软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(4):713−730 [doi: 10.13328/j.cnki.jos.004570]

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

位置大数据的价值提取与协同挖掘方法*

郭 迟[1], 刘经南[1], 方 媛[1], 罗 梦[2], 崔竞松[2,3]

[1](武汉大学 卫星定位导航技术研究中心,湖北 武汉 430079)

[2](武汉大学 计算机学院,湖北 武汉 430072)

[3](软件工程国家重点实验室(武汉大学 计算机学院),湖北 武汉 430072)

通讯作者: 郭迟,崔竞松, E-mail: {guochi, jscui}@whu.edu.cn

摘 要: 随着位置服务和车联网应用的不断普及,由地理数据、车辆轨迹和应用记录等所构成的位置大数据已成为当前用来感知人类社群活动规律、分析地理国情和构建智慧城市的重要战略性资源,是大数据科学研究极其重要的一部分.与传统小样统计不同,大规模位置数据存在明显的混杂性、复杂性和稀疏性,需要对其进行价值提取和协同挖掘,才能获得更为准确的移动行为模式和区域局部特征,从而还原和生成满足关联应用分析的整体数据模型.因此,着重从以下 3 个方面系统综述了针对位置大数据的分析方法,包括:(1) 针对数据混杂性,如何先从局部提取出移动对象的二阶行为模式和区域交通动力学特征;(2) 针对数据复杂性,如何从时间和空间尺度上分别对位置复杂网络进行降维分析,从而建立有关社群整体移动性的学习和推测方法;(3) 针对数据的稀疏性,如何通过协同过滤、概率图分析等方法构建位置大数据全局模型.最后,从软件工程角度提出了位置大数据分析的整体框架.在这一框架下,位置数据将不仅被用来进行交通问题的分析,还能够提升人们对更为广泛的人类社会经济活动和自然环境的认识,从而体现位置大数据的真正价值.

关键词: 大数据;轨迹移动模式;位置服务;泛在测绘;数据挖掘

中图法分类号: TP311 **文献标识码**: A

中文引用格式: 郭迟,刘经南,方媛,罗梦,崔竞松.位置大数据的价值提取与协同挖掘方法.软件学报,2014,25(4):713−730. http://www.jos.org.cn/1000-9825/4570.htm

英文引用格式: Guo C, Liu JN, Fang Y, Luo M, Cui JS. Value extraction and collaborative mining methods for location big data. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):713−730 (in Chinese). http://www.jos.org.cn/1000-9825/4570.htm

Value Extraction and Collaborative Mining Methods for Location Big Data

GUO Chi[1], LIU Jing-Nan[1], FANG Yuan[1], LUO Meng[2], CUI Jing-Song[2,3]

[1](Global Navigation Satellite System Research Center, Wuhan University, Wuhan 430079, China)

[2](Computer School, Wuhan University, Wuhan 430072, China)

[3](State Key Laboratory of Software Engineering (Computer School, Wuhan University), Wuhan 430072, China)

Corresponding author: GUO Chi, CUI Jing-Song, E-mail: {guochi, jscui}@whu.edu.cn

Abstract: Uncountable geographical location information, vehicle trajectories and users' application location records have been recorded from different location-based service (LBS) applications. These records are forming to a location big data resource which facilitates mining human migrating patterns, analyzing geographic conditions and building smart cities. Comparing with traditional data mining, location big data has its own characteristics, including the variety of resources, the complexity of data and the sparsity in its data space. To restore and recreate data analysis network model from location big data, this study applies data value extraction and cooperative mining on location big data to create trajectories behavior pattern and local geographical feature. In this paper, three major aspects of

* 基金项目: 国家自然科学基金(41104010); 国家高技术研究发展计划(863)(2013AA12A206, 2013AA12A204); 国家自然科学重大研究计划(9112002); 高等学校学科创新引智计划(B07037)

收稿时间: 2013-10-14; 修改时间: 2013-12-18; 定稿时间: 2014-01-27

analysis methods on location big data are systematically explained follows: (1) For the variety of resources, how to extract potential contents, generate behavior patterns and discover transferring features of moving objects in a partial region; (2) For complexity of data, how to conduct dimension reduction analysis on complex location networks in temporal and spatial scale, and thus to construct learning and inferential methods for mobility behavior of individuals in communities; (3) For sparsity, how to construct the global model of location big data by using collaborative filtering and probabilistic graphical model. Finally, an integral framework is provided to analyze location big data using software engineering approach. Under this framework, location data is used not only for analyzing traffic problems, but also for promoting cognition on a much wider-range of human social economic activities and mastering a better knowledge of nature. This study incarnates the practical value of location big data.

Key words: big data; trajectories mobility pattern; location based service; ubiquitous mapping; data mining

位置服务(location based service,简称 LBS)是近年来新兴的移动计算服务.发展位置服务主要需重视其两个方面的能力:提供位置的能力和理解位置的能力.在提供位置方面,随着室内外无缝定位技术和增强系统技术的发展,定位精度不断提高[1],在大众应用层面已经基本满足人们生产、生活的需要;然而在理解位置的能力方面,目前尚有很多挑战,是学术界和产业界关注的热点.理解位置其实就是理解位置背后所反映出来的人的活动、人的情感和人的环境,因此也被称为泛在测绘(ubiquitous mapping)或位置社会感知(location-based social awareness)[2].

位置大数据(location big data)是构成泛在测绘和位置社会感知的重要资源,具有相当大的体量.近几年,位置服务、数据挖掘和机器学习领域,已经涌现出一批针对位置大数据的优秀研究.其所使用的数据集在体量和复杂性上均已达到了“大”数据的层次,代表性实例见表 1.

Table 1 Instances of location big data

表 1 位置大数据实例

移动目标	目标数量|O|	持续时间|T|(天)	记录数量|P|	研究目的
出租车	12 000	110	577 000 000	寻找乘客和空闲出租车[3];推断交通异常[4]
	7 475	385	3 000 000 000	土地规划分类[5]
移动电话	50 000	90	10 000 000	研究人们移动行为的可预测性[6]
	1 500 000	450	/	研究人们移动行为的独一性[7]
	1 600 000	365	/	模拟灾后人们大规模移动行为[8]
社交网站	632 611	30	15 944 084	模拟疾病传播[9]

位置大数据主要来源于车联网(Internet of vehicles,简称 IOV)、移动社交网络、微博等新兴互联网应用,更新速度快且具有很大的混杂性(inaccurate).同时,往往受到数据采集技术等方面的客观制约,使得这些数据不能全面和正确地反映观察对象的整体全貌,因而具有“复杂但稀疏(complex yet sparse)”的特点.如何从位置大数据中获得价值,进而发现人类社群活动规律,是非常值得探讨的问题.本文将着重归纳和阐述这其中有关局部特征提取、数据降维、整体特征建模以及整体数据协同挖掘的方法.

本文的另一个贡献是从关联应用角度阐述了位置大数据的意义和价值.传统的诸如轨迹数据等往往仅被用以分析城市交通等直接且特定的问题.大量经典的大数据科学研究表明,通过价值提取和协同挖掘后的数据结果能够将一些看似无关的事件很好地联系在一起,从而从数据层面“直接”反映一些原本需要复杂因果建模才能得到的结果,且更加直观和准确[10−13].这些案例对位置大数据研究同样具有启发性.因此,我们在探讨位置大数据分析方法时,本身就应将其置于关联应用的大背景下,着重探讨如何将模型参与到社会经济活动、政治活动、自然环境、人类情感以及人口卫生等一系列社会学、人类学、经济学的研究中.这样的位置大数据才更有助于地理国情的分析和智慧城市的建设.

1 基本定义和预处理方法

首先,我们给出本文所面对的位置大数据的基本结构.前文已述,当前的位置大数据主要来源于 IOV、移动社交网络新兴互联网应用,有如下描述:

定义 1. 位置数据集记为 $LBD=\{O,T,P\}$,其中,$O=\{o_1,o_2,\ldots\}$表示数据集中的移动对象集合,包括了$|O|$个产生位置的移动目标;T 为观察数据集的时间;$|T|$天内总共获得$|P|$个位置记录.

定义 2. 单个位置数据记录 p 主要包含移动目标 o 和位置的地理坐标$\langle x,y\rangle$和记录时刻 t,可以用一个四元或五元组表示.如果是车辆轨迹数据,一般还包含车辆的速度 v 以及一组状态信息 $S=\langle S_1,S_2,\ldots\rangle$,如行驶方向、油耗值、载客状态等,记为 $p=\langle o,x,y,t,v,\boldsymbol{S}\rangle$,其中,一个具体的状态 S_i 可能有多个状态取值;如果是用户在社交网络等媒体上主动分享的位置数据,则还包括与位置相关的媒体信息 $\boldsymbol{I}$,可记为 $p=\langle o,x,y,t,\boldsymbol{I}\rangle$.一般地,将移动目标 o_i 的第 j 条位置记录记为 $p_j^{(o_i)}$,在不影响理解的情况下也可直接写作 p_j.

1.1 地图的预处理

位置大数据分析一般需要基于地图或路网数据展开.通常,平面地图被认为是一个连续的二维空间,为方便分析,需将其离散化,即,将地图划分为多个区域.这也是位置大数据预处理分析中普遍采用的方法,常见的包括:

(1) 网格化分区[14,15],如图 1(a)所示;

(2) 依道路网分区[16].这种分区方法能够很好地保留地图语义.为了精简操作,一般按照城市主干道进行划分,如图 1(b)所示;

(3) 依位置密度分区[5,17,18].这种方法主要依据 LBD 中 p 的密度,将在一定范围内的位置点进行聚集,继而将地图划分为大小不同的网格或不规则图形(凸包),如图 1(c)所示.常见的密度聚类算法如 DBSCAN[19]等;

(4) 依参考点分区[7].这种分区方法主要是选取 LBD 中若干位置点或地图上即有的若干兴趣点(point of interesting,简称 POI)作为参考点,按照 Voronoi 多边形(又称为泰森多边形)的方式划分区域.使其每个分区内的任意一点到相应参考点的距离比到其他参考点的都近,从而很好地保留了参考点的代表性位置语义,如图 1(d)所示.

(a) 网格 (b) 道路网 (c) 位置密度 (d) 参考点

Fig.1 Preprocessing method of map segmentation

图 1 地图分区预处理方法

通过地图离散化,将地图划分为多个区域,完成对位置大数据分析的第 1 步预处理.在后文的分析中,我们不强调具体的分区方式,统一表述.

定义 3. 一个地图区域集合记为 $\mathfrak{R}=\{r_1,r_2,\ldots,r_n\}$,其中,$\|\mathfrak{R}\|=n$ 表示一共划分有 n 个区域,$r_i.ref$ 为区域 r_i 的参考点.

1.2 位置轨迹数据的预处理

(1) 轨迹插值

在对位置大数据尤其是轨迹数据进行分析时,一般会要求数据集具有较高的采样率.当原始观察数据无法满足这样的要求时,可以对其进行简单线性插值[15].

(2) 地图匹配

地图匹配(map matching)是进行位置大数据研究的十分重要的预处理步骤,其目的是将原始观察数据与地

图中的道路网信息联系起来,在使定位更加精确的同时,获取移动目标的移动轨迹.目前已存在一些经典算法,将位置数据、路网数据以及道路特征(如限速)等信息加以融合,能够较为准确地还原移动对象的轨迹.如ST-Matching 算法、IVMM 算法、Passby 算法等[20−23].

定义 4. 移动对象 o_i 的一条轨迹 j 记为 $traj_j^{(o_i)}=\langle p_1\rightarrow p_2\rightarrow ...\rangle$,$||traj||$表示轨迹上位置数据 p 的数量,称为轨迹长度.在不影响理解的情况下也可直接写作 $traj_j$.

2 局部位置数据的特征提取

大数据分析的首要任务是从局部研究对象中提取出价值,建立单个区域 r_i 或单个移动对象 o_i 的若干特征模式.根据特征模式的提取方法,我们将其划分为如下两类:

(1) 一阶特征:是指从区域内的位置记录、地图数据或移动对象历史轨迹中可简单计算获得的特征,如均值、方差等;

(2) 二阶特征:是指需要经过一些高阶统计处理才能获得的模式特征.这些特征经过统计处理,能够在一定程度上消除原始观察数据混杂性所带来的影响,是本文归纳的重点.

2.1 区域静态特征 φ_s

区域静态特征主要统计的是区域内与地图地貌相关的一些指标,可用于对不同区域进行聚/分类处理,常见的区域静态特征包括[14]:

(1) 路网特征(road feature,f_{RN})

f_{RN} 是由区域内快速路的长度、普通路段的长度、道路交叉口数量、区域道路弯曲度、道路基质质量等特性所构成的特征向量.

(2) POI 特征(POI feature,f_{POI})

f_{POI} 可以为一个有关区域内 POI 信息的多维向量,包含各类型 POI 的数量及其变化率以及没有准确 POI 信息覆盖的区域面积等.

2.2 个体移动模式特征 φ_{mp}

个体移动模式(mobility pattern,简称 MP)以单个移动对象 o 为观察目标,包括其在一段时间内的移动独一性、随机性、周期性、转移性、动静间歇性和移动期望性等方面.

(1) 移动独一性(uniqueness feature,f_{uniq})

移动独一性可用来区别移动对象,定义为通过给定地图区域个数$||r||$、区域平均大小 $\overline{r.size}$ 和统计时间间隔 $\overline{r.time}$,唯一确定一条轨迹 $traj_j$ 的概率,即:

$$\Pr\{|traj_j|\leqslant 2\,|\,\overline{r.size},\overline{r.time},||r||\} \tag{1}$$

实证研究表明,当 $\overline{r.size}$ 和 $\overline{r.time}$ 相对合适时(比如当 $\overline{r.size}$=0.15km^2, $\overline{r.time}$=1 hour),仅需 4 个左右的区域(即$||r||$=4)便可以在茫茫人海中以很高的概率确定一条唯一的轨迹[7].当$||r||$固定时,这一概率与 $\overline{r.size}$ 和 $\overline{r.time}$ 分别呈现相似的幂律关系,即:

$$\begin{cases} f_{\mathrm{uniq}}=\alpha-(\overline{r.size})^{\beta} \\ f_{\mathrm{uniq}}=\alpha-(\overline{r.time})^{\beta} \end{cases} \tag{2}$$

其中,β为幂指数,与$||r||$呈线性关系,满足:

$$\beta=\lambda_1-\lambda_2\cdot||r|| \tag{3}$$

这说明:参与决策的区域越多,幂指数越小,越容易确定移动独一性.文献[7]通过实证拟合,计算得到λ_1=0.157,λ_2=0.007.

通过观察很少的区域,便能唯一确定一条用户轨迹.这既说明个体移动具有高度的规律性,也说明两两间移动行为具有很大的差异性.这种实证拟合的方法同样适用于分析其他任何 LBD 数据集.个体移动独一性 f_{uniq} 可

以用在大数据环境下的个人隐私发掘或保护工作.独一性大小反映出数据集所在人群的整齐划一程度,因此在不同数据集上分析个体移动独一性,将有助于通过位置大数据分析其背后人群的自由程度、政治体制和生活情态,这将是很有趣的.

(2) 移动随机性(randomness feature,f_{rand})

个体移动的随机性可用位置熵(location entropy)来度量.设 p_x 为访问一个位置的随机变量,参照信息熵的定义,可以给出多类位置熵:

① 假设移动对象 o_i 共访问了$||r||$个不同的位置区域,其随机熵有:

$$H_1=H_i(p_x)\equiv\log_2||r|| \tag{4}$$

② 进一步地,统计其在各个位置区域 $\mathfrak{R}=\{r_1,r_2,\ldots,r_{||r||}\}$ 上出现的概率,记为 $\{\Pr(r_1),\Pr(r_2),\ldots,\Pr(r_{||r||})\}$,则其标准位置熵[24]有:

$$H_2=H_i(p_x\mid\mathfrak{R})=-\sum_{\mathfrak{R}}\Pr(r_j)\log_2\Pr(r_j) \tag{5}$$

③ 再进一步,考虑到移动对象 o_i 位置记录的时序性 $traj=\langle p_1\rightarrow p_2\rightarrow\ldots\rangle$,则位置时序熵有:

$$H_3=H_i(p_x\mid traj)=-\sum_{tr\subset traj}\Pr(tr)\log_2\Pr(tr) \tag{6}$$

其中,$\Pr(tr)$为在 $traj$ 中找到一条特定的子时序序列的概率,其大小反映了该移动对象每移动一个位置的信息增益,即,移动对象每随机行走 1 步平均有 2^{H_3} 个选择.一般来说,这个值是很小的[6,7],说明个体的移动具有可预测性[2].文献[6]在个体移动随机性的基础上,对其可预测性进行了深入探讨.

通过比较不同移动对象的位置熵的相似性,可用来进行朋友关系预测等[24].此外,在 LBD 中,通过对 O 中所有的对象分别计算其位置熵 $H(\cdot)$,继而获得熵的概率分布 $\Pr\{H(\cdot)\}$,便可对整体数据集中的移动随机性进行度量.位置熵还可以在不同时间尺度下(如将工作日和休息日分开)分别计算,这样可以对混杂的位置数据提取更多、更准确的知识.

(3) 移动周期性(periodic feature,f_{peri})

对一个移动对象 o_i 来说,将其访问区域 r_j 的序列二值化(1 表示访问,0 表示未访问),继而将该二值化序列进行离散傅立叶变换(discrete Fourier transform,简称 DFT),通过观察傅立叶系数最大的频率,即可获得该位置点被访问的周期 TP_j^i [15].

假设一组位置区域 $\mathfrak{R}=\{r_1,r_2,\ldots,r_{||r||}\}$ 具有相同的被访问周期 $TP=\{t_1,t_2,\ldots,t_k\}$,划分到 k 个时间槽,从而可以得到一个个体移动详细的概率分布矩阵 $P=[\rho_1,\rho_2,\ldots,\rho_k]$,其中,每一个列概率向量$\rho_j=[\Pr(r_1|t=t_j),\Pr(r_2),\ldots,\Pr(r_{||r||})]$.将 LBD 中 T 时段的位置记录按照周期 TP 分别生成 $\left\lfloor\frac{T}{TP}\right\rfloor=m$ 个概率分布矩阵$\{P_1,P_2,\ldots,P_m\}$,则可通过计算两两间分布的 KL 散度(Kullback-Leibler divergence)来分析移动对象的周期行为.

将公式(6)细粒化,可以得到一个更为精细的标准位置熵:

$$H(P)=-\sum_{t_j=1}^{k}\sum_{\mathfrak{R}}\Pr(r_i\mid t=t_j)\log_2\Pr(r_i\mid t=t_j) \tag{7}$$

则两两分布的相对熵有:

$$KL(P_1\|P_2)=\sum_{t=1}^{k}\sum_{\mathfrak{R}}\Pr_{P_1}(r_j)\log_2\frac{\Pr_{P_1}(r_j)}{\Pr_{P_2}(r_j)} \tag{8}$$

对连续 m 个位置概率分布$\{P_1,P_2,\ldots,P_m\}$按照相对熵大小进行层次聚类,可以得到频繁集最大的几个簇,代表移动对象 o_i 的几个典型周期行为[18],如图 2 所示.在聚类过程中,合并两个簇 C_i 和 C_j 的位置概率分布可简单计算为

$$P^{new}=\frac{|C_i|}{|C_i|+|C_j|}P_i+\frac{|C_j|}{|C_i|+|C_j|}P_j \tag{9}$$

(4) 移动转移性(transition feature,f_{trans})

衡量移动对象 o_i 在两个相邻时间段 T^{bef} 和 T^{aft} 中是否存在区域集合 $\mathfrak{R}=\{r_1,r_2,\ldots,r_{\|r\|}\}$ 上的转移行为,可以很方便地借助 Jaccard 相似性来统计[8].

设移动对象 o_i 在 T^{bef} 和 T^{aft} 时间段内在 $\mathfrak{R}$ 上的访问概率分别为 $\Phi=\{\Pr(r_1),\Pr(r_2),\ldots,\Pr(r_{\|r\|})\}$ 和 Φ',则其 Jaccard 系数满足:

$$\alpha_i=\left(\sum_{i=1}^{\|r\|}\min\{\Pr(r_i),\Pr'(r_i)\}\right)\Big/\sum_{i=1}^{\|r\|}\max\{\Pr(r_i),\Pr'(r_i)\} \tag{10}$$

α_i 越小,说明移动对象的转移性越明显.如果进一步对 T 内前后相邻的两个时间段分别计算 α_i,其期望 $\bar{\alpha}_i$ 可表示移动对象 o_i 的整体转移性.

转移性的另外一种常见的度量方法是计算移动对象 o_i 在 $\mathfrak{R}=\{r_1,r_2,\ldots,r_{\|r\|}\}$ 的转移概率,使用马尔可夫过程对用户的行为进行预测,相关的方法在我们早前的论文中有详细描述[2].

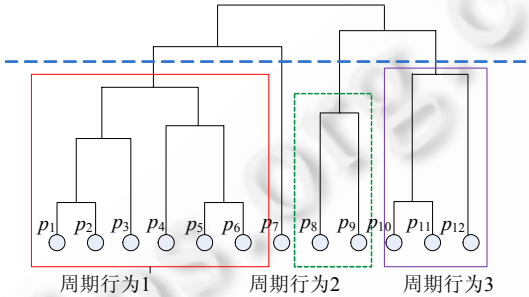


Fig.2 Hierarchical clustering of periodic behaviour

图 2 周期行为的层次聚类

(5) 移动间歇性(intermittent feature,f_{inte})

考虑到移动目标并不总是处于运动状态,则有必要从一段连续位置记录中发现其静止状态,以及估计造成这种静止状态的原因:分为被动静止(如车辆遇到红绿灯停车)和移动目标主动静止.

给定一条轨迹 $traj=\langle p_1\rightarrow p_2\rightarrow\ldots\rangle$,若其中存在静止状态,则会出现较密集的几个连续位置点 $\langle p_i\rightarrow p_{i+1}\rightarrow\ldots\rightarrow p_j\rangle$.用密度聚类的方式将其归并,再根据归并后位置点所在区域的时空特征 $\varphi_{\mathfrak{R}}$ 可分类判定[3,14].

(6) 移动期望性(expectation feature,f_{exp})

假设移动对象有一组伴随移动的状态 $\boldsymbol{S}=\langle S_1,S_2,\ldots\rangle$(见定义 1),每一类状态存在多种可能的状态值 $S_1\in\{s_1^1,s_2^1,\ldots\}$.通过历史位置记录获得移动对象发生状态转换的相关统计特征,称为移动期望性.为了方便起见,我们将只阐述移动目标 o_i 的状态 S_1 发生转换的情况.

定义 5. 移动目标从 T_0 时刻起,在路径 $path=\langle rd_1,rd_2,\ldots,rd_{|path|}\rangle$ 上发生 1 次状态转换 $s_1^1\rightarrow s_2^1$,记为事件 W.由于在一条路径 $path$ 上可能包括停顿和移动过程两部分,不妨设存在 m 个停顿点 $\{c_1,c_2,\ldots,c_m\}$.假设该状态转换恰好发生在路段 rd_i 上,记为子事件 w_{r_i};发生在停顿点 c_i 上记为事件 w_{c_i}.$rd_i.l$ 表示路段 rd_i 的长度,$rd_i.t$ 表示经过路段 rd_i 需花费的时间,$c_i.t$ 表示在停顿点 c_i 的平均停留时间.

设移动对象在路段 rd_i 上发生状态转换的概率为 $\Pr(rd_i)=\Pr\left(s_1^1\rightarrow s_2^1\middle|r_i,T_0+\sum_{j=1}^{i}r_j.t\right)$,则以其达到路段 rd_i 的时间 $T_0+\sum_{j=1}^{i}r_j.t$ 为参考,取其前后各 $k\times\tau$ 时间段内整体移动对象 o 的历史数据进行统计,则有:

$$\Pr(rd_i)=\left(\sum_{2k+1}\|(s_1^1\rightarrow s_2^1\,|_{T_0}^{\tau};rd_i)\|\right)\Big/\left(\sum_{2k+1}\|(s_1^1\,|_{T_0}^{\tau};rd_i)\|\right) \tag{11}$$

其中,$\|(s_1^1\rightarrow s_2^1\,|_{T_0}^{\tau};rd_i)\|$ 表示在 LBD 整体位置记录中,对应的那个时间段 τ 内在路段 rd_i 上发生子事件 w_i 的次数;

$\|(s_1^1|_{T_0}^{\tau};rd_i)\|$表示在路段 rd_i 上初状态为 s_1^1 的移动目标的个数.将待分析的路径 *path* 分为多个路段分别进行统计,再根据概率乘法原理求解移动对象在该 *path* 上事件发生的概率,是位置大数据分析中很常见的一种手段,被称为 partition-and-group[3,25].这样做可以有效地利用 LBD 中的位置记录,避免直接按照 *path* 进行数理统计时,由于 *path* 过于独特或复杂而带来的样本缺失的影响.

若移动对象在停顿点发生状态转换,则其概率不仅与停止时刻有关,与停止时间长短也有关.到达停顿点 c_i 前经过了 n_{c_i} 个路段以及 m_{c_i} 个停顿点,则到达 c_i 的时刻为 $t_{c_i}=T_0+\sum_{j=1}^{n_{c_i}}r_j.t+\sum_{j=1}^{m_{c_i}}c_j.t$, 以此为参考,取其前后各 $k\times\tau$时间段内整体移动对象 *o* 的历史数据进行统计,则有:

$$\Pr(c_i)=\left(\sum_{2k+1}\|(s_1^1\rightarrow s_2^1|_{T_0}^{\tau};c_i)\|\right)\Big/\sum_{2k+1}\|(s_1^1|_{T_0}^{\tau};c_i)\| \tag{12}$$

因此,事件 w_{c_i} 发生的概率为

$$\Pr(w_{c_i})=\Pr(c_i)\prod_{j=1}^{n_{c_i}}(1-\Pr(rd_j))\prod_{k=1}^{m_{c_i}}(1-\Pr(c_k)) \tag{13}$$

同理,可得 $\Pr(w_{r_i})=\Pr(r_i)\prod_{j=1}^{n_{r_i}}(1-\Pr(rd_j))\prod_{k=1}^{m_{r_i}}(1-\Pr(c_k))$,到达路段 rd_i 前经过了 n_{r_i} 个路段以及 m_{r_i} 个停顿点.

因此,在一个存在|*path*|路段和 *m* 个停顿点的路径上,发生事件 $s_1^1\rightarrow s_2^1$ 的概率为

$$\Pr(W)=1-\prod_{i=1}^{|path|}(1-\Pr(rd_i))\prod_{j=1}^{m}(1-\Pr(c_j)) \tag{14}$$

根据公式(14),可以进一步计算从 T_0 时刻开始到事件 *W* 发生时,移动对象所花费的时间期望 $E[T|W]$和距离期望 $E[L|W]$,以及事件 *W* 发生后到下次再次发生状态转换时的时间 $E[T_N|W]$和距离期望 $E[L_N|W]$:

$$E[T|W]=\left\{\sum_{i=1}^{|path|}\left[\Pr(w_{r_i})\cdot\left(\sum_{j=1}^{n_{r_i}}rd_j.t+\sum_{j=1}^{m_{r_i}}c_j.t\right)\right]+\sum_{i=1}^{m}\left[\Pr(w_{c_i})\cdot\left(\sum_{j=1}^{n_{c_i}}rd_j.t+\sum_{j=1}^{m_{c_i}}c_j.t\right)\right]\right\}\Big/\Pr(W) \tag{15}$$

$$E[L|W]=\left\{\sum_{i=1}^{|path|}\left[\Pr(w_{r_i})\cdot\sum_{j=1}^{n_{r_i}}rd_j.l\right]+\sum_{i=1}^{m}\left[\Pr(w_{c_i})\cdot\sum_{j=1}^{n_{c_i}}rd_j.l\right]\right\}\Big/\Pr(W) \tag{16}$$

在计算 $E[T_N|W]$和 $E[L_N|W]$时,仍然依照 partition-and-group 的思想,将距离$(0,l_{\max}]$以γ为单位划分为若干个连续的区间.设 $rc=\{rd_i,c_j|i=1,\ldots,|path|;j=1,\ldots,m\}$,则移动对象在 rc_i 上发生 $s_1^1\rightarrow s_2^1$ 后到再次发生状态转换,其间移动的距离为γ的 *j* 倍,其概率 $\Pr(rc_i,j\gamma)=\Pr\left\{L_N\in((j-1)\gamma,j\gamma]\middle|rc_i,T_0+\sum_{j=1}^{i}rc_j.t\right\}$.同样,以到达 rc_i 的时间 $T_0+\sum_{j=1}^{i}rc_j.t$ 为参考,取其前后各 $k\times\tau$时间段内整体移动对象 *o* 的历史数据进行统计,则有:

$$\Pr(rc_i,j\gamma)=\sum_{2k+1}\|(L_N\in((j-1)\gamma,j\gamma]|_{T_0}^{\tau};rc_i)\|\Big/\sum_{2k+1}\|L_N\in(0,l_{\max}]|_{T_0}^{\tau};rc_i\| \tag{17}$$

因此,事件 *W* 发生后,到下一次发生状态转换,其间移动的距离为 $j\gamma$的概率为

$$\Pr(j\gamma|W)=\sum_{i=1}^{|path|+m}\Pr(w_{rc_i})\Pr(rc_i,j\gamma)/\Pr(W) \tag{18}$$

移动距离的期望为

$$E[L_N|W]=\sum_j\left(\sum_{i=1}^{|path|+m}\Pr(w_{rc_i})\Pr(rc_i,j\gamma)\cdot j\gamma/\Pr(W)\right) \tag{19}$$

同理,也可以求得时间期望值 $E[T_N|W]$.

2.3 区域交通动力学特征φ_d

区域交通动力学特征以位置区域为观察对象,对区域内多个移动目标的动态移动行为进行抽取.

定义 6. 假设位置区域 r_i 内存在着多个移动对象的历史记录,记为 $O_{r_i}=\{o_j\in O: o_j在r_i处被观察到\}$,$\| O_{r_i} \|$ 为该区域内被观察到的移动对象的个数.

(1) 区域混杂性(diversity feature,f_{div})

统计一个区域内被访问的移动对象个体分布的差异,可以很好地区分不同位置区域的社会功能.设移动对象 o_j 在区域 r_i 内被观察到的次数为 $freq_i^j$,则其访问区域 r_i 的概率为

$$\Pr(o_j)=freq_i^j \Big/ \sum_{o_k\in O_{r_i}} freq_i^k \tag{20}$$

那么,参照公式(5)可以定义该区域的访问熵.很显然:f_{div} 越大,说明该区域人员流动越混杂,一般出现在商场、银行等公共区域;反之,则说明该区域更加具有私密性[24].

(2) 区域流动性(traffic feature,f_{traf})

① 路段上的转移花费时间

一个区域或一个路段上的转移时间花费是位置大数据分析中的重要特征.在定义 5 中,我们曾直接给出 $rd_i.t$ 为经过路段 rd_i 所有移动目标所花费的期望时间,但由于移动目标在移动工具、驾驶习惯等方面的差异巨大,导致简单的期望时间往往不够精确,因此需要引入一个特征时间向量$\langle tc_0,tc_1,tc_2,\ldots,tc_k\rangle$($tc_0$=0),用 k+1 个特征时间来反映路段 rd_i 的转移时间特性.设 $T_{rd_i}=\{t_1,t_2,...,t_m\}$ 为路段 rd_i 上观察到的所有转移时间,经过特征时间向量$\langle tc_0,tc_1,tc_2,\ldots,tc_k\rangle$可将 T_{rd_i} 划分为 k+1 个子集合,记为 $T^j=\{tc_{j-1}<t_i\leqslant tc_j\}$,$\nabla(T\cdot)$为其统计方差,$\langle tc_1,tc_2,\ldots,tc_k\rangle$满足:

$$\langle tc_0,tc_1,tc_2,...,tc_k\rangle=\arg\max\left\{\nabla(T_{rd_i})-\sum_{k+1}\| T^j \|\cdot\nabla(T^j)/\| T_{rd_i} \|\right\} \tag{21}$$

如何确定 k 的大小及各个特征时间,是其中的关键问题.一般来说,不同的路段有不同的 k 值.文献[26]采用了一种贪婪二分法去求取特征向量$\langle tc_0,tc_1,tc_2,\ldots,tc_k\rangle$,并将其所得到的特征转移时间分布来训练导航引擎,取得了很好的效果.

在各个路段转移时间特征提取的基础上,进一步对位置区域 r_i 内的流动性进行分析,几个常见的指标[14,27]包括:

② 区域内移动速度的期望 $E(v)$和方差 $D(v)$

假定在一个时间段 T 内,区域 r_i 内的交通特征稳定不变,则可通过对这个时段在区域上的所有轨迹数据$\{traj_1,traj_2,\ldots,traj_n\}$进行统计分析.设定轨迹 $traj_i=\langle p_1\to p_2\to\ldots\to p_{|traj_i|}\rangle$ 的总长度 $traj_i.l=\sum_{j=1}^{|traj_i|} dist(p_j,p_{j+1})$,时长 $traj_i.t=p_{|traj_i|}.t-p_i.t$,其中,$dist(p_j,p_{j+1})$表示 p_j 与 p_{j+1} 的距离,则有:

$$E(v)=\sum_{i=1}^{n} traj_i.l \Big/ \sum_{i=1}^{n} traj_i.t \tag{22}$$

进一步地,很容易计算方差 $D(v)$.

③ 区域交通流动性的时间演化性

进一步引入时间维度,可以对区域流动性的时间演化性[15]进行分析.

设区域流动性随时间变化的函数 $Traffic_i(t)$满足:

$$Traffic_i(t)=\alpha_i\cdot Traffic_i(t-\tau_i)+(1-\alpha_i)\cdot(1-\Pr\{o.\upsilon\leqslant E^{(t)}(v)\}) \tag{23}$$

其中,α_i 和 τ_i 是反映演化特性的两个参数.τ是区域交通流量的观察时间窗口,一般可以将一段时间内连续的 N 个位置记录 $p.v$ 构造成一个时序信号 $V(t)$,对其进行 DFT,获得其频率分布 $\hat{f}(\xi)=\sum_t \mathrm{e}^{-2\pi it\xi/N}V(t)$,有:

$$\begin{cases}\alpha_i=\max(\hat{f}(\xi))\\ \tau_i=1/\arg\max_{\xi}(\hat{f}(\xi))\end{cases} \tag{24}$$

(3) 区域聚散性(arriving-departure feature,$f_{\text{arr-dep}}$)

聚散性反映了一个位置区域中移动对象整体进入及离开的动力学模式,一般通过将一天划分为 k 个时段,继而统计 LBD 历史记录中该天每个时段“来到区域”和“离开区域”这两个行为的次数,以构造两个统计向量 $V_{\text{arr}}^{day_i}=[n_1^a,n_2^a,...,n_k^a]$ 和 $V_{dep}^{day_i}=[n_1^d,n_2^d,...,n_k^d]$,$day_i$ 表示第 i 天.在这两个向量统计的基础上,可以构造一批区域聚散性的二阶特征指标[5],如聚散比等:

$$ad\_radio=\left(\sum_{day_i}V_{\text{arr}}^{day_i}/V_{\text{dep}}^{day_i}\right)\Big/day_s \tag{25}$$

此外,我们也经常需要对移动目标离开(或到达)一个区域的时间进行估计.较为常见的方法是将其看作一个非齐次泊松过程(non-homogeneous Poisson process),对特征参数$\lambda(t)$进行估计[28].对路段 k 天内同一个时间片 ΔT 内的聚散事件 E 的数量进行观察,可得到似然函数:

$$L(\lambda)=\prod_{i=1}^{k}\frac{(\lambda\cdot\Delta T)^{N_i}}{N_i!}\mathrm{e}^{-\lambda\cdot\Delta T} \tag{26}$$

其中,N_i 表示第 i 天的时间片ΔT 内事件 E 发生的数量.令 $\mathrm{d}\ln(L(\lambda))/\mathrm{d}\lambda=0$,根据极大似然估计法求解$\lambda$,则可得:

$$\hat{\lambda}=\sum_{i-1}^{k}N_i/k\cdot\Delta T=\bar{N}/\Delta T \tag{27}$$

结果说明:可以根据多天内到达/离开区域的平均数量来对$\lambda(t)$进行建模,构建一个线性分段函数.

3 位置大数据降维分析及全局建模

3.1 位置大数据建模

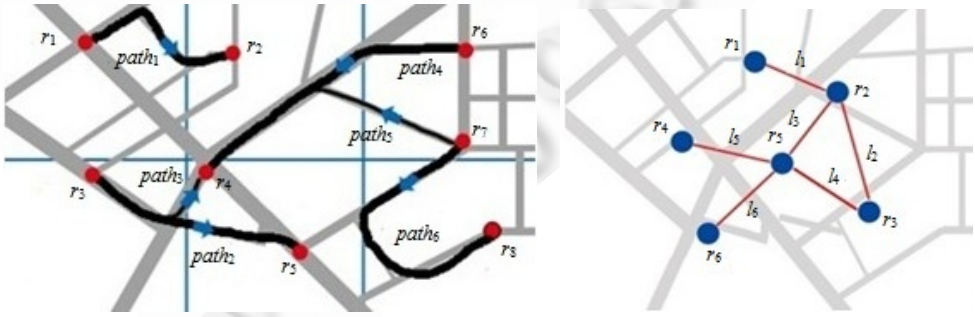
定义 7. 参考定义 3,我们可以将 LBD 所在的地图划分为 n 个区域,$\mathfrak{R}=\{r_1,r_2,\ldots,r_n\}$,分别计算 1 天内各个时间片 $Ts=\{t_1,t_2,\ldots,t_m\}$下的局部特征 $r.\varphi$,$\varphi=\langle\varphi_s,\varphi_{mp},\varphi_d\rangle$中的全部或部分特征,可以得到一个 $m\times n$ 的特征矩阵 M_1.再引入时间维度 $T=\{d_1,d_2,\ldots,d_D\}$,则可得到一个特征立方体 C_1.

M_1 和 C_1 是一种常见的 LBD 建模方式[27],其缺点是无法直观反映两两区域间的联系.因此将其变形,可以得到位置大数据网络 LBDN.这是一种利用网络科学[29,30]思想建立的有关位置大数据的全局模型,其主要目的就是为了通过网络全局特征建模[31],获得 LBD 数据的全局知识,从而参与到相关关联应用的分析中.

定义 8. 位置大数据网络为 $N=\{\mathfrak{R},L,\varphi\}$,其中,$\mathfrak{R}=\{r_1,r_2,\ldots,r_n\}$为 LBD 对应的地图的区域集合,$L$ 表示连接各区域的边的集合 $L=\{l_{ij}\}$.因此,N 也可以用一个 $n\times n$ 的邻接矩阵 $M_2=\{l_{ij}\}$表示,$l_{ij}=1$ 表示区域 r_i 和 r_j 之间存在移动性联系.

定义 9. l_{ij} 的权值反映了区域 r_i 和 r_j 之间移动性的关联,可根据具体情况定义不同的函数$\Omega\langle r_i.\varphi,r_j.\varphi\rangle$进行衡量,比如,$\Omega\langle r_i.\varphi,r_j.\varphi\rangle$可以表现为 r_i 和 r_j 之间的交通流量.

图 3 给出了用户轨迹、道路和 l_{ij} 的关系.一般地,在一个分区域地图上,一条用户轨迹 $traj=\langle p_1\rightarrow p_2\rightarrow\ldots\rangle$可以被粗粒化为 $traj=\langle r_i\rightarrow r_j\rightarrow\ldots\rangle$,如图 3(a)所示.依据轨迹的起止点可以建立 LBDN,如图 3(b)所示.



(a) 道路轨迹 (b) 位置大数据网络

Fig.3 Illustration of LBDN

图 3 LBDN 示意图

定义 10. 将时间维度 $Ts=\{t_1,t_2,\ldots,t_m\}$引入位置大数据网络为 $N=\{\mathfrak{R},L,\varphi\}$,则可得到一个 $n\times n\times m$ 位置立方体 C_2,其每个纵切面为一个时间尺度下的邻接矩阵,记为 $M_2^{(t_i)}$;每个横切面即 M_1,如图 4 所示.

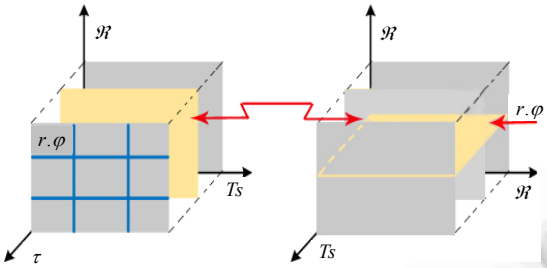


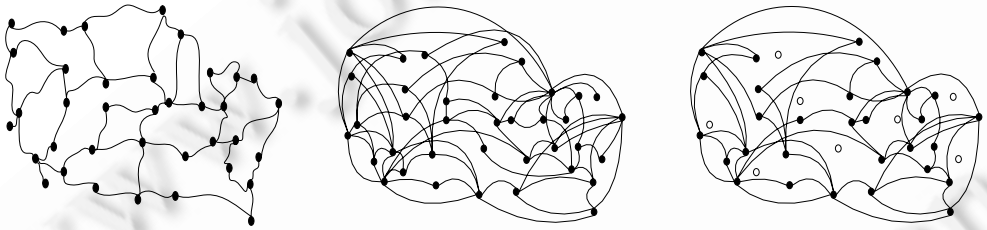
Fig.4 Global model of LBD (C_1,C_2)

图 4 位置大数据全局模型(C_1 和 C_2)

下面将阐述如何综合使用 M_1 和 M_2 这两类矩阵模型,构建 LBD 全局特征的方法.

3.2 空间尺度上的降维处理

降维分析(dimensionality reduction analysis)在大数据处理中是非常重要的环节,对于 LBDN 在空间尺度上的降维处理,其核心就是希望减少 $\mathfrak{R}=\{r_1,r_2,\ldots,r_n\}$中的区域或减少 L 中的边,通过关键分量的分析获得全局特征(如图 5 所示).



(a) 路网 (b) 降维前 (c) 降维后

Fig.5 Dimensionality reduction of LBDN

图 5 LBDN 的降维处理

一种直观的方法是比较 $r.\varphi$,选取特征显著的区域作为关键区域[32].考虑到LBDN主要是用来反映区域间移动特性的(这样的 LBDN 又被称为“交通网络”),下面我们着重介绍另外两种针对交通网络的空间降维方法.

① 依超介数进行降维

在交通网络中,结点 r_i 的介数(betweenness centrality)是反映结点交通重要性的标志特征[30],一般定义为网络中所有经过 r_i 的最短路径数量(因为大部分交通行为是按照最短路径原则展开的).

介数是以单条路径逐次计算而实现的,忽视了多条路径在交通中存在的关联性.所以我们对介数指标进行了改进,其核心思想是:

i) 如果网络中大量交通行为会同时选择两个结点 r_i 和 r_j 作为其最短路径的传播点,那么这两个结点的重要性是共生关系,记为 $\zeta_{ij}>0$.原始介数指标将这二者共同承担的那一部分重要性重复计算到各自结点中去,造成了重要性的高估;

ii) 如果对网络中结点 r_i 进行摘除后可以发现,原本那些以 r_i 为最短路径的链路大部分“取道”结点 r_j,说明结点 r_j 对结点 r_i 具有潜在的替代作用,记为 $\zeta_{ij}<0$.这种替代性在原始介数指标中未能体现,从而造成了结点交通重要性的低估.

定义 11. 结点 r_i 的重要性 $\overline{I_i}$ 一部分由自身的介数 I_i 所决定,而另一部分则通过其他结点与它的关联性来衡量,有:

$$\overline{I_i} = I_i + \lambda \cdot \left(\sum_{\{k|\zeta_{ki}>0\}} \alpha_{k\to i} + \sum_{\{k|\zeta_{ki}<0\}} \beta_{i\to k} \right) \tag{28}$$

其中,由于 r_i 在一定程度上与结点 $\{k|\zeta_{ki}>0\}$ 互有替代关系,因此,应该根据那些结点的重要性以及其与 r_i 的关联程度,提高对 $\overline{I_i}$ 的评估,记为 $\alpha_{k\to i}$,且 $\alpha_{k\to i}>0$;同理,由于 r_i 重要性的一部分是与之共生结点所共同承担的,应该在 I_i 中适当减除,记为 $\beta_{i\to k}$,且 $\beta_{i\to k}<0$. λ 是一个系数因子. $\overline{I_i}$ 又被称为结点的超介数.具体求解的方法可参考我们早期的论文[33,34].

② 依主分量进行降维

依超介数降维的方法针对的是 LBDN 中的 $\mathfrak{N}$,此外,还可以通过主分量分析(principal components analysis,简称 PCA)法对 L 进行空间降维分析[4].

3.3 时间尺度上的降维处理

在前文中,我们提及 $Ts=\{t_1,t_2,\ldots,t_m\}$,通常表示将一天划分到 m 个时间片去观察 LBD.事实上,只有观察到移动对象的整体移动模式在各自时间片下具有显著不同的差异时,划分的时间片才有意义.那么如何找寻和量化这种“显著差异”呢?不失一般性,我们在一天 24 小时内观察 O 中的每一个移动对象的某种移动行为特征 f.设 $f=\langle \varepsilon_1,\varepsilon_2,\ldots,\varepsilon_m\rangle$ 有 m 个取值,则对每个移动对象可以用二元组 $\langle time_i,\varepsilon_j\rangle$ 表示,得到全体样本集合 $S=\{\langle time_i,\varepsilon_j\rangle\}$,其中,$time_i$ 表示观察时刻.那么,在该样本下移动行为特征 f 的熵为

$$H(f) = -\sum_{i=1}^{m} \Pr(\varepsilon_i)\log_2 \Pr(\varepsilon_i) \tag{29}$$

假设将一天分为两个时间片,则按公式(29)分别计算各自时间片下各观察样本特征 f 的概率分布,记为 $P(f)$ 和 $Q(f)$.二者的距离仍然可以用 KL 散度来度量,有:

$$KL(P\|Q) = \frac{1}{2}\cdot\left[\sum_{i=1}^{m}\Pr_P(\varepsilon_i)\frac{\log_2 \Pr_P(\varepsilon_i)}{\log_2 \Pr_Q(\varepsilon_i)} + \sum_{i=1}^{m}\Pr_Q(\varepsilon_i)\frac{\log_2 \Pr_Q(\varepsilon_i)}{\log_2 \Pr_P(\varepsilon_i)}\right] \tag{30}$$

那么,只有当时间片由 $t_i=\text{argmax}_t KL(P\|Q)$ 划分时才是最有意义的.同样地,这种划分可以采用一种贪婪二分法继续深入下去,直到所划分的两个特征 f 的概率分布的距离差异小于阈值即可[26].

至此,可以得到一个灵活且相对精确的时间片划分 $Ts=\{t_1,t_2,\ldots,t_m\}$(所得到的每个时间片不一定与整点时刻对齐),同时使得 m 最小且有意义.

3.4 全局特征分布的模型生成

首先,在对 M_1 和 M_2(C_1 和 C_2)进行适当降维处理后,即可采用生成模型(generative model)的思路,从 LBD 观察样本数据构建整体的概率模型;然后,将这些模型中的参数当作一种更高阶的特征参与到具体的关联应用中,作为聚类、分类或关联挖掘的依据.

近几年来,随着自然语言处理(natural language processing,简称 NLP)技术的发展,其中很多经典的生成模型如隐含狄利克雷分配模型(latent Dirichlet allocation,简称 LDA)等已在社交网络计算、推荐系统等多个领域广泛应用[35−37].在位置大数据分析中,这一类生成模型也有很好的应用前景,比如,文献[16]将 LDA 模型用于通过用户行为模式发掘城市各区域的土地功能,取得了相当好的效果.

站在位置社会感知的角度,可以认为人们的移动行为(MP)都是与所在位置区域的社会语义上下文密切相关的,可以用 φ_{mp} 去度量.也就是说,每个 MP 的产生都可以看作是“先以一定概率选择了区域的某个社会语义特征(social semantic feature,简称 SSF),然后再从这个特征中以一定概率选择了特定 MP”而得到的,如图 6 所示.LDA 的目的之一就是通过观察样本去学习隐分布 θ 和 ϕ. $\theta_{r_i}=\langle\theta_{r_i,1},\theta_{r_i,2},\ldots,\theta_{r_i,Y}\rangle$ 是区域 r_i 的社会语义分布的非负归一化向量,表明该区域满足每种社会语义(共 Y 种)的概率.$\Pr(\theta|\alpha)$ 是 θ 的分布,被设置成为狄利克雷分布(Dirichlet distribution)形式,具体用来生成 θ,由参数 α 确定.ϕ 是一个条件分布 Pr(MP|SSF),表明各种社会语义下各

类移动行为发生的概率,是一个 $x \cdot Y$ 的矩阵,由参数 β 确定[35].

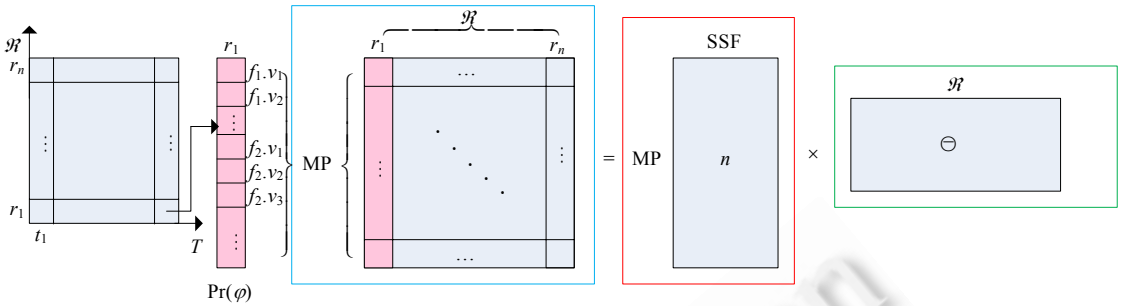


Fig.6 LDA based generative model of location data

图 6 基于 LDA 的位置数据生成模型

这种方法所最终确定的两个分布 θ 和 ϕ 在位置大数据分析中具有相当重要的作用:

- ϕ 表明了位置的社会语义与移动行为模式之间的关系.比如,文献[5]将移动模式行为(类似区域聚散性行为)与城市区域的功能用途联系起来.同样地,也可以使用其他移动模式特征与其他社会语义关联,从而达到通过位置大数据分析人类社会的目的;
- θ 是位置区域满足社会语义概率,这种概率向量其实也可以被进一步看作是区域移动模式特征.因此,θ 一方面可以作为位置社会感知的量化结果;另一方面,也可以作为一种区域位置特征的精炼表达形式,参与到其他数据挖掘的工作中去.

此外,通过对 LBDN 的观察,还可以获得社群整体移动的规律.尤其是在特殊时段或特殊事件发生前后,对社群整体移动行为建模,也是位置大数据建模的重要任务.

设在某特殊时段或事件发生时(记为事件 E)的社群移动行为是 $\Lambda=\langle r_s \to r_i \to \ldots \to r_d \rangle$,人群在各个位置区域上的移动轨迹称为 route ξ.各个位置区域上存在着一组特征向量 $\varphi=\langle f_1, f_2, \ldots \rangle$,每种特征都与人群的转移有内在联系,则在一个 route ξ 上移动的开销函数有:

$$\mathcal{C}(\Lambda,\varphi) \equiv \mathcal{C}(\xi \mid V_\lambda) = \sum_{r \in \xi} V_\lambda^T \cdot \varphi_r = \sum_{r \in \xi} (\lambda_1 \cdot f_1 + \lambda_2 \cdot f_2 + \ldots) = V_\lambda^T \varphi_\xi \tag{31}$$

其中,$V_\lambda=\langle \lambda_1, \lambda_2, \ldots \rangle$ 是一组权向量,用来确定人群在各个位置区域移动的成本开销.

假设事件 E 发生时,观察到人群具有初步的移动 $\xi_{r_s \to r_i}$,计算其最终达到区域 r_d 的概率满足贝叶斯公式:

$$\Pr(r_d \mid \xi_{r_s \to r_i}, V_\lambda) = \Pr(\xi_{r_s \to r_i} \mid r_d, V_\lambda) / \Pr(r_d) \tag{32}$$

其中,$\Pr(r_d)$ 是一个先验概率表明事件 E 发生后移动到 r_d 的人群比例. $\Pr(\xi_{r_s \to r_i} \mid r_d, V_\lambda)$ 由于 V_λ 不确定,则需通过最大熵原则(principle of maximum entropy),从观察的样本数据中进行学习.一般地:

$$P \equiv \arg\max_{\Pr(\xi \mid V_\lambda)} H(P) \tag{33}$$

运用拉格朗日乘数法,很容易求得满足条件最大熵原则的 $\Pr(\xi \mid V_\lambda)$,有:

$$\Pr(\xi \mid V_\lambda) = \exp\left(\sum_{r \in \xi} V_\lambda^T \cdot \varphi_r \right) \Big/ Z = \exp \mathcal{C}(\xi \mid V_\lambda) / Z \tag{34}$$

Z 是归一化因子,定义为

$$Z \equiv \sum_{\xi'} \exp\left(\sum_{r \in \xi'} V_\lambda^T \cdot \varphi_r \right) = \sum_{\xi'} \exp \mathcal{C}(\xi' \mid V_\lambda) \tag{35}$$

求解公式(34)的方法在马尔可夫决策过程(Markov decision process,简称 MDP)中经常会碰到,一般采用 GIS,IIS 等迭代算法[8,38,39],直到 $\Pr(\xi \mid V_\lambda)$ 收敛,获得 V_λ.那么,公式(32)中的 $\Pr(\xi_{r_s \to r_i} \mid r_d, V_\lambda)$ 也就可以很容易获得:

$$\Pr(\xi_{r_s \to r_i} \mid r_d, V_\lambda) = \sum_{\xi_{r_i \to r_d}} \exp \mathcal{C}(\xi \mid V_\lambda) \Big/ \sum_{\xi_{r_s \to r_d}} \exp \mathcal{C}(\xi \mid V_\lambda) \tag{36}$$

在 LDBN 实际运算过程中,可以用 Bellman-Ford 算法列举出 $r_s \to r_d$ 的若干最短路径即可.

通过上述MDP建模,我们可以观察多种事件E条件下社群的移动规律.在较早的论文中[2],我们还提到了使用隐马尔可夫模型(hidden Markov model,简称HMM)或动态贝叶斯网络(dynamic Bayesian network,简称DBN)的移动行为预测方法,对于社群行为推测仍然适用.

4 特征关联及协同挖掘

大数据研究中还有一个突出问题,即,数据稀疏性导致的结果失真.比如文献[15]发现,通过城市出租车轨迹密度所绘制的城市交通热点图与真实情况存在很大的偏差.原因是出租车群体往往比较喜欢在一些特定场所聚集,从而造成这些地方的观察数据密度过高.而真正需要密度数据的区域,由于缺少采集手段,却又无法获得真实的位置记录.

造成位置大数据稀疏的主要原因与采集手段和采集对象都有关:

(1) 一些位置区域天然地缺少移动对象访问(比如新修路段、偏远地区等);

(2) 特定的采集对象群体的移动偏好,使得有些位置区域很少被这一群体访问,而其他那些经常访问的群体又难以采集;

(3) 受到通信、存储等客观条件制约,使得采集数据在时间尺度上不连续,也会造成时间尺度上的稀疏.

针对这些问题,需要有效借助一些协同挖掘的方法,还原整体样本.

4.1 空间尺度上的协同挖掘

① 空间区域聚类

位置区域的关联聚类是解决空间数据稀疏的主要手段之一.主要采用区域的静态特征φ_s构成一组特征向量,用向量间的欧式距离(Euclidean distance)、马氏距离(Mahalanobis distance)、余弦相似性等常见的手段度量各特征向量的差异,并运用一种层级聚类方法,将特征相似的区域归于一类,从而用数据样本较多的区域去“替代”数据样本较少的区域[3,16].

② 协同过滤

与区域聚类思想不同,用协同过滤(collabora-tive filtering)算法对数据整体性进行补充.不妨设M_1矩阵记录了各时间片下各位置区域的交通流量(如某种f_{traf}),由于数据采集的问题造成了一定的区域数据缺失,如图7(a)所示.下面介绍用协同过滤的思想对缺失数据进行估计.

协同过滤实际上是在PCA的基础上基于矩阵分形而得到的一种数据项聚类(item clustering)方法,被广泛应用于推荐系统(recommender systems)中.常见的矩阵分形有UV分解(如图7(a)所示)、SVD分解等.

以UV分解为例,设$M_1=U\times V$.首先假设U,V矩阵内全为1,得到一个M_1',计算M_1和M_1'的方均根误差(root meansquare error,简称RMSE),有:

$$RMSE(M_1,M_1')=\sqrt{\sum_{m_{ij}\neq ?}(m_{ij}-m'_{ij})^2/\|M_1\|} \tag{37}$$

其中,$\|M_1\|$为原矩阵中非空数据的个数.随机调整U或V中的值,重新构造M_1',再次计算其与原矩阵的RMSE,…,反复以上步骤,控制$RMSE(M_1,M_1')$向越来越小的方向发展,从而用M_1'估计出M_1中缺失元素的值.

注意到:如果对u_{xy}进行猜测,则仅会影响M_1'的第x行元素.于是,只需对M_1'与M_1所有第x行的已知元素进行计算,便可得到由u_{xy}而带来的误差:

$$Error_{u_{xy}}=\sum_j(m_{xj}-m'_{xj})^2 \tag{38}$$

其中,$m'_{xj}=\sum_k u_{xk}v_{kj}=\sum_{k\neq y}u_{xk}v_{kj}+u_{xy}v_{yj}$.带入后对$Error_{u_{xy}}$求导,可得到使$RMSE$最小的$u_{xy}$为

$$u_{xy}=\left\{\sum_j v_{yj}\left(m_{xj}-\sum_{k\neq y}u_{xk}v_{kj}\right)\right\}\Big/\sum_j v_{yj}^2 \tag{39}$$

同理,对V的一个元素v_{xy}进行改变,使$RMSE$最小的值为

$$v_{xy}=\left\{\sum_i u_{ix}\left(m_{iy}-\sum_{k\neq x}u_{ik}v_{ky}\right)\right\}\Big/\sum_i u_{ix}^2 \tag{40}$$

依据公式(39)、公式(40)的显式解,可以直接完成最优解的搜索过程.

一般地,为了突出位置区域的局部特征在协同过滤中的作用[27],可以将空间区域聚类的思想引入协同过滤算法中.不妨构造一个辅助矩阵 $Z=\mathfrak{R}\times\varphi$,如图 7(b)所示,其每个行向量记为 φ_i.将各个位置区域的局部特征量化后形成一组特征向量,并构造一个过滤因子函数 $\Omega_i=V_\lambda^T\cdot\varphi_i$ (在实际计算中,不一定采用线性函数形式).将各个区域的 Ω_i 加入到 M_1' 中,得到 M_1''.反复调整参数向量 V_λ,并通过目标函数 $RMSE(M_1,M_1'')$ 控制优化搜索的方向,恢复更为准确的全局数据.

$$\underset{\boldsymbol{M}}{\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n_2-1} & ? \\ m_{21} & ? & & m_{2n_2-1} & m_{2n_2} \\ \vdots & & \ddots & & \vdots \\ m_{n_1-11} & & & & ? \\ ? & m_{n_12} & \cdots & m_{n_1n_2-1} & m_{n_1n_2} \end{bmatrix}}=\underset{\boldsymbol{U}}{\begin{bmatrix} u_{11} & \cdots & u_{1d} \\ u_{21} & & u_{2d} \\ \vdots & \ddots & \vdots \\ u_{n_11} & \cdots & u_{n_1d} \end{bmatrix}}\times\underset{\boldsymbol{V}}{\begin{bmatrix} v_{11} & v_{12} & & v_{1n_2} \\ \vdots & & \ddots & \vdots \\ v_{d1} & v_{d2} & & v_{dn_2} \end{bmatrix}}$$

(a) UV 分解

$$\begin{array}{c} \\ r_1 \\ r_2 \\ r_3 \end{array}\begin{array}{c} \begin{array}{cccc} F_1 & F_2 & F_3 & \end{array} \\ \begin{bmatrix} f_{11} & f_{12} & f_{13} & \cdots \\ f_{21} & f_{22} & & \\ f_{31} & & \ddots & \\ \vdots & & & \end{bmatrix} \end{array}\quad F$$

(b) 附加特征矩阵

Fig.7 Collaborative filtering of location data

图 7 位置数据的协同过滤

注意到,UV 分解或 SVD 分解得到的分形矩阵,在维度上要远远小于原矩阵,十分有利于大数据分析.最后,在位置大数据实际分析中,如果引入时间维度,则可能是对 C_1 或 C_2 进行协同过滤,这就需要使用高阶奇异值分解(high order singular decomposition,简称 HSVD)[27].同时,如果矩阵 M_1 或 M_2 中的“值”是一个由多个局部特征构成的特征向量,则公式(37)需要重新定义,引入向量距离的计算方法来度量整体差异.

4.2 时间尺度上的协同挖掘

将从 LBD 中提取的特征与某一特定应用关联,在时间尺度上很容易建立起概率图模型(graphical model).因此,时间尺度上的协同挖掘可以灵活地选择 HMM[40]、条件马尔可夫模型(conditional Markov model,简称 CMM)或条件随机场模型(conditional random fields,简称 CRF)[14]来处理.

在位置数据的概率图模型中,一般将状态序列 $s_1,s_2,\ldots$ 定义为需要透过位置大数据获得的知识(比如区域内人的情感、区域的环境污染情况等),将观测序列定义为位置特征 $\varphi_1,\varphi_2,\ldots$,其中,$\varphi_1,\varphi_2,\ldots,=f_1,f_2,\ldots$,运用图模型的 3 个基本问题,可以完成尺度上的协同挖掘.

5 位置大数据的应用框架

5.1 位置大数据的关联应用分析

在大数据科学研究中,一些与数据本身的来源和主体看似无关的对象,在经过数据价值提取和协同挖掘后往往会呈现出密切关联性.同样地,位置大数据的应用分析,将不仅仅被用来进行交通问题的分析,还能够提升我们对更为广泛的人类社会经济活动和自然环境的认识,从而体现其真正价值.事实上,LBD 完全可以与城市生活的经济运行情况(如 CPI 指数、绿色 GDP 指数、房价指数、城市投资与负债情况)、城市资源与环境情况(如 PM2.5)[14,27]、城市土地规划[5,16],甚至政府执政水平等联系起来,以位置数据特征作为观测值,以地理国情情况作为隐变量,进行全局建模和协同挖掘.

5.2 应用框架描述

我们在上述思路的基础上,提出了一种位置大数据管理应用分析的框架(location big data relational analysis framework,简称 LRAF),并将其实施到一个在线车联网位置服务系统 iWISE[41]中,作为一种位置数据的智能处

理中间件[42].图 8 描述了其基本结构和步骤:

i) 将社会经济及自然环境有关的数据与位置区域的特征数据进行相关分析,通过散点图绘制和 Pearson 相关等常见相关系数的计算和分析,找出人类行为和这些待分析事务间的关联性;

ii) 将这些关联性知识引入 LBDN 中,进行协同挖掘和全局数据建模,重点是对整体概率分布等参数进行学习,获得社会经济运行情况或自然环境分布情况的整体知识;

iii) 运用学习后得到的大数据整体知识,发现社会问题,引导社会行为.

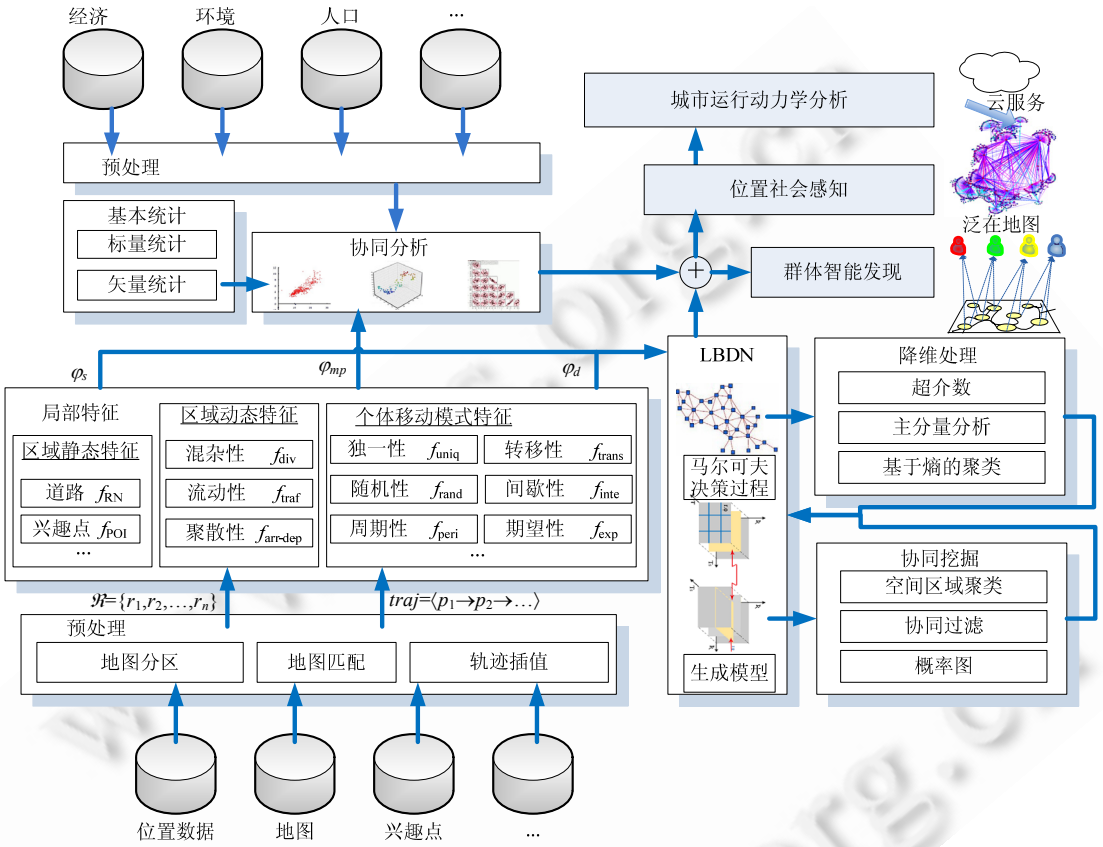


Fig.8 Location big data relational analysis framework

图 8 位置大数据关联应用分析的框架

6 结束语

本文以位置大数据为对象,针对其混杂性、复杂性和稀疏性,系统综述了如何进行价值提取和协同挖掘的方法.大数据科学研究强调“重关联,轻因果”,如何将位置数据与人类社会生活关联起来,是本文阐述的重点.这些方法被引入到一个位置服务软件框架中,将为智慧城市建设等提供支撑.

传统测绘强调对物理世界的测量.如果能够将对物理世界的测量结果(即位置)引申到对人类社会的测量中去,将极大地促进计算机和测绘科学的联系,形成一种智能化、社会化的泛在测绘计算[43].因此,本文所归纳和阐述的位置大数据分析方法具有重要的意义.

References:

[1] Liu JN. The recent progress on high precision applications of Beidou navigation satellite system. Report of the Stanford's 2012 PNT Challenges and Opportunities Symp. (SCPNT 2012), 2012. http://scpnt.stanford.edu/pnt/pnt2012.html

[2] Guo C, Fang Y, Liu JN, Wan Y. Study on social awareness computation methods for location-based services. Journal of Computer Research and Development, 2013,50(12):2531−2542 (in Chinese with English abstract).

[3] Yuan NJ, Zheng Y, Zhang LH, Xie X. T-Finder: A recommender system for finding passengers and vacant taxis. IEEE Trans. on Knowledge and Data Engineering, 2012,25(10):2390−2403. [doi: 10.1109/TKDE.2012.153]

[4] Chawla S, Zheng Y, Hu J. Inferring the root cause in road traffic anomalies. In: Proc. of the IEEE 12th Int'l Conf. on Data Mining (ICDM). Piscataway: IEEE, 2012. 141−150.

[5] Pan G, Qi GD, Wu ZH, Zhang DQ, Li SJ. Land-Use classification using taxi GPS traces. IEEE Trans. on Intelligent Transportation Systems, 2012, 14(1):113−123. [doi: 10.1109/TITS.2012.2209201]

[6] Song C, Qu Z, Blumm N, *et al*. Limits of predictability in human mobility. Science, 2010,327(5968):1018−1021. [doi: 10.1126/science.1177170]

[7] de Montjoye YA, Hidalgo CA, Verleysen M, Blondel UD. Unique in the CROWD: The privacy bounds of human mobility. Scientific Reports, 2013,3. [doi: 10.1038/srep01376]

[8] Song X, Zhang QS, Sekimoto Y, Horanont T, Ueyama S, Shibasaki R. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2013. 1231−1239. [doi: 10.1145/2487575.2488189]

[9] Sadilek A, Kautz HA, Silenzio V. Modeling spread of disease from social interactions. In: Proc. of the 6th Int'l AAAI Conf. on Weblogs and Social Media (ICWSM). Menlo Park: AAAI, 2012. 322−329. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4493%26lt%3B/4999

[10] Li GJ. The scientific value of the study in big data. Communications of the China Computer Federation, 2012,8(9):8−15 (in Chinese with English abstract).

[11] Wang S, Wang HJ, Qin XP, Zhou X. Architecting big data: Challenges, studies and forecasts. Chinese Journal of Computers, 2011, 34(10):1741−1752 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2011.01741]

[12] Wang YZ, Jin XL, Cheng XQ. Network big data: Present and future. Chinese Journal of Computers, 2013,36(6):1−15 (in Chinese with English abstract).

[13] Meng XF, Ci X. Big data management: Concepts, technique and challenges. Journal of Computer Research and Development, 2013, 50(1):146−169 (in Chinese with English abstract).

[14] Zheng Y, Liu F, Hsie HP. U-Air: When urban air quality inference meets big data. In: Proc. of the KDD. 2013. http://research.microsoft.com/pubs/193973/U-Air-KDD-camera-ready.pdf

[15] Liu SY, Liu YH, Ni LM, Fan JP, Li ML. Towards mobility-based clustering. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2010. 919−928. [doi: 10.1145/1835804.1835920]

[16] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2012. 186−194. [doi: 10.1145/2339530.2339561]

[17] Zhu B, Huang QX, Guibas L, Zhang L. Urban population migration pattern mining based on taxi trajectories. 2013. http://sensor.ee.tsinghua.edu.cn/mediawiki/images/b/b1/2013.Migration.BingZhu.pdf

[18] Li ZH, Ding BL, Han JW, Kays R, Nye P. Mining periodic behaviors for moving objects. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2010. 1099−1108. [doi: 10.1145/1835804.1835942]

[19] Ester M, Kriegel H-P, Sander J, Xu XW. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. of the KDD, Vol.96. 1996. 226−231. http://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

[20] Lou Y, Zhang CY, Zheng Y, Xie X, Wang W, Huang Y. Map-Matching for low-sampling-rate GPS trajectories. In: Proc. of the 17th ACM SIGSPATIAL Int'l Conf. on Advances in Geographic Information Systems. New York: ACM Press, 2009. 352−361. [doi: 10.1145/1653771.1653820]

[21] Yuan J, Zheng Y, Zhang CY, Xie X, Sun GZ. An interactive-voting based map matching algorithm. In: Proc. of 2010 the 11th Int'l Conf. on Mobile Data Management (MDM). Kansas City: IEEE, 2010. 43−52. [doi: 10.1109/MDM.2010.14]

[22] Liu KE, Li YG, He FC, Xu JJ, Ding ZM. Effective map-matching on the most simplified road network. In: Proc. of the 20th Int'l Conf. on Advances in Geographic Information Systems. New York: ACM Press, 2012. 609−612. [doi: 10.1145/2424321.2424429]

[23] Tang YZ, Zhu AD, Xiao XK. An efficient algorithm for mapping vehicle trajectories onto road networks. In: Proc. of the 20th Int'l Conf. on Advances in Geographic Information Systems. New York: ACM Press, 2012. 601−604. [doi: 10.1145/2424321.2424427]

[24] Cranshaw J, Toch E, Hong J, Kittur A, Sadeh N. Bridging the gap between physical location and online social networks. In: Proc. of the 12th ACM Int'l Conf. on Ubiquitous Computing. New York: ACM Press, 2010. 119−128. [doi: 10.1145/1864349.1864380]

[25] Lee JG, Han J, Whang KY. Trajectory clustering: A partition-and-group framework. In: Proc. of the 2007 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2007. 593−604. [doi: 10.1145/1247480.1247546]

[26] Yuan J, Zheng Y, Xie X, Sun GZ. T-Drive: Enhancing driving directions with taxi drivers' intelligence. IEEE Trans. on Knowledge and Data Engineering, 2013,25(1):220−232. [doi: 10.1109/TKDE.2011.200]

[27] Zhang FZ, Wilkie D, Zheng Y, Xie X. Sensing the pulse of urban refueling behavior. In: Proc. of the 2013 ACM Int'l Joint Conf. on Pervasive and Ubiquitous Computing. New York: ACM Press, 2013. 13−22. [doi: 10.1145/2493432.2493448]

[28] Zheng XD, Liang X, Xu K. Where to wait for a taxi? In: Proc. of the ACM SIGKDD Int'l Workshop on Urban Computing. New York: ACM Press, 2012. 149−156. [doi: 10.1145/2346496.2346520]

[29] Barabási AL, Albert R. Statistical mechanics of complex networks. Reviews of Modern Physics, 2002,74(1):47−97. [doi: 10.1103/RevModPhys.74.47]

[30] Newman MEJ. The structure and function of complex networks. Society for Industry and Applied Mathematics, 2003,45(2): 167−256.

[31] Yin CY, Wang BH, Wang WX, Zhou T, Yang HJ. Efficient routing on scale-free networks based on local information. Physics Letters A, 2006, 351(4):220−224. [doi: 10.1016/j.physleta.2005.10.104]

[32] Zheng Y, Liu YC, Yuan J, Xie X. Urban computing with taxicabs. In: Proc. of the 13th Int'l Conf. on Ubiquitous Computing. New York: ACM Press, 2011. 89−98. [doi: 10.1145/2030112.2030126]

[33] Guo C, Wang LN, Zhang XY. Study on network vulnerability identification and equilibrated network immunization strategy. IEICE on Information and System, 2012,E95-D(1):46−55. [doi: 10.1587/transinf.E95.D.46]

[34] Zhang XY, Wang LN, Guo C. Using game theory to reveal vulnerability for complex networks. In: Proc. of the 10th IEEE Int'l Conf. on Computer and Information Technology (CIT 2010). Bradford: IEEE, 2010. 978−984. [doi: 10.1109/CIT.2010.180]

[35] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003,3:993−1022.

[36] Wang Y, Agichtein E, Benzi M. Tm-Lda: Efficient online modeling of latent topic transitions in social media. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2012. 123−131. http://dl.acm.org/citation.cfm?id=2339552

[37] Porteous I, Newman D, Ihler A, Asuncion A, Smyth P, Welling M. Fast collapsed gibbs sampling for latent dirichlet allocation. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2008. 569−577. [doi: 10.1145/1401890.1401960]

[38] Song X, Zhang QS, Sekimoto Y, Horanont T, Ueyama S, Shibasaki R. An intelligent system for large-scale disaster behavior analysis and reasoning. IEEE, 2013. [doi: 10.1109/MIS.2013.35]

[39] Ziebart BD, Maas AL, Bagnell JA, Dey AK. Maximum entropy inverse reinforcement learning. In: Proc. of the 23th AAAI Conf. on Artificial Intelligence. AAAI, 2008. 1433−1438. http://www.aaai.org/Papers/AAAI/2008/AAAI08-227.pdf

[40] Sadilek A, Kautz H, Bigham JP. Finding your friends and following them to where you are. In: Proc. of the 5th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM Press, 2012. 723−732. [doi: 10.1145/2124295.2124380]

[41] Guo C, Liu JN. iWISE: A location-based service cloud computing system with content aggregation and social awareness. In: Proc. of the 10th Int'l Symp. on Location Based Services (LBS 2013). Shanghai, 2013.

[42] Wang ZH, Li JZ, Gao H. Data model for dirty databases. Ruan Jian Xue Bao/Journal of Software, 2012,23(3):539−549 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4042.htm [doi: 10.3724/SP.J.1001.2012.04042]

[43] Liu JN. The concept and progress of ubiquitous mapping and ubiquitous position. Digital Communication World, 2011,4:28−30 (in Chinese).

附中文参考文献:

[2] 郭迟,方媛,刘经南,万怡.位置服务中的社会感知计算方法研究.计算机研究与发展,2013,50(12):2531−2542.

[10] 李国杰.大数据研究的科学价值.中国计算机学会通讯,2012,8(9):8−15.

[11] 王珊,王会举,覃雄派,周烜.架构大数据:挑战、现状与展望.计算机学报,2011,34(10):1741−1752. [doi: 10.3724/SP.J.1016.2011.01741]

[12] 王元卓,靳小龙,程学旗.网络大数据:现状与发展.计算机学报,2013,36(6):1−15.

[13] 孟小峰,慈祥.大数据管理:概念、技术与挑战.计算机研究与发展,2013,50(1):146−169.

[42] 王志宏,李建中,高宏.一种非清洁数据库的数据模型.软件学报,2012,23(3):539−549. http://www.jos.org.cn/1000-9825/4042.htm [doi: 10.3724/SP.J.1001.2012.04042]

[43] 刘经南.泛在测绘与泛在定位的概念与发展.数字通信世界,2011,4:28−30.

郭迟(1983－),男,湖北武汉人,博士,讲师,CCF 会员,主要研究领域为位置服务,复杂网络,车联网系统.
E-mail: guochi@whu.edu.cn

刘经南(1943－),男,教授,博士生导师,中国工程院院士,主要研究领域为大地测量理论及应用.
E-mail: jnliu@whu.edu.cn

方媛(1990－),女,硕士生,CCF 学生会员,主要研究领域为位置服务.
E-mail: 315194877@qq.com

罗梦(1993－),女,本科生,主要研究领域为位置服务.
E-mail: 289523896@qq.com

崔竞松(1975－),男,博士,副教授,CCF 会员,主要研究领域为云计算,位置服务,智能汽车技术.
E-mail: cuijs@qq.com