TraDR:一种基于轨迹分解重构的移动社交网络位置预测方法

薛 迪 吴礼发 李华波 洪 征
(解放军理工大学指挥信息系统学院 南京 210007)

摘 要 随着移动社交网络的不断发展,利用用户发布的位置信息为其提供基于地域的个性化推荐服务不仅给用户提供了便利,也为商户带来了巨大的潜在利益。位置预测技术作为此类服务中的关键技术,是移动社交网络中的重要研究内容之一。结合移动社交网络的特点,提出了基于轨迹"分解-重构"的位置预测方法TraDR,利用公开易得的先验知识,为用户建立个性化的位置推理模型,有效解决了常见位置预测算法所面临的"轨迹数据稀疏问题"。基于真实数据集的实验验证了该预测方法在预测有效性及效率方面的优越性。

关键词 移动社交网络,位置预测,数据稀疏问题,签到轨迹
中图法分类号 TP393.08 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.3.019

TraDR:A Destination Prediction Method Based on Trajectory Decomposition and Reconstruction in Geo-social Networks

XUE Di WU Li-fa LI Hua-bo HONG Zheng
(Institute of Command Information System,PLA University of Science and Technology,Nanjing 210007,China)

Abstract With the development of geo-social networks,the practice of utilizing the locations published by GeoSN users to offer them personalized reference services not only benefits users,but also brings the business providers potential profits. As the fundamental enabling technology of the location-based reference services,destination prediction becomes one of the most significant research topics in GeoSNs. Considering the features of GeoSNs,this paper proposed a novel destination prediction method named TraDR specially for GeoSNs based on trajectory decomposition and reconstruction to construct personalized inference model for each target GeoSN user,which not only solves the "trajectory data sparsity problem" faced by common location inference models,but also takes advantages of the rich commonly available public background information. Experiments based on real world dataset were carried out,and results prove the high performance of the presented method both in prediction accuracy and running efficiency.

Keywords Geo-social network,Destination prediction,Data sparsity problem,Check-in trajectory

1 引言

随着智能手机、平板电脑等移动终端的普及以及无线通信与定位技术的进步,兼具位置感知与社交网络服务功能的移动社交网络(Geo-Social Network,GeoSN)迅速发展[1]。"带位置标签的内容分享"、"签到(Check-in)"等主流移动社交网络服务允许用户发布自己的位置信息,GeoSN自身或其他第三方应用可利用用户发布的位置信息对其目的地进行推测,并以此为基础向用户提供形式多样的个性化推荐服务[2],如基于地域的商户推荐、定向广告、自动路线设计等。此类服务将物理空间、信息空间和人的活动三者无缝地融合,为用户带来便利的同时也为商家带来巨大的潜在利益,受到了越来越多的关注。而位置预测作为影响其服务质量的重要基础环节,具有重要的研究意义。

位置预测主要是利用用户历史位置所形成的轨迹来分析挖掘用户的移动特征,并据此推算某一时间段内用户出现在某一个位置的概率。目前典型的位置预测算法大多以贝叶斯框架为基础[3-6],但此类算法普遍面临"轨迹数据稀疏"问题[7,8],即用户的历史轨迹样本集通常无法覆盖所有可能的轨迹,而当用户的当前轨迹不与数据集中的任何一条轨迹样本的整体或局部相匹配时,模型将失去预测能力,无法获知用户未来的位置。

近来虽然已有学者针对"轨迹数据稀疏问题"提出了相应的解决方案[7],但该方案主要是针对传统LBS服务设计的,其中涉及的用户通常是匿名的,缺乏GeoSN中公开易得的用户属性、社交关系等先验知识,算法只是单纯地利用用户位置信息本身进行预测。但实际上,用户的移动性不仅蕴含于历史位置信息中,而且会因受到社会因素(如社交关系、用户相似度等)、地理因素(如道路连通性、最大时速限制等)的影响而发生变化:首先,现实生活中,好友结伴出行是非常普遍的

到稿日期:2015-01-09 返修日期:2015-05-04 本文受江苏省自然科学基金项目(BK20131069)资助。
薛 迪(1990-),女,硕士生,主要研究方向为网络安全与隐私保护,E-mail:dixue_nj@126.com;吴礼发(1968-),男,教授,博士生导师,主要研究方向为网络安全;李华波(1981-),男,博士,讲师,主要研究方向为网络安全;洪 征(1978-),男,副教授,硕士生导师,主要研究方向为网络安全。

现象，Cho 等人在文献[9]中通过分析来自真实移动社交网络的签到数据发现，用户 40%以上的移动行为与其好友有关。文献[10,11]表明，通过分析与用户相关的人(如亲朋好友等)的行为来预测用户行为，比单纯分析用户本身所取得的预测效果更佳。其次，有相似移动特征的用户可能有相同的兴趣爱好[12]，进而导致更加相似的移动轨迹，因而，相似用户的轨迹数据对于目标用户的位置预测有重要的参考意义。再次，用户的移动性受限于现实地理因素(如路网连通性、最大时速限制等)的制约，在一定的时间间隔内，研究区域内的相当一部分地点是用户根本不可达的，据此可以过滤掉若干不可达兴趣点(Position of Interest，POI)，从而极大地减少冗余计算，提高算法效率。综上，单纯依据位置信息进行预测在一定程度上限制了预测结果的准确率及效率。

本文针对现有位置预测方法的缺陷和不足，将传统 LBS 中的位置预测算法进行扩展迁移，提出了适用于移动社交网络的基于子轨迹"分解-重构"的位置预测方法 TraDR (Trajectory Decomposition and Reconstruction)。该方法以贝叶斯推理框架为基础，主要分为离线建模和在线预测两个阶段。离线建模阶段首先利用用户相似度、社交关系等先验知识为目标用户构建个性化轨迹数据集，然后通过对数据集中轨迹的分解与重构，建立马尔科夫推理模型，计算得出目标用户在各 POI 之间的转移概率；在线预测阶段首先根据路网信息过滤不可达目的地，然后针对目标用户的当前轨迹，利用离线阶段得出的转移概率计算出其潜在目的地，从而完成位置预测。本方法的主要优势如下：

· 利用社交关系及用户相似度为用户建立个性化轨迹数据集，增强了位置预测的针对性，进而提高了预测结果的准确率。

· 引入道路网络信息(主要包括道路连通性及最大时速限制等)对不可达目的地进行初步过滤，大大减少了无意义的计算开支；以贝叶斯推理框架为基础，将大量计算转为线下进行，提高了在线预测效率。

· 通过对原始轨迹的"分解-重构"，有效应对了移动社交网络位置预测中所面临的"轨迹数据稀疏问题"，增强了模型的预测能力。

2 基于轨迹"分解-重构"的位置预测方法概述

2.1 基本概念及表示

· 签到(c)是指用户 u_k 在时刻 t 通过 GeoSN 发布其所在位置 l 的过程，对应的签到记录记为 $c^k=(l^k,t^k)$；

· 轨迹(tra)是指用户 u_k 在某一时间段内连续发布的签到位置序列，记为 $tra_k=\langle l_1,l_2,\cdots,l_i,\cdots,l_n\rangle$；

· 查询轨迹(tra^q)是指目的地需要预测的目标用户 u_k 未完成的当前轨迹，记为 $tra_k^q=\langle l_s,l_2,l_3,\cdots,l_c\rangle$，其中 l_s 和 l_c 分别为用户 u_k 的起始位置点和当前位置点。

鉴于本文所研究的位置预测方法主要用于个性化推荐服务，所需的用户位置并不一定要某个精准的 POI，为简化问题且不失一般性，本文将目标区域划分若干边长为 λ 的栅格，位于同一个栅格内的不同 POI 将由同一个位置对象——栅格 g 来表示，而相应的签到及轨迹将分别表示为 $c^k=(g^k,t^k)$，$tra_k=\langle g_1,g_2,\cdots,g_i,\cdots,g_n\rangle$，其中 $g_i=(x_i,y_i)$。例如图 1 中由位置序列$\langle l_1,l_5,l_6,l_9\rangle$组成的轨迹 tra_1 可表示为$\langle g_1,g_4,g_5,g_8\rangle$。

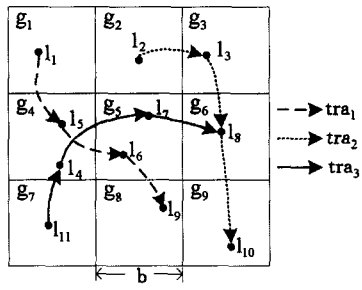


图 1 目标区域栅格划分及签到轨迹示意图

2.2 先验知识

合理利用各类先验知识对于提高预测算法的准确率及运算效率具有十分积极的作用[9-12]。移动社交网络中，用户主页包含丰富的位置与非位置信息，第三方开发者可通过调用 GeoSN 服务提供商提供的开放 API 来获取所需的信息，如用户的签到记录、好友关系等。此外，城市交通网络及路段限行速度等地理信息可通过各类公开的数据库轻易获得，因而本文主要考虑以下 3 类信息作为位置预测算法的先验知识。

· 历史信息：用户 $u_i(i\in[1,n])$的签到记录的集合 C_i，且有 $C=\bigcup_{i=1}^{n}C_i$；

· 社交信息：用户 $u_i(i\in[1,n])$的好友列表 F_i；

· 地理信息：目标区域的路网 $G(V,E,S)$，其中 G 为无向图，E 是代表路段的边集，V 是代表路段交叉点的点集，S 是表征各路段限行速度的权重集。

2.3 TraDR 基本思想

本文提出的基于轨迹"分解-重构"的位置预测方法 TraDR 以目前位置预测研究中应用最广泛的贝叶斯推理框架为基础。在给定查询轨迹 tra_k^q、最大移动时间 Δt_m 的前提下，条件概率 $p(l_d\in g_i\mid\Delta t_m,tra_k^q)$即表征目标用户 u_k 的目的地位于位置栅格 g_i 范围内的可能性，因而对目标用户的位置预测问题可转化为对该条件概率(称为目的地概率)的求解。

由贝叶斯定理可知

$$p(l_d\in g_i\mid\Delta t_m,tra_k^q)=\frac{p(l_d\in g_i,\Delta t_m\mid tra_k^q)}{p(\Delta t_m)} \tag{1}$$

其中，$p(\Delta t_m)$在给定 tra_k^q 和 Δt_m 的情况下为常数，因此有

$$\begin{aligned}&p(l_d\in g_i\mid\Delta t_m,tra_k^q)\propto p(l_d\in g_i,\Delta t_m\mid tra_k^q)\\&=p(\Delta t_m\mid l_d\in g_i,tra_k^q)\cdot p(l_d\in g_i\mid tra_k^q)\end{aligned} \tag{2}$$

上式的被乘数 $p(\Delta t_m\mid l_d\in g_i,tra_k^q)$(称为可达性概率)主要衡量在 Δt_m 时间内用户从当前位置 l_c 到目的位置 g_i 的可达性，用于对可能的目的区域进行初步筛选，从而减少不必要的计算。可达性概率可依据目标区域的路网连通性及交通情况(如最大限行速度等)计算得出，如式(3)所示，即只有从当前位置移动到目的位置所需时间不超过最大移动时间 Δt_m 时，取值为 1，否则为 0。

$$p(\Delta t_m\mid l_d\in g_i,tra_k^q)=\begin{cases}1, & \Delta t_m\geqslant t_d-t_c\\0, & \text{else}\end{cases} \tag{3}$$

利用贝叶斯定理进一步推导式 (2) 中的乘数 $p(l_d\in g_i\mid tra_k^q)$，可得

$$p(l_d\in g_i\mid tra_k^q)=\frac{p(tra_k^q\mid l_d\in g_i)p(l_d\in g_i)}{\sum_j p(tra_k^q\mid l_d\in g_j)p(l_d\in g_j)} \tag{4}$$

其中先验概率 $p(l_d\in g_i)$易由式(5)计算得出：

$$p(l_d \in g_i)=\frac{|\{tra|l_d \in g_i\}|}{|\{tra\}|} \tag{5}$$

而式(4)中的后验概率 $p(tra_k^q|l_d \in g_i)$ 衡量了在目的地为 g_i 的情况下，用户从 g_s 出发沿轨迹 tra_k^q 到达 g_c 并最终到达 g_i 的概率，其中 g_s 和 g_c 分别为查询轨迹 tra_k^q 的起始位置和当前位置。后验概率可由式(6)计算得出：

$$p(tra_k^q|l_d \in g_i)=\frac{[p(tra_k^q)\cdot p_{ci}^t]}{p_{si}^t} \tag{6}$$

其中，p_{ci}^t 和 p_{si}^t 分别是从栅格 g_s 和 g_c 出发到达 g_i 的转移概率；$p(tra_k^q)$ 是用户严格按照轨迹 $tra_k^q=\langle g_s,g_2,\cdots,g_{n-1},g_c\rangle$ 中所涉及的位置点移动的概率(称为轨迹概率)，其大小可由式(7)计算得出：

$$\begin{aligned}p(tra_k^q)&=p(\langle g_s,g_2,\cdots,g_{n-1},g_c\rangle)\\&=p_{s2}^t\cdot(\prod_{i=2}^{n-2}p_{i(i+1)}^t)\cdot p_{(n-1)c}^t\end{aligned} \tag{7}$$

其中，$p_{i(i+1)}^t$ 为用户从栅格 g_i 到 g_{i+1} 的转移概率。若已知用户在相关位置栅格间的转移概率，将其代入式(7)便可求得轨迹概率，进而利用式(2)—式(6)可求得位置 g_i 的目的地概率的相对大小。最终，目的地概率最大的前 κ 个位置便是用户最有可能到访的目的地。

以上即为本文位置预测方法 TraDR 的基本思想，其中转移概率 $p_{i(i+1)}^t$ 的求解将主要由 TraDR 的离线建模阶段完成，用户最有可能到访的 κ 个目的地主要由 TraDR 的在线预测阶段完成。下面第 3、4 节将分别对这两个阶段的算法进行具体介绍。

3 TraDR 离线建模算法

TraDR 的离线建模算法主要通过对用户的个性化建模来求解转移概率矩阵，从而获取用户在各个位置栅格间的转移概率。离线建模阶段主要包含个性化数据集筛选、马尔科夫模型建立及转移概率矩阵计算 3 个步骤。

3.1 个性化数据集筛选

人们的移动性具有一定的规律，且这些规律会因地理环境、社会因素等的影响和制约而存在个体差异，因此诸如文献[7]中利用所有用户的全部轨迹数据建立唯一的推理模型来预测所有人的位置的做法显然限制了预测结果的正确率。鉴于此，TraDR 结合 GeoSN 的特点，利用公开易得的用户信息及社会信息，首先对建模所需的历史轨迹数据进行筛选，为目标用户建立个性化轨迹数据集。所用的筛选指标及规则如下。

(1)好友亲密度

研究表明，好友关系对人们的移动行为有很大影响[9]，且关系越密切的好友，结伴出行的可能性越大，因而其移动特征越相似，这意味着好友的移动数据对于用户的移动特征挖掘及位置预测有十分重要的意义。鉴于此，本文引入变量 $Fclos$——好友亲密度[13]，用于衡量目标用户 u_t 与其好友 u_f 的亲密程度，如式(8)所示。

$$Fclos(u_t,u_f)=\frac{|F_t\cap F_f|}{\min(|F_t|,|F_f|)} \tag{8}$$

当 u_f 与 u_t 的好友亲密度高于阈值 φ 时，便将 u_f 的历史签到数据集 C_f 并入 u_t 的个性化轨迹数据集 T_t 中，即：

$\forall u_f \in F_t, t\in[1,n], f\in[1,n]$，若 $Fclos(u_t,u_f)\geqslant\varphi$，则 $T_t=T_t\cup C_f$。

(2)用户相似度

拥有相似移动行为的用户可能有共同的兴趣爱好，意味着未来有巨大可能访问相同 POI。鉴于此，本文引入变量 $Usim$——用户相似度，并采用以下相似度度量公式计算其值：

$$Usim(u_t,u_k)=\frac{\sum_{g\in I_{t,k}}P_{t,g}P_{k,g}}{\sqrt{\sum_{g\in I_{t,k}}P_{t,g}^2}\sqrt{\sum_{g\in I_{t,k}}P_{k,g}^2}} \tag{9}$$

其中，$I_{t,k}$ 为用户 u_t 和 u_k 到访位置的并集，$P_{t,g}$ 和 $P_{k,g}$ 分别为用户 u_t 和 u_k 对位置 g 的访问概率。当 u_k 与 u_t 的用户相似度高于阈值 θ 时，便将 u_k 的历史签到数据集 C_k 并入 u_t 的个性化轨迹数据集 T_t 中，即：

$\forall u_k \in F_t, t\in[1,n], f\in[1,n]$，若 $Usim(u_t,u_k)\geqslant\theta$，则 $T_t=T_t\cup C_k$。

利用上述两个规则对完整的数据集进行筛选，最后将获得目标用户 u_t 的个性化轨迹数据集 T_t，T_t 中涉及的签到位置栅格的集合记为 G_t。

3.2 马尔科夫模型建立

为充分利用 T_t 中的轨迹数据并克服“轨迹数据稀疏问题”，首先将 T_t 中的所有轨迹分解为长度为 2 的子轨迹，并据此建立一阶马尔科夫模型。具体方法如下：签到位置集 G_t 中的每个位置栅格对应马尔科夫模型中的一个状态，用户从一个位置栅格移动到相邻的另一个位置栅格则对应马尔科夫模型中的状态转移，由此，模型中的一步状态转移概率矩阵 M^1 可表示为一个二维矩阵，矩阵元素 p_{ij}^1 表示从栅格 g_i 直接移动到与其相邻的栅格 g_j 的概率，大小可由式(10)计算得出。

$$p_{ij}^1=p(g_j|g_i)=\frac{|\{tra|\langle g_i,g_j\rangle\subset tra,tra\in T_t\}|}{|\{tra|\langle g_i\rangle\subset tra,tra\in T_t\}|} \tag{10}$$

以图 1 为例，假设图上 3 条轨迹为数据集 T_t 中的全部轨迹，则与之相对应的栅格位置集合将包含图上所有栅格，即 $G_t=\{g_1,g_2,\cdots,g_9\}$。进而轨迹 $tra_1=\langle g_1,g_4,g_5,g_8\rangle$ 将被分解为 $\langle g_1,g_4\rangle$，$\langle g_4,g_5\rangle$，$\langle g_5,g_8\rangle$ 这 3 条子轨迹，并分别用于计算 p_{14}^1，p_{45}^1，p_{58}^1 这 3 个一步转移概率。利用同样的方式对轨迹 tra_2 和 tra_3 进行处理，最终得到的马尔科夫模型及一步状态转移概率 M^1 如图 2 所示。

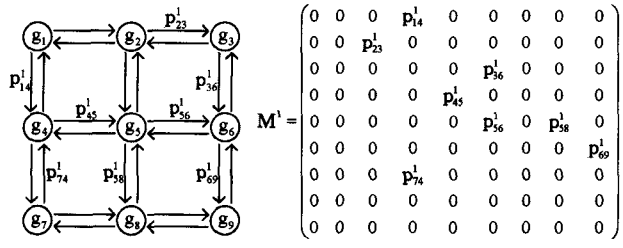


图 2 马尔科夫模型及一步状态转移概率矩阵(样例)

3.3 总转移概率矩阵计算

上文求得的一步状态转移概率矩阵 M^1 中的元素对应的是用户从一个栅格区域直接移动到与之相邻的栅格区域的概率，但如果两个位置区域在空间上不相邻，则二者之间无法通过一步转移到达，也就意味着 M^1 中对应的元素为 0。为求出用户从栅格 g_i 移动到栅格 g_j 的可能性，不妨设 g_i 和 g_j 之间间隔 $r(r>1)$ 个栅格距离，即从 g_i 移动到 g_j 至少需要 r 步，则与之对应的 r 步状态转移概率即代表用户从 g_i 经最短路径(r 步)移动到 g_j 的可能性，其大小与矩阵 M^1 自乘 r 次求

得的M^r中的元素p_{ij}^r相等。但在实际中，由于种种原因，用户从一个地方到另一个地方所选择的路径通常不总是最短的，即用户从g_i移动到g_j的概率并不直接等于p_{ij}^r的大小，而应该是用户经二者之间所有可能路径到达的概率总和，即总转移概率——p_{ij}^T，其计算公式为

$$p_{ij}^T=\sum_{r=ds}^{dl}M_{ij}^r=M_{ij}^{ds}\sum_{r=0}^{ds-dl}M_{ij}^r=M_{ij}^{ds}(M_{ij}^0+M_{ij}^1+\cdots+M_{ij}^{ds-dl}) \tag{11}$$

其中，ds为g_i和g_j之间最短路径的长度，dl为g_i和g_j之间用户可能选择的最长路径的长度，M_{ij}^0是单位矩阵。

利用式(10)和式(11)可求出G_t中每对栅格之间的总转移概率，从而构建总转移概率矩阵M^T。至此，TraDR离线建模阶段完成。

4 TraDR在线预测算法

TraDR的在线预测阶段基于离线建模阶段求得的签到位置栅格的集合G_t、总转移概率矩阵M^T、用户的实时状态(包括查询轨迹tra_k^q、最大移动时间Δt_m)及路网信息$G(V,E,S)$，对其目的地进行预测，具体过程如算法1所示。

算法1 TraDR_Prediction(G_t,M^T,tra_k^q,Δt_m,G)

输入：G_t,M^T,tra_k^q,Δt_m,$G(V,E,S)$

输出：des_list中前κ个元素

1. des_list←∅;
2. $p(tra_k^q)$←M^T;
3. for each g_i in G_t do
4. $p(\Delta t_m|l_d\in g_i,tra_k^q)$←$G(V,E,S)$,$\Delta t_m$
5. if $p(\Delta t_m|l_d\in g_i,tra_k^q)==1$
6. retrieve p_{ci}^t and p_{si}^t from M^T;
7. $p(tra_q^k|l_d\in g_i)$←$p(tra_k^q)$,p_{ci}^t,p_{si}^t;
8. $p(l_d\in g_i|\Delta t_m,tra_k^q)=1\cdot p(l_d\in g_i|tra_k^q)$
9. ←$p(tra_k^q|l_d\in g_i)$,$p(l_d\in g_i)$;
10. save $p(l_d\in g_i|\Delta t_m,tra_k^q)$ into des_list;
11. end if
12. sort des_list;
13. return the top-κ elements of des_list;

首先，依据式(7)，利用转移概率矩阵M^T计算用户当前轨迹tra_k^q对应的轨迹概率$p(tra_k^q)$(行2)；然后，利用路网信息$G(V,E,S)$及目标用户的最大移动时间Δt_m计算集合G_t中的每个栅格区域g_i的可达性概率$p(\Delta t_m|l_d\in g_i,tra_k^q)$(行4)：若$p(\Delta t_m|l_d\in g_i,tra_k^q)==1$，则表明在当前情况下$g_i$可能是用户的目的地，则利用式(6)进一步计算后验概率(行6、行7)，进而利用式(4)和式(5)计算$p(l_d\in g_i|tra_k^q)$，由于该情况下可达性概率为1(行2)，因而求得的$p(l_d\in g_i|tra_k^q)$等于$p(l_d\in g_i|\Delta t_m,tra_k^q)$，即为$g_i$对应的目的地概率(行10、行11)，将其存入*des_list*中；否则$p(\Delta t_m|l_d\in g_i,tra_k^q)==0$，意味着用户无法在当前状态下到达$g_i$，$g_i$对应的目的地概率必为0，因而无需进行额外的计算(行11)；最终对所有求得的目的地概率进行排序(行12)，返回最大的κ个，即为目标用户可能到达的目的地(行13)，至此，位置预测结束。

5 实验及分析

本文以两个典型的位置预测算法为基准模型，利用真实数据集对TraDR的有效性及效率进行评估。

5.1 实验设置

5.1.1 数据集

本实验采用来自移动社交网络Brightkite的真实签到数据集[9]及加利福尼亚州路网数据集[14]。

Brightkite数据集包含从2008年4月至2010年10月的用户签到记录及好友关系数据。本实验首先对其进行预处理，仅保留签到位置在加利福尼亚的9427个用户及541037条签到记录，并从中随机选取1000个用户作为目标用户，每个用户一天当中的所有签到记录组成的序列被看作是一条轨迹。然后，将所有轨迹数据以3个月为单位大致分为6组，其中每个组的前2.5个月的数据作为训练数据，用于离线建模；后半个月的数据作为测试数据，用于在线预测。针对每个目标用户从每组测试数据中随机选取5条签到记录，每条签到记录c^k对应的签到地点作为用户的当前位置g_c，紧邻该签到记录的下一个签到地点作为预测算法将要预测的用户目的地，与c^k同一天发布但发布时间早于c^k的该用户的所有签到位置组成的轨迹作为查询轨迹。此外，进一步的统计分析发现，用户的连续签到时间间隔的众数在3h内，因而本实验设置用户最大转移时间Δt_m为3h。

5.1.2 基准模型

(1)ZMDB模型

ZMDB模型[15]为典型的基于贝叶斯推理框架进行位置预测的算法，其通过轨迹匹配进行后验概率计算的思想为众多预测算法所采用，因而本实验将其作为基准模型对TraDR进行评估。

ZMDB模型中目的地概率的计算如式(12)所示：

$$p^{des}(l_d\in g_i|\Delta t_m,tra_k^q)=\frac{p(tra_k^q|l_d\in g_i)p(l_d\in g_i)}{\sum_j p(tra_k^q|l_d\in g_j)p(l_d\in g_j)} \tag{12}$$

其先验概率定义如式(5)所示，后验概率定义如式(13)所示，其中$tra_k^q\subset tra_{l_d\in g_i}$表示查询轨迹$tra_k^q$与样本集中终点为$g_i$的轨迹$tra_{l_d\in g_i}$部分匹配。

$$p(tra_k^q|l_d\in g_i)=\frac{|\{tra_{l_d\in g_i}|tra_k^q\subset tra_{l_d\in g_i}\}|}{|\{tra_{l_d\in g_i}\}|} \tag{13}$$

(2)SubSyn模型

SubSyn模型[7]为首次关注“数据稀疏问题”并提出解决方案的模型，但它单纯利用所有用户的全部轨迹数据建立唯一的推理模型来预测所有用户的位置。因其未考虑实际中地理因素的影响，所以其目的地概率的计算与ZMDB模型一致，如式(12)所示；先验概率的计算如式(5)所示。

5.2 有效性评估

5.2.1 评估指标

本实验引入召回率(Recall Rate,RR)和预测误差[7](Prediction Error,PE)两个指标来衡量预测方法的有效性。召回率是指能够找到至少一个目的地位置的查询轨迹数量占所有查询轨迹数量的比率，该指标主要是衡量位置预测算法应对“轨迹数据稀疏问题”的能力，召回率越高，算法鲁棒性越强，反之亦然；预测误差用于衡量预测得到的目的地与用户真正目的地之间的距离误差。对于单个查询轨迹来说，其距离误差定义为得到的κ个潜在目的地$g_i(i\in[1,\kappa])$与真实目的地g_d之间的距离的加权平均值，如式(14)所示：

$$PE=\frac{\sum_{i=1}^{k}(p_{g_i}^{des}\cdot\|g_i-g_d\|)}{\sum_{i=1}^{k}p_{g_i}^{des}} \quad (14)$$

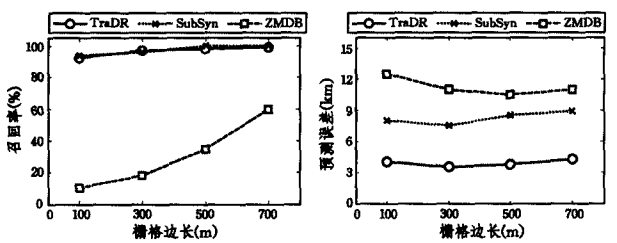
其中，$p_{g_i}^{des}$ 为 g_i 的目的地概率。因而，预测误差定义为一次实验中每条查询轨迹对应的距离误差的算术平均值。预测误差越小，算法性能越好，反之亦然。

5.2.2 实验评估

为充分测试预测算法的有效性，本文分别改变位置栅格边长 λ 及查询轨迹 tra_k^q 的长度这两个关键变量的大小，并统计分析算法召回率及预测误差的变化情况。

实验 1 位置栅格的大小对算法有效性的影响

位置栅格的大小对预测算法的召回率和预测误差都有重要的影响。具体来讲，位置栅格越大，意味着同一个位置栅格内包含的 POI 数目越多，则包含不同 POI 的轨迹间的区分度越小，因此查询轨迹与样本轨迹匹配的概率越大，即召回率越高，但与此同时预测误差也会越高；反之亦然。为最大限度地提高预测算法的有效性(即高召回率、低预测误差)，研究位置栅格的大小对预测算法有效性的影响十分必要。因此，本实验设计 4 组子实验寻找最佳位置栅格的大小，每组子实验的位置栅格边长 λ 从 100m 到 700m 以 200m 为单位递增，其他参数的设置与 5.1 节一致，分别统计各模型在不同子实验中的召回率和预测误差，实验结果如图 3 所示。



(a)召回率随栅格边长的变化　　(b)预测误差随栅格边长的变化

图 3 预测算法的有效性随栅格边长的变化情况

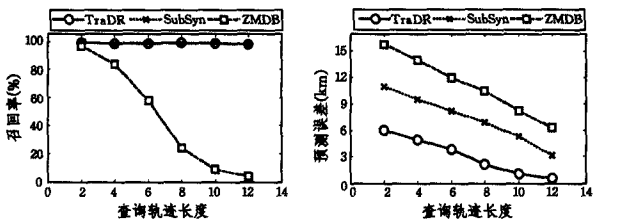
由图 3(a)可知，预测算法的召回率随栅格边长的增大而呈现出整体增大的趋势，其中 TraDR 与 SubSyn 无论栅格边长大小，其召回率均在 90%以上，而 ZMDB 在栅格边长为 100m 时召回率仅为 10%左右，即使栅格边长增大为 700m 时，其召回率也仅为 60%。如前文所述，栅格越小，意味着轨迹区分度越大，轨迹数据稀疏问题也就越明显，因此，由实验结果可知，TraDR 和 SubSyn 均能较好地应对“轨迹数据稀疏”问题。此外可以注意到，当栅格较小时，SubSyn 的召回率略高于 TraDR，这主要是由于 SubSyn 在建模过程中考虑了轨迹数据全集中的所有样本，因而轨迹覆盖范围相对更广；而 TraDR 利用的是筛选出的用户相关的个性化轨迹数据集，由于用户移动行为的不确定性，可能偶尔会访问一些“奇异”的新位置，因此在栅格粒度很小时 TraDR 单在召回率方面略微逊于 SubSyn。

由图 3(b)可知，TraDR 和 SubSyn 的预测误差大致呈现出随 λ 的增大而增大的趋势，其中 TraDR 较 SubSyn 表现出了明显的优势，并在 λ=300m 时取得最小值。而 ZMDB 由于“轨迹数据稀疏问题”影响，预测误差随 λ 的增大而减小，但这种变化仅仅是由于实验中人为增大栅格粒度而引起的，并未从本质上提高算法的预测精度，因此 ZMDB 的性能仍远低于其他模型。

综上，当栅格大小变化时，TraDR 无论在召回率还是在预测精度方面都表现出了较明显的优势，且当 λ=300m 时性能最优。

实验 2 查询轨迹长度对算法有效性的影响

查询轨迹的长度不同，预测算法的结果可能不尽相同。本实验首先对所有查询轨迹的长度进行统计，并结合 λ=300m 时的预测结果，得到召回率和预测误差分别随查询轨迹长度的变化情况，如图 4 所示。



(a)召回率变化情况　　(b)预测误差变化情况

图 4 预测算法的有效性随查询轨迹长度的变化情况

由图 4(a)可知，TraDR 和 SubSyn 的召回率基本不受查询轨迹长度的影响，一直维持在 100%左右；而 ZMDB 却随着查询轨迹长度的增大而骤降，究其原因，还是“轨迹稀疏问题”的影响：由于查询轨迹越长，在样本集中找到与查询轨迹相匹配的轨迹的概率越小，因此当查询轨迹较长，比如等于 12 时，ZMDB 召回率几乎为 0。

由图 4(b)可知，预测算法的预测误差均随着查询轨迹长度的增加而减小，其中 TraDR 表现最佳，SubSyn 次之，ZMDB 最差。TraDR 与 SubSyn 的变化趋势可以解释为随着查询轨迹长度的增大，其中蕴含的用户移动特征越明显，因此算法能够更好地利用用户历史移动特征对其当前运行行为进行预测。由于 TraDR 是对目标用户进行个性化建模及预测，相对于 SubSyn 利用唯一的模型对所有用户进行预测的做法而言更有针对性，因此预测效果更佳。此外，由于在算法失去预测能力时，本实验将用户的当前位置作为目的地进行统计分析，因此 ZMDB 在查询轨迹长度较大，召回率几乎为零(即基本失去预测能力)时，仍有较低的预测误差。这可以解释为：用户查询轨迹长度大，通常是由于用户的签到频率较高而导致的，即当前签到位置点与下一个签到位置点的距离相对较短，而这恰好使得 ZMDB 预测误差也随查询轨迹长度的增大而降低。

综上，当查询轨迹长度变化时，TraDR 无论在召回率还是在预测精度方面都表现出了较明显的优势。

5.3 效率评估

由于移动社交网络中应用场景的限制，推荐算法通常需要实时运行并能对用户查询作出迅速响应，因此算法的运行效率作为一项重要评估指标，直接影响用户体验及算法的可用性。本文分别统计分析各预测算法的运行时间，据此评估 TraDR 的运算效率。实验硬件配置为：Intel Xeon CPU E5-2650，64GB RAM。

通过对 TraDR 在线预测算法运行时间的统计及与 SubSyn 与 ZMDB 的对比发现：TraDR 对用户目的地单次预测的平均响应时间明显短于其他两个算法。以栅格边长为 300m 的统计结果为例，TraDR 的平均运行时间约为 10^2 ms，SubSyn 约为 10^{-1} ms，而 ZMDB 表现最差，约为 10^2 ms。针对上

述结果，结合各算法的原理及流程分析可知，ZMDB由于每次预测都需要进行大量的轨迹匹配运算，这是十分耗时的，因此运行效率低下；SubSyn为了选出访问概率最大的目的地，需要对研究区域内的所有样本位置的目的地概率进行计算，其中包含那些因时空限制完全不可达的位置点，因而其效率也逊于TraDR；而TraDR通过离线建模阶段计算得出了在线阶段所有可能用到的位置样本间的转移概率，在线预测时仅需要简单的查询即可获得，此外，在线预测过程中，在计算目的地概率之前首先利用路网信息及时空限制对不可达目的地进行过滤，大大减少了不必要的计算开支，因而表现出良好的运算效率。

虽然预测算法的在线预测效率是影响用户体验的最主要因素，但由于TraDR包含离线建模和在线预测两个阶段，且离线建模的时间耗费对算法可用性也有重要影响，为保证评估的完备性，本文统计了离线建模阶段的平均时间耗费，结果发现当栅格大小取300m时，平均时间耗费约为2h。考虑到一般情况下，用户的总体移动特征在相对较长的时间内是不会发生大的改变的，离线建模阶段结果无需实时更新，一次计算可满足几周甚至几个月的需求，因而TraDR的离线建模效率是完全可接受的。

结束语 本文提出了一种基于轨迹“分解-重构”的移动社交网络位置预测方法TraDR。该方法合理利用公开易得的先验知识，通过对轨迹的“分解-重构”为目标用户建立个性化位置推理模型，增强了位置预测的针对性，提高了预测结果的准确率，并有效应对了“轨迹数据稀疏问题”。道路网络信息的引入实现了对不可达目的地的初步过滤，大大减少了无意义的计算开支，提高了算法效率。基于真实数据集的实验验证了TraDR在预测能力及预测效率方面的优势。

TraDR方法目前主要是利用移动社交网络中签到服务产生的结构化位置数据对用户位置进行预测，实际上GeoSN中包含用户位置信息的数据并不局限于此：一方面，用户在分享资源时可能会添加位置标签，在近邻发现时会通过提交其位置数据查询附近的好友、POI等，这两类服务中都包含了明确的位置数据；另一方面，GeoSN用户发布的文字内容中可能包含具有空间特征的语言描述，此类描述虽然不包含结构化的位置数据（如经纬度），但从中同样能够有效提取出用户所在的地理位置[16]。若将这些位置相关的数据与用户的签到位置数据混杂，并按时间排序，便可构成与用户真实的移动路线拟合度更高的轨迹，因而，后续研究中将广泛收集此类位置信息，并将其作为签到位置数据的有效补充，通过增强TraDR对位置相关信息的识别能力、丰富算法的先验知识来扩展算法，提高预测的准确性。

参考文献

[1] iResearch Consulting Group. China Mobile social application market research reports[R]. Beijing, 2014(in Chinese)
艾瑞咨询公司. 中国移动社交应用市场研究报告[R]. 北京, 2014

[2] Zhai Hong-sheng, YU Hai-peng. Present situation and trend of research of location-based service on online social networks[J]. Application Research of Computers, 2013, 30(11): 3221-3227(in Chinese)
翟红生，于海鹏. 在线社交网络中的位置服务研究进展与趋势[J]. 计算机应用研究, 2013, 30(11): 3221-3227

[3] Krumm J, Horvitz E. Predestination: Where do you want to go today? [J]. Computer, 2007, 40(4): 105-107

[4] Krumm J, Horvitz E. Predestination: Inferring destinations from partial trajectories[M]// UbiComp 2006: Ubiquitous Computing. Springer Berlin Heidelberg, 2006: 243-260

[5] Wei Ling-yin, Zheng Yu, Peng W C. Constructing popular routes from uncertain trajectories[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012: 195-203

[6] Horvitz E, Krumm J. Some help on the way: Opportunistic routing under uncertainty[C]// Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, 2012: 371-380

[7] Xue A Y, Zhang Rui, Zheng Yu, et al. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction[C]// Proceedings of the 29th International Conference on Data Engineering (ICDE). IEEE, 2013: 254-265

[8] Ye Mao, Yin Pei-feng, Lee Wang-chien, et al. Exploiting geographical influence for collaborative point-of-interest recommendation[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011: 325-334

[9] Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in location-based social networks[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011: 1082-1090

[10] Bell R, Koren Y, Volinsky C. Modeling relationships at multiple scales to improve accuracy of large recommender systems[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007: 95-104

[11] Pentland A S. Honest signals[M]. MIT press, 2010

[12] Lian De-fu. Data Mining on Location based Social Networks [D]. Hefei: University of Science and Technology of China, 2014 (in Chinese)
连德富. 基于位置社交网络的数据挖掘[D]. 合肥：中国科学技术大学, 2014

[13] Sadilek A, Kautz H, Bigham J P. Finding your friends and following them to where you are[C]// Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. ACM, 2012: 723-732

[14] Li Fei-fei, Cheng Di-han, Hadjieleftheriou M, et al. On trip planning queries in spatial databases[M]// Advances in Spatial and Temporal Databases. Springer Berlin Heidelberg, 2005: 273-290

[15] Ziebart B D, Maas A L, Dey A K, et al. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior [C]// Proceedings of the 10th International Conference on Ubiquitous Computing. ACM, 2008: 322-331

[16] Cheng Zhi-yuan, Caverlee J, Lee K. You are where you tweet: a content-based approach to geo-locating twitter users[C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM). 2010: 759-768