第47卷第7期 中国科学技术大学学报 Vol.47,No.7
2017年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2017

文章编号:0253-2778(2017)7-0087-08

面向多维稀疏时空数据的可视化研究

赵 凡[1,2,3],蒋同海[1,3],周 喜[1,3],马 博[1,3],程 力[1,3]

(1.中国科学院新疆理化技术研究所,乌鲁木齐 830011;2.中国科学院大学,北京 100049;
3.新疆民族语音语言信息处理实验室,乌鲁木齐 830011)

摘要:时空数据的多维属性和稀疏分布特征是数据分析的主要难点.利用数据可视化技术实现多维稀疏时空数据的表现和辅助分析是当前一个研究热点.基于此,提出一种多模态数据可视化方法,利用多层次视图表现模型和人机交互方式,直观展示稀疏时空数据的多维属性,进而分析数据的统计群组特征和典型个体行为模式,最终实现对异常行为的识别.针对覆盖新疆全区的车辆加油数据,融合多种相关数据源,利用该可视化方法,实现了一个车辆行为可视化数据分析系统,使用平行坐标、地图、日历矩阵、桑基图、散点图等视图模型,实现了对个体行为特征和群体行为模式的可视化表现,进而实现了对异常行为的识别、确认和预警等功能.

关键词:可视化;时空数据;多维数据;稀疏数据

中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.07.010

引用格式:赵凡,蒋同海,周喜,等. 面向多维稀疏时空数据的可视化研究[J]. 中国科学技术大学学报,2017,47(7):323-330.

ZHAO Fan, JIANG Tonghai, ZHOU Xi, et al. Visualization of multi-dimensional sparse spatial-temporal data[J]. Journal of University of Science and Technology of China, 2017,47(7):323-330.

Visualization of multi-dimensional sparse spatial-temporal data

ZHAO Fan[1,2,3], JIANG Tonghai[1,3], ZHOU Xi[1,3], MA Bo[1,3], CHENG Li[1,3]

(1. *Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumchi* 830011, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*;
3. *Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumchi* 830011, *China*)

Abstract: Multi-dimensionality and sparseness of spatial-temporal data are major challenges for data analysis. Data visualization can effectively address certain data analysis challenges and has increasingly drawn attention from both industry and academia. A hybrid approach for the visualization of multi-dimensional sparse spatial-temporal data was proposed. The method combined multiple data view models and human-machine interaction mechanisms in order to intuitively express the multi-dimensional features, statistical group features, as well as typical individual behavior patterns. Furthermore, a visual analysis method was introduced for the identification and detection of abnormal individual behaviors. A data visualization system based on gas filling data gathered from gas stations in Xinjiang Province was implemented. By using different view models (parallel coordinates, map view, calendar matrix, Sankey

收稿日期:2016-08-28;**修回日期**:;2016-12-08

基金项目:新疆维吾尔自治区高技术计划项目(201512103),新疆维吾尔自治区重点实验室项目(2016D03019),中国科学院科技服务网络计划项目(KFJ-EW-STS-129)资助.

作者简介:赵凡,男,博士/副研究员. 研究方向:数据分析、数据可视化及可视分析. E-mail: zhaofan@ms.xjb.ac.cn.

通讯作者:程力,博士/研究员.E-mail: chengli@ms.xjb.ac.cn.

Diagram, scatter plots), the system can perform various tasks for the visualization of individual and group behavior patterns. It also supports the detection, confirmation, and early alert for abnormal gas filling behaviors.

Key words: visualization; spatial-temporal data; multi-dimensional data; sparse data

0　引言

数据可视化指使用图形图像处理、计算机视觉等技术，通过建模及对平面、立体、属性的显示，提供一种直观的挖掘、分析与展示数据内部蕴含规律的手段.可视分析结合了可视化、人机交互和自动分析，可以把数据通过直观的视觉形式展现给用户，并提供有效的交互方式支持用户对数据进行探索[1].数据可视化当前已经成为当前大数据分析的重要研究领域，在各行各业都得到了广泛的应用.

数据可视化研究同互联网、物联网、地理信息系统、社交网络等应用领域都有着十分紧密的联系，其中多维时空数据是数据可视化研究热点之一.多维数据指的是具有多个维度属性的数据变量，时空数据是同时带有地理位置属性和时间信息属性的数据集合.分布在一个区域内的加油站产生的车辆加油记录数据就是典型的多维时空数据，本文以此类数据为重点研究目标开展可视化研究.

本文的主要目标是通过数据可视化的分析方法，利用个体的人(或车)在不同的时间点产生的加油记录数据以及在相同或不同加油站地点之间的行驶轨迹数据，来总结和发现个体的行为模式，并通过分析典型行为模式来预警个体异常的行为.

1　相关工作

%1.1 由于数据采集方式的不同，时空数据可以是连续的，比如通过车载 GPS 设备采集到的车辆信息，可以描述精确的车辆行驶轨迹，从而分析出准确的行为模式[2]；数据也可能是稀疏的，比如加油站或停车场采集到的信息只包括相对孤立的车辆位置和时间信息.由于目前通过 GPS 等设备采集的数据大多只有公共交通设施如公交车、出租车等，缺少私家车等信息，而对这些车辆的运动行为模式分析能帮助人们发现和解决一些重要的问题，所以对稀疏的时空数据进行分析，找出相似个体的行动规律，分析出相应的行为模式，进而发现问题和异常，具有现实意义.

1.1　稀疏时空数据的处理

时空数据的稀疏性是指采样数据在时间和空间范围内分布不稠密；Wang 等[3]提出了使用稀疏数据集用来研究大量车辆轨迹形成的宏观交通模式，Deka 等[4]使用稀疏的 GPS 数据来设计基于地图的轨迹匹配算法；Sanaullah 等[5]使用稀疏的 GPS 数据来估算出行时间；Wang 等[6]利用稀疏轨迹数据来估算在城市任意道路行驶的时间.以上基于稀疏数据的研究大都是通过研究大量样本数据如出租车、公交车等的运动数据，研究群体的共同行为模式，同时通过增加样本数据的互补性来解决数据的稀疏性问题，针对个体的车辆的运动轨迹模式的识别目前尚未见较有成效的研究.本文使用的数据集中，虽然加油记录采集点覆盖了全疆大部分加油站，但数据本身仍属于稀疏数据.车辆在某个加油站发生加油的动作时会记录其位置和时间信息，但该车辆在两次加油期间(通常间隔较长)的运行轨迹则无法获知.这种数据特点为车辆的行为模式分析带来了一定的难度.

1.2　时空数据可视化

随着物联网的发展、传感器与移动终端的迅速普及，使得时空数据成为大数据时代典型的数据类型[7-8].时空数据可视化与传统的地理制图学相结合，重点对时间与空间维度以及与之相关的信息对象属性建立可视化表征，对与时间和空间密切相关的模式及规律进行展示.为了反映信息对象随时间进展与空间位置所发生的行为变化，通常通过信息对象的属性可视化来展现.将时间事件流与地图进行融合是一种典型的方法，多维数据可视化方法也常与时空数据可视化进行融合，如使用多维平行坐标轴与传统地图制图方法结合[9]，还有采用时空立方体(space-time cube)[10]以三维方式对时间、空间及事件做直观展现.本文设计了一种多模态数据可视化方法，将平行坐标[7,9,11]、地图和日历矩阵[12]等多种视图综合起来表现数据的多维属性，这种设计可以完整地展现数据的时空特征.

1.3　轨迹可视化

轨迹数据作为时空数据中的代表类型，已经有很多相关可视化研究.Andrienko 等[13]将轨迹可视化技术可分为三大类：直接可视化，聚集可视化和特

征可视化.这三种类型的技术都可以适用于稀疏时空数据可视化.

直接可视化是将轨迹数据直接显现在视图上的方法.包括将轨迹作为动画来表现物体的移动[14],将轨迹表示为折线[15]或堆叠路径带[16],将轨迹显示在时间轴上[16-17],将空间和时间信息一起展示的时空信息立方体[18].如果应用直接可视化来表现数据稀疏的轨迹,我们需要解决的问题是轨迹重建的不确定性.例如,车辆在两个点之间行驶的精确轨迹实际上是未知的,所以对稀疏数据轨迹进行可视化需要先假设经过的路径,还要假设车辆是匀速直线运动的,这些假设可能是有问题的.为了解决这样的问题,Stoll 等[19]提出了在轨迹动画和路径线顶端加上色带来表现轨迹的不确定性.聚集可视化和特征可视化的相关研究也比较多.北京大学可视化及可视分析小组开发的 TripVista 系统[20]集成了主题河流图[21]和平行坐标视图[11],帮助用户有效地发现交通流的规律和异常特征,使用密度图[22]组件来分析交通流量的密度以及流量图组件来分析大规模的交通流[23].Liu 等[24]设计了一个带 Trip-View 视图的可视分析系统,用户可通过在外圈与内圈上的选择、过滤等交互操作进行出租车路线多样性的时空分析.Wang 等[25]设计了一种城市道路交通拥堵可视分析工具用以研究交通阻塞的时空传播规律.Krüger 等[26]设计的 TrajectoryLenses 系统基于交互技术,支持对长轨迹数据的复杂过滤表达及长时间区间分析.Pu 等[27]构建了一类依据出租车 GPS 轨迹数据的城市道路流量交互式可视分析系统,通过车辆指纹、道路指纹和区域指纹等可视化组件从车辆、道路和区域多个层面进行监测和分析城市交通流量模式.

上述工作都能够很好地展示出轨迹路线,但是在轨迹数据稀疏、路线不确定的情况下,需要提供一种交互方式,让用户能够参与到轨迹细节调整的工作中,使系统与人工相结合,描绘出更加精确合理的轨迹图.

2 方法概述

2.1 问题描述

本文的目标是要通过数据可视化展示及可视分析工作,观察并总结数据的一些普遍特征,识别一些典型模式,同时能够及时发现和预警作为个体的人或车的不正常行为.例如,同一辆车在短时间内在一个或多个加油站多次加油、或者同一人多次更换车辆加油等行为都是异常的,可能会有人通过提取和储备车中的汽油来进行可能的暴恐或违法行为.为实现这个目标,我们需要做到以下几点:

(Ⅰ)数据可视化展示

每条活动数据中都含有时空信息,且数据不是孤立的.同一个体在不同时间点产生的活动数据具有某种关联.将同一个体的全部或部分相关数据的所有属性用可视化方式直接展示出来,既可以完整保留数据信息,又还可以方便地展示数据中隐含的特征.

(Ⅱ)特征分析

通过直接观察可视化后的数据视图,找出数据中直观或隐含的各种特征,选取合适的特征并进行分析,包括系统自动分析和人工分析相结合,从分析结果中识别出典型的行为模式.

(Ⅲ)异常发现及预测

一旦确定了正常的行为模式,就可以制定对应的规则,不符合这个规则的活动数据就可以判断为异常数据.进一步,若能分析出典型的异常模式,就可以对可能发生的异常行为作出预测.

2.2 数据描述

本文采用的数据主要是加油记录数据,每条数据记录都可以表示为:人(p)在时间(t)驾驶车(c)在加油站(g)加了数量(l)的汽油,生成了一条加油记录 d,于是 N 条加油记录可以描述为

$$d_i=\{p_i,c_i,l_i,g_i,t_i\}\mid i=1,\cdots,N.$$

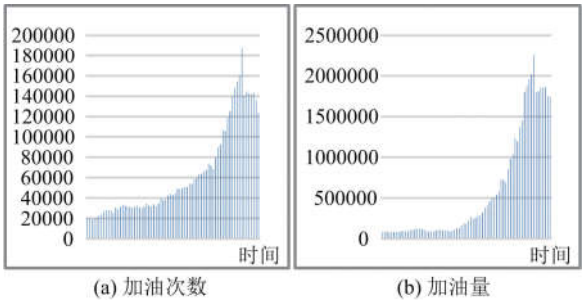
式中,g 可以表示为

$$g_i=\{x_i,y_i,\mathrm{area}_i,\mathrm{type}_i\}.$$

其中,x、y 代表加油站的经纬度坐标,area 是加油站所属行政区域的信息,type 代表该加油站的类别,比如中石油、中石化等.

本文分析的数据主要来源是“自治区汽油销售信息采集监管平台”.图 1 是该平台的数据记录统计图,目前每天产生的加油记录超过 10 000 条.图 2 展示了平台内所有加油站位置的概况,其中的标注点代表加油站,各个行政区的涂色颜色深浅代表该区接入监管平台的加油站数量的多少,颜色越深代表接入的加油站数量越多.从图 2 可以看出,目前接入系统的加油站已基本覆盖了全疆范围.本文从平台数据库中选择了 2015 年 7 月至 2016 年 6 月共 12

个月的数据记录，包含全疆 130 万辆车（包括摩托车）在近 1 000 个加油站的加油记录，总记录数达到了 600 万条.



(a) 加油次数　　(b) 加油量

图 1　平台近两个月加油数据统计

Fig.1　Refueling data statistics in the system for two months

图 2　全疆加油站数量及位置分布统计

Fig.2　The amount and distribution of gas stations in the province

2.3　工作流程

本文所设计的可视化研究的工作流程如图 3 所示，可分为 4 个部分.

（Ⅰ）数据预处理

将数据从其他业务系统中汇集到平台，保留数据源的数据格式，具有快速、完整地从多个异构数据源采集数据的功能；然后是数据清洗，需要去除原始数据的噪音和一些脏数据.最后是数据融合，将多源异构数据融合为一体，并转换为统一的数据表达方式，将处理后的数据存入图数据库及文档数据库中准备进行后续的分析工作.

（Ⅱ）直接可视化

系统通过多模态数据视图将数据源中的数据特征及分析结果提供视觉上直观的可视化展示，并提供交互式的操作界面，满足包括数据分析员在内的各类系统用户的具体使用需求.

（Ⅲ）可视化分析

根据特征属性及最终的分析需求设计可视分析方案，通过人机交互操作结合数据分析算法，得到最终分析结果.

（Ⅳ）异常处理

识别并验证不符合典型或正常行为模式的异常数据，并计算异常数据判定的可信度与预警规则的匹配度，向相关用户发出异常预警信息.

该系统利用可视化工具识别数据的特征和典型的群组和个体行为模式，进而发现和检测异常的行为数据，为用户完成复杂的数据分析工作提供辅助支持.

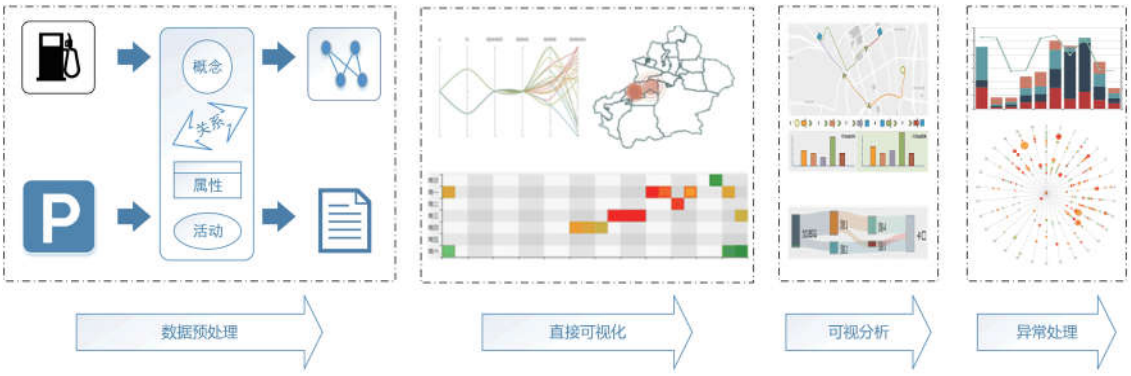


图 3　工作流程

Fig.3　System pipeline

3 多维稀疏时空数据可视化

表 1 中列出了本节中使用的所有可视化视图与数据属性特征的关系.

表 1 可视化视图列表

Tab.1 The list about visual views

图例编号	可视化视图	表现的数据特征
图 5	平行坐标	多维属性展示
图 6	地图	地理属性展示
图 7	日历矩阵	时间属性展示
图 8	地图＋柱形图	不确定轨迹调整
图 9	桑基图	两点间路线选择
图 10	堆叠柱形图＋折线图	两组变量的比较
图 11	极坐标散点图	时间分布和多维属性展示

3.1 数据预处理

在进行数据可视化及可视分析工作之前,首先要对来自多个系统的数据进行采集、清洗、融合,最终形成平台定义的统一数据表达.

由于许多数据集在类型、结构、语义、组织、粒度等方面是异构的,因此需要一种方法来实现跨数据集的有效操作,并设计一种合适的数据表示方法来反映数据的结构、层次和多样性.数据融合的工作包括构建一种机制,在少量人工参与的情况下将多源分散的数据有机整合起来,有效实现不同数据源的海量数据共享,具体内容包括模型管理、模式映射与匹配、相似数据合并与关键节点联通几个方面.

本文采用 RDF 的表示方法作为数据融合结果的表示形式.采用概念(Concept)、关系(Relations)、活动(Activity)、属性(Property)4 要素的数据模型

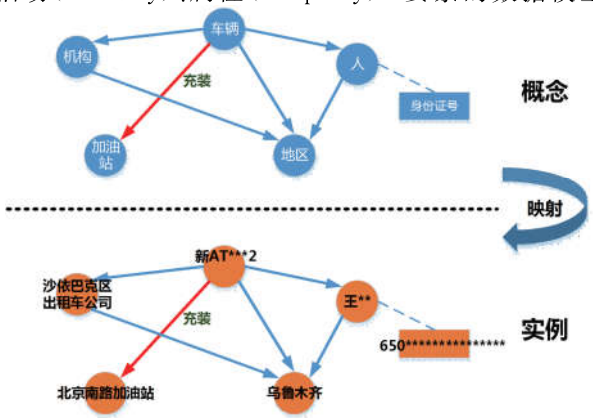


图 4 概念与实例的映射示意

Fig.4 The mapping of concept and instance

的表示形式,数据映射如图 4 所示,其中图上部的圆形代表概念,充装线条代表活动,其余线条代表关系,图上部的矩形代表属性,图下部是实例,该图说明了概念到实例的映射方法.

本文使用 NOSQL 数据库来存储经过清洗和融合的数据.考虑到不同的 NOSQL 数据库处理不同数据类型的特性,我们选择了两种 NOSQL 数据库,使用 Neo4j 图数据库来存储知识层面的数据,包括概念、关系、属性数据,最终形成领域知识图谱;使用 MongoDB 来存储活动层面的数据,例如加油记录等数据.此外,本文使用 Elasticsearch 作为全文索引搜索引擎,配合以上两种 NOSQL 数据库做下一步的数据分析及可视化工作.

3.2 直接可视化

直接可视化是对经过清洗的原始数据的直接展示,数据的所有特征都可以表现在视图中.直接可视化的目的是为了利用人的视觉分析判断能力,发现或总结出数据的一些隐含的特征,从而为进一步的数据分析提供基础.

本文关注的重点数据特征是其时空属性,在设计可视化方案中首先从时间和空间两个维度来展示数据,并且设计了可以同时展现时空属性的视图来表现数据的特征.本系统一共设计了 3 种视图来实现原始数据的直接可视化:平行坐标、地图、日历矩阵图.

3.2.1 平行坐标

平行坐标是一种可以在二维平面上展示多维数据关系和趋势的一种经典可视化方法,该方法使用平行的多条竖直坐标轴来代表多个维度,在轴上刻划某一维度数据的数值或分类,用曲线相连某一数据项在所有轴上的坐标点.本文使用平行坐标表现各种关系,如人车关系、加油站和所属地区的关系、车辆和加油站的关系等.还可以表现随时间变化加油数据的变化趋势,包括加油数量的变化、所选加油站的变化等.图 5 画出了由实际数据形成的平行坐标视图,展现了各种不同特点的加油数据集合.

设计平行坐标的时候,考虑要展示数据集的所有属性,包含了人(p)、时间(t)、车(c)、加油站(g)、加油数量(l),因此坐标轴就按照这些属性来设计,分别为人、车、加油地区、加油站、加油量、加油时间.然后取每个人一年内的所有加油数据绘制图形,可从图中看到的特征有:在某个区域加油的频度,加油时间的间隔,每次加油的数量等.

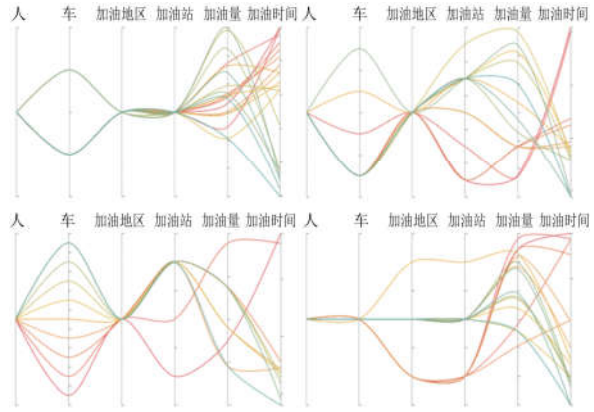


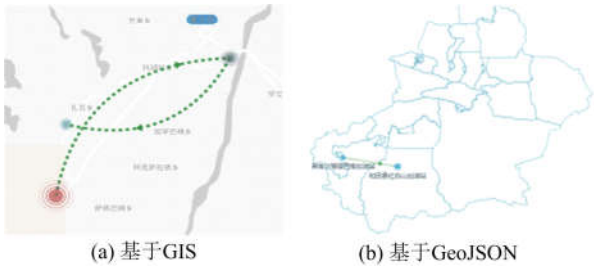
图 5　不同个体多次加油数据的平行坐标

Fig.5　The parallel coordinates of the individual refueling data

平行坐标视图能够直接反映出特定的数据特征，如车辆相关驾驶员的数量、驾驶员所驾驶过的车辆的数量、个体驾驶员或车辆的加油活动的地理范围及加油时间间隔等.

3.2.2　地图

使用地图可以方便直观地展示具有地理信息的数据，并可以从中分析出数据集相关的空间特征.本文使用了两种地图绘制方法，一种是在引用 GIS 系统数据绘制出的详细地图上添加表现数据地理空间属性的点及连接线，可以直观地表现距离相近的数据点之间的关系，如图 6(a)所示，这里只显示不同数据点之间的移动方向.另一种是使用 GeoJSON 数据绘制出的表现行政区划和轮廓线的地图，用来表现数据的空间分布特征以及随时间的进展数据的空间属性发生的改变，如图 6(b)所示.



(a) 基于GIS　　(b) 基于GeoJSON

图 6　加油站地理分布

Fig.6　Geographical distribution of gas stations

地图中表现信息的主要是点和线，其中点代表了加油站，位置可以直接在地图上看到，加油量(统计的总量)则用点的面积大小表示，不同的加油站用不同的颜色区分，为了不和 GIS 地图上原有的地标描点相混淆，在加油站的描点上使用了动画效果以便更加突出加油站的位置特征.两点之间的连线主要表现随时间的变化，前后两次加油地点变化的轨迹，图中只能表现出前后两次加油所在加油站之间的轨迹，并不能反映出车辆真实的行驶轨迹.如果前后两次加油站相同则经过表示该加油站的点绘制一条随机的闭合曲线.从视图上可以观察到个人在一段时间内的活动区域和个人习惯加油的地点.

3.2.3　日历矩阵图

反映时间特征属性的可视化方案使用时间轴、多快照或动画的形式比较常见.本文采用数据集的时间特征有个明显的特点，即同一辆车两次加油的时间间隔比较长，所以连续的时间序列表现方法在此并不适用，使用日历矩阵可以方便地表现更大范围时间尺度的离散数据，从而集中反映一段时间内的加油记录，示例数据如图 7 所示.

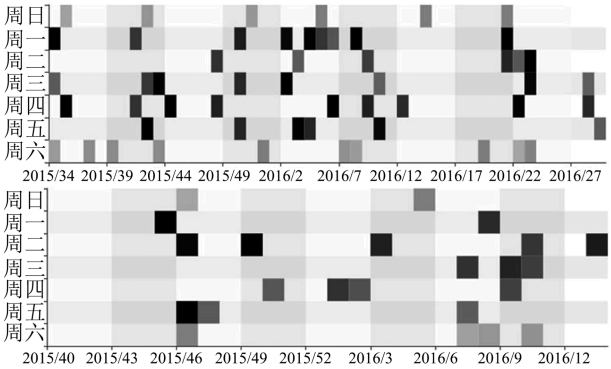


图 7　不同个体加油数据的日历矩阵

Fig.7　The calendar matrix of the individual refueling data

日历矩阵的纵轴表示每周的周日到周六，横轴表示周数，并固定分为 24 周约半年的时间.这样全图就划分为 168 个小格，每一格代表一天.将对应日期的加油行为绘制在图上，灰度深浅代表了加油量的多少，而且周末加油和工作日加油的表示灰度也有区分，如果同一天加油次数多于一次，则当天的加油量为多次加油量之和.

观察日历矩阵图可以获取的信息主要有加油频次的疏密程度，习惯加油时间的分布，还可以判断出使用车辆的频度.

进一步分析该数据集的特征，可以看出同一个体的活动数据数量不多，不适于使用通常的数据分析和机器学习方法分析个体特征和模式.由于不同个体之间的差异，如人的职业、收入、习惯等，所以通过分析群体特征和模式来判断个体的异常状况的方法也不适用.对单一的加油记录数据源分析很难达到最终目的，为此我们引入了其他数据源来辅助分析结果，包括卡口数据、停车场数据以及一些辅助的相关数据源，如历史道路拥堵状况数据、节假日信息

等,将这些数据汇总起来,就可以分析出具体的行驶轨迹,从而判断个体异常状况是否发生.

3.3 OD(origin/destination)数据轨迹分析

加油数据、卡口数据和停车场数据虽然来源不同,但数据特征的相似度很高,都包含了车辆信息和带有时间戳的空间信息.卡口数据 e 可以表示为

$$e=\{c,\mathrm{dir},x,y,t\}.$$

式中,dir 代表了车辆 c 经过卡口的方向,比如是自西向东还是自东向西,x、y 代表卡口的经纬度坐标,t 是车辆经过卡口的时间点.同样的停车场数据 p 可以表示为

$$p=\{c,x,y,t_{\mathrm{in}},t_{\mathrm{out}}\}.$$

式中,x、y 代表了停车场的经纬度坐标,t_{in}、t_{out}则表示车辆 c 进入和离开停车场的时间点.将上述几个数据源的数据融合后,可以得到一种新的数据表达形式如下:

$$\{\mathrm{car},\mathrm{origin},\mathrm{destination}\}.$$

式中,car 代表车辆信息,origin 和 destination 代表出发点和目的地信息,其中包含了地理空间信息和时间信息,可以看出这就是一条典型的 OD 数据[2].OD 数据描述的是物体在出发点、目的地之间移动,但不记录具体的移动路径.通常 OD 数据的可视化方法包括流向图、OD 矩阵和 OD 图,但这些方法大多是描述一定区域中群体移动的特征,如出租车、地铁等交通工具的轨迹数据.本文关注的主要是基于个体行为特征来识别异常,所以使用 OD 数据的目的是为了分析车辆可能的真实行驶轨迹.具体分析过程如下:

(Ⅰ)首先选取两次时间相连的加油记录数据,将这两次数据中的加油站地理位置作为起点和终点,车辆 c 的行驶轨迹为:$g_{\mathrm{start}}\rightarrow g_{\mathrm{end}}$,时间范围为 t.

(Ⅱ)找出 t 范围内的所有与 c 相关的卡口数据和停车场数据,取得卡口数据集合 $e=[e_1,e_2,\cdots,e_m]$ 以及停车场数据集合 $p=[p_1,p_2,\cdots,p_n]$.

(Ⅲ)按照时间先后顺序组合集合中的数据,可得到一组有序的带有时间戳的车辆行驶轨迹序列,例如可表示为如下序列:

$$g_{\mathrm{start}}\rightarrow e_1\rightarrow p_1\rightarrow e_2\rightarrow e_3\rightarrow p_2\rightarrow e_4\rightarrow g_{\mathrm{end}}$$

这个序列代表着车 c 在时间 t 的范围内共经过了 4 个卡口和 2 个停车场.

(Ⅳ)将第(Ⅲ)步取得的轨迹序列转换为一组 OD 数据集合,例如其中一组 OD 数据为 $\{c,p_1,e_2\}$,这组数据代表的含义就是车 c 从出发点 p_1 出发,经过一段时间后到达目的地 e_2.因为 OD 数据中的出发点和目的地都是带有时间信息的,所以在两点间行驶的时间是已知的.

(Ⅴ)对每一组 OD 数据,提供车 c 可能的轨迹路线,并标注好每条轨迹路线的置信度.实际轨迹路线的生成方法如下:通过 OD 两点的地理坐标,从公共地图 API 中获取几组备选路线的经纬度坐标集合,根据每条路线的距离、该路线的限行速度、该路线的拥堵情况和实际车辆的行驶时间计算选择本路线的置信度,计算过程如算法 3.1 所示.

算法 3.1 OD 轨迹生成算法

输入:$C_1=\{x_1,y_1\}$, $C_2=\{x_2,y_2\}$//OD 两点经纬度坐标

t_o, t_d//经过起始点和结束点的时间

输出:$R=[\{R_1,cf_1\},\{R_2,cf_2\},\cdots\{R_n,cf_n\}]$

//R_n是路线坐标集合,cf_n是该路线的置信度

```
L←getLines(C1, C2)
R←new array
FOR EACH li∈L
P←all path of li
FOR EACH pi∈P
d←distance of pi
v←speed limit of pi
tj←traffic jam status of pi at to
v←getVelocity(v, tj)
ti= d/v
t+=ti
END FOR
cfi←getCf(t, td-to)
R.push({li, cfi})
END FOR
```

(Ⅵ)默认选取每组 OD 数据中置信度最高的轨迹路线,组合成一条经过所有卡口和停车场的轨迹路线,即为车 c 从 g_{start} 到 g_{end} 的具体的行驶轨迹.这一步除了使用系统默认的轨迹路线方案之外,用户还可以通过交互的方式自己选择其余的轨迹路线组合成最终的行驶轨迹.

由 OD 数据分析出来的轨迹路线是分布在真实路网之上的,所以可视分析设计的主要视图就是基于 GIS 的地理视图,如图 8 所示.

系统在轨迹分析图的地图视图上以实际地理坐标标注每组 OD 数据的出发点和目的地,并用不同的形状和颜色来区分不同类型的点,如圆形代表加

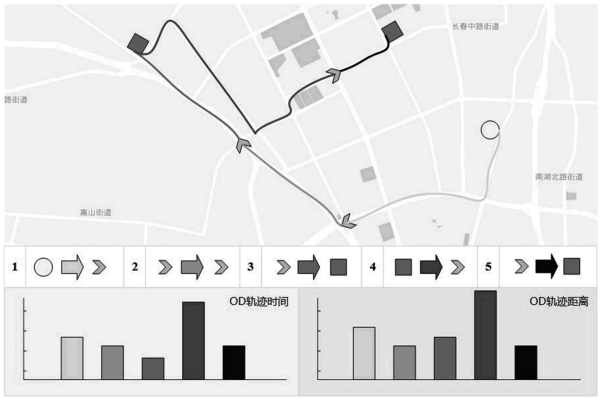


图 8　轨迹分析

Fig.8　Trajectory analysis

油站，箭头代表卡口并标明了车辆经过卡口的方向，矩形代表停车场.在每组 OD 之间使用不同的灰度绘制轨迹路线，并在地图视图的下方排列出一个时间阶段（如一天）之内所有的行驶轨迹 OD 示意图.分析图最下方则以柱形图的方式展现出这个时间阶段所有 OD 数据的精确时间统计及所选轨迹的实际距离，一般情况下这两项参数应该是正相关的，可以作为用户选择轨迹的参考.用户可以使用交互的方式，通过点击不同的 OD 示意图，展开车辆可能行驶的 3 条轨迹选择如图 9 所示，该图以桑基图[28]表示，起始点是相同的，中间经过的道路是变化的.用户通过点击操作选择其中最有可能的轨迹路线，选择结束后轨迹分析图上的路线就会发生变化，同时 OD 轨迹距离统计图上的柱形图中对应的 OD 轨迹也会随着路线的改变调整距离，用户通过直接观察对比 OD 轨迹时间统计图和 OD 轨迹距离统计图的相关度，并且通过日常经验及参考当时的路况判断该时间点地图上的路线是否合理（例如某些道路在这个时间段是禁止通行的，则经过该道路的路线是不合理的），判断本次操作的结果是否正确.

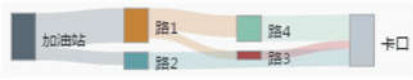


图 9　轨迹选择

Fig.9　Trajectory selection

3.4　异常检测与预警

在行驶轨迹确定后，就可以采用一些算法来计算异常出现的概率.在判断异常之前，首先根据经验和常识提出一些前提假设条件，我们认为符合这些假设条件的行为是正常的行为，反之就可能是异常行为.假设如下：

（Ⅰ）根据日常经验得知，作为个体的汽车，每次加油后如果没有抽取汽油的行为，所有汽油应全部消耗在行驶过程中，因此假设每次车 c 的加油量 l 和两次加油时间之间行驶的总距离 k 是正相关的.

（Ⅱ）取 n 次加油的加油量集合 $s_1=[l_1,l_2,\cdots,l_n]$ 以及对应计算得到的 n 次行驶轨迹总距离集合 $s_2=[k_1,k_2,\cdots,k_n]$，由假设（Ⅰ）可得出，s_1 和 s_2 这两组数据为线性的，且变化趋势是相似的.

（Ⅲ）在融合数据中获取车 c 的型号，可以计算出该型号的车每公里理论耗油量 fpk，则需要满足如下条件：$k\times fpk\approx l$

（Ⅳ）同时满足假设（Ⅱ）和（Ⅲ）的情况下，判断车 c 的行为是正常的，反之为异常.

根据以上前提假设条件，我们选择使用皮尔逊积矩相关系数（Pearson correlation coefficient）[29]来计算假设（Ⅱ）中 s_1 和 s_2 的相似度，公式如下：

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})2}}\tag{1}$$

式中，x_i 代表 s_1 集合中的元素，y_i 代表 s_2 集合中的元素，$\overline{x}$、$\overline{y}$ 分别代表 s_1 和 s_2 集合所有元素的平均值.最后的计算结果 r 是一个介于[−1,1]之间的值，用来描述两组线性的数据一同变化移动的趋势，如果 $r\geqslant 0$，说明 s_1 和 s_2 是正相关的.

利用以上的前提假设条件，我们可以统计出当天所有加油信息和历史数据的变化曲线，计算出相对应的行驶轨迹距离并通过公式（1）计算比较两组线性数据的相关度，如果相关度小于 0，则根据该值的大小发出不同等级的异常警报.

所有的轨迹通过系统自动选择及手动调整生成后，系统就开始计算该车辆的加油量和行驶距离，同时与历史数据进行对比，绘制出加油量和行驶轨迹总距离的关系如图 10 所示.在双 y 轴坐标系中使用堆叠柱形图和折线图，x 轴代表加油的次数，左 y 轴是行驶距离，右 y 轴是加油量.系统按照公式（1）

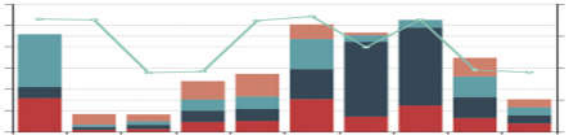


图 10　加油量与行驶距离的比较

Fig.10　Comparison of refueling number and travel distance

计算柱形图和折线图是否正相关,用户也可以直观地看出是否有数据异常的情况出现.

实际系统在运行过程中,对重点关注人群的加油轨迹实时分析监控,对一般人员的加油轨迹不作预先人工调整,该可视分析模块作为异常预警之后的分析确认使用.用户可以通过该模块直观地查看预警的异常发生原因,人工判断是否因为系统自动计算的轨迹出现较大偏差而导致异常的发生.每次人工调整轨迹的操作记录都会保留并作为学习样本,供轨迹计算自学习模块使用,从而不断地提升系统自动选择轨迹算法的准确性.

当系统按计划任务定时分析出当天的异常后就会发出预警信号,提醒相关人员来检查该异常是否为真.本文设计了基于极坐标的散点图来表现一个月内的所有异常分布,如图11所示.使用极坐标而不是笛卡尔坐标的原因是为了解决大量散点可能遮盖的问题.一个月中的每一天按极坐标轴均匀分布,每天的24小时均匀分布在每条极坐标轴上,所有的异常预警以散点方式分布在图中,灰度的深浅代表了异常发生的可信度大小,用c表示该值,c的取值范围在[0,1]之间,由系统选择轨迹的平均概率决定;点的大小代表了异常与预警规则的匹配度,用m表示该值,m的范围在[0,100]之间,由式(1)计算得到的r值和其余的异常预警规则计算结果共同决定.

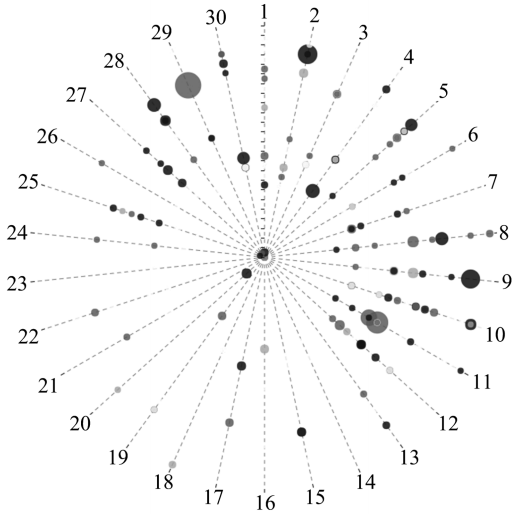


图11 异常预警图

Fig.11 Abnormal warning

3.5 小结

本节共设计使用了7种不同类型的可视化视图,在研究工作的不同阶段从多种维度展示数据的属性.表2总结了每个阶段需要展示的数据特征及对应视图的关系.

表2 各阶段数据特征对应可视化视图总结

Tab.2 The summary about the data features for visual views of each step

阶段	数据	数据特征或操作	可视化视图
直接可视化	个体加油情况总体历史数据	人与车的关系	平行坐标
		车与加油站的关系	平行坐标
		加油地点	平行坐标(名称对应) 地图(实际地理位置)
		加油时间	平行坐标(先后顺序) 日历矩阵(时间分布)
		加油量	平行坐标(坐标轴) 地图(散点大小) 日历矩阵(颜色深浅)
可视分析	个体单次加油后的行驶轨迹数据	OD行驶轨迹	地图(点、连线)
		OD距离和时间	柱形图
		OD路线选择	桑基图
异常处理	加油数据	加油量	折线图
	计算结果	行驶距离	堆叠柱形图
	异常数据	异常信息时间分布	极坐标图
		异常信息m、c值	散点图(颜色、大小)

通过表2可以看出,设计出不同的可视化视图是为了方便表现不同类型数据的多层次特征,或者是方便用户的交互式操作.其中直接可视化阶段使用的平行坐标视图虽然能够展示出全部数据特征,但从用户的角度看,数据的时空特性表现的还不够明显直观,使用地图和日历矩阵来补充能完整地展现时空数据多维、稀疏的特性.

4　案例分析

我们选择实际应用场景中的 3 个典型案例来对本系统的方法进行验证，采用的数据分别来自车辆加油记录数据、卡口数据和停车场数据.案例的车辆加油记录是从监管平台中获取的实际数据.由于卡口和停车场数据尚未完全接入，案例分析使用了仿真数据.加油记录的时间范围从 2015 年 7 月至 2016 年 6 月，车辆记录总数为 1 283 952，覆盖加油站 963 个，加油记录总数 6 307 595 条.

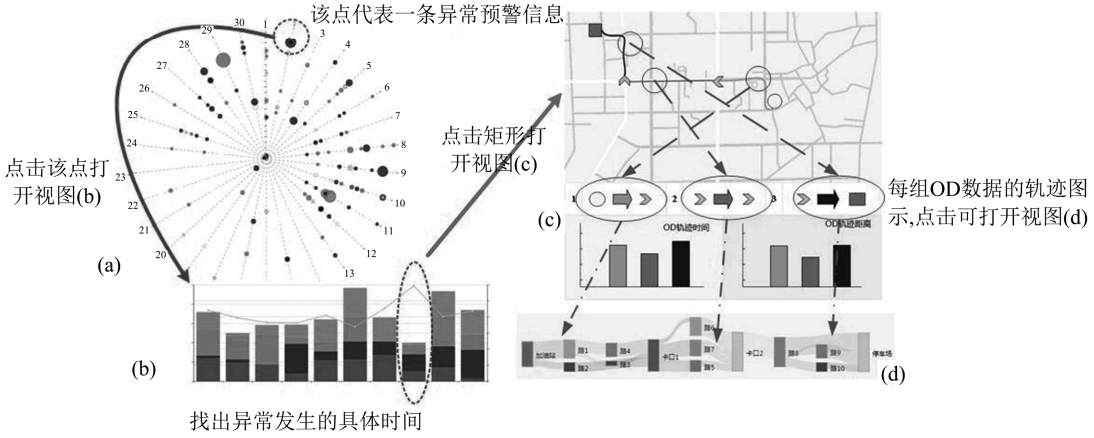
4.1　案例 1：异常预警与分析

本案例来源自某一天系统发布的异常预警信息，图 12(a)展示了一个月内所有的异常预警信息，案例中选择的信息对应绘制点较大，灰度较深，该异常的 c 值大小为 0.9，m 值大小为 82.5，发生时间为 2016 年 6 月 2 日 21 时.

图 12(b)展示了目标车辆最近 n 次(本案例中采用 $n=10$)的加油记录以及系统关联多个数据源之后自动匹配的行驶轨迹数据统计.图中柱形的堆叠块代表了目标车辆一天的行驶距离，折线上的数据点代表了上一次加油总量.通过图中的信息特征可以直观地看出，第 8 次的数据(即图中圈出的部分)存在异常，加油的数量呈上升趋势，行驶总距离却下降很多.异常发现的时间是第 9 次加油的时间，说明系统通过模拟发现了上一次的加油量同行驶总距离两者不是正相关的关系.

接下来需要通过轨迹分析工具详细分析第 8 次的行驶轨迹，系统自动生成的轨迹图如图 12(c)所示，这是第 8 次加油后第一天的 OD 数据集，可以看出车辆以加油站为起点，经过两个卡口后，终点是一个停车场.为防止系统误判，需要由人工调整目标车辆的轨迹路线.图 12(d)展示了在 3 组 OD 数据中所有可能的行驶路径，操作员通过点选来调整轨迹路线，并和行驶时间做比较.最终得到人工校对后的轨迹路线，然后再计算 r 值，发现校对后的轨迹对应的行驶总距离依然同加油量不是正相关的，于是判断该异常可能是实际发生的，目标车辆加了足够多的油却没有行驶相应的距离，可能存在从车中抽取油品的现象，需要报告给相关人员做下一步处理.

从以上案例的分析过程中可以看出，系统提供的异常检测方法能够发现现实存在的异常现象并及时预警，系统提供的可视分析模块也可以方便用户直观地发现异常并对异常数据来源进行进一步分析.



(a)包括异常预警图；(b)加油量与行驶距离对比图；(c)异常轨迹分析图；(d)轨迹选择调整图

图 12　异常预警与分析图示

Fig.12　Abnormal warning and analysis

4.2　案例 2：特定行为模式可视化展示

我们从车辆数据库中选择一辆出租车来观察相关的加油记录特征，目的是要通过可视化方法全面展示这种特定类型的车辆的典型行为模式.

从图 13(a)平行坐标的曲线分布可以直观地看出，目标车辆对应的驾驶员有两个，该车只在一个地区加过油，经常加油的站点有 5 个，每次的加油量比较稳定，加油时间分布比较均匀，再结合图 13(b)和图 13(c)，可以更加准确地观察到该车加油的时空分布特征.我们可以得到的信息有：这辆出租车的驾驶员有两人，符合目前部分同一出租车拥有两名司机，分早晚班互相倒班来驾驶车辆的现状；该车加油次数非常频繁，基本每天都会去加油，从图 13(b)的方格灰度分布上可以判断出该车的运营状况，还可

以看到该车的加油记录只有最近 14 周的,在查看系统记录后发现出现这种现象是有原因的,因为之前该车所在地的加油站还没有接入监管平台.图 13(c)展示出了该车在固定 5 个加油站之间行驶的趋向,说明该车基本都在这一个地区活动,没有跑过长途,该车在 5 个加油站行驶的趋向路线图是成闭环的,行驶趋向线上箭头的数量代表了行驶至目标加油站的次数,可以看出该车在城南的活动次数要多于城市的其余位置.从图 13(d)中展示的 20 次加油量和行驶距离对比图来看,该车的加油量和行驶距离呈现出明显的正相关,说明该车出现异常的可能性比较小.

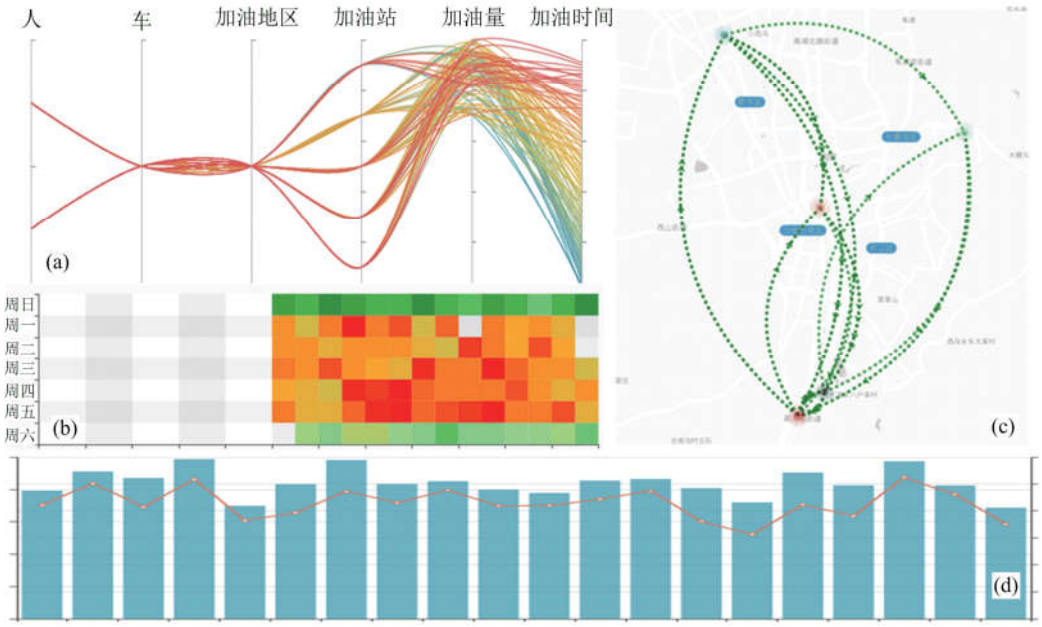


图 13 出租车加油记录可视化视图

Fig.13 The visualization of taxi refilling records

4.3 案例 3:群体行为模式分析

我们从数据库中随机选取了多个样本,期望能够发现一些能够适用多数人群的一些数据特征.我们采用最简单的多快照可视化方法:小型多重方格图(small multiple)的方式来查看是否有相似的行为特征和差异.

图 14 是一部分样本的数据特征可视化视图,我们预先对不同种类的车辆进行分类,相同类型的视图聚合在一起可以更好地辨别数据特征,图 14 中展示了 3 种类型的车,分别为私家汽车、摩托车、公务车.为方便区分,我们对平行坐标视图做了一些调整,使用灰度代表车辆类型,其中图(a)至(f)中深灰曲线代表摩托车加油记录,浅灰曲线代表汽车加油记录,图(j)至(l)中是汽车加油的记录.为了能够快速准确地获取数据特征,这里只展示平行坐标图和日历矩阵图.

观察图 14 中的样本数据,我们可以得到一些比较明显的数据.例如图 14(c)和(d)的平行坐标深灰曲线比较类似,都记录了多个驾驶员驾驶该摩托车去加过油,加油记录也多分布在距今比较近的时间段内.图 14(g),(j),(k)中的公务车都是属于工作比较繁忙的类型,区别在于忙碌的时间段不一样,(g)中的车辆半年中都比较忙,(j)中的车辆只在中间两三个月比较忙,(k)中的车辆则是最近很忙.

从可视化视图中能够提取数据特征,还可以验证一些符合人们日常经验判断的典型行为,具体如下:

(Ⅰ)大部分人的习惯是在同一地区(基本是在居住地附近)的一个或几个加油站加油.

(Ⅱ)公务车的加油记录在工作日分布较多,且加油频率较高,私家车的加油记录分布比较平均,频率较低.

(Ⅲ)汽车加油频率较低,每次加油数量较多;摩托车加油频率较高,每次加油数量较少.

(Ⅳ)多数人只有一辆车的加油记录,某些特别关注人群(例如在逃人员)则有多辆车的加油记录,而部分摩托车则有多位驾驶员加油的记录.

上述行为特征说明,这些行为是典型的正常行为,如果该群体中的某个个体行为与这些行为的偏

差过大，则认为该个体的行为是异常的.

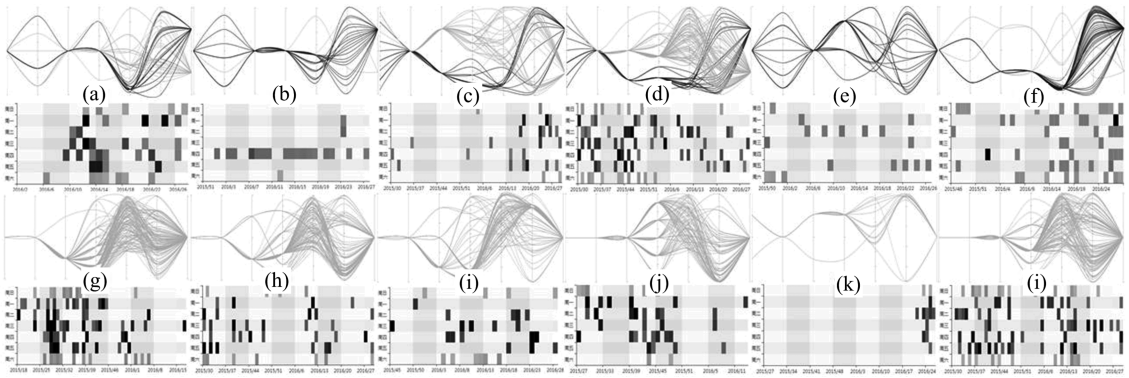


图 14　部分样本数据特征图

Fig.14　Data features of some samples

5　结论

本文提出了使用数据可视化分析方法来分析具有多维稀疏时空属性的车辆加油记录数据的方案.为了更好的发现、分析、总结多维稀疏时空数据的各种特征，我们采用多模态数据可视化方法，即使用多种互相关联的可视化视图相结合的方法解决数据属性维度众多、特征不明显的问题，使用多个相关数据源的融合方法解决由于数据稀疏造成的各种问题，设计轨迹生成及可视化方法来完整地展现数据的时空属性，并能够直观地发现相关特征.文中共设计使用 7 种不同类型的可视化视图从不同层次全面展示数据的属性，提出了一种面向个体 OD 数据的车辆行驶轨迹自动生成及可视化的方法，并依据生成的车辆行驶轨迹计算该车辆的行为是否异常.文章最后列举了 3 种典型案例，对文中提出的方法在实际场景中的应用加以说明，并演示了可视化方法的实用性和有效性.本文提出的分析方法基于实际加油站产生的车辆加油记录数据集，也可以用于其他系统所产生的多维稀疏时空数据集的分析，如智能交通站、加气站、公交站、公共自行车系统等.

本文所述研究基于当前项目所涉及到的数据集，随着项目的进一步延伸，更多具有不同时空特性的数据将接入系统，为进一步的群体行为模式的特征分析提供数据基础.在下一步的工作中，我们将结合不断完备的数据源，如汽车里程表数据、公共交通车辆 GPS 数据等，验证我们的数据模型，设计新的可视化分析方法，将分析对象从个体转为群体，特别是针对小团体共同行为模式的发现和识别，如人群的流动方向和行为模式、关注人群流动目的地的预测、具有相同特征和行为模式的人群划分等，将可视化分析与其他数据分析方法结合（如人员关系图谱分析），从而实现对群体关系和行为的分析和异常识别.

参考文献（References）

[1] KEIM D, ANDRIENKO G, FEKETE J D, et al. Visual Analytics: Defi-Nition, Process, and Challenge [M] //Lecture Notes in Computer Science. Heidelberg: Springer, 2008, 4950: 154-175.

[2] 姜晓睿，郑春益，蒋莉，等. 大规模出租车起止点数据可视分析[J]. 计算机辅助设计与图形学学报，2015，27(10)：1907-1917.
JIANGXiaorui, ZHWNG Chunyi, JIANG Li, et al. Visual analysis of large taxi origin-destination data[J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(10): 1907-1917.

[3] WANG Z C, YE T Z, LU M, et al. Visual exploration of sparse traffic trajectory data[J]. IEEE Transactions on Visualization & Computer Graphics, 2014, 20(12): 1813-1822.

[4] DEKA L, QUDDUS M. Trip-based weighted trajectory matching algorithm for sparse GPS data [R]. Transportation Research Board Annual Meeting, http://transp-or. epfl. ch/heart/2014/abstracts/268. pdf,2015.

[5] SANAULLAH I, QUDDUS M, ENOCH M P. Developing travel time estimation methods using sparse GPS data [J]. Journal of Intelligent Transportation Systems, 2016, 20(6): 532-544.

[6] WANG Y, ZHENG Y, XUE Y. Travel time estimation of a path using sparse trajectories [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2014: 25-34.

[7] 任磊，杜一，马帅，等. 大数据可视分析综述[J]. 软件学报，2014，25(9)：1909-1936.

[8] HEY T，GANNON D，PINKELMAN J. The future of data-intensive science[J]. Computer，2012，45(5)：81-82.

[9] PEUQUET D J，KRAAK M J. Geobrowsing：Creative thinking and knowledge discovery using geographic visualization[J]. Information Visualization，2002，1(1)：80-91.

[10] SLINGSBY A，DYKES J，WOOD J. Exploring uncertainty in geodemographics with interactive graphics[J]. IEEE Transactions on Visualization and Computer Graphics，2011，17(12)：2545-2554.

[11] INSELBERG A. The plane with parallel coordinates [J]. The Visual Computer，1985，1(2)：69-91.

[12] VON LANDESBERGER T，BRODKORB F，ROSKOSCH P，et al. MobilityGraphs：Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering[J]. IEEE Transactions on Visualization & Computer Graphics，2016，22(1)：11-20.

[13] ANDRIENKO G L，ANDRIENKO N V，DYKES J，et al. Geovisualization of dynamics，movement and change：Key issues and developing approaches in visualization research[J]. Information Visualization，2008，7(3)：173-180.

[14] LUNDBLAD P，EURENIUS O，HELDRING T. Interactive visualization of weather and ship data[C]// Proceedings of the 13th International Conference on Information Visualisation. Barcelona，Spain：IEEE Press，2009：379-386.

[15] OpenDataCity. Visitor flow analysis by public wireless [EB/OL]. http://apps.opendatacity.de/relog/，2013.

[16] THUDT A，BAUR D，CARPENDALE S. Visits：A spatiotemporal visualization of location histories[J]. Journal of Heat Transfer，2013，114(1)：255-163.

[17] TOMINSKI C，SCHUMANN H，ANDRIENKO G，et al. Stacking-based visualization of trajectory attribute data[J]. IEEE Transactions on Visualization Computer Graphics，2012，18(12)：2565-2574.

[18] KAPLER T，WRIGHT W. GeoTime information visualization [C]// Proceedings of the IEEE Symposium on Information Visualization. Austin，USA：IEEE Press，2004：25-32.

[19] STOLL M，KRÜGER R，ERTL T，et al. Racecar tracking and its visualization using sparse data[C/OL]// Proceedings of the Workshop on Sports Data Visualization，http://www.vis.uni-stuttgart.de/~cvis/publications/stoll_sportvis2013.pdf.

[20] GUO H Q，WANG Z C，YU B W，et al. TripVista：Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection[C]// Proceedings of Pacific Visualization Symposium. Los Alamitos：IEEE Computer Society，2011：163-170.

[21] BYRON L，WATTENBERG M. Stacked graphs-geometry & aesthetics[J]. IEEE Transactions on Visualization and Computer Graphics，2008，14(6)：1245-1252.

[22] WILLEMS N，VAN DE WETERING H，VAN WIJK J J. Visualization of vessel movements[J]. Computer Graph Forum，2009，28(3)：959-966.

[23] WANG Z C，GUO H Q，YUAN X R，et al. Discovery exhibition：Visual analysis on traffic trajectory data [EB/OL]. [2014-09-24]. http://discovery exhibition.org/uploads/Main/2011Wang.pdf.

[24] LIU H，GAO Y，LU L，et al. Visual analysis of route diversity[C]// Proceedings of the IEEE Conference on Visual Analytics Science and Technology. Los Alamitos：IEEE Computer Society，2011：171-180.

[25] WANG Z C，LU M，YUAN X R，et al. Visual traffic jam analysis based on trajectory data[J]. IEEE Transactions on Visualization and Computer Graphics，2013，19(12)：2159-2168.

[26] KRÜGER R，THOM D，WÖRNER M，et al. TrajectoryLenses—A set-based filtering and exploration technique for long-term trajectory data[J]. Computer Graphics Forum，2013，32(3/4)：451-460.

[27] PU J S，LIU S Y，DING Y，et al. T-Watcher：A new visual analytic system for effective traffic surveillance [C]// Proceedings IEEE14th International Conference on Mobile Data Management. Los Alamitos：IEEE Computer Society，2013，1：127-136.

[28] RIEHMANN P，HANFLER M，FROEHLICH B. Interactive sankey diagrams[C/OL]. Proceedings of the IEEE Symposium on Information Visualization. Minneapolis，USA：IEEE Press，2005. https://static.aminer.org/pdf/PDF/000/404/817/interactive_sankey_diagrams.pdf.

[29] SZMIDT E，KACPRZYK J. On an enhanced method for a more meaningful Pearson's correlation coefficient between intuitionistic fuzzy sets[C]// Proceedings of the 11th International Conference on Artificial Intelligence & Soft Computing. Zakopane，Poland：Springer-Verlag，2012：334-341.