基于前向搜索和投票的移动群智感知动态用户招募方法

纪圣堪[1)] 郑 宇[1),2)] 王诏远[1)] 李天瑞[1)]

[1)](西南交通大学计算机与人工智能学院,成都 611756)

[2)](京东智能城市研究院,北京 100176)

摘 要 随着移动传感设备的快速发展,移动群智感知已经成为城市数据收集的一项重要模式.用户招募——招募进行数据收集的用户——是移动群智感知中的重要一环.现有的研究主要集中在静态的用户招募,对动态的用户招募研究还比较少.因此,本文提出了一个基于前向搜索和投票的动态用户招募方法,该方法能够很好地综合多个因素,从而做出更优的招募决策.除此之外,本文还设计了一个新的数据均匀程度指标来更好地评估数据的质量.基于真实世界的Foursquare用户数据的实验结果表明,相比于基准方法,本文的方法能够收集到均匀程度更高的数据.

关键词 移动群智感知;动态用户招募;前向搜索;爬山算法;数据均匀指标

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2021.01998

Look-ahead Search and Voting-based Dynamic Participant Recruitment in Mobile Crowd Sensing

JI Sheng-Gong[1)] ZHENG Yu[1),2)] WANG Zhao-Yuan[1)] LI Tian-Rui[1)]

[1)](*School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu* 611756)

[2)](*JD Intelligent Cities Research, Beijing* 100176)

Abstract With the rapid development of mobile crowd sensing devices and mobile crowd sensing technologies, in recent years mobile crowd sensing has become one of the important modes to collect the data in urban areas. Different with the traditional static sensing methods, mobile crowd sensing leverages mobile devices, such as mobile phones, wearable devices and so on, to collect data in urban areas. Participant recruitment, i.e., recruiting participants to collect urban data, is of great importance to collecting high-quality data and thus plays an essential role in mobile crowd sensing. Given its importance, participant recruitment has achieved extensive attention, and a large number of related studies have been conducted. However, most of the existing research focuses on the static participant recruitment, rather than the dynamic participant recruitment. In contrast to the static participant recruitment, the dynamic participant recruitment can dynamically adjust its recruitment strategy, based on the real-time status in the system, e.g., the data already collected, the budget left. As a result, the dynamic participant recruitment can be much more flexible than the static participant recruitment and is more likely to collect higher-quality data. To this end, in this paper we study the dynamic participant recruitment problem in mobile crowd sensing. However, dynamic participant recruitment is more challenging than static participant recruitment, too. In

收稿日期:2020-02-19;在线发布日期:2021-01-11.本课题得到国家重点研发计划(No. 2019YFB2101802)资助.**纪圣堪**,博士,主要研究领域为城市资源优化、社交网络.E-mail:shenggongji@163.com. **郑 宇**(通信作者),博士,教授,中国计算机学会(CCF)会员,IEEE Fellow,主要研究领域为人工智能、城市计算.E-mail:msyuzheng@outlook.com. **王诏远**,博士,主要研究领域为多源数据融合、数据挖掘;**李天瑞**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能、大数据.E-mail:trli@swjtu.edu.cn.

detail，in dynamic participant recruitment we need to consider and fuse more factors，for example the budget left，the data already collected，the data that a participant can collect，and the potential participants in the future. To deal with this challenge，in this paper we design a novel look-ahead search and voting-based dynamic participant recruitment method for mobile crowd sensing. Specifically，the proposed method consists of two main components. In the first component，we propose a look-ahead search method to model the potential participants in the future，which are unknown and should be estimated. In the second component，we mathematically formulate the dynamic participant recruitment problem as a combinatorial optimization problem and then we propose a heuristic hill climbing-based voting method to solve the combinatorial optimization problem. With these two components，the proposed method can well fuse the multiple related factors and can make better participant recruitment decisions. In addition，we provide a new hierarchical entropy-based data balance indicator to better measure the quality of the collected urban data. Existing data balance indicators see each grid in a city equally important which however in many scenarios is not reasonable，and existing data balance indicators cannot adapt to the complex situations in urban areas，either. Our proposed hierarchical entropy-based data balance indicator can well handle these problems. Extensive experiments using check-in data from real-world Foursquare users demonstrate that comparing with existing methods，our method can significantly improve the balance degree of the collected data. Specifically，compared with the best baseline method，our method can reduce 13.33%（for dataset 1）and 19.64%（for dataset 2）of the hierarchical coefficient of variation，can improve 4.65%（for dataset 1）and 11.82%（for dataset 2）of the hierarchical coverage，and can reduce 13.08%（for dataset 1）and 23.10%（for dataset 2）of the average minimum distance.

Keywords mobile crowd sensing；dynamic participant recruitment；look-ahead search；hill climbing；data balance indicator

1 引 言

移动群智感知（Mobile Crowd Sensing）用于城市数据的收集[1-4]，是城市计算（Urban Computing）领域的一类重要应用[5-6]. 与传统的固定传感器不同，移动群智感知主要基于移动传感器对城市数据进行收集，例如：出租车的GPS系统、移动手机内置传感器、可穿戴设备等. 基于众多的移动传感器，移动群智感知可以收集到各式各样的城市数据，包括出租车的轨迹数据、噪声数据、空气质量数据等[7-11]. 由于收集城市大数据的重要性，移动群智感知已经受到国内外学者的广泛关注[1-4].

用户招募（Participant Recruitment）是移动群智感知的重要部分，其旨在在一定经费的前提下招募合适的用户使得收集到的数据质量最优，亦或是在保证数据质量达到一定标准的前提下最小化招募用户所需要的经费[10,12-13]. 虽然用户招募已经得到较为广泛的研究，但现有的研究大多集中在静态用户招募，对动态用户招募的研究还较少[12]. 如图1a所示，静态用户招募是指在数据收集开始之前就完成对用户的招募，数据收集开始后不再进行用户招募. 动态用户招募则在数据收集过程中对于每一个实时到来的用户进行招募（图1b）. 动态用户招募比静态用户招募更加灵活，可以根据实时收集到的数据以及实时参与的用户，动态地调整招募策略[12]. 现有的研究大多集中在静态用户招募，原因主要有两点. 第一，在现有的许多移动群智感知中，用户招募和数据收集是分开的两个阶段（图1a），在数据收集开始之前就需要完成对用户的招募. 因此，所提出的用户招募算法便属于静态用户招募算法. 第二，正如下文将要介绍的，动态用户招募比静态用户招募难度更大，因此现阶段的研究更多的是相对较为容易的静态用户招募，对动态用户招募的研究才刚刚起步.

动态用户招募的难度更大，原因在于其决定是否招募一个实时到来的用户时需要综合考虑四个方面的因素：（1）剩余的经费（总经费减去已用去的经费）；（2）已经收集到的数据；（3）该实时用户所能够

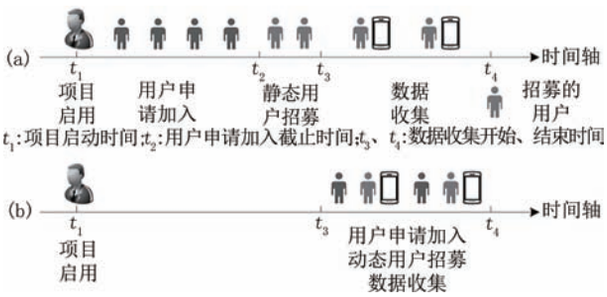

图1 静态用户招募与动态用户招募

收集到的数据;(4)未来的潜在用户.如何有效地融合这四个因素并做出更优的招募决策从而最终收集到更高质量的数据是一个巨大的挑战.需要注意的是,未来的潜在用户对当前用户的招募决策有巨大影响.例如,若能预见到未来的用户能够收集到质量更高的数据,就可以不必招募当前用户.然而未来的潜在用户是未知的,具有较大的不确定性,因此有效地对其进行考虑并不容易.静态用户招募不对实时到来的用户进行招募,从而不需要考虑未来的潜在用户,因此相对较为容易.

数据质量评估是移动群智感知的另一重要部分[14-18].一个好的数据质量评估指标对于用户招募和数据收集具有重要的指导意义.数据的均匀程度是一项重要的数据质量评估指标[10],虽然已有研究提出了较为有效的数据均匀程度指标,但是其适用性不够,不能适用于复杂的城市环境[10].因此,有必要提出一个新的数据均匀程度指标.

为解决以上两个问题,本文提出了一个新的移动群智感知动态用户招募方法.具体而言,本文的贡献有以下三个方面:

(1)提出了一个高效的动态人员招募方法,该方法能够很好地综合所需考虑的四种因素,从而做出更优的动态用户招募决策.

(2)提出了一个新的基于层次信息熵的数据均匀程度指标,该指标能够更好地反映城市数据的均匀程度,具有更大的适用性.

(3)运用真实世界的数据进行实验,验证了所提出的动态人员招募方法的高效性,与基准方法相比,能够大幅提升收集到的数据的均匀程度.

2 问题定义

2.1 名词定义

定义1. (用户)一个用户定义为一个元组$r=(lat,lon,time)$,其中,lat和lon代表该用户将进行数据收集的位置(经度和纬度),$time$代表该用户将进行数据收集的时间.

当一个用户$r=(lat,lon,time)$实时到来时,本文的动态用户招募方法将进行实时的招募决策:招募或不招募.

定义2. (格子)一个格子定义为一个矩形区域,例如大小为300 m×400 m的矩形区域.

本文将进行数据收集的空间分割成许多大小一样的$I\times J$个格子,其中I代表纬度方向的格子数,J代表经度方向的格子数.

定义3. (时间段)一个时间段定义为一定长度的时间,例如一个小时作为一个时间段.

本文将进行数据收集的时间范围分割为等长的T个时间段.例如,若数据收集的时间范围为上午8点到晚上20点,而一个小时为一个时间段,那么总共有$T=12$个时间段.

定义4. (数据矩阵)数据矩阵指收集到的数据矩阵,是一个大小为$I\times J\times T$的三维矩阵D,其中$D(i,j,t)$记录的是在格子(i,j)、时间段t收集到的数据数量.

若一个用户$r=(lat,lon,time)$被招募,则需将$(lat,lon,time)$映射到对应的格子和时间段,记为$(lat_{id},lon_{id},time_{id})$.因此,对于用户$r$,定义$D_r$为用户$r$能够收集到的三维数据矩阵:

$$D_r(i,j,t)=\begin{cases}1 & if\ (i,j,t)=(lat_{id},lon_{id},time_{id})\\ 0 & otherwise\end{cases} \tag{1}$$

D_r除了在$(lat_{id},lon_{id},time_{id})$等于1以外,其他位置均为0.

2.2 问题定义

动态用户招募:给定总经费B、数据收集的时间段T和空间大小I、J,当用户$r=(lat,lon,time)$实时到来时,动态用户招募旨在实时地决定是否招募该用户r,使得最终收集到的总数据D的均匀程度指标$f(D)$最大.

具体地,在决定是否招募用户r时,需要考虑以下四个因素:

(1)剩余的经费B_{left}(总经费减去已用去的经费);

(2)已经收集到的数据矩阵D_{col};

(3)用户r所能收集到的数据矩阵D_r;

(4)未来潜在的用户集合R_{fut}.

表1展示了本文将用到的一些主要符号及其含义.除了动态用户招募方法(第四节),本文还提出了一个新的基于层次信息熵的数据均匀程度指标$f(D)$(第五节).

表1　主要符号及其释义

符号	释义
B	总经费
I	纬度方向的格子数
J	经度方向的格子数
T	时间段个数
D_{col}	已经收集到的数据矩阵
r	一个用户
D_r	用户r所能收集到的数据矩阵
$I(r)$	对用户r的招募决策；$I(r)=1$代表招募该用户；$I(r)=0$代表不招募该用户
R_{arrived}	已经到来的用户集合
R_{fut}	未来将要到来的用户集合

3　相关文献

由于移动群智感知对城市数据收集的重要性，现已有很多研究从不同角度对其进行研究[1-4]，包括用户招募[9-10,12-13]、数据质量[14-18]、能源消耗[19-22]、用户隐私[23-25]、激励机制[26-33]等[34-35]. 用户招募、数据质量和激励机制与本文的研究最为相关，因此本文将在3.1、3.2和3.3节中对其进行详细介绍. 对能源消耗的研究旨在降低移动群智感知过程中所需的能源消耗，如电量、传输带宽、计算资源等[19-22]. 降低能源消耗能够提升移动感知设备的使用时长以及避免资源的过度浪费. 用户隐私和安全是进行移动群智感知过程中必须考虑的[23-25]. 现有的研究提出一些能够保护用户隐私的移动群智感知策略来避免用户敏感数据的泄露.

3.1　用户招募

用户招募可以分为静态用户招募和动态用户招募[12]. 由于现有的许多移动群智感知都把用户招募和数据收集分为两个阶段(见图1a)，因此对静态用户招募的研究较多[9-13,36-39]. Zhang等[9]在进行静态用户招募时考虑用户的历史移动轨迹数据，提出迭代算法进行用户招募使得在数据覆盖率达到一定标准的前提下，其所需招募的用户数量最小. 与文献[9]类似，Xiong等[13]也考虑了用户历史的移动轨迹数据，并提出了一个基于嵌套贪心搜索算法的静态用户招募方法. Ji等[10]针对用户提交的移动信息数据，在数据收集之前对用户进行招募，旨在在有限经费资源下最大化所能收集到的数据的均匀程度.

动态用户招募则是在数据收集过程中实时地对用户进行招募[12,22,37](见图1b). 由于需要考虑未来的潜在用户，动态用户招募会比静态用户招募更难一些. 对动态用户招募方法的研究才刚刚起步，包括文献[37]的EEMC方法和文献[22]的EMC^{3}方法，具体见综述论文[12]. 表2展示了本文的动态用户招募方法与文献[22,37]中的方法的异同点. 本文的方法考虑的是整个城市细粒度的数据收集，将城市划分为较小的格子，例如，在本文实验中，纬度方向有$I=76$个格子，经度方向有$J=43$个格子. 因此每个时间段需要收集大量的数据($I\times J=3268$). 但同时本文考虑的是有限的经费，且经费数量远远小于所要收集的数据数量. 因此本文的方法只收集部分的数据，旨在最大化收集到的数据的均匀程度. 均匀程度越大，收集到的数据越能代表未收集到的数据，越有助于未来的数据分析和使用.

文献[37]的方法考虑的则是粗粒度的数据收集，应用于较小数量数据的收集. 在其实验中，仅收集13个大区域的数据，每个时间段只需收集130个数据. 由于所要收集的数据量较少，其在用户招募过程中要求所有数据都必须收集到，并旨在最小化招募到的冗余用户数量.

文献[22]的方法与文献[37]的方法较为类似，主要差别为两点. 第一，其增加了一个要求，要求收集到的数据覆盖全部13个大区域. 第二，将目标改为最小化招募到的用户数量.

从表2可以看出，本文的方法和文献[37,22]中的方法具有不同的应用场景，是相互补充的关系，能够满足不同的数据收集需求.

3.2　数据质量

数据质量是指导移动群智感知的重要指标. 现有的数据质量指标主要包括以下四种. 首先是数据覆盖率，其用于评估收集到的数据占所需要收集的总数据的比例[9,14-15]. 其通常用于在给定数据覆盖率前提下最小化所需要的总经费. 其次是数据的推断

表2　现有的动态用户招募方法

方法	需要收集的数据量	经费	实际收集的数据量	覆盖全部区域	目标
本文	巨大	给定	小部分	不需要	最大化收集到的数据的均匀程度
EEMC[37]	较小	不给定	全部	不需要	最小化招募到的冗余用户数
EMC^3[22]	较小	不给定	全部	需要	最小化招募到的用户数

能力，其用于评估收集到的数据对未收集到的数据的推断能力[17-18]. 城市数据具有较为明显的时空相关性，因此收集部分数据就可能可以较为准确地推断全部的数据. 推断的准确性就是数据的推断能力. 接着是数据的均匀程度[10]，其评估的是收集到的数据的时间和空间均匀程度. 很显然，越均匀的数据越具有代表性，越有利于后续的数据分析和应用. 本文提出了一个新的、更有效的、应用面更广的数据均匀性指标. 最后是基于专家经验定义的数据质量[15]. 在一些移动群智感知应用中，数据的质量可以由领域专家直接定义.

3.3 激励机制

激励是用户参与移动群智感知项目的重要动力，最常见的激励就是金钱. 如何合理地设计激励机制来提升用户的积极性和收集到的数据质量是重要的研究内容[26-33]. 城市中用户的分布在空间上是很不均匀的[31]，因此为了收集到更加均匀的数据，有必要为不同区域的数据设置不同的奖励. 南文倩等[29]提出了一个基于跨空间多元交互的激励模型，利用逆向拍卖方法计算每个数据的合理奖励，提高了用户接受任务的意愿以及数据收集的完成率. 相似地，Lee等[33]也提出了一个逆向拍卖方法，既能用较低的奖励维持较高的用户参与度，又能提升收集到的数据的均匀程度. Kawajiri等[32]提出了一个激励机制模型，通过假设用户对激励的反应函数，为不同区域的数据设置不同的奖励. 如果一个区域收集到的数据越少，数据价值越高，奖励也就越高，以此来鼓励用户去数据较少的区域进行数据收集. 更多的关于移动群智感知激励设计方法可以参考综述论文[26-27].

激励机制主要通过激励来引导用户去收集高价值数据，而用户招募则是招募能够收集高价值数据的用户来最大化收集到的数据质量. 本质上，它们的目标是相近的，只是运用的手段不同.

4 动态用户招募方法

4.1 整体思路

在决定是否招募用户r时所需要考虑的四个因素中，只有第四个因素R_{fut}是未知的. 为此，本文提出运用前向搜索来预估未来的R_{fut}. 前向搜索的思想在很多决策问题中得到应用，例如：在用人工智能算法下围棋时，决定当前的落子决策就需要考虑该决策下未来可能出现的情形，即用前向搜索未来的可能情形[40-41].

基于前向搜索的思想，如算法1所示，当一个用户r实时到来时，本文的动态用户招募方法的招募过程如下：(1)运用前向搜索算法生成M次未来的潜在用户集合R_{fut}，每一次生成的用户集合标记为R^m_{fut}，$m=1,\cdots,M$；(2)对于每一个生成的R^m_{fut}以及前三个因素，建立数学优化模型，求解该模型得到当前用户r的招募决策$I^m(r)$，即基于生成的R^m_{fut}对是否招募当前用户r进行投票；(3)经过M次的投票后，如果投票招募该用户的次数n_{rec}大于$M/2$，则招募该用户r，即$I(r)=1$；否则不招募，即$I(r)=0$. 需要注意的是，算法1中的M次投票是独立的，因此是可以并行的，在计算资源充足的情况下，该招募算法可以并行，可以大幅提升该招募算法的运行效率.

算法1. 动态用户招募算法

输入：B_{left}、D_{col}、D_r、M

输出：$I(r)$

$n_{\text{rec}}=0$

FOR $m=1,2,\cdots,M$ DO

　　前向搜索生成R^m_{fut}　　%% 4.4节

　　求解优化模型3并得到解$I^m(r)$　　%% 算法2

END

计算n_{rec}　　%%公式4

IF $n_{\text{rec}}>M/2$：$I(r)=1$

ELSE：$I(r)=0$

在实时招募的过程中，根据B_{left}和D_{col}的定义，B_{left}和D_{col}也会不断地更新，其更新过程如下：

$$\begin{cases} B_{\text{left}}=B-\sum_{r\in R_{\text{arrived}}} I(r) \\ D_{\text{col}}=\sum_{r\in R_{\text{arrived}}} I(r)D_r \end{cases} \tag{2}$$

4.2 数学优化模型

现在假定R^m_{fut}已经由前向搜索算法生成. 给定R^m_{fut}以及前三个因素B_{left}、D_{col}和D_r，可以建立以下数学优化模型来综合考虑这四个因素：

$$\begin{aligned} &\max\ f\left(D_{\text{col}}+I^m(r)D_r+\sum_{r'\in R^m_{\text{fut}}} I^m(r')D_{r'}\right) \\ &s.t.\ \begin{cases} I^m(r)+\sum_{r'\in R^m_{\text{fut}}} I^m(r')=B_{\text{left}} \\ I^m(r),I^m(r')\in\{0,1\},r'\in R^m_{\text{fut}} \end{cases} \end{aligned} \tag{3}$$

这里f是数据质量指标，用于评估收集到的数据的质量. 本文提出了一个新的数据均匀程度指标f来评估数据质量(5.3节公式7). f可以是任意的数据质量指标，模型3对所有的数据质量指标都是通用的. $I^m(r)$和$I^m(r')$是0-1决策变量，$I^m(r)=1$代表招募当前用户r，$I^m(r)=0$代表不招募当前用户r；

同理，$I^m(r')=1$代表招募用户r'，$I^m(r')=0$代表不招募用户r'.

该优化模型的目标是对当前用户r和未来用户R^m_{fut}进行整体招募，从而最优化所能收集到的所有数据的整体均匀程度. 未来用户是基于前向搜索得到的R^m_{fut}，因此可以看成R^m_{fut}对是否招募当前用户r的一次投票. 该模型背后的含义是，如果我们能够精确地知道未来用户R_{fut}，那么就可以进行整体的优化；但是由于我们并不知道精确的R_{fut}，因此基于前向搜索生成的R^m_{fut}只能作为是否招募当前用户r的一次投票. 第一个约束是指所有招募的人员数量要等于剩余的经费.

基于优化模型3的答案$I^m(r)$，$m=1,\cdots,M$，计算投票招募用户r的次数n_{rec}为

$$n_{\text{rec}}=\sum_{m=1}^{M}I^m(r) \tag{4}$$

如果$n_{\text{rec}}>M/2$，则基于前向搜索生成的M次R^m_{fut}，招募用户r是一个更好的决策，因此就决定招募用户r，即$I(r)=1$. 注意，M最好取奇数，以避免出现$n_{\text{rec}}=M/2$的情况.

4.3　优化模型的解法

由于本文的数据均匀程度指标$f(D)$是一个复杂的非线性方程(公式7)，因此优化模型3其实是一个复杂的非线性0-1整数优化，是一个典型的NP难问题[42]. 对于每个用户，都有两种可能的决策方案：招募或者不招募，因此该问题的搜索空间是$2^{(1+|R^m_{\text{fut}}|)}$，其中$|R^m_{\text{fut}}|$是$R^m_{\text{fut}}$中用户的个数. 由于NP难问题很难被高效地、快速地求解，因此本文设计了一个启发式算法来求解优化模型3. 具体而言，如算法2所示，本文设计了一个爬山算法来求解模型3[43-44].

整体来说，本文的爬山算法分别考虑$I^m(r)=1$和$I^m(r)=0$的情况. 对于$I^m(r)=1$的情况，就有$B_{\text{left}}-1$的经费来招募R^m_{fut}中的用户. 具体地，可以分两步从R^m_{fut}招募$B_{\text{left}}-1$用户. 第一，从R^m_{fut}中随机选取$B_{\text{left}}-1$个用户作为初始解，记为R_{sel}. 第二，尝试用$R^m_{\text{fut}}\backslash R_{\text{sel}}$中的一个随机用户替换$R_{\text{sel}}$中的一个随机用户，看是否能够提升所收集到的数据的均匀程度，如果能够提升，则保留这个替换；否则不保留. 第二步就是爬山的过程，通过N_{attempt}次的替换尝试，可以不断改进初始解R_{sel}，最后得到$I^m(r)=1$的情况下的数据均匀程度f_{rec}. f_{rec}代表当招募当前用户r时，所能够获得的最大的数据均匀程度.

同理，对于$I^m(r)=0$的情况，也可以用相同的过程求解$I^m(r)=0$的情况下的数据均匀程度f_{not}，但是有两点不同. 首先，由于$I^m(r)=0$，我们可以随机从R^m_{fut}中随机选取B_{left}个用户作为初始解；其次，同样由于$I^m(r)=0$，在计算数据均匀程度时，D_r不能计算在内.

最后，通过对比f_{rec}和f_{not}的结果来决定$I^m(r)$的值，即如果$f_{\text{rec}}>f_{\text{not}}$，$I^m(r)=1$；否则$I^m(r)=0$.

算法2.　爬山算法求解优化模型3

输入：B_{left}、D_{col}、D_r、R^m_{fut}、N_{attempt}

输出：$I^m(r)$

$R_{\text{sel}}\leftarrow$随机选取R^m_{fut}中的$B_{\text{left}}-1$个　%% 如果$I^m(r)=1$

FOR $q=1,2,\cdots,N_{\text{attempt}}$

　$r_{\text{in}}\leftarrow$随机选取R_{sel}中的一个

　$r_{\text{out}}\leftarrow$随机选取$R^m_{\text{fut}}\backslash R_{\text{sel}}$中的一个

　$R'_{\text{sel}}=R_{\text{sel}}\setminus\{r_{\text{in}}\}\cup\{r_{\text{out}}\}$

　IF $f\left(D_{\text{col}}+D_r+\sum_{r'\in R'_{\text{sel}}}D_{r'}\right)>f\left(D_{\text{col}}+D_r+\sum_{r'\in R_{\text{sel}}}D_{r'}\right)$

　　$R_{\text{sel}}=R'_{\text{sel}}$

　END

END

计算$f_{\text{rec}}=f\left(D_{\text{col}}+D_r+\sum_{r'\in R_{\text{sel}}}D_{r'}\right)$

$R_{\text{sel}}\leftarrow$随机选取R^m_{fut}中的B_{left}个　%% 如果$I^m(r)=0$

FOR $q=1,2,\cdots,N_{\text{attempt}}$ DO

　$r_{\text{in}}\leftarrow$随机选取R_{sel}中的一个

　$r_{\text{out}}\leftarrow$随机选取$R^m_{\text{fut}}\backslash R_{\text{sel}}$中的一个

　$R'_{\text{sel}}=R_{\text{sel}}\setminus\{r_{\text{in}}\}\cup\{r_{\text{out}}\}$

　IF $f\left(D_{\text{col}}+\sum_{r'\in R'_{\text{sel}}}D_{r'}\right)>f\left(D_{\text{col}}+\sum_{r'\in R_{\text{sel}}}D_{r'}\right)$

　　$R_{\text{sel}}=R'_{\text{sel}}$

　END

END

计算$f_{\text{not}}=f\left(D_{\text{col}}+\sum_{r'\in R_{\text{sel}}}D_{r'}\right)$

IF $f_{\text{rec}}>f_{\text{not}}$：$I^m(r)=1$

ELSE：$I^m(r)=0$

4.4　前向搜索

正如前文介绍的，本文需要生成M次的未来潜在用户集合R^m_{fut}. 但是，因为正如模型3中所示的，对于每个未来的潜在用户r'，我们需要的并不是具体的$r'=(lat',lon',time')$，而是数据矩阵Dr'. 因此，可以直接生成未来每个格子(i,j)在每个时间段t有多少用户，记为$R(i,j,t)$，其中$i=1,\cdots,I$，$j=1,\cdots,J$，$t=time_{id},\cdots,T$. $time_{id}$是当前到来的用户r所在的时间段. 当将生成的$R(i,j,t)$带入算法2时，需要将$R(i,j,t)$转换成用户集合R^m_{fut}，例如：$R(i_0,j_0,t_0)=2$

时，就有两个(i_0,j_0,t_0)在R^m_{fut}中.

所以本质上，我们是要根据历史的用户数据来生成预测未来的用户.现在已经有很多现成的生成预测模型可以直接应用，因此可以直接应用这些算法来进行生成预测[45-46]，本文的动态用户招募方法框架能够支持所有的生成预测模型.由于在移动群智感知中，通常收集数据往往只有几天时间，没有太多的训练数据来支持训练复杂的模型，再加上本文更注重动态用户招募的整体策略，因此本文选用一个较为简单的生成预测模型.具体而言，假设用户的到来服从一个泊松过程，每个格子(i,j)在每个时间段t出现的用户数量$R(i,j,t)$是一个均值为$\lambda(i,j,t)$泊松分布，即$R(i,j,t)$的概率分布为

$$\mathrm{Prob}\{R(i,j,t)=l\}=\frac{e^{-\lambda(i,j,t)}\lambda(i,j,t)^l}{l!} \quad (5)$$

根据泊松分布的性质，可以用过去几天格子(i,j)在时间段t出现的用户数量的平均值作为$\lambda(i,j,t)$.因此，根据上述概率分布，就可以生成$R(i,j,t)$，然后将其转换成R^m_{fut}.

5 数据均匀程度指标

5.1 研究动机

文献[10]提出了一个基于层次信息熵的数据均匀程度指标，其核心思想是考虑数据D在不同时空粒度下的信息熵作为数据D的整体均匀程度[10,47].具体而言，其首先考虑k_{max}个不同的时空粒度，将数据D映射到k_{max}个不同的时空粒度中，从而得到数据$D^1,\cdots,D^{kmax}$.例如，在图2a中(为了更好地展示，这里展示2维数据，3维数据的情况类似)，原始的2维数据D的粒度是4×4，现将其转换成图2b的2×2的粒度.接着，计算不同粒度下的数据$D^1,\cdots,D^{kmax}$的信息熵，取其加权平均值得到数据D的均匀程度指标.综上，该指标认为原始粒度$I\times J\times T$的数据信息熵并不能反映该数据的真实均匀程度，因此需要综合考虑该数据在不同粒度下的信息熵.

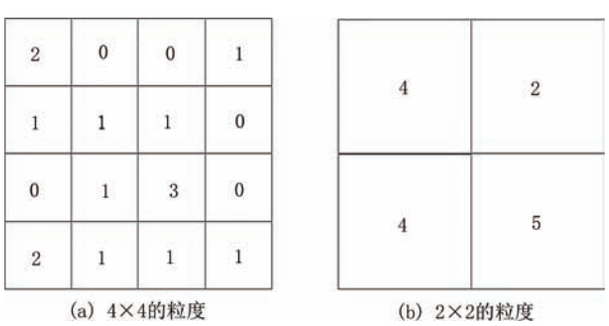

图2 数据的不同粒度表示

虽然文献[10]的指标取得了较好的效果，但是其存在以下两个问题.第一，该指标将城市中的每个格子视为同等重要，这并不总是成立.例如在图3a中，城市中有一些区域是用户无法到达的，是无法收集到数据的，例如河流、湖泊等.因此，计算数据均匀程度时须将这些区域排除掉.然而论文[10]中的指标没有考虑到这种情况.第二，论文[10]中的指标要求I、J和T必须能够被约分.例如，如果$T=13$，其不能被约分，那么该指标就无法使用.因此，有必要对论文[10]中的数据均匀程度指标进行改进和优化，拓展其可用性.

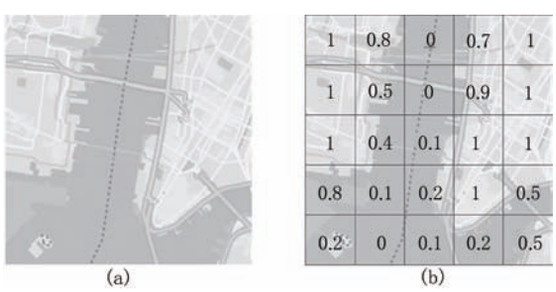

图3 论文[10]中的指标存在的不足

5.2 整体思路

与论文[10]中的数据均匀程度指标类似，本文也考虑k_{max}个不同的时空粒度，将数据D映射到k_{max}个不同的时空粒度，得到数据$D^1,\cdots,D^{kmax}$.但是本文提出以下三点重要的改进.

第一，将城市中的每个格子视为不同等重要的，赋予每个格子一个权重.如图3b所示，河流所在的格子的权重都很小，甚至有些全部是0.一个格子权重的大小取决于这个格子有多大比例的区域是用户能够达到的.基于此思路，就可以构造一个三维权重矩阵W，其大小也是$I\times J\times T$.由于本文暂时只考虑格子之间具有不同权重，因此W在同个格子不同时间段的权重是相等的.实际上，本文的权重矩阵W也能够直接拓展到不同时间段具有不同权重的情况.

第二，本文运用卷积的思想来生成不同的粒度.如图4所示，原本5×5的粒度，通过大小为2×2的卷积核，变成了4×4的粒度.在新生成的粒度中，每个格子的权重等于原始粒度下进行卷积的格子的权重之和，例如3.3 = 1 + 0.8 + 1 + 0.5、1.6 = 0 + 0.7 + 0 + 0.9等.

对于k_{max}个不同的时空粒度，本文将第k个粒度的卷积核大小记为$n^k_i\times n^k_j\times n^k_t$.根据卷积核的大小，

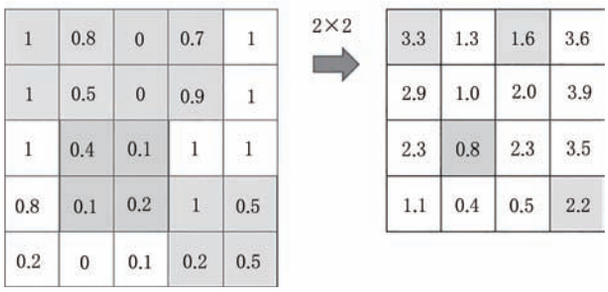

图4　基于卷积思想生成不同的粒度及其权重矩阵

可以计算得到第k个粒度下的数据矩阵D^k和权重矩阵W^k的大小都是：

$$I^k=I-n_i^k+1, J^k=J-n_j^k+1, T^k=T-n_t^k+1 \tag{6}$$

运用卷积的思想，本文的指标就不必要求I、J和T必须能够被约分，从而极大地拓展了层次信息熵的可用性.

第三，论文[10]中只考虑了层次信息熵，本文则将数据均匀指标拓展到层次离散系数[48-49]和层次覆盖率[9,14-15]. 需要注意的是，离散系数和覆盖率也能很好地评估数据分布的均匀程度. 本文将层次信息熵作为目标，并将层次离散系数、层次覆盖率等作为实验的评价指标. 因此，下面详细进行层次信息熵的推导，层次离散系数和层次覆盖率则在实验中具体介绍.

5.3　基于层次信息熵的数据均匀程度指标

与论文[10]类似，本文的基于层次信息熵的数据均匀程度指标$f(D)$定义如下：

$$f(D)=\frac{1}{k_{max}}\sum_{k=1}^{k_{max}}\frac{Ent\left(D^k, W^k\right)}{\log_2 I^k J^k T^k} \tag{7}$$

其中，$Ent(D^k, W^k)$表示的是数据在第k个粒度下的信息熵. $\log_2 I^k J^k T^k$表示的是数据第k个粒度下的理论最大信息熵. 因此，$Ent(D^k, W^k)/\log_2 I^k J^k T^k$就表示收集到的数据的信息熵与理论最大信息熵的比值. 之所以运用比值是因为不同粒度k的I^k、J^k和T^k不同. 本文的层次信息熵就是不同粒度下的信息熵的加权平均值. 具体来说，第k个粒度下数据的信息熵$Ent(D^k, W^k)$的定义是

$$Ent\left(D^k, W^k\right)=-\sum_{i,j,t}p^k(i,j,t)\log_2 p^k(i,j,t) \tag{8}$$

其中，$p^k(i,j,t)$为：

$$p^k(i,j,t)=\begin{cases}0, \ if \ W^k(i,j,t)=0\\ \dfrac{D_w^k(i,j,t)}{\sum\limits_{i',j',t',W^k(i',j',t')\neq 0}D_w^k(i',j',t')}, \ otherwise\end{cases} \tag{9}$$

$D_w^k(i,j,t)$为(i,j,t)中的单位权重数据数量，即数据数量$D^k(i,j,t)$除以权重$W^k(i,j,t)$：

$$D_w^k(i,j,t)=\frac{D^k(i,j,t)}{W^k(i,j,t)} \tag{10}$$

因此，$p^k(i,j,t)$是(i,j,t)中的单位权重数据数量占所有格子的总单位权重数据数量的比例.

5.4　时间复杂度

计算$f(D)$的过程包括两步：获取D^k和计算$Ent(D^k, W^k)$. 首先，获取D^k的过程是卷积计算的过程. 由于D^k中的每一个元素是原始D中$n_i^k\times n_j^k\times n_t^k$个数据之和，因此，其获取每个元素需要的计算量是$O(n_i^k\times n_j^k\times n_t^k)$. 而$D^k$的总元素个数是$I^k\times J^k\times T^k$，因此，获取$D^k$的总计算量是$O(n_i^k\times n_j^k\times n_t^k\times I^k\times J^k\times T^k)$. 计算$Ent(D^k, W^k)$只需要对$D^k$进行两次循环，因此时间复杂度是$O(I^k\times J^k\times T^k\times 2)$. 因此，计算$f(D)$的总时间复杂度是：

$$O\left(\sum_{k=1}^{k_{max}}I^k J^k T^k\times\left(n_i^k n_j^k n_t^k+2\right)\right) \tag{11}$$

由于W^k是固定的，因此并不需要计算在时间复杂度里面.

6　实　验

6.1　数据集和实验设置

本文用真实世界的用户数据来评估所提出的动态用户招募方法. 实验需要两部分的数据：一是用户数据；二是城市地图数据. 对于用户数据，本文运用美国纽约市曼哈顿区域的Foursquare用户打卡(check-in)数据[50]. 每个Foursquare用户的打卡数据包括其打卡的时间和地点(经纬度). 因此Foursquare用户数据符合本文对用户的定义，同时其也能够反映该城市中用户真实的时间和空间分布. 对于地图数据，纽约市曼哈顿区域的地图数据可以从公开的地图数据网站OpenStreetMap上获取[51].

本文随机选取2012年7月2日-2012年7月7日(数据集1)和2012年5月11日-2012年5月16日(数据集2)两个时间段的Foursquare用户数据进行实验. 数据集1和数据集2中申请加入的用户总数量分别为342和776人. 用户数据的时间和空间分布如图5所示. 总经费B设置为$B=200$，最多可以招募200个用户来收集200个数据. 两个时间段的数据都是6天，前5天的数据用来训练前向搜索模型(第4.4节)，第6天的数据作为测试数据来评估所提出的动态用户招募算法. 对于每个实验设置(M和N_{attempt})，本文都进行10次的实验，取其平均值作为最终的实验结果.

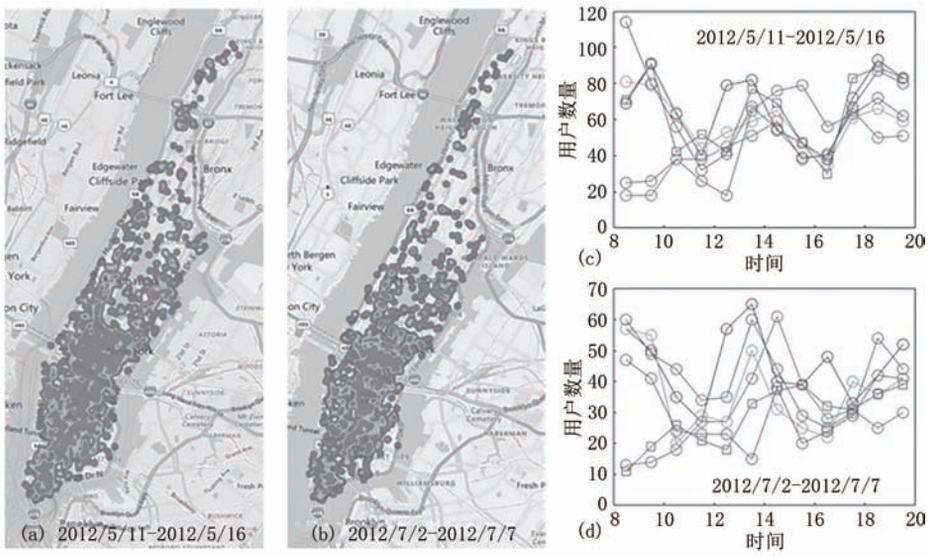

图5 Foursquare用户数据的时间和空间分布

本文将纽约市曼哈顿区分割成76×43个格子，每个格子的大小为300 m×300 m. 运用地图数据，根据格子中用户可以到达的面积比例对每个格子进行赋权. 收集数据的时间区间设为上午8点到下午20点，每个小时作为一个时间段，共12个时间段. 因此，有$I=76$、$J=43$、$T=12$. 数据均匀程度指标考虑4个不同的空间粒度和时间粒度，每个粒度的卷积核大小及其数据大小设置如表3所示. 第一粒度的卷积核大小为5×3×2，数据矩阵的大小为72×41×11. 第二粒度的卷积核大小为第一粒度的两倍，第三粒度的则是第二粒度的两倍，依此类推. 在现实应用中，可以根据需要设置不同的卷积核大小.

表3 每个粒度的卷积核大小及数据矩阵大小

	$k=1$	$k=2$	$k=3$	$k=4$
(n^k_i, I^k)	(5,72)	(10,67)	(20,57)	(40,27)
(n^k_j, J^k)	(3,41)	(6,38)	(12,32)	(24,20)
(n^k_t, T^k)	(2,11)	(4,9)	(6,7)	(8,5)

6.2 评价指标

对于收集到的数据D，本文选用以下几个指标对实验结果进行评估.

离散系数是描述数据离散程度的指标，离散系数越小，说明数据值离平均值越近[50-51]，数据值之间差距也就越小，数据值的分布就越均匀. 定义第k粒度离散系数$CV(D^k, W^k)$为第k粒度单位权重数据矩阵$D_w{}^k$的标准差除以平均值[50-51]，即：

$$CV\left(D^k, W^k\right)=\frac{STD\left(D^k, W^k\right)}{\overline{D^k_w}} \tag{12}$$

其中，$\overline{D^k_w}$为单位权重数据数量$D^k_w(i,j,t)$的均值(公式10)，而$STD(D^k, W^k)$为$D_w{}^k$的标准差，即

$$STD\left(D^k, W^k\right)=\sqrt{\frac{\sum\limits_{i,j,t,\,W^k(i,j,t)\neq 0}\left[D^k_w(i,j,t)-\overline{D^k_w}\right]^2}{\sum\limits_{i,j,t,\,W^k(i,j,t)\neq 0}1}} \tag{13}$$

定义层次离散系数$CV(D)$为k_{max}个粒度离散系数的平均值，即

$$CV(D)=\frac{1}{k_{max}}\sum_{k=1}^{k_{max}}CV\left(D^k, W^k\right) \tag{14}$$

覆盖率是常用的数据质量指标之一，其评估的是收集到数据的位置数量占所有需要收集数据的位置数量的比例[9,14-15]. 覆盖率越高，数据质量越高[9,14-15]. 定义第k粒度覆盖率$C(D^k, W^k)$为第k粒度数据矩阵D^k中有数据的比例，即

$$C\left(D^k, W^k\right)=\frac{\sum\limits_{i,j,t,\,W^k(i,j,t)\neq 0}\left|D^k(i,j,t)\geqslant 1\right|}{\sum\limits_{i,j,t,\,W^k(i,j,t)\neq 0}1} \tag{15}$$

其中，如果$D^k(i,j,t)$中有数据，则$\left|D^k(i,j,t)\geqslant 1\right|=1$；否则$\left|D^k(i,j,t)\geqslant 1\right|=0$. 定义层次覆盖率$C(D)$为$k_{max}$个粒度覆盖率的平均值，即：

$$C(D)=\frac{1}{k_{max}}\sum_{k=1}^{k_{max}}C\left(D^k, W^k\right) \tag{16}$$

根据文献[52]的观察，城市时空数据具有显著的时空相关性，基于时空相关性，可以对缺失的数据进行推断. 两个位置之间的时空距离越小，这两个位置的数据的相关性越大[52]. 因此，对于一个未收集到

数据的位置(i,j,t)，定义其离最近的收集到数据的位置的距离为：

$$d(i,j,t)=\min_{(i_0,j_0,t_0),D(i_0,j_0,t_0)>0}|i-i_0|+|j-j_0|+|t-t_0| \tag{17}$$

距离越小，推断该缺失数据的能力越大[52]. 因此，可以定义所有未收集到数据的位置距离收集到数据的位置的平均最短距离为：

$$d(D)=\frac{1}{\sum_{(i,j,t),D(i,j,t)=0}1}\sum_{(i,j,t),D(i,j,t)=0}d(i,j,t) \tag{18}$$

需要注意的是，当所有位置都收集到数据时，可以定义$d(D)=0$，以表明此时平均最短距离最小. 然而，在本文的方法中，收集到的数据量是远小于未收集到的数据量，一般不会出现所有位置都收集到数据的情况.

数据的数量$Q(D)$定义为收集到的数据总数量，即：

$$Q(D)=\sum_{i,j,t}D(i,j,t) \tag{19}$$

图6给出了两个二维权重数据矩阵D_w^k的示例，用于展示评价指标的计算. 图6a是完全均匀分布的数据，其信息熵为最大值$\log_2 16$，即4. 标准差为0，因此其离散系数为0. 由于16个格子都有数据，因此覆盖率为1，平均最短距离为0，数据量为16. 同理，可以计算图6b中数据的各项指标. 可以发现，数据越均匀，信息熵越大，离散系数越小，覆盖率越高，平均最短距离越小(表4).

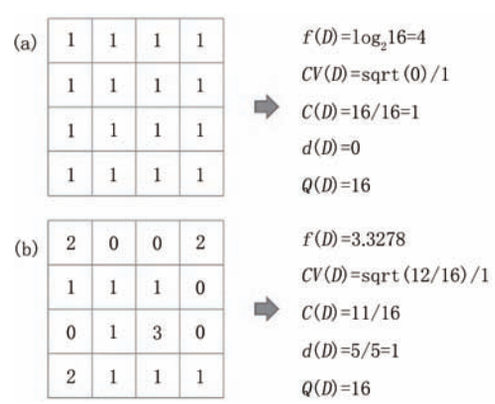

图6　各评价指标计算示例

6.3　基准方法

正如表2中所示，由于本文的方法与现有方法解决的是不同的数据收集需求，应用场景不同，因此无法与现有方法进行公平的对比实验. 为此，本文与以下三个方法进行对比，这三个方法是现有用户招募方法常用的基准方法[10,22,37].

表4　评价指标及其释义

评价指标	释义	期望
层次离散系数	$CV(D)$，见公式14	↓
层次覆盖率	$C(D)$，见公式16	↑
平均最短距离	$d(D)$，见公式18	↓
数量	$Q(D)$，见公式19	↑
第k粒度离散系数	$CV(D^k,W^k)$，见公式12	↓
第k粒度覆盖率	$C(D^k,W^k)$，见公式15	↑

(1) 随机招募策略[10]：随机招募策略是文献[10]中的一个对比方法，本文将其推广到动态用户招募中，即对于每个到来的用户，该方法都以50%的概率招募该用户.

(2) 贪心策略[37,22]：与文献[37]和[22]中的贪心策略类似，本文的贪心策略招募每个实时到来的用户，直到所有经费用完.

(3) 改进的随机招募策略[10]：在随机招募策略中，招募概率是固定的50%，这不够合理，招募概率应该随着招募过程而进行动态调整. 为此，本文对其进行改进，提出改进的随机招募策略. 首先计算剩余经费比例$B_{\text{left_ratio}}=B_{\text{left}}/B$以及剩余时间比例$time_{\text{left_ratio}}=(time_{end}-time)/(time_{end}-time_{start})$. 如果剩余经费比例$B_{\text{left_ratio}}$大于剩余时间比例$time_{\text{left_ratio}}$，则应该增大招募概率. 因此，设置改进的招募概率为$50\%\times B_{\text{left_ratio}}/time_{\text{left_ratio}}$.

6.4　与基准算法的对比

本小节将所提出的动态用户招募算法与所有基准算法进行对比实验. 设定本文招募方法的$M=11$和$N_{\text{attempt}}=1000$. 运用不同的招募方法，在两个数据集上进行10次实验并取平均值，得到表5的实验结果. 从实验结果可以发现，对于收集到的数据，本文的方法在各个指标上都是最好的. 具体而言，从离散系数来看，与最优的基准方法相比，本文的方法能够降低13.33%(数据集1)和19.64%(数据集2)的层次离散系数，各个粒度的离散系数也均有较大比例的降低. 正如6.2节介绍的，离散系数越小，收集到的数据越均匀[50-51]. 从层次覆盖率来看，同样与最优的基准方法相比，本文的方法能够提升4.65%(数据集1)和11.82%(数据集2)的层次覆盖率，第一、第二和第三粒度覆盖率都有较大提升. 覆盖率越大，说明本文的方法收集到的数据质量越高[9,14-15]. 从平均最短距离来看，本文的方法在数据集1和数据集2能够分别降低13.08%和23.10%，这大幅降低了未收集到数据的

位置与收集到数据的位置的距离，能够大幅提升推断未收集到数据的能力[52]. 这些结果都表明，与基准方法相比，本文的方法收集到的数据更加均匀，更具有代表性，更有利于未来的数据挖掘和分析.

表5 所提出的方法与基准方法的对比结果

评价指标	数据集1					数据集2				
	随机招募	贪心	改进的随机招募	前向搜索和投票	优化比例（%）	随机招募	贪心	改进的随机招募	前向搜索和投票	优化比例（%）
层次离散系数	1.365	1.384	1.305	**1.131**	13.33	1.814	3.218	1.600	**1.286**	19.64
层次覆盖率	0.620	0.635	0.645	**0.675**	4.65	0.576	0.291	0.609	**0.681**	11.82
平均最短距离	4.556	4.115	4.074	**3.541**	13.08	5.029	7.368	4.697	**3.612**	23.10
数量	170.1	**200.0**	**200.0**	**200.0**	0	200.0	200.0	200.0	**200.0**	0
第一粒度离散系数	2.202	2.314	2.074	**1.891**	8.82	2.782	4.597	2.478	**2.176**	12.19
第二粒度离散系数	1.483	1.523	1.415	**1.223**	13.57	2.062	3.824	1.774	**1.445**	18.55
第三粒度离散系数	1.208	1.153	1.171	**0.976**	15.35	1.593	2.888	1.395	**1.031**	26.09
第四粒度离散系数	0.564	0.547	0.559	**0.434**	20.66	0.820	1.563	0.753	**0.493**	34.53
第一粒度覆盖率	0.251	0.254	0.281	**0.315**	12.10	0.199	0.108	0.241	**0.280**	16.18
第二粒度覆盖率	0.522	0.534	0.557	**0.605**	8.62	0.414	0.174	0.493	**0.619**	25.56
第三粒度覆盖率	0.709	0.750	0.742	**0.781**	4.13	0.693	0.281	0.703	**0.825**	17.35
第四粒度覆盖率	**1.000**	**1.000**	**1.000**	**1.000**	0	1.000	0.600	1.000	**1.000**	0

6.5 不同M对招募方法效果的影响

在本文所提出的动态用户招募方法中，M是一个可以设定的参数，其控制着前向搜索生成未来用户的次数. 正常来说，M越大，生成未来用户次数越多，本文的招募算法的效果越好. 本小节通过实验研究M的变化对所提出招募方法效果的影响. 设定$N_{\text{attempt}}=0$，设置M分别为1、11和101，得到实验结果如图7. 如图所示，M越大，所提出的动态招募算法的效果越好.

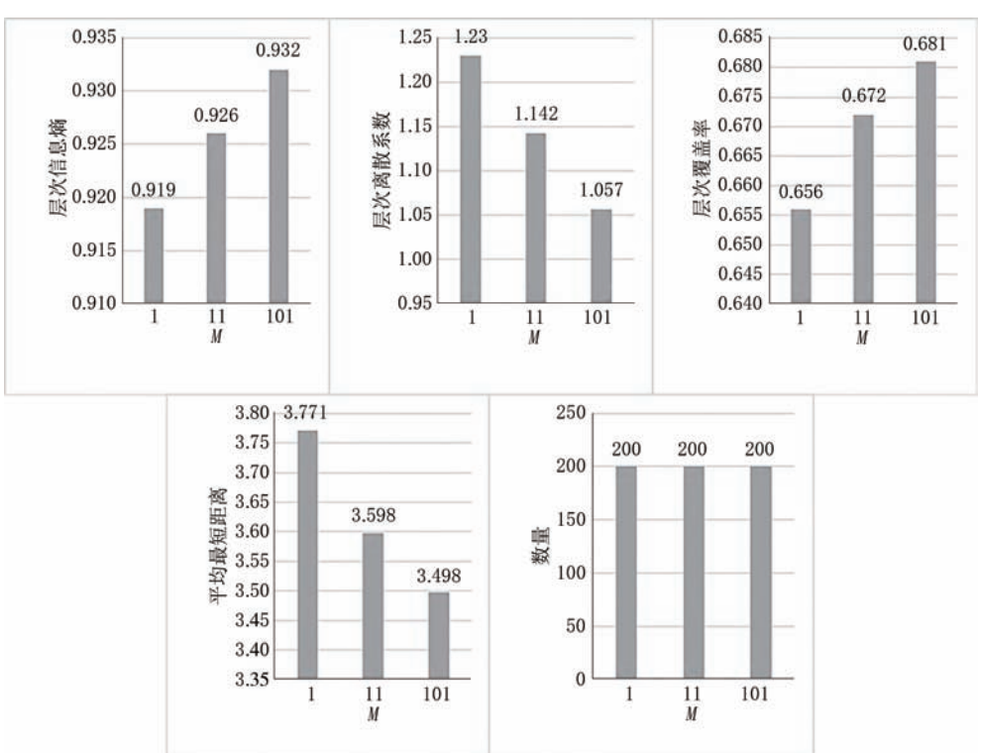

图7 M对招募方法效果的影响（$N_{\text{attempt}}=0$）

6.6 不同N_{attempt}对招募方法效果的影响

同M类似，N_{attempt}也是所提出的招募算法中的一个重要参数，其是算法2求解模型3时的一个参数. 理论上，N_{attempt}越大，所提出的爬山算法做的替换

尝试越多，模型3的解就越优，从而招募效果就越好. 因此，本小节通过实验来验证$N_{attempt}$对招募方法效果的影响. 设定$M=11$，$N_{attempt}$分别设置为100和1000，得到实验结果见图8. 从图8可见，所提出的招募效果确实随着$N_{attempt}$变大而变得更好，符合本文对参数$N_{attempt}$的设定.

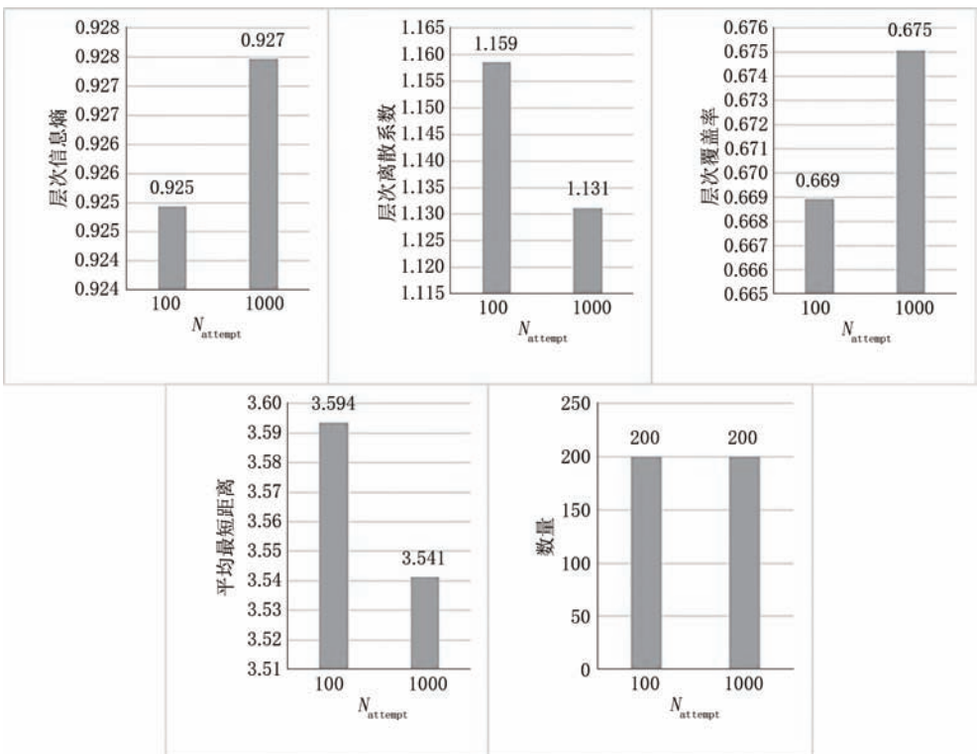

图8　$N_{attempt}$对招募方法效果的影响($M=11$)

6.7　运算时间

正如前文所述，M和$N_{attempt}$越大，招募效果越佳，但是其也可能带来越大的运算时间. 为此，需要研究不同的M和$N_{attempt}$对运算时间的影响. 本文的实验所用电脑配置如下：Intel(R) Xeon(R)的CPU；运算频率3.70GHz；内存为15G，用Python 3实现所提出的动态用户招募算法. 图9a展示的是所提出的招募算法在$N_{attempt}=0$、参数M变化时单次招募所需要的平均时间. 图9b展示的是所提出的招募算法在$M=11$、参数$N_{attempt}$变化时单次招募所需要的平均时间. 如图所示，随着M和$N_{attempt}$的增大，所提出的算法所需要的运算时间就越大. 因此，M和$N_{attempt}$的选取需要综合考虑效果和运算时间. 另外，所提出的动态用户招募算法(算法1)中的M次投票是独立，即M次投票是可以并行的，因此实际所需要的运行时间能够大幅降低.

6.8　考虑层次信息熵的必要性

本文提出了一个新的数据均匀指标来评估城市数据的质量. 为了验证考虑层次信息熵的必要性，本文与以下四种指标(作为模型3中的$f(D)$)进行比较：(1)第一粒度数据的信息熵；(2)第二粒度数据的信息熵；(3)第三粒度数据的信息熵；(4)第四粒度数据的信息熵. 由于本文提出的数据均匀指标是在论文[10]的数据均匀指标基础上进行改进和推广得到的，因此其性能与论文[10]是接近的. 但是，正如5.1和5.2节所叙述的，相比于论文[10]的指标，本文的数据均匀指标有三大改进，能够适用于更复杂的城市环境. 设定$M=11$、$N_{attempt}=1000$，用不同的数据均匀指标作为模型3中的$f(D)$，进行10次实验取平均值，得到表6中的实验结果. 从结果中可以看出，四种均匀指标都倾向于在各自粒度下取得最大的信息上，而本文的数据均匀指标能够在所有粒度都取得接近最大的信息熵，并且层次信息熵最大、层次离散系数最小、层次覆盖率最大、平均最短距离最小，因此综合效果最好.

6.9　与静态用户招募方法的对比和结合

本小节的研究包含三个部分：第一是静态用户招募存在的两个潜在不足，第二是动态用户招募与静态用户招募的对比，第三是动态用户招募与静态用户招募的结合. 由于文献[10]中的静态用户招募方

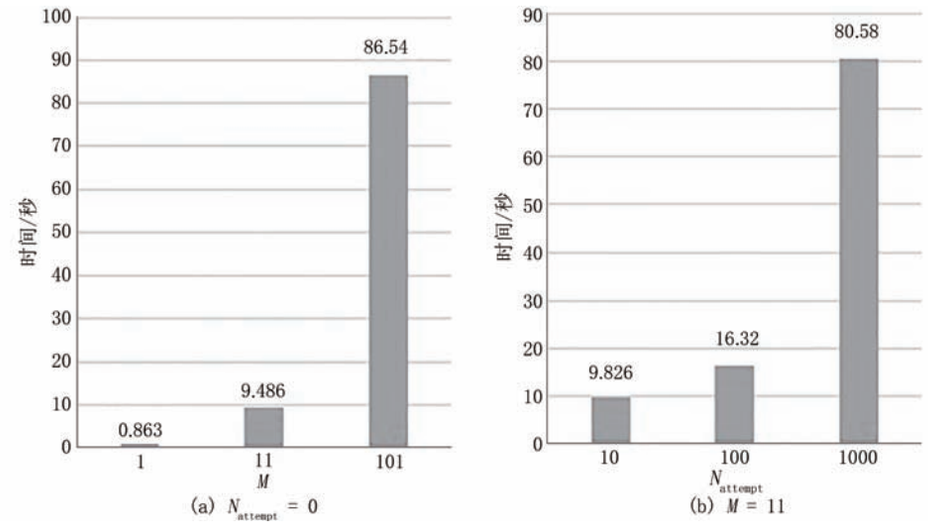

图9 不同的M和$N_{attempt}$对应的运行时间

表6 不同数据均匀指标的对比结果

评价指标	均匀指标1	均匀指标2	均匀指标3	均匀指标4	本文的指标
第一粒度信息熵	**11.382**	11.315	11.282	11.288	11.354
第二粒度信息熵	12.227	**12.238**	12.214	12.226	12.259
第三粒度信息熵	12.453	12.474	12.473	**12.513**	12.507
第四粒度信息熵	11.700	11.709	11.703	**11.737**	11.720
层次信息熵	0.926	0.925	0.924	0.926	**0.927**
层次标准差率	1.146	1.154	1.174	1.144	**1.131**
层次覆盖率	0.672	0.670	0.671	0.669	**0.675**
平均最短距离	3.583	3.602	3.602	3.598	**3.541**
数量	**200.0**	**200.0**	**200.0**	**200.0**	**200.0**

法与本文的研究设定最为相似，故与其进行对比.

6.9.1 静态用户招募存在的潜在不足

静态用户招募存在以下两个潜在不足.第一是其只对数据收集开始之前申请加入的用户进行招募，而不考虑数据收集开始后实时申请加入的用户(图1a).然而现实中总是存在着数据收集开始后实时申请加入的用户.设定参数α为在数据收集开始之前申请加入的用户数量占总数量的比例，1-α则为数据收集开始后实时申请加入的用户比例.第二是当招募的用户无法完成数据收集时，由于数据收集已经开始，静态用户招募便无法招募新用户或无法调整现有的招募用户，因此无法进行有效的调整.假定β为招募一个用户后该用户能够完成数据收集的概率.注意，α和β的取值范围均为0到1之间.直观上，当$\alpha=1$和$\beta=1$的情况，静态用户招募取得最佳的效果.这部分的实验设置分为两部分：(1)固定$\alpha=1$，研究β变化对静态用户招募结果的影响和(2)固定固定$\beta=1$，研究α变化的影响.实验结果见表7.

表7 静态用户招募在不同参数下的效果

评价指标	α=1.0 β=1.0	α=1.0 β=0.8	α=1.0 β=0.6	α=0.8 β=1.0	α=0.6 β=1.0	α=0.4 β=1.0
层次离散系数	0.948	0.998	1.088	1.054	1.312	1.392
层次覆盖率	0.692	0.672	0.637	0.675	0.638	0.605
平均最短距离	2.372	2.527	2.866	2.483	2.764	3.130
数量	200.0	161.6	120.9	200.0	200.0	136.0

表7首先展示了当$\alpha=1$的情况，即所有用户都在数据收集开始之前申请加入时，β对静态用户招募[10]结果的影响.随着的β减小，收集到的数据质量不断减低，层次离散系数不断增大、层次覆盖率不断降低、平均距离不断上升以及数量不断降低.这说明当招募到的用户完成数据收集的概率β越低，静态用户招募效果越差.其原因在于当有用户无法完成数据收集时，静态用户招募无法做出及时有效的调整.

当$\beta=1$固定时，从表7可以看出，随着数据收集开始之前申请加入的用户比例α的下降，静态用户招募[10]所能收集到的数据质量也不断下降.原因是因为其能够考虑的用户数量在不断减少，能够招募的用户质量必然下降.

6.9.2 与静态用户招募存在的对比

本小节将本文的动态用户招募方法与静态用户招募方法[10]进行对比.固定$\beta=0.8$，对于静态用户招募方法[10]，考虑$\alpha=1$、0.9、…、0.5的情况；对于动态用户招募，所有用户都是实时申请加入.实验结果如图10所示.从层次离散系数来看，当α从1

降到0.7时，本文的方法就超过了静态用户招募的方法；从层次覆盖率和平均最短距离来看，本文的方法与$\alpha=1$时的静态用户招募方法取得了非常接近的结果；从数量的角度来看，即使招募到的用户中有20%($\beta=0.8$)的用户未能完成数据收集，本文的方法也可能进行动态的调整，从而收集到200个数据(总经费为200)．而静态用户招募方法[10]则无法进行动态调整，总有大约20%的经费闲置，未能收集到数据．

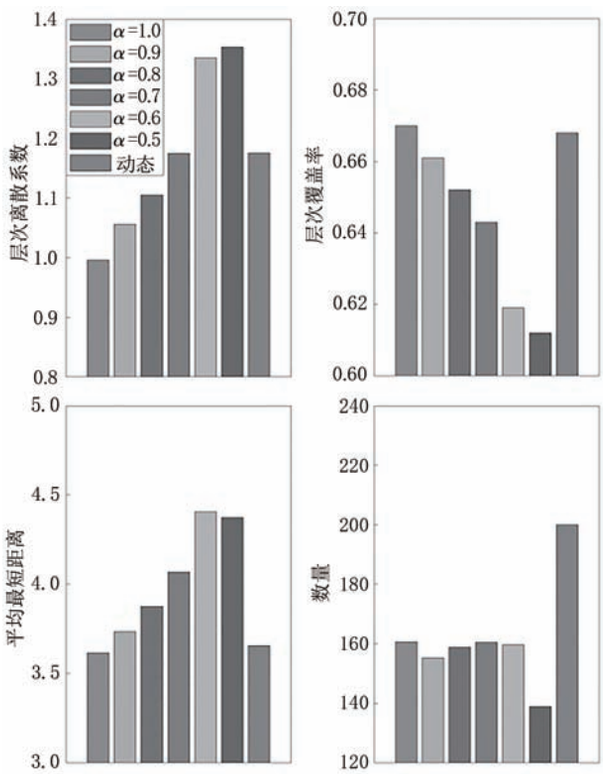

图10　动态用户招募与静态用户招募的对比

6.9.3　与静态用户招募的结合

本小节将静态用户招募方法[10]与本文所提出的动态用户招募方法进行结合，这也是探讨进行动态用户招募的有效性和必要性的一种方式．具体而言，结合后的方法对在数据收集开始之前申请加入的用户进行静态用户招募，对数据收集开始之后实时申请的用户进行动态用户招募．设定三组数据：$\alpha=0.75$，$\beta=0.8$、$\alpha=0.5$，$\beta=0.8$和$\alpha=0.25$，$\beta=0.8$，得到表8中的实验结果．从实验结果可以看出，相比于静态用户招募，静态和动态结合后，各项指标都得到较大幅度优化．因此，将动态用户招募与静态用户招募相结合确实能够提升整体的数据收集质量，这也验证了进行动态用户招募的有效性和必要性．

表8　静态用户招募与动态用户招募的结合

评价指标	α=0.75 β=0.8		α=0.5 β=0.8		α=0.25 β=0.8	
	静态	静-动	静态	静-动	静态	静-动
层次离散系数	1.129	1.109	1.354	1.225	1.599	1.313
层次覆盖率	0.647	0.663	0.612	0.657	0.536	0.645
平均最短距离	2.775	2.581	3.029	2.611	4.223	2.744
数量	159.9	185.6	138.9	197.3	68.2	199.1

7　讨　论

7.1　用户隐私

在本文的动态用户招募中，所需要用户提供的隐私数据是用户能够进行数据收集的地点和时间，即(lat,lon,$time$)．这是一个单点数据，并不需要用户的GPS轨迹数据，因此无法推断出用户的其他任何敏感数据．除此之外，也无需记录用户的任何身份信息．由此可见，本文的方法对用户隐私的泄露程度是较低的．

7.2　差异化的数据质量

在真实应用场景中，招募到的用户所能收集到的数据质量可能是不同的．用户设备越先进、越智能，其所能收集的数据质量可能就越高．基于该情况，可以将本文现有的模型进行推广．对于一个用户r，定义其所能收集到的数据的质量为z，z属于[0,1]，即$z=1$代表数据质量最高，$z=0$代表最低．由于本文的模型考虑的是收集到的数据数量矩阵，因此需要将数据质量z映射到数据数量，即以数量代表质量，通常质量越大，数量也越大，假设该映射为$g(z)$：

$$g(z):[0,1]\rightarrow[0,1] \tag{20}$$

接着，公式1中的用户r能够收集到的数据矩阵D_r也应修改为

$$D_r(i,j,t)=\begin{cases}g(r.z), \ if\ (i,j,t)=(lat_{id},lon_{id},time_{id})\\0 \quad , \ otherwise\end{cases} \tag{21}$$

其中，$r.z$为用户r所能收集到的数据质量．特别地，当$g(z)=z$时，有

$$D_r(i,j,t)=\begin{cases}r.z, \ if\ (i,j,t)=(lat_{id},lon_{id},time_{id})\\0, \ otherwise\end{cases} \tag{22}$$

当所招募到的用户收集到的数据数量$g(r.z)$不为1时，其所能获得的奖励存在两种可能：(1)仍旧为1个单位的奖励；(2)改为获得$g(r.z)$个单位的奖励．具体的设置方式可以根据真实应用场景进行修

改.当改为获得$g(r.z)$个单位的奖励时,公式2中的B_{left}需改为

$$B_{\text{left}}=B-\sum_{r\in R_{\text{arrived}}}I(r)g(r.z) \tag{23}$$

7.3 不规则的区域划分

本文所提出的基于层次信息熵的均匀程度指标不仅可以应用到栅格化的区域划分中,其也可以应用到不规则区域中.如图11所示,对于不规则的区域,依旧可以考虑不同的粒度,计算不同粒度下的不规则区域的数据数量,从而可以计算不同粒度下的数据的信息熵.因此,本文的方法也可以在不规则区域中运用层次信息熵作为数据均匀程度指标.

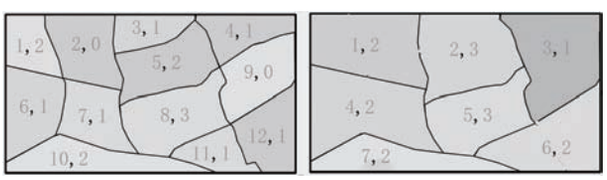

图11 不同粒度下的不规则区域;左:细粒度;右:粗粒度(图中的两个数字分别是区域编号和收集到的数据数量).

7.4 所提出方法的实用性

本文所提出方法的实用性具体体现在以下三点.(1)在现实的移动群智感知应用场景中,存在这样一类典型问题:用于数据收集的经费有限、数据收集的空间范围大且时间范围长(需要收集的数据量大)、数据之间具有显著的时间和空间相关性.例如,有限经费下城市空气质量数据的收集、有限经费下城市人流量数据的收集等.本文所述的方法能够很好地应用于这类场景,其能够基于有限的经费(例如只够收集200个数据),进行合理的用户招募,使得收集到的数据的均匀程度最高.数据的均匀程度越高,收集到的数据越具有代表性,覆盖率越高,并且基于数据的时间和空间相关性,越能准确地推断出未收集到的数据.详见3.1节的讨论以及6.2节和6.3节的实验.

(2)针对本文提出的新的能够综合考虑四种因素的高效动态用户招募方法,正如第1节引言中所介绍的,对动态用户招募方法的研究才刚刚起步,因此本文的方法能够为未来的动态用户招募方法研究起到较好的借鉴作用.

(3)针对本文提出的基于层次信息熵的均匀程度指标,虽然其在模型的推导过程中主要考虑栅格化区域(方块),但是正如7.3节所述,其也能够很好地推广到不规则区域中(路网和街区).

8 结论和未来工作

本文提出了一个新的移动群智感知动态招募框架,该框架包含了一个新的城市数据均匀程度指标以及一个基于前向搜索和投票的动态招募算法.基于所提出的方法,本文能够收集到更均匀的数据,从而提升移动群智感知的质量.具体而言,基于美国纽约市曼哈顿区域的Foursquare用户数据和地图数据的实验表明,相比于基准方法,本文的方法收集到的数据显著地更加均匀.未来的研究方向包括两个方面:一是将本文的方法推广到既有动态招募又有静态招募的模式中,即动态招募和静态招募的结合;二是将本文的框架运用到不同的数据质量定义的问题中.

参 考 文 献

[1] Ganti R K, Ye F, Lei H. Mobile crowdsensing: current state and future challenges. IEEE Communications Magazine, 2011, 49(11): 32-39

[2] Guo B, Yu Z, Zhou X, Zhang D. From participatory sensing to mobile crowd sensing. Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communication Workshops. Zurich, Switzerland, 2014: 593-598

[3] Ma H, Zhao D, Yuan P. Opportunities in mobile crowd sensing. IEEE Communications Magazine, 2014, 52(8): 29-35

[4] Yu Z W, Yu Z Y, Zhou X S. Socially Aware Computing. Chinese Journal of Computers, 2012, 35(01): 16-26 (in Chinese)
(於志文,於志勇,周兴社. 社会感知计算: 概念、问题及其研究进展. 计算机学报, 2012, 35(01): 16-26)

[5] Zheng Y, Capra L, Wolfson O, Yang H. Urban computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology, 2014, 5(3), 1-55

[6] Zheng Y. Introduction to urban computing. Geomatics and Information Science of Wuhan University, 2015, 40(1): 1-13 (in Chinese)
(郑宇. 城市计算概述. 武汉大学学报, 2015, 40(1): 1-13)

[7] Zheng Y, Liu Y, Yuan J, Xie X. Urban computing with taxicabs. Proceedings of the 13th International Conference on Ubiquitous Computing. Beijing, China, 2011: 89-98

[8] Chen L, Zhang D, Pan G, Ma X, Yang D, Kushlev K, Zhang W S, Li, S. Bike sharing station placement leveraging heterogeneous urban open data. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Osaka, Japan, 2015: 571-575

[9] Zhang D, Xiong H, Wang L, Chen G. CrowdRecruiter: selecting participants for piggyback crowdsensing under probabilistic coverage constraint. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.

Seattle，USA，2014：703-714

[10] Ji S，Zheng Y，Li T. Urban sensing based on human mobility. Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Heidelberg，Germany，2016：1040-1051

[11] An J，Peng Z L，Gui X L，Xiang L L，Liang D W. Chinese Journal of Computers，2019，42(02)：65-78(in Chinese)
(安健，彭振龙，桂小林，向乐乐，梁丹薇. 群智感知中基于公交系统的任务分发机制研究. 计算机学报，2019，42(02)：65-78)

[12] Wang J，Wang L，Wang Y，Zhang D，Kong L. Task allocation in mobile crowd sensing：State-of-the-art and future opportunities. IEEE Internet of Things Journal，2018，5(5)：3747-3757

[13] Xiong H，Zhang D，Chen G，Wang L，Gauthier V，Barnes L E. iCrowd：Near-optimal task allocation for piggyback crowdsensing. IEEE Transactions on Mobile Computing，2015，15(8)：2010-2022

[14] Ahmed A，Yasumoto K，Yamauchi Y，Ito M. Distance and time based node selection for probabilistic coverage in people-centric sensing. Proceedings of the 2011 Annual IEEE Communications Society Conference on Sensor，Mesh and Ad Hoc Communications and Networks. Salt Lake City，USA，2011：134-142

[15] Hachem S，Pathak A，Issarny V. Probabilistic registration for large-scale mobile participatory sensing. Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications. San Diego，USA，2013：132-140

[16] Reddy S，Estrin D，Srivastava M. Recruitment framework for participatory sensing data collections. Proceedings of the 2010 International Conference on Pervasive Computing. Helsinki，Finland，2010：138-155

[17] Wang L，Zhang D，Pathak A，Chen C，Xiong H，Yang D，Wang Y. CCS-TA：Quality-guaranteed online task allocation in compressive crowdsensing. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Osaka，Japan，2015：683-694

[18] Roughan M，Zhang Y，Willinger W，Qiu L. Spatio-temporal compressive sensing and internet traffic matrices (extended version). IEEE/ACM Transactions on Networking，2011，20(3)：662-676

[19] Wang J，Wang Y，Zhang D，Helal S. Energy saving techniques in mobile crowd sensing：Current state and future opportunities. IEEE Communications Magazine，2018，56(5)：164-169.

[20] Lane N D，Chon Y，Zhou L，Zhang Y，Li F，Kimz D，Ding G Z，Zhao F，Cha H. Piggyback crowdsensing (pcs) energy efficient crowdsourcing of mobile sensor data by exploiting smartphone app opportunities. Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems. Roma，Italy，2013：1-14

[21] Liu C H，Zhang B，Su X，Ma J，Wang W，Leung K K. Energy-aware participant selection for smartphone-enabled mobile crowd sensing. IEEE Systems Journal，2015，11(3)：1435-1446

[22] Xiong H，Zhang D，Wang L，Chaouchi H. EMC^3：Energy-efficient data transfer in mobile crowdsensing under full coverage constraint. IEEE Transactions on Mobile Computing，2014，14(7)：1355-1368

[23] Wang L，Yang D，Han X，Wang T，Zhang D，Ma X. Location privacy-preserving task allocation for mobile crowdsensing with differential geo-obfuscation. Proceedings of the 26th International Conference on World Wide Web. Perth，Australia，2017：627-636

[24] Jin H，Su L，Xiao H，Nahrstedt K. Inception：Incentivizing privacy-preserving data aggregation for mobile crowd sensing systems. Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing. Paderborn，Germany，2016：341-350

[25] Ni J，Zhang A，Lin X，Shen X S. Security，privacy，and fairness in fog-based vehicular crowdsensing. IEEE Communications Magazine，2017，55(6)：146-152

[26] Wu Y，Zeng JR，Peng H，Chen H，Li C P. Survey on incentive mechanisms for crowd sensing. Ruan Jian Xue Bao/Journal of Software，2016，27(8)：2025-2047 (in Chinese)
(吴垚，曾菊儒，彭辉，陈红，李翠平. 群智感知激励机制研究综述. 软件学报，2016，27(08)：2025-2047)

[27] Zhang X，Yang Z，Sun W，Liu Y，Tang S，Xing K，Mao X. Incentives for mobile crowd sensing：A survey. IEEE Communications Surveys & Tutorials，2015，18(1)：54-67

[28] Xiong H，Zhang D，Guo Z，Chen G，Barnes L E. Near-optimal incentive allocation for piggyback crowdsensing. IEEE Communications Magazine，2017，55(6)：120-125

[29] Nan W Q，Guo B，Chen H H，Yu Z W，Wu W L. A Cross-Space，Multi-Interaction-Based Dynamic Incentive Mechanism for Mobile Crowd Sensing. 2015，38(12)：2412-2425 (in Chinese)
(南文倩，郭斌，陈荟慧，於志文，吴文乐，周兴社. 基于跨空间多元交互的群智感知动态激励模型. 计算机学报，2015，38(12)：2412-2425)

[30] Guo B，Chen H，Yu Z，Nan W，Xie X，Zhang D，Zhou X. Taskme：Toward a dynamic and quality-enhanced incentive mechanism for mobile crowd sensing. International Journal of Human-Computer Studies，2017，102：14-26

[31] Chon Y，Lane N D，Kim Y，Zhao F，Cha H. Understanding the coverage and scalability of place-centric crowdsensing. Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing. Zurich，Switzerland，2013：3-12

[32] Kawajiri R，Shimosaka M，Kashima H. Steered crowdsensing：Incentive design towards quality-oriented place-centric crowdsensing. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Seattle，USA，2014：691-701

[33] Lee J S，Hoh B. Dynamic pricing incentive for participatory sensing. Pervasive and Mobile Computing，2010，6(6)：693-708

[34] Zhang J T，Zhao Z H，Zhou S W. Vector Task Map：Progressive Task Allocation in Crowd-Sensing. Chinese Journal of Computers，2017，40(8)：1946-1960 (in Chinese)
(张君涛，赵智慧，周四望. 矢量任务地图：群智感知任务渐进式分发方法. 计算机学报，2017，40(08)：1946-1960)

[35] Li Z，Xu Z，Chen X，Li S Q. Location-related Online Multi-task Assignment Algorithm for Mobile Crowd Sensing，2019，46

(06): 102-106 (in Chinese)
(李卓,徐哲,陈昕,李淑琴. 面向移动群智感知的位置相关在线多任务分配算法. 计算机科学,2019,46(06): 102-106)
[36] Wang J, Wang F, Wang Y, Wang L, Qiu Z, Zhang D, Guo B, Lv, Q. HyTasker: Hybrid task allocation in mobile crowd sensing. IEEE Transactions on Mobile Computing, 2020, 19 (3): 598-611
[37] Xiong H, Zhang D, Wang L, Gibson J P, Zhu J. EEMC: Enabling energy-efficient mobile crowdsensing with anonymous participants. ACM Transactions on Intelligent Systems and Technology, 2015, 6(3): 1-26
[38] Liu Y, Guo B, Wu W L, Yu Z W, Zhang D Q. Multitask-Oriented Participant Selection in Mobile Crowd Sensing. Chinese Journal of Computers, 2017, 40 (8): 1872-1887 (in Chinese)
(刘琰,郭斌,吴文乐,於志文,张大庆. 移动群智感知多任务参与者优选方法研究,计算机学报,2017,40(8): 1872-1887)
[39] Wang J, Wang Y, Zhang D, Wang F, He Y, Ma L. PSAllocator: Multi-task allocation for participatory sensing with sensing capability constraints. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. Portland, USA, 2017: 1139-1151
[40] Silver D, Schrittwieser J, et al. Mastering the game of go without human knowledge. Nature, 2017, 550(7676): 354-359
[41] Silver D, Huang A, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529 (7587): 484-489
[42] Hochba D S. Approximation algorithms for NP-hard problems. ACM SIGACT News, 1997, 28(2): 40-52
[43] Xi B, Liu Z, Raghavachari M, Xia C H, Zhang L. A smart hill-climbing algorithm for application server configuration. Proceedings of the 13th International Conference on World Wide Web. New York, USA, 2004: 287-296
[44] Mitchell M, Holland J H, Forrest S. When will a genetic algorithm outperform hill climbing. Proceedings of the 1994 Advances in Neural Information Processing Systems. Denver, USA, 1994: 51-58
[45] Durbin J, Koopman S J. Time series analysis by state space methods. England: Oxford university press, 2012
[46] Zhang J, Zheng Y, Qi D. Deep spatio-temporal residual networks for citywide crowd flows prediction. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 1655 - 1661
[47] Wang W, Yin H, Chen L, Sun Y, Sadiq S, Zhou X. Geo-SAGE: A geographical sparse additive generative model for spatial item recommendation. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 1255-1264
[48] Mood A M. Introduction to the Theory of Statistics. 1950
[49] Feller W. An introduction to probability theory and its applications. John Wiley & Sons, 2008.
[50] Yang D, Zhang D, Zheng V W, Yu Z. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014, 45(1): 129-142
[51] Open Street Map, https://www.openstreetmap.org/ 2020.2.11.
[52] Yi X, Zheng Y, Zhang J, Li T. ST-MVL: filling missing values in geo-sensory time series data. Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, USA, 2016: 9-15

JI Sheng-Gong, Ph. D. His research interests include urban resource optimization and social networks.

ZHENG Yu, Ph. D., professor. His research interests include artificial intelligence and urban computing.

WANG Zhao-Yuan, Ph. D. His research interests include cross domain data fusion and data mining.

LI Tian-Rui, Ph. D., professor. His research interests include artificial intelligence and big data.

Background

This paper is supported by the National Key R&D Program of China (Grant No. 2019YFB2101802).

Advance in mobile sensing technologies and devices is making mobile crowd sensing one of important modes to collect urban big data. Mobile crowd sensing has been widely used to collect many types of urban data, such as noise, traffic flow, air quality, temperature, and so on. Participant recruitment plays an essential role in mobile crowd sensing, which aims to recruit proper participants so as to collect urban data with the higher quality.

Although there have already been a number of works on participant recruitment, most of them focus on static participant recruitment, instead of dynamic participant recruitment. Static participant recruitment methods recruit participants before the data collection, while dynamic participant recruitment methods recruit participants during the data collection. Specifically, dynamic participant recruitment real time recruits coming participants according to the data already collected, the budget left, the data that the coming participants can collect, and the potential participants in the future. Obviously, comparing with

the static participant recruitment, the dynamic participant recruitment is more challenging and is more effective, too.

Data quality indicators are of great importance to mobile crowd sensing, since collecting high-quality data is the core. Among different data quality indicators, data balance is an effective one to measure the urban data quality. However, existing data balance indicators cannot work well in complex urban scenarios. Thus, to provide a new data balance indicator is necessary.

To deal with the above two issues, in this paper, we firstly propose a novel and effective dynamic participant recruitment method. Through the look-ahead search and voting strategy, the proposed recruitment method can well combine the four factors into the real-time recruitment of each coming participant. Specifically, we use the look-ahead search to estimate the participants in the future. Then, we propose to optimize the recruitment decisions for both the current participant and the estimated future participants. Next, we provide a hill-climbing algorithm to solve the mathematical optimization problem. Finally, we can get the recruitment decision for the current participant using the voting strategy.

Secondly, we propose a new data balance indicator to measure the quality of the collected urban data. The proposed indicator leverages the hierarchical entropy to measure the urban data quality, which can better indicate the true balance of the data collected. Thus, our data balance indicator can better guide the dynamic recruitment of the participants. Besides, the proposed indicator can better adapt to the complex scenarios in a city.