doi:10.11959/j.issn.1000-436x.2015318

轨迹大数据：数据、应用与技术现状

许佳捷[1,2]，郑凯[1,2]，池明旻[3]，朱扬勇[3]，禹晓辉[4]，周晓方[1,2]

(1. 苏州大学 计算机科学与技术学院，江苏 苏州 215006；2. 江苏省软件新技术与产业化协同创新中心，江苏 南京 211102；
3. 复旦大学 计算机科学技术学院 上海市数据科学重点实验室，上海 201203；4. 山东大学 计算机科学与技术学院，山东 济南 250101)

摘 要：移动互联技术的飞速发展催生了大量的移动对象轨迹数据。这些数据刻画了个体和群体的时空动态性，蕴含着人类、车辆、动物的行为信息，对交通导航、城市规划、车辆监控等应用具有重要的价值。为了实现有效的轨迹数据价值提取，近年来学术界和工业界针对轨迹管理问题开展了大量研究工作，包括轨迹数据预处理，以解决数据冗余高、精度差、不一致等问题；轨迹数据库技术，以支持有效的数据组织和高效的查询处理；轨迹数据仓库，支持大规模轨迹的统计、理解和分析；最后是知识提取，从数据中挖掘有价值的模式与规律。因此，综述轨迹大数据分析，从企业数据、企业应用、前沿技术这 3 个角度揭示该领域的现状。

关键词：时空数据库；轨迹数据管理；数据索引；查询优化

中图分类号：TP392 **文献标识码**：A

Trajectory big data: data, applications and techniques

XU Jia-jie[1,2], ZHENG Kai[1,2], CHI Ming-min[3], ZHU Yang-yong[3], YU Xiao-hui[4], ZHOU Xiao-fang[1,2]

(1. School of Computer Science and Technology, Soochow University, Suzhou 215006, China;
2. Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211102, China;
3. Dept. of Computer Science, Shanghai Key Laboratory of Data Science, Shanghai 201203, China;
4. School of Computer Science and Technology, Shandong University, Jinan 250101, China)

Abstract: The fast development of mobile internet has given rise to an extremely large volume of moving objects trajectory data. These data not only reflect the spatio-temporal mobility of individuals and groups, but may also contain the behavior information of people, vehicles animals, and other objects of interest. They are invaluable for route planning, urban planning and vehicle monitoring, etc., and tremendous efforts have been made to support effective trajectory data management, including trajectory data pre-processing, which handles issues such as high redundancy, low precision and inconsistency of sampling; trajectory database technologies, concerning the efficient and effective storage of trajectory data and query processing; trajectory data warehousing, which supports the analytics on large-scale trajectory data; knowledge discovery, by which useful patterns can be extracted from trajectory data. A survey of trajectory big data analytics from three different aspects: data, applications and techniques is provided.

Key words: spatio-temporal database; trajectory data management; indexing structure; query processing

1 引言

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，形成了时空轨迹（spatio-temporal 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。每条轨迹由一序列时空采样点构成，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，刻画了人们在

收稿日期：2015-10-21；**修回日期**：2015-11-27

基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目(2015CB352500)；国家自然科学基金资助项目(61232006, 61402312, 71331005, 61272092)；山东省科技发展计划基金资助项目(2014GGE27178)

Foundation Item: The National Basic Research Program of China (973 Program) (2015CB352500); The National Natural Science Foundation of China(61232006，61402312, 71331005，61272092); High-tech R&D Program of Shandong Province(2014GGE27178)

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。

高质量的轨迹数据具有重要的社会和应用价值，不仅为解决交通拥堵、改善交通服务、监控道路环境、缓解能源紧缺等社会问题提供了新的机遇，而且对认知人们的社会活动、优化公共资源配置有着特殊意义，成为各政府与企业的重要财富并受到广泛重视。在此背景下，轨迹大数据管理被学术、工业界大量研究，轨迹数据分析与挖掘已经成为数据挖掘领域的一个重要的新兴分支。工业界和学术界针对大规模轨迹数据存储与分析技术开展了大量的理论和系统探索工作，包括轨迹预处理、索引、查询优化、轨迹分析与挖掘等。这些成果的使用显著提升了政府管理、社会服务、企业盈利能力，并深入影响了人们的生活方式。但是随着数据规模的指数级增长，应用需求的飞速提升，现有的轨迹数据存储、计算和分析方法面临诸多局限，亟需突破轨迹数据的处理架构、分布式算法等关键技术。

本文将从轨迹分析的需求入手，从数据、应用、技术 3 个方面阐述该领域现状和发展。在数据方面，介绍轨迹数据的类型、规模、频率等指标，并分析它们对轨迹数据管理的影响；在应用方面，将介绍各种轨迹数据的典型应用及其场景，并分析其现状和发展趋势；在轨迹管理技术方面，将分类介绍轨迹数据存储与分析领域的科学问题和前沿技术，最后展望大数据环境下的轨迹管理技术存在的问题和发展方向。

2 企业级轨迹数据现状

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，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。

本文首先介绍不同采集方式下真实的企业轨迹数据。表 1 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 CellID 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。

由表 1 可知，轨迹数据继承了大数据的经典“3V”特征，即量大（volume）、实时（velocity）、多样（variance）。此外，移动对象轨迹数据库的一

表 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	车载 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
公交轨迹	公交卡	小时级	百万–千万级	TB、PB 级
卡口数据	卡口抓拍	分钟级	千万级别	TB 级
行为轨迹	社交媒体	分钟、小时级	百万–千万级	PB 级

些特有特征可以总结如下。

1) 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，显著加大了搜索与分析的处理复杂度。

2) 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，这种差异性极大影响了轨迹的相似性度量与分析。

3) 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。

4) 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。

综上，轨迹数据语义丰富，蕴含着各种移动对象的时空和行为信息，被广泛应用在诸多企业级应用中。而轨迹数据的上述特征给轨迹数据处理与分析提出了一系列要求与挑战。

3　企业级轨迹应用

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。如表 2 所示，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，具体应用主要包括以下几方面。

1) 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在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泛在的移动模式，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。

2) 交通路况预测。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。

3) 城市规划。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，提炼城市内、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。

4) 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和个性化景点、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。

5) 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

表 2　　**代表性轨迹分析应用**

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、历史事故等数据	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、社交网络与社交媒体数据、兴趣点和签到、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
出租车服务	出租车 GPS 轨迹、私家车移动轨迹、公交线路与轨迹等数据	应用于相关业务优化，有进一步提升空间

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）；为行人提示就近的有效打车地点，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。

在上述应用系统中，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。下面将介绍轨迹数据管理与分析技术的前沿成果与研究现状。

4 轨迹搜索与分析技术现状

从轨迹数据的生命周期来看，图 1 展示了轨迹数据金字塔模型，代表了轨迹数据的不同认知程度和可用性层次。各层之间紧密联系，相互依托。最底层是原始轨迹数据，存在很多冗余和噪音，无法被直接使用。通过数据预处理，由一系列操作将其转换为校准轨迹。校准轨迹是可用数据，但是无法被有效检索和分析，需要通过数据库管理技术对其有效组织，成为能够有效存取的数据库轨迹。在此基础上，需要进一步对时空、文本等属性的理解分析，构成轨迹数据仓库，形成语义轨迹。最后通过挖掘和分析处理等手段从语义轨迹中得到有用的轨迹知识，服务各类应用。轨迹金字塔模型体现了轨迹数据的知识化过程。

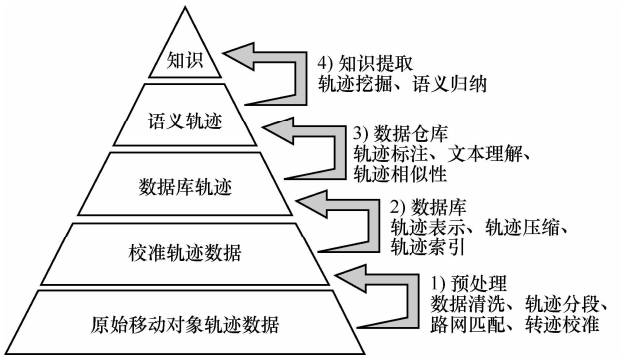

图 1 轨迹金字塔

过去 10 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，信息和知识能够被从中提取。这些技术按照轨迹金字塔模型分层展开。如图 1 所示，它们的目标是使轨迹从底层向高层转化，可以被大致归纳为数据预处理（data preprocessing）、轨迹数据库（trajectory database）、轨迹数据仓库（trajectory data warehouse）、知识提取（trajectory knowledge discovery）。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。本节首先探讨这些关键问题，结合一些有影响力的科研成果阐述研究现状。

4.1 数据预处理

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（位置）；由设备、传输故障或误操作等因素导致的数据不完整，使部分（通常是一段时间和区域）数据缺失；由不同坐标表示更新策略和语境变换（例如轨迹数据集参照了多个地图或地图版本等）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。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，首先需要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以下 4 类操作。

1) 轨迹数据清洗（data cleaning）旨在去除轨迹中的冗余点（redundant points）和噪音点（noisy points）。冗余点是指可以通过插值等计算导出的采样，它们显著增加了系统的存储和计算开销，移动对象在静止和匀速运行状态中都会产生大量的冗余点；噪音点是指由软硬件设备异常导致的错误采样，它们会极大影响轨迹挖掘和分析结果。现有方法主要是从单条轨迹的角度清洗数据。借鉴曲线平滑思想，轨迹清洗算法[1, 2]智能选取少量的“代表性”采样点，去除大量的冗余采样，使该条轨迹的完整时空投影依然能够被有效表示（基于线性拟合算法）。结合一些时空规则，异常的噪音点得以被准确识别并去除。针对实时化的轨迹清洗要求，文献[3]提出了一种基于时间窗的在线清洗算法。

2) 轨迹分段（trajectory segmentation）是指对长时段轨迹（例如以天、月为单位）的合理切分与标注，切分后的每个子轨迹段代表一次出行记录，是原子级的轨迹分析对象。轨迹分段的核心问题是理解时空移动特征，主流的轨迹分段方式包括基于时间阈值、几何拓扑和轨迹语义这 3 种基本策略。以基于语义的轨迹分段为目标，郑宇在文献[4]中提出了一种基于 GPS 轨迹数据的停留点检测（stay points detection）方法，停留点是通过学习得到的经常作为起点或终点的位置或区域，例如天安门广场、首都机场等，是轨迹分段的重要参考对象；文献[5]提出

了一种基于轨迹聚类的停留点抽取方法。

3) 路网匹配（map-matching）是关联轨迹与数字地图，将 GPS 坐标下采样序列转换为路网坐标序列。路网匹配后的每个轨迹采样点都映射到一个路网位置，难点在于采样的位置误差、低采样频率、地图对路段连续拓扑的离散化表示等，使每个采样点坐标无法准确匹配路段。路网匹配算法的核心思想是利用轨迹点之间的时空可达性做匹配校正。为了实现高效的路网匹配，文献[6]提出了一种基于空间几何度量的匹配算法；文献[7]提出了一种基于拓扑 Frechet 距离的路网匹配算法；文献[8]采用隐马尔可夫模型，通过动态规划算法最大化匹配状态的转移概率，实现向地图空间的精准映射。文献[9~11]解决了基于简化地图的高效路网匹配问题。

4) 轨迹校准（trajectory calibration）是保证轨迹数据可用性的重要技术。面向低频采样轨迹，文献[12]提出了一种基于轨迹移动模式学习的位置不确定消减机制。轨迹数据的质量问题很大程度上是由于采样率差异过大导致的。当 2 个轨迹的采样率差别较大时，直接基于它们的采样点进行相似性比较没有意义，需要对原始轨迹校准以便合理评估轨迹相似性。文献[13]提出了 2 种考虑空间特性的校准模型，通过机器学习算法训练得到参照系统，在该系统中去除冗余数据并补全重要缺失采样。文献[14]进一步同时考虑了时间和空间属性，满足时空双重受限的轨迹相似性分析要求，大幅提升轨迹校准效果并得到高质量轨迹数据。

4.2 轨迹数据库

轨迹数据库是轨迹大数据管理的核心，是数据搜索与处理性能的保证。传统数据库技术不适用于管理高度冗余、非结构化变长的轨迹数据。为了实现大规模轨迹数据在数据库的有效管理与组织，人们针对轨迹数据模型、轨迹压缩、轨迹索引等核心问题展开了大量研究。

轨迹数据模型（trajectory data model）是轨迹数据在数据库中的表示方法，是数据组织与管理的基础。轨迹数据模型起源于移动对象数据库，早期工作主要是基于关系模型扩展，包括 Wolfson 等提出的 MOST 模型[15]，将移动对象的位置信息表示为动态属性；Guting 等在文献[16]中，设计了一种基于抽象数据类型 ADT 的模型和类似 SQL 的查询语言，以移动点和移动区域为基本抽象。近些年随着 Hadoop 的兴起，轨迹数据模型被极大简化，通常是以移动对象为中心，以序列化轨迹点来灵活表示，数据在 HDFS 中持久化保存。面向实时轨迹数据分析的需求，文献[17]借鉴视频数据表示机制，提出了一种以时间为中心的轨迹模型，基于该模型的轨迹数据库 SharkDB 适用于内存计算环境、对时间受限的轨迹分析（如实时交通流、拥堵识别与趋势分析等）具有天然优势。

轨迹压缩（trajectory compression）。由于轨迹数据的低价值密度和存储设备限制，数据库无法保存全部轨迹数据，通常需要对轨迹数据集进行压缩存储。现有轨迹压缩方法主要分为 3 类。

1) 基于路网（road network based）的压缩[18, 19]。对基于路网表示的轨迹通过路段拓扑和编码等方法来压缩轨迹数据存储空间。

2) 基于轨迹（trajectory based）的压缩[2, 3, 20]，主流研究侧重于单条轨迹的压缩，通过对移动对象轨迹建模，去除可通过模型（插值）还原的轨迹点，例如路网最短路径上的轨迹点，保证与原轨迹的误差符合精度范围。

3) 基于帧编码（frame encoding based）的压缩[17]，每个对象在关键帧中记录精确的采样信息，在非关键帧中仅记录采样与上一帧中的偏移值，从而大幅压缩数据量。

轨迹查询与索引（trajectory indexing）。轨迹分析依赖于大量的查询操作，根据用户给定的移动对象、时空范围、移动属性（如平均/瞬时速度、轨迹长度、采样频率等）值域等条件，返回用户或分析所需的相关轨迹。

轨迹数据的高效检索依赖于数据索引。在空间数据库中最经典的索引结构是 R-tree[21]及其改进版本，若直接使用三维的 3D R-tree 对轨迹索引将导致诸多问题，如死区(dead space)过大、轨迹的完整性被破坏等。针对这些问题 Pfoster 等在文献[22]中提出了 TB-tree 索引结构，严格地让每个叶节点只包含属于同一轨迹的线段以最大限度地保证轨迹的完整性。有些学者尝试把轨迹的时间和空间维度分别进行索引，把数据在时间维度上进行划分然后分别用 R-tree 组织起来，每个 R-tree 对应的是一个时间点（HR-tree[23]）或时间段（MV3R-tree[24]），这种结构可以更好地处理基于时间的轨迹查询。文献[25]提出了一种空间优先划分的格栅索引结构，对基于空间区域的查询有很好的效果。

除了上述通用轨迹索引，还有一些工作研究重

点在于面向定制查询的索引设计，例如面向路网受限轨迹查询的 NDTR-tree[26]以及基于 LCSS[27]、ERP[28]、EDR[29]、*k*-BCQ[30]等轨迹相似性度量的索引结构[27~30]。结合相应的查询优化技术，这些索引支持了各种类型的轨迹精准查询与个性化轨迹分析的快速处理。

4.3 轨迹数据仓库

轨迹知识发现以对数据的深刻理解为前提。学者们为此展开大量研究，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 trajectories），构建轨迹数据仓库。

移动性理解（mobility understanding）。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。文献[31]研究了基于轨迹运动方式（如步行、骑车、公交、自驾等）的轨迹分段与标注，设计了一种基于条件随机场模型的算法最大化分段精度，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，现有方法主要通过 2 种方式。第 1 种是将轨迹数据与兴趣点（point of interests）和签到（check-ins）数据结合[32, 33]，丰富用户在轨迹停留点可能的行为内容`。第 2 种是从社交媒体、签到数据中爬取行为轨迹（activity trajectories），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，郑凯等在文献[34]中提出了一种高效的检索框架。

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。在基于空间的轨迹相似性度量函数方面，除了经典 DTW、LCSS，陈雷等在文献[29]中定义一种基于轨迹编辑距离的 EDR 函数，文献[35]定义了考虑连续性的度量指标 OWD。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[27]。针对包含用户行为信息的语义轨迹，文献[36]定义了一种融合文本相似度的轨迹距离。文献[37]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。

4.4 轨迹知识提取

轨迹数据挖掘旨在从轨迹中发现有价值的知识和模型，已经成为数据挖掘领域的一个重要新兴分支[38]，被广泛使用在各类应用之中。现有的轨迹知识提取工作主要从基于轨迹的数据挖掘和语义归纳 2 个角度展开。

1) 频繁模式挖掘（frequent patterns）旨在从大规模轨迹中发现时序模式，例如超过一定数量的对象在给定时间间隔内行驶的公共路径，对目的地预测、路径推荐、行为理解有重要的价值。文献[39]将空间格栅化，根据 GPS 采样点密度将格栅组合成为区域，通过频繁项挖掘算法提取频繁模式；文献[40]提出了一种基于前缀树的频繁轨迹高效挖掘方法，避免过量的子轨迹组合验证；文献[41]定义了体现出行共性规律的频繁路径，并设计了一种高效的频繁路径搜索策略。在此基础上，文献[42]设计了一种基于离群检测机制的异常轨迹提取方法。文献[43, 44]通过轨迹学习实现有效的目的地预测，此外，文献[45]研究了面向大规模轨迹的周期性模式挖掘。

2) 伴行模式挖掘（moving together patterns）在轨迹数据中提取伴行的移动对象，用于事故调查、军事监控等应用。代表性的轨迹模式主要包括 Flock[46]、Convey[47]、Swarm[48]、Gathering[49,50]等。其中，Flock 模式挖掘[46, 47]旨在发现给定时间间隔内可被给定面积覆盖的一组移动对象；Convey 模式[46,47]则是根据密度来挖掘紧密伴行的移动对象，避免过于机械化的空间阈值限定；而 Swarm[48]是一种更为通用化的轨迹模式。文献[51]提出了面向轨迹数据流的伴行模式在线挖掘方法。

3) 轨迹聚类、分类（trajectory clustering and classification）。对移动轨迹的时空聚类可以帮助发现具有代表性和趋势性的移动模式。早期的轨迹聚类是以整条轨迹为对象。但是由于移动对象轨迹通常不完全重叠，文献[52]设计了一种基于分段轨迹、以豪斯多夫距离为度量的聚类方法，显著提升了聚类效果。文献[53]提出了增量式的轨迹聚类算法，降低轨迹聚类处理的时间和空间开销。而轨迹分类问题[31]则是根据行为、交通方式等特征来区别不同类型轨迹。此外，文献[54]提出了一种基于轨迹的高影响力位置挖掘算法。袁晶等[55]所研发了 T-Drive 系统从轨迹数据中学习出租车的运行规律和经验，为用户推荐更为通畅、便捷的路径以及出发时间的

合理推荐。

4) 轨迹摘要（trajectory summarization）是以文本的方式来概要一条轨迹所包含的信息，使轨迹数据更加易于理解。文献[54]提出了一种基于轨迹切分的摘要方法，使轨迹段内语义相似、轨迹段间语义不同，最后通过短文本总结蕴含在轨迹数据中的时空、交通、行为等各维度信息。

5 结束语

移动对象轨迹数据已经成为一种基本的数据资源，相关应用具有无限的潜力。因此，分析移动轨迹所蕴含的知识是认知用户、人群和城市行为的重要手段。轨迹数据与其他数据、特别是用户行为数据的叠加将产生更为巨大的商业和社会价值。企业级的轨迹数据已经积累到相当大规模，且增量数据以极快速度产生。因此，多用途、易扩展的轨迹数据管理系统已经成为上述应用的共性需求。

为了满足上述需求，人们对轨迹预处理、数据库、数据仓库和知识提取等一系列问题展开研究，取得了丰硕的成果。通过这些技术，轨迹数据能够被有效处理，从中提取的知识被应用于车辆导航、行程推荐、城市规划等中。但是随着轨迹数据规模的快速增长，各类位置服务的涌现，现有的轨迹处理技术在处理性能、分析能力等方面已经无法满足实际应用需求。

未来，轨迹数据管理需要重点突破以下关键技术。在数据库/数据仓库技术方面，现有的轨迹数据库主要面对集中式的轨迹管理，无法支持企业级大规模、高增量轨迹数据的高效处理，因此分布式环境下、可动态扩容的轨迹存储与计算框架将成为热点问题；在轨迹数据的理解与分析方面，为了充分发掘数据背后的巨大价值，如何集成轨迹与其他网络数据业务，从中准确地刻画用户行为并挖掘交通模式将受到更为广泛的关注。同时，轨迹数据管理作为一个跨学科研究问题，涉及地理信息系统、智能交通、数据库、数据挖掘等不同领域，需要这些不同领域的研究人员共同协作，最终实现轨迹数据完整生命周期的有效管理和价值发现。

参考文献：

[1] SCHUESSLER N, AXHAUSEN K. Processing raw data from global positioning systems without additional information[J]. Transportation Research Record: Journal of the Transportation Research Board, 2009(2105): 28-36.

[2] DOUGLAS D H, PEUCKER T K. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature[J]. Cartographica: The International Journal for Geographic Information and Geovisualization, 1973, 10(2): 112-122.

[3] MERATNIA N, ROLF A. Spatiotemporal compression techniques for moving point objects[A]. Advances in Database Technology-EDBT [C]. 2004. 765-782.

[4] ZHENG Y, et al. Mining interesting locations and travel sequences from GPS trajectories[A]. Proceedings of the 18th International Conference on World Wide Web[C]. 2009.

[5] PALMA A T, et al. A clustering-based approach for discovering interesting places in trajectories[A]. Proceedings of the 2008 ACM Symposium on Applied Computing[C]. 2008.

[6] GREENFELD J S, Matching GPS observations to locations on a digital map[A]. Transportation Research Board 81st Annual Meeting[C]. 2002.

[7] BRAKATSOULAS S, et al. On map-matching vehicle tracking data[A]. Proceedings of the 31st International Conference on Very Large Data Bases[C]. 2005.

[8] NEWSON P, KRUMM J. Hidden Markov map matching through noise and sparseness[A]. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems[C]. 2009.

[9] LIU K, et al. Effective map-matching on the most simplified road network[A]. Proceedings of the 20th International Conference on Advances in Geographic Information Systems[C]. 2012. ACM.

[10] LI S, et al. Quick geo-fencing using trajectory partitioning and boundary simplification[A]. Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems[C]. 2013.

[11] TANG Y, ZHU A D, XIAO X. An efficient algorithm for mapping vehicle trajectories onto road networks[A]. Proceedings of the 20th International Conference on Advances in Geographic Information Systems[C]. 2012.

[12] ZHENG K, et al. Reducing uncertainty of low-sampling-rate trajectories[A]. Data Engineering (ICDE), 2012 IEEE 28th International Conference on[C]. 2012.

[13] SU H, et al. Calibrating trajectory data for similarity-based analysis[A]. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data[C]. 2013.

[14] SU H, et al. Calibrating trajectory data for spatio-temporal similarity analysis[J]. The VLDB Journal, 2015, 24(1): 93-116.

[15] SISTLA A P, et al. Modeling and querying moving objects[A]. ICDE[C]. 1997.

[16] GÜTING R H, et al. A foundation for representing and querying moving objects[J]. ACM Transactions on Database Systems (TODS), 2000, 25(1): 1-42.

[17] WANG H, et al. SharkDB: An in-memory column-oriented trajectory

storage[A]. Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management[C]. 2014.

[18] KELLARIS PELEKIS G N, THEODORIDIS Y. Trajectory compression under network constraints[A]. Advances in Spatial and Temporal Databases[C]. 2009. 392-398.

[19] SONG R, et al. PRESS: A novel framework of trajectory compression in road networks[J]. Proceedings of the VLDB Endowment, 2014, 7(9): 661-672.

[20] CHAN W S, CHIN F, Approximation of polygonal curves with minimum number of line segments or minimum error[J]. International Journal of Computational Geometry & Applications, 1996, 6(1): 59-77.

[21] GUTTMAN A, R-trees: a dynamic index structure for spatial searching[A]. SIGMOD[C].1984.47-57.

[22] PFOSER D, JENSEN C S, THEODORIDIS Y. Novel approaches to the indexing of moving object trajectories[A]. Proceedings of VLDB[C]. 2000.

[23] NASCIMENTO M A, SILVA J R. Towards historical R-trees[A]. Proceedings of the 1998 ACM Symposium on Applied Computing[C]. 1998.

[24] YUFEI T, PAPADIAS D. MV3R-tree: a spatio-temporal access method for timestamp and interval queries[A].VLDB[C].2001.431-440.

[25] CHAKKA V P, EVERSPAUGH A C, PATEL J M. Indexing large trajectory data sets with SETI[A]. CIDR[C]. 2003.

[26] 丁治明. 一种适合于频繁位置更新的网络受限移动对象轨迹索引[J]. 计算机学报, 2012, 35(7): 1448-1461.
DING Z M. An index structure for frequently updated network-constrainted moving object trajectories[J]. Chinese Journal of Computers, 2012, 35(7): 1448-1461.

[27] VLACHOS M, KOLLIOS G, GUNOPULOS D. Discovering similar multidimensional trajectories[A]. Data Engineering, 2002 Proceedings 18th International Conference on[C]. 2002.

[28] CHEN L, NG R. On the marriage of lp-norms and edit distance[A]. Proceedings of the Thirtieth International Conference on Very Large Data Bases-Volume[C]. 2004.

[29] CHEN L, ÖZSU M T, ORIA V. Robust and fast similarity search for moving object trajectories[A]. Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data[C]. 2005. ACM.

[30] CHEN Z, et al. Searching trajectories by locations: an efficiency study[A]. Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data[C]. 2010.

[31] ZHENG Y, et al. Learning transportation mode from raw gps data for geographic applications on the web[A]. Proceedings of the 17th International Conference on World Wide Web[C]. 2008.

[32] YAN Z, et al. Semantic trajectories: Mobility data computation and annotation[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2013, 4(3): 49.

[33] ALVARES L O, et al. A model for enriching trajectories with semantic geographical information[A]. Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems[C]. 2007.

[34] ZHENG K, et al. Towards efficient search for activity trajectories[A]. Data Engineering (ICDE), 2013 IEEE 29th International Conference[C]. 2013.

[35] LIN B, SU J. One way distance: For shape based similarity search of moving object trajectories[J]. Geoinformatica, 2008, 12(2): 117-142.

[36] ZHENG B, et al. Approximate keyword search in semantic trajectory database[A]. Data Engineering (ICDE), 2015 IEEE 31st International Conference[C]. 2015.

[37] MA C, et al. KSQ: Top-*k* similarity query on uncertain trajectories[J]. Knowledge and Data Engineering, IEEE Transactions 2013, 25(9): 2049-2062.

[38] ZHENG Y. Trajectory data mining: an overview[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 6(3): 29.

[39] GIANNOTTI F. et al. Trajectory pattern mining[A]. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. 2007.

[40] WANG Y, ZHENG Y, XUE Y. Travel time estimation of a path using sparse trajectories[A]. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. 2014.

[41] CHEN Z, SHEN H T, ZHOU X. Discovering popular routes from trajectories[A]. Data Engineering (ICDE), 2011 IEEE 27th International Conference[C]. 2011.

[42] LEE J G, HAN J, LI X. Trajectory outlier detection: A partition-and-detect framework[A]. Data Engineering, ICDE 2008 IEEE 24th International Conference[C]. 2008.

[43] KRUMM J, HORVITZ E. Predestination: Inferring destinations from partial trajectories[A]. UbiComp 2006: Ubiquitous Computing[C]. 2006.243-260.

[44] LIAO L, et al. Learning and inferring transportation routines[J]. Artificial Intelligence, 2007, 171(5): 311-331.

[45] CAO H, MAMOULIS N, CHEUNG D W. Discovery of periodic patterns in spatiotemporal sequences[J]. Knowledge and Data Engineering, IEEE Transactions on, 2007, 19(4): 453-467.

[46] GUDMUNDSSON J, KREVELD M V. Computing longest duration flocks in trajectory data[A]. Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems[C]. 2006. ACM.

[47] JEUNG H, et al. Discovery of convoys in trajectory databases[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 1068-1080.

[48] LI Z, et al. Swarm: Mining relaxed temporal moving object clusters[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 723-734.

[49] ZHENG K, ZHENG Y, et al. On discovery of gathering patterns from trajectories[A]. ICDE[C]. 2013. 242-253

[50] ZHENG K, ZHENG Y, et al. Online discovery of gathering patterns over trajectories[J]. IEEE Trans Knowl Data Eng, 2004 26(8): 1974-1988.

[51] TANG L A, et al. On discovery of traveling companions from streaming trajectories[A]. Data Engineering (ICDE), IEEE 28th International Conference[C]. 2012.

[52] LEE J G, HAN J, WHANG K Y. Trajectory clustering: a partition-and-group framework[A]. Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data[C]. 2007.

[53] LI Z, et al. Incremental clustering for trajectories[A]. Database Systems for Advanced Applications[C]. 2010.

[54] CAO X, CONG G, JENSEN C S. Mining significant semantic locations from GPS data[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2): 1009-1020.

[55] YUAN J, et al. T-drive: driving directions based on taxi trajectories[A]. Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems[C]. 2010.

[56] SU H, et al. Making sense of trajectory data: a partition-and-summarization approach[A]. Data Engineering (ICDE), IEEE 31st International Conference[C]. 2015.

作者简介：

许佳捷（1983-），男，北京人，博士，苏州大学副教授、硕士生导师，主要研究方向为大数据管理、时空数据库、海量数据存储、分布式计算、工作流系统。

郑凯（1983-），男，山东淄博人，博士，苏州大学特聘教授、博士生导师，主要研究方向为大数据管理、社交媒体数据分析、时空数据库、不确定数据库、内存数据库、数据挖掘等。

池明旻（1977-），女，福建三明人，博士，复旦大学副教授、硕士生导师，主要研究方向为数据科学、大数据、机器学习。

朱扬勇（1963-），男，浙江金华人，博士，复旦大学教授、博士生导师，主要研究方向为数据科学、大数据、数据挖掘。

禹晓辉（1977-），男，山东德州人，博士，山东大学教授、博士生导师，主要研究方向为大数据管理与分析，包括分布式流数据管理、时空数据挖掘、社会媒体数据分析等。

周晓方（1963-），男，江苏无锡人，博士，苏州大学特聘教授，主要研究方向为空间数据库、多媒体数据库、数据质量、高性能数据处理及网络信息系统，以及这些技术在生物信息学、地理信息系统、移动对象管理、水文信息系统、医疗卫生系统、Web 查询及视频数据检索等方面的应用。