城市移动数据知微探秘

陆 旻 王祖超 袁晓如 等
北京大学

关键词：城市 移动数据 轨迹可视分析

引言

城市也是一套复杂且动态变化的系统，人们的行为构成了城市的血脉。研究城市演变的目的在于研究人的行为。《科学》(*Science*) 杂志近期推出了一组专栏文章“城市星球——城市是人类的未来”[1]，其中收录的文章探讨了城市化进程中的各种因子、现状、结果与期望。其核心思想是研究人，研究个体或群体的行为活动是如何影响城市的，同时每个个体又如何受到城市的影响。以关乎人类未来为题，研究城市的意义不言而喻。不同区域、不同人群在城市中有着自己运行的规律。随着传感器技术的发展与大数据时代的到来，人群的移动逐渐被各种数据源所记录。然而，随着数据的爆发式增长，挑战也随之而来。采用何种方式让人们理解城市这一复杂的主题，进而探索城市中不同的人群行为如何对城市造成影响？可视化与可视分析技术通过将数据转换为图形的方式，让城市的脉动跃然屏中。用户可以探索不同时空尺度下的城市数据，从而挖掘出城市数据背后蕴含的移动规律。

城市移动数据

城市如同生命，有自己的脉动规律。脉动的背后是人群移动的规律。市民早出晚归让城市的数据特征如同灯光一般均匀地“呼吸”。对移动特征的探索可以揭示城市运行规律。这一规律，从微观的一个街区、到一条路径、再到一个城市，都有其不同的特性。

另一方面，城市中的每个个体也被城市的运行所影响。例如不法分子利用伪造基站发送广告、诈骗短信，普通人却无法了解他们的真实行踪。但通过可视分析方法，可以将原本离散的轨迹采样数据进行重建，从而挖掘出恶意伪基站的行为轨迹，为公安部门打击犯罪提供有利的线索。通过结合更多的城市数据源，包括那些原本稀疏且不被认为是“轨迹”的移动采样数据，我们能更深入地了解城市运行的规律，从而对城市的脉动进行感知和预判。

数据采样按照稀疏程度可分为密集采样（全球定位系统 (GPS) 车辆轨迹）、较密集采样（信号基站轨迹）、稀疏采样（射频与摄像头探测数据）和不定采样（社交媒体移动轨迹）。

城市数据的深度探索——以出租车轨迹为例

我们使用北京市 24 天的出租车 GPS 轨迹数据，大小为 34.5GB，包括 28519 辆车、3.79 亿个采样点，每 30 秒采样一次（但有大量采样点缺失）。该数据包含了北京市 43% 的出租车，大约对应北京市城区 7% 的交通流量，因此根据该数据可以大致估算出北京的交通流 [2]。基于此，我们在微观的路口交互选择、单个路段的时变探索以及城市中任意一条乃至多条路径进行深入分析，描述出人们出行的规律，找出相关的因素，进一步了解了城市整体的运行规律。

城市路口的交互选择

对城市的探索，需要从用户的细节出发，比如对区域路口的选择。可视化的方式能让用户直观地感知并且操作轨迹数据，基于此还可以建立起用户和数据之间的桥梁。区域的选择是用户进一步探索路段、路径以及多条路径的基础，也赋予用户探索城市不同区域、不同空间层面的灵活性[3]。

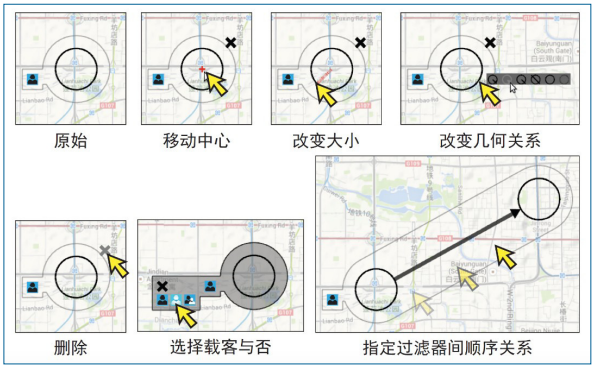


图1　用户对城市路口区域的交互选择探索设计

具体而言，用户可以在城市的不同区域放置过滤器（如图 1 所示），原始的过滤器是一个圆形的区域，用户可以直接拖拽中心进行移动，拖拽边缘来改变大小，并且可以设置多个过滤器。过滤器的逻辑操作包括交、并、差等，可有起始、终点或经过等模式。用户还可以选择经过该区域的出租车是否载客等高维属性。经过对起点、终点以及进一步的路段、路径与多条路径的分析，我们希望能从大量的轨迹中方便地筛选出特定的轨迹，以便比较它们的时间分布、属性区别和路径选择。

城市路段的时变特征探索

基于以上的路口区域选择，我们进一步探索了某个路段的特征。在一条路段中车来车往，我们将其周期性的行为进行整合，并用设计密度图的方式进行可视化[2]。图 2 表示选出的路段在一段时间（24 天）的交通状况，其中每行代表一天，每列代表 10 分钟。通过将每天的交通状况排列显示，用户可以直观地感受到早晚高峰以及一些异常事件的发生。

单条路径的时间花销稳定性分析

基于上述两方面的探索，我们对单条选定起始 - 终点的路径进行了细节分析，目的是帮助用户了解路上的拥堵情况、拥堵发生路段、潮汐规律以及不同路段的时间花销比例等。

我们设计了一个探索单条路径不同路段需要花费时间的排名与变化的可视化系统[3]。

以北京北四环的交通为例，如图 3 所示，每条竖直的轴表示一个路段，由左往右按照道路上的顺序排列。在每根轴上，矩形块表示一个通行时间的类，从上往下按升序排列。轨迹表示为连接轴间的折线。折线的颜色表示其所属类别。通过颜色的分布，我们能够总览道路的时间花销情况。另外，为了避免产生视觉混淆，我们将属于相同类的轨迹进行合并，获得聚集后的条带。按照路的岔口，将路径分为 10 个路段进行分析。所有路段的平均通行时间相似，但发现东边前 2 个路段的通行时间波动较大。我们将历史轨迹数据按路段进行积分排名后，挖掘出 8 种通行模式（颜色区分）。该 8 种通行模式随出发时间的分布也不同，不良通行模式多分布在工作日的早高峰期间，傍晚只

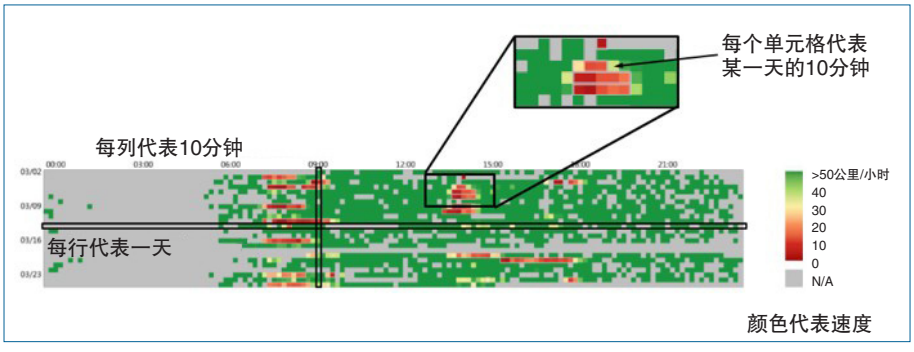
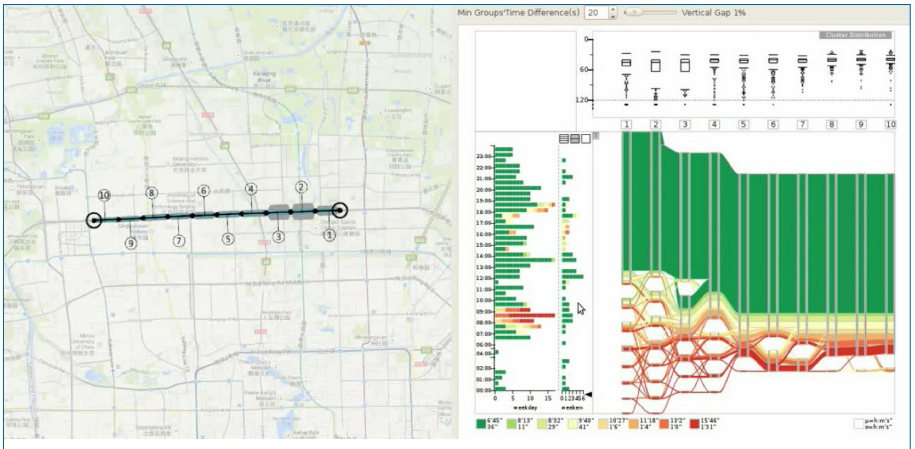


图2　对路段时变特征可视化



(a) (b)

(a)路径的空间分布，时间花销的变化程度以路径上的灰色宽度表示
(b)对单条路径不同路段的时间花销排名可视化

图3 对单条路径的细节分析

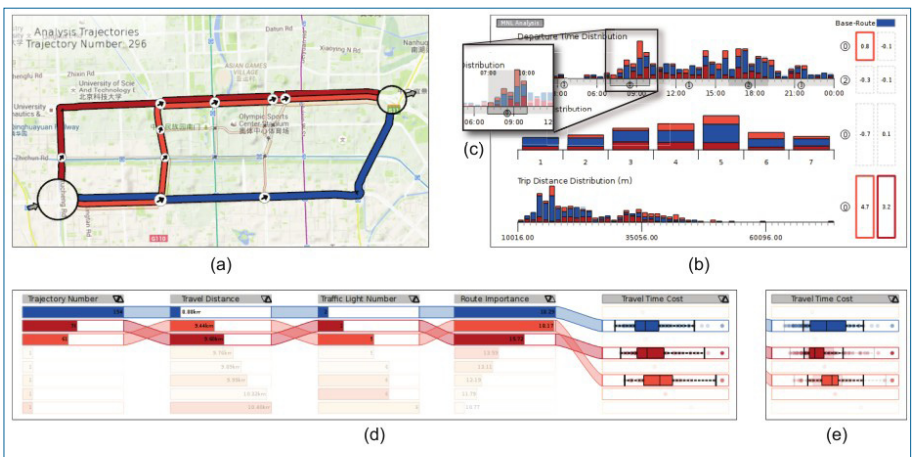
有少许出现。因此，我们建议经常出行北四环的市民，尽量在早高峰避开自东向西的路线，尤其是鸟巢与北辰路附近。

多条路径选择因素探究

在实际应用中，有哪些因素影响人们的移动行为？是红绿灯数量、路径长度，抑或环路？它们的影响和不同时间段是否有关系？每个影响因子是否起了决定性作用？带着这些问题，我们设计了一套对多条路径选择因素探究的可视分析系统（图4）[4]。首先根据基于网格的路径提取算法，得到两个区域间的可达道路，并构建道路拓扑结构图。基于该拓扑结构，计算道路的编辑距离与道路之间的相似度，并对其进行聚类，从而获得不同的道路类并赋予它们不同的颜色。同时我们计算每一条提取出来的道路的属性，包括道路车流量、道路长度、红绿灯数、道路的重要程度、道路通行时间的分布等。为了研究选择因子的影响在当前选择条件下是否显著，我们同时支持交互式配置多项选择模型的输入，使用多项式选择模式对轨迹因素对路径选择的影响进行建模，计算出P值。

还是以北四环为例，我们这次关心的是从北三环到北四环的多路径的选择行为。首先，将过滤器放置在其上的两个路口，其中有三条主要被选择的路径。对比这三条道路的多种属性，包括路程、红绿灯数、道路等级等，可以发现被选择最多的道路（蓝色）具有较短的路程、红绿灯数也较少。同时，在总时间花销的分布中也可以看到，蓝色路径的平均通行时间最少，方差最小。

我们进一步研究了不同时间段影响因子的变化。根据出行时间不同，观察路径选择的不同情况。



(a)空间视图 (b)时间花销分布 (c)时间选择 (d)各影响因子 (e)高亮影响因子

图4 多条路径选择

可以发现在早晨 9 点左右，橘色路径的选择量有所增加，由此提出了此段时间会影响司机对路径的选择的假设。基于离散选择模型，将此段时间作为一个因素，通过验证发现该假设成立，早晨 9 点左右，司机在北辰路鸟巢附近提前拐入四环的概率会有显著增加。

城市数据的广度探索——不同采样粒度的轨迹可视分析

不同的数据源反映出城市运行的不同方面。对连续的 GPS 轨迹的时变特征与路径选择的研究有助于我们了解城市的交通状况。而经常被大众和分析人员忽略的是那些看似稀疏的零散轨迹采样，不同的稀疏特征反映了城市移动不同方面的特征。其实，只要设计合理的分析方法，就可以重构出原有的实际轨迹，这些轨迹蕴含的信息往往对我们了解城市十分重要。

城市伪基站轨迹揭秘

手机移动伪基站的可视分析就是一个典型的对稀疏轨迹重建、探索的例子。犯罪者通常把伪基站放置在车内，扩大伪基站的影响范围，同时也极大地增加了追捕难度。追踪移动伪基站的位置，重构其移动轨迹，能够帮助相关法律部门抓住罪犯，从根源上杜绝此类问题的发生。

然而，定位移动伪基站并非易事。由于无法直接获得移动伪基站的 GPS 定位信息，我们不得不从受害者手机中收集间接信息。间接信息包括恶意短信的记录和手机连接基站的日志，系统通过结合真实基站的位置信息来重构轨迹。然而间接信息给原本就稀疏的轨迹带来了极大的时空不确定性。在时间上，受害人手机收到恶意短信的时间与其最后一次连接真实基站的时间存在偏移，从几分钟到几小时不等。在空间上（如图 5 所示），通常在两公里范围内，受害手机以及伪基站的移动也会增加不确定性。我们的系统结合自动算法以及人工分析，通过交互的方式重构手机移动伪基站的轨迹，包括重构恶意短信的传播轨迹和合并恶意短信两个步骤。

首先根据恶意短信的文本内容对用户上报的所有记录分组。具有完全相同文本内容的短信很可能来自同一个手机移动伪基站。对于每一条恶意短信，系统通过时空聚类的方法提取出短信传播的主要路径。之后使用滑动窗口对原始记录进行二次采样，来平滑轨迹。根据时空采样的密度，分别计算出重构的轨迹中不确定性较高和不确定性较低的部分。如图 5(b) 所示，是一条关于代开发票的恶意短信的传播轨迹，其中实线表示不确定性较低，虚线表示不确定性较高。

一个移动伪基站可能同时发送多条不同内容的恶意短信，为了有效地合并来自同一伪基站的恶意短信，系统从文本相似性和时空相似性两方面进行分析，给用户推荐潜在能合并的恶意短信。针对一条色情类的恶意短信（图 6 中的 4 号信息），我们分

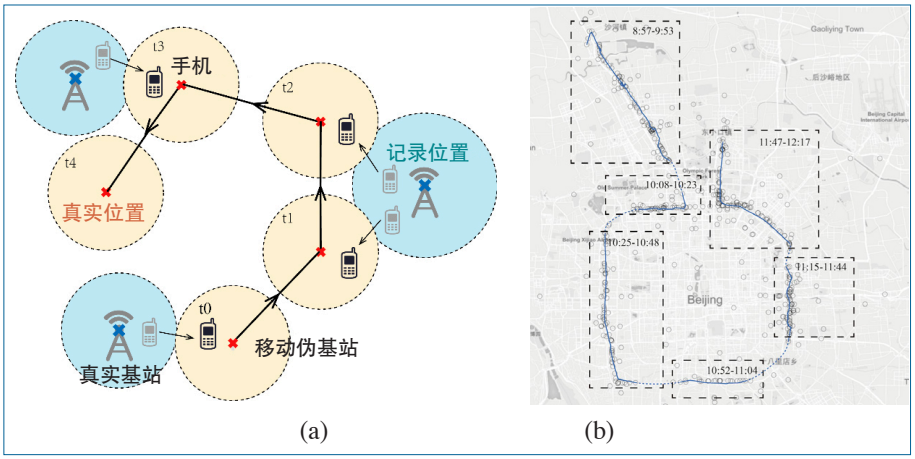
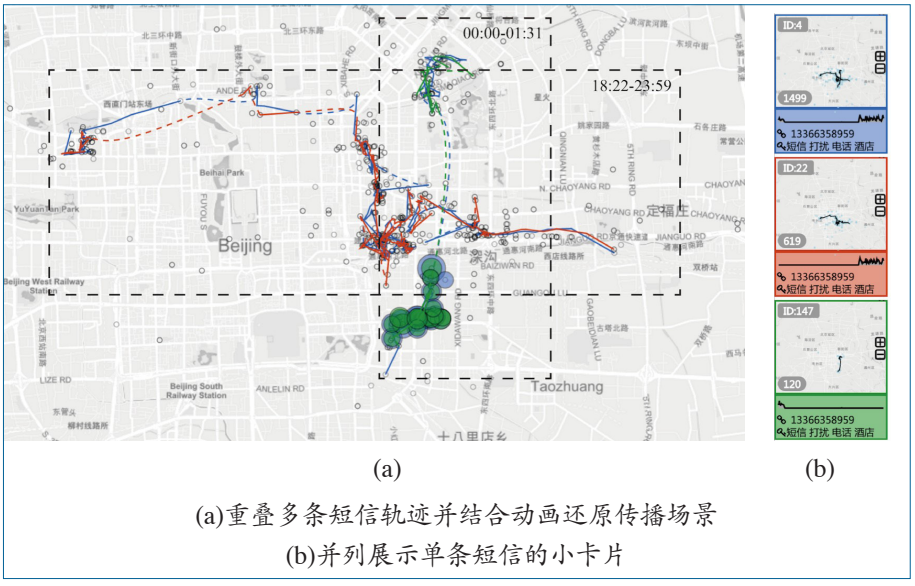


图5 (a)手机移动伪基站数据记录了手机最后一次连接真实基站的位置，数据较为稀疏并且带有极大的不确定性
(b)在四环上散播代开发票短信的手机移动伪基站

析了系统推荐的另外两条恶意短信（即22号和147号），发现它们在内容上非常相似。将这些短信以并列卡片的形式展示（见图6(b)），便于用户快速对比。针对这些信息，我们比较了其传播的轨迹，发现它们都在深夜或凌晨活跃在北京三里屯一带，路径非常相似。通过分析文本和时空的相似性，我们确定是由同一个伪基站发出了多条短信，最终成功还原出该伪基站的活动模式。



(a) (b)

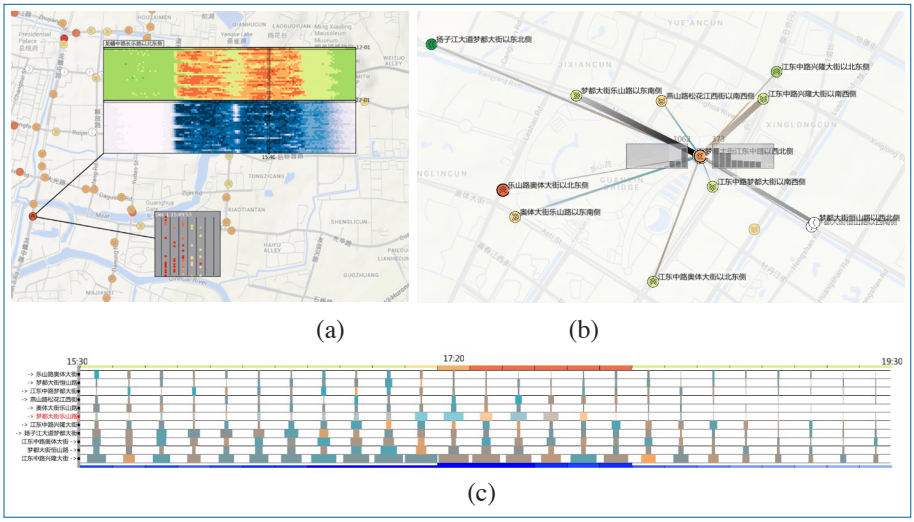
(a)重叠多条短信轨迹并结合动画还原传播场景
(b)并列展示单条短信的小卡片

图6 在三里屯附近散播多条色情短信的移动伪基站

城市智能交通基站探索城市运行

对伪基站的重建是基于移动基站，移动基站在城市中相对密集，可以通过多点定位来确定最近的位置，进而重建出其轨迹。然而，对于稀疏轨迹（比如智能交通基站分布稀疏），需要探索一套完全不同的分析方法。

智能交通基站是智能交通系统中的重要组成单元。它使用射频识别传感器技术、摄像头视频检测技术以及数据通信技术，来收集道路上的车辆通行记录。每个基站监控单一方向的多个车道上的车辆通行记录。每个车道上都安置了摄像头和射频识别装置，用于检测车辆经过基站时的车牌号、速度、通过时间等。智能交通基站对单辆车的轨迹采样是十分稀疏的，车辆只有在经过基站时才有轨迹记录，而在相邻两个基站之间的轨迹则无法得到。以南京市为例，整个南京市有472个智能交通基站，相邻两个基站之间的距离通常在1公里以上。而智能交通基站对城市内的车辆采样却十分稠密，数据记录了城市主要道路上几乎所有车辆的轨迹。南京市的智能交通基站每天能记录超过100万辆车，涵盖了90%以上的活跃车辆，包括货车、出租车、私家车等多种车辆类型。因此，此类数据能很好地反应整个



(a) (b)

(c)

图7 (a)单个智能交通基站的交通状况分析 (b)和(c)局部区域的交通状况分析

城市的交通状况，非常适合做宏观交通分析。

我们从单个智能交通基站、局部区域以及整个城市三个层次分析了南京市的交通情况[5]。

单个智能交通基站 我们分析该基站的交通情况，以每条道路一个月的交通情况为例，其模式如图 7(a) 所示，所呈现的方式如同一个表格。其中每一行代表 1 天，144 列表示一天的 144 个 10 分钟。每个单元格表示某一天的某个 10 分钟时间段，颜色表示统计的结果。上方的表格表示 10 分钟内所有通过该智能交通基站车辆的平均速度，从红色到绿色表示速度由低到高，而下方的表格则表示 10 分钟内的交通流大小，颜色越深表示交通流越大。我们通过动画还原了车辆通过该智能交通基站的场景，可以看到除了最右侧的非机动车道外，其他车道的负载都比较重，并且每辆车的通行速度都比较慢，由此可以确定该拥堵事件是因早晚高峰交通流量过大造成的。城市规划专家据此可以考虑提高附近道路的通行能力来缓解拥堵。

局部区域的交通状况 选定一个中心基站，如图 7(b) 所示，系统通过统计以及排序的方式给出与中心基站最相关的 10 个上游基站（左侧）和 10 个下游基站（右侧）。用户通过交互的方式筛选并过滤出相关的基站，来对交通流进行时序分析。如图 7(c) 所示，过滤出的上下游基站和中心基站的交通流在时间轴上展开。时间轴的每一行代表与中心基站相关的一条链路，以 10 分钟为粒度统计该链路的交通流，以矩形的方式映射到时间轴上，矩形的宽度代表流量，颜色则表示流量的变化趋势。为了易于分析链路流量的变化和中心基站状态变化的相关性，我们在时间轴的顶部和底部分别显示了通过中心基站车辆的平均速度和交通流量随时间的变化情况。

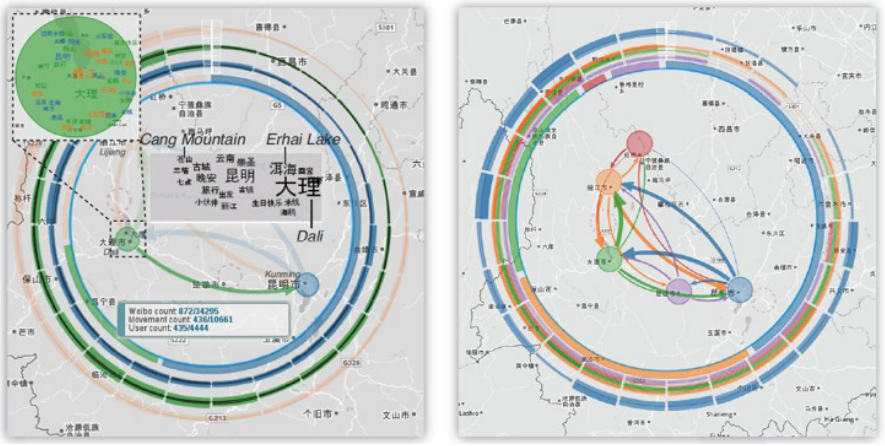
整个城市的交通分析 我们将整个城市的所有智能交通基站看成一张网络，每个智能交通基站是网络中的节点，基站之间的交通流是网络的边。通过分析图的变化情况来查看城市的交通情况。

社交媒体移动轨迹可视分析

除了传感器记录的数据，还有一大类城市数据是城市中的人贡献的——带有地理位置的社交媒体数据，此类数据的采样稀疏性是根据不同的人在网络上发布信息的习惯而变化的。该类数据不仅覆盖面广，而且覆盖的人群数目也很庞大。据统计，2014 年平均每天约有 100 万条带有地理标签的微博产生。由于其包含了时间、空间、文本等丰富的信息，因此我们可以从中挖掘出大量重要且有特征的行为模式。以新浪微博为例，将用户带有地理信息的微博按照时间顺序连接起来，就可以构造出用户在实际物理空间中的稀疏轨迹。通过合理的可视化设计方案，可以构造出每个社交媒体用户带有的明显个人特征的轨迹，例如旅行爱好者、商务白领、学者等，每个人的轨迹不尽相同。这些轨迹的每个采样往往都含有时间、文本、图片等丰富的信息，可以讲述出一个个精彩的故事。我们的可视化系统允许用户探索移动的群体行为，通过交互的方式发现群体行为中空间、时间以及多维属性上的规律[6]。

我们的研究针对的是群体行为，因为单体的不确定性可能是由某种随机因素决定，而群体的行为往往能反映其背后的移动规律。我们针对两地之间的移动进行建模，并认为多种交通方式（例如飞机、火车、汽车等）并存。在交通领域的研究中，某种具体交通方式的时间花销被认为呈正态分布。而且，不同的交通方式产生的时间花销分布不同。由此，我们提出了基于高斯混合模型的不确定性建模方法，通过对移动数据花费时间的建模，可以找到两地之间移动的不同类型，并结合微博的语义（关键词）进行分析，来判断其使用的交通方式，以及计算每种方式的时间花销均值和可信区间。但仅仅有模型还是不够的，用户在之后的可视分析流程中，还要对输入的数据进行过滤，或对模型的参数（k 个类型）进行控制，并且选择不同类型以及置信区间做进一步的可信数据探索。我们这项研究的一大特色是将不确定性模型引入可视分析系统中，允许用户探索稀疏采样的交互轨迹数据。

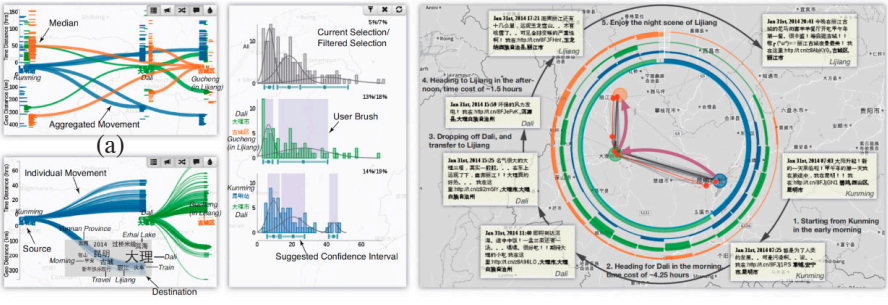
我们对云南省内带有地理标签的数据进行了分析。首先，用户会面对一个时空过滤视图，上



(a) (b)

(a)兴趣点(带颜色小圆圈)、流量(箭头)、关键词分布以及时间与数量分布(外围)
(b)用户可以增加POI区域，观察各个流量分布，并且外围的时间视图是细节级别(level of detail)分布的

图8 空间视图与过滤兴趣点



(a) (b) (c) (d)

图9 (a)聚集的人群移动分布，(b)具体人群移动分布与关键词分布，(c)模型视图，包含了两两兴趣点之间的移动时间花销的分布，以及计算出的不确定性模型与置信区间。(d)频繁模式与具体代表案例，从昆明→大理→丽江

面的轨迹显示了原始的微博移动数据，用户可以选择一个时空区域进行细节分析。我们提供了基于密度的方法，可检测分析区域中微博较多的点。在这个例子中，我们找出大理、丽江、昆明三个城市,通过外环的圆圈看到其数量比例大概是1:1:2(图8)。用户可以利用交互方式来调节、增加、删除、缩放相应的分析区域，数量外环之外展示了对应地区(颜色)的时间分布规律。此外，用户还可以通过筛选方式选出早上从昆明出发、下午到达大理的人群，或点击相应地点的圆圈，观察其关键词的分布。

在筛选出相应的兴趣点(Point of Interest, POI)位置之后，兴趣点之间的相互流量会被自动过滤出来，图9中的箭头粗细表示流量。鼠标悬浮在上面时会看到细节的流量分布信息，在筛选出兴趣点以及相应的移动序列之后，用户可以看到相应的非空间属性的变换。系统还提供了ST-Matrix视图，其中横轴是移动距离，纵轴是花费时间的一个二维直方图，颜色的深浅代表微博数量的多少。用户可以在上面进行刷选，也可以选择多个属性区域空间做进一步分析，例如我们可以删去大于100小时的移动，或者超过理论上往返距离的数据。在这个例子中，我们选择筛选距离大于600公里的数据。

结语

我们从深度和广度两个层面研究了城市数据，探索了不同层次的分析对象以及不同类型的稀疏城市轨迹。许多城市数据原本包含的移动特征被其稀疏性掩盖，通过不同层次的可视分析方法，可以抽取其中关键的特征，发现整体性的规律。通过城市数据，可视分析为用户架起了一座直观理解城市、分析城市、最终达到感知城市的桥梁。■

陆　旻

CCF专业会员。北京大学博士生。主要研究方向为可视化与可视分析等。
lumin.vis@gmail.com

王祖超

CCF专业会员。奇虎360公司。北京大学博士。主要研究方向为时空轨迹、社会媒体可视分析。
zuchao.wang@gmail.com

袁晓如

CCF理事、本刊编委。北京大学研究员。主要研究方向为可视化与可视分析等。xiaoru.yuan@pku.edu.cn

其他作者：陈思明　叶唐陟

参考文献

[1] Wigginton N S, Fahrenkamp-Uppenbrink J, Wible B, et al. Cities are the Future[J]. *Science*, 2016, 352(6288):904-905.

[2] Wang Z, Lu M, Yuan X, et al. Visual traffic jam analysis based on trajectory data[J]. *IEEE Transactions on Visualization & Computer Graphics*, 2013, 19(12):2159-2168.

[3] Lu M, Wang Z, Yuan X. TrajRank: Exploring travel behaviour on a route by trajectory ranking[C]// *Proceedings of IEEE Pacific Visualization Symposium*. IEEE, 2015:311-318.

[4] Lu M, Lai C, Ye T, et al. Visual analysis of route choice behaviour based on GPS trajectories[C]//*Proceedings of Visual Analytics Science and Technology*. IEEE, 2015:203-204..

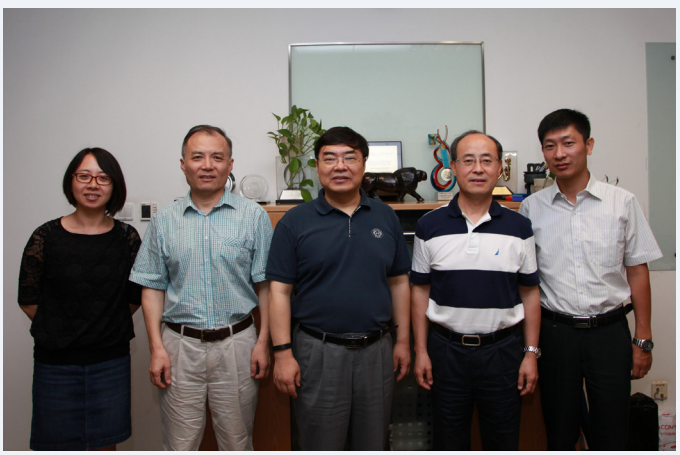
[5] Wang Z, Ye T, Lu M, et al. Visual Exploration of Sparse Traffic Trajectory Data[J]. *IEEE Transactions on Visualization & Computer Graphics*, 2014, 20(12):1813-1822.

更多参考文献：www.ccf.org.cn/cccf

CCF职业教育委员会筹备工作全面展开

2016 年 7 月 20 日，CCF 职业教育发展委员会第一次筹备会议在 CCF 总部举行，筹备委员会主任陈钟、委员温涛、秘书董本清、CCF 会员部部长戴丽霞参加会议，CCF 秘书长杜子德作为筹备委员会委员出席会议。筹备委员会就职业教育发展发展委员会的定位、组织机构设置、活动的组织、财务制度以及成立时间、工作计划等事宜进行了热烈讨论。

CCF 职业教育委员会筹备工作目前已经全面展开，将在 CNCC 同期举行职业教育发展委员会成立大会。职业教育发展委员会既不同于专业委员会，也有别于工作委员会，要联合职业教育院校代表、专家、企业等群体共同开展工作。CCF 职业教育发展委员会运行过程中既要借鉴各个专业委员会和工作委员会实践中取得的良好经验，也要成为改革的试验区，各项工作组织运作均要相对独立。



CCF 职业教育发展委员会将紧密围绕自身定位，以协助高职院校提升办学质量、培养合格实战型人才为核心来开展工作。