轨迹预测技术及其应用
——从上海外滩踩踏事件说起

孙未未,毛江云

复旦大学计算机科学与技术学院;上海市数据科学重点实验室,上海 200433

轨迹预测问题一直是社会管理学、地理学、交通和计算机等各大领域关注的热点。相当长的一段时间内轨迹一直被怀疑“是否可以预测”,直到2010年2月Song等在《Science》上发表了“预测轨迹准确度上限”的论文,轨迹预测问题在理论上得到了可行性保障。本文首先将介绍轨迹和基于位置的服务(LBS)的概念与关系,从基础、工具、驱动力3个方面介绍轨迹预测代表性应用和成果,并分析面临的关键挑战和困难。

社会学家在总结人类社会发展的特征时,一个重要的指标就是人类活动的范围和速度。虽然也许我们并不享受每天奔波数十公里在上下班路上、乘高铁飞机长途旅行,但是现代人类活动的大规模、高频次、大范围、高速等特点,确确实实反映了现代社会对比过往任何社会阶段的巨大优越性。

从时空角度看,人类活动的发展还体现出显著的时空聚集性。如工作日的上下班高峰,是因为在同一时间有大量的人或车出现在同一地点造成;每年春节主要大城市来往海南三亚的机票卖空,是因为大量的人想在三亚温暖的海滩上过一个舒适的春节假期。

人群过度聚集势必会带来多方面的问题,典型的如下。

1) 安全隐患。

当人流过度聚集时,极易发生群体性安全事故。如2014年12月31日跨年夜上海外滩踩踏事件致36人死亡、49人受伤,2015年麦加朝觐踩踏事故致717人死亡、863人受伤。这些惨剧均和人流短时间内在某个空间范围内过度聚集,而组织方预判不够、现场疏导不足有直接关系。

2) 公共服务短缺。

当人流在过度聚集时,即使组织方做足了安全防范工作,但是由于聚集区域内各种软硬件条件所限,会出现各种公共服务短缺的问题。以2012年国庆节期间丽江地区的旅馆爆满导致近万名游客无法住宿的事件为例,每逢长节长假,旅游景区总会出现就餐难、如厕难、住宿难等各种公共服务的短缺,导致游客的旅游体验大打折扣。

那么有没有可能预测人群活动的趋势,提前发现存在的隐患呢?答案是肯定的。在智能手机普及的移动互联网时代,每个人都是一个定位传感器,每时每刻都产生定位数据,一系列时空数据组成时空轨迹。轨迹数据蕴含丰富的语义,单一移动对象的轨迹反映了其自身的行为特征,群体的轨迹反映了该群体共同的行为特征,而同一城市大量移动对象的轨迹则反映了该城市的社会活动特征。它已成为“智慧城市”的研究对象和研究热点,基于轨迹分析人群活动的规律,不但可以发现异常聚集的隐患,而且在城市规划、智能交通、环境保护和旅游路线推荐等诸多领域具有广阔应用前景。

在介绍基于轨迹的人群活动分析技术之前,以上海为例,首先了解一下在现实生活中如何应对以避免类似2014年底外滩踩踏事件的悲剧重演。在事件发生后,上海市政府快速响应,2015年初开始陆续出台了一系列应对手段。

首先,取消一批大型活动。2015年初上海市取消了作为国家级非物质文化遗产已经举办了20年的“豫园元宵灯会”引起很大社会反响。同样命运的还有:松江方塔园元宵灯会、嘉定古猗园元宵灯会等传统活动,多个签售会、演唱会,甚至还有读书会。

收稿日期:2016-04-28

作者简介:孙未未,副教授,研究方向为轨迹大数据,电子信箱:wwsun@fudan.edu.cn;毛江云(共同第一作者),硕士研究生,研究方向为轨迹数据挖掘,电子信箱:jymao14@fudan.edu.cn

引用格式:孙未未, 毛江云. 轨迹预测技术及其应用——从上海外滩踩踏事件说起[J]. 科技导报, 2016, 34(9): 48-54; doi:10.3981/j.issn.1000-7857.2016.09.006

其次,加大应急预防的人力投入,加强实时监控的技术和设施投入。2015年国庆节期间,除了民警外还组织大批武警在外滩排成人墙隔离人流以备不测,确保国庆期间重点区域游客安全。同时,上海警方利用手机基站信号、WiFi嗅探这两种技术手段对外滩、南京路等重点区域加强监控,每半小时更新客流数据,加上已有的视频监控技术,通过增强技术手段保障安全。

政府的压力和苦衷不言而喻,态度积极,措施得力,但现有措施存在负面影响,有效性也存在不足。社会对"取消活动消除隐患"的做法争议很大,央视《新闻1+1》在2015年1月13日以"大型活动管控:叫停不会进步"为题讨论了这一社会热点话题。超规格的警力投入无法持续。依赖实时监控技术而没有警力准备,即使发现也为时已晚,2014年底上海外滩踩踏事件在发生惨剧前现场的政府工作人员已经发现并上报,但限于现场安保人力有限,无法采取有效措施(图1)。

这个例子反映出利用人群活动规律研究问题的重要性和迫切性,同时也看到这是个很有挑战性的难题。轨迹预测就是用来解决这一问题的,本文针对现代社会高度动态化的人群活动现状,介绍基于时空轨迹的预测问题、技术和挑战。

1 什么是轨迹?

"轨迹"是一个既古老又新颖的名词。它是早就存在于数学中的一个概念,维基百科的解释是"含有某种性质的所有点的集合";在移动互联网时代,提到"轨迹"时一般指时空轨迹,记录了一个移动对象的一系列位置点和对应时间戳。轨迹属于LBS(基于位置的服务,Location Based Service)领域的一个专业名词,轨迹分析是LBS领域的一个重要方向。

虽然很多人并不熟悉LBS这个英文缩写,或者基于位置的服务是指什么,但是几乎每个智能手机用户都在使用LBS,如天气预报App告诉所在城市的天气预报,通过手机地图查询附近的地铁站。这些App都需要定位用户的位置,这些和位置相关的App都属于LBS。

早在智能手机概念之前就有LBS了,而1993年发生在美国的一起绑架案使得LBS受到社会广泛关注和重视。1993年11月一个名叫詹妮弗·库恩的女孩遭遇绑匪绑架,库恩在过程中悄悄用手机拨打了911报警电话,但是当时911呼救中心没有办法通过手机信号确定库恩的位置,因为援救不及时,最终导致库恩被杀害。该事件当时在美国掀起热议,美国FCC(美国通信委员会)在压力之下于1996年推出了一个行政性命令,要求强制性构建一个公众安全网络,即无论在任何时间和地点,都能通过无线信号追踪到用户的位置。这就是有名的E911,促使移动运营商投入大量的资金和力量来研究位置服务,从而开启了现代LBS的"大门"。

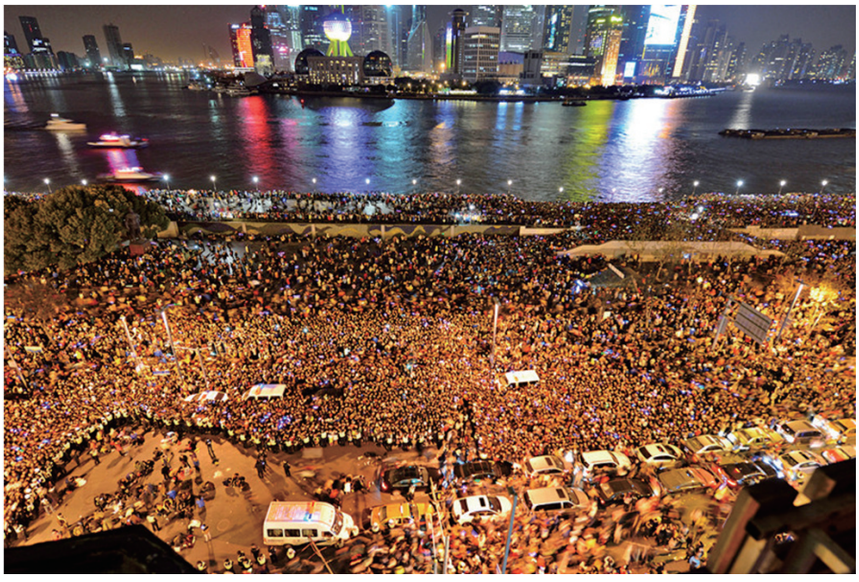

图1 2014年12月31日晚外滩踩踏事件(图片来源:《财新周刊》)

今天,LBS的工作形式,是通过电信移动运营商(如国内的移动、联通、电信等公司)的无线电通信网络或者其他方式(GPS、WiFi等)获取移动终端用户的位置信息,依托于电子地图平台,为移动终端用户提供相应服务。

轨迹数据,也是通过电信移动运营商或者诸如GPS、WiFi等方式采集得到的。通过电信运营商或者其他方式采集得到位置数据的过程,被称为定位,也就是说,可以通信的手机、携带GPS的仪器、可以连接到WiFi的仪器等所有携带位置采集传感器的设备,都可以通过某种技术手段查询到用户所在的物理位置。当这些设备孜孜不倦地工作了一定的时长之后,可能是一个小时、一个下午甚至是一整天,在这段时间里用户每时每刻的位置都被记录下来了,如果把这些位置都铺开放到到一张电子地图上的话,看到的就是用户这段时间的全部踪迹,这就是轨迹。当然,实际情况要更复杂一些,真正由位置采集设备得到的轨迹是一段由离散的位置点连接而成的歪歪扭扭的曲线(图2)。

在深入了解轨迹预测工作之前,先来窥探这些采集设备的工作原理(图3)。手机越来越成为人们生活不可或缺的必需品之一,而在使用手机打电话、发信息的时候,背后究竟发生了哪些不为人知的事?电信移动运营商究竟获取和记录了哪些数据?手机通信的原理,简单直观来说,手机发射信号搜索附近的基站,并且告诉这个基站"请帮我把我的信号M转达给对方B",当基站收到这个请求后会根据请求中的接收方B的位置搜自己区域内的所有对象,当找不到这个对象时则选择转达给其他基站,这个传递过程会一直进行下去直到到某个基站找到了B。整

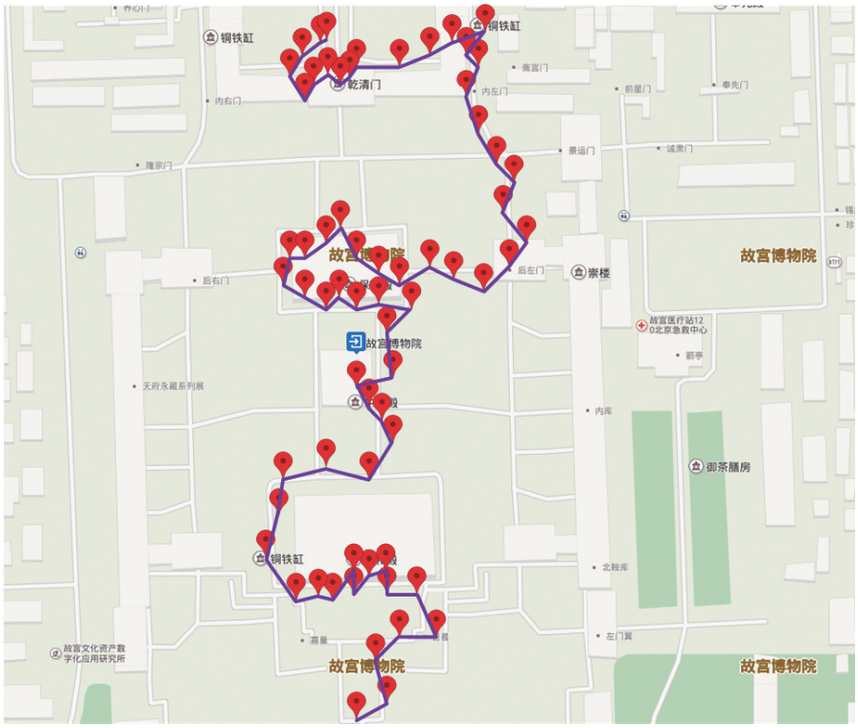

图2　手持位置采集设备采集得到的GPS位置点在地图上的显示
（北京故宫博物馆保和殿到乾清宫区域，以百度地图为基础）

个过程中，基站似乎扮演了一个非常忠实的“传话者”的角色，然而在这一系列接收、转达和送达手机信号的过程中，这个看似忠实的“传话者”却“悄悄”地记录下了接收到信号的时间，并且解析数据还记录了手机号码、手机型号，最重要的是，它还计算出手机用户当时的地理位置，并且把这些信息传给了电信移动运营商。不过目前不必太过担心自己的位置隐私会被完全泄露，一个原因是目前基站定位的误差大约在50~500 m不等，也就是说运营商并不能从位置数据中得到手机用户的精确位置，另一个原因是电信移动运营商们非常注意保护用户数据的隐私。

接着来看GPS的做法。GPS定位则显得更为“光明正大”，它直接告诉用户“我的工作是定位你的物理位置”。与基站定位技术不同，GPS依赖卫星组成的星网进行定位。目前广泛使用的GPS是美国GPS全球定位系统，出于种种隐私安全保护的原因，美国GPS所提供的民用卫星信号中都加入了干扰码，也就是说民用GPS采集得到的位置数据有5~10 m的精度误差，鉴于此，中国、俄罗斯和欧盟都投入研究开发中，目前备受关注的全球定位系统还有中国的北斗定位系统、俄罗斯的Glonass定位系统和欧盟正在建设的伽利略定位系统。

近几年兴起的其他定位技术，如蓝牙、WiFi、RFID等，受限于采集设备部署成本、信号波长特点等，往往应用于公共场所（如电影院、购物商厦等）的室内定位。

2　轨迹预测

对轨迹数据有了最基本的认识之后，再来讨论轨迹预测的工作就更自然了。

预测的问题自古有之，中国民间有一种非常流行的帮人预测诸如未来仕途、姻缘的职业，俗称“算命先生”，学术上称为“周易预测”。而在西方，塔罗牌和星座占星也是一种算命方式。一种说法算命最早起源可追溯到中国的伏羲氏，另一种说法算命始于中国先秦，无论哪种说法，都表明了人们对于未发生的事物能够提前预测的愿景历史悠久。不过问题放到当今大数据技术和产业蓬勃发展的时代，预测应当是通过对历史和当前的数据本身进行科学的统计和分析，从而去推测未来的发展趋势和规律。同样，轨迹预测工作也应该从轨迹数据本身出发，借助大数据分析技术，推测移动对象的将来位置。

轨迹预测，一直是社会管理学、地理学、交通和计算机等领域关注的热点，在相当长一段时间内，由于移动物体的运动具有“过程随机性，目的不确定性”等特点，普遍被怀疑是不是可以完成的工作。

轨迹预测究竟是否可行呢？在2010年2月的《Science》期刊上，Song等[1]发表了“预测轨迹准确度上限”的论文，利用移动通信运营商收集到的大量人流轨迹数据进行信息熵检验，发现接近93%的人类出行轨迹信息是可以预测的，人类在城市中的行为具有规律性、稳定性和集群性，同时也会存在大量的外部因素来干扰甚至改变人类的

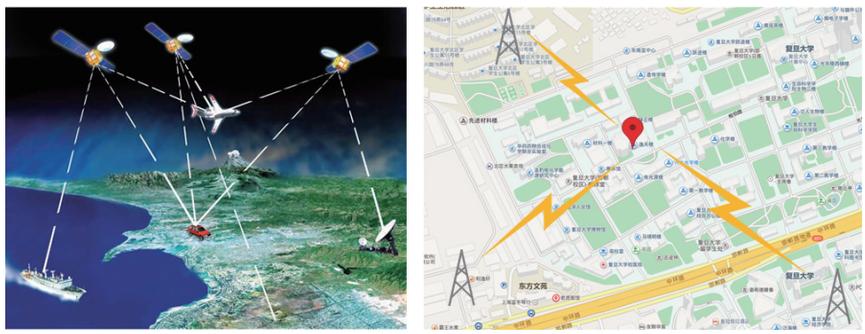

图3　左图为GPS定位（图片来源于环球网），
右图为基站定位图（基站位置为虚构，仅作示意）

图4　轨迹预测研究工作框架图,可以看到在轨迹预测工作中,大数据是基础,前沿技术是工具,而多样应用是驱动力

出行行为。这一结论对轨迹预测研究具有里程碑意义,确立了轨迹预测的理论基础。通俗地说,轨迹是可以预测的,测得准不准看各人本事。

此后,对轨迹预测工作的研究陆续开展起来。香港理工大学Pan等[2]提出了基于多变元正态分布的最佳线性预测器以预测轨迹;乔少杰[3,4]从2010年开始尝试了贝叶斯网络、隐马尔科夫模型等多种方法完善轨迹预测工作。轨迹预测领域,之所以能够催生大量优秀工作,可以总结如下3点(图4):1) 大数据基础,即以城市多源大数据为基础,催生越来越多的原来看似"不可能"的工作;2) 技术前沿性,依靠前沿的技术手段作为强有力的"工具"推动着信息时代背景下大部分的工作开展;3) 应用多样性,消费需求是社会生产力发生发展的永恒内在动力,对美好城市生活的应用需求驱动了一批又一批的杰出研究工作和应用产品的诞生。

2.1　大数据基础

轨迹数据是一种典型的大数据,符合"四V"特征:大量(Volume),数据量达到PB级别以上;快速(Velocity),需要在短时间内完成对新产生数据的分析处理;多样性(Variety),数据来源和形式各不相同;价值(Value),价值巨大同时价值密度低。以国内最大的两家地图公司为例,高德拥有5亿多高德地图App用户的导航轨迹,以及超过80%的出租车、长途客车和物流车的车辆轨迹;四维图新旗下从事智能交通业务的子公司世纪高通每分钟实时处理峰值超过1000万个GPS采样点,每天约40亿条、1 TB的新增轨迹数据。这些数据不但规模大、来源多样,而且主要价值体现在实时分析上,如实时路况分析,导航路径推荐等。

轨迹数据蕴含了人的移动信息,蕴含丰富的价值。随着大数据技术逐渐迅速发展,人们掌握了越来越多、越来越有效的"工具"把深藏于海量数据中的"宝藏"发掘出来。在城市管理和新形态城市构建体系中,以2014年12月31日的上海外滩事件为例,该事件给中国政府和社会拉响警钟,迫使政府利用信息技术手段对于"大人流,大车流"的进行科学管理,而"大人流,大车流"工作的核心就是轨迹预测;在商业应用中,轨迹预测可以指导酒店、餐馆科学浮动制定价格以增加收益,人流密度和人群肖像数据也是商铺选址的重要参考指标之一;在气候学和自然生态保护领域,轨迹预测工作的研究手段可以尝试"移植"到台风、飓风、动物等移动物体轨迹预测上,指导灾害防护工作以及生态动植物保护工作。

维克托·迈尔·舍恩伯格(图5)最早洞见大数据时代发展趋势的数据科学家之一,在《大数据时代:生活、工作与思维的大变革》著作中指出,大数据的方式出现了3个变化:第一,人们处理的数据从样本数据变成全部数据;第二,由于是全样本数据,人们不得不接受数据的混杂性,而放弃对精确性的追求;第三,人类通过对大数据的处理,放弃对因果关系的渴求,转而关注相互联系[5]。这代表着以大数据为基础,从低质量、粗粒度的轨迹数据中挖掘人群或其他移动物体未来的运动趋势和过去的运动的潜在关系是可行且必将成为研究的趋势。

2.2　技术前沿性

轨迹预测工作是一项具有很大挑战的任务,庆幸的是,在信息时代大背景下,大数据等技术逐渐迅速发展,使得越来越多深藏于数据"沙漠"底部的"宝藏"逐渐被发掘出来。

轨迹数据研究的技术手段,经历了从无到有,由少至多的发展。以时空数据索引作为研究轨迹数据的基础技术手段,各种索引技术和产品,克服了传统数据库技术维度表示能力弱、时空检索低效等不足,趋于成熟。时空查询技

图5　维克托·迈尔·舍恩伯格(图片来源:《时代周报》)

术更是在一代又一代杰出的研究学者的努力下,变得更为高效,可满足更多复杂的查询需求。轨迹数据清洗在整个与轨迹相关工作的发展过程中发挥巨大作用,技术手段也逐步发展。由于GPS卫星信号受地面接收端所在区域气候条件的影响较大,所以当地面接收端上方的云层较厚、大气中颗粒较多或者湿度过大时,均会影响GPS定位的精度。在城市环境中,如果道路两侧建筑物较高,也会因为多路径效应(multi-path effect)极大地影响道路中车辆的GPS定位精度。当GPS定位得到的经纬度坐标偏离车辆所在路段时,这种现象就被称为定位误差,以地图匹配工作为消除定位误差的代表工作,从曲线拟合技术慢慢发展,引入了诸如隐马尔科夫等模型,大大提高了准确性和速率,这些技术就是隐藏在人们日常生活中使用的导航软件背后的"智慧大脑";处于运动中的位置采集设备可能丢失与卫星的连接而无法定位、或者丢失与后台服务器的连接而无法回传导航数据、又或者用户出于流量/能耗的考虑减少向后台服务器发送采样数据的频率和规模,这些情况均会导致车辆的轨迹数据不完整,在时间和空间维度上产生较长的断裂和缺失,使得两个相邻采样之间的路线变得不确定,为了降低两点之间的不确定性需要运用路径还原技术,该技术也从轨迹插值起步,引入了更多更有效的技术,如反向增强学习算法。而在轨迹挖掘研究的技术手段,不再局限于统计学,更多引入了数据挖掘、机器学习甚至深度学习等技术手段。深度学习已经不是一个陌生的词汇了,2016年3月在Google的人工智能围棋AlphaGo以4:1打败了围棋九段高手李世乭之后,人工智能、深度学习成为学术界和工业界热议的话题。其实早在2015年,Yoshua Bengio团队就在以"出租车线路预测(Taxi Trajectory Prediction)"为主题的Kaggle竞赛中凭借他们在深度学习的应用优势斩获第一,这也标志着深度学习在轨迹预测领域获得了成功。

2.3 应用多样性

轨迹预测工作在城市发展各大领域有着强烈而迫切的需求。

未来新形态城市构建体系发展和城市公共安全管理一直是各国政府高度重视的问题。21世纪初,美国为了应对2008年华尔街金融危机以及规划人类未来城市蓝图,在2009年1月28日召集美国工商业领袖举行了一次"圆桌会议",会上IBM首席执行官彭明盛(Sam Palmisano)针对城市建设规划首次提出"智慧地球"(Smarter Planet)的理念。随后于2010年,IBM正式提出"智慧城市"(Smarter Cities)的概念,并于同年将该概念引入中国。人群轨迹预测工作无疑将为智慧城市起到添砖加瓦的推动作用。东京大学宋轩团队在2013年发表了他们对于160万人在日本大地震以及福岛核事故之后的移动规律的研究,以期为城市大灾难之后的应急救援提供理论支持[6]。悉尼大学教授Sanjay Chawla的团队通过轨迹研究发现交通流的紊乱程度进而判断城市的突发事件,为城市公共安全管理提供科学指导[7]。

城市环境和城市生态恶化是伴随城镇化迅速发展而带来的"副产物",2015年2月柴静公布空气污染深度调查《穹顶之下》,引起全社会对于城市环境和生态,尤其是空气质量问题的关注。轨迹预测在提升城市环境与生态问题上同样能贡献不小的力量,2013年微软亚洲研究院郑宇研究员就提出利用已有空气质量监测站点读数,结合气象、交通流、路网和兴趣点等多种数据源来实时分析细粒度的空气质量,郑宇团队还表示之后的工作方向将从空气质量的监控转向空气质量预测[8]。该团队还对纽约的311市民投诉数据进行了整合,通过与道路结构、兴趣点(Point of Interest,POI)语义化信息、人的社交媒体数据进行融合,计算出每个区域每段时间噪音指数,反映了公众对噪音的整体感受情况[9]。更有微软亚洲研究院张福铮通过出租车轨迹数据,计算在加油站点的排队等候时间,估算其能耗及汽车尾气中$PM_{2.5}$的排放量[10]。

基于位置的商业创新模式是近几年备受关注的商业性话题。以广告投放为例,传统的广告投放依托于纸质媒介、普通电台媒体等定时投放,随着人们的聚焦点逐渐转向互联网后,以及移动端设备逐渐赶超PC甚至占据大部分市场份额,广告投放领域出现了"移动广告"、"广告精准投放"等概念,广告投放策略发生了革命性的变化。据国外媒体报道,移动设备广告投入在2012年增长111%,2011年增长达到惊人的149%。正是凭借着依托互联网平台覆盖率广、结合用户位置信息定位精准、采用移动开发等技术互动性强等原因,后者迅速崛起。谷歌执行董事长施密特十分看好该类广告前景,他在接受采访时说道:"移动广告的价值肯定还会提升,因为我们将获取更多有关消费者的信息,还因为我们将与他们建立连接。我们知道他们大概位置。如果他们同意,便可与我们分享他们的历史。他们还可以指定希望看到广告的区域。"

3 挑战及展望

以城市大数据为基础、前沿的技术手段为保障、多样的应用驱动催生的轨迹预测工作在城市公共安全管理、城市环境治理、商业决策和创新模式等诸多领域逐渐"崭露头角"。

轨迹预测是绘制未来美好城市生活蓝图的重要"工具"。虽然目前在这个领域涌现了许许多多优秀的研究工作和优秀学者,但是想要取得更准确的预测结果,以及从研究理论的绘图纸上真正打造出可用可行的工具,还面临着艰巨的挑战,而这些挑战也将作为轨迹预测工作今后的研究重点:轨迹数据海量性,轨迹数据质量低下,城市异构数据源融合,用户个体的隐私保护等(图6)。

3.1 轨迹数据海量性

大数据是轨迹预测工作的基础,在驱动轨迹数据采集、整合、挖掘和分析发展的同时也带来了存储管理数据的负担、快速处理数据的压力和高效计算能力的挑战。2008年9月《Nature》专刊就大数据的问题提出:人类已经进入

图6　轨迹预测未来展望(图片来源于互联网)

海量存储PB时代,并预测下一个IT巨头的主营业务将会是大数据管理。华东师范大学周傲英教授在2012年明确提出了存储与组织数据密集型科学与工程的大数据存在的挑战[11]。

在存储和管理数据方面,传统的关系型数据库的管理模式和索引技术在大规模数据面前显得力不从心,轨迹数据体量大、信息碎片化、结构化程度差,时空多维度依赖性强等特点都对现有的数据处理、分析技术提出了更高要求。

在快速处理方面,轨迹预测工作在特定应用场景下,如"大人流、大车流"管理,非常讲求高效性和实时性,对实时到来的数据进行快速处理得到预测结果,进而支持管理和决策,如果不能保证快速处理当前新产生的数据并产生结果,将会导致人流和车流的管理和决策的滞后,严重时甚至发生因措施采取不及时造成的人群或车流拥堵、对冲惨案。

大数据计算中普遍采用分布式并行计算的方式,通过网络将分散的计算机连接组成完整的系统,接着将大量的数据分散"派发"到各个部分并交由该部分的计算机计算,最后将各部分计算的结果回收整合得到最终的结果。Hadoop、Spark和Storm是目前最流行的3种分布式计算机系统,Hadoop是雅虎工程师在2005年合作开发成功的产物,Spark是由加州大学伯克利分校的实验室开发,Storm则是由BackType团队开发。而基于轨迹这类时空数据的分布式系统,目前也陆陆续续提出了一些解决方案,如明尼苏达大学开发的SpatialHadoop,是专门用于在Apache-Hadoop集群上处理空间数据。但是目前时空分布式计算技术还不够完善,如难以支持复杂的算法逻辑,还有很长的路要走。

3.2　轨迹数据质量低下

轨迹数据作为轨迹预测工作的主要基础数据,其质量的优劣直接影响预测准确度精准与否。然而,真相却是:真实生活中约有60%的轨迹数据是低质量、粗粒度的。除了前文提到的定位误差和路径缺失问题(也可以理解为低采样问题),数据稀疏性也是最大的问题之一。

稀疏性看上去似乎不能够被接受,不少人的第一印象是:"现在不是大数据吗?既然已经有海量的数据了,怎么还会存在数据稀疏性的问题?"其实,大数据与数据稀疏性并不矛盾,由于位置采集设备本身误差或信号传输误差造成不可用的错误数据,城市人群在时间和空间上的不均匀分布导致部分区域轨迹数量稀疏甚至缺失,以及设备采集在时间和空间上的不连续性等诸多原因,导致看似庞大的轨迹数据在多维(时间和空间,甚至在某些应用中会被划分成更高维度)划分后出现严重的稀疏性现象。如2014年在数据挖掘顶级会议SIGKDD上的一项预测路径时间的工作[12],它将轨迹按"车辆、时间单位、道路单位"分割成三维张量,随后发现只有约3%的三元组中包含足量的数据,数据稀疏性异常严重。

随着数据稀疏程度增加,预测的工作的结果不确定性也随之增加。特别是对于对数据依赖性特别强的技术手段,数据稀疏性问题直接影响最后的效果。时空数据稀疏性问题逐渐被研究学者们意识到了,但是着重解决时空稀疏性问题的工作寥寥无几。一个原因是问题表现性多样,即来源不同的轨迹其稀疏性特征不同;另一个原因是技术手段不统一,对于不同来源数据和不同应用的问题,解决数据稀疏性的技术手段通常是不同的。

轨迹数据还面临着其他诸如数据错误、采样率不一致、数据异常等问题。因此,提高轨迹数据质量工作任重而道远。

3.3　城市异构数据源融合

城市数据包罗万象,不同来源的数据刻画了城市不同方面的"肖像"。城市兴趣点数据/POI,反映了城市功能的语义化信息,如医院、住宅区等;城市居民在社交网络上的活动,不仅是一张城市人群关系图,也间接反映了城市情感和活动动态;城市摄像头采集到的路况图像数据等,则反映了城市动态活动。

浙江大学潘纲教授在2013年提出可以依据城市人群在空间上的活动强度和时空序列模式,通过将模式相同的轨迹聚类和被聚类的区域再次分类的方法,能有效地识别出不同的活动单元[13]。北京市城市规划设计研究院高级工程师龙瀛根据北京市850万张公交卡在2008年连续一周的刷卡记录结合北京市2005年居民出行调查和地块级别的土地利用图数据,分析了北京市公交持卡人的居住地、就业地和通勤出行特征[14]。法国学者罗斯(Roth)利用英国伦敦203万人1122万条地铁刷卡

数据分析了城市多中心的空间结构特征[15]。

同时,政府和企业都看到了多源异构数据融合的应用前景。上海政府在2015年举办了上海开放数据创新应用大赛(Shanghai Open Data Apps, SODA),大赛开放了上海城市道路交通指数、地铁运行数据、一卡通乘客刷卡数据、浦东公交车实时数据、强生出租车行车数据、空气质量状况、气象数据、道路事故数据、高架匝道关闭数据、新浪微博交通数据等数TB数据,吸引了2914人参赛。同年"中国好创意"CCF全国青年大数据创新大赛中,亚信公司单独以"基于位置的应用及商业模式创新"为题设立子赛题,复旦大学计算机学院MDM217团队首次提出了利用多源异构数据来解决城市人群轨迹预测问题。中国电信上海公司举办"理想杯"大学生大数据创新应用与建模大赛,提供35 TB网络历史数据,包括用户上网数据和基站定位数据。

可以预见,多源异构的城市数据能够帮助提升轨迹预测的准确度。但是多源异构数据五花八门,属性相差很大,如轨迹是时空数据,社交网络以关系网络数据为主,监控采集得到的则是图片数据等。如何管理和融合大规模的多源异构数据是轨迹预测工作面临的挑战,只有建立好不同数据之间的关联,才能为后续工作提供保障。

3.4 用户个体的隐私保护问题

隐私保护一直是数据处理、挖掘与分析过程中不可回避的问题。在轨迹数据中,即使进行用户身份脱敏,通过移动个体位置信息的时空特征表现进行隐私攻击,仍然可以根据个人爱好、习惯、行为模式等逆向推测出移动个体的身份信息。而设计一套有效的基于轨迹的隐私保护技术极其困难,因为需要同时考虑保护轨迹数据的隐私和保证轨迹数据有较高的可用性。而在轨迹预测工作中,融合了多源异构数据,这将大大增加了用户隐私被暴露的可能性,以上海SODA大赛开放的数据为例,仅仅通过"交通一卡通"一天的刷卡记录就能关联到一天内乘坐了公交车、出租车和地铁的用户,并推测早上的出发地点和晚上的终点如果一致那么该用户的家位于这个地点所在的区域,早高峰的终点和晚高峰的起点一致则推测该用户的上班单位地点位于对应地点的区域,这样简单的规则就已经能够发现不少满足这个规律的用户。轨迹隐私还表现在访问敏感位置、家庭地址或单位地址被泄漏、甚至是个人健康状况等私密信息被泄漏。如发现个人轨迹表现为在某家医院附近停留,通过规律的通勤轨迹可以知道用户的家庭区域和上班单位区域等。隐私保护问题,不仅仅是一个技术问题,也是一个道德和法律的问题,应当给予进一步的关注。

轨迹预测工作对实际生活究竟能够产生多大影响呢?就如同以IBM的深蓝、Google的AlphaGo为代表的人工智能在未来社会中将发挥怎样的作用一样,值得期待!

参考文献(References)

[1] Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility[J]. Science, 2010, 327(5968): 1018-1021.

[2] Pan T L, Sumalee A, Zhong R X, et al. Short-Term traffic state prediction based on temporal-spatial correlation[J]. Intelligent Transportation Systems, IEEE Transactions on, 2013, 14(3): 1242-1254.

[3] Qiao S, Tang C, Jin H, et al. PutMode: Prediction of uncertain trajectories in moving objects databases[J]. Applied Intelligence,2010, 33(3): 370-386.

[4] Qiao S, Shen D, Wang X,et al. A self-adaptive parameter selection trajectory prediction approach via hidden Markov models[J]. Intelligent Transportation Systems, IEEE Transactions on, 2015, 16(1): 284-296.

[5] 荆林波. 信息技术时代: 哲学社会科学研究面临的挑战及其应对措施[J]. 学术探索, 2015(1): 1-6.

[6] Song X, Zhang Q, Sekimoto Y, et al. Modeling and probabilistic reasoning of population evacuation during large-scale disaster[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 1231-1239.

[7] Pang L X, Chawla S, Liu W, et al. On detection of emerging anomalous traffic patterns using GPS data[J]. Data & Knowledge Engineering, 2013(87): 357-373.

[8] Zheng Y, Liu F, Hsieh H P. U-Air: When urban air quality inference meets big data[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 1436-1444.

[9] Zheng Y, Liu T, Wang Y, et al. Diagnosing New York city's noises with ubiquitous data[C]//Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014: 715-725.

[10] Zhang F, Yuan N J, Wilkie D, et al. Sensing the pulse of urban refueling behavior: a perspective from taxi mobility[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 6(3): 37.

[11] 宫学庆, 金澈清, 王晓玲, 等.数据密集型科学与工程:需求和挑战[J]. 计算机学报, 2012, 35(8): 1563-1578.

[12] 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. 2014: 25-34.

[13] 潘纲, 李石坚, 齐观德, 等. 移动轨迹数据分析与智慧城市. 中国计算机学会通讯, 2012, 8(5): 31-37.

[14] 龙瀛, 张宇, 崔承印. 利用公交刷卡数据分析北京职住关系和通勤出行[J]. 地理学报, 2012, 67(10): 1339-1352.

[15] Roth C, Kang S M, Batty M, et al. Structure of urban movements: polycentric activity and entangled hierarchical flows[J]. PloS ONE, 2011, 6(1): e15923.

(责任编辑 刘志远)