第37卷第3期 东北大学学报(自然科学版) Vol.37,No.3
2016年3月 Journal of Northeastern University(Natural Science) Mar. 2016

doi:10.3969/j.issn.1005-3026.2016.03.005

基于用户移动行为相似性聚类的 Markov 位置预测

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

摘 要:由于采集点丢失或出现新用户等原因,GPS 轨迹数据往往具有稀疏性,使得基于单个用户数据的位置预测准确率较低.针对这种情况,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,基于 Voronoi 图和原始 GPS 轨迹进行区域划分,位置预测基于区域轨迹进行;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.真实 GPS 轨迹数据上的实验表明了所提出方法的有效性.

关 键 词:移动行为相似性;转移概率矩阵;区域向量;聚类概率向量;位置预测
中图分类号:TP 391 **文献标志码**:A **文章编号**:1005-3026(2016)03-0323-04

Markov Location Prediction Based on User Mobile Behavior Similarity Clustering

LIN Shu-kuan, LI Sheng-zhi, QIAO Jian-zhong, YANG Di
(School of Computer Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: LIN Shu-kuan, E-mail: linshukuan@mail.neu.edu.cn)

Abstract: GPS trajectories are often sparse due to the sampling points lost or new users appearing, which makes the accuracy of location prediction low based on the data of a single user. To solve this problem, a novel Markov location prediction approach was proposed based on user mobile behavior similarity and user clustering. First, the map was partitioned into various regions based on Voronoi diagram and original GPS trajectories. And then locations were predicted over region trajectories. Second, a new approach was proposed to measure the similarity of users' mobile behaviors by considering users' transferring features and regional features. Third, based on the mobile behavior similarity, users were divided into various groups and the first-order Markov model was applied on the groups to predict users' locations. Therefore, the accuracy of location prediction was improved. The experiments over real GPS trajectory dataset indicate that the proposed method is effective for location prediction.

Key words: mobile behavior similarity; transition probability matrix; region vector; clustering probability vector; location prediction

随着移动设备、无线网络和定位技术的发展和广泛应用,可以获得越来越多的位置信息,基于位置服务[1-3]逐渐成为研究热点.位置预测是基于位置服务的重要组成部分.Markov 模型由于符合移动对象移动规律,被广泛用于位置建模和预测[4-6].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,使得依据单个用户的历史轨迹构建状态转移矩阵的 Markov 位置预测方法准确率不高.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的 Markov 位置预测方法,通过建立用户转移概率矩阵和区域向量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

收稿日期:2015-02-05
基金项目:国家自然科学基金资助项目(61272177).
作者简介:林树宽(1966-),女,吉林长春人,东北大学教授;乔建忠(1964-),男,辽宁兴城人,东北大学教授,博士生导师.

进行位置预测,从而提高 Markov 模型的预测准确性.

1 数据预处理及问题定义

在位置预测之前,进行数据预处理,将原始 GPS 轨迹转化为区域轨迹. 首先,采用文献[7]中的方法提取具有一定规模和访问频繁程度的十字路口作为关键交通枢纽,在此基础上,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[8]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,并将原始 GPS 轨迹数据转化为区域轨迹.

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 S_{next}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

$$S_{\text{pre}}=\underset{S_{\text{next}}}{\operatorname{argmax}}\{p(S_{\text{next}}\mid S_l)\}\ . \tag{1}$$

2 基于用户移动行为相似性聚类的 Markov 位置预测

2.1 区域向量和用户转移概率矩阵的建立

定义 1 (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 $\boldsymbol{M}_i$ 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 $\boldsymbol{M}_i(r,c)$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 2 (区域向量)用户 i 的区域向量 $\boldsymbol{V}_i$ 是一个 N 维向量,其第 r 个元素 $\boldsymbol{V}_i(r)$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 3 (用户转移概率矩阵)用户 i 的转移概率矩阵 $\boldsymbol{P}_i$ 是一个 $N\times N$ 矩阵,其中第 r 行、第 c 列的元素 $\boldsymbol{P}_i(r,c)$为用户 i 在当前位置为区域 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从用户历史轨迹数据统计得出,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(2)所示,从而建立用户转移概率矩阵:

$$\boldsymbol{P}_i(r,c)=\boldsymbol{M}_i(r,c)/\boldsymbol{V}_i(r)\ . \tag{2}$$

2.2 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

给定用户 i、用户 j 的用户转移概率矩阵 $\boldsymbol{P}_i$,$\boldsymbol{P}_j$,以及它们的区域向量 $\boldsymbol{V}_i,\boldsymbol{V}_j$. 用户 i 相对于用户 j 的移动行为的差异性可表示为式(3):

$$D_{i\to j}=\sum_{1\leqslant r\leqslant N}W_i^rD_{ij}^r=\sum_{1\leqslant r\leqslant N}\frac{\boldsymbol{V}_i(r)}{\sum\limits_{1\leqslant k\leqslant N}\boldsymbol{V}_i(k)}\times\left(\sum_{1\leqslant c\leqslant N}\boldsymbol{P}_i(r,c)\cdot\lg\frac{\boldsymbol{P}_i(r,c)}{\boldsymbol{P}_j(r,c)}\right). \tag{3}$$

其中:括号中的量是基于相对熵的思想,利用用户转移概率矩阵计算的用户 i 与用户 j 同处于区域 r 时转移特性的概率分布差异;括号前的量是基于区域向量计算的用户 i 对区域 r 的偏好(区域特性).

同理,用户 j 相对于用户 i 移动行为的差异性,可计算如式(4):

$$D_{j\to i}=\sum_{1\leqslant r\leqslant N}W_j^rD_{ji}^r=\sum_{1\leqslant r\leqslant N}\frac{\boldsymbol{V}_j(r)}{\sum\limits_{1\leqslant k\leqslant N}\boldsymbol{V}_j(k)}\times\left(\sum_{1\leqslant c\leqslant N}\boldsymbol{P}_j(r,c)\cdot\lg\frac{\boldsymbol{P}_j(r,c)}{\boldsymbol{P}_i(r,c)}\right). \tag{4}$$

综合式(3)与式(4),用户 i 和用户 j 的移动行为差异性 D_{ij}定义为

$$D_{ij}=\sum_{1\leqslant r\leqslant N}\left[\frac{\boldsymbol{V}_i(r)}{\sum\limits_{1\leqslant k\leqslant N}\boldsymbol{V}_i(k)}\cdot\left(\sum_{1\leqslant c\leqslant N}\boldsymbol{P}_i(r,c)\cdot\lg\frac{\boldsymbol{P}_i(r,c)}{\boldsymbol{P}_j(r,c)}\right)+\frac{\boldsymbol{V}_j(r)}{\sum\limits_{1\leqslant k\leqslant N}\boldsymbol{V}_j(k)}\cdot\left(\sum_{1\leqslant c\leqslant N}\boldsymbol{P}_j(r,c)\cdot\lg\frac{\boldsymbol{P}_j(r,c)}{\boldsymbol{P}_i(r,c)}\right)\right]. \tag{5}$$

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 $\text{sim}_{ij}=1/D_{ij}$.

2.3 基于移动行为相似性的用户聚类

基于 2.2 节计算的用户移动行为相似性,本文借鉴文献[9]中的方法,通过最大化类的内聚程度进行聚类.

定义 4 (用户相似矩阵)设 K 为参加聚类的用户总数,用户相似矩阵 $\boldsymbol{A}$ 是一个 $K\times K$ 矩阵,元素 $\boldsymbol{A}_{ij}$为用户 i 和用户 j 间的移动行为相似度($K\geqslant i,j\geqslant 1$). 为了方便,将对角线上元素置为 0.

定义 5 (聚类概率向量)聚类概率向量 $\boldsymbol{z}$ 是一个 K 维向量. 其元素 z_i 为用户 i 出现在聚类 $\boldsymbol{z}$ 的概率($K\geqslant i\geqslant 1$).

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,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 $\boldsymbol{z}$ 相对应的类的内聚程度[9]:

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

$g(\boldsymbol{z})$越大则用户之间的关系越紧密,则越有可能成为一类,这样就把用户聚类的问题转化为寻找合适的向量 $\boldsymbol{z}$,使得类的内聚程度 $g(\boldsymbol{z})$达到

最大值，如式(7)所示：

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

具体聚类过程如算法 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 $\boldsymbol{D}$[ ]

1)计算用户相似矩阵 $\boldsymbol{A}$；

2)$c=1$；$i=0$；

3)while 用户集 U 非空 do

4)　对式(7)进行求解，得到向量 z；

5)　向量 z 中非零的元素对应的用户放入聚类 Cluster[$c-1$]；

6)　删掉用户集 U、向量 z 和矩阵 $\boldsymbol{A}$ 中已聚类用户的信息，得到新的用户集 U、向量 z 和矩阵 $\boldsymbol{A}$；

7)　$c=c+1$；

8)while $i<c$ do

9)　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；

10)　由 CSet[i]和 CVSet[i]利用式(2)计算用户类 i 的转移概率矩阵 $\boldsymbol{D}$[i]；

11)　$i++$；

12)return Cluster，$\boldsymbol{D}$

2.4　基于用户聚类的 Markov 位置预测

经过 2.3 节的聚类算法得到多个用户聚类，在基于聚类的位置预测中，首先基于贝叶斯的思想为用户确定所属聚类，即计算用户当前轨迹为 $S_1,S_2,\cdots,S_l$ 时属于类别 C_k 的概率，如式(8)所示，选择概率最大者作为用户所属类别.

$$p(C=C_k \mid S_1,S_2,\cdots,S_l) = \frac{p(S_1,S_2,\cdots,S_l \mid C=C_k)\cdot p(C=C_k)}{p(S_1,S_2,\cdots,S_l)} \ . \tag{8}$$

式(8)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. 其中：$p(C=C_k)$表示类别为 C_k 的先验概率，可由该聚类中用户的数量和总用户数量求出；$p(S_1,S_2,\cdots,S_l \mid C=C_k)$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 $\boldsymbol{D}_{C_k}$求出.

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 $\boldsymbol{D}_{C_k}$如式(9)所示：

$$\boldsymbol{D}_{C_k} = \begin{array}{c} \\ S_1 \\ \vdots \\ S_i \\ \vdots \\ S_N \end{array} \begin{array}{c} \begin{array}{cccccc} S_1 & \cdots & S_j & \cdots & S_N \end{array} \\ \begin{pmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1N} \\ \vdots & & \vdots & & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{iN} \\ \vdots & & \vdots & & \vdots \\ p_{N1} & \cdots & p_{Nj} & \cdots & p_{NN} \end{pmatrix} \end{array} . \tag{9}$$

设用户当前所在的位置是区域 S_i ($1 \leqslant i \leqslant N$)，则 S_i 所在行概率最大者所对应的列就是用户最可能前往的下一个位置.

3　实验分析

本文实验环境为 Intel(R) Core(TM2) Duo E8500 CPU，4 GB 内存，500 GB 硬盘，操作系统为 Windows XP. 所用的数据集是北京市 10 357 辆出租车的 GPS 轨迹真实数据集[10].

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.

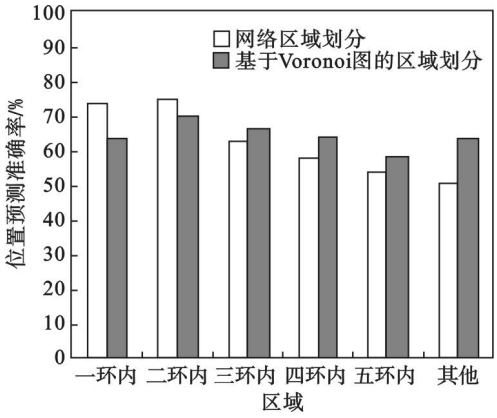

图 1　区域划分的有效性

Fig. 1　The effectiveness of region partitioning

从图 1 可以看出，在二环以外，基于 Voronoi 图进行区域划分的位置预测准确率高于网格区域划分，而在二环以内，出现相反的情况. 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，被划为了预测错误的情况.

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，DBSACN)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.

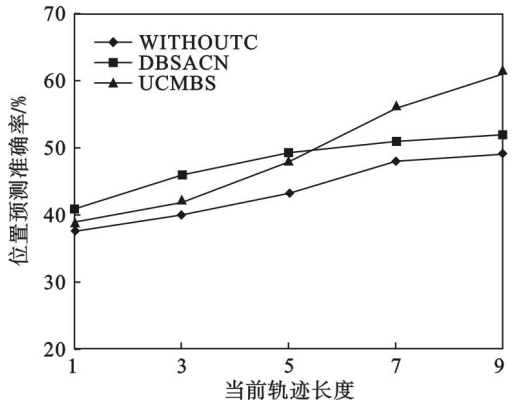

图 2 用户聚类对于位置预测准确性的影响

Fig. 2 The effect of clustering on location prediction accuracy

4 结 语

由于采集点丢失或出现新用户等原因,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,并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. 真实数据上的实验表明了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.

参考文献:

[1] 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.
(Zhou Ao-ying, Yang Bin, Jin Che-qing, et al. Location-based services: architecture and progress [J]. *Chinese Journal of Computers*, 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. Redondo Beach, California, 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. Xi'an, 2013: 398 - 401.

[4] 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. Redondo Beach, California, 2012: 57 - 64.

[5] 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.
(Lyu Ming-qi, Chen Ling, Chen Gen-cai. Position prediction based on adaptive multi-order Markov model[J]. *Journal of Computer Research and Development*, 2010, 47 (10): 1764 - 1770.)

[6] 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.
(Yu Xue-gang, Liu Yan-heng, Wei Da, et al. Hybrid Markov mode for mobile path prediction [J]. *Journal on Communications*, 2006, 27 (12): 61 - 69.)

[7] 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. Hannover, 2011: 900 - 911.

[8] 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.
(Chen Chun. The establishment and application of Voronoi diagram in computer mapping [J]. *Acta Geodaetica et Cartographica Sinica*, 1987, 16(3): 223 - 231.)

[9] Pavan M, Pelillo M. Dominant sets and pairwise clustering [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(1): 167 - 172.

[10] 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. San Diego, 2011: 316 - 324.