·复杂性科学·

签到行为的可预测性及影响因素分析

卢 扬[1]，樊 超[1,2]，韩筱璞[3]，荣智海[1]

(1. 电子科技大学互联网科学中心CompleX实验室 成都 611731; 2. 山西农业大学文理学院 山西 太谷 030801;

3. 杭州师范大学信息经济研究所和阿里巴巴商学院 杭州 310036)

【摘要】人类日常的出行行为受到多种因素的制约和影响。本文通过两组手机用户上报的位置信息分析人类签到行为的空间规律特征，着重分析了访问地点数、平均跳转距离、回转半径和最常访问地点等因素对签到行为的可预测性的影响。研究表明签到行为具有一定的记忆性，用户访问的地点数、对最常访问地点的访问规律是影响可预测性和规律性的主要因素，用户的活动范围和平均跳转距离对可预测性的影响则微弱得多。

关 键 词 签到行为; 熵; 人类动力学; 可预测性; 空间运动规律

中图分类号 N94 **文献标志码** A doi:10.3969/j.issn.1001-0548.2015.02.001

Predictability and Influential Factors on Check-in Behaviors

LU Yang[1], FAN Chao[1,2], HAN Xiao-pu[3], and RONG Zhi-hai[1]

(1. CompleX Lab, Web Sciences Center, University of Electronic Science and Technology of China Chengdu 611731;

2. College of Arts and Sciences, Shanxi Agricultural University Taigu Shanxi 030801;

3. Institute of Information Economy and Alibaba Business College, Hangzhou Normal University Hangzhou 310036)

Abstract The human mobility pattern in ordinary life is influenced by various factors. Two datasets of location information reported by mobile phones are utilized to analyze the spatial mobility pattern of check-in behavior. Our research focuses on the impacts of the numbers of visited locations, average jump distances, radiuses of gyration and most frequent visited locations on the predictability of check-in behavior. It is found that the check-in behavior shows certain memory effect. The numbers of visited locations and the visiting patterns to the most frequent visited locations have more significant influence on the predictability and regularity of check-in behavior, meanwhile the impacts of radiuses of gyration and the average jump distances are obviously unremarkable.

Key words check-in behavior; entropy; human dynamics; predictability; spatial mobility pattern

对人类行为规律的探索长久以来一直是自然、经济、社会等各个学科领域的学者关注的研究方向。近年来，随着越来越多的人类行为的数据资料被精确记录，学者得以从定量角度分析人类行为的时空规律及其动力学机制，并由此改变了很多对人类行为的传统认识。如在时间规律上，过去人们假设人类行为的产生是具有均匀特性的泊松过程，而近年来大量实证结果显示人类行为在很多方面具有明显的阵发和重尾特征[1-5]，即表现为长时间静默和短时间爆发交织，且时间间隔服从重尾分布。

研究人类行为的空间规律在疾病传播[6-8]、交通流控制[9-11]、异常行为监测[12]、人口迁移[13]等方面具有重大的理论和应用价值。过去，人们假设人类的出行行为可以用随机游走或者列维飞行刻画，但近年来的一系列研究却证实人类出行的时间间隔分布和位移距离分布都服从重尾分布，表现为阵发性、有界性、周期性和规律性综合的特征[14-20]。为此，学者相继从不同角度提出了统计模型来解释上述特征产生的原因[14,16-17]。在实证和建模的基础上，更具有理论和商业价值的位置预测[21-29]也是人类出行行为研究的重点之一。文献[21]用熵的方法得到人类出行的理论可预测性最高可达93%，该结果受到了广泛关注。

过去对人类出行规律的研究所采用的数据多来源于钞票或者手机通信，这些数据都可视为被动签到行为的结果，并非用户主动上传。随着GPS设备

收稿日期：2014 – 08 – 21；修回日期：2015 – 01 – 26
基金项目：国家自然科学基金(61473060，11205040)；CCF-腾讯犀牛鸟科研基金.
作者简介：卢扬(1991 –)，女，硕士生，主要从事人类动力学方面的研究.

的微型化，更能反映用户的主观愿望的即时通讯(instant messaging, IM)和基于位置的服务(location based services, LBS)工具变得更加普及，从而为研究人们的出行行为提供了更好的媒介。

本文通过两组由手机收集的地点签到数据(包括基于IM的QQ和基于LBS的Gowalla)研究人们在日常生活中的签到行为，总结了签到行为的基本特征，利用熵和Fano不等式计算了用户的平均最大可预测性，重点分析了影响可预测性的因素，包括访问地点数、平均跳转距离、回转半径和最常访问地点。发现人们的签到行为具有明显的非均匀特征和一定的记忆效应，可预测性和规律性受用户访问的地点数的影响明显，而与用户的活动范围和平均跳转距离关系不大，更进一步，可预测性会随着用户最常访问地点的删除而呈现先减小后增大的趋势。同时还发现，与被动签到行为相比，主动签到行为具有更大的熵值，因而也更难预测。相比于地点分享行为，日常出行行为的记忆性、规律性和可预测性都更强一些。

1 数据描述

本文研究所采用的数据集来源于两组由手机收集到的经过匿名化处理的地点签到信息：数据集*D*1来自LBS社交网站Gowalla，全球范围内的用户可通过移动端的应用程序或者浏览器进行主动签到，从而与好友分享新的地点、活动和旅行线路；数据集*D*2来自国内某沿海城市的手机QQ用户使用涉及地图服务的应用时被动记录下的地点信息。因此，两组数据都是用户发生空间移动行为时记录的位置信息，包括了用户ID、地点经纬度、时间等属性，且相比于*D*2，*D*1由于是用户主动上传分享的，故其主动性更强一些。为了保证用户轨迹信息量具有统计意义，本文在计算可预测性时去掉了地点签到量不足100条的用户，在去掉不活跃的用户之后，*D*1、*D*2的用户数量分别为全部用户的8.35%和28.92%，但轨迹量却能分别达到65.59%和79.33%，地点数目分别达到全量数据的76.56%和81.00%。两组数据的概述如表1所示。

表1 数据集简介

数据集		时间跨度	用户数量	签到数量	地点数量
Gowalla	全部用户	2009.02～2010.10	196 591	6 442 892	1 256 693
	活跃用户		16 412	4 225 868	962 142
QQ	全部用户	2013.07～2013.12	352 018	37 752 170	889 800
	活跃用户		101 797	29 948 589	720 762

这两组数据都源于手机用户上网、签到或查询地图等行为，文中将用户在某个地点产生一条轨迹信息记录的行为统称为“签到”，若相邻两次签到的地点发生变化，则称为一次“跳转”，若地点没有发生变化，则称之为“停留”。由于数据集中存在短时间内产生多条签到记录的现象，使得数据在时间上会显得非常频繁，但在空间地点信息上又显得不够丰富。为了更好地分析用户的空间移动行为特征，将极短时间内在同一地点的多条签到记录合并为一条，最后保留的数据集中仍然存在一定时间间隔下的有意义的地点停留。从而获得每个用户 i 的签到轨迹集合 $D_i=\{d_1^i,d_2^i,\cdots,d_j^i,\cdots,d_n^i\}$，其中 d_j^i 代表用户 i 访问的第 j 个地点。同时定义用户 i 的跳转距离集合为 $L_i=\{l_1^i,l_2^i,\cdots,l_j^i,\cdots,l_{n-1}^i\}$，其中 l_j^i 代表用户 i 在签到地点 d_j^i 和 d_{j+1}^i 之间的跳转距离，可以根据签到地点的经纬度信息计算获得。

2 签到行为的基本特征

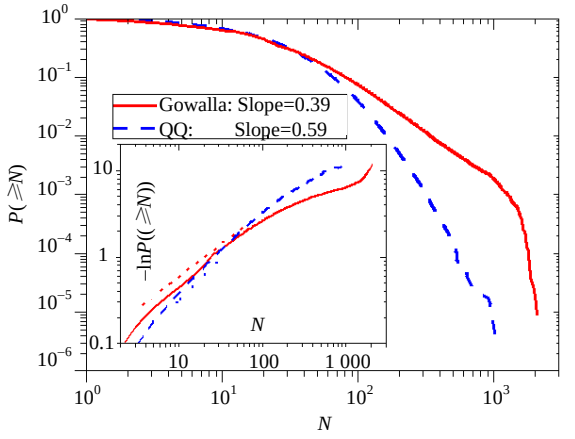
2.1 用户和地点的活跃度分布

统计结果显示，本文所研究的签到行为的时间间隔分布和跳转距离分布都表现出幂律特征，与文献[13-15]的结果类似。那么，在人们的日常生活中，每个人会访问多少个不同的地点？每个地点又会有多少不同的人来访问呢？为了回答这两个问题，定义用户的活跃度为用户去过的地点集的大小 N，定义地点的活跃度为去过该地点的用户集的大小 U。统计两个数据集中全部用户和地点的活跃度分布，结果如图1所示。

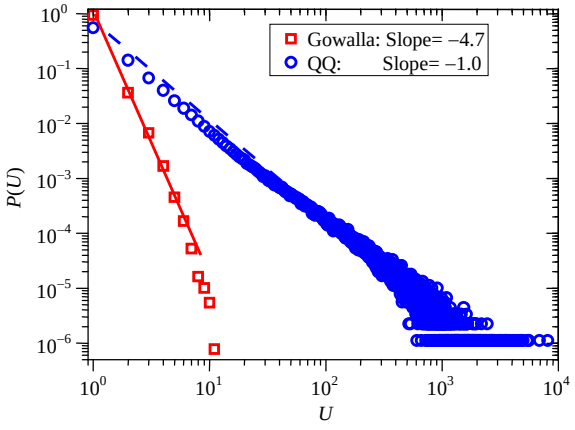
用户活跃度 N 的累积分布如图1a所示，两个数据集中用户比例均在大约30个地点处开始明显下降，这说明在人们的日常生活中，大多数人经常访问的地点数是有限的，对这些有限数量地点的访问是较为均匀的。如图1a插图所示，曲线在双对数坐标下近似为直线，即 $-\ln(P(\geqslant N))\sim N$，故两个数据集中用户的活跃度分布的累积形式表现为广延指数分布形式[30-31]：$P_c(x)=\exp[-(x/x_0)^c]$，其中 x_0 为特征标度，指数 c 即为图1a插图中近似直线的斜率。

而由图1b知地点的活跃度分布则为幂律分布。这说明在特定地点签到的人数具有较强的异质性，即日常生活中大部分地点的访问人数较少，同时存在少数热门地点具有大量的访问人数。这样的现象与购物、点评等典型二部图网络的度分布研究结果类似[32]，说明在真实系统中，行为的主动发出者所覆盖的受众是有限而较为均匀的，而行为的被动接

收者却可以接受大量而异质的访问。由于Gowalla数据的地点精确度非常高，故大部分地点的访问量非常少，因而其U曲线的衰减速度比QQ的U曲线要快得多，后者的异质性更强。



a. 用户的活跃度分布



b. 地点的活跃度分布

图1　用户和地点的活跃度分布

2.2　签到行为的统计特征

根据签到记录中的经纬度信息，计算用户 i 在签到过程中的跳转距离，并进一步得到每个用户的平均跳转距离为：

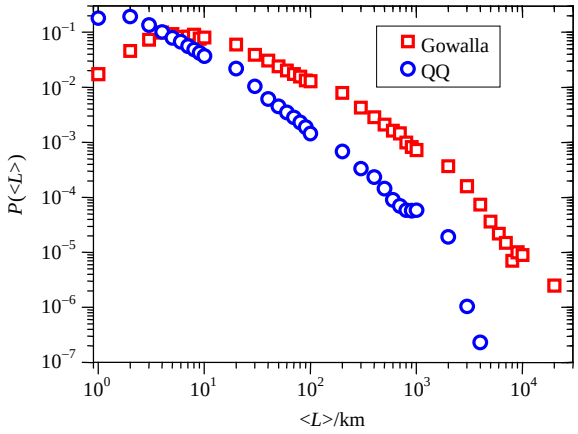
$$\bar{l}^i = \frac{1}{|L^i|}\sum_{k=1}^{|L^i|} l_k^i$$

式中，$|L^i|$ 代表集合 L^i 的大小，该值反映了个体用户在签到过程中发生跳转时的平均距离。绘制全部用户的平均跳转距离分布，如图2a所示。从图中可以看到该分布的主体部分具有明显的重尾特征，大部分用户的平均跳转距离在几十公里范围内，极少部分用户的平均跳转距离能达到数千公里以上。

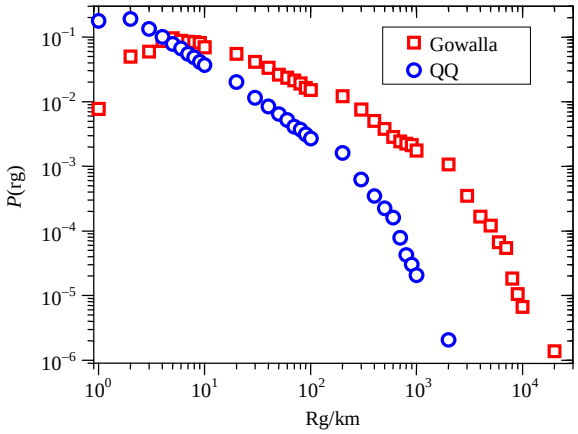
为了考察用户日常活动范围的大小，定义回转半径[15]为：

$$\mathrm{Rg}^i = \sqrt{\frac{1}{|L^i|}\sum_{k=1}^{|L^i|} |l_k^i - \bar{m}^i|^2}$$

式中，$\bar{m}^i$ 表示该用户全部轨迹点的质心。计算每个用户的回转半径，其概率分布如图2b所示，该分布同样表现为幂律形式，说明大多数人在日常生活中的活动半径是有限的，只有少数人的活动半径能达到数百、甚至数千公里。进一步计算每个用户的平均跳转距离和回转半径之间的Pearson相关系数，结果在$D1$和$D2$中分别为0.630和0.556，即二者表现为较强的正相关关系。



a. 平均跳转距离分布



b. 回转半径分布

图2　平均跳转距离和回转半径分布

2.3　跳转距离相关性

用户相邻的两次跳转之间是否存在内在联系，是否一次长距离的跳转也预示着下一步也是长距离的跳转？为了研究这个问题，本文采用文献[33]中定义的记忆性指标，研究所有个体用户跳转距离的相关性。

若某用户 i 的跳转距离序列共有 $n_\tau^i(=|L^i|)$ 个元素(即有 n_τ^i+1 次签到)，则将原序列分为2个子序列，分别包含前 n_τ^i-1 个元素和后 n_τ^i-1 个元素。用 M' 表示用户 i 的记忆性指标，则该用户的记忆性可以用上述两个子序列的Pearson相关系数衡量，有：

$$M' = \frac{1}{n_\tau^i - 1}\sum_{k=1}^{n_\tau^i-1}\frac{(\tau_k - m_1)(\tau_{k+1} - m_2)}{\sigma_1\sigma_2}$$

式中，m_1 和 m_2、σ_1 和 σ_2 分别是两个子序列的均值和标准差。该值在–1～1之间，$M'>0$ 意味着记忆效应，$M'<0$ 意味着反记忆性。

本文计算每一个用户的跳转距离序列的 M' 值。结果显示，所有用户 M' 值的平均值 $\langle M'\rangle$ 在Gowalla和QQ中分别为0.134±0.163和0.249±0.186。从该结果可以看出，对于大多数用户来说，长距离的跳转之后仍然倾向于长距离的跳转，反之亦然，即跳转距离具有一定的弱记忆性和正相关性。考虑在日常生活中，人们大部分的出行是在以家和公司为焦点的椭圆范围之内活动[34]，连续出行距离都比较短；但一旦有出差、旅行或探亲活动，则很容易伴随一系列的长短距离交替的跳转活动。相比于数据集 $D1$，$D2$更多是日常生活中城内和城际范围内的活动，因而后者的签到行为更集中，$\langle M'\rangle$ 更大，即日常生活中签到行为的记忆性更强。

3 签到行为可预测性分析

3.1 签到行为的可预测性度量

本文采用文献[21]中的方法定义签到行为的熵和可预测性，包括三种熵的度量指标。

随机熵：$S^i_{\text{rand}}=\log_2 N^i$，其中 N^i 表示用户 i 去过的地点集的大小。该指标只考虑用户访问过的唯一地点数，默认用户以相同概率访问这些地点。

香农熵：$S^i_{\text{unc}}=-\sum_{d\in D^i} p^i(d)\log_2 p'(d)$，该指标进一步考虑访问过某一历史地点 d 的概率 $p'(d)$，其中 $p'(d)=n^i(d)/n^i$，$n^i(d)$ 表示用户 i 在地点 d 的签到次数，n^i 表示用户 i 的总签到次数。在这种情况下，用户选择下一个地点的概率和历史概率分布一致。

真实熵：$S^i_{\text{real}}=-\sum_{D^{i'}\in D^i} p(D^{i'})\log_2[p(D^{i'})]$，其中 $D^{i'}$ 是 D^i 的子序列，$p(D^{i'})$ 表示 $D^{i'}$ 出现在轨迹序列 D^i 中的概率。该指标不仅考虑用户访问地点的不同概率，同时也考虑地点访问的顺序。

根据Fano不等式可得到每个用户的可预测性：

$$\Pi_i \leqslant \Pi^i_{\max}(S^i,N^i)$$

式中，$S^i=H(\Pi^i_{\max})+(1-\Pi^i_{\max})\log_2(N^i-1)$，$H(\Pi^i_{\max})=-\Pi^i_{\max}\log_2(\Pi^i_{\max})-(1-\Pi^i_{\max})\log_2(1-\Pi^i_{\max})$，$\Pi^i_{\max}$ 刻画的是用户 i 的可预测性上限值。由于该不等式很难求出解析解，故可以采用不断逼近的方法找到 $\Pi^i_{\max}$ 的近似最优解。

同时定义用户地点访问的规律性。将一周的时间分成24小时*7天=168个时段，用 R^i_{real} 表示在真实情况下每个时段的最常访问地点找到该用户的概率，其中用户在某时段的最常访问地点为用户的历史签到轨迹中在该时段签到次数最多的地点。同时用 R^i_{rand} 表示用户随机访问任意地点的规律性，则 $R^i_{\text{rand}}=1/N^i$。规律性刻画的是可预测性的一个不严格下限。

根据上述指标计算了两个数据集中活跃用户的3种熵 S_{rand}、S_{unc}、S_{real} 和3种熵分别对应的最大可预测性值 Π^{rand}、Π^{unc}、$\Pi^{\max}$ 的分布情况，结果如表2和图3所示。

表2 签到行为的可预测性度量指标计算结果

数据集	熵指标	熵	最大可预测性	规律性
Gowalla	随机熵	6.67	0.001	0.012
	香农熵	6.10	0.205	1
	真实熵	3.53	0.622	0.289
QQ	随机熵	5.67	0.002	0.024
	香农熵	4.18	0.436	1
	真实熵	2.11	0.767	0.399

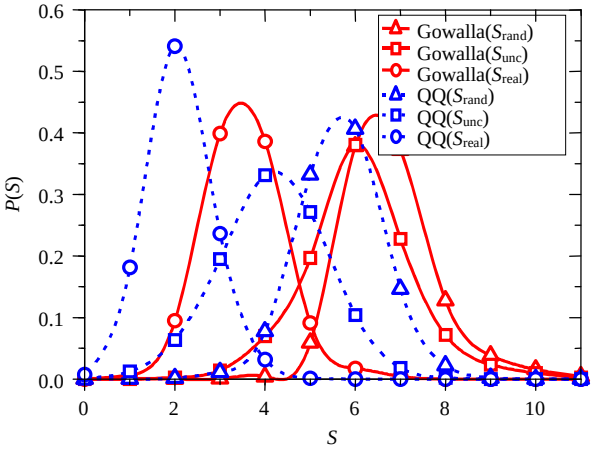
如图3a所示，对于数据集$D1$，从用户的随机熵 S_{rand}、香农熵 S_{unc} 以及真实熵 S_{real} 的分布情况可以发现，当同时考虑用户地点签到的时空特性时，熵值将大幅度降低。用户的 S_{rand} 和 S_{unc} 都分布较广，均值 $<S_{\text{rand}}>\approx 6.67$，说明用户平均每次都从 $2^{6.67}\approx 102$ 个曾经去过的地点中选择一个地点进行跳转，而均值 $<S_{\text{unc}}>\approx 6.1$，即用户在每一次跳转时有 $2^{6.1}\approx 69$ 种选择。当同时考虑地点的签到频率以及签到的顺序时，其均值 $<S_{\text{real}}>\approx 3.53$，说明用户跳转的不确定性为约 $2^{3.53}\approx 12$ 个地点。

如图3b所示，熵值的减少导致了最大可预测性的增长，$\Pi^{\max}$ 的分布也较为宽广。如果仅知道用户在多少个地点签到过，那么任何预测算法的准确度都将不会超过0.1；如果能够进一步利用用户在每个地点签到的频次信息，预测算法的平均最大可预测性将达到 $<\Pi^{\text{unc}}>\approx 0.205$；如果又能进一步了解用户具体的地点签到序列，此时最大可预测性将达到 $<\Pi^{\max}>\approx 0.622$。

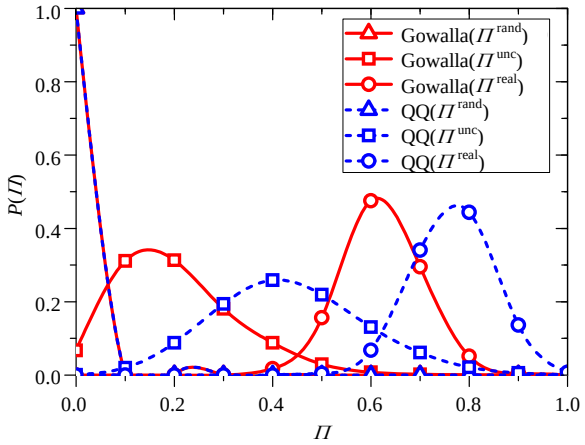
图3c揭示了用户的地点访问的规律性分布，在用户的签到行为中，约28.9%的时间里都是位于该时段最常签到的地点。故对于某个特定时段，只要猜测用户位于在其最常访问的地点，就至少能够获得28.9%左右的准确度。

对于数据集$D2$，熵、可预测性和规律性等指标表现为与$D1$类似的情况。二者的差别表现在：$D2$的3种类别的熵值都比$D1$低，可预测性则要高。这是

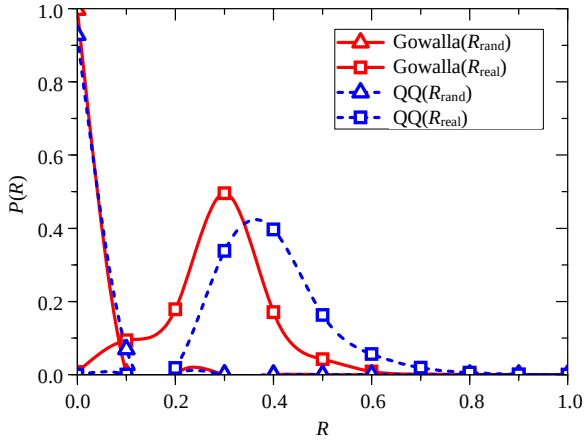
由于*D*2数据中地点的经纬度精度要低于*D*1，且*D*2的数据中地点的范围相对较小(*D*1中的签到地点遍布全世界，而*D*2大部分局限在该城市及周边)，使得*D*2中的地点重合度高达97.6%，而*D*1中只有80.5%。对于数据集*D*1，在每个用户的签到序列中新地点的比例更大，总地点个数更多，每个地点访问的概率更小，因此熵值也必然更大。



a. 熵的分布



b. 可预测性的分布



c. 规律性的分布

图3　签到行为的熵、可预测性与规律性分布

如图3a所示，*D*2的真实熵值 $<S_{real}>\approx 2.11$，即用户跳转的不确定性约为 $2^{2.11}\approx 4$，与文献[20]一致。说明在日常生活中，用户的被动签到行为在下一时刻可能访问的地点数是非常有限，而主动签到行为可能访问的地点数要大得多，即用户行为的主动性会大大提高熵值，同时降低可预测性。

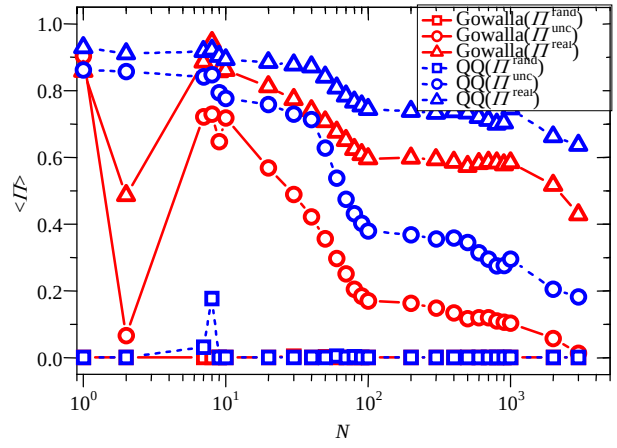
3.2　影响可预测性和规律性的因素分析

从前文的统计结果可以看出，人类的日常签到行为具有复杂性和规律性交织的特征。那么，规律性越强的用户是否更容易预测？访问过更多地点的用户、活动半径更大的用户是否更难预测？计算每个用户的可预测性 Π^i_{max} 和规律性 R^i_{real} 之间的Pearson系数，结果显示该值在*D*1和*D*2中分别为0.057和0.027，即规律性与可预测性之间并无明显的相关关系，并不是行为越规律的用户越容易预测。此外，计算可预测性 Π^i_{max} 和跳转距离记忆性 M' 之间的Pearson系数，结果在*D*1和*D*2中分别为0.111和0.096，说明可预测与跳转距离也没有显著关联。下面本文分析影响用户签到行为可预测性和规律性的因素。

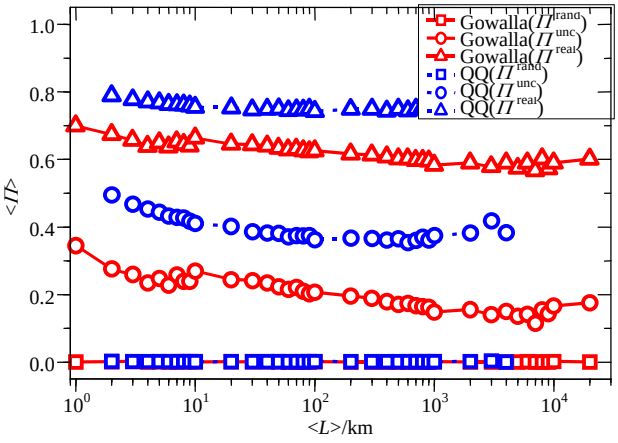
3.2.1　可预测性的影响因素分析

统计用户去过的地点数和去过该地点数的全部用户的平均可预测性值，考察二者之间的关系，结果如图4a所示。访问地点数与可预测性的关系在两个数据集中表现出了相同的规律，即先在一段小范围内减小，然后迅速变得平缓，在波动中缓慢下降。由于Gowalla数据的观测期更长，故其用户访问的地点数也更多。这说明在一定范围内，确实存在用户访问过的地点数越多，其行为就更难预测的现象。但是随着地点数持续增多其真实可预测性开始趋于平缓，即地点数的影响作用变小。总体上看，用户去过的地点数与用户的可预测性存在一定的负相关性。

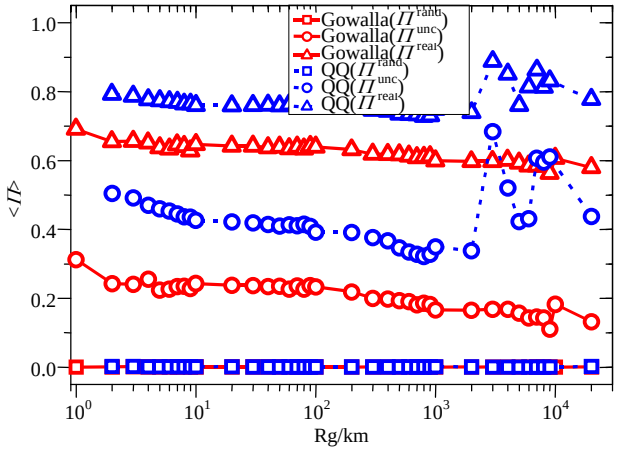
根据2.2节得到的每个用户的平均跳转距离和回转半径分析二者和可预测性的关系，如图4b和4c所示，不论是回转半径还是平均跳转距离对于可预测性的影响都表现出了相似的规律，即随着用户活动范围和出行距离的增大，$<\Pi^{max}>$ 和 $<\Pi^{unc}>$ 会在一定的范围内迅速降低，随后保持比较平稳的波动过程，而 $<\Pi^{rand}>$ 由于其计算方式导致其损失了过多的信息故数值接近于零，因此没有明显变化。相对于回转半径，平均跳转距离对可预测性的影响作用更小。



a. 访问地点数对可预测性的影响



b. 平均跳转距离对可预测性的影响

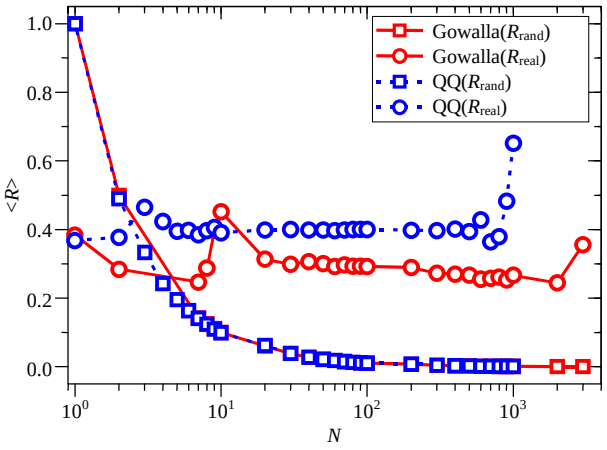


c. 回转半径对可预测性的影响

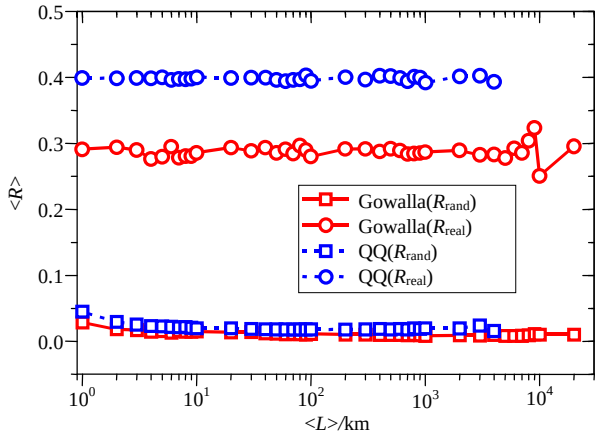
图4 可预测性的影响因素分析

3.2.2 规律性的影响因素分析

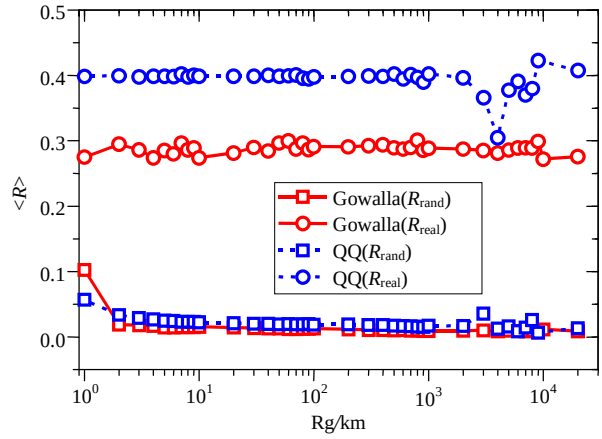
规律性反映了用户在特定时段出现在最常访问地点的概率，那么上述三个统计量对用户签到行为的规律性是否有影响呢？计算结果显示，随着用户访问地点数的增大，$<R^{rand}>$ 快速衰减并趋近于零，而 $<R^{real}>$ 在很大范围内保持缓慢的下降，说明仅仅是地点数的增大并不会对用户签到的规律性产生太大影响。而回转半径和平均跳转距离对规律性几乎没有影响。



a. 访问地点数对规律性的影响



b. 平均跳转距离对规律性的影响



c. 回转半径对规律性的影响

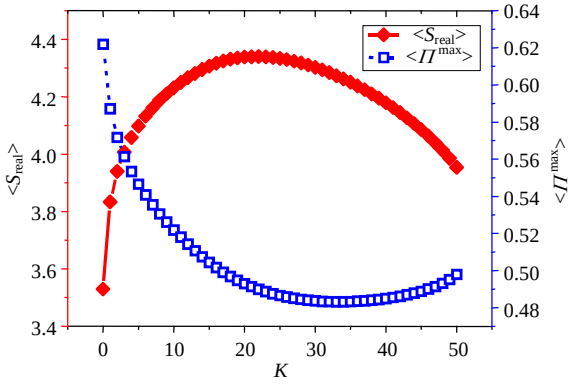
图5 规律性的影响因素分析

3.2.3 最常访问地点的影响

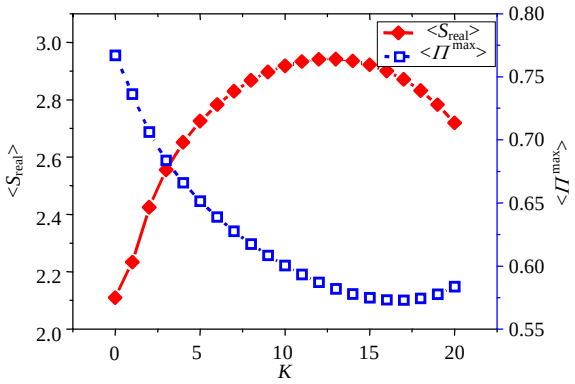
在人们的日常生活中，不论是个体还是群体用户对某个特定地点的访问量都具有显著的异质性，少数地点具有极高的访问量，而大多数地点极少被光顾。那么这些访问量大的地点是否对可预测性产生影响呢？为了回答这个问题，逐步删除用户移动

轨迹中访问量最大的K个地点，查看用户最大真实熵和可预测性的变化情况。在实验前首先挑选访问过的唯一地点数大于最大删除量(在数据集D1和D2中分别是50和20)的用户，以保证在删除访问量大的地点时用户仍然访问过多于1个不同的地点。

实验结果如图6所示，平均最大真实可预测性$<\Pi^{max}>$曲线的变化规律大致可以分为两个阶段。当删除的地点数N不超过某一阈值时，整体可预测性呈下降趋势；当N继续增长超过该阈值后，整体可预测性反而呈上升趋势。而$<S_{real}>$的变化趋势则正好相反，在小于阈值范围区间内随着K的增大而变大，在大于阈值范围内则慢慢变小。并且，在删除前面几个访问量特别大的地点时，曲线的斜率都比较大，且熵曲线变化的阈值要小于最大可预测性曲线变化的阈值。



a. Gowalla 数据



b. QQ 数据

图6　删除最常访问地点对熵和可预测性的影响

可以从以下方面理解这种非平凡现象：一般情况下，对地点访问信息丰富的用户来说，随着最常访问地点的删除，用户的地点签到序列会慢慢变得随机化，此时熵值将慢慢增大，最大可预测性也随之降低。但当轨迹点被删除到一定程度时，用户访问序列中的轨迹点都逐渐趋近于被访问极少的次数，几乎成为一个完全随机的地点访问序列，可预测性下降趋势逐渐变缓。当全部的轨迹点的访问次数都为1的时候，熵值达到最大，此时可预测性曲线也慢慢趋向最小值。当继续删除轨迹点时，熵值随着N的增大而逐渐变小，此时最大可预测性则因为随机序列中地点数的减少而缓慢增长。由此说明，用户经常访问的地点是带来签到行为高可预测性的一个重要因素。而可预测性曲线的最值点比熵曲线滞后则是Fano不等式中二者的非线性关系造成的。

4　结语和讨论

本文通过两组手机用户的签到数据研究人类日常的出行行为，总结了签到行为的一般规律，用熵的方法分析了签到行为的可预测性，并重点分析了影响可预测性的几个因素。发现人们的签到行为具有一定的记忆效应，对地点的访问具有明显的异质性。总体来看，用户访问的地点数和对最常访问地点的访问规律对可预测性和规律性有明显影响。具体而言，用户访问过的地点的数量与可预测性和规律性都具有一定的反相关关系，而回转半径和平均跳转距离对二者的影响则微弱的多。用户经常访问的地点对可预测性具有显著影响，随着这些地点被逐个删除，可预测性表现为先下降再略微上升的形态。进一步研究还发现，可预测性和规律性是人们日常生活的普遍规律，与性别、年龄等属性无关[21]，因而该性质是人类空间运动的普遍规律，在人口统计学属性上无个体差异。

研究表明，当用户访问的地点数逐步增大时，以及当用户最常访问的地点被逐步删除时，其可预测性都会下降，说明用户对地点的访问次数和访问模式对可预测性有重要影响。一方面，当用户访问的地点逐渐增多时，其访问序列会变得混乱，因而熵值增大，可预测性下降；另一方面，当用户经常访问的地点被删掉时，可预测性曲线的非线性的下降速率说明不同地点对可预测性的影响程度是不同的，访问量大的地点的影响程度也更大。这些结果都说明用户对不同地点的访问量是非均匀的。因此，用户对地点访问的异质性是影响其可预测性的重要因素。

从研究结果可以看到，数据集D2得到的可预测性数值要高于D1，这样的差别反应了两组数据集的不同。如前文介绍所说，Gowalla是一个鼓励用户主动上报地理位置的LBS网站，其行为更多源自旅游、美食、娱乐等活动的分享；而QQ数据是在用户日常生活中使用地图服务时记录的位置信息，日常生活

中出行的记忆性和规律性更强，地点重合度也更高，因而其可整体可预测性也更高。

人类行为动力学研究的是人类行为的宏观统计规律，而熵的方法分析可预测性得到的则是预测准确度的理论上限，并不是真正意义上的预测算法。由于人类行为的高度复杂性，对于个体出行行为的精确预测并不是一件容易的事情，预测的准确度也受到多种客观条件和数据本身的质量等因素制约。社会学、物理学、计算机科学等领域的学者都在从多方面关注影响人们出行的因素并探索提高预测算法的准确度的方式。本文有助于理解人类的出行规律，为寻找制约预测准确度的因素、改进利用熵和Fano不等式计算可预测性的方法提供一定的参考和借鉴。

本文的研究工作得到了山西农业大学科技创新基金(201208)的资助，在此表示感谢！

参 考 文 献

[1] BARABÁSI A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 435(2005): 207-211.

[2] ZHOU T, KIET H A T, KIM B J, et al. Role of activity in human dynamics[J]. Europhys Lett, 2008, 82(2): 28002.

[3] 周涛, 韩筱璞, 闫小勇, 等. 人类行为时空特性的统计力学[J]. 电子科技大学学报, 2013, 42(4): 481-540.
ZHOU Tao, HAN Xiao-pu, YAN Xiao-yong, et al. Statistical mechanics on temporal and spatial activities of human[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(4): 481-540.

[4] 樊超, 郭进利, 韩筱璞, 等. 人类行为动力学研究综述[J]. 复杂系统与复杂性科学, 2011, 8(2): 1-17.
FAN Chao, GUO Jin-li, HAN Xiao-pu, et al. A review of research on human dynamics[J]. Complex Systems and Complexity Science, 2011, 8(2): 1-17.

[5] ZHAO Z D, CAI S M, HUANG J, et al. Scaling behavior of online human activity[J]. Europhys Lett, 2012, 100(4): 48004.

[6] HUFNAGEL L, BROCKMANN D, GEISEL T. Forecast and control of epidemics in a globalized world[J]. Proc Natl Acad Sci, 2004(101): 15124-15129.

[7] EUBANK S, GUCLU H, KUMAR V S A, et al. Modelling disease outbreaks in realistic urban social networks[J]. Nature, 2004, 429(6988): 180-184.

[8] HAN X P, WANG B H, ZHOU C S, et al. Scaling in the global spreading patterns of pandemic Influenza A (H1N1) and the role of control: empirical statistics and modeling [EB/OL]. [2014-09-23]. http://arxiv.org/pdf/ 0912.1390.

[9] MEYER M D, MILLER E J. Urban transportation planning: a decision-oriented approach[M]. New York: McGraw-Hill, 2001.

[10] MOKHTARIAN P L, SALOMON I. In perpetual motion: Travel behavior research opportunities and application challenges[M]. Amsterdam: Elsevier Science Press, 2002.

[11] CHON Y, LANE N D, KIM Y, et al. Understanding the coverage and scalability of place-centric crowdsensing[C]// Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. [S.l]: ACM, 2013: 3-12.

[12] BARABASI A L. Bursts: the hidden patterns behind Everything we do, from your E-mail to bloody crusades[M]. New York: Plume Books, 2010.

[13] YANG Z, YUAN N J, XIE X, et al. Indigenization of Urban Mobility[EB/OL]. [2014-10-12]. http://arxiv.org/pdf/1405.7769.

[14] BROCKMANN D, HUFNAGEL L, GEISEL T. The scaling laws of human travel[J]. Nature, 2006(439): 462-465.

[15] GONZÁLEZ M C, HIDALGO C A, BARABÁSI A L. Understanding individual human mobility patterns[J]. Nature, 2008, 453(7196): 779-782.

[16] SONG C, KOREN T, WANG P, et al. Modelling the scaling properties of human mobility[J]. Nat Phys, 2010(6): 818-823.

[17] 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. [S.l]: ACM, 2011: 1082-1090.

[18] HAN Xiao-pu, HAO Qiang, WANG Bing-hong, et al. Origin of the scaling law in human mobility: Hierarchy of traffic systems[J]. Phys Rev E, 2011, 83(3): 036117.

[19] YAN X Y, HAN X P, WANG B H, et al. Diversity of individual mobility patterns and emergence of aggregated scaling laws[J]. Scientific Reports, 2013, 3: 2678.

[20] SCHNEIDER C M, BELIK V, COURONNE T, et al. Unravelling daily human mobility motifs[J]. Journal of The Royal Society Interface, 2013, 10(84): 20130246.

[21] SONG C, QU Z, BLUMM N, et al. Limits of predictability in human mobility[J]. Science, 2010, 327(5968): 1018-1021.

[22] MONREALE A, PINELLI F, TRASARTI R, et al. WhereNext: a location predictor on trajectory pattern mining[C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. [S.l]: ACM, 2009: 637-646.

[23] 朱寅, 杨强. 诺基亚移动数据挖掘竞赛[J]. 中国计算机学会通讯, 2012, 8(8): 67-70.
ZHU Yin, YANG Qiang. Nokia mobile data challenge[J]. Communications of the Chinese Computer Federation, 2012, 8(8): 67-70.

[24] GAMBS S, KILLIJIAN M O, DEL PRADO CORTEZ M N. Next place prediction using mobility markov chains[C] //Proceedings of the First Workshop on Measurement, Privacy, and Mobility. [S.l]: ACM, 2012: 3.

[25] NOULAS A, SCELLATO S, LATHIA N, et al. Mining user mobility features for next place prediction in location-based services[C]//ICDM. [S.l.]: [s.n.], 2012, 12: 1038-1043.

[26] LU Xin, BENGTSSON L, HOLME P. Predictability of population displacement after the 2010 Haiti earthquake[J]. Proc Natl Acad Sci, 2012, 109(29): 11576-11581.
[27] GALLOTTI R, BAZZANI A, ESPOSTI M D, et al. Entropic measures of individual mobility patterns[J]. Journal of Statistical Mechanics: Theory and Experiment, 2013(10): P10022.
[28] BAUMANN P, KLEIMINGER W, SANTINI S. The influence of temporal and spatial features on the performance of next-place prediction algorithms[C] //Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. [S.l]: ACM, 2013: 449-458.
[29] LU Xin, WETTER E, BHARTI N, et al. Approaching the limit of predictability in human mobility[J]. Scientific Report, 2013(3): 2923.
[30] LAHERRERE L, SORNETTE D. Stretched exponential distributions in nature and economy: “fat tails” with characteristic scales[J]. Euro Phys J B, 1998, 2: 525.
[31] ZHOU T, WANG B H, JIN Y D, et al. Modelling collaboration networks based on nonlinear preferential attachment[J]. Int J Mod Phys C, 2007, 18: 297-314.
[32] SHANG Ming-sheng, LÜ Lin-yuan, ZHANG Yi-cheng, et al. Empirical analysis of web-based user-object bipartite networks[J]. Europhys Lett, 2010(90): 48006.
[33] GOH K I, BARABASI A L. Burstiness and memory in complex systems[J]. Europhys Lett, 2008(81): 48002.
[34] YAN Xiao-yong, HAN Xiao-pu, ZHOU Tao, et al. Exact solution of the gyration radius of an individual’s trajectory for a simplified human regular mobility model[J]. Chin Phys Lett, 2011, 28(12): 120506.

编 辑 蒋 晓