一种基于用户移动行为相似性的位置预测方法

李昇智 乔建忠 林树宽
(东北大学计算机科学与工程学院 沈阳 110819)

摘 要 随着移动通信技术和车载定位系统的发展和广泛应用,基于位置服务越来越受到人们的关注。位置预测技术是其重要组成部分,并有着广泛的应用。在实际应用中,由于采集点丢失或新用户出现等,GPS 轨迹数据往往具有稀疏特性,使得基于单个用户数据的位置预测的准确率较低。针对这种情况,文中提出了基于移动行为相似性和用户聚类的 Markov 位置预测方法。首先,为使预测的位置具有物理意义,提出了基于 Voronoi 图的区域划分方法,并基于区域轨迹进行位置预测;其次,提出了同时考虑用户转移特性和用户区域特性的移动行为相似性计算方法;再次,根据移动行为相似性对用户进行聚类,并在聚类的用户组上采用一阶 Markov 模型进行位置预测,提高了位置预测的准确性。在真实 GPS 轨迹数据上的实验表明了所提方法的有效性。

关键词 移动行为相似性,转移概率矩阵,区域向量,位置预测
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.12.046

Location Prediction Method Based on Similarity of Users Moving Behavior

LI Sheng-zhi QIAO Jian-zhong LIN Shu-kuan
(School of Computer Science & Engineering, Northeastern University, Shenyang 110819, China)

Abstract With the development and widespread use of mobile communication technology and global positioning system, location-based service (LBS) receives extensive attention. Location prediction technology is an important part of LBS, and has wide application. In practical application, GPS trajectories are often sparse due to the sampling points lost or a new user appearing, which makes the accuracy of location prediction based on data of a single user low. To solve this problem, this paper proposed a novel Markov location prediction approach based on similarity of moving behavior and clustering of users. Firstly, in order to endow the locations with physical meaning, this paper proposed a region partitioning method based on Voronoi diagram. Then, the paper transformed the GPS trajectories into region trajectories and predicted the locations over region trajectories. Secondly, this paper put forward a new approach to measure the similarity of users' moving behavior by considering users' transfer features and regional features. Thirdly, based on the similarity of moving behavior, this paper divided users into various groups and applied the first-order Markov model on the groups for location prediction, which improves the accuracy of location prediction. The experiments over real GPS trajectory dataset indicate that the proposed method is effective for location prediction.

Keywords Moving behavior similarity, Transition probability matrix, Region vector, Location prediction

1 引言

随着移动设备、无线网络和定位技术的发展和广泛应用,可以获得的位置信息越来越多,基于位置服务[1-3]逐渐成为研究热点。位置预测不仅是基于位置服务必不可少的组成部分,还可以应用于其他基于位置服务中。例如,移动物体使用无线信号进行数据传输并发生基站切换时,可以通过位置预测提前预知下一个基站,以便提前配置资源,使得传输更加平稳,提高传输质量;在基于位置的广告服务中,服务提供商可以根据用户当前位置预测其运动趋势,在其进入某个区域之前,提供该区域的相关服务信息,不但有针对性地使用户及时接收到广告,而且可以节省信息传输的开销;在交通运输方面,对车辆的定位和预测可以使交通协调变得更加有效和快捷。可见,位置预测使得基于位置服务得到了丰富和扩展,对基于位置服务具有重要意义。Markov 模型由于符合移动对象的移动规律,被广泛用于位置建模和预测。然而,在实际应用中,GPS 设备采集的轨迹数据由于采集点丢失或面向新用户等原因往往具有稀疏性,影响了 Markov 位置预测模型的预测精度,这也是目前位置预测的准确度不高的原因之一。针对这种情况,本文提出一种基于用户移动行为相似性聚类

到稿日期:2017-07-20 返修日期:2017-10-18

李昇智(1974-),男,博士生,主要研究方向为智能计算,E-mail:qiaojianzhong@mail.neu.edu.cn(通信作者);**乔建忠**(1964-),男,博士,教授,CCF 会员,主要研究方向为人工智能、操作系统;**林树宽**(1966-),女,博士,教授,CCF 会员,主要研究方向为人工智能,E-mail:linshukuang@mail.neu.edu.cn。

的 Markov 位置预测方法，并考虑了用户转移特性和用户区域特性，提出了基于区域向量和用户转移概率矩阵的移动行为相似性计算方法，提高了位置预测的准确性。

2　相关工作

随着基于位置服务成为研究热点，越来越多的学者对位置预测方法进行了研究。目前常见的位置预测方法可归纳为 3 类。

(1)基于线性或非线性数学模型进行有效计算的位置预测方法[4-5]。这类方法主要通过建立数学模型来模拟移动对象的运动轨迹，根据当前的运行速度、运行时间，采用数学模型计算得到预测结果。这种方法假定移动对象的速度和方向在一定时间内不发生变化，对直线运行轨迹有较好的适用性。但是，在实际应用中，大多数移动对象在路网上运动，因此移动对象的移动方向和速度会受到路网条件的限制，很难满足此类模型的假设条件。

(2)基于频繁轨迹模式挖掘的预测方法。它是从大量的历史轨迹中找出频繁的轨迹模式，再将当前的查询轨迹与轨迹模式匹配，从而进行位置预测，可有效避免第一类方法出现的问题。文献[6]采用 PrefixSpan[7] 频繁模式挖掘算法，将搜索空间限制在一组投影数据库中，通过扫描投影数据库获得序列模式，并根据频繁序列模式获取相应的规则，从而进行位置预测。文献[8]采用一种类似 Apriori[9] 的 AprioriTraj 方法对历史轨迹进行挖掘，获得频繁轨迹模式并建立移动关联规则，然后通过模式匹配进行预测。此类方法在面向短轨迹数据或稀疏轨迹数据时可能存在无法挖掘频繁模式的情况，预测效果不理想。

(3)基于 Markov 模型的位置预测方法。Markov 模型由于可以很好地表示时序数据，而且经过统计和计算可以有效获得轨迹间的转换关系，因此被广泛用于位置建模和预测。文献[10]以 GPS 轨迹作为历史数据，通过建立一阶 Markov 模型对用户的移动位置进行预测。文献[11]基于历史模式树对 Markov 模型的阶数进行自动调节，并在此基础上进行位置预测。文献[12]融合一阶和多阶的 Markov 模型，通过建立混合模型进行位置预测。上述方法均依据单个用户的历史轨迹构建状态转移矩阵，未考虑新用户或数据稀疏的情况，影响了位置预测的准确性。为解决上述问题，本文通过建立区域向量和用户转移概率矩阵进行用户移动行为相似性计算，并在此基础上进行聚类，发现行为相似的用户，用行为相似的一类用户的历史轨迹进行位置预测，从而提高 Markov 模型的预测准确性。

3　问题定义及数据预处理

车载定位系统采集的原始 GPS 轨迹模式为$\langle p_1, p_2, \cdots, p_n\rangle$，$p_i=\langle id, time, longitude, latitude\rangle$表示编号为 id 的用户在 $time$ 时刻位于($longitude$, $latitude$)经纬度坐标的位置。因此，原始 GPS 轨迹是由没有实际意义的离散点构成的序列。为了预测具有物理意义的位置，本文并不对未来到达的轨迹点进行预测，而是基于关键交通枢纽进行地图区域划分，并将原始 GPS 轨迹数据转化为由区域构成的序列，位置预测则基于区域序列进行。

3.1　问题定义

定义 1(区域)　区域是指关键交通枢纽所覆盖的范围，表示为多边形 $S_c=(\langle f_c\rangle, \langle f_1, f_2, \cdots, f_n\rangle)$，其中，$f_c$ 是关键交通枢纽的位置，$\langle f_1, f_2, \cdots, f_n\rangle$是环绕 f_c 的多边形的 n 个顶点。

定义 2(区域轨迹)　区域轨迹是用户走过的区域序列，表示为 $Traj=\langle (S_1, t_1), \cdots, (S_i, t_i), \cdots, (S_m, t_m)\rangle$，其中 S_i ($1\leqslant i\leqslant m$)为区域，t_i 为用户初次到达区域 S_i 的时间。

定义 1 中的关键交通枢纽是采用文献[13]中的方法提取的具有一定规模和访问频繁程度的十字路口。区域和区域轨迹可分别由关键交通枢纽扩展而成以及由原始轨迹序列转化得到，具体的生成方法将在 3.2 节中介绍。

为方便表达，本文将区域轨迹简单表示为 $Traj=S_1, S_2, \cdots, S_i, \cdots, S_m$。

定义 3(位置预测)　给定用户已经走过的区域轨迹 $S_1, S_2, \cdots, S_l$，位置预测即在现有轨迹条件下，计算到达下一个区域的概率 $P(S_{\text{next}} \mid S_1, S_2, \cdots, S_l)$，并求得使该概率最大的区域 S_{next}。

本文采用一阶 Markov 模型进行位置预测，因此，预测结果 S_{pre} 可表示为：

$$S_{pre}=\arg\max_{S_{\text{next}}}\{P(S_{\text{next}} \mid S_l)\} \tag{1}$$

3.2　数据预处理

合理的区域划分将赋予预测出的位置一定的物理意义，从而提高位置预测的实用性。因此本文首先将地图划分成区域，然后将原始 GPS 轨迹转化为区域轨迹。

3.2.1　基于 Voronoi 图进行地图区域划分

传统的区域划分方法是将地图网格化，但是网格化的区域没有实际的物理意义。为此，本文提出将关键交通枢纽这样具有物理意义的位置作为顶点，基于 Voronoi 图[14]进行地图区域划分，将地图划分为以关键交通枢纽为中心的若干区域。由 Voronoi 图的性质[14-15]可知，每个 Voronoi 图内仅包含一个中心点(这里为关键交通枢纽)，且 Voronoi 图内的点到其中心点的距离最近，因此，以 Voronoi 图表示的区域是以关键交通枢纽为中心，由距离该关键交通枢纽最近(与到其他交通枢纽的距离相比)的点构成的区域。

将关键交通枢纽进行编号，以关键交通枢纽集合为输入进行区域划分的具体步骤如下。

步骤 1　以关键交通枢纽为顶点构成三角网；

步骤 2　对于任一关键交通枢纽 f_c，在已构建的三角网中找出以该关键交通枢纽为公共顶点的所有三角形；

步骤 3　计算每个三角形的外接圆圆心，设为 $f_1, f_2, \cdots, f_n$；

步骤 4　连接所有外接圆的圆心，即可得到以交通枢纽 f_c 为中心的 Voronoi 图，即以交通枢纽 f_c 为中心的区域 $S_c=(\langle f_c\rangle, \langle f_1, f_2, \cdots, f_n\rangle)$。

经过上述步骤，可以得到以每个交通枢纽为中心的区域。区域包括的信息有区域的编号(即对应的关键交通枢纽的编号)，以及按顺时针方向记录的多边形各个顶点的坐标。

3.2.2　将原始 GPS 轨迹转化为区域轨迹

为了使预测的位置具有物理意义，本文将原始 GPS 轨迹

转化为区域轨迹。将地图进行区域划分后，通过判断原始轨迹中的 GPS 点处于哪个区域内，就可以将原始 GPS 轨迹转化为区域轨迹。根据 Voronoi 图的性质[14-15]可知，GPS 点距离哪个交通枢纽最近，它就属于哪个区域。因此，轨迹转化问题转化为判断轨迹点是否处于 Voronoi 图的问题。为此，对于每一个原始轨迹中的 GPS 点，需要计算其到所有交通枢纽的距离，选择距离最近的交通枢纽。也就是说，必须遍历完整个区域集合后才能确定轨迹点最终属于哪个区域，即使遍历过程中已经得到最近的区域，也不能停止计算，因为无法确定后续计算是否还存在更近的区域。为了解决这一问题，考虑到 Voronoi 图是一种凸多边形[15]，本文利用性质 1 即可在遍历区域的过程中判断出 GPS 点所属的区域，无需再遍历整个区域集合。

性质 1 给定具有 n 个顶点的凸多边形 O 和轨迹点 $p_o(x_o,y_o)$，已知 O 的 n 个顶点按顺时针方向分别为 $f_1(x_1,y_1),f_2(x_2,y_2),\cdots,f_n(x_n,y_n)$，将 O 的 n 条边看作 n 个边向量，即 $\overrightarrow{f_1f_2},\overrightarrow{f_2f_3},\cdots,\overrightarrow{f_{n-1}f_n},\overrightarrow{f_nf_1}$。则点 $p_o(x_o,y_o)$ 在多边形 O 内，当且仅当 $p_o(x_o,y_o)$ 在 n 个向量的右侧。

因为多边形的顶点按顺时针方向记录，所以对于任意一个多边形边向量 $\overrightarrow{f_if_j}$，若要判断 $p_o(x_o,y_o)$ 是否位于向量 $\overrightarrow{f_if_j}$ 的右侧，只需判断向量 $\overrightarrow{f_ip_o}$ 与向量 $\overrightarrow{f_if_j}$ 的夹角 θ 是否位于区间 $[0,\pi]$，即 $\sin\theta\geqslant 0$ 是否成立。$\sin\theta$ 可由式(2)计算：

$$\sin\theta=\frac{\overrightarrow{f_ip_o}\times\overrightarrow{f_if_j}}{|\overrightarrow{f_ip_o}|\cdot|\overrightarrow{f_if_j}|} \tag{2}$$

其中，$|\overrightarrow{f_ip_o}|\cdot|\overrightarrow{f_if_j}|$ 是恒正的数，因此，$\sin\theta$ 的正负只取决于分子，故可得到性质 2。

性质 2 给定轨迹点 $p_o(x_o,y_o)$ 和由 n 个顶点 $f_1(x_1,y_1),f_2(x_2,y_2),\cdots,f_n(x_n,y_n)$(按顺时针顺序)表示的区域 O，若对于 O 的任一边向量 $\overrightarrow{f_if_j}$，式(3)均成立，则轨迹点 $p_o(x_o,y_o)$ 在区域 O 内。

$$\overrightarrow{f_ip_o}\times\overrightarrow{f_if_j}=(x_o-x_i)(y_j-y_i)-(y_o-y_i)(y_j-y_i)\geqslant 0 \tag{3}$$

其中，f_i 和 f_j 的坐标分别为 (x_i,y_i) 和 (x_j,y_j)。

依据性质 2 判断轨迹点是否属于某个区域时，只需要考查轨迹点对于区域的每条边来说是否都满足式(3)即可，这样就可以提前发现轨迹点的所属区域，而无需遍历整个区域集合。

在得到原始 GPS 轨迹上的轨迹点所属的区域后，就可以将原始 GPS 轨迹转化成区域轨迹。本文的位置预测在区域轨迹上进行。

4 基于用户移动行为相似性聚类的 Markov 位置预测

传统的 Markov 位置预测只依据单个用户的历史轨迹构建状态转移矩阵，在出现新用户或历史数据稀疏时，预测精度会降低。针对这种情况，本文对用户移动行为的相似性进行计算，并在此基础上进行用户聚类，利用一类用户的历史信息进行 Markov 位置预测，以提高预测精度。

4.1 基于区域向量和用户转移概率矩阵的移动行为相似性计算

为了计算用户移动行为的相似性，首先要生成区域向量和用户转移概率矩阵。相关定义如下：

定义 4(用户转移矩阵) 设地图划分的区域总数为 N，用户 i 的转移矩阵 M_i 是一个 $N\times N$ 矩阵，其中第 r 行、第 c 列($1\leqslant r\leqslant N,1\leqslant c\leqslant N$)的元素 $M_i(r,c)$ 是用户 i 的历史轨迹中从区域 r 转移到区域 c 的轨迹数目。

定义 5(区域向量) 用户 i 的区域向量 V_i 是一个 N 维向量，其第 r 个元素 $V_i(r)$ 表示用户 i 的历史轨迹中从区域 r 出发向所有区域转移的计数总和。

定义 6(用户转移概率矩阵) 用户 i 的转移概率矩阵 P_i 是一个 $N\times N$ 矩阵，其中第 r 行、第 c 列的元素 $P_i(r,c)$ 为用户 i 在当前位置是区域 r 的条件下转移到区域 c 的概率。

用户转移矩阵和区域向量可通过统计用户历史轨迹数据得出，用户转移概率可通过用户转移矩阵和区域向量计算得到(如式(4)所示)，从而建立用户转移概率矩阵。

$$P_i(r,c)=M_i(r,c)/V_i(r) \tag{4}$$

用户转移概率矩阵的每一行 $r(1\leqslant r\leqslant N)$ 可看作用户处于区域 r 时的转移特性的概率分布。

通过分析和观察可以发现，移动行为相似的用户具有以下特征：1)移动行为相似的用户频繁访问的区域也相似(本文称为用户区域特性)；2)移动行为相似的用户在相同区域选择下一步区域的概率接近(称为用户转移特性)。常用的相似性度量方法(如欧氏距离、马氏距离、余弦相似度等)难以度量用户移动行为的相似性。基于用户历史轨迹建立的区域向量和用户转移概率矩阵包含用户的历史移动信息，可以较为全面地代表用户的移动行为习惯并衡量用户间的移动行为相似性。其中，区域向量可表示用户对每个区域的偏好，从而体现用户区域特性；而用户转移概率矩阵可表示用户在区域间的转移概率，从而体现用户转移特性。因此，本文在度量用户移动行为相似性时，同时考虑了用户区域特性和用户转移特性，提出了基于区域向量和用户转移概率矩阵的相似性度量方法。

定义 7(移动行为相似度) 给定用户 i、用户 j 的用户转移概率矩阵 P_i 和 P_j，以及它们的区域向量 V_i 和 V_j。用户 i 和用户 j 的移动行为差异性 D_{ij} 定义为式(9)。用户 i 和用户 j 的移动行为相似度 Sim_{ij} 为差异性 D_{ij} 的倒数，即 $Sim_{ij}=1/D_{ij}$。约定 $0\log(0/0)=0,0\log(0/q)=0,p\log(p/0)=0$。

$$D_{ij}=\sum_{1\leqslant r\leqslant N}\Big[\frac{v_i(r)}{\sum_{1\leqslant k\leqslant N}v_i(k)}\cdot\Big(\sum_{1\leqslant c\leqslant N}p_i(r,c)\cdot\log\frac{p_i(r,c)}{p_j(r,c)}\Big)+\frac{v_j(r)}{\sum_{1\leqslant k\leqslant N}v_j(k)}\cdot\Big(\sum_{1\leqslant c\leqslant N}p_j(r,c)\cdot\log\frac{p_j(r,c)}{p_i(r,c)}\Big)\Big] \tag{5}$$

式(5)满足以下的相似性度量规范，因此可以作为距离度量方式。

1)当且仅当 $i=j$ 时，$D_{ij}=0$(自反性)；

2)$\forall i,j,i\neq j,D_{ij}>0$(非负性)；

3)$D_{ij}=D_{ji}$(对称性)；

4)$D_{ij}\leqslant D_{ik}+D_{kj}$(三角不等式)。

从式(5)可以看出，本文定义的用户移动行为相似性同时考虑了用户的区域特性和转移特性，对于移动用户聚类更有意义。

4.2 基于移动行为相似性的用户聚类

每个用户聚类可以代表一类典型用户，他们的移动行为

可以代表一类典型的移动行为。因此，将移动行为相似的用户聚类，可以有效提高位置预测的准确率。基于 4.1 节计算的用户移动行为相似性，本文借鉴文献[16]的方法，通过最大化类的内聚程度进行聚类，并采用进化博弈论中的“模拟者动态”(replicator dynamics)进行迭代求解。

定义 8(用户相似矩阵)　设 K 为参加聚类的用户总数，用户相似矩阵 A 是一个 $K*K$ 矩阵，元素 A_{ij} 为用户 i 和用户 j 之间的移动行为相似度($1\leqslant i\leqslant K, 1\leqslant j\leqslant K$)。为了方便，将对角线上元素的数值置 0。

定义 9(聚类概率向量)　聚类概率向量 z 是一个 K 维向量。其元素 z_i 为用户 i 出现在聚类 z 的概率($1\leqslant i\leqslant K$)。

特别地，当 $z_i=1, \forall j\neq i, z_j=0$ 时，意味着只有用户 i 包含在聚类中，而其他用户都不属于该聚类。

好的聚类应保证同一类中的用户具有好的内聚性，体现在用户相似矩阵中，即这些用户应该具有较大的移动行为相似度。式(6)可表示与向量 z 相对应的类的内聚程度[16]。

$$g(z)=z^{\mathrm{T}}\cdot A\cdot z \tag{6}$$

其中，$g(z)$越大则用户之间的关系越紧密，越有可能成为一类。这样就把用户聚类的问题转化为寻找合适的向量 z 使得类的内聚程度 $g(z)$达到最大值的问题 ，如式(7)所示：

$$\max_{z} g(z)=z^{\mathrm{T}}\cdot A\cdot z \tag{7}$$

为了求解式(7)，采用进化博弈论中的“模拟者动态”(replicator dynamics)方法，即利用迭代式(8)对式(7)进行迭代求解。

$$z_i(t+1)=z_i(t)\cdot\frac{(A\cdot z(t))_i}{z(t)^{\mathrm{T}}\cdot A\cdot z(t)}, i=1,\cdots,K \tag{8}$$

其中，t 为迭代次数，$z_i(t)$是用户 i 在第 t 次迭代中属于类 z 的概率，$(A\cdot z(t))_i$ 表示第 t 次迭代中用户 i 和其他用户的相似程度，分母表示所有用户间的相似程度。式(7)随着迭代过程不断递增并收敛于一个稳定值[16]，该值所对应的 z 就是所求的聚类概率向量。此时，向量 z 中的非零元素就是属于该类的用户。去掉已完成聚类的用户，建立新的聚类概率向量和用户相似矩阵，重复上述过程，直到所有用户都聚类为止。具体的聚类过程如算法 1 所示。

算法 1　基于移动行为相似性的用户聚类算法

输入：用户转移概率矩阵集 MPSet[]、区域向量集 MVSet[]、用户转移矩阵集 MSet[]

输出：聚类集合 Cluster[]，用户类的转移概率矩阵集 D[]

1. for each matrix m∈MPSet do
2.　for each matrix n∈MPSet do
3.　　按照式(5)计算矩阵 m 和 n 之间的相似度，并存入用户相似矩阵 A 中；
4. c=1，i=0；
5. while 用户集 U 非空 do
6.　初始化 z；
7.　利用 A 和 z 计算迭代式(8)直至收敛，得到新的向量 z；
8.　将向量 z 中的非零元素对应的用户放入聚类 Cluster[c−1]；
9.　删掉用户集 U、向量 z 和矩阵 A 中已聚类用户的信息，得到新的用户集 U、向量 z 和矩阵 A；
10.　c=c+1；
11. while i<c do
12.　将 Cluster[i]中所有用户的转移矩阵和区域向量加和保存到 CSet[i]和 CVSet[i]中；
13.　由 CSet[i]和 CVSet[i]，利用式(4)计算用户类 i 的转移概率矩阵 D[i]；
14.　i++；
15. return Cluster，D.

算法 1 中，第 1−3 行是计算用户移动行为相似性，建立用户相似矩阵的过程；第 4−10 是基于移动行为相似性进行用户聚类的过程；第 11−14 行是生成各用户类的转移概率矩阵的过程。

4.3　基于用户聚类的 Markov 位置预测

经过聚类算法虽然得到多个用户聚类，但是对于新用户而言则难以获得所属的用户聚类。因此本文基于贝叶斯的思想，在只考虑当前轨迹的情况下选择用户聚类。首先为当前轨迹引入代表类别的隐变量 C，然后根据贝叶斯公式计算用户当前轨迹 $S_1, S_2, \cdots, S_l$ 属于类别 C_k 的概率，如式(13)所示，选择概率最大者作为用户所属类别。

$$P(C=C_k\mid S_1,S_2,\cdots,S_l)=\frac{P(S_1,S_2,\cdots,S_l\mid C=C_k)\cdot P(C=C_k)}{P(S_1,S_2,\cdots,S_l)} \tag{9}$$

其中，分母对于所有类别都是相同的，因此只需要比较分子的大小。$P(C=C_k)$表示类别为 C_k 的先验概率，可由该聚类中用户的数量 N_{C_k} 和总用户数量 N_{total}求出，如式(10)所示：

$$P(C=C_k)=N_{C_k}/N_{\text{total}} \tag{10}$$

$P(S_1,S_2,\cdots,S_l\mid C=C_k)$是类别为 C_k 的聚类中轨迹 $S_1, S_2, \cdots, S_l$ 出现的概率，可通过聚类 C_k 的转移概率矩阵 D_{C_k}求出，如式(11)所示：

$$P(S_1,S_2,\cdots,S_l\mid C=C_k)=D_{C_k}(S_1,S_2)\cdot D_{C_k}(S_2,S_3)\cdot\cdots\cdot D_{C_k}(S_{l-1},S_l) \tag{11}$$

由上述过程可得出当前用户所属的聚类，基于此，Markov 位置预测可基于一类用户的转移概率矩阵进行。设用户聚类 C_k 的转移概率矩阵 D_{C_k} 如式(12)所示：

$$D_{C_k}=\begin{array}{c} \\ S_1\\ \vdots\\ S_i\\ \vdots\\ S_N\end{array}\begin{array}{c}\begin{array}{cccccc} S_1 & \cdots & S_j & \cdots & S_N\end{array}\\ \begin{pmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1N}\\ \vdots & \ddots & \vdots & \iddots & \vdots\\ p_{i1} & \cdots & p_{ij} & \cdots & p_{iN}\\ \vdots & \iddots & \vdots & \ddots & \vdots\\ p_{N1} & \cdots & p_{Nj} & \cdots & p_{NN}\end{pmatrix}\end{array} \tag{12}$$

其中，$S_1\sim S_N$ 为所有的区域，元素 p_{ij} 是从区域 S_i 到区域 S_j 的转移概率。按照本文的位置预测定义(见定义 3)，设用户当前所在的位置是区域 S_i($1\leqslant i\leqslant N$)，则矩阵 D_{C_k} 的第 i 行中概率最大的列 $j_{\max}$所对应的区域 $S_{j_{\max}}$ 就是用户最可能前往的下一个位置，如式(13)所示：

$$j_{\max}=\arg\max_{k}\{p_{ik}\}, 1\leqslant k\leqslant N \tag{13}$$

5　实验分析

为了验证所提位置预测方法的性能，本文在真实 GPS 轨迹数据集上进行了实验。所有程序的开发环境为 Intel(R) Core(TM2) Duo E8500 CPU，4GB 内存，500GB 硬盘，操作系统为 Windows XP。

实验所用的数据集是北京市 10357 辆出租车的 GPS 轨迹真实数据集[17]。本文将 GPS 轨迹数据分为两部分，一部分作为训练集，用来进行地图区域划分、建立用户转移概率矩阵以及进行用户聚类；另一部分作为测试集，用于评估位置预测的准确率。

本文基于 Voronoi 图，围绕重要交通枢纽进行区域划分，与通常的网格区域划分相比，所划分的区域更具物理意义。此外，不同的区域划分方法也会影响位置预测的准确性。图 1 对比了本文基于 Voronoi 图的区域划分方法与通常的网格区域划分方法对于位置预测准确率的影响。

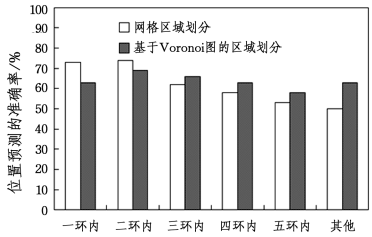

图 1 区域划分的有效性

Fig. 1 Effectiveness of region partitioning

从图 1 可以看出，不考虑所划分区域是否具有物理意义，单就区域划分对于位置预测准确性的影响而言，在二环以外，基于 Voronoi 图进行区域划分的位置预测的准确率高于网格区域划分，而在二环以内，情况相反。其原因在于，在二环以内，重要交通枢纽非常多，导致基于 Voronoi 图所划分的区域粒度过细，所预测的区域即使离实际区域非常近，但是可能分属于两个不同的区域，这种情况被认为是预测错误。在一般的区域粒度下(二环以外)，基于 Voronoi 图的区域划分所进行的位置预测，其准确率均高于通常的网格划分方法。

图 2 对比了不带用户聚类(Without Clustering，WITHOUTC)、基于密度的聚类(Density-Based Spatial Clustering of Applications with Noise，DBSCAN)以及本文基于移动行为相似性的用户聚类(User Clustering Based on Mobile Behavior Similarity，UCMBS)在预测结果上的准确率。整体来看，对用户进行聚类比不带有用户聚类的位置预测准确率更高。轨迹长度较小时，基于 UCMBS 的位置预测准确率低于基于 DBSCAN 的位置预测准确率，但是，随着轨迹长度的增加，基于 UCMBS 的预测准确率增长迅速，超过了 DBSCAN，这是因为轨迹长度越长，当前用户的轨迹信息越充分，得出用户所属的聚类也就越准确，从而预测准确性越高，而 DBSCAN 随着轨迹长度的增加一直采用相同的聚类阈值，使得预测准确率增加不明显。

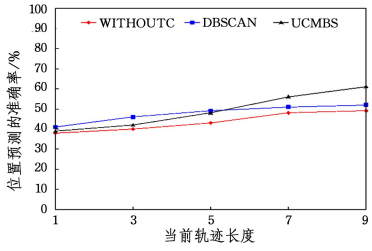

图 2 用户聚类对于位置预测准确性的影响

Fig. 2 Effect of clustering on location prediction accuracy

结束语 由于采集点丢失或新用户出现等原因，GPS 数据往往具有稀疏性，以往基于单个用户轨迹进行位置预测的准确率较低。针对这种情况，本文提出了基于用户移动行为相似性聚类的 Markov 位置预测方法。为了增加位置预测的实用性，本文在预处理阶段提出了基于 Voronoi 图的区域划分方法，并将原始 GPS 轨迹转化为区域轨迹，基于区域轨迹进行位置预测；在用户聚类的过程中，通过对用户移动行为进行分析，提出了同时考虑用户转移特性和区域特性的移动行为相似性测度，并在此基础上进行了用户聚类和 Markov 位置预测。真实数据上的实验表明，本文所提出的位置预测方法相比于不进行用户聚类和基于其他聚类方法的位置预测具有更高的准确率。

从实验结果可以看出，目前位置预测的准确率还不高，本文在预测环节只采用了简单的一阶 Markov 模型，因此，在未来的工作中，需对 Markov 预测环节进行改进，例如可考虑在 Markov 预测的过程中进行多阶或混合 Markov 位置预测，从而进一步提高预测准确性；另外，区域划分的过程中，在保证所划分区域具有物理意义的前提下，可考虑对区域划分粒度进行自适应的调整，避免划分粒度过细导致预测准确率降低的情况。

参 考 文 献

[1] ZHOU A Y，YANG B，JIN C Q，et al. Location-based services：architecture and progress [J]. Chinese Journal of computers，2011，34(7)：1155-1171. (in Chinese)
周傲英，杨彬，金澈清，等. 基于位置的服务：架构与进展 [J]. 计算机学报，2011，34(7)：1155-1171.

[2] BAO J，ZHENG Y，MOKBEL M F. Location-based and preference-aware recommendation using sparse geo-social networking data[C]// Proceedings of the 20th International Conference on Advances in Geographic Information Systems. USA：ACM，2012：199-208.

[3] LI H F，DONG L H，HAN J F. A Mobile Ordering Scheme Based on LBS[C]// Proceedings of the 4th International Conference on Emerging Intelligent Data and Web Technologies. China：IEEE，2013：398-401.

[4] TAO Y，FALOUTSOS C，PAPADIAS D，et al. Prediction and indexing of moving objects with unknown motion patterns[C]// Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. New York：ACM，2004：611-622.

[5] AGGAREAL C C，AGRAWAL D. On nearest neighbor indexing of nonlinear trajectories[C] // Proceedings of the 22th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. New York：ACM，2003：252-259.

[6] CHENG X L，XU X L，WANG Z Y，et al. Location Prediction Based On Sequenlial Mining [J]. Industrial Control Computer，2013，26(3)：70-72. (in Chinese)
程贤亮，徐小良，王中友，等. 基于序列挖掘的用户移动位置预测 [J]. 工业控制计算机，2013，26(3)：70-72.

[7] PEI J，HAN J W，ASL B M，et al. PrefixSpan：Mining Sequential Patterns Efficiently by Prefix Projected Pattern Growth[C]// Proceedings of the 17th International Conference on Data Engineering. 2001：215-224.

(下转第 307 页)

tain Graph Database[J]. Journal of Software, 2018, 29(10): 3150-3163. (in Chinese)
缪丰羽,王宏志. 一种不确定图数据库上的相似性连接方法[J]. 软件学报,2018,29(10):3150-3163.

[7] LIN J, SCHATZ M. Design patterns for efficient graph algorithms in MapReduce[C]// Eighth Workshop on Mining & Learning with Graphs(Mlg 10). 2010:78-85.

[8] PLIMPTON S J, DEVINE K D. MapReduce in MPI for Large-scale graph algorithms[J]. Parallel Computing, 2011, 37(9): 610-632.

[9] QIN L, YU J X, CHANG L, et al. Scalable big graph processing in MapReduce[C]// Proceedings of the ACM SIGMOD International Conference on Mangement of Data. 2014:22-27.

[10] KOLDA T G, PINAR A, PLANTENGA T, et al. Counting Triangles in Massive Graphs with MapReduce [J]. Siam Journal on Scientific Computing, 2013, 36(5): 48-77.

[11] CHEN Y, ZHAO X, XIAO C, et al. Efficient and Scalable Graph Similarity Joins in MapReduce[J]. The Scientific World Journal, 2014, 2014(3): 749028.

[12] CHEN Y F, ZHAO X, HE P J, et al. BMGSJoin: A MapReduce Based Graph Similarity Join Algorithm[J]. Pattem Recognition and Artificial Intelligence, 2015, 28(5): 472-480. (in Chinese)
陈一帆,赵翔,何培俊,等. BMGSJoin:一种基于 MapReduce 的图相似度连接算法[J]. 模式识别与人工智能,2015,28(5):472-480.

[13] PANG J, GU Y, XU J, et al. Efficient Graph Similarity Join with Scalable Prefix-Filtering Using MapReduce[J]. Lecture Notes in Computer Science, 2014, 8485: 415-418.

[14] SANJAY G, HOWARD G, SHAN-TAK L. The Google file system[J]. ACM SIGOPS Operationg Systems Review, 2003, 37(5): 29-43.

[15] AFRATI F N, ULLMAN J D. Optimizing joins in a map-reduce environment[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(9): 1282-1298.

[16] OKCAN A, RIEDEWALD M. Processing theta-joins using MapReduce[C]// ACM SIGMOD International Conference on Management of Data(SIGMOD 2011). Athens, Greece, DBLP, 2011: 949-960.

[17] YANG H C, DASDAN A, HSIAO R L, et al. Map-reduce-merge: simplified relational data processing on large clusters[C]// ACM SIGMOD International Conference on Management of Data. ACM, 2007: 1029-1040.

[18] AFRATI F N, SARMA A D, MENESTRINA D, et al. Fuzzy Joins Using MapReduce[C]// IEEE International Conference on Data Engineering. IEEE, 2012: 498-509.

[19] LUO W, TAN H, MAO H, et al. Efficient Similarity Joins on Massive High-Dimensional Datasets Using MapReduce[C]// IEEE International Conference on Mobile Data Management. IEEE Computer Society, 2012: 1-10.

[20] WANG Y, WANG H, LI J, et al. Efficient subgraph join based on connectivity similarity[J]. World Wide Web-internet & Web Information Systems, 2015, 18(4): 871-887.

[21] KIM Y, SHIM K. Parallel Top-K Similarity Join Algorithms Using MapReduce[C]// IEEE International Conference on Data Engineering. IEEE Computer Society, 2012: 510-521.

[22] MORRIS R, MORRIS R. ExOR: opportunistic multi-hop routing for wireless networks[C]// Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. ACM, 2005: 133-144.

[23] CEOL A, CHATR ARYAMONTRI A, LICATA L, et al. MINT, the molecular interaction database: 2009 update. [J]. Nucleic Acids Research, 2010, 38(suppl1): D532-D539.

[24] CHUA H N, SUNG W K, WONG L. Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions[M]// Data Mining for Biomedical Applications. Springer Berlin Heidelberg, 2006: 1623-1630.

(上接第 292 页)

[8] MORZY M. Prediction of moving object location based on frequent trajectories[C]// The 21st International Symposium on Computer and Information Sciences(ISCIS'2006). 2006: 583-592.

[9] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C]// Proceedings of the 20th International Conference on Very Large Data Bases. VLDB, 1994: 487-499.

[10] GIDÓFALVI G, DONG F. When and where next: individual mobility prediction [C] // Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems. ACM, 2012: 57-64.

[11] LV M Q, CHEN L, CHEN G C. Position Prediction Based on Adaptive Multi-Order Markov Model[J]. Journal of Computer Research and Development, 2010, 47(10): 1764-1770. (in Chinese)
吕明琪,陈岭,陈根才. 基于自适应多阶 Markov 模型的位置预测[J]. 计算机研究与发展,2010,47(10):1764-1770.

[12] YU X G, LIU Y H, WEI D, et al. Hybrid Markov mode for mobile path prediction [J]. Journal on Communications, 2006, 27(12): 61-69. (in Chinese)
余雪岗,刘衍珩,魏达,等. 用于移动路径预测的混合 Markov 模型[J]. 通信学报,2006,27(12):61-69.

[13] CHEN Z B, SHEN H T, ZHOU X F. Discovering popular routes from trajectories[C]// Proceedings of the 27th ICDE International Conference on Data Engineering. Germany, IEEE, 2011: 900-911.

[14] DEMIRYUREK U, SHAHABI C. Indexing network voronoi diagrams[C]// Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2012: 526-543.

[15] CHEN C. The establishment and application of voronoi diagram in computer mapping[J]. Acta Geodaetica et Cartographica Sinica, 1987, 16(3): 223-231. (in Chinese)
陈春. 泰森多边形的建立及其在计算机制图中的应用[J]. 测绘学报,1987,16(3):223-231.

[16] PAVAN M, PELILLO M. Dominant sets and pairwise clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(1): 167-172.

[17] YUAN J, ZHENG Y, XIE X, et al. Driving with knowledge from the physical world[C]// Proceedings of International Conference on the 17th ACM SIGKDD Knowledge Discovery and Data Mining. USA, ACM, 2011: 316-324.