）

3.轨迹索引和检索

有两种主要类型的查询：K-最近邻（KNN）查询和范围查询，如图9所示。范围查询检索落入（或相交）空间（或时空）范围的轨迹。例如，如图9（a）所示，范围查询可帮助我们检索过去一个月下午2点至4点之间通过给定矩形区域R的车辆轨迹。然后，可以使用检索的轨迹（或段）来导出诸如分类和预测的数据挖掘任务的特征，例如行进速度和交通流量。 回答这种时空范围查询有三种方法。

第一种是构建3DTree，即将时间视为二维以外的第三维；第二种是划分为多个时间间隔，每隔间隔中生成的轨迹构建一个独立的空间索引；第三种是KNN。

4.轨迹不确定性

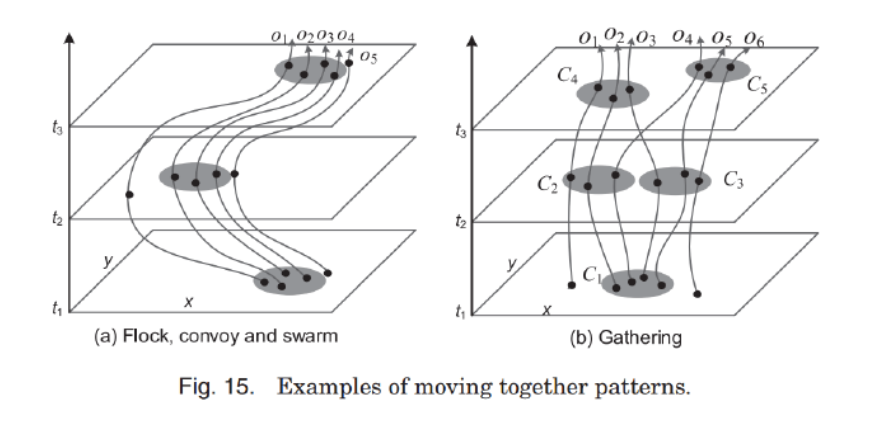
随着移动物体的位置以一定的时间间隔被记录，我们获得的轨迹数据通常是物体的真实运动的样本。一方面，物体在两个连续采样点之间的移动变得未知（或称为不确定）。为此，我们期望减少轨迹的不确定性。另一方面，在某些应用中，为了保护用户的隐私，为了防止轨迹泄漏，我们需要使轨迹更加不确定。

5.轨迹数据挖掘

在本节中，我们研究了可以从单个轨迹或一组轨迹中发现的四种主要类型的模式：伴行模式，轨迹聚类，序列模式和周期模式。

5.1伴行模式

这个分支的研究是发现一组在一段时间内一起移动的对象，如flock、convoy、swarm、traveling companion、gathering。这些模式可以帮助研究物种的迁移，军事监视和交通事件检测等。这些模式可以基于以下因素彼此区分组的形状或密度，组中的对象的数量和模式的持续时间。



5.2轨迹聚类

即找到不同移动物体共享的代表性路径或共同趋势。

5.3轨迹挖掘序列模式

研究的一个分支是从单个轨迹或多个轨迹中找到序列模式。这里，序列模式是指以相似的时间间隔行进公共位置序列的一定数量的移动物体。旅行顺序中的位置不一定是连续的。为了从轨迹中检测序列模式，我们首先需要在序列中定义一个（公共）位置。理想情况下，在轨迹数据中，像来自社交网络服务的用户签入序列一样，每个位置被标记为唯一的身份（例如餐厅的名称）。如果两个位置共享相同的身份，那么它们是相似的。然而，在许多GPS轨迹中，每个点的特征在于一对GPS坐标，其在每个模式实例中都不会重复。这使得来自两个不同轨迹的点不能直接比较。此外，GPS轨迹可以由数千个点组成。没有妥善处理，这些点将导致巨大的计算成本。

5.4周期模式

移动物体通常具有周期性的活动模式。例如，人们每个月都要去购物，动物从一个地方逐年迁移到另一个地方。这种周期性行为为长时间的历史提供了深刻而简明的解释，有助于压缩轨迹数据并预测移动物体的未来移动。

6.轨迹分类

轨迹分类旨在区分不同状态的轨迹（或其部分），如运动，交通模式和人类活动。用语义标签标记原始轨迹（或其段）将轨迹的价值提升到下一层次，这可以促进诸如旅行建议，生活经验共享和上下文感知计算之类的许多应用。

7.轨迹异常检测

1. 检测异常轨迹，一般的想法是利用现有的轨迹聚类或频繁模式挖掘方法。如果轨迹（或段）不能适应任何（基于密度的）集群，或者不是频繁的，则它可能是异常值。
2. 通过轨迹识别异常现象，通过使用许多轨迹来检测交通异常的原因，分析其可能是由事故，控制，抗议，体育，庆典，灾难等事件造成的。

8.轨迹的其他表示

包括图、矩阵和张量等。

三、结论

本文系统地介绍了轨迹数据挖掘的基本框架，简要说明了轨迹数据挖掘的研究邻域以及所需用到的一些研究方法，在文末还详细阐述了轨迹数据的不同表示形式，并为读者提供了一些必要的公共数据资料，对轨迹数据挖掘的前景进行了展望。数据挖掘是在大量轨迹数据的基础上运用机器学习的相关算法对数据进行处理，发现数据背后的价值，为生产生活和管理提供便利，为决策提供依据。同时，它涉及到很多机器学习的算法，牵涉的知识面较广，对于没有基础的初学者存在很多难点，通过阅读这篇文章可以起到启蒙的作用。让我深刻认识到在轨迹数据挖掘学习研究上道阻且长，任重而道远。