网络出版时间：2016-11-24 13:41:15
网络出版地址：http://www.cnki.net/kcms/detail/11.2560.TP.20161124.1341.004.html

软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.005143]

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

轨迹大数据：数据处理关键技术研究综述*

高 强[1], 张凤荔[1], 王瑞锦[1,2], 周 帆[1]

[1](电子科技大学 信息与软件工程学院,四川 成都 610054)

[2](美国西北大学 电子工程与计算机科学系,芝加哥 伊利诺伊州 60208)

通讯作者: 张凤荔, E-mail: fzhang@uestc.edu.cn

摘 要: 大数据时代下移动互联网发展与移动终端的普及形成了海量移动对象轨迹数据.轨迹数据含有丰富的时空特征信息,通过轨迹数据处理技术可以挖掘人类活动规律与行为特征、城市车辆移动特征、大气环境变化规律等信息.海量的轨迹数据也潜在性地暴露移动对象行为特征、兴趣爱好和社会习惯等隐私信息,攻击者可以根据轨迹数据挖掘出移动对象的活动场景、位置等属性信息.另外,量子计算因其强大的存储和计算能力成为大数据挖掘重要的理论研究方向,用量子计算技术处理轨迹大数据可以使一些复杂的问题得到解决并实现更高的效率.本文对轨迹大数据中数据处理关键技术进行综述.首先,介绍轨迹数据概念和特征,并且总结了轨迹数据预处理方法包括噪声滤波、轨迹压缩等.其次,归纳轨迹索引与查询技术,以及轨迹数据挖掘已有的研究成果包括模式挖掘、轨迹分类等.总结了轨迹数据隐私保护技术基本原理和特点,介绍了轨迹大数据支撑技术如处理框架、数据可视化.本文也讨论了轨迹数据处理中应用量子计算的可能方式,并且介绍了目前轨迹数据处理中所使用的核心算法所对应的量子算法实现.最后,对轨迹数据处理面临的挑战与未来研究方向进行了总结与展望.

关键词: 轨迹大数据;轨迹数据挖掘;隐私保护;支撑技术;量子计算;

中图法分类号: TP311

中文引用格式: 高强,张凤荔,王瑞锦,周帆.轨迹大数据：数据处理关键技术研究综述.软件学报. http://www.jos.org.cn/1000-9825/ 0000.htm

英文引用格式: Gao Q, Zhang FL, Wang RJ, Zhou F. Trajectory Big Data: A Review of Key Technologies in Data Processing. Ruan Jian Xue Bao/Journal of Software, 2016 (in Chinese). http://www.jos.org.cn/1000-9825/0000.htm

Trajectory Big Data: A Review of Key Technologies in Data Processing

GAO Qiang[1], ZHANG Feng-Li[1], WANG Rui-Jin[1,2], ZHOU Fan[1]

[1](School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

[2](Department of Electrical Engineering and Computer Science, Northwestern University, Chicago 60208, USA)

Abstract: The development of mobile internet and the popularity of mobile terminals produce massive trajectory data of moving objects under the era of big data. Trajectory data has spatio-temporal characteristic and rich information. Using trajectory data processing techniques can mine the patterns of human activities and behaviors, the moving patterns of vehicles in the city and the changes of atmospheric environment. However, trajectory data also can be exploited to disclose moving objects' privacy information, e.g., behaviors, hobbies and social relationships. Thus, attackers can easily access moving objects' privacy information by digging into their trajectory data such as activities and check-in locations. Moreover, quantum computation is an important theoretical research direction to mine big

* 基金项目: 国家自然科学基金项目(61602097,61272527); 四川省科技厅计划项目(2015JY0178); 四川省科技支撑计划(2016GZ0065,2016GZ0063);中央高校基本科研业务费(ZYGX2014J051,ZYGX2011J066,ZYGX2015J072); 博士后基金(2015M572464)

Foundation item: National Science Foundation of China(61602097,61272527); Sichuan Provincial Science and Technology Department Project(2015JY0178); Sichuan Sicence-Technology Support Plan Program(2016GZ0065, 2016GZ0063); the Fundamental Research Funds for the Central Universities(ZYGX2014J051,ZYGX2011J066,ZYGX2015J072); China Postdoctoral Science Foundation(2015M572464)

收稿时间: 2016-06-19; 修改时间: 2016-08-18, 2016-10-14; 采用时间: 2016-11-05; jos 在线出版时间: 2016-11-24

data due to its scalable and powerful storage and computing capacity. Applying quantum computing approaches to handle trajectory big data could make some complex problem solvable and achieve higher efficiency. This paper reviews the key technologies of processing trajectory data. We first introduce the concept and characteristic of trajectory data and summarize the pre-processing methods, including noise filtering and data compression, etc. Then, we review the trajectory indexing and querying techniques, and the current achievements of mining trajectory data, such as pattern mining and trajectory classification. An overview of the basic theories and characteristics of privacy preserving with respect to trajectory data is followed. The supporting techniques of trajectory big data mining, such as processing framework and data visualization, are presented in detail. We also discuss the possible ways of applying quantum computation into trajectory data processing, as well as the implementing some core trajectory mining algorithms by quantum computation. Finally, we conclude this paper by summarizing the challenges of trajectory data processing and outlining promising future research directions.

Key words: trajectory big data; trajectory data mining; privacy protection; the supporting techniques; quantum computation

在移动互联网、卫星定位技术、LBS 技术高速发展的背景下,无时无刻不在产生轨迹数据,轨迹数据包括交通数据、人类移动数据、动物迁移轨迹数据和自然现象轨迹数据等.海量的轨迹数据具有很大的研究价值,一般来说,轨迹数据都具有时空序列性、异频采样性、数据质量较差的特征[1].通过对轨迹数据的分析,可以挖掘人类活动和迁移规律,分析车辆、大气环境等的移动特征[2].对于轨迹数据的分析与挖掘是研究人员重点研究领域之一.例如基于城市交通的轨迹数据处理能够为优化交通路线[3]、个性化推荐路线[4]、路网预测[5]、城市规划[6]等提供很好的解决方案.

另一方面,海量的轨迹数据潜在性地暴露了个人的行为特征、兴趣爱好和社会关系等隐私信息.例如,大量基于位置服务的轨迹数据存在着相互关联的轨迹行为,这些数据的关联性与时空特征使得攻击者更容易挖掘用户的隐私信息[7],[8].连续变化的时空轨迹数据将更容易推断出用户行为模式与习惯,同时攻击者根据用户的历史移动轨迹数据特征,可以挖掘出其活动范围和活动场景[9],所以对于轨迹数据的隐私保护也是目前重点研究内容.

大数据的迅猛发展和海量轨迹数据的出现,对于计算性能和存储性能都提出了更高的要求.量子系统的独特特征,使其具有经典计算不具有的量子超并行计算的能力,同时使其具有强大的信息存储能力.虽然量子计算与量子存储硬件设备研究还是处于起步阶段,但是对于量子计算与量子存储的理论研究已经非常活跃[10].对于轨迹数据的挖掘过程使用的挖掘算法可以用量子算法进行代替,很好地实现大数据与量子计算的具体应用场景结合.基于轨迹数据特征设计或改进量子算法,有助于促进量子计算领域的发展.

本文关于轨迹数据处理关键技术重点介绍了轨迹数据挖掘、轨迹数据隐私保护、轨迹大数据支撑技术,以及轨迹数据挖掘核心算法与量子算法四个方面内容,目前许多研究人员针对轨迹数据处理技术进行了相应的总结与综述性阐述.许佳捷[1]等人从企业数据、企业应用和前沿技术三个角度总结了目前轨迹数据处理的现状,但是对于基于轨迹生命周期的轨迹数据处理方法相对缺少基本原理和基本概念的描述,没有涉及到轨迹数据中轨迹隐私保护方面.郑宇[2]系统地阐述了轨迹数据挖掘中相关问题包括预处理、轨迹数据管理、隐私保护、轨迹挖掘、轨迹分类和轨迹异常检测等方面,帮助研究人员较为全面地了解轨迹数据处理相关领域知识与方法,但是对于轨迹数据隐私保护方面介绍偏少,没有进行相应处理方法归纳与总结.特别在大数据时代,海量的轨迹数据和异构特性需要采用新的思路与方法来提高轨迹数据处理能力,目前针对轨迹数据处理综述特别是轨迹挖掘方面的综述侧重于传统轨迹数据方法介绍,对轨迹大数据支撑技术没有相应的自底向上的总结.同时目前对于轨迹数据处理中某一特定领域研究的综述也很多,如 Elgendy[11]主要阐述了结合语义特征对轨迹数据进行挖掘的方法,文献[7][12]则主要关注轨迹数据隐私保护从基于位置服务的隐私保护和轨迹数据发布两个方面进行归纳与总结,分析了目前所采取的隐私保护主要方法.王书浩[10]等人分类总结了大数据与量子计算相关算法研究,并分析了未来量子计算在加速经典算法效率研究的发展趋势.本文通过整理与归纳目前关于轨迹数据挖掘、轨迹数据隐私保护方面的综述性文章与最近研究文献,介绍了轨迹大数据处理相关支撑技术,并且结合目前轨迹数据挖掘核心算法归纳其对应的通用量子算法方案,为研究人员从轨迹数据挖掘、轨迹数据隐私保护,以及轨迹数据挖掘核心算法与量子算法三个方面的研究提供方法总结和研究思路,并对这三个方面未来可能的研究方向进行探讨.

本文第 1 节对轨迹数据进行概要性描述,概述轨迹数据特征与轨迹数据处理过程.第 2 节对已有的轨迹数据预处理方法进行总结.第 3 节对针对轨迹数据查询类型和索引结构进行总结与对比.第 4 节对已有的轨迹数据模式挖掘进行总结.第 5 节对于轨迹数据隐私保护从基于位置服务的轨迹数据隐私保护和基于轨迹数据发布的隐私保护方法进行总结.第 6 节针对大数据环境下的轨迹数据处理介绍了轨迹大数据存储技术、轨迹大数据处理新技术和可视化技术.第 7 节对轨迹数据挖掘中采用的经典算法相对应的量子算法进行总结,本文最后总结全文和展望未来轨迹数据处理需要解决的问题和热点研究方向.

1 轨迹数据概述

轨迹大数据作为大数据的一种,其丰富的数据来源与多样化结构符合大数据的"3V"特征,即量大(volume)、实时(velocity)、多样(variety).轨迹数据作为轨迹大数据处理的对象,需要在充分了解其来源、特征与处理技术架构的情况下挖掘其中有价值的信息.本节介绍了轨迹数据概念与分类、轨迹数据特征和轨迹数据关键技术架构,从整体上了解轨迹数据处理架构与方法.

1.1 轨迹数据

轨迹数据是具有时空特征的,通过对一个或多个移动对象运动过程的采样所形成的数据信息,一般包括采样点位置信息、采样时间信息、速度等.轨迹数据来源多样并且复杂,可以通过 GPS 定位器、手机服务、通信基站、信用卡、公交卡等,也包括射频识别、图像识别技术、卫星遥感和社交媒体数据等不同方式获取.轨迹数据一般包括人类活动轨迹、交通工具活动轨迹、动物活动轨迹和自然规律活动轨迹[2].

(1) 人类活动轨迹数据包括主动式记录数据和被动式记录.主动式记录数据是人们在基于 GPS 定位技术的主动分享位置信息,同时通过社交网络的照片分享、邮件来往等一系列活动轨迹也是属于轨迹数据的一类;被动式记录是用户在无意间开启基于基站定位服务而暴露了具有时空特征的轨迹数据,同样用户的信用卡消费行为、公交刷卡记录等也可以汇聚成具有时空特征的轨迹数据.

(2) 交通工具活动轨迹是在城市环境中基于车载 GPS 技术的移动轨迹数据.例如出租车、公交车上的车载 GPS 记录了在城市的活动范围轨迹.

(3) 动物活动轨迹数据是为了研究动物迁徙特征、行为特征和生活习惯而通过传感器获取动物的活动轨迹数据.

(4) 自然现象活动轨迹数据是很多研究研究人员关注的研究领域之一,通过收集的台风活动轨迹、海洋事件等来探索自然现象活动规律.

1.2 轨迹数据特征

轨迹数据符合大数据时代的 3V 特征即量大、实时、多样.轨迹数据采样由于受设备、采样、采样频率、存储方式等因素影响,其具有如下特征:

(1) 时空序列性.轨迹数据是具有位置、时间信息的采样序列,轨迹点蕴含了对象的时空动态性,时空序列性是轨迹数据最基本特征.

(2) 异频采样性.由于活动轨迹的随机性、时间差异较大的特征,轨迹的采样间隔差异显著,例如导航服务的秒级或者分钟级的采样,社交媒体行为轨迹是以小时或者以天作为间隔的采样.差异性的轨迹增加了轨迹数据分析的难度.

(3) 数据质量差.由于连续性的运动轨迹被离散化表示,受到采样精度影响、位置的不确定与预处理方式影响,给基于轨迹数据的分析带来一定的困难.

1.3 轨迹数据处理关键技术

轨迹数据模型中,各层次之间紧密联系,原始轨迹数据存在很多数据冗余与噪音,需要通过数据清理(data cleaning)、轨迹压缩(trajectory compression)、轨迹分段(trajectory segmentation)等预处理方式转化为校准轨迹.校准轨迹数据需要通过数据库管理技术进行轨迹索引与检索,能够有效地存取.最后对处理后的轨迹数据进行

模式挖掘、隐私保护等操作获取有价值知识.其采用的关键技术如下图 1 所示.

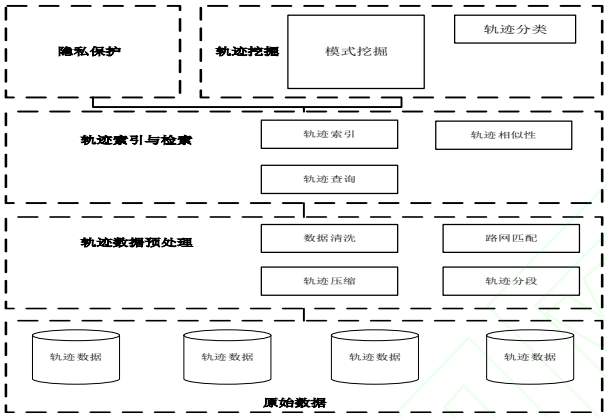


Fig.1 Trajectory big data processing key technology

图 1 轨迹大数据处理关键技术

2 轨迹数据预处理

轨迹数据作为轨迹大数据处理对象,其预处理效果将直接影响轨迹数据挖掘的效果,同时针对不同的应用场景与挖掘目标,采用不同的预处理方式具有重要作用.本节介绍了轨迹数据预处理中主要的方法,包括原始数据的数据清洗、轨迹压缩和针对应用场景与挖掘目标的轨迹分段与路网匹配预处理方法.特别地提出停留点检测技术,一般对于轨迹数据的挖掘研究很多都是围绕轨迹停留点展开研究,同时轨迹压缩技术在海量数据的分析中消减无用数据、提升数据分析能力具有重要作用.

2.1 数据清洗(data cleaning)

轨迹数据清洗主要是为了剔除数据中的冗余点和噪音点.移动对象在静止和匀速运行状态下都会产生大量的冗余点,其中移动对象在某一时间段驻留时间较长,称之为停留点(stay point),人们往往在数据分析时关心的是数据中停留区域,而不是整个轨迹数据,例如文献[13]中对旅游路线热点区域和路线进行推荐,挖掘其中热点区域和剔除冗余采样点.在基于速度的行为模式挖掘中往往关心的是基于速度划分的行为模式[14],文献[15]通过基于速度的出行模式划分,获取了用户行为模式与习惯.

噪音点是指由软硬件设备异常导致的错误采样,例如移动对象进入室内或其他干扰 GPS 接收信号而导致定位误差,噪音点会极大影响轨迹数据挖掘和分析的结果.本文定义轨迹数据清洗处理方法包括噪音滤波(noise filtering)、停留点检测(stay point detection)等.

2.1.1 噪音滤波(noise flitering)

受到传感器噪声、物体遮挡等因素而产生轨迹数据噪音,如图 2 所示,在轨迹中 p_5 点相对轨迹距离过大,如果此类噪音点不进行处理将会影响后续的轨迹数据分析.噪音点的过滤一般来说有三种处理方式:基于中值或均值滤波、基于卡尔曼滤波和基于粒子滤波方式.

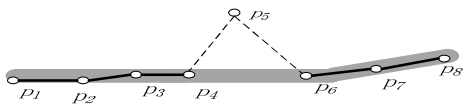


Fig.2 A noise point in a trajectory

图 2 轨迹中的噪音点

- **中值或均值滤波**. 对于一个测量点 z_i,处理单个噪音点真值的估计是 z_i 和其前 n-1 个轨迹点的均值或

中值[2].中值滤波处理过程与均值滤波类似,但是在处理极端错误值时,中值滤波要比均值滤波具有更强的鲁棒性,在密集轨迹中均值滤波或中值滤波能够很好地处理单个噪音点.但是在处理多个连续噪音点时,均值滤波或中值滤波并不能取得很好地效果.

- **卡尔曼滤波.** Lee 等人在文献[16]中详细地介绍了卡尔曼滤波在轨迹数据预处理中的应用,卡尔曼滤波包含测量模型与动态模型,卡尔曼滤波相对中值或均值滤波处理连续噪音点具有一定的优势,对于当前点的估计取决于先前的测量,它可以包含更多的状态向量例如速度.文献[17]中对于 GPS 的汽车防撞系统,利用扩展卡尔曼滤波处理不同情况下车辆行驶轨迹估计.但是卡尔曼滤波也存在一定的局限性,例如它的动态模型是线性的,对于移动对象受限于其预定义的路径,很难用卡尔曼滤波处理.
- **粒子滤波.**粒子滤波也同样有测量模型和动态模型,但是粒子滤波不需要卡尔曼滤波那样要求动态模型是线性的,文献[18]利用粒子滤波处理基于移动终端的位置估计,但是一般来说,粒子滤波对于噪声滤波处理效率不明显,文献[19]中详细地介绍了 Rao-Blackwellised 粒子滤波器在处理复杂系统中的优势,通过实验证明 Rao-Blackwellised 粒子滤波器在轨迹数据处理中的良好适应性,相对于卡尔曼滤波具有更好的性能优势.

2.1.2 停留点检测(stay point detection)

时空轨迹中存在的轨迹点其重要性并不相同,往往轨迹中某些点反映了人们一段时间的行为,例如购物、观光某个旅游景点等,在轨迹数据中,在某一时间区域或空间区域内产生某种行为的轨迹数据定义为停留点,一般将停留点分为轨迹中停留点和环绕轨迹停留点,如下图 3 所示的停留点 1 和停留点 2.文献[20][21]利用停留点检测技术探索城市领域中驾驶员的加油行为,挖掘兴趣点与加油事件,减少驾驶员等待时间和优化加油站布局.文献[22]结合停留点检测、频繁模式挖掘估计轨迹路径旅游时间.通过将轨迹数据转化为具有场景的停留点,更加有助于对轨迹数据的分析,文献[23]结合停留点检测技术利用出租车轨迹数据优化乘客等待时间,根据乘客到达时间和出租车的抵达与离开时间建立模型,提出了在历史出租车轨迹数据中给定时间和地点预测乘客等待时间,文献[24]结合停留点检测技术通过不同的轨迹策略,提升出租车运行效率,节约成本.在文献[3]中提出的 T-Drive 系统中也好地应用停留点检测技术.

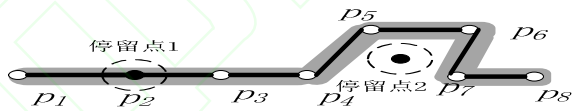


Fig.3 Stay points

图 3 停留点

2.2 轨迹压缩(trajectory compression)

对于移动对象基于时间戳的轨迹记录可以采用秒级记录,但是由于存储设备、计算能力等限制,轨迹数据挖掘不需要如此精细的位置定位,通常需要采用轨迹压缩的方法来处理轨迹数据.

2.2.1 距离测量

- **垂直欧式距离(perpendicular Euclidean distance).**文献[2][25]都提出使用垂直欧式距离进行轨迹压缩,如下图 4 所示,采用垂直欧式距离,将$\{p_0,p_1,..,p_{16}\}$轨迹压缩成三个轨迹点,每一个轨迹点 p_i 都有一个时间戳 t_i.

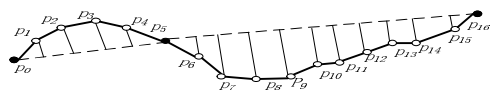


Fig.4 perpendicular Euclidean distance

图 4 垂直欧式距离

- **时间同步欧式距离(time synchronized Euclidian distance).**是一种从轨迹数据约简算法中产生近似轨迹的新的误差方法,其基于时间同步的轨迹计算方式来产生近似轨迹.假设一个初始轨迹有 n 个采样点,当然可以将其视为有 n-1 个分段.如下图 5 所示,A,B 和 C 是三个连续的时空位置点,根据其前一个位置点 A 和后一个位置点 C 来计算 B 的时间同步欧式距离[26].定义如下:

  $sed(A,B,C)=\sqrt{(x_B^{'}-x_B)^2+(y_B^{'}-y_B)^2}$,其中 $x_B^{'}=x_A+v_{AC}^{x}\times(t_B-t_A)$, $y_B^{'}=y_A+v_{AC}^{y}\times(t_B-t_A)$, v_{AC}^{x} 和 v_{AC}^{y} 表示速度向量.其中 v_{AC}^{x} 和 v_{AC}^{y} :

  $$v_{AC}^{x}=\frac{x_C-x_A}{t_C-t_A}\quad v_{AC}^{y}=\frac{y_C-y_A}{t_C-t_A}\ 。$$

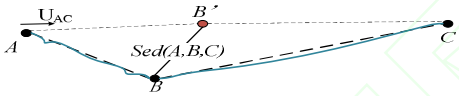


Fig.5 time synchronized Euclidian distance

图 5 时间同步欧式距离

2.2.2 离线压缩(offline compression)

对于离线压缩方法,给定一个轨迹,轨迹由一个一系列完整的基于时间戳的点,通过批量压缩算法丢弃一些可以忽略的轨迹点来获得近似轨迹,著名的算法有 Douglas-Peucker 算法.文献[27]基于 Douglas-Peucker 模型建立轨迹点批处理模式,但是在很多真实轨迹数据压缩处理中,却很难实现批处理模式.

2.2.3 其他

郑宇在文献[2]提到在线数据约简(online data reduction),方法包括基于滑动窗口算法和开放窗口算法,同时归纳了语义压缩(compression with semantic meaning)算法,语义压缩需要结合具体的应用场景保留关键位置点,例如在基于位置的社交网络中,用户可能在某些位置点拍照、改变轨迹方向等[28].关于语义压缩文献[29]也进行了深入研究.

关于轨迹压缩采用的方法很多[30]-[32],根据不同的场景需求与处理目标,针对性地建立轨迹压缩方法,例如 Song[33]等人提出的基于路网的轨迹压缩算法.

2.3 轨迹分段(trajectory segment)

轨迹分段是对长时段轨迹(例如以天或月为单位)的合理切分与标注,切分后的子轨迹段代表一次出行的记录,轨迹分段不仅降低了计算复杂度,同时提供了更加丰富的知识,例如子轨迹模式挖掘,轨迹分段的核心是理解时空移动特征,轨迹分段的主要方式有基于时间阈值、几何拓扑和轨迹语义三种基本策略[1].文献[13]和[34]利用旅游者的经验和地点兴趣度推荐热门旅游景点,利用历史数据学习获得停留点,将原始轨迹数据分段,大大提高了挖掘效率.文献[35]融入频率权重来优化兴趣点挖掘模型,结合轨迹分段思想挖掘兴趣点区域.

2.4 路网匹配(map matching)

路网匹配[36]是结合轨迹与数字地图,将 GPS 坐标下的采样序列转换成路网坐标序列.经过路网匹配后,每个轨迹采样点都映射到一个路网位置.路网匹配对于评估交通流量、引导车辆导航、预测车辆行驶路线、从出发地到目的地之间寻找频繁旅游路线等具有重要的作用,但是路网匹配是较为复杂的问题.路网匹配主要有在线和离线两种方式,在线匹配更强调实时性,当每一个 GPS 位置信息被接收时算法需要在路网中进行识别匹配,该方法主要应用于车辆导航、追踪等,而离线匹配主要针对静态历史数据匹配,其精度要求更高,只有整个行程数据匹配完成才提供匹配结果,并且尽可能准确地匹配车辆行驶路线.对于离线路网匹配问题,文献[37]利用 Dijkstra 最短路径算法来确定轨迹与地图序列距离从而实现轨迹匹配,对于在线路网匹配问题,文献[38]提出了使用坐标系观测用户位置的拓扑分析方法,该方法假定没有预期行驶路线和任何速度信息.文献[39]提出基于模糊逻辑理论的路网匹配算法,针对不同密度和不同复杂程度的路网进行匹配表明该算法精度较高.关于路网匹配研究也是目前研究人员研究的热点[40]-[43].

3 轨迹索引与检索

轨迹数据索引与检索对于轨迹数据库管理具有重要作用,不仅在传统关系型数据库包括目前主流的分布式存储系统都需要建立高效的索引结构和查询策略提升存储和查询效率,同时不同模式的查询需要对应的查询算法与相似性轨迹测量技术,本节阐述分为轨迹查询类型和轨迹索引两方面,以表格形式列出目前采取的主要轨迹距离测量方法和轨迹索引构建模型.

3.1 轨迹查询类型

对于轨迹数据的分析都依赖于大量的查询操作,根据用户定义的输出条件如轨迹长度、采样频率等,返回用户所关心的轨迹数据,典型的轨迹查询请求信息包括轨迹点信息查询(P-Query)、轨迹区域信息查询(R-Query)和轨迹查询(T-Query)[44].

- **P-Query**.轨迹点查询即请求获取满足预期时空关系的轨迹或轨迹段中的兴趣点(POI)信息.一般来说,点查询是获取在轨迹段中每个点的邻近点,例如文献[45]利用 P-Query 获取轨迹段中轨迹点 a 到轨迹点 b 的加油站信息.典型的查询算法有 NN(Nearest Neighbor)查询算法[46],点查询也可以获取轨迹段中满足预期条件的所有兴趣点信息.除此之外,当给出特定轨迹点,点查询可以请求获取轨迹中的段信息[47].
- **R-Query**.区域查询是属于特定轨迹区域内查询轨迹段,此类查询在交通分析、LBS 服务中应用广泛,在轨迹聚类中 R-Query 也尤为重要[48].
- **T-Query**.对于轨迹分类或聚类处理中,T-Query 可以从轨迹数据中获取相似轨迹,相似轨迹测量基于轨迹点距离即轨迹之间的相似性通常是通过某种聚集轨迹点之间的距离来衡量的.常用的方法有 CPD(Closet-Pair Distance)[49]、SPD(Sum of Pairs Distance)[49]、DTWD(Dynamic Time Warping Distance)[50]、LCSS(Longest Common Subsequence)[34]、EDR(Edit Distance on Real Sequence)[45]和 ERP(Distance with Reak Penalty)[51]等,具体描述如下表 1 所示.

- **Table 1** Trajectory distance measurement
- **表 1** 轨迹距离测量

名称	描述
CPD	测量轨迹的最小距离.
SPD	测量所有对应点的距离之和.
DTWD	处理轨迹长度不一致具有更好地灵活性,为了让轨迹点对齐,允许某些点多次出现.
LCSS	为了减少噪音点影响,允许忽略一些轨迹点而不是重新排列,其中孤立点会被排除,对于存在噪音点的相似性计算更具有鲁棒性.
EDR	相对不受异常值或噪音值影响,更好地处理两个具有相似子轨迹的轨迹.
ERP	结合 Li-norm 和 ED(Edit Distance)方法,支持本地时间转换.它是标准度量,满足三角不等式.

3.2 轨迹索引

在移动数据库中,海量的数据无法通过遍历数据库查询方式满足性能需求,轨迹数据的高效检索依赖于数据索引,索引结构可以提供快速和有选择地存取移动数据库中的对象,一般来说空间数据库中常用的索引结构有 R-tree 及其扩展版本[52]、B-tree 及其扩展版本[53],[54]、基于 K-D 树的索引技术[55]、四叉树索引结构[56]等.下表 2 列举了相关树结构及其介绍.

- **Table 2** Indexing structure
- **表 2** 索引结构

分类	索引名称	描述

增强式 R-tree	R-tree[57]	最早由 Guttman 提出的,是 B-树在多维空间的扩展.R-tree 使用范围广,查询性能好,并且支持对高维空间数据对象的查询,R-tree 最初是存储多维对象的最小边界框(MMB,Minimum Bounding Box).
	3D R-tree[58]	轨迹数据库查询中,往往需要考虑时间因素,3D R-tree 对于空间数据加入了时间维,有存储空间消耗低、时间间隔查询效率高的优点.但是往往受 MMB 覆盖、重叠和节点形状影响,导致轨迹不完整、死角过大.
	STR-tree[53]	采用相对于R-tree不同的插入/拆分策略,STR-tree考虑空间邻近性来保证分段属于相同的轨迹.
	TB-tree[53]	严格地让叶子节点只包含属于同一轨迹的线段来最大限度保证轨迹的完整性,TB-tree 侧重于移动对象轨迹的保存,索引移动对象过去的轨迹数据.
多版本 R-tree	MR-tree[59],[60]	各个R-tree子树对应不同的时刻,而各个子树并不是孤立的,采取共享节点的方式节省空间开销.
	HR-tree[61]	采用共享节点方式,节省存储空间,较好地支持时间点查询,但是时间段查询的效率不高.
	HR+-tree[62]	为了有效避免通过不同的根节点多次访问 HR-tree 同一节点,HR+-tree 时空索引区分了非共享与共享节点.
	MV3R-tree[63]	主要思想是构建两棵树,其中一棵 MVR-tree 用于处理时间片查询,而另一棵 3DR-tree 用于处理时间间隔查询,这样可以有效处理时间片查询和时间间隔查询,但是需要更多地存储空间.
基于网格索引	SETI[64]	将空间维度分成不重叠分区,在每个分区中的轨迹段使用一个 R-tree 索引.
	MTSB-tree[65]	类似于 SETI 索引方式,将空间分割成单元格,对应每一个单元建立时间访问方法,不同于 SETI 的是其基于 TSB-tree[66]建立时间维度的索引,适用于轨迹在时间和空间紧密的查询处理.

4 轨迹数据挖掘

轨迹数据挖掘作为轨迹数据处理的重要组成部分,为了分析与获取有价值信息,需要通过融合多种方法挖掘经过预处理的数据,本节将从轨迹数据模式挖掘与轨迹数据分类两个方面进行归纳,其中轨迹聚类是众多模式挖掘快速提取有价值信息的基础,轨迹聚类在相似性测量、轨迹压缩等方面有很大的作用,同时往往在数据挖掘中需要多种技术的融合,例如轨迹聚类与轨迹分类的结合来标签未分类数据.

4.1 轨迹数据模式挖掘

4.1.1 伴随模式挖掘

伴随模式挖掘是在轨迹数据中提取伴随的移动对象,通过分析移动对象群体的行为特征和规律,可以实现时空环境中的事故调查、群体跟踪等.代表性的轨迹模式主要包括 Flock[67]、Convoy[68]、Swarm[69]、Gathering[70]等.

- **Flock.**早期研究群体模式中只考虑单一时刻移动对象的行为特征,要求在某时刻至少有 m 个移动对象在同一区域内并且移动方向是相同,但是并不适应实际应用.Gudmundsson[67]等人对于群体模式挖掘提出了新定义即 Flock 模型,其中 $flock(m,k,r)$表示在一定数量的移动对象在给定半径的圆形区域内持续地移动,m 表示群体移动对象的最小数量,k 表示移动对象持续运动的最短时间,r 表示移动对象所在圆形区域的最大半径.但是 Flock 模型预先定义了区域形状和群体大小,不能很好地适应实际.
- **Convoy.**为了解决在伴随模式挖掘中对于移动对象群体大小和形状的限制,Convoy 模型定义为获取基于密度聚类的任意形状的轨迹挖掘,避免预先定义的空间阈值限定.模型要求一定数量的移动对象在 k 个持续时间内密度相连.
- **Swarm.** Flock 和 Convoy 模型对于移动对象群体的定义上有很大的限制,要求移动对象在持续时间内同时移动,而 Swarm 是更加通用化的模型,模型定义在一定时间内移动对象在任意形状区域内一起移动而时间不要求连续.
- **Gathering.**Gathering 模式是轨迹中模拟各种群体性事件,如庆祝活动、游行、抗议等.通过有效的索引结构、基于位向量的快速模式检测算法等解决群体性模式挖掘.

- **Traveling companion.** Flock 和 Convoy 模型需要将整个轨迹数据导入进行模式挖掘,而对于轨迹数据流的分析并不适用,Traveling companion 利用 traveling buddy[71]数据结构只存储移动对象间的关系,实现了 Flock 和 Convoy 模型的在线分析功能,极大减少了计算量.

4.1.2 频繁模式挖掘

频繁模式挖掘是从大规模轨迹数据中发现频繁时序模式,例如挖掘公共性规律或公共性频繁路径等.一般来说轨迹数据中包含了位置、时间和语义信息,所以时空轨迹频繁模式挖掘与传统的频繁序列挖掘有一定的区别.频繁模式挖掘在旅游推荐[72]、生活模式挖掘[73]、地点预测[74]、用户相似性估计[75],[76]等方面有很多应用.一般来说,对于轨迹数据频繁模式挖掘分为：

- **基于简单分段轨迹挖掘方式**.这种模式贴近于对于个体移动对象行为分析,文献[77]利用 Douglas-Peucker 算法处理轨迹数据,通过简单线性方式分割轨迹,利用最小支持度挖掘频繁模式轨迹.但是一般来说轨迹数据的处理往往关心的是轨迹中的兴趣区域,而不是具体的轨迹采样点信息.
- **基于聚类的兴趣区域挖掘方式**.该方式在旅游推荐等方面应用广泛[78],首先通过不同轨迹聚类,从中挖掘兴趣区域.文献[72]在数据预处理阶段,将轨迹的采样点序列转化为兴趣区域序列,预设兴趣区域的分类,并且提出三种不同模式的频繁模式挖掘,通过实验证明了可以更精确识别轨迹模式.
- **基于路网匹配的频繁模式挖掘**.这种模式更加贴近于在城市交通中的应用,首先通过路网匹配算法将轨迹数据转化为具有道路网络信息的语义轨迹,结合路网轨迹进行频繁模式挖掘[22],[33].

对于频繁模式挖掘算法主要分为两种挖掘模式,一种是基于 Apriori 的频繁模式挖掘,另外一种是基于树结构的频繁模式挖掘.

- **基于 Apriori 的频繁模式挖掘**.主要针对经典的 Apriori 算法进行改进加入时序特征,并且根据不同应用场景进行相应的优化,例如 GSP[79]、PrefixSpan[80]等.
- **基于树结构的频繁模式挖掘**.此类算法主要包括基于 Suffix Tree[22],[33]、SubString Tree[77]的结构树模式挖掘.

4.1.3 周期模式挖掘

移动对象经常有一些周期性的活动行为,例如购物行为、动物的定期迁移行为等,通过挖掘此类轨迹可以预测移动对象未来的行为.一般将周期模式挖掘分为完全周期模式、部分周期模式、同步周期模式、异步周期模式等.

- **完全周期模式**.周期模式更加强调全局性,强调整个行为过程中呈现周期性,周期内每个时间点都影响整个周期[81].
- **部分周期模式.**周期模式相对于完全周期模式更加关注局部周期特征[82],在时间序列中部分轨迹点表现出某种行为或特征,但是在部分周期模式挖掘经常受到冗余模式和计算效率问题而影响挖掘效果,文献[83]提出了并行性部分周期模式挖掘算法,有效避免产生冗余周期模式,同时解决海量数据计算问题,提升处理速度.
- **同步周期模式**.周期模式按照周期间隔发生或者在周期间隔的整数倍出发生[82].
- **异步周期模式.**受到噪音数据的影响,使得周期模式发生变化称之为异步周期模式.文献[84]提出了异步周期模式概念,有效地解决了由于噪音数据影响对周期同步性带来的破坏.但是该算法不能挖掘多事件时间序列周期模式,即一个时间点上不能够同时发生多事件行为,而且算法只能挖掘最长子序列.文献[85]提出了更加广义的异步周期模式挖掘算法,实现多事件的时间序列的周期模式挖掘.
- **其他.**文献[86]提出惊异周期模式(Surprising Periodic Patterns),其主要基于信息增益理论的带有概率的周期模式,时空序列中所有时间的每次发生具有不同的权重信息.文献[87]提出了广义信息增益方法和空位惩罚处理部分周期模式挖掘,在时间序列和空位惩罚方面,这个新的指标可以无缝地适应不同频率的事件发生,三角不等式保留的广义信息增益能够设计一个线性算法来挖掘任何一个重要的模式序列组合.由于周期性行为的复杂性,存在时空噪声和离群点等,Li[88]等人提出了两阶段检测方

法,首先通过聚类算法检测少量频繁访问点,移动对象轨迹被转换成二进制时间序列,通过利用傅里叶变换和自相关方法处理每一个时间序列,每一个周期性点的值可以计算得出,其接下来证明了可以通过从历史数据中使用分层聚类算法分析周期性行为.

4.1.4 轨迹聚类

为了通过不同的移动对象获得代表性路径或公共倾向行为,需要聚合具有相似轨迹作为集群,一般的聚类方法是利用一个特征向量代表一条轨迹,通过他们之间的特征向量距离来确定其相似性.Yuan[3]等人研发了T-Drive 系统,通过从轨迹数据中学习出租车的运行规律和经验,为普通用户推荐更为便捷的路径与出发时间规划.

轨迹数据聚类关键在于根据时空数据特征,设计与定义不同轨迹数据之间的相似性度量.关于轨迹数据距离测量在 3.1 节已经列表给出.轨迹聚类在整个轨迹数据处理中具有很大的作用,例如在预处理中轨迹压缩.根据轨迹数据时间维度要求的严格程度和轨迹相似性测量将轨迹聚类处理模式分为:

- **基于时间维度的相似轨迹聚类.**挖掘整体轨迹相似性,同时需要满足在轨迹点的时间维度必须一一对应相似.主要处理方法有基于欧式距离的轨迹聚类和基于最小边界矩形(Minmum Boundary Rectangle)的轨迹聚类.1993 年,Agrawal[49]等人提出了基于欧式距离的轨迹相似性测量,文献[89][90]通过采取离散傅里叶变换、离散小波变换预处理基于欧式距离的轨迹相似性测量.文献[91]改进MBR 表示,提出 MBB(Minmum Boundary Box)平滑轨迹,更好地处理噪声影响.
- **基于轨迹相似性的聚类.**侧重于整个轨迹特征的相似性挖掘,降低时间维度要求,只要求轨迹记录点的时间顺序.一般利用 DTWD[50]方式处理此类聚类.
- **局部多子轨迹聚类.**侧重于整条轨迹中寻找多条子轨迹的相似性聚类而不是整条轨迹的相似性聚类.一般处理此类模式聚类采用基于 LCSS、EDR、ERP 测量方法.
- **局部单子轨迹聚类.**侧重于处理轨迹中的最大相似子轨迹,降低时间维度限制.Lee[48]等人提出先划分后聚合的框架,按照最小描述长度原则划为子轨迹,利用基于密度的聚类方法处理.除此之外此类模式处理还有基于密度的聚类方法(OPTICS)、DBSCAN 等[92].
- **轨迹点聚类.**侧重于轨迹点的相似性处理,进一步不依赖于时间变化的轨迹,即用一个时间点代表整个轨迹时间.一般处理方法有历史最近距离、Frechet 距离[93],[94].

轨迹聚类在整个轨迹数据处理中尤为重要,在轨迹压缩过程中通过轨迹聚类的方式减少轨迹数据挖掘的计算资源和存储资源的消耗,在无监督轨迹数据分类学习中,需要首先通过轨迹聚类提取特征,然后根据提取的特征进行数据分类.以上根据轨迹数据聚类的不同要求列出了主要的处理模式,文献[95]中对轨迹聚类算法和轨迹相似度测量进行了很好地总结,首先对数据聚类算法进行了分类与归纳,同时系统化地介绍了轨迹数据算法,从基于空间聚类、基于时间聚类、基于路网匹配聚类和语义轨迹聚类等方面进行介绍.轨迹数据聚类依赖于高效的原始数据存储索引结构和根据需要提供相应的查询算法,在本文第 3 节已经进行了阐述与归纳.轨迹聚类在轨迹数据隐私保护中也有交叉应用,如文献[96]将轨迹聚类应用于基于轨迹数据发布的隐私保护策略中,其首先分析了轨迹数据隐私保护(k,δ)-anoymity 模型的局限性即无法获取轨迹数据动态变化不确定性特征和难以适用于真实环境,其通过将轨迹数据转化为不确定区域进行聚类,很好地融合两种算法,提升隐私保护效果.

到目前为止,传统的基于轨迹点或轨迹相似性的研究已经比较完善,目前研究人员对于轨迹数据聚类从场景上来说主要从语义轨迹聚类和基于路网匹配聚类方面进行研究.

- **语义轨迹聚类.**语义轨迹通常在于人为地赋予其场景特征或行为模式特征,在轨迹数据处理过程中,人们关心的是轨迹数据代表的丰富语义信息如其所在场景、轨迹速度、在某个轨迹点的运行模式等,而轨迹数据本身不具备这样的信息特征,通过语义信息能够更好地适应于真实环境.如目前人们通过语义轨迹聚类挖掘相似性用户[97]、推荐用户下一个目的地[98]以及轨迹数据中热点区域识别[99].
- **基于轨迹聚类的路网匹配.**原始数据中的轨迹点映射在路网中会形成一个基于地图的用户行为轨迹

集合.在真实的生活环境中存在着三种轨迹形态类型,第一种是自由空间的轨迹如鸟的飞行活动轨迹;第二种是不确定环境的轨迹环境如公园或广场散步轨迹;第三种是基于空间约束的轨迹即路网轨迹.在城市轨迹数据处理中,往往轨迹聚类需要考虑复杂的路网信息.在本文 2.4 节介绍了路网匹配相关技术,文献[100]结合路网匹配和轨迹聚类算法提供了交通轨迹数据聚类和分类的框架即相似性测量、轨迹聚类、产生聚类后的代表性子轨迹、轨迹分类.文献[101]提出了 NEAT 聚类框架,考虑路网环境限制、路网邻近度和交通流情况进行轨迹聚类,方法性能优于一般的基于密度的轨迹聚类方法.

在大数据时代,如何快速高效地进行轨迹聚类是研究人员关注的重要问题,对于轨迹大数据处理中轨迹数据聚类采用的工具和算法模型将在第 6 节进行归纳总结,本小节不再赘述.

4.2 轨迹数据分类

轨迹数据分类主要目的是区分轨迹或段的不同状态,例如交通出行方式、人类活动等.一般来说,轨迹数据种类繁多,通过轨迹数据分类可以挖掘更多趋势性、价值性规律.一般轨迹分类主要分为三步骤:

(1) 使用分段方法将轨迹分割成段;

(2) 从段中或者采样点中提取特征;

(3) 通过建立模型划分段或者采样点;

常用方法有动态贝叶斯网络(Dynamic Bayesian Network,DBN)、隐马尔科夫模型(Hidden Markov Model,HMM)、条件随机场(Conditional Random Field,CRF)等.文献[102]首先通过基于主成分分析(Principal Component Analysis,PCA)估计多变量概率密度函数,使用高斯混合模型(Gaussian Mixture Model,GMM)处理轨迹数据,但是GMM无法获取实体之间的时序关系和次序,基于此种缺陷研究人员使用基于HMM算法处理轨迹数据,文献[103]提出了改进的 HMM 模型即 SD-HMM 模型在处理处理人类活动认知方面具有更好地适应性,降低了整体误差率.使用 HMM 处理轨迹数据的研究还有很多如文献 [104],[105].

文献[106]利用 DBN 提出两层认知模型,提出的模型相对于仅仅使用 DBN 模型算法在精确度方面显著.文献[107]提出了使用滑动窗口的人类动作轨迹分析算法,通过 DBN 作为分类器进行分类推理,相对于固定窗口的动作分割具有很好地适应性.

除此之外,文献[108]提出了层次马尔科夫模型来推导用户日常行为特征,使用 Rao-Blackwellised 粒子滤波器处理多层次模型.Abidine[109]等人针对挖掘人类活动行为比较了三种算法分类精确度即C-SVM、CRF 和 LDA 算法.通过比较发现 C-SVM 识别效率更高.

5 轨迹数据隐私保护

海量的轨迹数据出现特别是个人用户轨迹数据必然会带来隐私泄露的风险,目前对于轨迹数据隐私保护的策略有很多,在参考大量文献之后进行了相应的总结与分析,目前针对轨迹数据的隐私保护主要从两个方面考虑,第一是基于位置服务的轨迹隐私保护,此类研究主要集中在基于 LBSN(Location Based Service Network)环境下的隐私保护,在时间层面提出了更高的要求;第二是基于轨迹数据发布的隐私保护,此类数据保护侧重于数据发布之前的数据处理,相对于前者在时间层面要求相对宽松.

本节首先对轨迹数据的隐私与分类进行了总结,5.2 节主要阐述隐私保护的度量标准与技术分类,为隐私保护策略选取提供衡量标准.在 5.3 节与 5.4 节分别对基于位置服务的轨迹隐私保护和基于轨迹数据发布的隐私保护进行概括和总结,两方面的隐私保护技术主要集中于假数据法、泛化法和抑制法等进行阐述.

5.1 轨迹数据隐私概述

移动终端、全球定位系统(GPS、AGPS)的普及和移动互联网的发展促进了基于位置服务(Location Based Services,LBS)的发展,LBS 应用领域包括社交网络、兴趣点检索、交通检测等.对于移动用户来说,LBS 提供了有价值的服务,一般 LBS 分为快照 LBS 和持续 LBS[110].对于快照 LBS,用户通过提交当前位置信息获取服务信息,而对于持续 LBS 是在一定周期或按需获取基于位置的连续服务信息,在轨迹数据隐私保护中持续 LBS 保护

一般来说尤为重要,连续性的轨迹数据暴露将会被攻击者获取其行为习惯或者其日常工作场景.

同时发布轨迹数据给第三方或者公众进行数据分析,同样存在隐私泄露的风险.综合分析多种文献,给出一些定义.

定义 1(隐私). 隐私是指个人、机构等实体不愿意被外部知晓的信息,例如,个人的行为模式、兴趣爱好、健康状况等.

定义 2(个人隐私).个人隐私是数据所有者不愿意透露的敏感信息,如个人的收入水平、旅游计划、活动场景、生活习惯等.一般来说,用户不愿意透露的或者容易暴露用户行为特征的信息都可以称之为隐私.

一般来说,将时空轨迹隐私分为数据隐私、位置隐私和轨迹隐私三种,下面将给出对三种隐私进行描述和采取的保护办法.

5.1.1 数据隐私

很多机构或者组织经常需要公布微数据(microdata),微数据是指在使用匿名方法后能被发布的数据集.为了防止个人信息泄露,一些明显标识符被移除或者代替,例如姓名、号码等信息.但是通过利用其它个人信息数据集与特定信息数据集结合处理往往可以推导出隐藏信息.如下图 6 所示,通过链接投票注册信息数据集与医疗信息数据集相关性别信息、出生年月和邮政编码就可以挖掘出用户的医疗信息隐私.这样的链接信息称之为准标识别符(quasi-identifiers)[111],即属性的值加在一起可能会找出一个单独的记录.

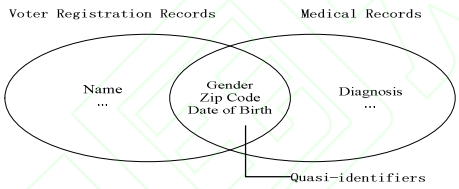


Fig.6 Two datasets are connected by using quasi-identifiers

图 6 使用准标识符连接两个数据集

因此数据隐私保护主要为了实现微数据的匿名化,目前研究人员提出了相应的数据隐私保护原则,用来限制匿名化微数据的泄露包括 *k*-匿名原则(k-anonymity)、*l*-多样性原则(l-diversity)和 *t*-相近性(t-closeness)原则.

- ***k*-anonymity.**L.Sweeney[112]等人提出 *k*-匿名(k-anonymity)方法,该方法主要通过概括和隐藏技术,发布精度较低的数据,使得每一条记录至少与数据表中其他 *k*-1 条记录具有完全相同的准标示符属性值.从而减少链接攻击所导致的隐私泄露.但是文献[113]发现除了链接攻击之外,发布数据表还存在同质攻击和背景知识攻击的隐私泄露风险.
- ***l*-diversity.**为了避免 *k*-匿名中存在的同质攻击和背景知识攻击,Machanavajjhala[113]等人提出 *l*-多样性原则,要求在每个等价类中至少包含有 *l* 个"well-represented"敏感值,而且给出了熵 *l*-多样性和递归(*c*,*l*)-多样性模型,有效地保证了单敏感属性数据集的隐私泄露问题.
- ***t*-closeness.**如果敏感属性在一个等价类中的分布和该敏感属性在整个表中的分布不超过某一个阈值 *t*,那么称该等价类是 *t*-相近性的,如果表中所有等价类都满足 *t*-相近性原则,则表示该表满足 *t*-相近性[114].

5.1.2 位置隐私

在请求基于位置服务中,移动用户向 LBS 服务商发出位置请求服务来获取物理定点位置,但是在持续性地请求位置服务将会暴露用户的移动轨迹和相关行为特征,为了在连续性请求服务中保护隐私,需要匿名化用户和对象数据集.基于位置的隐私保护主要有:

- **假位置(False Location).**用户发送一个关联真实地点的假地点信息[115],[116],这样在 LBS 服务提供商记录的是假地点而真实位置信息被隐藏,但是隐私保护程度和服务质量与假位置与真实位置的距离相

关.

- **地点转换(Space Transformation).**将地点信息转化为在查询和数据之间的空间关系已经被编码的其他空间[117].
- **空间匿名(Spatial Cloaking).**空间混淆主要是为了把用户位置信息隐藏在空间区域[118],用一个空间区域来表示用户的真实精确位置,区域形状不限.同时需要保证满足 k-匿名原则或者满足隐私保护要求.
- **时空匿名(Spatio-Temporal Cloaking).**在空间匿名基础上加入时间维度,在扩大位置区域的同时设置安全时间距离或者延迟相应时间[119].

5.1.3 轨迹隐私

一个空间轨迹是一个移动的路径或一个移动对象的一系列地理空间位置组成,一般一个轨迹 $T: p_1 \to p_2 ... \to p_n$ 包含位置信息和时间戳即 $p_i = (x_i, y_i, t_i)$.具有空间和时间属性的空间轨迹可以被认为是很强的准标识符,可以连接到各种其他类型的物理数据对象.

RID	Trajectory	Disease	...
1	(1,5,2) → (6,7,4) → (8,10,5) → (11,8,8)	HIV	...
2	(5,6,1) → (3,7,2) → (1,5,6) → (7,8,7) → (1,11,8) → (6,5,10)	Flu	...
3	(4,7,2) → (4,6,3) → (5,1,6) → (11,8,8) → (5,8,9)	Flu	...
4	(10,3,5) → (7,3,7) → (4,6,10)	HIV	...
5	(7,6,3) → (6,7,4) → (6,10,6) → (4,6,9)	Fever	...

Fig.7 Patient trajectory data

图 7 病人轨迹数据

例如上图 7 医学数据[12],医院将病人的轨迹数据公布给第三方研究机构,轨迹数据中不包含任何明显标识信息,但是却包含了敏感属性信息,如 HIV、Flu 等.图中 RID=1 的记录标识轨迹包含(1,5)、(6,7)、(8,10)和(11,8).假设一个病人 Alice 在时间戳分别为 2、5 去了位置(1,5)、(8,10),那么只有 RID=1 满足此轨迹,所以可以 100%确定其患了 HIV.所以即使处理过的轨迹数据,仍有可能暴露用户的隐私.在 LBS 中,用户的活动轨迹往往容易暴露其工作场景,例如其家庭地址等.因为即使其轨迹经过匿名化,但是可以通过其他人的轨迹推导出.

5.2 轨迹隐私保护度量标准与方法分类

5.2.1 轨迹隐私保护度量标准

(1)**隐私保护度.**通过轨迹隐私披露风险反映,披露风险越小,隐私保护度越高,披露风险是在一定情况下,轨迹数据隐私泄露的概率,披露风险依赖于攻击者掌握的背景知识,攻击者掌握的背景背景知识越多,隐私披露风险越大[7][7] .

(2)**数据质量/服务质量.**在轨迹数据发布中,经过隐私保护技术处理后发布数据的可用性越好,数据质量越好,一般采用信息扭曲度衡量数据质量的好坏,在基于位置服务中,一般用服务质量衡量,反映用户的 LBS 查询经隐私保护技术处理后获得服务结果的好坏,一般由查询响应时间和查询准确性来度量.在相同的隐私保护强度下,用户获得的服务质量越高,说明隐私保护技术越好[7],[8].

(3)**开销.**采用隐私保护技术所带来的代价,包括预处理和运行时存储、计算和传输代价.一般采用时间复杂度和通信协议的通信复杂度来度量.

5.2.2 轨迹隐私保护技术分类

综合基于位置服务的隐私保护和基于轨迹数据发布的隐私保护采取的技术,将轨迹隐私保护分为三类：

(1)**基于假数据的轨迹隐私保护.**通过添加假轨迹对原始数据进行干扰,同时保证被干扰的轨迹数据不发生严重失真.假数据方法简答、开销小,但是容易造成数据可用性低.假数据分为假位置点和假轨迹,假位置即不发布真实位置,用假位置获得相应服务.假轨迹是通过假轨迹降低敏感信息披露风险.一般生成假轨迹有两种方法.

- **随机生成法.** 随机生成一条连接起点与终点、连续运行且运行模式一致的假轨迹.
- **旋转模式生成法.**以真实轨迹为基础,利用真实轨迹中某些采样点为轴进行旋转,旋转后的轨迹为生成的假轨迹.

(2)**基于泛化法的轨迹隐私保护.**将轨迹上所有的采样点都泛化为对应的匿名区域,达到隐私保护的目的.可以保护数据真实性的同时却造成计算开销较大.

(3)**基于抑制法的轨迹隐私保护.**根据具体情况有条件的发布轨迹数据,不发布轨迹数据中的敏感位置或频繁访问位置.此类方法减少了开销,但是会导致某些信息丢失.

5.3 基于位置服务的轨迹隐私保护

用户在获取 LBS 位置服务时,需要提供自己的位置信息,虽然可以通过匿名化等方式处理位置信息,但是不一定能够保护移动对象的实时轨迹隐私.攻击者可以通过其他手段获得实时轨迹信息,例如利用位置 *k*-匿名模型对发出连续查询的用户进行位置隐私保护时,移动对象的匿名框(Cloaking Region)位置和大小产生连续性更新,那么将这些连续性位置信息连接起来就可以获得实时轨迹路线。

1.假数据法

- **Path Protection**[115][120]**.**利用假数据法思想处理在 LBS 中从起始地到目的地的查询,通过混合真假起始地和目的地,来获取所需要查询服务信息,一定程度上保护了轨迹隐私.
- **Dummy**[115]**.**综合考虑多项指标产生与真实用户移动模式相近的哑元位置,将哑元位置与真实查询一起发给 LBS 提供商,以达到保护真实位置的目的.但是往往哑元位置与真实位置特征差距较大,攻击者很容易区分出来,文献[121]提出 DUMMY-Q 方法考虑查询上下文与用户的运动模型将真实查询隐藏.

2.泛化法

- **KAT**[122]**.**结合用户的历史位置数据来增强用户的位置隐私,确保每个位置提供给服务商至少已经被 K 个不同的人访问过.
- **Path Confusion**[123]**.**利用期望距离误差量化攻击者估计用户精确位置的准确度,通过路径干扰获取隐私保护,定义应用程序对一组用户路径容忍的平均位置错误,每两个用户路径接近时,就干扰用户路径使其交叉以达到迷惑攻击者的跟踪.
- **Split-generalization**[124]**.**将二维空间划分为大小相等的正方形“格”,根据用户隐私需求将一个或者多个“格”定义成一个“划分”,在划分交界处的位置采用边界位置延时发布匿名区域策略处理,提升了数据服务质量.
- **其他.** 为了增强连续查询隐私保护力度与基于攻击者先验知识的隐私保护的有效性,需要尽可能减小隐形区域,同时考虑用户个性化隐私需求问题,文献[125]提出位置模糊模型(FBP),允许用户通过制定一个用户所在的公共区域来表示隐私需求,并将该公共区域作为用户的位置发给服务器,降低攻击者攻击风险.利用泛化技术处理位置服务的轨迹隐私有很多方法,如文献[126]提出了 GLON 模型.

3.抑制法

- **Location tracking**[127]**.**基于 LBS 特征和区域访问对象的多少将地图上的区域分成敏感区域和非敏感区域,当移动对象进入敏感区域时,将抑制或者推迟其位置更新,而对于非敏感区域,算法并不限制移动对象的位置更新.

4.用户身份替换

- **假名.**一般主要使用 *k*-匿名技术[112],使得某个位置的位置用户有 *k* 个,而且这 *k* 个用户无法用 ID 区分,即使而已攻击者得到某个用户位置信息也无法准确地辨别真正攻击对象.
- **混合区(Mix-Zone).**在移动过程中攻击基于位置查询服务的用户的前驱位置或者后续位置时,为了防止攻击,通过在某一攻击者无法监控的区域内秘密进行身份标识的更换,使得其在离开该区域时,攻击者无法将用户与掌握边界数据的用户相关联,以此阻断攻击者的位置攻击.所以 mix-zone [128]可以

看做是一个黑盒区域,虽然 minx-zone 能够抵抗以假名为关联的位置轨迹攻击,对于通过历史位置信息进行推断的攻击算法存在不足.后续人们也提出了相应的改进算法[129],[130].

5.空间加密技术

通过对位置加密达到匿名效果,一般来说其位置服务的效果与加密方法有关系.空间加密技术在确保服务可用性的情况下不会泄露任何用户的位置信息,实现了更加严格的隐私保护,根据文献[8]部分介绍,列出其中空间加密技术.

- **PIR**[8]**.**隐私信息检索协议(PIR)允许用户在服务器不知道其任何查询请求的情况下,从数据库中秘密地检索所需信息.按照隐私保护度的强弱分为基于计算能力的 PIR 协议和基于信息论的 PIR 协议.基于计算能力的 PIR 协议通过降低理论上不可解或计算上不可行难题的复杂度,来保证攻击者无法区分用户对不同数据项的访问.基于信息论的 PIR 协议是保证攻击者无法区分用户对不同数据项的访问,即使无论攻击者的计算能力多强,但是基于信息论的 PIR 协议存在开销太大的缺点.
- **HilCloak**[117]**.**在 K 最近邻居查询处理中,使用基于 Hilbert 曲线的位置匿名 HilCloak 将整个空间旋转一个角度,在旋转后的空间用密钥 H 建立 Hilbert 曲线,密钥只有用户和可信实体知道.用 SDK(Space Decryption Key)和 TDK(Textual Decryption Key)可信实体把每个兴趣点转化为 Hilbert 值并且上传至服务器,用户将基于位置服务的信息经防篡改终端生成密钥 H 并且上传服务器请求服务,服务器返回离 H 最近的 Hilbert 值.

5.4 基于轨迹数据发布隐私保护

基于轨迹数据发布隐私保护主要是轨迹数据发布前的数据处理保护其中的敏感信息,利用假数据、泛化法、抑制法、关联规则隐藏等方式变换或剔除敏感数据.轨迹数据发布之前的匿名处理主要遵循的是 k-anonymity 原则等,关于此类匿名处理在 5.1.1 节已有阐述.

1.假数据法

- **Dummy**[131]**.**对于轨迹数据发布前的敏感数据处理,可以通过添加假轨迹方式降低隐私泄露风险,同时也要保证被干扰的轨迹数据某些统计属性不发生严重的信息失真.在 5.2.2 节介绍了基于假数据隐私保护的两种方法,文献[131]详细描述了如何通过两种方式进行轨迹数据的隐私保护.

2.泛化法

- **k^m-匿名**[132]**.**该算法是基于 k-匿名的改进算法,用于降低数据发布中用户身份信息泄露风险,k 表示隐私保护强度,m 表示攻击者已经掌握用户之前访问过的 m 个位置点相关信息,通过基于距离的计算方式实现匿名,不需要在轨迹数据发布前了解准标识符属性详细信息,不需要区分敏感信息与非敏感信息.
- **NWA**[133]**.**用不确定轨迹数据在相同的时间周期内组合 k 个共定位轨迹(co-localized trajectories)来形成一个 k-匿名骨干轨迹.将同一时间区间内处于同区域的至少 k-1 条其他轨迹聚类成一个 k-匿名集,一个轨迹 k-匿名集合至少应包含 k 条处于同区域的轨迹,并转换为一条聚合轨迹,这条轨迹中每个位置节点都是该时刻所有轨迹节点的算术平均值.
- **Anonymity-reconstruction.**文献[119]提出 Anonymity-reconstruction 算法,将原始轨迹数据集泛化为一组 k-匿名轨迹集合,使得 k-匿名轨迹集合每条轨迹都是多个 k-匿名区域组成的有序序列,接着对于每条 k-匿名轨迹,在每个时刻所对应的匿名区域中通过均等地选择 k 个原子节点不重复地选取一个节点并连接起来,实现 k 条轨迹的重构.

3.抑制法

- **Data-suppression**[134]**.**根据攻击者掌握的移动对象部分轨迹情况,提出了抑制部分信息来保护用户隐私,该方法主要是将轨迹数据库 D 转换为 D^*,使得攻击者 A 不能以高于 P_{br} 的概率推导出轨迹上的位置属于某个移动对象.
- **Time-To-Confusion.**文献[135]提出了 time-to-confusion(TTC)度量的基于信息论对时序性位置隐私保

护,通过抑制某些节点,使得相邻节点间的不确定性增加从而增强保护能力.

4.关联规则隐藏

利用数据挖掘的方式进行敏感信息隐藏也是目前研究的热点[136], Ataiiah[137]等人提出了重要规则、敏感规则、数据清理等概念,他们认为问题的解决可以通过降低给定规则或规则集的重要性来实现,同时尽可能地不对其他规则的重要性造成影响.

6 轨迹大数据支撑技术

在大数据时代,轨迹数据面对多种数据源、数据质量低、数据量庞大、挖掘需求多样化等问题,需要采用新技术提高处理能力,在过去几年以 MapReduce 模型为代表的分布式并行处理架构成为处理海量数据的新手段.面对海量数据,轨迹数据从原始数据预处理、数据模式挖掘、数据隐私保护以分布式并行思想处理将更加快速高效地获取处理结果,轨迹数据处理经过多年的发展,其算法模型、存储结构、隐私保护策略模型等已经十分丰富,研究人员更加注重将轨迹数据当作对象处理,注重解决轨迹数据的运算效率与挖掘效果以及赋予其更多的场景特征.因此,按照轨迹大数据中轨迹数据的存储、处理到可视化层次进行支撑技术说明和归纳.并且列举了一些用作轨迹数据研究的公开数据资源.

6.1 轨迹大数据存储技术

目前成熟的轨迹数据存储主要依赖于关系型数据库例如 Oracle、PostgreSQL 等,但是面对海量轨迹数据管理、高并发读写和扩展性等方面存在一定的局限性,目前作为大数据支撑技术即云计算技术对于大数据存储和管理提供了很好地解决方案,传统的轨迹数据存储主要存储的是静态数据,大数据时代数据的异构性和动态性都对数据的存储格式和存储方法提出了很大的挑战,以 NoSQL 技术作为支撑的一系列数据库的出现为解决海量数据的存储提供了新模式[138].目前一些传统的关系型数据库已经开始支持 NoSQL 技术,例如 PostgreSQL.当然轨迹数据的存储依赖于高效的索引结构,关于轨迹数据索引结构介绍在第 3 节已经进行了总结,关于大数据环境下的轨迹数据存储本节主要介绍几种常用的非关系型数据库和相关文献.

- **MongoDB.**作为分布式文件存储数据库,其具有高性能、易部署、高并发读写等特点.MongoDB 不仅扩展了传统关系型数据库功能,同时支持 MapReduce 模型和地理空间索引,可以很好地应用在轨迹数据存储方面,例如文献[139]利用 MongoDB 存储轨迹数据和索引结构,文献介绍了 MongoDB 中如何对 R-tree 结构命名、定义存储内容和存储结构.再如文献[140]针对危险品运输路径规划处理问题采用 MongoDB 存储危险品事件和运输轨迹数据.
- **HBase.**作为一个分布式、面向列的开源数据库系统,由于其开源性在海量数据存储中使用广泛,并且 HBase 是 Hadoop 的一个子项目.因此 HBase 的使用与使用 MapReduce 模型进行数据处理有很大关系,文献[141]针对轨迹数据特点提出了基于 HBase 的可扩展空间数据存储模型 HBaseSpatial,并且设计分布式存储和索引模型,其架构如图 8 所示,并通过与 MySQL、MongoDB 数据存储进行对比,随着数据量的增大,其优势更加明显.
- **其他.**目前针对海量数据存储开源化的分布式存储系统越来越受到欢迎,HBase 和 MongoDB 都是开源化存储系统,再如 Cassandra 数据库系统作为混合型的非关系型数据库,也能够很好地支持海量轨迹数据的存储需求,在轨迹数据存储管理方面,最重要的是对轨迹数据建立高效的索引结构和查询算法,如何能够快速建立轨迹数据索引结构,如何能够快速地查询目标轨迹区域是评价数据存储管理模型的重要参考条件,以上介绍的是基于分布式磁盘存储的数据库系统,目前针对多核处理器的分布式内存计算也一定程度上解决轨迹数据的高性能查询需求.文献[142]提出了针对轨迹数据特征的基于列的内存数据库存储模型,分析了传统关系型数据库对轨迹数据存储存在的缺陷,将数据库划分为帧的形式,把同一时刻的所有移动对象位置存储在一起,同时将内存数据对齐.很好地提高了内存吞吐量,减轻对 CPU 高速缓存压力.

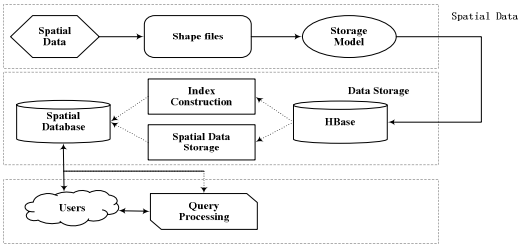


Fig.8 Architecture of Vector Spatial Data Storage

图 8 向量空间数据存储的体系结构

6.2 轨迹数据处理新技术

6.2.1 基于 MapReduce 模型的 Hadoop 分布式处理架构

MapReduce 在处理大规模数据集的具有很大的优势,MapRuduce 通过分割数据到每个节点,并且通过周期性地返回各个节点结果来更新全局状态,实现分布式并行计算和提高数据处理能力.Hadoop 分布式系统基础架构促进了传统轨迹数据处理向大数据平台处理的跨越.目前基于 Hadoop 平台下的轨迹数据处理获得了很大的发展,尤其在基于轨迹数据的地点推荐和轨迹数据聚类方面有众多文献.

(1) 基于 Hadoop 的轨迹聚类

轨迹聚类在轨迹数据处理中十分重要,关于轨迹聚类相关介绍已经在 4.1.4 节给出,而通过基于 Hadoop 分布式架构改进传统聚类方法越来越受到关注,文献[143]通过 Hadoop 平台优化 *k*-means 算法处理海量出租车轨迹数据,其详细给出了 MapReduce 下的三阶段轨迹数据处理流程和获取目标结果的处理框架,利用平方误差准则评判聚类效果的质量,其算法效率和准确率方面都优于传统的 *k*-means 算法和两阶段并行算法[144].文献[145]通过改进传统 *k*-means 算法将轨迹聚类分成 Map 阶段、Shuffle 阶段和 Reduce 阶段,聚类过程结合改进的动态时间规整算法(Dynamic Time Warping,DTW)即 FCDTW 通过实验证明其算法性能随着数据量的增大优于传统聚类算法.文献[146]则通过改进基于密度的 DBSCAN 算法进行轨迹聚类,实现了基于 MapReduce 架构的四阶段聚类方法,即预处理、局部 DBSCAN、获取需要合并的集群、全局集群处理四个阶段.目前研究人员集中于改进传统聚类算法来更好地适用于在 MapReduce 架构下的轨迹聚类处理,同时轨迹聚类的效果也会受到轨迹数据预处理结果的影响.

(2) 基于 Hadoop 的模式挖掘

相对于传统的轨迹数据模式挖掘,结合 Hadoop 平台的模式挖掘将更加适合处理海量轨迹数据,提高数据处理能力,一般基于 MapReduce 模型的模式挖掘主要解决基于分布式架构的数据索引结构和传统算法的分布式解决策略,文献[147]提出基于 MapReduce 编程模型的频繁移动模式挖掘,为了解决海量轨迹数据下的网格固定性分辨率的局限性和资源消耗过大的问题,提出使用基于网格的四叉树搜索方法,利用 MapReduce 模型加速计算.文献[148]提出基于 Hadoop 平台从大量不确定序列数据中挖掘隐藏模式算法,文献首先分析了传统的 Probability Suffix Tree(PST)无法处理海量的不确定轨迹数据的局限性,提出了基于改进 PST 的 $uPST_{MR}$ 算法和 $uPST^{+}_{MR}$. 基于 Hadoop 平台构建 PST 的目的为了计算每个节点的条件概率分布同时需要检测其是否符合条件,为了避免挖掘不符合条件的模式和平衡分布式计算过载问题,文献构造一个渐进、多层次和迭代的 $uPST_{MR}$ 处理海量数据,减少了迭代次数和运算时间,在 $uPST_{MR}$ 基础上提出了 $uPST^{+}_{MR}$ 算法,其使用二维数据结构存储临时数据,有效地减少数据扫描次数,通过实验证明提出的算法具有很好的扩展性和稳定性.

(3) 其他

Hadoop 在轨迹数据其他方面也有一定应用,例如轨迹查询,文献[149]介绍了一种针对空间数据的

SpatialHadoop 架构,采用简单的空间高层次语言、双层空间索引结构、基于 MapReduce 层的基本空间组件和采用一个三层基础空间操作即范围查询、k-NN 查询和空间连接操作,文献也进行了相应的结果对比演示.

6.2.2 类 MapReduece 模型的其他分布式处理架构

- **Spark 平台.**美国加州博客里大学开发的类 Hadoop MapReduce 的通用并行计算框架,Spark 启用了内存分布数据集,能够很好地优化迭代工作负载,而轨迹数据对于迭代式计算能力要求更高.其能够更好地处理轨迹数据模式挖掘.
- **Storm 平台.**Twitter 发布的分布式、容错实时计算系统,更加适用于流式计算,提升了对实时性的计算要求.流式计算目前也越来越受到重视,在轨迹大数据处理领域,对于实时性要求高、无法事先存储轨迹数据的轨迹数据,采用流式计算框架很好地解决此类轨迹数据处理需求,如智能交通、LBSN 位置推荐等.

6.2.3 基于 GPU 的并行架构

NVIDIA 推出的 CUDA 并行计算架构推动人们在基于 GPU 上实现并行计算,目前研究人员也使用基于 GPU 的分布式架构进行轨迹数据处理,其在轨迹数据索引、模式挖掘等方面有诸多应用[150].

- **轨迹索引与查询.**文献[151]将 GPU 计算应用到轨迹的距离阈值相似度搜索,改进传统索引结构即空间索引、时间索引和时空索引从而适用于 GPU 计算,通过实验证明其计算速度超过传统 CPU 计算的 15.2 倍.文献[152]使用 R-tree 实现了基于 GPU 架构的并行空间查询处理.GPU 主要用来加速处理 R-tree 的批量加载和空间窗口查询处理,设计了适用于 GPU 上 R-tree 处理的数据分布模式,实验证明其处理能力超过基于 CPU 计算的 10 倍.文献[153]同样针对构建并行 R-tree,受到树的非线性拓扑结构和 GPU 的单指令多线程架构(single-instruction multiple-thread,SIMT)在实际工程应用中的局限性,对比目前存在的基于 GPU 的算法加速性能最快约 20 倍,文献通过构建一个非平凡的自上而下树结构和设计有效的空间数据结构其加速性能远远超过原始的加速能力,进一步提升了 GPU 实际应用能力.
- **轨迹聚类.**文献[154]结合 Fréchet 距离提出基于 GPU 的子轨迹聚类算法,在轨迹曲线之间利用连续 Fréchet 距离作为衡量相似度标准,该算法性能优势明显.文献[155]提出改进的 OPTICS 轨迹聚类算法 Tra-POPTICS,并且提出基于 GPU 计算的 G-Tra-POPTICS 算法,基于 STR-tree 轨迹索引结构实现 CPU 线程数据分配,每个 CPU 线程同时产生基于 Hyper-Q 任务队列,这样 Hyper-Q 能够执行 GPU 并行化操作,G-Tra-POPTICS 能够并行化 Tra-POPTICS 三个计算过程即计算本地最小生成树、形成全局最小生成树和从全局最小生成树中获取聚类结果.
- **模式挖掘.** 文献[156]中提出基于 GPU 挖掘热门区域,采用弱准则和强准则进行轨迹热门区域处理,同时在数据处理中允许受约束的输入和输出轨迹路径.文献[157]利用 GPU 挖掘轨迹数据 flock 模式,讨论了三种变异 flock 模式即 maximal flock、largest flock 和 longest flock.通过真实实验数据运行和对比目前的算法复杂度,其优势明显,即使 longest flock 是 NP 问题,但是在并行化计算下,算法处理速度十分优秀.

6.3 轨迹大数据可视化技术

轨迹数据描述移动对象的空间位置和时间属性,同时在语义轨迹中也存在移动轨迹点的相关属性信息,如图 9 轨迹数据可视化技术可以更加形象地展示轨迹数据挖掘效果,数据可视化对于快速了解数据结果之间的关系具有直观效果,目前很多学者开展轨迹数据可视化研究,轨迹数据种类众多,处理方法也不尽相同,海量的轨迹数据从产生到后期的处理结果,结合一定的可视化系统将会帮助研究人员更好地找出移动对象规律或行为特征信息.

在大数据时代,交互式可视化对于帮助研究人员快速理解数据特征和解释具有重要作用,交互式可视化更加注重在可视化系统中参与调整参数、数据排序、数据提取、模型选择、结果展示选择等方面,结合可视化系统,能够辅助研究人员从轨迹数据预处理、模式挖掘、结果表示等过程有更加清晰的了解与交互式参与.关于

轨迹数据可视化的研究有很多,本节根据轨迹数据的处理过程并结合文献[158]将轨迹数据可视化分为直接可视化、抽象可视化和特征提取可视化.

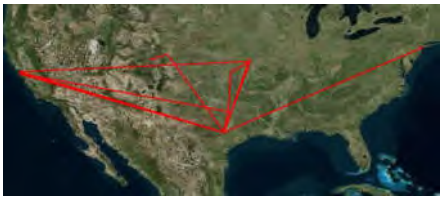


Fig.9 Trajectory visualization

图 9 轨迹可视化

6.3.1 直接可视化

直接可视化是最为基础的轨迹可视化方法即轨迹数据直接导入后绘制每条轨迹.对于 GPS 原始数据集来说,其有固定的数据格式,当数据量不是很大时,可以借助 Google 地图或 OziExplorer 实现原始数据的可视化功能,如上图 9 所示.文献[159]主要解决轨迹路径的可视化,以突出显示轨迹的空间位置信息.文献[160]实现了基于时间维度的二维轨迹可视化,突出轨迹数据更加丰富的时间维度信息.文献[161]提供了一种三维语义轨迹可视化框架,该框架可以查询轨迹数据和实现轨迹数据的可视化功能,同时该框架实现语义注释和自动视觉提示匹配,该可视化工具对于语义轨迹挖掘具有很大的用处.直接可视化可以直接表达原始轨迹数据特点,但是却不适用于大数据可视化或者难以实现对原始数据的可视化.

6.3.2 抽象可视化

先对轨迹数据进行聚集或聚类,保留重要轨迹数据,剔除冗余数据,根据处理结果后可视化显示.轨迹数据有时间维度属性、空间维度属性和描述轨迹数据移动对象的各种属性特征,根据处理目标需求不同,可以将概要可视化分为时间维度可视化、空间维度可视化等.

- **时间维度.**可视化目的是为了关注轨迹数据基于时间维度上的特征信息,文献[162]利用可视化技术关注基于时间维度的轨迹数据类别全局和局部特征,用户选择移动对象、地点并且关注其随着时间变化如何变化,但是其主要解决少量分类轨迹数据的可视化动态显示,对于大量分类数据并不适用,文献也针对数据量大提出了针对性的解决方案.文献[163]基于时间维度对历史气候变化数据进行了可视化展示与分析,利用全球经向地图标记气候变化状态,其提供的 Time Series Discs 可视化分析工具包括基本三角模块和旋转三角模块组成,基本三角模块从四个方面展示气候变化情况即行数据、列数据、块数据、斜边数据,旋转三角模块能更好地展示区域整体气候状态变化情况.
- **空间维度.**轨迹的空间属性一般指移动对象在空间中的位置信息和周边地理情况,将原始数据抽象化后联合轨迹数据空间分布特征形象化展示轨迹空间特征,同时根据空间划分策略,采用不同的图形表示方式辅助研究人员进行特征分析.文献[164]针对城市轨迹数据中的出租车轨迹进行可视化分析,建立出租车轨迹可查询模式,很好地处理大量轨迹数据的可视化处理,而涉及到大规模数据存储,文献分析了 SQLite 和 PostgreSQL 两种数据库在构建空间索引需要消耗大量时间的缺陷,提出了基于 K-D 树的空间分割索引结构,大幅度减少出租车数据索引构建时间消耗,并且结合 OpenGL 和 Qt 技术进行可视化界面渲染.
- **其他.**在轨迹抽象化表示时,根据特定的需求而可视化显示不同的结果,例如只关心移动轨迹的起始位置和目的位置而忽略中间过程[165],或者在路网匹配处理中,往往关心的抽象化轨迹路径而不是具体的轨迹点[166].

6.3.3 特征提取可视化

可视化应用最终的目的是辅助研究人员发现和分析经过挖掘后的数据特征结果,能够更加形象地展示研

究人员所关注的局部特征.按照特征提取可视化所关注的局部特征或全局特征分成局部特征可视化和轨迹模式分析可视化.

- **局部特征可视化.**注重基于轨迹段的事件特征分析,其中满足一定条件的时间段称之为事件,一般轨迹段包含时间和空间信息,研究人员首先提取符合特定条件的轨迹段构成事件数据集,再对基于轨迹段的事件数据进行可视化显示与分析,如文献[167]首先将出租车轨迹数据预处理后匹配到路网中,在面对多条路径选择问题上,根据始发地和目的地之间的路径历史数据分析出热门路径候选集,在根据相关参数如路径时间花费推荐最优路径,并且利用不同颜色在可视化系统中标识.
- **群体特征可视化.**相对于局部和个体特征数据分析,其主要从多目标多轨迹数据中提取行为特征,同时可以挖掘出多轨迹之间的相互关系,如文献[168]结合可视化系统从时间维度分析群体中个体行为特征、群体内的行为特征和群体之间的群体特征,可视化系统提供的时间折线图很好地展现群体性轨迹特征.

6.4 公开数据资源

在轨迹数据处理中,首先需要针对所要处理的轨迹数据集进行数据预处理、建模、分析与结果展示.因此不同的轨迹数据集所采用的方法是不相同的,同时不同的数据集也有不同的数据特征.在大数据环境下,原始数据的收集越来越受到重视,一般的原始轨迹数据量基本上达到 TB 级、甚至 PB 级.因此轨迹数据已经成为一种重要的数据资源.目前已经有不少公开轨迹数据集用于轨迹数据处理研究,郑宇在文献[2]中已经列举了七种公开数据集包括人类轨迹、出租车轨迹、签到轨迹等方面,本小节补充几条公开数据集,帮助研究人员根据不同需求快速获取所需要的公开数据集.

1.Gowalla 签到数据

Gowalla 数据集是一个基于 LBSN 网站的数据集(http://snap.stanford.edu/data/loc-gowalla.html),由社交网络数据和签到地点数据集组成.社交网络数据包括 196591 个节点和 950327 条边数据,签到地点数据收集了 196591 个用户从 2009 年 2 月到 2010 年 10 月期间 6442890 个签到点数据.其签到数据格式如下表 3:

- **Table 3** Gowalla Data
- **表 3** Gowalla 数据

用户 ID	时间戳	纬度	经度	签到点 ID
0	2010-10-19T23:55:27Z	30.2359091167	-97.7951395833	22847
0	2010-10-18T22:17:43Z	30.2691029532	-97.7493953705	420315
…	…	…	…	…

2.Brightkite 签到数据

Brightkite 签到数据是基于 LBSN 网站的公开数据集(http://snap.stanford.edu/data/loc-brightkite.html),由社交网络数据和签到数据组成.社交网络数据包括 58228 个节点和 214078 条边,签到地点数据收集了 58228 个用户从 2008 年 4 月到 2010 年 10 月共 4491143 个签到点数据,其签到数据格式如下表 4:

- **Table 4** Brightkite Data
- **表 4** Brightkite 数据

用户 ID	时间戳	纬度	经度	签到点编码
0	2010-10-17T01:48:53Z	39.747652	-104.99251	88c46bf20db295831bd2d1718ad7e6f5
0	2010-10-16T06:02:04Z	39.891383	-105.070814	7a0f88982aa015062b95e3b4843f9ca2
…	…	…	…	…

3. FOILing NYC's Taxi Trip Data

该数据集由两部分组成即 Trip Data 和 Fare Data(http://chriswhong.com/open-data/foil_nyc_taxi/).两部分数据文件下的单个文件数据均在 1.5GB-2.5GB 之间.Fare Data 存储的是出租车收费记录包括消费金额、付款方式、时间等信息.Trip Data 中每个文件有 1400 行数据,每一行存储了经纬度信息、上下车时间和乘坐距离等信

息.

4.Taxi Service Trajectory

kaggle 竞赛平台发布的一个出租车轨迹数据,该数据集包含 2013 年 7 月 1 日到 2014 年 6 月 30 日 442 个出租车在 Porto 城运行的轨迹数据(https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/data).该轨迹数据集有包括 GPS 位置信息、时间戳、出租车标签在内的 9 个特征描述.

7 轨迹计算与量子计算

轨迹数据处理很重要的处理方式包括轨迹分类、轨迹聚类、轨迹频繁集挖掘等,对于轨迹数据的挖掘算法理论研究不仅要考虑挖掘方式与挖掘目的,更重要地受存储空间和计算资源的限制往往无法快速地获取轨迹数据结果,而量子计算的出现逐步从理论层面解决了海量数据存储与计算对于空间资源与计算资源的需求,目前关于量子计算算法研究主要分为基于量子独有特性的量子算法研究和基于经典算法的量子衍生算法研究.其主要区别在于量子衍生算法主要针对目前已有基于经典计算机的算法通过引入量子信息理论优化算法结构,提升算法效率.

而目前对于量子计算与大数据算法研究主要集中于通用算法研究,而没有建立针对性的背景理论体系,轨迹数据由于其具有时空特征在大数据研究领域是重要的分支之一,所以轨迹数据的海量数据特征通过量子计算技术处理很好地实现大数据与量子计算的具体应用场景结合将丰富轨迹数据处理理论.目前还没有针对轨迹数据背景下的量子算法系统研究,本节将归纳总结在轨迹数据挖掘中常用的挖掘算法利用量子计算方式的实现.主要从数据分类、数据聚类、频繁模式挖掘和量子衍生算法四个方面进行阐述,帮助研究人员结合轨迹数据特征设计或改进基于通用量子算法的量子轨迹数据处理算法.

7.1 轨迹分类算法与量子分类算法

轨迹数据数据分类就是确定目标对象属于哪一个预定的目标类,具体来说分类是通过学习得到一个目标函数 f,把每一个属性集 x 映射到一个预先定义的类标号 y 上去.在轨迹数据分类算法中中常用决策树、贝叶斯网络等算法,在此列出目前研究人员对于基于量子理论的此类算法研究情况.

- **量子决策树(Quantum Decision Tree).**作为最经典的数据分类算法,决策树在轨迹数据处理中具有很重要的应用.文献[169]对于决策树算法与量子计算的结合提出基于量子信息熵的决策树算法,针对经典熵不纯度评价提出了相应的量子熵不纯度评价准则.同时运用 Grover 算法完成对节点信息搜索,实现量子决策树分类.
- **量子最近邻算法(QNN)/量子 k-最近邻算法(QKNN).** NN/KNN 算法是经典的轨迹数据处理算法,根据待分类数据和样本数据的相似度进行数据分类,一般常用的相似度标准有内积、欧式距离、汉明距离等.对于相似度测量,文献[170]提出用保真度 $Fid(|a\rangle,|b\rangle)=|\langle a|b\rangle|^2$ 表示量子态 $|a\rangle$ 和 $|b\rangle$ 的相似度测量,文献[171]基于文献[170]提出高维向量空间计算的量子算法实现,降低了计算时间,并且提出了邻近质心方法,针对邻近质心算法在真实环境中的不适应性,文献[172]提出了最近邻算法(nearest-neighbor algorithm,NN),在噪声环境和高维特征环境因素下具有更好地鲁棒性.关于量子 KNN 文献[173][174]进行了相应的研究.
- **量子贝叶斯网络.**量子贝叶斯网络相对于传统贝叶斯网络利用量子态概率幅和酉算子构建模型[175],量子贝叶斯网络由标记图和节点矩阵集合组成,通过基于量子理论构建的贝叶斯网络和结合 Grover 算法提升了构建效率,同时贝叶斯方法在量子态分类方面也具有独特的作用.
- **隐量子马尔科夫模型.**文献[176]首次提出了量子隐马尔科夫模型,相对于经典隐马尔科夫模型,融入量子理论后的模型不仅包括经典模型同时提供了一个更加丰富的动态处理过程,减少了针对一些随机过程的内在状态的必要数量.

7.2 轨迹聚类算法与量子聚类算法

将聚类分析的问题应用量子的方法进行优化,主要有两种方法,第一种方法是直接在经典聚类方法中加入Grover搜索算法来改进计算性能.第二种方法是利用统计力学知识或量子力学的原理来优化算法.其中*k*-means算法是最经典的聚类算法.

由于 Grover 搜索算法相对经典搜索算法具有很高的效率,所以对于 *k*-means 聚类算法中距离计算利用 Grover 算法通过黑盒算子计算量子态之间的距离提升了运算效率[177],[178].文献[171]讨论了基于绝热量子计算的*k*-means算法,绝热量子计算实施统一的门操作,不断调整量子系统中的参数,降低了*k*-means算法中对于向量计算的复杂度.

根据量子力学原理,Horn[179],[180]等人提出了新的聚类分析算法,在基于传统的尺度空间算法(scale-space algorithm)中,使用 Parzen 窗估算量(Parzen-window estimator)的极大值来确定聚类中心.

7.3 轨迹频繁模式算法与量子频繁模式挖掘算法

关联规则与量子计算的结合将极大地减少时间消耗与空间消耗,传统的Apriori算法需要很大的计算资源,Yu[181]等人提出了对于关联规则中频繁 2-项集的基于纯态的量子化算法.利用密度矩阵处理频繁数据集矩阵计算.文献[182]利用量子关联规则进行隐私保护研究,将原始垂直数据库分割成 Alice 拥有和 Bob 拥有,利用量子随机访问存储(Quantum Random Access Memory,QRAM)实现数据库模型,通过构建的量子协议进行关联规则挖掘,在时间效率和存储效率方面有显著提升.

7.4 轨迹计算与量子衍生算法

目前关于量子衍生算法的研究是非常多,其基于经典计算机的非线性处理能力对于改进传统算法具有重要作用,例如量子遗传算法其相对于传统遗传算法,有效地避免了收敛速度慢、易陷入早熟的缺点.量子衍生算法侧重于对于传统算法的局部优化,对于轨迹数据挖掘来说,在轨迹数据隐私保护、轨迹数据模式挖掘中融入量子衍生算法,并且结合具体应用场景将会进一步优化和提升算法效率.但是一般来说对于量子衍生算法的使用需要结合具体背景,某些情况下量子衍生算法并不能比传统算法取得更优的结果.

- **量子遗传/进化算法.**1996 年 Narayanan[183]等人正式提出量子遗传算法,用量子概率编码表示染色体,利用量子更新门优化种群,使种群朝着最优方向进行演化.2002 年 Han[184]等人对量子遗传算法进行进一步优化,引入种群迁移机制,提出了量子进化算法,对于轨迹数据处理来说,量子遗传算法和量子进化算法在轨迹数据聚类方法优化和轨迹数据模式挖掘中具有重要的作用,结合具体轨迹数据特征利用此类算法进行优化具有更好地效果.文献[185]将量子遗传算法应用于室内轨迹规划中,利用量子叠加态编码控制规则和利用量子旋转门实现染色体的进化,实验证明其具有更好地寻优精度.
- **量子粒子群算法.**粒子群算法[186]作为一种优化工具,其在数据分类聚类、数值优化计算等方面有重要的应用[187],[188],量子粒子群算法是 Sun[189]等人在标准的粒子群算法中提出的可以在整个可行解的空间进行搜索,从而可以寻求全局最优解的方法,其具有更好地全局收敛性和搜索能力量子粒子群算法在数据聚类方面有很好地应用,如 Zhang[190]等人针对有障碍约束的空间数据聚类,提出了基于量子粒子群算法和蚁群优化算法的空间聚类方法,并且通过实验证明其优良的聚类特性.

8 总结与展望

轨迹数据处理是在移动互联网发展与 LBS 技术进步的背景下催生出的热门研究领域之一.本文从轨迹数据挖掘、轨迹数据隐私保护、轨迹大数据支撑技术以及轨迹数据挖掘核心算法与量子算法四个方面归纳总结了近年来轨迹数据预处理方法、索引与检索技术、数据挖掘技术、隐私保护方式等方面的主要技术和算法,重点阐述了目前关于轨迹数据挖掘方案和隐私保护策略,分析了在大数据时代海量轨迹数据存储、分布式处理和可视化技术方面的发展情况,并且延伸了未来量子计算在轨迹数据处理领域中的应用,目前还没有针对轨迹数据特性进行量子算法设计的研究,因此本文介绍了轨迹数据处理中所使用的核心算法所对应的量子方案实

现.本文旨在为广大研究人员快速和较为全面了解轨迹数据处理关键技术提供帮助,为研究人员从轨迹数据挖掘、轨迹数据隐私保护和针对轨迹数据特性的量子算法设计三个方面研究提供思路与方法.在大数据时代背景下,面对海量轨迹数据的处理是一项复杂而系统性工作,未来对于轨迹数据研究所面临的挑战可以从以下几个方面进行总结与展望.

8.1 轨迹大数据挖掘方向

(1)基于多源数据融合的轨迹大数据处理

轨迹数据作为轨迹大数据研究的对象,研究人员已经从轨迹数据预处理、索引建立与查询、模式挖掘、分类等方面开展了数十年的研究,随着以 MapReduce 为代表的分布式架构的出现,研究人员希望更加快速地处理海量轨迹数据.同时,不仅仅单一的研究轨迹数据处理,需要融合其他领域数据来扩大轨迹数据处理的应用范围和研究轨迹数据背后的语义信息.因此,大数据时代下的多源数据融合,对于全面刻画数据特征和深入挖掘数据本质具有重要作用.例如,在 LBSN 的快速发展下,传统单一的分析某一类数据集已经不能全面刻画移动用户特征,融合用户社交网络数据和轨迹数据将在用户推荐和地点推荐方面发挥更大作用.

(2)基于分布式的轨迹大数据处理

大数据时代下,传统的单一处理模式和处理方法已经无法适应多源、异构、海量的轨迹数据,分布式处理是目前提升海量数据处理能力的重要手段.根据不同的轨迹数据处理要求,如何选择合适的分布式处理框架与处理算法是未来研究人员关注的重点.

- 首先,针对静态轨迹数据,以经典 Hadoop、Spark 为代表的大数据批量计算架构在处理静态轨迹数据具有较大优势,静态轨迹数据处理是先存储后计算的计算过程,通过历史轨迹数据的挖掘来获取有价值信息,目前工业界在此领域研究比较多.
- 其次,针对动态轨迹数据,以 Storm 为代表的大数据流式计算架构处理实时轨迹数据具有重要研究意义.动态轨迹数据无法进行数据事先存储,但要求近乎实时的计算,因而在基于 LBSN 的位置推荐、朋友推荐和智能交通领域具有重要应用前景.
- 另外,基于 GPU 并行计算架构在高性能计算方面具有很大优势.随着 GPU 的可编程性的增强,并且其多线程机制和内置的大量运算单元,使得基于 GPU 的数据处理能力甚至超越了基于 CPU 的数据处理能力,数据计算在 GPU 中的运算效率远远高于传统基于 CPU 上的计算效率,目前基于 GPU 的轨迹数据处理也是研究的热点之一.

(3)基于深度学习的轨迹大数据处理

深度学习最近几年成为计算机多项顶级会议讨论的重点话题,特别 2016 年 3 月 AlphaGo 击败韩国世界围棋冠军李世石,研究人员越来越发现人工智能在数据处理中的重要性,因而深度学习从传统的自然语言处理(Natural Language Processing,NLP)延伸到各个领域探索的研究方向.随着计算机性能的不断提升,尤其 NVIDIA 推出的 CUDA 并行计算架构促进了深度学习在 GPU 并行架构中研究的发展,深度学习成为 NLP、计算机视觉、图像处理等领域研究的重点,推动了神经网络在 GPU 中的应用.目深度学习已经很好地应用于社交网络社团划分[191],以 RNN(Recurrent Neural Network)为代表的深度学习模型在处理时序数据方面展现了优秀的性能优势,例如文献[192]基于 RNN 模型实现对轨迹兴趣点的预测,由于传统 RNN 对于长序列轨迹数据处理效率低,文献[193]融合社交网络数据和轨迹数据利用 RNN、GRU(Gated Recurrent Unit)模型更加精准地推荐相似用户和兴趣点.未来以 RNN 为代表的深度学习模型如 LSTM(Long-Short Term Memory)、Attention 机制在轨迹数据聚类、地点推荐和基于轨迹数据的相似用户推荐方面有重大研究前景.同样,CNN(Convolutional Neural Network)和 DNN(Deep Neural Network)深度学习模型在基于轨迹数据特征分析也具有重要研究前景如城市人流预测.

8.2 轨迹大数据隐私保护方向

(1)完善隐私保护评价体系与个性化保护

关于轨迹数据的隐私保护主要分为基于位置服务的轨迹数据保护和基于轨迹数据发布的隐私保护.在基

于 LBS 技术的轨迹数据隐私保护方法或技术中,缺乏系统化的隐私保护度评价体系与性能质量保障体系,对于轨迹数据的隐私保护处理力度很大程度影响到请求位置服务质量.同时,不同的轨迹数据语义环境下,如何提供个性化的隐私保护模式也是需要充分考虑的问题.

(2)复杂环境下的位置信息加密技术

在基于位置服务的隐私保护方法中,可以利用加密技术处理敏感位置信息,但是面对多种类、多数量移动终端和频繁位置更新数据的情况下,加密性能必然受到影响,因此,考虑优化加密技术的位置服务隐私保护策略也是目前需要关注的方向之一.另外,某些 LBS 服务商不一定拥有自己的云平台,而是将位置数据外包给云服务商,普通用户向云服务商请求查询服务,云服务商返回满足条件的查询结果,这种模式可能导致非可信云服务商窃取、备份、篡改用户位置数据.传统的加密技术在范围查询、KNN 查询方面有诸多限制,所以对于对于基于外包云服务商特征的位置信息加密技术和检索结果完整性验证的隐私保护技术也值得深入研究.

(3)完善敏感数据与非敏感数据处理机制

在传统轨迹隐私保护中,使用最广泛的是 *k*-匿名模型,通过泛化技术、抑制技术等方式对轨迹数据发布前的处理需要考虑如何提高敏感数据的保护效果与发布数据的质量之间的平衡关系.与基于位置服务的隐私保护策略相似,需要考虑轨迹数据所具有的语义特征与攻击者的不同背景知识下,较好地处理敏感数据与非敏感数据,提升保护力度的同时保障数据发布质量也是重点研究方向.

(4)基于多源数据融合的轨迹数据挖掘隐私保护

在大数据环境下面对大规模的轨迹数据,攻击者往往可以通过多数据源的融合,再通过连接攻击对匿名后的数据源进行攻击与推测个人敏感信息,这方面也是需要解决的安全问题之一.目前利用数据挖掘方式处理数据隐私保护也是研究方向之一.如本文介绍的利用关联规则隐藏算法降低轨迹数据中敏感信息支持度,达到保护效果等.

8.3 轨迹大数据与量子计算方向

目前越来越多的研究人员关注大数据与量子计算的结合即量子机器学习,而轨迹数据的海量特征需要消耗很大的空间资源与计算资源,量子计算的超并行特征与高存储能力对于轨迹数据处理的研究是值得关注的领域,轨迹数据的海量数据特征通过量子计算技术处理可以实现大数据与量子计算在具体应用场景的结合,丰富轨迹数据处理理论.目前对于量子算法的研究主要集中于经典算法的量子方案实现,而没有考虑到数据本身所具有的特征,对于具体的应用场景没有系统地介绍算法的存储过程与计算过程,所以针对轨迹数据的时空特征进行量子算法设计(包括量子衍生算法的研究)将丰富轨迹数据处理技术与方法理论,也是未来量子领域在轨迹大数据发展的重要研究方向.

- 基于量子特性的轨迹大数据量子算法设计,主要利用量子固有特性设计量子算法.轨迹大数据处理算法中,很多基本算法可以通过量子算法实现海量信息存储与计算能力加速,量子算法设计需要从量子态制备、量子态存储、量子态测量和量子态超并行计算等设计完整的方案才能真正将经典处理算法量子化,因此实现轨迹数据特征信息的存储与计算是量子计算与轨迹数据结合的难点也是未来的发展方向之一.
- 以量子遗传算法为代表的量子衍生算法,主要采用量子信息编码方式来优化数据结果,在进行轨迹数据处理中融合量子衍生算法来进行数据处理,一定程度上会促进数据处理优势与结果反馈.同时也促进了量子衍生算法在轨迹数据处理领域的延伸.

致谢 在此,我们向对本文的工作给予支持和建议的老师和同学表示感谢.

References:

[1] Xu JJ, Zheng K, Chi MM, Zhu YY, Yu XH, Zhou XF. Trajectory big data: data, applications and techniques.Journal on Communications, 2015,36(12):97−105 (in Chinese with English abstract). [doi: 10.11959/j.issn.1000-436x.2015318]

[2] Zheng Y. Trajectory Data Mining: An Overview. Acm Transactions on Intelligent Systems & Technology, 2015,6(3):1-41. [doi:10.1145/2743025]

[3] Yuan J, Zheng Y, Xie X, Sun G. T-Drive: Enhancing Driving Directions with Taxi Drivers' Intelligence. IEEE Transactions on Knowledge & Data Engineering, 2013,25(1):220-232. [doi:10.1109/TKDE.2011.200]

[4] Zheng Y, Xie X, Ma WY. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. Bulletin of the Technical Committee on Data Engineering, 2010,33(2):32-39. [doi:10.1.1.165.4216]

[5] Lu F, Duan Y, Zheng N. A practical route guidance approach based on historical and real-time traffic effects, International Conference on Geoinformatics. 2009:1-6. [doi:10.1109/GEOINFORMATICS.2009.5293444]

[6] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012:186-194. [doi:10.1145/2339530.2339561]

[7] Zheng H, Meng XF. A Survey of Trajectory Privacy-Preserving Techniques. Chinese Journal of Computers, 2011, 34:1820-1830 (in Chinese with English abstract). http://cjc.ict.ac.cn/quanwenjiansuo/2011-10/hz.pdf [doi: 10.3724/SP.J.1016.2011.01820]

[8] Zhang XJ, Gui XL, Wu ZD.Privacy preservation for location-based services: A survey. Ruan Jian Xue Bao/Journal of Software, 2015,26(9):2373-2395(in Chinese). http://www.jos.org.cn/1000-9825/4857.htm [doi: 10.13328/j.cnki.jos.004857]

[9] Zheng H, Meng XF, Hu HB, Yi H. You Can Walk Alone: Trajectory Privacy-Preserving through Significant Stays Protection, Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2012:351-366. [doi:10.1007/978-3-642-29038-1_26]

[10] Wang SH, LONG GL. Big data and quantum computation. Chinese Science Bulletin, 2015,60(5-6):499-508.[doi: 10.1360/N972014-00803]

[11] Elragal A, El-Gendy N. Trajectory data mining: integrating semantics. Journal of Enterprise Information Management, 2013, 26(5):516-535. [doi:https://doi.org/10.1108/JEIM-07-2013-0038]

[12] Chow CY, Mokbel MF. Trajectory privacy in location-based services and data publication. Acm Sigkdd Explorations Newsletter, 2011, 13(1):19-29. [doi:10.1145/2031331.2031335]

[13] Zheng Y, Xie X. Learning travel recommendations from user-generated GPS traces. Acm Transactions on Intelligent Systems & Technology, 2011, 2(1):389-396. [doi:10.1145/1889681.1889683]

[14] Li S, Peter RS. Review of GPS Travel Survey and GPS Data-Processing Methods. Transport Reviews, 2014, 34(3):316-334. [doi:10.1080/01441647.2014.903530]

[15] L Stenneth, Ouri Wolfson, Philip S. Yu, Bo Xu. Transportation mode detection using mobile phones and GIS information, ACM Sigspatial International Symposium on Advances in Geographic Information Systems, Acm-Gis 2011, November 1-4, 2011, Chicago, Il, Usa, Proceedings. 2011:54-63. [doi:10.1145/2093973.2093982]

[16] Lee WC, Krumm J. Trajectory Preprocessing. Computing with Spatial Trajectories. Springer New York,2011:3-33. [doi:10.1007/978-1-4614-1629-6_1]

[17] Barrios C, Himberg H, Motai Y, Sadek A. Multiple model framework of adaptive extended kalman filtering for predicting vehicle location, Intelligent Transportation Systems Conference. ITSC '06.IEEE, 2006:1053-1059. [doi:10.1109/ITSC.2006.1707361]

[18] Hightower J, Borriello G. Particle Filters for Location Estimation in Ubiquitous Computing: A Case Study，UbiComp 2004: Ubiquitous Computing: 6th International Conference, Nottingham, UK: Proceedings. 2004:88-106. [doi:10.1007/978-3-540-30119-6_6]

[19] Gustafsson F. Particle filter theory and practice with positioning applications, IEEE Aerospace & Electronic Systems Magazine, 2010, 25(7):53-82. [doi:10.1109/MAES.2010.5546308]

[20] Zhang F, Wilkie D, Zheng Y, Xie X. Sensing the pulse of urban refueling behavior. Acm Transactions on Intelligent Systems & Technology, 2013, 6(3):13-22. [doi:10.1145/2493432.2493448]

[21] Zhang F, Yuan NJ, Wilkie D, Xie X. Sensing the Pulse of Urban Refueling Behavior: A Perspective from Taxi Mobility. Acm Transactions on Intelligent Systems & Technology, 2015, 6(3):1-23. [doi:10.1145/2644828]

[22] Wang Y, Zheng Y, Xue Y. Travel time estimation of a path using sparse trajectories. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014:25-34. [doi:10.1145/2623330.2623656]

[23] Qi G, Pan G, Li S, Wu Z, Zhang D, Sun L, Yang LT. How Long a Passenger Waits for a Vacant Taxi -- Large-Scale Taxi Trace Mining for Smart Cities. Green Computing and Communications. IEEE, 2013:1029-1036. [doi:10.1109/GreenCom-iThings-CPSCom.2013.175]

[24] Zhang D, Sun L, Li B, Chen C, Pan G, Li S, Wu Z. Understanding Taxi Service Strategies from Taxi GPS Traces. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1):123-135. [doi:10.1109/TITS.2014.2328231]

[25] Meratnia N, By R. A. D. Spatiotemporal Compression Techniques for Moving Point Objects. Advances in Database Technology - EDBT 2004. Springer Berlin Heidelberg, 2004:765-782. [doi:10.1007/978-3-540-24741-8_44]

[26] Muckell J, Olsen PW, Hwang JH, Lawson CT, Ravi SS. Compression of trajectory data: a comprehensive evaluation and new approach, Geoinformatica, 2014, 18(3):435-460. [doi:10.1007/s10707-013-0184-0]

[27] Hershberger J, Snoeyink J. Speeding Up the Douglas-Peucker Line-Simplification Algorithm, Proc.intl.symp.on Spatial Data Handling. 2000:134--143.

[28] Zheng Y. Location-Based Social Networks: Users, Computing with Spatial Trajectories. Springer New York, 2010:243-276. [doi:10.1007/978-1-4614-1629-6_8]

[29] Richter KF, Schmid F, Laube P. Semantic trajectory compression: Representing urban movement in a nutshell. Journal of Spatial Information Science, 2012, 4(4):3-30. [doi:10.5311/josis.2012.4.62]

[30] Liu J, Zhao K, Sommer P, S Shang. A Novel Framework for Online Amnesic Trajectory Compression in Resource-constrained Environments. IEEE Transactions on Knowledge & Data Engineering, 2016:2827-2841.[doi:10.1109/TKDE.2016.2598171]

[31] Muckell J, Hwang JH, Patil V, Lawson CT, Fan P, Ravi SS. SQUISH: an online approach for GPS trajectory compression. International Conference on Computing for Geospatial Research & Applications. ACM, 2011:1-8. [doi:10.1145/1999320.1999333]

[32] Muckell J, Olsen P W, Hwang J H, Ravi SS, Lawson CT. A Framework for Efficient and Convenient Evaluation of Trajectory Compression Algorithms. Fourth International Conference on Computing for Geospatial Research and Application. 2013:24-31.[doi: 10.1109/COMGEO.2013.5]

[33] Song R, Sun W, Zheng B, Y Zheng. PRESS: A novel framework of trajectory compression in road networks. Proceedings of the VLDB Endowment, 2014, 7(9): 661-672. [doi:10.14778/2732939.2732940]

[34] Y Zheng, LZ Zhang, X Xie, WY Ma. Mining interesting locations and travel sequences from GPS trajectories. International Conference on World Wide Web. ACM, 2009:791-800. [doi:10.1145/1526709.1526816]

[35] Tiwari S, Kaushik S. Popularity estimation of interesting locations from visitor'strajectories using fuzzy inference system. Open Computer Science, 2016, 6(1):8-24. [doi:10.1515/comp-2016-0002]

[36] Jensen CS, Tradišauskas N. Map Matching. Encyclopedia of Database Systems, 2009:1692-1696. [doi:10.1007/978-0-387-39940-9_215]

[37] Yin H, Wolfson O. A weight-based map matching method in moving objects databases. Proc Ssdbm Conf,2004,16:437-438. [doi:10.1109/SSDBM.2004.10]

[38] Greenfeld JS. Matching GPS Observations to Locations on a Digital Map. Transportation Research Board 81st Annual Meeting. 2002.

[39] Mohammed A. Quddus, Robert B. Noland, Washington Y. Ochieng. A High Accuracy Fuzzy Logic Based Map Matching Algorithm for Road Transport. Journal of Intelligent Transportation Systems, 2006, 10(3):103-115.[doi:10.1080/15472450600793560]

[40] Zampella F, Jimenez Ruiz AR, Seco Granja F. Indoor Positioning Using Efficient Map Matching, RSS Measurements, and an Improved Motion Model. IEEE Transactions on Vehicular Technology, 2015, 64(4):1304-1317. [doi:10.1109/TVT.2015.2391296]

[41] Chen BY, Yuan H, Li QQ, Shaw SL. Map-matching algorithm for large-scale low-frequency floating car data. International Journal of Geographical Information Science, 2014, 28(1):22-38. [doi:10.1080/13658816.2013.816427]

[42] Abdallah F, Nassreddine G, Denoeux T. A Multiple-Hypothesis Map-Matching Method Suitable for Weighted and Box-Shaped State Estimation for Localization. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(4):1495-1510. [doi:10.1109/TITS.2011.2160856]

[43] Fouque C, Bonnifait P. Matching Raw GPS Measurements on a Navigable Map Without Computing a Global Position. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(2):887-898. [doi:10.1109/TITS.2012.2186295]

[44] Deng K, Xie K, Zheng K, Zhou X. Trajectory Indexing and Retrieval. Computing with Spatial Trajectories. Springer New York, 2011:35-59. [doi:10.1007/978-1-4614-1629-6_2]

[45] Chen L, zsu, M. Tamer, Oria V. Robust and fast similarity search for moving object trajectories. SIGMOD. ACM, 2005:491-502. [doi:10.1145/1066157.1066213]

[46] Tao Y, Papadias D, Shen Q. Continuous Reverse Nearest Neighbor Search. Proceedings of the 28th international conference on Very Large Data Bases. 2002:287-298.

[47] Frentzos E, Gratsias K, Pelekis N, Theodoridis Y. Algorithms for Nearest Neighbor Search on Moving Object Trajectories. Geoinformatica, 2007, 11(2):159-193. [doi:10.1007/s10707-006-0007-7]

[48] Lee J G, Han J, Whang KY. Trajectory clustering: a partition-and-group framework. ACM SIGMOD International Conference on Management of Data. ACM, 2007:593-604. [doi:10.1145/1247480.1247546]

[49] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases. Proc.of Intl Conf.on Data Organazation, 1993, 730:69--84. [doi: 10.1007/3-540-57301-1_5]

[50] Vries GD, Someren MV. Clustering Vessel Trajectories with Alignment Kernels under Trajectory Compression. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2010:296-311. [doi:10.1007/978-3-642-15880-3_25]

[51] ZB Chen, HT Shen, XF Zhou, Y Zheng, X Xie. Searching trajectories by locations: An efficiency study. Association for Computing Machinery. Special Interest Group on Management of Data. International Conference Proceedings. Association for Computing Machinery, 2010:255-266. [doi:10.1145/1807167.1807197]

[52] Zhu Q, Gong J, Zhang Y. An efficient 3D R-tree spatial index method for virtual geographic environments. Isprs Journal of Photogrammetry & Remote Sensing, 2007, 62(3):217-224. [doi:10.1016/j.isprsjprs.2007.05.007]

[53] Pfoser D, Jensen CS, Theodoridis Y. Novel approaches to the indexing of moving object trajectories. Proceedings of VLDB. 2000: 395-406.

[54] Jae LE, Ryu KH, Nam KW. Indexing for Efficient Managing Current and Past Trajectory of Moving Object Advanced Web Technologies and Applications. Springer Berlin Heidelberg, 2004:782-787. [doi:10.1007/978-3-540-24655-8_85]

[55] Bentley JL. Multidimensional Binary Search Trees Used for Associative Searching. Communications of the Acm, 1975, 18(9):509-517. [doi:10.1145/361002.361007]

[56] Corral A, Vassilakopoulos M, Manolopoulos Y. Algorithms for Joining R-Trees and Linear Region Quadtrees. Proceedings Ssd Symposium, 2010, 1651:251-269. [doi:10.1007/3-540-48482-5_16]

[57] Guttman A. R-trees: a dynamic index structure for spatial searching. Acm Sigmod Record, 1984, 14(2):47-57. [doi:10.1145/971697.602266]

[58] Theoderidis Y, Vazirgiannis M, Sellis T. Spatio-temporal indexing for large multimedia applications. IEEE International Conference on Multimedia Computing and Systems. IEEE, 1996:441-448. [doi:10.1109/MMCS.1996.535011]

[59] Abraham T, Roddick JF. Survey of Spatio-Temporal Databases. GeoInformatica, 1999, 3(1):61-99. [doi: 10.1023/A:1009800916313]

[60] Theodoridis Y, Sellis T, Papadopoulos AN, Manolopoulos Y. Specifications for efficient indexing in spatiotemporal databases.Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference on. IEEE, 1998: 123-132. [doi:10.1109/SSDM.1998.688117]

[61] Nascimento M A, Silva J R O. Towards historical R-trees. Proceedings of the 1998 ACM symposium on Applied Computing. ACM, 1998:235-240. [doi:10.1145/330560.330692]

[62] Tao Y, Papadias D. Efficient historical R-trees. Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings. Thirteenth International Conference on. IEEE, 2001:223-232. [doi:10.1109/SSDM.2001.938554]

[63] Tao Y, Papadias D. MV3R-Tree: A Spatio-Temporal Access Method for Timestamp and Interval Queries. VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases, September 11-14, 2001, Roma, Italy. 2001:431-440.

[64] Chakka VP, Everspaugh AC, Patel JM. Indexing large trajectory data sets with SETI. Ann Arbor, 2003, 1001(48109-2122): 12. [doi:10.1.1.12.182]

[65] Zhou P, Zhang D, Salzberg B, Cooperman G, Kollios G. Close pair queries in moving object databases. ACM International Workshop on Geographic Information Systems, Acm-Gis 2005, November 4-5, 2005, Bremen, Germany, Proceedings. 2005:2-11. [doi:10.1145/1097064.1097067]

[66] Lomet, David, Salzberg, Betty. The performance of a multiversion access method. ACM SIGMOD Record, 1990, 19(2):353-363. [doi:10.1145/93597.98744]

[67] Gudmundsson J. Computing longest duration flocks in trajectory data. 14th ACM International Symposium on Geographic Information Systems, ACM-GIS 2006,2006,Arlington,Virginia,USA,Proceedings.2006:35-42. [doi:10.1145/1183471.1183479]

[68] Jeung H, Yiu ML, Zhou X, Jensen CS, Shen HT. Discovery of Convoys in Trajectory Databases. Proceedings of the Vldb Endowment, 2010, 1(1):1068-1080. [doi:10.1007/3-540-48482-5_16]

[69] Li Z, Ding B, Han J, Kays R. Swarm: Mining Relaxed Temporal Moving Object Clusters. Proceedings of the Vldb Endowment, 2010, 3(1):723-734. [doi:10.14778/1920841.1920934]

[70] Zhou X. Online Discovery of Gathering Patterns over Trajectories. 2013 IEEE 29th International Conference on Data Engineering (ICDE). IEEE Computer Society, 2013:242-253. [doi:10.1109/TKDE.2013.160]

[71] Tang L A, Zheng Y, Yuan J, Han J. A Framework of Traveling Companion Discovery on Trajectory Data Streams. Acm Transactions on Intelligent Systems & Technology, 2013, 5(1):992-999. [doi:10.1145/2542182.2542185]

[72] Giannotti F, Nanni M, Predreschi D, Pinelli F. Trajectory Pattern Mining. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. San Jose,California, USA:ACM, 2007:330-339[doi:10.1007/978-1-4614-1629-6_5]

[73] Ye Y, Zheng Y, Chen Y, Feng J, Xie X. Mining Individual Life Pattern Based on Location History. Tenth International Conference on Mobile Data Management: Systems, Services and MIDDLEWARE. IEEE Computer Society, 2009:1-10. [doi:10.1109/MDM.2009.11]

[74] Monreale A, Pinelli F, Trasarti R, Giannotti F. WhereNext: a location predictor on trajectory pattern mining. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July. 2009:637-646. [doi:10.1145/1557019.1557091]

[75] Xiao X, Zheng Y, Luo Q, Xie X. Inferring Social Ties between Users with Human Location History. Journal of Ambient Intelligence & Humanized Computing, 2014, 5(1):3-19. [doi:10.1007/s12652-012-0117-z]

[76] Lv M, Chen L, Chen G. Mining user similarity based on routine activities. Information Sciences, 2013, 236(1):17-32. [doi:10.1016/j.ins.2013.02.050]

[77] Cao H, Mamoulis N, Cheung DW. Mining frequent spatio-temporal sequential patterns. Icdm, 2005:82--89. [doi:10.1109/ICDM.2005.95]

[78] Chen Z, Shen HT, Zhou X. Discovering popular routes from trajectories. IEEE, International Conference on Data Engineering. IEEE Computer Society, 2011:900-911. [doi:10.1109/ICDE.2011.5767890]

[79] Srikant R, Agrawal R. Mining sequential patterns: Generalizations and performance improvements. Advances in Database Technology — EDBT '96. Springer Berlin Heidelberg, 1996:1-17.

[80] Pei J, Han J, Mortazavi-Asl B, Pinto H. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. International Conference on Data Engineering. IEEE Computer Society, 2001:215-224. [doi:10.1109/ICDE.2001.914830]

[81] Ozden B, Ramaswamy S, Silberschatz A. Cyclic association rules. IEEE International Conference on Data Engineering. IEEE, 1998:412-421. [doi:10.1109/ICDE.1998.655804]

[82] Han J, Dong G, Yin Y. Efficient Mining of Partial Periodic Patterns in Time Series Database. Proc.int.conf.on Data Engineering, 1999:106-115. [doi:10.1109/ICDE.1999.754913]

[83] Lee G, Yang W, Lee JM. A parallel algorithm for mining multiple partial periodic patterns. Information Sciences, 2006, 176(24):3591-3609. [doi:10.1016/j.ins.2006.02.010]

[84] Yang J, Wang W, Yu PS. Mining Asynchronous Periodic Patterns in Time Series Data. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2000:275-279.[doi: 10.1145/347090.347150]

[85] Huang KY, Chang CH. SMCA: A general model for mining asynchronous periodic patterns in temporal databases. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(6):774-785. [doi:10.1109/TKDE.2005.98]

[86] Yang J, Wang W, Yu PS. Mining Surprising Periodic Patterns. Data Mining & Knowledge Discovery, 2004, 9(2):189-216. [doi:10.1023/B:DAMI.0000031631.84034.af]

[87] Yang J, Wang W, Yu PS. InfoMiner+: Mining Partial Periodic Patterns with Gap Penalties. 2013 IEEE 13th International Conference on Data Mining. IEEE Computer Society, 2002:725-728. [doi:10.1109/ICDM.2002.1184039]

[88] Li Z, Wang J, Han J. Mining event periodicity from incomplete observations. Knowledge & Data Engineering IEEE Transactions on, 2015, 27(5):444-452. [doi:10.1145/2339530.2339604]

[89] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast subsequence matchingin time-series databases. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1994:419-429. [doi:10.1145/191843.191925]

[90] Chan KP, Fu WC. Efficient Time Series Matching by Wavelets. IEEE International Conference on Data Engineering. IEEE, 1999:126-133. [doi:10.1109/ICDE.1999.754915]

[91] Elnekave S, Last M, Maimon O. Incremental Clustering of Mobile Objects. IEEE, International Conference on Data Engineering Workshop. IEEE, 2007:585-592. [doi:10.1109/ICDEW.2007.4401044]

[92] Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: ordering points to identify the clustering structure. SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, Usa. 1999:49-60. [doi:10.1145/304181.304187]

[93] Gao Y, Zheng B, Chen G, Li Q. Algorithms for constrained k-nearest neighbor queries over moving object trajectories. Geoinformatica, 2010, 14(2):241-276. [doi: 10.1007/s10707-009-0084-5]

[94] Gudmundsson J, Valladares N. A GPU Approach to Subtrajectory Clustering Using the Fréchet Distance. IEEE Transactions on Parallel & Distributed Systems, 2015, 26(4):924-937. [doi:10.1109/TPDS.2014.2317713]

[95] Yuan G, Sun P, Zhao J, Li D, Wang C. A review of moving object trajectory clustering algorithms. Artificial Intelligence Review, 2016:1-22. [doi:10.1007/s10462-016-9477-7]

[96] Cai ZF, Yang HX, Shuang W, Xu J, Wei WM, Na WL. A Clustering-Based Privacy-Preserving Method for Uncertain Trajectory Data. 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications. IEEE, 2014: 1-8. [doi:10.1109/TrustCom.2014.5]

[97] Xiao X, Zheng Y, Luo Q, Xie X. Finding similar users using category-based location history. Sigspatial International Conference on Advances in Geographic Information Systems. ACM, 2010:442-445. [doi:10.1145/1869790.1869857]

[98] Ying JC, Lee WC, Weng TC, Tseng VS. Semantic trajectory mining for location prediction. ACM Sigspatial International Symposium on Advances in Geographic Information Systems, Acm-Gis 2011,Chicago,Proceedings.2011:34-43. [doi:10.1145/2093973.2093980]

[99] Liu S, Liu Y, Ni LM, Fan J, Li M. Towards mobility-based clustering. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010:919-928. [doi:10.1145/1835804.1835920]

[100] Kim J, Mahmassani HS. Spatial and Temporal Characterization of Travel Patterns in a Traffic Network Using Vehicle Trajectories. Transportation Research Part C Emerging Technologies, 2015, 9:164-184. [doi:10.1016/j.trpro.2015.07.010]

[101] Han B, Liu L, Omiecinski E. NEAT: Road Network Aware Trajectory Clustering. IEEE International Conference on Distributed Computing Systems. 2012:142-151. [doi:10.1109/ICDCS.2012.31]

[102] Bashir F I, Khokhar AA, Schonfeld D. Object trajectory-based activity classification and recognition using hidden Markov models. IEEE Transactions on Image Processing, 2007, 16(7):1912-9. [doi:10.1109/TIP.2007.898960]

[103] Nascimento J C, Figueiredo M A T, Marques J S. Trajectory Classification Using Switched Dynamical Hidden Markov Models. Image Processing IEEE Transactions on, 2010, 19(5):1338-48. [doi:10.1109/TIP.2009.2039664]

[104] Mlích J,Chmelař. Trajectory classification based on Hidden Markov Models. Proceedings of 18th International Conference on Computer Graphics and Vision, 2008,101-105.

[105] Sun S, Zhao J, Gao Q. Modeling and recognizing human trajectories with beta process hidden Markov models. Pattern Recognition, 2015, 48(8):2407-2417. [doi:10.1016/j.patcog.2015.02.028]

[106] Yin J, Chai X, Yang Q. High-Level Goal Recognition in a Wireless LAN. Proceedings of the Nineteenth National Conference on Artificial Intelligence,Sixteenth Conference on Innovative Applications of Artificial Intelligence,San Jose,California,USA. 2004:578-584.

[107] Santos L, Khoshhal K, Dias J. Trajectory-based human action segmentation. Pattern Recognition, 2015,48(2):568-579. [doi:10.1016/j.patcog.2014.08.015]

[108] Liao L, Patterson DJ, Fox D, Kautz H. Learning and inferring transportation routines. Artificial Intelligence, 2007, 171(5-6):311-331. [doi:10.1016/j.artint.2007.01.006]

[109] Abidine MB, Fergani B. Evaluating C-SVM, CRF and LDA classification for daily activity recognition. International Conference on Multimedia Computing and Systems. 2012:272-277. [doi:10.1109/ICMCS.2012.6320300]

[110] Chi-Yin Chow, Mohemad F. Mokbel. Privacy of Spatial Trajectories.Computing with Spatial trajectories.Springer New York, 2011,109-138. [doi:10.1007/978-1-4614-1629-6_4]

[111] Motwani R, Xu Y. Efficient algorithms for masking and finding quasi-identifiers. Proceedings of the Conference on Very Large Data Bases (VLDB). 2007: 83-93.

[112] Sweeney L. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2008, 10(5):557-570. [doi:10.1142/S0218488502001648]

[113] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 3. [doi:10.1145/1217299.1217302]

[114] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE, International Conference on Data Engineering. IEEE, 2007:106 - 115. [doi:10.1109/ICDE.2007.367856]

[115] Kido H, Yanagisawa Y, Satoh T. An anonymous communication technique using dummies for location-based services. Pervasive Services, 2005. ICPS '05. Proceedings. International Conference on. 2005:88-97. [doi:10.1109/PERSER.2005.1506394]

[116] Yiu ML, Jensen C S, Huang X, Lu H. Spacetwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. IEEE,International Conference on Data Engineering. IEEE, 2008:366-375. [doi:10.1109/ICDE.2008.4497445]

[117] Khoshgozaran A, Shahabi C. Blind evaluation of nearest neighbor queries using space transformation to preserve location privacy. Bioscience Reports, 2010, 17(4):239-257. [doi:10.1007/978-3-540-73540-3_14]

[118] Chow CY, Mokbel MF, He T. A Privacy-Preserving Location Monitoring System for Wireless Sensor Networks. Mobile Computing IEEE Transactions on, 2010, 10(1):94-107. [doi:10.1109/TMC.2010.145]

[119] Nergiz ME, Atzori M, Saygin Y. Towards trajectory anonymization: a generalization-based approach. Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS. ACM, 2008: 52-61. [doi:10.1145/1503402.1503413]

[120] Lee K CK, Lee WC, Leong HV, Zheng B. Navigational path privacy protection: navigational path privacy protection. ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, 2009:691-700. [doi:10.1145/1645953.1646041]

[121] Pingley A, Zhang N, Fu X, Choi H A, Subrammanianm S, Zhao W. Protection of query privacy for continuous location based services. Proceedings - IEEE INFOCOM, 2011, 8(1):1710-1718.[doi: 10.1109/infcom.2011.5934968]

[122] Xu T, Cai Y. Exploring Historical Location Data for Anonymity Preservation in Location-Based Services. Dissertations & Theses - Gradworks, 2008:547-555.[doi: 10.1109/infocom.2007.103]

[123] Hoh B, Gruteser M. Protecting Location Privacy Through Path Confusion. International Conference on Security and Privacy for Emerging Areas in Communications Networks, 2005. SECURECOMM. 2005:194-205.[doi: 10.1109/securecomm.2005.33]

[124] Gidofalvi G, Huang X, Pedersen T B. Privacy-Preserving Data Mining on Moving Object Trajectories. International Conference on Mobile Data Management. IEEE, 2007:60-68.[doi: 10.1109/mdm.2007.18]

[125] Xu T, Cai Y. Feeling-based location privacy protection for location-based services. ACM Conference on Computer and Communications Security, CCS 2009, Chicago, Illinois, Usa, November. 2009:348-357.[doi: 10.1145/1653662.1653704]

[126] Matt Duckham, Lars Kulik. A Formal Model of Obfuscation and Negotiation for Location Privacy. Lecture Notes in Computer Science, 2010, 3468:152-170.[doi: 10.1007/11428572_10]

[127] Gruteser M, Liu X. Protecting privacy in continuous location-tracking applications. IEEE Security & Privacy Magazine, 2004, 2(2):28-34.[doi: 10.1109/MSECP.2004.1281242]

[128] Beresford AR, Stajano F. Location Privacy in Pervasive Computing. IEEE Pervasive Computing, 2003, 2(1):46-55.[doi: 10.1109/mprv.2003.1186725]

[129] PALANISAMY B, LING L. MobiMix: Protecting location privacy with mix-zones over road networks. the Data Engineering (ICDE), 2011 IEEE 27th International Conference on, Hannover, 2011: 494-505.[doi: 10.1109/icde.2011.5767898]

[130] Gao S, Ma J, Shi W, Zhan G. TrPF: A Trajectory Privacy-Preserving Framework for Participatory Sensing. IEEE Transactions on Information Forensics and Security, 2013, 8(6): 874-887.[doi: 10.1109/tifs.2013.2252618]

[131] You TH, Peng WC, Lee WC. Protecting Moving Trajectories with Dummies. International Conference on Mobile Data Management. IEEE, 2007:278-282.[doi: 10.1109/mdm.2007.58]

[132] Poulis G, Skiadopoulos S, Loukides G, et al. Distance-Based k^m-Anonymization of Trajectory Data. IEEE, International Conference on Mobile Data Management. 2013:57-62.[doi: 10.1109/MDM.2013.66]

[133] Abul O, Bonchi F, Nanni M. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. IEEE Computer Society, 2008:376-385.[doi: 10.1109/icde.2008.4497446]

[134] Terrovitis M, Mamoulis N. Privacy Preservation in the Publication of Trajectories. International Conference on Mobile Data Management. 2008:65-72.[doi: 10.1109/mdm.2008.29]

[135] Hoh B, Gruteser M, Xiong H, Alrabady A. Achieving Guaranteed Anonymity in GPS Traces via Uncertainty-Aware Path Cloaking. IEEE Transactions on Mobile Computing, 2010, 9(8):1089-1107.[doi: 10.1109/tmc.2010.62]

[136] Gkoulalas-Divanis A, Verykios VS. Association Rule Hiding for Data Mining.Springer US.2010,41.[doi: 10.4018/9781605660103.ch012]

[137] Atallah M, Elmagarmid A, Ibrahim M, Bertino E, Verykios V. Disclosure limitation of sensitive rules. The Workshop on Knowledge & Data Engineering Exchange. IEEE, 1999:45-52.[doi: 10.1109/kdex.1999.836532]

[138] Aydin B, Akkineni V, Angryk RA. Modeling and Indexing Spatiotemporal Trajectory Data in Non-Relational Databases. Managing Big Data in Cloud Computing Environments, 2016,6:133-163.[doi: 10.4018/978-1-4666-9834-5.ch006]

[139] Ke S, Gong J, Li S, Zhu Q, Liu X, Zhang Y. A hybrid spatio-temporal data indexing method for trajectory databases. Sensors, 2014, 14(7):12990-13005.[doi: 10.3390/s140712990]

[140] Boulmakoul A, Karim L, Laarabi M H, Sacile R, Garbolino E. MongoDB-Hadoop Distributed and Scalable Framework for Spatio-Temporal Hazardous Materials Data Warehousing. Ames, D.p. Quinn, N.w.t. Rizzoli, A.e. 2014.

[141] Zhang N, Zheng G, Chen H, Chen J, Chen X. Hbasespatial: A scalable spatial data storage based on hbase. 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications. IEEE, 2014: 644-651.[doi: 10.1109/trustcom.2014.83]

[142] Wang H, Zheng K, Xu J, Zheng B, Zhou X, Sadig S. SharkDB: An In-Memory Column-Oriented Trajectory Storage. ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014:1409-1418.[doi: 10.1145/2661829.2661878]

[143] Xia D, Wang B, Li Y, Rong Z, Zhang Z. An Efficient MapReduce-Based Parallel Clustering Algorithm for Distributed Traffic Subarea Division. Discrete Dynamics in Nature & Society, 2015, 2015(6018):1-18.[doi: 10.1155/2015/793010]

[144] Nguyen CD, Nguyen DT, Pham VH. Parallel Two-Phase K-Means. Computational Science and Its Applications – ICCSA 2013. Springer Berlin Heidelberg, 2013:224-231.[doi；10.1007/978-3-642-39640-3_16]

[145] Hu C, Kang X, Luo N, Zhao Q. Parallel clustering of big data of spatio-temporal trajectory. Natural Computation (ICNC), 2015 11th International Conference on. IEEE, 2015: 769-774.[doi: 10.1109/icnc.2015.7378088]

[146] He Y, Tan H, Luo W, Mao H, Ma D, Feng S, Fan J. MR-DBSCAN: An Efficient Parallel Density-based Clustering Algorithm using MapReduce. 2011 IEEE 17th International Conference on Parallel and Distributed Systems. IEEE Computer Society, 2011:473-480.[doi: 10.1109/icpads.2011.83]

[147] Jinno R, Seki K, Uehara K. Parallel distributed trajectory pattern mining using MapReduce[C]// IEEE, International Conference on Cloud Computing Technology and Science. 2012:269-273.[doi: 10.1109/cloudcom.2012.6427526]

[148] Sun ZY, Tsai MC, Tsai HP. Mining Uncertain Sequence Data on Hadoop Platform. Trends and Applications in Knowledge Discovery and Data Mining. Springer International Publishing, 2014:204-215.[doi: 10.1007/978-3-319-13186-3_20]

[149] Eldawy A, Mokbel M F. A demonstration of SpatialHadoop: an efficient mapreduce framework for spatial data. Proceedings of the Vldb Endowment, 2013, 6(12):1230-1233.[doi: 10.14778/2536274.2536283]

[150] Huang P, Yuan B. Mining Massive-Scale Spatiotemporal Trajectories in Parallel: A Survey. Trends and Applications in Knowledge Discovery and Data Mining. Springer International Publishing, 2015: 41-52.[doi: 10.1007/978-3-319-25660-3_4]

[151] Gowanlock M, Casanova H. Indexing of spatiotemporal trajectories for efficient distance threshold similarity searches on the GPU. Parallel and Distributed Processing Symposium (IPDPS),2015 IEEE International. IEEE,2015:387-396.[doi: 10.1109/ipdps.2015.24]

[152] You S, Zhang J, Gruenwald L. Parallel spatial query processing on GPUs using R-trees. ACM Sigspatial International Workshop on Analytics for Big Geospatial Data. 2013:23-31.[doi: 10.1145/2534921.2534949]

[153] Prasad SK, McDermott M, He X, Puri S. GPU-based Parallel R-tree Construction and Querying. Parallel and Distributed Processing Symposium Workshop (IPDPSW), 2015 IEEE International. IEEE, 2015: 618-627.[doi: 10.1109/ipdpsw.2015.127]

[154] Gudmundsson J, Valladares N. A GPU Approach to Subtrajectory Clustering Using the Fréchet Distance. IEEE Transactions on Parallel & Distributed Systems, 2015, 26(4):924-937.[doi: 10.1109/tpds.2014.2317713]

[155] Deng Z, Hu Y, Zhu M, Huang X, Du B. A scalable and fast OPTICS for clustering trajectory big data. Cluster Computing, 2015, 18(2): 549-562.[doi: 10.1007/s10586-014-0413-9]

[156] Valladares Cereceda I. GPU parallel algorithms for reporting movement behaviour patterns in spatiotemporal databases. 2013.

[157] Sellarès J A. A parallel GPU-based approach for reporting flock patterns. International Journal of Geographical Information Science, 2014, 28(9):1877-1903.[doi: 10.1080/13658816.2014.902949]

[158] Wang Z, Yuan X. Visual analysis of trajectory data. Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/Journal of Computer-Aided Design and Computer Graphics, 2015, 27(1):9-25(in Chinese with English abstract). [doi: 10.3969/j.issn.1003-9775.2015.01.002]

[159] Guo H, Wang Z, Yu B, Zhao H, Yuan X. TripVista: Triple Perspective Visual Trajectory Analytics and its application on microscopic traffic data at a road intersection. IEEE Pacific Visualization Symposium, PACIFICVIS 2011, Hong Kong, China, 1-4 March. 2011:163-170.[doi: 10.1109/pacificvis.2011.5742386]

[160] Wang Z, Yuan X. Urban trajectory timeline visualization. International Conference on Big Data and Smart Computing. 2014:13-18.[doi: 10.1109/bigcomp.2014.6741397]

[161] Bakshev S, Spinsanti L, Vidal C, Casanova MA. Trajectory Semantic Visualization. Iceis 2011 - Proceedings of the, International Conference on Enterprise Information Systems, Volume 1, Beijing, China, 8-11 June. 2011:326-332.[doi: 10.5220/0003565603260332]

[162] Von Landesberger T, Andrienko G, Andrienko N, et al. Visual Analytics Methods for Categoric Spatio-Temporal Data. Visual Analytics Science and Technology. IEEE, 2012:183-192.[doi: 10.1109/vast.2012.6400553]

[163] Li J, Zhang K, Meng Z P. Vismate: Interactive visual analysis of station-based observation data on climate changes. Visual Analytics Science and Technology. IEEE, 2015:133-142.[doi: 10.1109/vast.2014.7042489]

[164] Ferreira N, Poco J, Vo HT, Freire J, Silva CT. Visual exploration of big spatio-temporal urban data: a study of New York City taxi trips. IEEE Transactions on Visualization & Computer Graphics, 2013, 19(12):2149-58.[doi: 10.1109/tvcg.2013.226]

[165] Boyandin I, Bertini E, Bak P, Lalanne D. Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data. Computer Graphics Forum, 2011, 30(3):971–980.[doi: 10.1111/j.1467-8659.2011.01946.x]

[166] Spretke D, Stein M, Sharalieva L, Warta A, Licht V, Schreck T, Keim D A. Visual Analysis of Car Fleet Trajectories to Find Representative Routes for Automotive Research. 2015 19th International Conference on Information Visualisation. IEEE, 2015: 322-329.[doi: 10.1109/iv.2015.63]

[167] Lu M, Lai C, Ye T, Liang J. Visual analysis of route choice behaviour based on GPS trajectories. IEEE Conference on Visual Analytics Science and Technology. IEEE Computer Society, 2015:203-204.[doi: 10.1109/VAST.2015.7347679]

[168] Landesberger TV, Bremm S, Schreck T, Fellner D W. Feature-Based Automatic Identification of Interesting Data Segments in Group Movement Data. Information Visualization, 2013, 13(3):190-212.[doi: 10.1177/1473871613477851]

[169] Lu S, Braunstein SL. Quantum decision tree classifier. Quantum Information Processing, 2013,13(3):757-770.[doi: 10.1007/s11128-013-0687-5]

[170] Aïmeur E, Brassard G, Gambs S. Machine Learning in a Quantum World. Lecture Notes in Computer Science, 2014, 4013:431-442.

[171] Lloyd S, Mohseni M, Rebentrost P. Quantum algorithms for supervised and unsupervised machine learning. arXiv preprint arXiv:1307.0411, 2013.

[172] Wiebe N, Kapoor A, Svore K. Quantum nearest-neighbor algorithms for machine learning. arXiv preprint arXiv:1401.2142, 2014.

[173] Li Q, Jiang JP. Quantum K nearest neighbor algorithm,System Engineering and Electronics,2008,30(5),940-943(in Chinese with English abstract).

[174] Chen HW,Gao Y,Zhang J. Quantum K-nearest neighbor algorithm, JOURNAL OF SOUTHEAST UNIVERSITY(Natual Science Edition),2015, 45(4):647-651(in Chinese with English abstract). [doi: 10.3969/j.issn.1001-0505.2015.04.006]

[175] Schuld M, Sinayskiy I, Petruccione F. An introduction to quantum machine learning. Contemporary Physics, 2014, 56(2):172-185.[doi: 10.1080/00107514.2014.964942]

[176] Monras A, Beige A, Wiesner K. Hidden Quantum Markov Models and non-adaptive read-out of many-body states. Physics, 2010.

[177] Aïmeur E, Brassard G, Gambs S. Quantum Clustering Algorithms. Machine Learning, Proceedings of the Twenty-Fourth International Conference. 2007:1-8.[doi: 10.1145/1273496.1273497]

[178] Aïmeur E, Brassard G, Gambs S. Quantum speed-up for unsupervised learning. Machine Learning, 2013, 90(2):261-287.[doi: 10.1007/s10994-012-5316-5]

[179] Horn D, Gottlieb A.The Method of Quantum Clustering.In Proceedings of the neural information processing systems.2002:769-776.

[180] Horn D, Gottlieb A. Algorithm for data clustering in pattern recognition problems based on quantum mechanics. Physical Review Letters, 2002, 88(1):261-268.[doi: 10.1103/physrevlett.88.018702]

[181] Yu CH, Gao F, Wang QL, Wen QY. Quantum algorithm for association rules mining. PHYSICAL REVIEW A.2016,94(4).[doi: 10.1103/PhysRevA.94.042311]

[182] Ying S, Ying M, Feng Y. Quantum Privacy-Preserving Data Mining. arXiv preprint arXiv:1512.04009, 2015.

[183] Narayanan A, Moore M. Quantum-inspired genetic algorithms. IEEE International Conference on Evolutionary Computation. IEEE, 1996:61-66.[doi: 10.1109/icec.1996.542334]

[184] Han KH, Kim JH. Genetic quantum algorithm and its application to combinatorial optimization problem. Evolutionary Computation, 2000. Proceedings of the 2000 Congress on. IEEE, 2003:1354-1360 vol.2.[doi: 10.1109/cec.2000.870809]

[185] Tao J, Sun Q, Zhu E, Chen Z. Quantum genetic algorithm based homing trajectory planning of parafoil system. Control Conference (CCC), 2015 34th Chinese. IEEE, 2015: 2523-2528.[doi: 10.1109/chicc.2015.7260028]

[186] J. Kennedy, R. Eberhart. Particle swarm optimization. IEEE International Conference on Neural Networks, 1995. Proceedings. 1995:1942-1948 vol.4.[doi: 10.1007/978-0-387-30164-8_630]

[187] Juang CF, Hsiao CM, Hsu CH. Hierarchical Cluster-Based Multispecies Particle-Swarm Optimization for Fuzzy-System Optimization. IEEE Transactions on Fuzzy Systems, 2010, 18(1):14-26.[doi: 10.1109/tfuzz.2009.2034529]

[188] Xue B, Zhang M, Browne WN. Particle swarm optimization for feature selection in classification: a multi-objective approach. IEEE Transactions on Cybernetics, 2013, 43(6):1656-1671.[doi: 10.1109/tsmcb.2012.2227469]

[189] Sun J, Feng B, Xu W. Particle swarm optimization with particles having quantum behavior. Evolutionary Computation, 2004. CEC2004. Congress on. IEEE, 2004:1571-1580.[doi: 10.1109/cec.2004.1330875]

[190] X Zhang, J Wang, H Du, T Yang, Y Liu. A Quantum Particle Swarm Optimization Used for Spatial Clustering with Obstacles Constraints. International Conference on Artificial Intelligence and Computational Intelligence. IEEE,2009:424-433.[doi: 10.1007/978-3-642-04020-7_45]

[191] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 701-710.[doi: 10.1145/2623330.2623732]

[192] Liu Q, Wu S, Wang L, Tan T. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[193] Yang C, Sun M, Zhao W X, Liu ZY. A Neural Network Approach to Joint Modeling Social Networks and Mobile Trajectories. arXiv preprint arXiv:1606.08154, 2016.

附中文参考文献:

[1] 许佳捷,郑凯,池明旻,朱扬勇,禹晓辉,周晓方.轨迹大数据:数据、应用与技术现状.通信学报,2015,(12):97-105. [doi:10.11959/ j.issn. 1000-436x.2015318]

[7] 霍峥,孟小峰.轨迹隐私保护技术研究.计算机学报, 2011, 34(10):1820-1830. [doi: 10.3724/SP.J.1016.2011.01820]

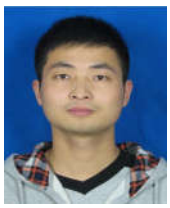
[8] 张学军,桂小林,伍忠东.位置服务隐私保护研究综述.软件学报,2015,26(9):2373-2395. [doi: 10.13328/j.cnki.jos.004857]

[10] 王书浩,龙桂鲁.大数据与量子计算.科学通报, 2015(Z1):499-508. [doi:10.1360/N972014-00803]

[158] 王祖超,袁晓如.轨迹数据可视分析研究.计算机辅助设计与图形学学报,2015(1):9-25.[doi: 10.3969/j.issn.1003-9775.2015.01.002]

[173] 李强,蒋静坪.量子 K 最近邻算法.系统工程与电子技术,2008,30(5):940-943. [doi: 10.3321/j.issn:1001-506X.2008.05.040]

[174] 陈汉武,高越,张军.量子 K-近邻算法.东南大学学报(自然科学版),2015,45(4):647-651. [doi: 10.3969/j.issn.1001-0505.2015.04.006]



高强(1993-),男,安徽合肥人,博士生,CCF 学生会员,主要研究领域为量子计算与大数据，轨迹数据挖掘.

王瑞锦(1980-),男,博士,实验师,美国西北大学访问学者,CCF 会员,主要研究领域量子通信，量子计算与大数据.

张凤荔(1963-),女,博士,教授,博士生导师,CCF 会员,主要研究领域空间数据库,信息安全.

周帆(1981-),男,博士,副教授,CCF 会员,主要研究领域移动互联网、社交网络、机器学习和深度学习等.