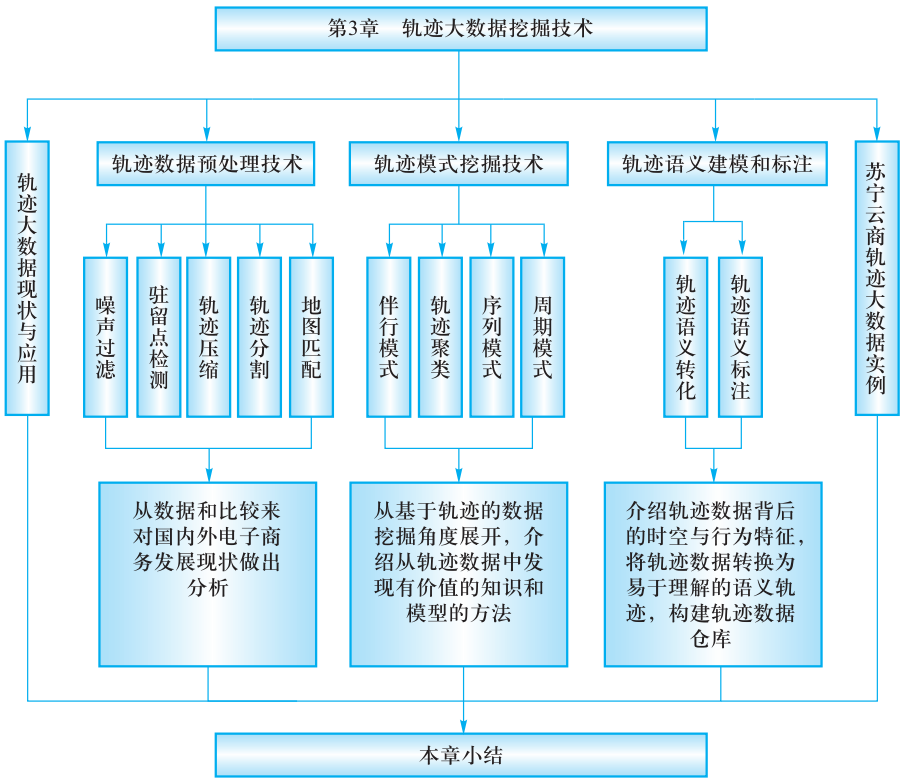
第3章

轨迹大数据挖掘技术



随着卫星导航、无线通信、普适计算技术的不断发展，带有定位功能的移动智能设备被广泛使用，人们在使用这些设备的同时也主动或被动地记录了大量的历史移动轨迹并被持久化保存，这形成了时空轨迹（time-space trajectories）数据。时空轨迹是地理空间加上时间轴所形成的多维空间中的一条曲线，可以表示移动对象在一段较长时间范围内的位置变化。每条轨迹由一系列时空采样点构成，其中每个采样点记录了位置、时间、方向、速度，甚至人与社会交互活动等信息，它刻画了人们在时空环境下的个体移动和行为历史。从宏观角度来看，海量的轨迹数据中不仅蕴含了群体对象的泛在移动模式与规律，例如人群的移动与活动特征、交通拥堵规律等，还揭示了交通演化的内在机理。在大数据时代，企业级的轨迹数据采集、存储已经普遍达到相当规模并得以有效利用。人们通过轨迹分析等手段进行知识发现，并将它们运用在各种交通和位置服务应用系统中，包括交通导航、城市规划、服务推荐、军事调度、交通指挥、物流配送、车辆监控等。

3.1　轨迹大数据现状与应用

卫星定位和移动互联技术在近年来的快速发展催生了海量的轨迹数据。它们记录了移动对象在时空环境下的位置采样序列。轨迹数据的来源多样复杂，可以通过车载 GPS、手机服务、通信基站、公交卡，甚至通过射频识别、图像识别、卫星遥感、社交媒体数据等不同方式获取，不同的回传轨迹遵循不同的数据格式和坐标系统。同时，轨迹数据以极快的速度产生并呈指数级增长，调查显示导航服务公司所接入的移动对象数量可达千万，它以高速数据流的形态进入存储和处理系统。轨迹数据的一些关键属性（例如更新频率、数据总量、每日增量、时空分布等）对数据处理和分析平台搭建有着直接的影响。

如表 3-1 所示，它汇总了不同应用中由 GPS、地图服务、基站、公交卡、道路卡口所采集的轨迹数据及其关键属性。在企业应用中，对象采样频率在秒级、分钟甚至小时级不等，每天所采集的轨迹数据在千万至百亿个采样点的规模区间，最终积累成为 TB 甚至 PB 规模的轨迹数据。其中基站定位的轨迹精度较差，通过 Cell ID 所对应的基站坐标转换获取位置信息，因此精度通常在数百米误差范围。而车载 GPS 和地图 App 所采集的轨迹采样精度较高，误差通常在数米以内。轨迹库已经成为各大地图、导航等服务公司的重要数据资源，单库的原始轨迹规模通常在百亿条以上。目前已经有一些公开的真实轨迹数据集可用于研究工作，如 GeoLife、T-Drive 等。

表 3-1　代表性轨迹数据

数据种类	采集方式	采样频率	日均数据量（采样点）	数据总量
车辆轨迹	车辆 GPS	秒级、分钟级	千万～亿级	TB 级
移动轨迹	地图 App	秒级、分钟级	千万～百亿级	TB、PB 级
手机轨迹	蜂窝基站	分钟级	十亿～百亿级	TB、PB 级

续表

数据种类	采集方式	采样频率	日均数据量（采样点）	数据总量
公交轨迹	公交卡	小时级	百万～千万级	TB、PB 级
卡口数据	卡扣抓拍	分钟级	千万级	TB 级
行为轨迹	社交媒体	分钟、小时级	百万～千万级	PB 级

由表 3-1 可知，轨迹数据继承了大数据的经典“4V”特征，即大规模性（volume）、实时高速性（velocity）、多样性（variety）、高价值性（value）。此外，移动对象轨迹数据库的特有特征可以总结如下。

时空序列性：轨迹是时空环境下的采样序列，这些轨迹点序列蕴含了对象的时空动态性，数据操作是以序列为基本单位的，显著加大了搜索与分析的处理复杂度。

异频采样性：轨迹的采样间隔差异显著，从导航服务的秒级或分钟级采样，到社交媒体行为轨迹的小时甚至以天为间隔的采样，这种差异性极大地影响了轨迹的相似性度量与分析。

数据质量差：由于连续的运动轨迹被离散化表示，特别是当采样间隔达到数分钟以上或设备的采样精度较差时，位置不确定性对轨迹数据分析构成极大挑战。

路网相关性：在交通类应用中，轨迹的运行状态通常限于交通路网，因此，数据分析需要首先完成 GPS 空间向路网空间的映射，并利用路网的时空拓扑信息优化数据处理。

轨迹数据记录了人类的活动和行为历史，蕴含了群体性的移动模式和规律。轨迹数据搜索与分析已经被广泛应用在智能交通、位置服务等系统，具体应用如表 3-2 所示。

表 3-2 代表性轨迹数据应用领域

应用	所用数据	应用现状
大众化经验路径推荐	出租车 GPS 轨迹、私家车移动轨迹数据、气象数据、交通路网数据、历史事故数据等	广泛应用在地图服务公司，显著提升服务水平
交通路况精准预测	GPS 数据（流）、路网路况数据、气象数据、大型活动记录、重大事故数据等	用于地图服务和交通指挥系统，但精度尚需提高
城市规划智能决策	轨迹数据、地图数据、兴趣点数据、消费数据、价格数据、公交线路数据、历史事故数据等	用于数据驱动的规划决策，多源数据集成与融合是难点
个性化服务与活动推荐	车辆与手机轨迹数据、社交网络与社交媒体数据、兴趣点和签到数据、评论数据等	用于基于位置的服务推荐，需提高语义理解和推荐算法的准确性
出租车服务	出租车 GPS 轨迹数据、私家车移动轨迹数据、公交线路与轨迹数据等	应用于相关业务优化，有进一步提升的空间

大众化经验路径推荐：路径搜索和导航服务的核心挑战是难以实时综合各种因素有效地评估并搜索路径。一些地图服务公司借助轨迹分析手段改进路径推荐策略，从大规模轨迹中提取移动模式，并挖掘不同环境下的高质量“经验”路径，根据实时的背景模式匹配（例如，根据气候、车辆类型、交通、匝道开放状态等因素），为用户推荐更为合理、多样化的经验路径，结果显示用这种方式显著提升了用户体验。

交通路况精准预测：通过轨迹流统计的方式评估不同区域的进出流量，检测施工或故障路段，获取实时的交通态势，为用户提供道路预警；通过轨迹数据分析来深入理解交通路况特征和拥堵的演化模式，综合运用历史事件、时空、活动、天气等多维信息，辅助构建数据驱动的城市交通指挥体系，做到指挥决策的先知先觉，警力的优化部署，指挥调度的及时主动；以此引导智能化的交通导航，为导航用户提供准确的行驶时间预测，并根据用户对到达时间的要求推荐路况敏感的合理出行时间。

城市规划智能决策：通过轨迹计算来分析城市不同区域的社会功能、热度特征，确定这些城市区域的性质、规模和发展方向，分析城市内、城市间的交通流模式。这些信息被用于指导城市开发、建设和管理，使有关部门能够合理利用土地资源，协调城市的空间布局，为城市建设、重大施工提供决策辅助；为机构、商家和各类活动的选址需求提供解决方案；优化城市公交、地铁等公共服务线路。

个性化服务与活动推荐：社交媒体中的轨迹数据记录了用户的位置行为，能够更加深入地分析轨迹，包括对轨迹行为的理解、用户特征的刻画、用户行为模式的挖掘等。针对用户对多个目的区域的活动描述，搜索引擎将为用户推荐能够满足查询意图的商家或个性化的服务与活动；考虑轨迹行为和用户体验（基于情感分析），为观光旅客推荐符合用户兴趣的个性化景点、路线。根据用户的驾驶路线推测目的地和出行意图，进行基于位置的精准广告投放。

出租车服务：轨迹数据被用来监控出租车的行驶路线，提供对绕路欺客等现象的检测功能。通过对海量出租车轨迹的分析，系统可以为空驶的出租车优化行驶路线（避免交通拥堵区域、最大化行驶中遇到客户的概率）。它也可以为行人提示附近的有效打车地点以及实时的、最优的公共交通出行路线。一些企业尝试通过轨迹挖掘寻找具有相似出行模式的用户，实现智能拼车等个性化推荐。

在上述应用中，对轨迹数据在完整生命周期内的有效处理成为共性需求。学术界和工业界开展了大量的研究工作，这些技术使原始轨迹数据逐步可用，最后变成所需要的信息与知识。

过去十几年中，人们对轨迹数据处理技术进行了大量的探索，使海量轨迹数据能够被及时处理，从中提取出想要的信息和知识。这些技术按照轨迹金字塔模型分层展开，如图 3-1 所示。

这些技术的目标是使轨迹从底层向高层转化，它们大致被归纳为数据预处理（data preprocessing）、轨迹数据库（trajectory database）、轨迹数据仓库（trajectory data warehouse）、知识提取（knowledge extraction）。4 种技术环环相扣，使轨迹由原始数据转变为规范化数据、信息、知识，形成完整的生命周期。

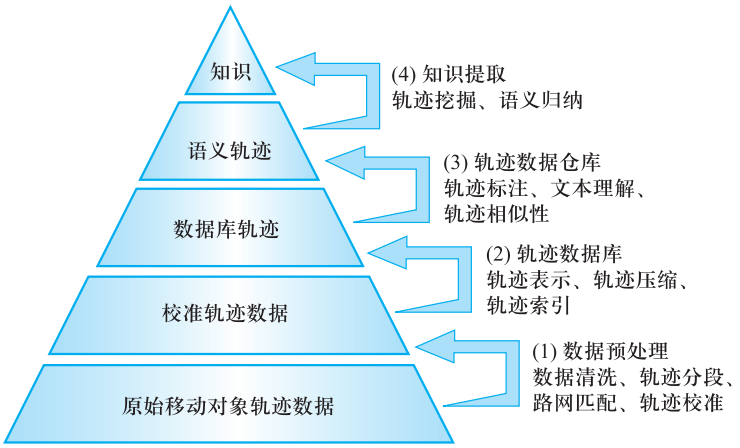


图 3-1 轨迹数据金字塔

3.2 轨迹数据预处理技术

日益积累的大量轨迹数据已经成为大数据范畴中体量最大、变化最快的数据类型。与其他大数据相似，轨迹数据存在着一系列的数据质量问题，主要包括由定位装置和物理环境导致的数据不准确、由设备和传输故障或误操作等因素导致的部分数据缺失、由不同坐标表示更新策略和语境变换导致的数据不一致、由部分轨迹数据导出备份导致的数据冗余等。

这些数据质量问题使原始轨迹数据不能直接用于分析和挖掘，需要先通过预处理技术进行数据转换与校准。一般来说，轨迹数据的预处理主要包括噪声过滤（noise filter）、驻留点检测（stay point detection）、轨迹压缩（trajectory compression）、轨迹分割（trajectory segmentation）和地图匹配（map matching）等。

3.2.1 噪声过滤

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，空间轨迹不会完全准确，有些错误可以接受，如车辆的几个 GPS 点落在实际驾驶车辆的道路之外，可以通过地图匹配算法来修复。在其他情况下，如图 3-2 所示，像 p_5 这样的噪声点的误差太大，距离其真实位置几百米，就难以得出诸如行进速度等有用的信息。

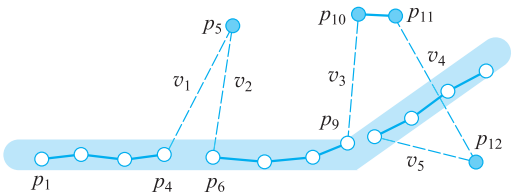


图 3-2 轨迹噪声点

因此，需要从轨迹中滤除这些噪点。虽然这个问题还没有完全解决，但现有的处理方法大体分为三大类。

均值（或中值）滤波器（mean filter 或 median filter）：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）。均值（中值）滤波器可以被认为是覆盖时间上相邻 z_i 值的滑动窗口（sliding window）。在图 3-2 所示的例子中，如果使用滑动窗口大小为 5 的均值滤波器，则 $p_5 \times z = \sum_{i=1}^{5} p_i \times z/5$。处理极端误差时，中值滤波器比均值滤波器鲁棒性强。均值和中值滤波器适用于处理具有密集表示的轨迹中的各个噪声点，如 p_5。然而，当处理多个连续的噪声点时，例如 p_{10}、p_{11} 和 p_{12}，需要较大尺寸的滑动窗口。否则会导致计算的均值（或中值）和点的真实位置之间的误差更大。当轨迹的采样率非常低（即两个连续点之间的距离可能长于几百米）时，均值和中值滤波器不再是很好的选择。

卡尔曼和粒子滤波器（Kalman and particle filter）：从卡尔曼滤波器估计的轨迹是测量和运动模型之间的折中。除了给出符合物理学规律的估计之外，卡尔曼滤波器还给出了诸如速度等高阶运动状态的原理估计。虽然卡尔曼滤波器通过假设线性模型和高斯噪声来提高效率，但是粒子滤波器放宽了这些假设，以获得更一般但效率较低的算法。

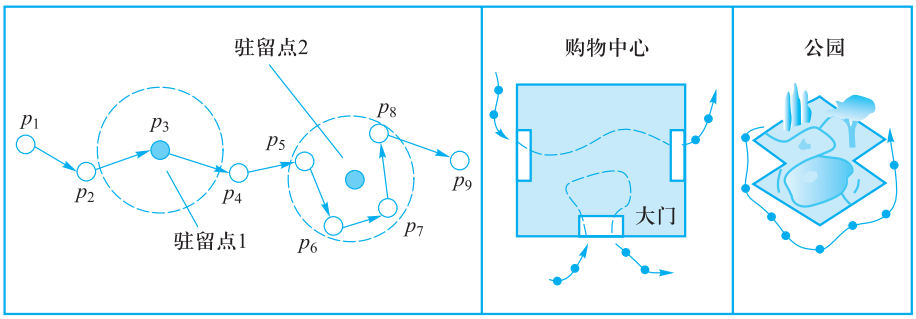
粒子滤波的初始化步骤是从初始分布生成 P 粒子 $x_i^{(j)}$，$j=1$，2，…，p。例如，这些粒子将具有零速度并且在高斯分布的初始位置测量周围聚集。第二步是“重要性抽样”，它使用动态模型 $P(x_i|x_{i-1})$，模拟粒子在一个时间步长上的变化。第三步使用测量模型 $w_i^{(j)}=P(z_i|\hat{x}_i^{(j)})$ 计算所有粒子的“重要性权重”。更重要的权重对应于更好地被测量支持的粒子。然后重要性权重被归一化，当从已归一化重要性权重成正比的粒子中选择一组新的 P 粒子 $x_i^{(j)}$ 时，循环中的最后一步是“选择步骤”。最后，可以通过 $\hat{x}_i=\sum_{i=1}^{p} w_i^{(j)}\hat{x}_i^{(j)}$ 来计算权重和。卡尔曼和粒子滤波器模拟测量噪声和轨迹的动力学。然而，它们取决于初始位置的测量。如果轨迹中的第一点为噪声点，则两个滤镜的有效性会显著下降。

基于启发式的异常点检测（heuristic-based outlier detection）：上文中提到的滤波器在轨迹中用估计值替代噪声测量，而第三类方法通过使用异常值检测算法从轨迹中直接去除噪声点。首先根据点与其后继者之间的时间间隔和距离（称为段），计算轨迹中每个点的行进速度。切断速度大于阈值（例如 300 km/h）的片段，例如 $p_4 \rightarrow p_5$，$p_5 \rightarrow p_6$ 和 $p_9 \rightarrow p_{10}$（如图 3-2 中虚线所示）。假设噪声点的数量比普通点少得多，像 p_5 和 p_{10} 这样的分离点可以被认为是异常值。一些基于距离的异常值检测可以很容易地找出在距离 d 内的 p_5 的邻居的数量小于整个轨迹中的点的比例。同样，可以过滤 p_{10}、p_{11} 和 p_{12}。虽然这样的算法可以处理轨迹中的初始误差和数据稀疏问题，但是设置阈值仍然是基于启发式的。

3.2.2　驻留点检测

空间点在轨迹上并不是同等重要，有些点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站，称这种点为“驻留点”。如图 3-3（a）所示，轨迹中出现两种驻留点。一个是单点位置，例如，驻留点 1 表示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 3-3（a）所示的驻留点 2 更为普遍，能观察到轨迹，这表示人们移动的地方或保持静止但定位读数会转移，如图 3-3（b）和图 3-3（c）所示。



(a) 驻留点　(b) 轨迹示例1　(c) 轨迹示例2

图 3-3　轨迹驻留点

有了这样的驻留点，可以将一系列时间戳—空间点的轨迹转化为有意义的地方 S，$P=p_1\rightarrow p_2\rightarrow\cdots\rightarrow p_n$，推导出 $S=s_1\xrightarrow{\Delta t_1}s_2\xrightarrow{\Delta t_2}\cdots\xrightarrow{\Delta t_{n-1}}s_n$，以此促进各种应用，如旅游建议、目的地预测、出租车推荐和天然气消费量估计。另一方面，在一些应用中，例如估计路径的行进时间和行车路线建议，这样的驻留点应该在预处理期间从轨迹中移除。

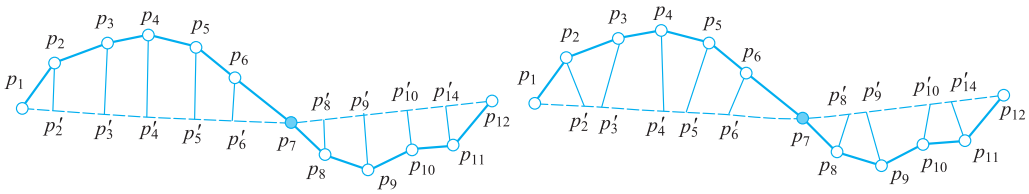
如图 3-3 所示，驻留点算法首先检查定位点（例如 p_5）与其后继者之间的距离是否大于给定阈值（例如 100 m）的轨迹。然后测量定位点和距离阈值内的最后一个后继（即 p_8）之间的时间间隔。如果时间间隔大于给定的阈值，则检测到驻留点（p_5、p_6、p_7 和 p_8），该算法开始从 p_9 检测下一个驻留点。Yuan 等人基于密度聚类的思想改进了这种驻留点检测算法，在找到 p_5 到 p_8 是候选驻留点（使用 p_5 作为定位点）之后，他们的算法进一步检查 p_6 的后继点。例如，如果从 p_9 到 p_6 的距离小于阈值，则 p_9 将被添加到驻留点。

3.2.3　轨迹压缩

轨迹数据可以记录每秒移动物体的时间戳和地理坐标，但是这需要大量的电池电量、通信、计算和数据存储成本。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，旨在减少轨迹大小的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。

除了以上两种策略之外，还有两个距离度量来测量压缩误差：垂直欧氏距离和时间同步欧氏距离。假设将具有 12 个点的轨迹压缩成 3 个点（即 p_1、p_7 和 p_{12}）的表示，则两个距离度量是连接 p_i 的段长总和，如图 3-4（a）和图 3-4（b）所示。

后一距离假定在 p_1 和 p_7 之间以恒定速度行进，通过时间间隔计算每个原始点的投影。

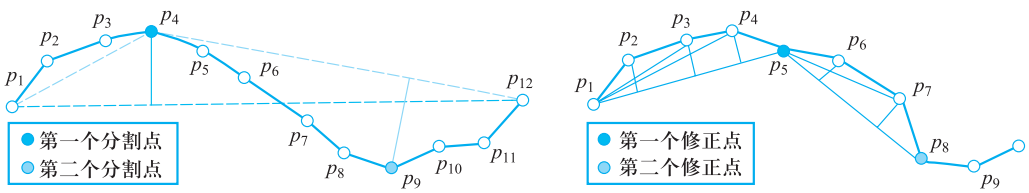


(a) 垂直欧氏距离　　(b) 时间同步欧氏距离

图 3-4　测量压缩误差的距离度量

著名的 Douglas-Peucker 算法被用于近似原始轨迹。如图 3-5（a）所示，它是用近似的线段代替原始轨迹，例如，如果替换不符合指定要求的错误（在本例中使用垂直欧氏距离），则通过选择贡献最大误差的点作为分割点（例如 p_4），将原始问题递归地分解为两个子问题。该过程一直持续到近似值和原始轨迹之间的误差低于指定误差。

随着许多应用程序需要及时传输轨迹数据，一系列在线轨迹压缩技术已经被提出，来确定新获取的空间点是否应当保留在轨迹中。在线压缩方法有两大类。一类是基于窗口的算法，例如滑动窗口算法和开放窗口算法。另一类是基于移动物体的速度和方向的算法。滑动窗口算法是使具有有效线段的滑动窗口中的空间点适应，并继续增长滑动窗口，直到近似误差超过某个误差界限，如图 3-5（b）所示。开放窗口算法应用 Douglas-Peucker 算法的启发式来选择窗口中最大误差的点（如图 3-5（b）中的 p_3）到近似轨迹段，然后将此点用作新的定位点来近似其后继点。

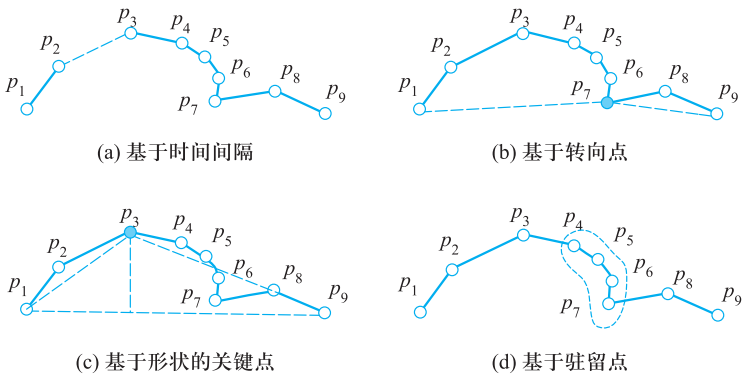


(a) Douglas-Peucker算法　　(b) 滑动窗口算法

图 3-5　Douglas-Peucker 算法和滑动窗口算法

3.2.4　轨迹分割

在许多情况下，例如轨迹聚类和分类，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，而且能够挖掘更丰富的知识，如子轨迹模式，从而超出了从整个轨迹中学到的知识。一般来说，如图 3-6 所示，有 3 种类型的分割方法。



(a) 基于时间间隔　(b) 基于转向点

(c) 基于形状的关键点　(d) 基于驻留点

图 3-6　轨迹分割方法

第一类是基于时间间隔，如图 3-6（a）所示，如果两个连续采样点之间的时间间隔大于给定的阈值，则将轨迹分为两部分，即 $p_1 \rightarrow p_2$ 和 $p_3 \rightarrow \cdots \rightarrow p_9$。有时可以将轨迹划分成相同时间长度的段。

第二类方法是基于轨迹的形状，如图 3-6（b）和图 3-6（c）所示。其中，图 3-6（b）是通过转向点来划分轨迹，即方向在阈值上改变幅度。图 3-6（c）是使用线简化算法，如 Douglas-Peucker 算法，来识别保持轨迹形状的关键点，然后通过这些关键点将轨迹分割成段。类似地，可以基于最小描述语言（MDL）的概念来划分轨迹，该概念由两个部分组成：$L(H)$ 和 $L(D|H)$。$L(H)$ 是假设 H 的描述的长度（以位为单位）；而 $L(D|H)$ 是借助于假设对数据的描述的长度（以位为单位）。解释 D 的最佳假设 H 是最小化 $L(H)$ 和 $L(D|H)$ 之和。更具体地说，它们使用 $L(H)$ 表示分割段的总长度（如和），而让 $L(D|H)$ 表示原始轨迹与新轨迹之间的总距离（垂直和角度）分区段。使用近似算法，它们找到从轨迹最小化 $L(H)+L(D|H)$ 的特征点的列表。通过这些特征点将轨迹划分成段。

第三类方法是基于轨迹中点的语义含义，如图 3-6（d）所示，基于其包含的驻留点，可以将轨迹分成段，即 $p_1 \rightarrow p_2 \rightarrow p_3$ 和 $p_8 \rightarrow p_9$。是否应该在分割结果中保留驻留点取决于应用程序。例如，在旅行速度估计的任务中，应该删除出租车停放等待乘客的驻留点（出租车的轨迹）。相反，为了估计两个用户之间的相似性，只能关注驻留点序列，同时跳过两个连续驻留点之间的其他原始轨迹点。

3.2.5　地图匹配

地图匹配是将原始纬度 / 经度坐标序列转换为路段序列的过程。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来说，了解车辆所在的道路十分重要。基于所使用的附加信息或轨迹中采样点的范围，有两种标准来对地图匹配算法进行分类。

根据所使用的附加信息，地图匹配算法可以分为 4 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；拓扑地图匹配算法注意道路网络的连通性，代表性算法是使用弗雷歇距离来测量 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，概率地图匹配算法明确规定了 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；还有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，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 / 增量和全局算法。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，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）。

3.3　轨迹模式挖掘技术

轨迹数据挖掘旨在从轨迹中发现有价值的知识和模型，这已经成为数据挖掘领域的一个重要分支，且被广泛使用在各类应用中。现有的轨迹知识提取工作主要从基于轨迹的数据挖掘角度展开，包括本节将研究的可以从单个轨迹或一组轨迹中发现的 4 种主要类型的模式，分别为伴行模式、轨迹聚类、序列模式和周期模式。

3.3.1　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、traveling companion 和 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，也可以基于以下几种因素彼此区分，如组的形状或密度，组中的对象的数量和模式的持续时间。

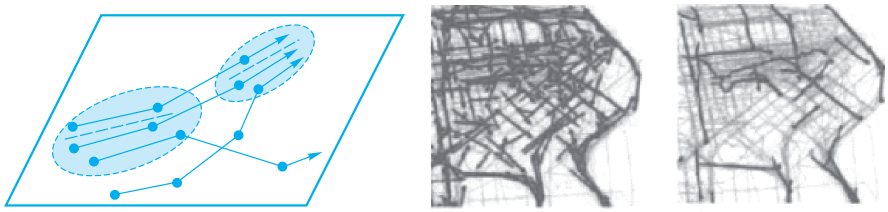
具体来说，flock 是一组在一些用户指定大小的盘中一起行进至少 k 个连续时间戳的对象。flock 的一个主要问题是预定义的圆盘，这不能很好地描述一个群体在现实中的形状，因此可能会导致所谓的 lossy-flock 问题。为了避免对移动组的尺寸和形状的刚性限制，提出了通过采用基于密度的聚类来捕获任何形状的通用轨迹挖掘的 convoy。代替使用磁盘，convoy 需要在 k 个连续时间点内密集连接一组对象，然而在连续的时间段内对 flock 和 convoy 都有严格的要求。swarm 是一种更通用的轨迹模式，它是持续至少 k 个（可能不是连续的）时间戳的对象簇。然而 convoy 和 swarm 需要将整个轨迹加载到内存中进行模式挖掘时，伴行模式使用数据结构从正在流式传输到系统的轨迹中不断地发现 convoy/swarm 样式。所以，伴行模式可以被认为是 convoy 和 swarm 的在线（和增量）检测方式。

为了发现一些（如庆祝活动和游行）经常有对象加入并离开的事件，gathering 模式进一步减少了上述模式的限制，允许一个群体的成员逐渐演变。gathering 的每个聚类应包含若干个参与者，它们是出现在该 gathering 的若干个聚类中的对象。由于 gathering 模式用于检测事件，因此还要求检测到的图案的几何属性（如位置和形状）相对稳定。

3.3.2 轨迹聚类

为了找到不同移动物体共享的代表性路径或共同趋势，通常需要将类似的轨迹组合成聚类（cluster）。一般的聚类方法是用特征向量表示轨迹，表示两个轨迹之间的相似度与它们的特征向量之间的距离。然而，由于不同的轨迹包含不同和复杂的属性，例如，长度、形状、采样率、点数和它们的顺序，所以不同的轨迹生成具有均匀长度的特征向量并不容易。此外，难以将轨迹中的点的顺序和空间属性编码为其特征向量。因此下面将重点介绍为轨迹提出的聚类方法。请注意，本节中讨论的聚类方法专用于自由空间中的轨迹（即没有道路网络约束）。虽然有一些书中讨论了道路网络设置中的轨迹聚类，但是这个问题实际上可以通过地图匹配和图聚类算法的组合来解决。也就是说，可以首先使用地图匹配算法将轨迹投影到道路网络上，然后使用图聚类算法在路网上找到子图（即道路集合）。

Gaffney、Smyth 和 Cadez 等提出通过使用回归混合模型和期望最大化算法将类似轨迹组合成聚类。该算法针对两个整个轨迹之间的总距离聚类轨迹。然而，在现实世界中移动的物体很少一起旅行整个路径。为此，Lee 等人提出将轨迹划分为线段，并使用 Trajectory-Hausdorff 距离构建近距离轨迹段，如图 3-7（a）所示，之后会为每个分组集合找到一个代表性的路径。由于轨迹数据经常被逐渐接收，所以 Li 等人进一步提出了增量聚类算法，旨在降低接收轨迹的计算成本和存储。Lee 和 Li 都采用了微簇 - 宏簇框架。该框架是 Aggarwal 等人处理聚类数据流时提出的。也就是说，他们的方法首先找到轨迹段的微簇（如图 3-7（b）所示），然后将微簇组成宏簇（如图 3-7（c）所示）。Li 得出的一个重要见解是新数据只会影响接收新数据的地方，而不是遥远的地区。



(a) 轨迹段聚类　(b) 微簇　(c) 宏簇

图 3-7　轨迹聚类

3.3.3 序列模式

这里研究的一个分支是从单个或多个轨迹中找到序列模式。序列模式是指以相似的时间间隔行进的公共位置序列，其中包含了一定数量的移动物体。序列中的位置不一定是连续的。例如两个轨迹 A 和 B：

$$\text{A：}\ l_1 \xrightarrow{1.5\ \text{h}} l_2 \xrightarrow{1\ \text{h}} l_3 \xrightarrow{1.2\ \text{h}} l_4,\quad \text{B：}\ l_1 \xrightarrow{1.2\ \text{h}} l_2 \xrightarrow{2\ \text{h}} l_4$$

它们共享一个序列，这是因为访问次数和行进时间是相似的（虽然 l_2 和 l_4 在轨迹 A 中不是连续的）。当语料库中出现这样的公共序列超过阈值时，就会检测到

序列模式。寻找这种模式可以用于旅游推荐和生活模式理解，下一个位置预测可以用于估计用户相似度和轨迹压缩。

为了从轨迹中检测序列模式，首先需要在序列中定义一个（公共）位置。理想情况下，轨迹数据像来自社交网络服务的用户签入序列一样，每个位置都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，每个点的特征是一对 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。如果没有妥善处理这些点，将导致巨大的计算成本。下面将介绍几种具体的序列模式挖掘技术。

（1）自由空间中的序列模式挖掘

基于线简化的方法（line-simplification-based method）：2005 年就有人提出了旨在应对上述问题的早期解决方案。该解决方案首先通过使用像 Douglas-Peucker 的线简化算法来识别轨迹的关键点。然后将轨迹的碎片组合到每个简化的线段附近，以便计算每个线段的支撑，不考虑轨迹中两点之间的行进时间。

基于聚类的方法（clustering-based method）：解决上述问题的更一般的方法是将不同轨迹的点聚类到感兴趣的区域。然后用点所属的簇 ID 表示轨迹中的点。因此，轨迹被重新形成在不同轨迹之间可比较的簇 ID 序列。

（2）道路网络中的序列模式挖掘

当将序列模式挖掘问题应用于道路网络设置时，可以首先使用地图匹配算法将每个轨迹映射到道路网络上。然后，轨迹由路段 ID 的序列表示，可以将其视为字符串。因此，针对字符串设计的某些序列模式挖掘算法可以适应于寻找序列轨迹模式。当轨迹数据集非常大时，需要在其后缀树的深度上设置约束。另外，从后缀树导出的顺序模式必须是连续的。虽然没有明确考虑时间约束，但考虑到路径的速度约束，两个对象在同一路径上的行进时间应该相似。

3.3.4　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活动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移动。

周期模式挖掘已被广泛地用于时间序列数据。例如，Yang 等人试图从（分类）时间序列中发现异步模式、令人惊讶的周期性模式和差距罚分的模式。由于空间位置的模糊性，为时间序列数据设计的现有方法不能直接适用于轨迹。为此，Cao 等人提出了一种从轨迹中检索最大周期模式的有效算法。该算法遵循类似于频繁模式挖掘的范例，其中需要（全局）最小支持阈值。然而，在现实世界中，周期性行为可能更复杂，涉及多个交错周期、部分时间跨度和时空噪声以及异常值。

为了处理这些问题，Li 等人提出了一种用于轨迹数据的两阶段检测方法。在第一阶段，该方法通过使用基于密度的聚类算法来检测运动对象频繁访问的几个参考点，然后将移动物体的轨迹变换为一些二进制时间序列，每个时间序列指示移动物体在参考点处的“in”（1）和“out”（0）状态，接着通过对每个时间序列应用傅里

叶变换和自相关方法，可以计算每个参考点的周期值。第二阶段通过使用层次聚类算法总结了部分移动序列的周期性行为。2012 年，Li 等人进一步研究了从不完整和稀疏的数据源里挖掘周期性模式。

3.4 轨迹语义建模和标注

3.4.1 轨迹语义转化

轨迹知识发现是指学者们以对数据的深刻理解为前提进行研究，试图融合各种相关信息，理解轨迹数据背后的时空与行为特征，将轨迹数据转换为易于理解的语义轨迹（semantic trajectory），构建轨迹数据仓库。

移动性理解（mobility understanding）表示对轨迹的认知首先是从时空角度对用户运动方式进行分析，研究基于轨迹运动方式（如步行、骑车、公交、自驾等）的轨迹分段与标注，设计一种基于条件随机场模型的算法来最大化分段精度，使对轨迹运动方式的精准标注成为可能。近年来，人们越来越多地关注如何通过时空统计的方法理解移动对象的共性移动，汇总趋势性信息。

行为理解（activity understanding）的目标是理解用户在轨迹中的行为或可能的行为。对轨迹行为的理解需要在时空维度之外引入文本描述，现有方法主要有两种。第一种是将轨迹数据与兴趣点（point of interest）和签到（check-in）数据结合，丰富用户在轨迹驻留点可能的行为内容。第二种是从社交媒体、签到数据中爬取行为轨迹（activity trajectory），其中每个轨迹点包含时空、文本和其他信息，表示了用户在不同位置的状态和行为。与传统时空轨迹相比，上述轨迹包含了更多维度信息，因此难于管理。

轨迹相似性（trajectory similarity）用于评估不同轨迹之间的时空曲线和语义相似程度，是轨迹搜索与挖掘的核心。针对时空环境下的轨迹相似性度量，人们在此基础上进行了时间维度的扩展。针对包含用户行为信息的语义轨迹，定义了一种融合文本相似度的轨迹距离，并针对轨迹不确定性定义了一种基于概率的相似性评估机制。

3.4.2 轨迹语义标注

传统的轨迹挖掘研究主要关注轨迹时空特征的提取，往往是从轨迹数据自身出发自下而上进行挖掘分析，片面强调计算模型的形式化，导致信息得不到有效利用。因此，时空信息和领域知识的有效融合是推动轨迹挖掘研究继续发展的重要途径。轨迹语义标注旨在利用时空信息和领域知识对原始轨迹数据进行语义丰富处理，其本质上属于轨迹分类问题，即根据行为、交通方式等特征来区别不同类型轨迹。可以采用专用算法对不同类型的空间对象进行语义标注，主要包括区域标注、路段标注和位置标注。

区域标注专注于计算轨迹与空间区域的拓扑相关性，如基于轨迹与特定兴趣点集合，建立空间关联模型来计算公交站点间的频繁移动；通过采用相似的数据抽象

概念隐藏行人位置来保护个人隐私。路段标注专注于设计有效的地图匹配算法，旨在准确地识别车辆行驶路段，并近似地计算车辆在路段中的位置。如基于路段特征（如路段长度、平均速度和停止率）研究交通模型。已有的地图匹配算法侧重于优化匹配精度，通常每种算法仅适用于一种交通工具（如自行车、汽车和卡车等）。综合考虑不同类型的交通方式（如公交车、地铁等），很大程度上优化了地图匹配算法的匹配精度。位置标注旨在基于轨迹聚类和强化推理技术精准识别轨迹兴趣点。如基于预先定义的地理热点集合，设计一种语义时空关联模型推测移动对象的行为，专注于工作、家庭位置，挖掘轨迹数据中的周期性行为。

3.5　苏宁云商轨迹大数据实例

线上电子商务企业与线下零售企业正互相渗透融合，O2O 模式已经成为极具前景的电子商务发展模式，本节介绍的苏宁云商案例正是基于 O2O 的轨迹大数据实例化应用。在苏宁云商打造的有显示度的实例化应用中，以 O2O 电子商务决策支持为导向，构建 O2O 商务大数据融合框架，突破轨迹大数据挖掘方法，深入研究支持 O2O 一体化的多渠道知识融合方法。通过研制 O2O 商务大数据分析平台，实现线上线下资源互补和应用协同，提升企业管理效率与经营绩效，推动大数据环境下商业模式创新。

3.5.1　研究思路

本节主要围绕基于室内轨迹数据挖掘的用户线下行为分析及其商务应用展开深入研究，并针对若干实践领域开发系统原型和实例化应用。具体而言，主要包括如下 3 个方面。

① 以商务场景为例，研究室内轨迹数据的基础支撑技术，需要包含室内空间建模（如兴趣区域划分及其连接路径建模）、用户驻留点定位及行走路径推测。

② 依托用户轨迹数据，在不同类型的室内场景（如购物商场、综合商业中心等）里研究用户线下行为模式，结合时空特性，分析线下行为与购买动机、购买决策、策划活动之间的关联及相互影响。

③ 研究基于线下行为的用户偏好建模与预测方法，尤其需探索线上线下数据的知识融合技术，并利用这些技术和采集的数据，构建室内综合导航和推荐系统实例。

3.5.2　数据采集

本案例中轨迹大数据包含网页端、移动端访问日志及线下用户移动轨迹，是 O2O 商务中体量最大、变化最快的数据类型。

对于线下轨迹大数据，在苏宁易购南京商贸店 4 层楼共部署了 23 个 WiFi 传感器，覆盖了该店的主要营业区域，如图 3-8 所示为部署传感器的柜台照片，如图 3-9 所示为 2 层的地图以及传感器部署的位置。

图 3-8 部署传感器的柜台照片

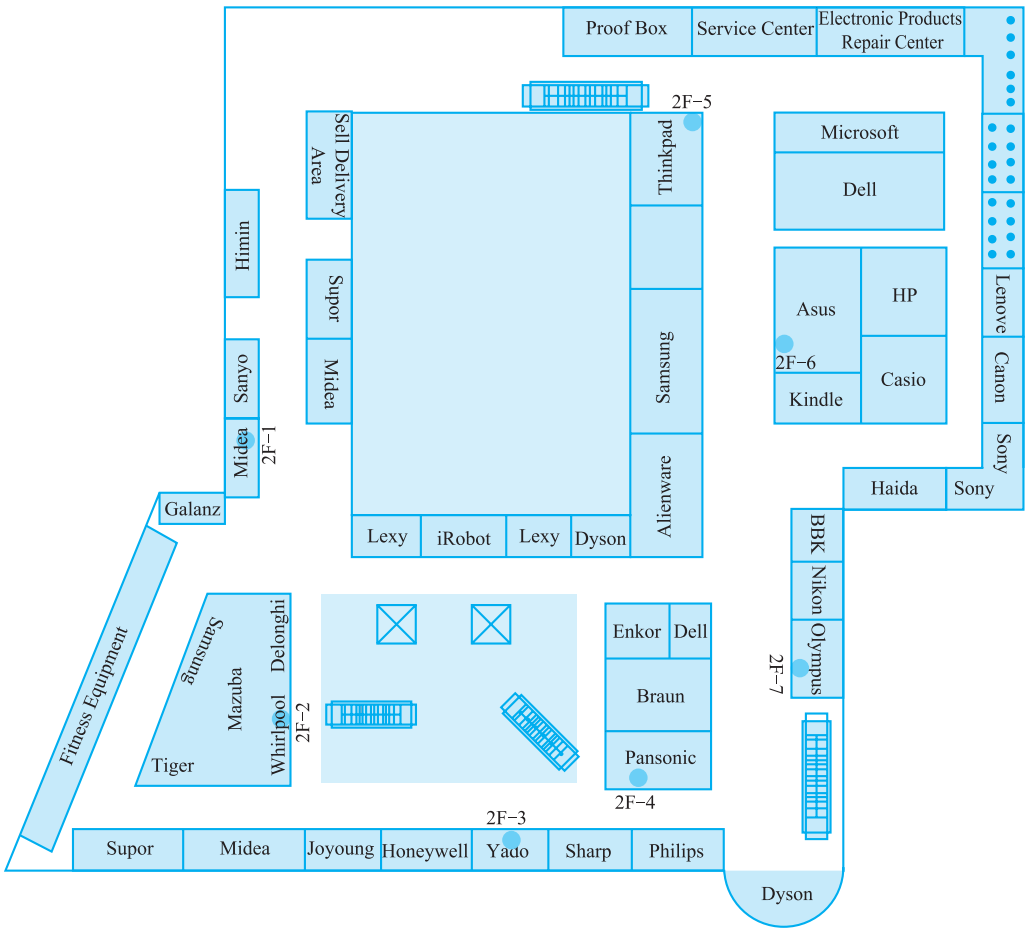


图 3-9 2 层平面图及传感器位置

当顾客进入苏宁并且携带打开 WiFi 的手机时，该客户周围的传感器就可以收集一个或多个数据记录，部署的传感器每隔 30 s 进行一次探测，每条探测记录都包含具有以下属性的信息：MAC、sensor_ID、timestamp、phone_brand 和 RSSI。

MAC 是用于识别手机的全球唯一编号，在本研究中它被视为用户 ID；sensor_ID 为传感器的编号；phone_brand 为手机品牌，品牌由手机的 MAC 产生，例如 iPhone、三星、华为等；RSSI 为接收信号强度指示器，范围是［0，99］，以指示客户和传感器之间的粗略距离。

线上轨迹数据通过苏宁易购 App 日志数据获取，如果用户注册为会员，将进一步关联用户的更多数据，比如将手机 MAC 与 PC 端日志关联。然后通过手机 MAC，能够将线下轨迹数据与苏宁易购 App 日志数据，即线上轨迹数据相关联，获取到商品信息、用户评论、交易数据和用户关系等信息，从而做进一步研究，如图 3-10 所示。

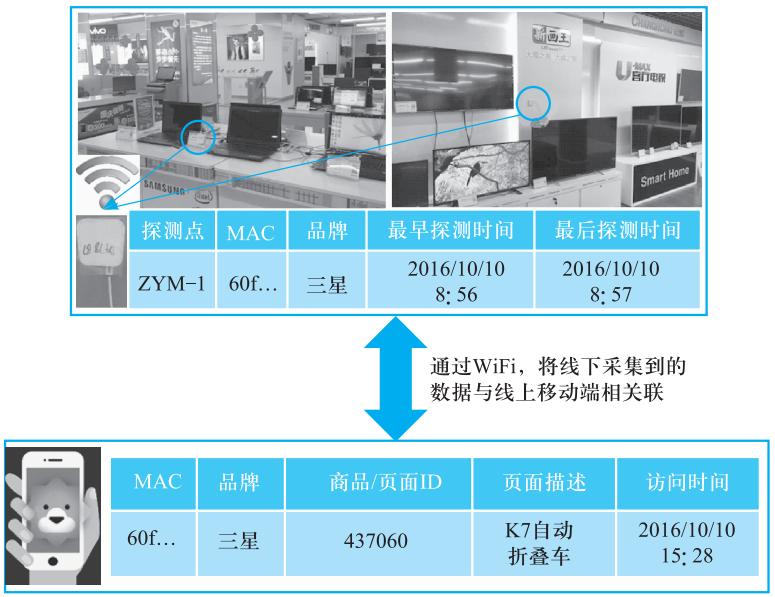


图 3-10　线上线下数据联系

3.5.3　数据预处理

线下轨迹数据数据量庞大，仅 2017 年 10 月 18 日—12 月 25 日这段时间内，已经通过传感器采集了超过 3 049 万条记录，但是其中有许多无用数据。比如由于 WiFi 传感器的检测范围是圆形区域，因此会检测到经过商场的用户。此外，短时间内在购物中心的用户的轨迹在分析客户行为方面几乎没有作用。因此，有必要对原始数据进行预处理，以保证对真正进入和漫步在购物中心周围的用户进行后续数据分析。

针对商场这个特定营业环境噪声数据的特点，使用三阶段法来过滤原始跟踪数据。

① 去除非营业时间探测到的记录。这也就是保持时间戳［10am，9pm］的记录，目的是将所有跟踪记录保留在工作时间内，并保留绝大多数原始数据。

② 去除距离传感器很远的手机产生的记录。这也就是保留 RSSI ≤ 60 的记录，用于过滤远离传感器检测到的噪声记录，比如路过商场的人。通过将 RSSI 的阈值

设置为 60，仅保留了 8.51%的用户和 9.62%的记录。

③ 去除在卖场中停留时间很短的顾客产生的记录。这也就是保持用户至少被检测到两次，并且两次检测之间的时间间隔大于 10 分钟，目的是过滤在商场中短时间停留的用户，例如少于 10 分钟的记录。

表 3-3 显示了逐步使用上述三个条件过滤的用户和记录的比例。

表 3-3 数据预处理的结果

类别	原始数据	过滤器 1	过滤器 2	过滤器 3
用户数	201 621	171 620	14 602	5 587
记录数	6 510 307	5 504 721	529 579	501 427

根据传感器收集到的用户轨迹数据，可以得出数据的基本统计信息如图 3-11 所示。

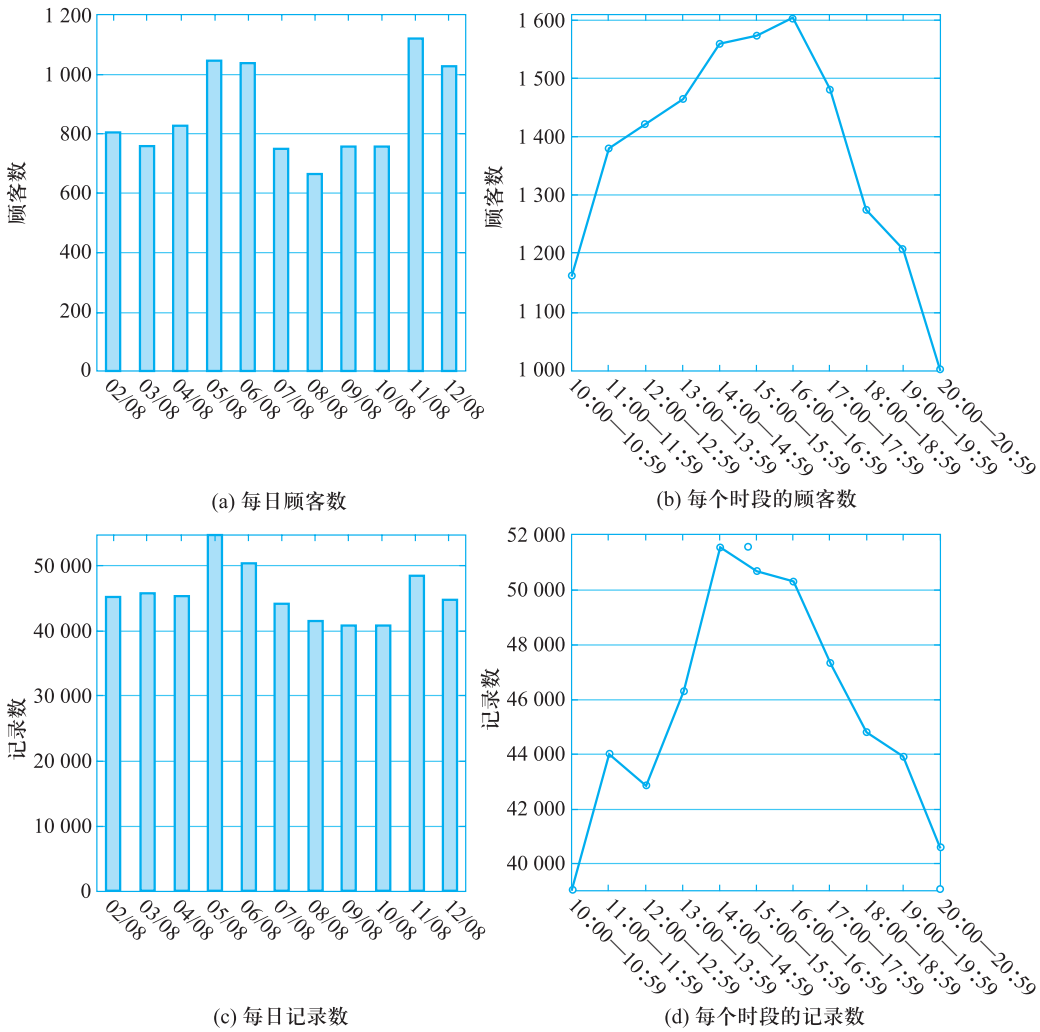


图 3-11 基本统计信息

3.5.4　顾客行为分析

根据收集到的用户轨迹数据，聚焦于小范围内轨迹数据挖掘及时序价值模式的定义和挖掘方法，可以对顾客的行为进行分析，主要分为两个方面。

第一个方面是从群体的角度对顾客的行为进行分析，主要任务如下。

划分顾客类型：利用聚类技术将用户划分为 5 类。

分析行为特点：根据每类用户在不同特征上的取值，分析他们的行为特点。

对于群体顾客行为分析，首先是特征构建，基于获取的数据集，为每个用户构建一个四维向量，以反映各种手机的行为特征。每个用户的四维向量有如下属性。

访问天数（NoVD）：NoVD 是用户在 2017 年 8 月 2 日和 12 日期间访问购物中心的天数。因此 NoVD ∈(0，11)。

访问时间段数（NoVT）：将上午 10 点到晚上 9 点之间的时间间隔平均分为 11 个时段，每个时段一个小时。NoVT 是至少一个传感器检测到用户的时间段数，因此 NoVT ∈(0，11)。

记录数（NoR）：NoR 是从所有传感器收集的用户记录的编号。将上限设置为 500 以避免值的范围变得过大，因此 NoR ∈(0，500)。

传感器数量（NoS）：NoS 表示至少检测一次用户的传感器数量。由于部署了总共 23 个 WiFi 传感器，因此 NoS ∈(0，23)。

将每个用户表示为四维特征向量，并使用归一化方法将每个特征值转换为区间[0，1]。然后，使用带有余弦距离的 Kmeans 方法将 5 587 个顾客划分为 5 个集群，也就是将顾客分为了 5 类。在表 3-4 中给出了每个集群中的顾客数。从中可以清楚地观察到，集群 1（C#1）和集群 3（C#3）是大集群，其余 3 个集群则相对较小。

表 3-4　5 个集群的规模和重心

类别	C#1	C#2	C#3	C#4	C#5
顾客数	4 015	135	1 116	210	111
NoVD	0.008	0.961	0.03	0.62	0.629
NoVT	0.008	0.795	0.017	0.237	0.381
NoR	0.021	0.936	0.062	0.229	0.915
NoS	0.084	0.192	0.309	0.256	0.388

注：2~5 行是每个集群重心的特征值。

根据划分出的 5 个集群，对其进行行为分析，图 3-12 描绘了 5 个集群内的每个顾客的特征值。

首先分析两个最大的集群，即 C#1 和 C#3，如表 3-4 所示，这两个集群中的用户具有极低的 NoVD、NoVT 和 NoR 值，这意味着他们很少去购物中心。同时，两个集群包括约 92% 的用户。因此，可以相信这两个集群中的大多数用户都是具有购物意向的真正客户。如图 3-12（a）和图 3-12（c）所示，C#3 的 NoS 和 NoVD

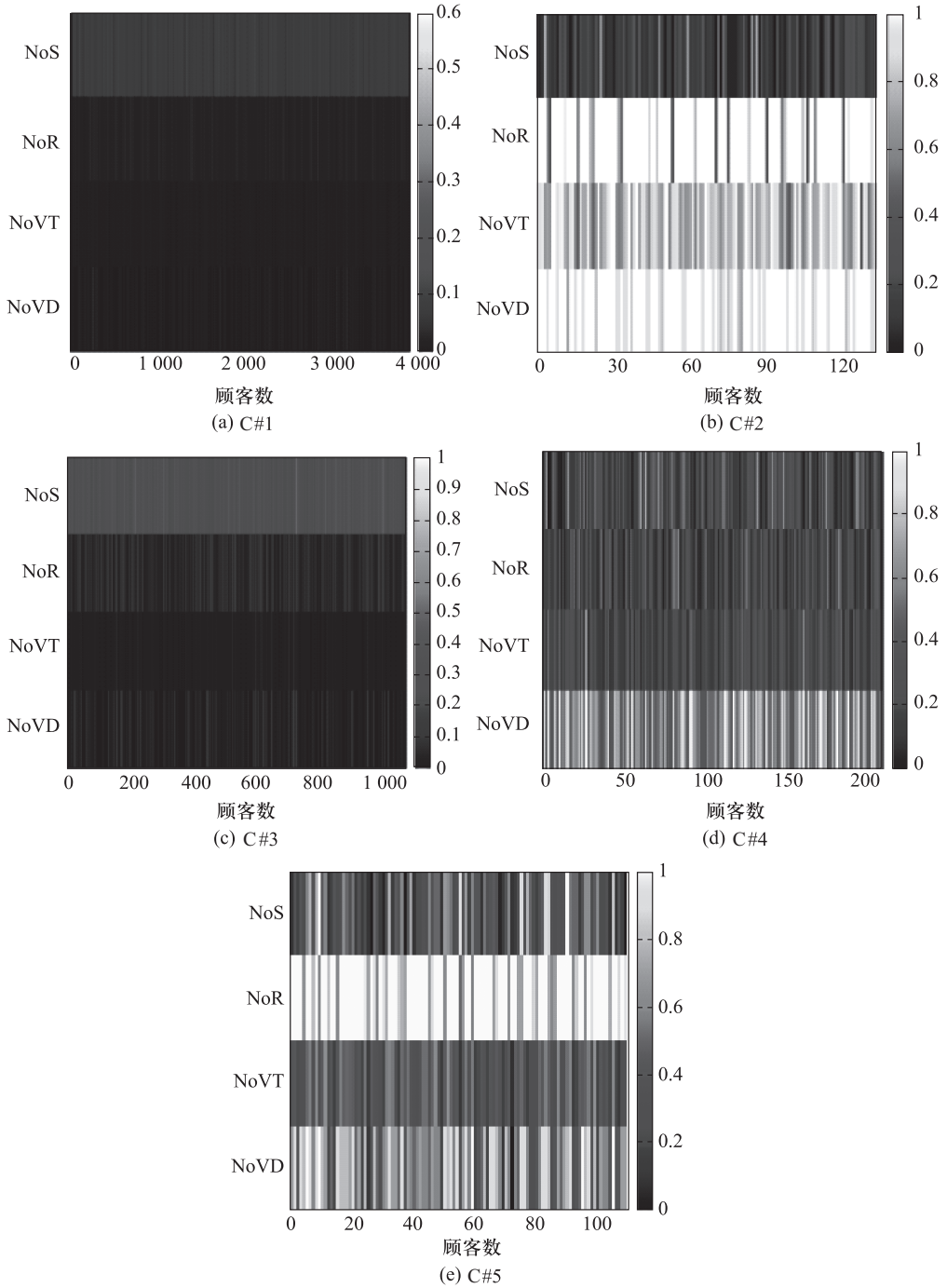


图 3-12 5 个集群的 4 个特征热点图

值均高于 C#1。这表明 C#3 中的用户更频繁地访问商场（即在更多天内检测到），并且在商场内的顾客比 C#1 中的更为广泛（即由更多传感器检测到）。直观地说，可以将 C#1 标记为偶尔客户群，并将 C#3 标记为忠诚客户群。

除了两个大型集群外，可以发现三个小集群也非常有趣。对于 C#2，可以从图 3-12（b）中观察到 NoVD、NoVT 和 NoR 值非常大，但 NoS 值非常小，这意味着这些手机总是留在商场里，几乎一动不动。在整个服务时间内，将 C#2 称为用于展览目的的智能手机。因此出现了大量的记录，这些手机的位置相对固定。当仔细观察 C#4 和 C#5 时，发现它们的 NoVD、NoVT 和 NoR 值远远高于 C#1 和 C#3，但略低于 C#2。此外，它们的 NoS 值显然高于 C#2。因此，将这两个群集中的用户视为工作人员，因为他们通常在工作时间留在购物中心并且往往在某个地区移动。如果进一步比较 C#4 和 C#5 之间的 NoS 值，可以发现 C#5 的 NoS 值平均更高，这可能是由工作人员的不同角色造成的。例如，一些工作人员应该总是在一个固定的柜台，而另一些工作人员可能会参观专柜，比如经理。基于上述分析，将 C#4 标记为定点员工，将 C#5 标记为巡航员工。综上，将 C#1 标记为偶尔客户群，C#3 标记为忠诚客户群，C#4 标记为定点员工，C#5 标记为巡航员工。

以上均是从群体的角度分析顾客行为，下面从个体的角度对顾客的行为进行分析，主要任务如下。

识别热点区域：利用回归模型，找出卖场的热门区域，也就是顾客通常长期停留的区域。

识别热点路径：利用马尔可夫模型，找出卖场的热点路径，也就是顾客走进购物中心的最可能路径。

首先是对热门区域的分析，人们经常观察到购物中心的一些地区挤满了人。具体而言，客户往往会在他们感兴趣的产品展示区域逗留并花费更多时间，然后匆匆走过不感兴趣的区域。将客户可能在较长时间内停留的区域称为热门区域，通过对商场内热门区域仔细检查，可以发现顾客购物兴趣的整体特征。正如在上文群体顾客行为分析中所讨论的，C#1 和 C#3 中的绝大多数用户都是真正的客户，而其他用户可能不是。因此，分析主要集中在 C#1 和 C#3 上，而忽略另外三个。

使用以下回归模型识别电子商城中的热门区域：

$$\log T_{i,d}=\beta \log r_{i,d}+\varepsilon_{i,d}$$

其中，$T_{i,d}$ 是第 i 天客户的停留时间，$r_{i,d}$ 是一个衡量客户在不同时间的停留时间的区域，β 是回归系数，$\varepsilon_{i,d}$ 表示误差。在 $T_{i,d}$ 和 $r_{i,d}$ 上加对数是为了减轻极大值的影响。

根据以上回归模型，识别出电子商城中的热门区域，结果如表 3-5 所示。

表 3-5　顾客停留的热门区域

Sensor_ID	系数	热门区域描述
1F-1	0.11	入口处和星巴克附近
1F-3	0.088	入口处和中国移动附近
1F-7	0.073	销售中国手机的柜台附近
1F-2	0.068	电梯附近和销售中国手机的柜台附近

续表

Sensor_ID	系数	热门区域描述
2F−5	0.098	自动扶梯附近和笔记本销售区
2F−2	0.059	小家电销售区
3F−1	0.095	电视机销售区，例如 LG
3F−3	0.084	自动扶梯附近和创维电视专区
4F−2	0.075	自动扶梯附近和小家电销售区
4F−1	0.07	空调销售区，例如海尔、大金

第一，可以看到三个入口的附近都是热门区域，即传感器 1F−1、1F−3 和 1F−7 周围的区域。此外，几个热区也位于电梯或自动扶梯附近，例如，由传感器 1F−2、2F−5、3F−3 和 4F−2 覆盖的范围。可以推断这两种类型的区域对客户非常有吸引力。第二，传感器 1F−1 的系数估计是最大的，表明它对客户的总停留时间产生重大影响。这是因为这个地区不仅靠近入口而且靠近星巴克。人们可能会留在星巴克并喝一杯咖啡，这使得逗留的时间显然比其他地区长。第三，非常有趣的是观察到销售中国手机的几个柜台，如华为、努比亚和魅族，被认为是热门区域，而出售国际品牌手机（如苹果）的柜台则不是热门区域。部分原因在于各种中国手机品牌的竞争日趋激烈，越来越多的中国消费者愿意购买国内厂商生产的手机。第四，可以发现，热门区域包含两个销售小家电的区域，即传感器 2F−2 和 4F−2 附近的区域，这意味着很多客户可能会更加关注以及有意购买此类产品。最后，空调销售区也是热门区域，即传感器 4F−1 的周围区域。推测这主要是因为碰巧在炎热时期收集的数据，这时期正是空调需求旺盛时期。

对于电子商城中的每对非交换的地点 i 和 j，当然存在一个或多个路径，顾客可以从 i 到 j 进行。通常，不同的客户倾向选择不同的路径，并且一些路径明显比其他路径更受欢迎。定义从一个地方到另一个地方的路径为热点路径，前提是它是所有候选路径中最可能的路径。通过对热点路径的详细分析，可以揭示客户运动的鲜明特征。与热门区域研究相同，主要关注 C#1 和 C#3 中的用户。使用马尔可夫模型，分析得出商场的热点路径如表 3−6 所示。

表 3−6 热 点 路 径

起点	中间点	目的地
1F−1	1F−2	1F−4
	1F−2 → 1F−4	1F−5
	1F−3 → 1F−7	1F−6
	1F−3	1F−7

续表

起点	中间点	目的地
1F−7	1F−5 → 1F−4 → 1F−2	3F−1
	1F−5 → 1F−4 → 1F−2 → 3F−4	3F−2
	1F−5 → 1F−4 → 1F−2	3F−3
	1F−5 → 1F−4 → 1F−2	3F−4
	1F−5 → 1F−4 → 1F−2 → 3F−3	3F−5
4F−4	1F−2	1F−4
	1F−2 → 1F−4	1F−5
	1F−2 → 1F−4 → 1F−5	1F−6
	1F−2 → 1F−4	1F−7

可以观察到许多热门路径穿过传感器 1F−2 覆盖的区域，例如，从传感器 1F−7 到部署在第三层的所有传感器的路径以及从传感器 4F−4 到传感器 1F−4、1F−5、1F−6 和 1F−7 的路径。换句话说，如果从图论的角度来看待这个问题，传感器 1F−2 的中介中心性明显大于其他传感器。这主要是因为在传感器 1F−2 附近有两个电梯，便于客户直接去其他楼层。此外，似乎有一些热点路径是“奇怪的”。作为示例，将传感器 1F−7 连接到传感器 3F−5 的热点路径越过传感器 3F−3 的检测范围，而不是直接到达传感器 3F−5。这是因为当客户从传感器 1F−7 开始时，几乎没有任何客户选择直接进入传感器 3F−5，这使得相应的转换概率接近 0。另一方面，相当多的客户选择在转到传感器 3F−5 之前到传感器 3F−3。

综上所述，本节详细介绍了从苏宁商城收集到的轨迹数据，详细解释了特殊的预处理过程和数据统计。然后，将用户划分为多个集群，并详细阐述每个集群的基本特征。此外，从个人角度审视客户的行为，具体而言，引入回归模型以寻找商场中的热门区域，并计算马尔可夫转移矩阵以及连通矩阵，以便从跟踪数据中识别热点路径。

3.6　本章小结

在当今的大数据时代下，移动互联网的普及发展形成了海量的移动对象轨迹数据，这些数据含有大量的时空特征信息，来源也多种多样，人类活动规律、行为特征、城市车辆移动路线等轨迹数据可以反映人们的个人行为、兴趣爱好和社会环境。由本章前文提供的表格可以看到，代表性轨迹数据量庞大，日均采样量都达到千万甚至百亿级，数据总量达到 TB、PB 级。这体现了大规模性、实时高速性、多样性、高价值性的“4V”特征。例如，导航服务公司每天都会存储处理数千万的

数据，而数据的不同属性也会对数据的分析处理结果产生影响。

轨迹数据被广泛应用到智能交通、位置服务等系统。本章介绍的大众化经验路径推荐、交通状况精准预测、城市规划智能决策、个性化服务与活动推荐和出租车服务都与人们的生活息息相关，这些系统应用则要求对轨迹数据进行有效处理，让原始数据逐步提取可用。其中主要的步骤就是轨迹数据预处理、轨迹模式挖掘、轨迹语义建模和标注。

由调查可知每天采集到的轨迹数据都是海量的，日积月累，庞大的数据集都会存在杂乱无序、数据不准确、数据缺失等问题，从而导致数据无法使用，因此数据预处理就成了首要任务。由于噪声的影响，定位服务系统无法准确探测到使用者所在的位置，定位系统中反馈出来的结果与实际位置产生巨大偏差，这对使用者方向的确定和路线的寻找带来了很大困扰，通过均值（或中值）滤波器、卡尔曼和粒子滤波器、异常检测的方法对噪声进行过滤，从而排除它对空间轨迹信号的干扰；空间轨迹中点的作用也各不相同，应用系统更加关注人们停留在了哪几个点上，这些驻留点某些情况下可以体现出人们的个人兴趣和重要程度，但是部分时刻只是显示该对象正在进行途中，因此通过驻留点的检测可以剔除无用的部分；对物体对象轨迹的精确记录需要耗费大量的成本，然而在很多应用情况下并不需要对轨迹精度有很高的要求，因此通过轨迹压缩的方式，在不损害轨迹数据精度的前提下减小轨迹的大小从而节约开销；对于轨迹数据中的聚类部分，则可以通过轨迹分割的方式将其分为若干个部分，在减少计算复杂度的同时可以在不同方面进行不同的挖掘研究。

数据预处理完成后总离不开数据挖掘的研究，从基于轨迹的数据挖掘角度展开，发现了其中 4 种主要模式：伴行模式、轨迹聚类、序列模式和周期模式，这几种主要模式都可以帮助人们对数据进行更深一步的研究。当对数据内容有了充足的认识后，试图理解数据背后的含义，因此将轨迹数据进行语义转化和标注，通过对移动和行为理解，分析出目标用户在整个过程的行为目的和接下来的可能行为，并对其进行精准标注。

本章的苏宁云商轨迹大数据实例是通过部署多个 WiFi 传感器，收集多个轨迹数据信息，将线上数据和线下轨迹数据相结合，从而进行研究。首先是将传感器中收集到的信息进行预处理操作，由于传感器采集的是经过商场的用户信息数据，然而其中很多用户可能只是中途经过或短暂停留，这类的数据对后续的分析没有帮助，因此通过去除非营业时间内的用户、距离传感器较远的用户和停留时间短的用户，筛选出了真正进入商城的用户信息。接下来通过轨迹数据分析用户行为，通过群体和个体用户的行为分析建立模型，从而根据热点图确定商场中的热门区域，通过分析热门区域的主要结构和部署，确定在这个时间段内顾客大多会处于哪一块区域以及这样的原因。通过这个实例，发现入口附近、中国手机销售区附近、小家电销售区附近、空调销售区附近属于热门区域，这也可以和社会背景及客观现实相联系起来，对于在实际生活的应用研究也能起到很大的帮助。

习　题

1. 简述噪声过滤的常用方法。
2. 简述驻留点检测算法的意义和应用。
3. 比较轨迹压缩中的 Douglas-Peucker 算法和滑动窗口算法。
4. 简述常用的伴行模式，并比较不同方法的区别。
5. 简述轨迹序列模式的定义和常用方法。
6. 简述轨迹语义标注的概念和意义。