从数字脚印到城市计算

张大庆 陈 超 杨丁奇 熊昊一
法国国立电信学院

关键词：数字脚印 城市计算

当前，随着感知、计算、通讯技术的日新月异，记录人类日常行为轨迹、物理世界的动态变化以及人类与虚拟世界交互等的数字印迹正以前所未有的规模积累和扩张，形成大数据。我们把这些数据称为“数字脚印”[1]。数字脚印能从不同的角度反映城市的动态变化及存在的问题。利用和分析这些数字脚印可为揭示隐含的各种城市现象、构建智能城市提供一种新的思路。本文主要研究如何从大量的数字脚印中挖掘和理解个人和群体活动模式、大规模人类活动和城市动态规律，并把这些信息服务于改善人类的城市生活、提升城市的整体服务质量。下面我们通过一个例子来展示通过分析数字脚印而实现的城市计算给人们的城市生活带来的变化。

北京的王先生第一次到巴黎。刚下飞机，他用手机的移动出租车平台打到了一辆去酒店的出租车。看着计价器上跳动的欧元数字和并不熟悉的巴黎街道，他一点也不担心，因为移动出租车平台显示出出租车一直在最优的路线上行驶。他开始考虑当天的晚餐，第一次来法国的他难免不知所措。于是，他打开了个性化美食推荐的移动应用，迅速找到了下榻酒店附近的一家特色海鲜餐馆。酷爱海鲜的王先生毫不犹豫地订下了当晚的座位。享受完美食，王先生打开手机，了解当地的空气质量后，选择了一条空气清新的路线，散步在夜色中的巴黎街头。

即使在不熟悉的城市，王先生仍然享受到了令自己满意的服务。而这些正是由城市计算带来的。具体来讲，城市计算通过挖掘出租车 GPS 数字脚印为王先生提供了智能叫车、自动绕路检测的服务[2]；通过挖掘基于位置的社交服务数字脚印，向王先生推荐了符合他口味的餐馆[3]；通过采集基于传感的智能电话数字脚印，向王先生提供城域的空气质量状况及路线规划服务。“服务无所不在”，正在给城市人们的生活方式等带来前所未有的变革。

数字脚印与城市计算

城市计算可简单概括为：通过城市感知、数据挖掘、智能提取和服务提供四大环节来建立一个生态循环系统[4]。本文主要是从大量原始的数字脚印入手，通过统计分析和数据挖掘等技术揭示隐藏在“大数据”背后的智能。不同的数字脚印蕴含着不同的社群智能[1]，可为城市带来迥然不同的智能服务。数字脚印除包括出租车 GPS 轨迹、基于位置的移动社交网络数据和移动智能电话记录等外，常见的还有城市公共自行车租借记录、乘客公共交通刷卡记录、城市居民家庭和机构用电用水记录等。截至目前，已有不少利用数字脚印在城市计算方面开展的研究工作，包括：

出租车 GPS 数字脚印

目前许多大城市的出租车上都装载有 GPS 设备，用于记录出租车在城市的驾驶轨迹。如何将出租车 GPS 数字脚印用于城市计算？它能为人们提供怎样的服务？近年来，涌现了一批有代表性的前期研究工作。已开展的研究问题有不同时刻的城市热点检测[5]、城市区域的功能特性分类[6,7]、路径规划[8,9]、出租车司机寻客策

略[10]、异常轨迹检测[2,11~13]、城市道路交通流量预测[14]等。例如，美国麻省理工学院SENSEable City实验室的研究人员通过分析上万辆装有GPS传感器的出租车的载客点和下客点数据，揭示了一天中整个城市不同时刻的热点区域；浙江大学的研究团队根据热点区域的时空特性为出租车司机推荐下一个可能的乘客载客点[15]；微软亚洲研究院的研究团队通过挖掘北京上万辆出租车的历史行驶轨迹，基于出租车司机对城市道路的丰富知识和驾驶经验，为驾车人员个性化推荐快速的行车线路[8]。

移动社交网络数字脚印 随着智能手机的普及，越来越多的用户开始使用移动社交网络服务。用户在使用移动社交网络服务时留下的数字脚印包含了大量的个体和群体的行为信息。通过分析和挖掘这些数字脚印，可以了解个人和群体基于位置的行为，从而设计出更好的基于位置的服务。例如，英国剑桥大学的研究团队通过探索个人和群体移动模式，来预测用户将来的位置[16]；美国德州农工大学的研究团队利用用户数字脚印进行群体事件监测[17]；法国国立电信学院及微软亚洲研究院的研究团队通过分析这些数字脚印中所包含的用户偏好信息，为用户提供个性化的兴趣点推荐[3,18]和搜索服务[19]。

移动电话数字脚印 移动电话已成为大众生活中不可或缺的通讯工具。人们在城市的大街小巷和各个角落广泛地使用手机，手机记录了大量的关于用户的数字脚印。这些数字脚印亦为大规模城市计算提供了新的视角。如IBM都柏林研究所通过挖掘多个城市的用户通话记录来测量该城市的交通系统的效率，并提出了优化城市道路的模型与工具[20]。美国麻省理工学院的多媒体实验室通过挖掘手机通话记录和基站数据，发现了人的移动性与地区经济发展[21]和传染病蔓延[22]之间的关联关系，并设计了预测模型。美国东北大学的巴拉巴斯(Barabasi)教授领导的团队从复杂系统的角度出发，通过挖掘即时的移动电话通信数据，研究社会群体对大规模紧急事件（如爆炸、坠机以及地震等等）的反应，并基于此提出了移动通话数据监测群体性事件的模型和工具[23]。

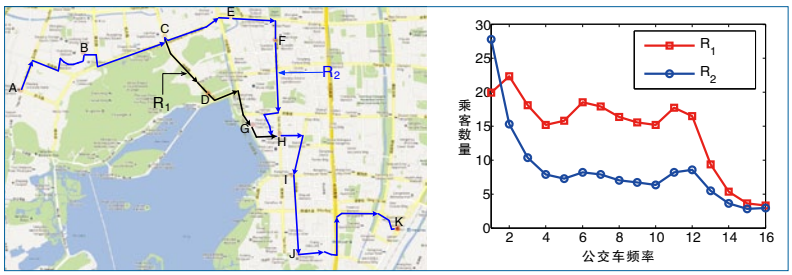
出租车GPS数字脚印与城市计算

作为城市中一种常用的交通工具，出租车在城市车辆总量占有较大的比重。出租车GPS数字脚印记录了出租车司机载客、寻客等行为在时间和空间两个维度的轨迹，构成了对城市中部分人群社会活动的独特采样。我们利用采自杭州市7600多辆出租车一年的GPS数字脚印，在出租车司机寻客策略发现、路径规划、异常检测、交通预测[2,9~14]等方面进行了系统的研究。以下着重介绍其中的两个工作。

夜班通宵公交车路线规划[9]

人们通常利用出租车或者私家车来满足夜间出行需求。与公交车出行相比，这两种方式既昂贵又会带来较为严重的汽车尾气排放。为了更好地保护城市环境和实现城市可持续发展，很多城市计划或已经推出通宵公交车来满足人们的夜间出行需求。因而，如何设计合理的夜间公交路线，既能满足部分人群的夜间出行需求，亦能通过运载较多的乘客使公交营运收支平衡成为问题的关键。通过挖掘夜间出租车GPS的历史数据，我们可以规划出合理的公交车线路，既确保公交满足相应的时间要求，又能最大限度地载客。

为此，我们提出B-Planner分两步来实现夜间公交车路线的规划。第一步，先将乘客夜间乘出租车的上客点和下客点进行聚类，通过对乘客密集区进行均匀切分来确定候选公交车站；第二步，给定公交路线的起始站和终点站，基于启发式规则和算法，来选择既满足时间约束又能达到载客人数最多的最优公交路线。图1(a)所示为由我们提出的B-Planner得到的通宵公交路线(R_1)与现实中一条由人工设计的公交路线(R_2)比较。值得注意的是，现实中R_2的开通晚于B-Planner规划R_1时所使用的出租车GPS数字脚印的产生时间，即规划



(a) (b)

图1 我们算法得到的公交路线和现实生活中的公交路线对比

R_1 使用的出租车 GPS 数字脚印反映了 R_2 开通前的情况。R_1 与 R_2 之间最大的区别是从 C 站到 H 站之间的路线。R_2 贯穿著名的商业街，而 R_1 贯穿夜生活聚集区域。图 1(b) 所示为两条公交路线在不同时段各班次的载客人数比较。可以看到，除了第一个班次外，R_1 的载客量都要大于 R_2。这是因为：第一班次的起始时间为晚上 10 点，此时商业街的客流量大于夜生活区的客流量；随后，夜生活区的客流量一直大于商业区的客流量。

出租车异常轨迹检测[2,11]

在乘坐出租车时，特别是当人们对城市道路不熟悉时，人们经常会担心自己被司机欺骗绕路。定义出租车司机是否“宰客”十分困难，需要考虑诸多因素。如有些路线虽然路程比较远，但行驶时间可能较短。我们根据历史轨迹，假定任意两地之间绕路的司机是少数的、不频繁的。我们将那些不频繁的轨迹定义为异常轨迹。这些轨迹可能是由那些绕路司机产生的，也可能是由经验丰富的司机找到的捷径。我们将两地之间的所有历史轨迹堆积（包括正常和异常的轨迹），提出了 iBAT 和 iBOAT 两个算法。它们能实时检测出异常轨迹，甚至轨迹中的异常片段。允许乘客在未到达目的地之前，预先了解到出租车的行驶路线是否正常。

在利用 iBOAT 算法检测出异常轨迹之后，我们进一步分析了以下三个问题：(1) 有多大比例的异常轨迹是由司机故意绕路产生的？ (2) 出租车司机在哪些起始点载客时，发生异常行为的可能性比较高？ (3) 那些爱绕路的出租车司机是不是比不绕路的司机赚的钱更多？为了回答这三个问题，我们先对 7350000 条出租车轨迹进行了异常检测，这些轨迹对应着杭州城 7600 多辆出租车一个月的载客轨迹。然后利用 iBOAT 算法检测出 438000 条异常轨迹。我们将同时满足如下两个条件的异常轨迹定义为绕路轨迹：(1) 该轨迹的长度大于正常轨迹的最大距离；(2) 该轨迹的行驶时间大于正常轨迹需要的最大行驶时间。从表 1 中发现，大约有 61% 的异常轨迹满足上述两个条件，表明绕路是出租车司机产生异常轨迹最主要的动机。图 2(a) 为绕路轨迹起始点分布图。从中可以看出，相比城市中其他区域，绕路轨迹起始点主要集中在长途汽车站、火车站，这说明出租车司机选择绕路的对象倾向于不熟悉城市道路的乘客。

表1 异常轨迹的行驶路程和时间分布

行驶距离	行驶时间 $[0, minT)$	$[minT, maxT]$	$(maxT, \infty)$
$[0, minD)$	0.0013	0.0137	0.0117
$[minD, maxD]$	0.0062	0.1063	0.0881
$(maxD, \infty)$	0.0045	0.1522	0.6162

图 2(b) 给出的是出租车司机的收入与绕路偏好的关系图。图中每一个点对应一辆出租车，纵坐标表示的是该出租车对应的司机产生的绕路轨迹数与其所有载客轨迹数目的比值，横坐标对应的是与该出租车对应的司机月收入。从中可明显地看出，偏爱绕路的司机所对应的月收入并不比不绕路司机的平均值高。

移动社交网络数字脚印与城市计算

移动社交网络最大的特色在于将用户位置标签添加到传统的交互媒体中，如用地理位置标记的消息、照片、视频等。通过使