



面向全局优化的时空众包任务分配算法

聂茜婵¹, 张 阳¹, 余敦辉^{1,2*}, 张兴盛¹

(1. 湖北大学 计算机与信息工程学院, 武汉 430062; 2. 湖北省教育信息化工程技术研究中心(湖北大学), 武汉 430062)

(* 通信作者电子邮箱 yumhy@hubu.edu.cn)

摘 要:针对时空众包任务分配研究中未考虑多方参与对象的效益和连续任务分配的全局优化, 导致分配效果不佳的问题, 提出一种面向三方综合效益全局优化的在线任务分配算法。首先, 基于在线随机森林和门控循环单元网络预测出下一时间戳内众包对象(众包任务和工人)的分布情况, 进而结合当前时间戳内众包对象的情况构造二分图模型, 最后采用带权二分图最优匹配算法完成任务分配。实验结果证明了所提算法在连续任务分配过程中实现了综合效益的全局优化。与贪心算法对比, 该算法在任务分配成功率方面提升 25.7%, 在平均综合效益方面提升 32.2%, 在工人平均机会成本方面提升 37.8%; 与随机阈值算法对比, 该算法在任务分配成功率方面提升 27.4%, 在平均综合效益方面提升 34.7%, 在工人平均机会成本方面 40.2%。

关键词: 时空众包; 预测分析; 在线随机森林; KM 算法

文献标志码: A

Spatial crowdsourcing task allocation algorithm for global optimization

NIE Xichan¹, ZHANG Yang¹, YU Dunhui^{1,2*}, ZHANG Xingsheng¹

(1. School of Computer Science and Information Engineering, Hubei University, Wuhan Hubei 430062, China;

2. Hubei Provincial Engineering Technology Research Center for Education Informatization (Hubei University), Wuhan Hubei 430062, China)

Abstract: Concerning the problem that in the research of spatial crowdsourcing task allocation, the benefits of multiple participants and the global optimization of continuous task allocation are not considered, which leads to the problem of poor allocation effect, an online task allocation algorithm was proposed for the global optimization of tripartite comprehensive benefit. Firstly, the distribution of crowdsourcing objects (crowdsourcing tasks and workers) in the next time stamp was predicted based on online random forest and gated recurrent unit network. Then, a bipartite graph model was constructed based on the situation of crowdsourcing objects in the current time stamp. Finally, the optimal matching algorithm of weighted bipartite graph was used to complete the task allocation. The experimental results show that the proposed algorithm realize the global optimization of continuous task allocation. Compared with greedy algorithm, this algorithm improves the success rate of task allocation by 25.7%, the average comprehensive benefit by 32.2% and the average opportunity cost of workers by 37.8%; compared with random threshold algorithm, the algorithm improves the success rate of task allocation by 27.4%, the average comprehensive benefit by 34.7% and the average opportunity cost of workers by 40.2%.

Key words: spatial crowdsourcing; predictive analysis; online random forest; KM (Kuhn-Munkres) algorithm

0 引言

随着 Web2.0 技术的兴起, 大量在线 Web 应用催生了众包^[1-3]这种通过群体智慧求解问题的新兴商业生产模式。众包是指“一种把过去由专职员工执行的工作任务通过公开的 Web 平台, 以自愿的形式外包给非特定的解决方案提供者群体来完成的分布式问题求解模式”^[4]。而随着智能移动设备的普及和共享经济模式的迅猛发展, 赋予了众包任务更多的时空属性, 衍生出一种全新的众包类型——时空众包^[5-7]。时空众包是指在时空约束的条件下, 将具有时空特性的众包任务分配给非特定的众包工人, 并要求众包工人以主动或被动的方式来完成众包任务。近年来, 全国流行的各类实时专车类服务平台, 例如滴滴出行、Uber 等, 均采用时空众包方

式提供服务。而在很多线上到线下 (Online-to-Offline, O2O) 应用、灾情监控和物流管理等领域, 也将时空众包技术运用其中以提高其服务质量。然而, 现有的研究大多局限从众包平台或工人的单一角度出发进行优化, 且没有满足实际应用中连续分配任务的需求。因此, 本文提出一种基于预测分析的全局优化在线任务分配算法 (Global Optimization online Mission Assignment algorithm based on predictive analysis, GOMA), 该算法基于在线随机森林和门控循环单元网络预测出下一时间戳内众包对象 (众包任务和工人) 的分布情况, 进而结合当前时间戳内众包对象的情况构造二分图模型, 最后采用带权二分图最优匹配算法完成任务分配, 从而实现众包平台、众包任务、众包工人三方综合效益的全局优化。

收稿日期: 2019-11-28; 修回日期: 2020-01-07; 录用日期: 2020-01-10。

基金项目: 国家重点研发计划项目 (2017YFB1400602); 国家自然科学基金资助项目 (61572371, 61832014)。

作者简介: 聂茜婵 (1999—), 女, 湖北武汉人, 主要研究方向: 时空众包、知识图谱; 张阳 (1999—), 男, 湖北恩施人, 主要研究方向: 数据挖掘; 余敦辉 (1974—), 男, 湖北武汉人, 教授, 博士, CCF 会员, 主要研究方向: 服务计算、大数据; 张兴盛 (1998—), 男, 湖北襄阳人, 主要研究方向: 时空众包。



本文主要工作在于:

1) 提出一种综合考虑众包平台、众包任务、众包工人三方效益的在线任务分配算法思想,进行多目标优化,更加切合实际应用中的需要。

2) 在任务分配中,考虑众包工人动态移动的特性,对任务进行连续动态分配。

1 相关研究

任务分配^[8-10]作为时空众包领域研究的核心问题之一^[11],众多学者对其展开了深入的讨论。

文献[12]在随机阈值算法的基础上,提出了面向三类众包对象的自适应阈值算法,验证了算法在提高分配效用方面的有效性;文献[13]以贪心算法作为基线算法,提出了一种基于两阶段框架模型的全球微任务分配算法,在确保算法执行效率的同时提高任务分配的效用;文献[14]以最大化任务的成功分配数为优化目标,提出了一种基于多臂赌博机的在线任务分配算法,为优化任务分配效用提供了新思路;文献[15]在众包工人查询算法的基础上,提出了一种基于动态效用的阈值选择算法,通过动态效用对比以提升分配效用;文献[16]提出一种基于众包对象预测的在线任务分配算法,在现有众包对象的基础上,采用线性回归模型预测动态出现的众包对象,进行任务分配,实现任务分配效用的全局优化。这类方法仅局限于众包平台的角度,以任务分配总效用最大化为目标进行优化,而未考虑众包工人的差旅成本以及众包任务的等待时间。

文献[17]在贪心算法的基础上,提出时间空间预测器对众包任务进行预测,进而辅助任务分配,旨在最小化一段时间内工人的总体差旅成本;文献[18]针对最小化最大匹配距离的问题,提出了一种交换链算法,有效地解决了最差匹配情况下的匹配距离最小化的问题;文献[19]面向最小化匹配距离的问题,进行综合性对比实验,证明了贪心算法解决该问题的有效性。这类方法仅局限于众包工人的角度,以工人平均差旅成本最小化为目标进行优化,而忽略了众包任务等待时间以及任务分配的总效用,同时也没有考虑众包平台、任务发布者的利益最大化。

文献[20-21]从众包平台和工人的角度出发,文献[20]根据众包工人的密度进行众包任务的范围调节,提出一种基于统计概率进行预测分析的方法,能够在提高任务分配总效用的同时降低工人的差旅成本;文献[21]基于分而治之的思想,提出了一种基于二分法框架模型的在线任务分配算法,力求最大化任务分配数量以提高分配效用,同时最小化工人的差旅成本;文献[22]从众包平台和众包任务的角度出发,综合考虑任务分配总效用和任务等待时间,提出一种基于分配时间因子的动态阈值算法,提升分配总效用的同时降低任务的平均等待时间。这类方法从众包平台、工人、任务中某两方的角度出发进行双目标优化,但没有从众包平台、任务和工人三方角度出发,综合考虑三方效益,所以仍然存在改进空间。

综上,不难看出现有研究中存在以下几点不足:

1) 现有研究大多以单目标优化为主,从众包平台或工人单一角度出发,聚焦于提升任务分配总效用或降低工人的差旅成本,而少量研究从双方角度出发,进行双目标优化。对于从众包平台、任务和工人三方角度出发,综合考虑三方效益的任务分配算法研究尚未出现。而考虑不完善的任务分配算法无法满足实际应用中的需要。

2) 大多在线任务分配算法仅依据当前时间戳内已有众

包对象进行任务分配,对当前时间戳内的任务分配进行局部优化,而在实际应用中,任务分配是连续动态进行的,现有大多任务分配方案无法满足连续分配中全局优化的目标。而已有基于预测方法进行任务分配的算法,尚未考虑众包工人动态移动的特性,无法满足实际应用环境的需要。

本文以实时专车类时空众包应用作为背景,针对上述问题,从众包平台、众包任务、众包工人三方的角度出发,以提高众包任务分配三方的综合效益指标作为优化目标,针对连续任务分配问题,提出了一种采用随机森林和门控制循环单元(Gated Recurrent Unit, GRU)网络进行预测分析的全局优化在线任务分配算法,在当前时间戳内众包对象分布情况的基础上,结合下一时间戳内众包对象的预测情况进行任务分配,实现连续任务分配过程中众包任务分配三方综合效益全局优化的目标。

2 问题定义

定义1 众包任务 m 。 m 由任务请求者在众包平台上发布,通常被定义为如下四元组的形式 $m = \langle l_m, p_m, d_m, g_m \rangle$ 。其中: l_m 是任务 m 在二维坐标系中所处的位置,任务 m 的发布时间是 p_m ,任务 m 的截止时间是 d_m , g_m 表示完成该任务 m 的工人所获得的报酬。

定义2 众包工人 w 。 w 是 m 的完成者,通常被定义为如下六元组的形式 $w = \langle l_w, r_w, dir_w, p_w, d_w, s_w \rangle$ 。其中 l_w 是工人 w 在二维坐标系中所处的位置; r_w 是工人 w 的范围半径,工人 w 的接单范围 $Rang_w$ 表示工人 w 愿意接受的任务的范围,即以 l_w 为圆心, r_w 为半径的区域; dir_w 表示工人的移动方向 $dir_w = \{N, S, W, E\}$; 工人 w 在平台上线时间是 p_w ,工人 w 在平台下线的时间是 d_w ; s_w 表示工人在平台上所完成的历史任务的成功率。对于众包工人,开始执行一个众包任务视为工人从平台下线;执行完一个众包任务后,若继续接单则视为工人在众包平台重新上线。为简化计算,本文定义工人的行驶速度均为 $speed$,工人单位路程所花费的成本为 per 。

定义3 任务工人匹配对 mp 。 mp 定义为 $mp = \langle m, w \rangle$,表示众包任务 m 和工人 w 的组合。

定义4 分配效用 U_p 。 U_p 表示将众包任务 m 分配给众包工人 w 后所产生的效用,即匹配对 $\langle m, w \rangle$ 的效用,定义为任务回报值 g_m 与工人成功率 s_w 的乘积,即:

$$U_p = g_m \times s_w \quad (1)$$

其中: $g_m \in (0, 1)$, $s_w \in (0, 1)$, 故 $U_p \in (0, 1)$ 。

定义5 任务等待分配时间 AT_m 。 AT_m 表示众包任务 m 在平台发布后,等待分配的时间,用于衡量任务的等待分配程度,等待程度高的任务优先分配。 AT_m 即为任务发布时刻 P_m 到任务分配时刻 T_a 的时间差:

$$AT_m = T_a - p_m \quad (2)$$

定义6 任务差旅时间 MT_m 。 MT_m 表示任务工人分配后,工人到达任务地点的时间,即为任务分配时刻到工人到达时刻的时间差。当工人速度一定时,任务与工人距离小的优先分配。则 MT_m 表示如下:

$$MT_m = MD_{mw} / speed \quad (3)$$

其中: MD_{mw} 表示任务和工人间的曼哈顿距离。

定义7 工人机会成本 CO_w 。工人的机会成本 CO_w 表示工人 w 在平台上线后处于空闲状态到工人被分配处于工作状态所花费的成本,当前已花费机会成本高的工人优先分配。



CO_w 表示为:

$$CO_w = L_f \times per \quad (4)$$

其中 L_f 表示工人处于空闲状态所行驶的总距离,即工人从平台上线到下线时间内所行驶的距离(曼哈顿距离):

$$L_f = (d_w - p_w) \times speed \quad (5)$$

定义 8 任务工人匹配对综合效益 CB (Comprehensive Benefits)。 mp 具有分配效用 U_p 、任务等待分配时间 AT_m 、任务差旅时间 MT_m 、工人机会成本 CO_w 四项属性,从众包三方角度考虑,定义综合效益指标 CB 用于综合衡量 U_p 、 AT_m 、 MT_m 、 CO_w 四项指标:

$$CB = \alpha U_p + \beta AT'_m + \frac{\gamma}{MT'_m} + \eta CO'_w \quad (6)$$

其中:权重系数 α 、 β 、 γ 、 η 根据熵权法确定, AT'_m 、 MT'_m 和 CO'_w 是离差法标准化处理后的取值。标准化处理方法如下:

$$x'_i = \frac{x_i - \min(X_i)}{\max(X_i) - \min(X_i)} \quad (7)$$

其中: x_i 为该指标当前样本的原值, $\max(X_i)$ 为该指标的最大值, $\min(X_i)$ 为该指标的最小值, x'_i 为该指标当前样本处理后的取值。

定义 9 综合效益最大化的众包在线任务分配问题 Crowdsourcing online Mission Assignment to Maximize the Comprehensive Benefits, CMA-MCB)。在时空众包环境下,给定众包任务集合 M 、众包工人集合 W 和任务工人匹配对综合效益 CB 的计算函数,寻求一个任务分配的结果集 R 使得三方综合效益最大化,即最大化任务工人匹配对综合效益 CB :

$$Maxsum(R) = \sum_{mp \in R} CB_{mp} \quad (8)$$

分配结果集 R 由匹配对 mp 组成,其中每个匹配对 mp 需满足以下基本约束:

时间约束 只有众包任务 m 和众包工人 w 同时平台在线时,才能实现分配,且众包任务 m 必须在任务的截止时间 d_m 前分配,否则无法完成分配。

不变性约束 众包任务一旦完成分配,分配结果则不能改变。

空间约束 众包任务 m 的空间位置 l_m 必须在分配的众包工人 w 的接单范围内,即 $|l_m - l_w| < r_w$ 。

3 GOMA

针对提出的 CMA-MCB 问题,首先将一个任务分配周期划分为 n 个固定时间长短的时间戳,在每个时间戳内进行任务分配。每轮分配首先设置当前待分配工人的接单范围,进而执行基于预测分析的全局优化在线任务分配算法(GOMA),引入下一时间戳内众包对象的分布情况作为当前分配的依据,基于在线随机森林和 GRU 两种预测模型,预测出下一时间戳内众包对象分布,结合当前时间戳内众包对象的分布情况,执行带权二分图最优匹配算法(KM 算法),完成本轮任务分配。

GOMA 主要可分为以下三步:

1) 执行基于在线随机森林的众包对象动态预测算法(Dynamic Prediction algorithm for crowdsourcing objects based on online Random Forest, DPRF)和基于在线随机森林回归预测模型,预测下一时间戳内众包对象动态出现的情况。

2) 执行基于 GRU 的工人移动轨迹预测算法(Worker Movement Trajectory Prediction algorithm based GRU, WMTF)和基于 GRU 循环神经网络预测出当前已有众包工人在下一

时间戳时的空间分布。

3) 在当前时间戳的众包对象分布的基础上,结合 1)、2) 预测的众包对象的分布情况,执行带权二分图最优匹配算法,完成任务分配。

3.1 DPRF

对于下一时间戳内新出现众包对象的预测,DPRF 将时空众包地理场景拟合成 $n \times n$ 的网格图,基于分而治之的思想,分别对网格图中每一个小室 $cell$ 进行预测。每个小室 $cell$ 预测可分为众包对象的空间分布预测和众包对象的属性参数预测。

对于空间分布预测,DPRF 首先预测每个小室内众包对象动态出现个数,进而基于均匀分布的策略预测出该小室内众包对象空间分布。

对于众包对象的属性参数预测,DPRF 基于当前时间戳内已有众包对象的属性参数,采用正态分布的策略完成对应的众包对象属性参数预测。

3.1.1 众包对象空间分布预测

时空众包中众包任务和工人的动态出现具有相似性,本文采用相同的方法预测众包任务和工人。DPRF 中众包对象的空间分布预测可分为三步:

1) 基于在线随机森林模型初步预测小室 $cell$ 内的众包对象个数。

在时空众包环境下,一个区域内众包对象的动态出现与该区域的时空属性息息相关,本文选取了星期 WEEK、一天中的时间段 TQ、天气 WEA、区域的繁华程度 BQ,作为训练样本的四类特征,即随机森林中决策树分支的影响因素。其中对于时间段 TQ,本文以一个小时为间隔;对于天气 WEA,划分为晴、大雨、中雨、小雨、阴五种状况;对于区域的繁华程度 BQ,本文也将其划分为 5 种等级。

基于上述四类特征本文选取分类回归树(Classification and Regression Tree, CART)作为随机森林回归预测算法的基本单元,对于 CART 则采用最小均方差原则作为决策树节点划分的依据。在决策树生成过程中,对应的任意划分点 s 两边划分成的数据集 D_1 和 D_2 ,使得 D_1 和 D_2 各自对应的均方差最小,同时 D_1 和 D_2 的均方差之和最小。即节点选择条件为:

$$\min_{A,s} \left[\min_{x_i \in D_1(A,s)} \sum (y_i - c_1)^2 + \min_{x_i \in D_2(A,s)} \sum (y_i - c_2)^2 \right] \quad (9)$$

其中: c_1 为 D_1 数据集的样本输出均值, c_2 为 D_2 数据集的样本输出均值。

进而确定随机森林中决策树的个数 num_{DT} 、每个决策树的特征数量 num_{SF} 以及建立决策树时递归次数即树的深度 num_{REC} ,据此,从样本数据集中有放回的随机抽取 num_{DT} 次,每次从抽取的样本中选取 num_{SF} 个特征用于构建决策树,进而构建基于 CART 的随机森林。

对于小室 $cell$,随机森林算法将所有决策树的预测结果的平均值作为最终预测结果,即:

$$num_{cell} = \frac{1}{num_{DT}} \sum_{i=1}^{num_{DT}} pre_{DTi} \quad (10)$$

其中 pre_{DTi} 表示第 i 棵决策树的预测结果。

最终可得到 $n \times n$ 的网格图中小室 $cell_i$ 的初始预测结果 num_{cell_i} ,其中 $num_{w_{cell_i}}$ 表示该小室内众包工人的初始预测数目, $num_{m_{cell_i}}$ 表示该小室内众包任务的初始预测数目。

2) 基于滑动窗口确定每个小室 $cell$ 内的众包对象个数。

时空众包中空间位置相近的区域,众包对象的出现往往

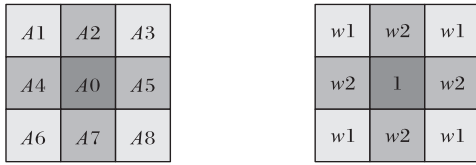


具有相似性,由此 DPRF 采用滑动窗口的方式,在初始预测的基础上,针对每个小室,结合滑动窗口对应相邻小室的预测结果,根据滑动窗口的权重,进行加权计算,完成众包对象数目的预测(对于网格图边缘的小室,无法满足滑动窗口,则取其剩余相邻的小室进行预测)。

以图 1(a)为例,小室 $cell_{A0}$ 作为待预测区域, $S_{cell} = \{cell_{A1}, cell_{A2}, cell_{A3}, cell_{A4}, cell_{A5}, cell_{A6}, cell_{A7}, cell_{A8}\}$ 作为滑动窗口涉及的相邻小室;图 1(b)表示滑动窗口对应的权重分布, $cell_{A1}, cell_{A3}, cell_{A6}, cell_{A8}$ 小室的权重为 $w1$, $cell_{A2}, cell_{A4}, cell_{A5}, cell_{A7}$ 小室的权重为 $w2$, 权重分布满足 $0 < w1 < w2 < 1$ 。预测小室 $cell_{A0}$ 的众包对象数目为:

$$n_{cell_{A0}} = \sum_{i=1}^{N_w} num_{cell_{Ai}} \times w_{cell_{Ai}} \quad (11)$$

其中: N_w 为滑动窗口中小室的数目(本例中 $N_w = 9$), $w_{cell_{Ai}}$ 表示滑动窗口对应小室的权重。 $cell_{A0}$ 中预测的众包工人数目为 $n_{w_{cell_{A0}}}$, 众包任务数目为 $n_{m_{cell_{A0}}}$ 。



(a) 小室的待预测区域 (b) 滑动窗口对应的权重分布

图 1 小室 $cell$ 的待预测区域和滑动窗口对应的权重分布示意图

Fig. 1 Area to be predicted of a $cell$ and weight distribution corresponding to sliding window

3)采用均匀分布的策略完成众包对象的空间分布预测。

对于网格图中每个小室 $cell$ 内预测的众包对象,均匀分布在四周的网格线上,完成空间分布预测。

3.1.2 众包对象参数属性预测

对于网格图中小室 $cell_i$ 内预测的众包对象的参数属性,即众包工人的任务成功率 s_w 和众包任务的回报值 g_m , ORFRP 算法采用正态分布的方式为小室 $cell_i$ 内预测的对象设置属性参数。以当前时间戳小室 $cell_i$ 内众包对象平均属性参数值作为正态分布的均值,以 σ_w 作为工人任务成功率的方差, σ_m 作为任务回报值的方差。公式表达如下:

$$g_{m_{cell_i}} \sim N(ag_{m_{cell_i}}, \sigma_m^2) \quad (12)$$

$$s_{w_{cell_i}} \sim N(as_{w_{cell_i}}, \sigma_w^2) \quad (13)$$

其中: $ag_{m_{cell_i}}$ 表示小室 $cell_i$ 当前时间戳内众包任务的平均回报值, $as_{w_{cell_i}}$ 表示小室 $cell_i$ 当前时间戳内众包工人的平均任务成功率。

综上,根据每个小室 $cell_i$ 内的众包对象的空间分布和属性参数分布完成该小室内众包对象的预测,进而基于每个小室的预测情况完成整个 $n \times n$ 网格图中下一时间戳内的动态出现众包对象预测。

3.2 基于 GRU 的工人移动轨迹预测算法(WMTP 算法)

首先定义众包工人 w 在时间戳 T_i 时的轨迹点:

$$p_{T_i} = \langle T_i, dir_w, x_c, y_c \rangle \quad (14)$$

其中: dir_w 表示工人的移动方向, x_c 和 y_c 表示工人在二维网格图中的横纵坐标值。

WMTP 算法基于 RNN-GRU 循环神经网络回归预测模型,根据众包工人 w 的移动轨迹序列 Seq_w 预测下一时间戳 T_{c+1} 任

务分配时工人 w 所处的空间轨迹点 $p_{T_{c+1}}$ 。工人移动轨迹序列 Seq_w 由连续 n 个时间戳的工人的移动轨迹点组成,即 $Seq_w = \{P_{T_{c-n+1}}, \dots, P_{T_{c-1}}, P_{T_c}\}$ 作为 RNN-GRU 预测网络的输入,预测网络输出下一时间戳 T_{c+1} 的轨迹点 $p_{T_{c+1}}$ 。据此,工人移动轨迹预测模型的表达式为:

$$p_{T_{c+1}} = f(\{P_{T_{c-n+1}}, \dots, P_{T_{c-1}}, P_{T_c}\}) \quad (15)$$

本文采用 many to one 类型的单层 RNN-GRU 循环神经网络,其展开图如图 2 所示。在模型训练和预测的过程中, RNN-GRU 模型每次输入样本为连续 n 个时间戳的工人移动轨迹点序列 Seq_w ($step = n$), 每个轨迹点 p_{T_i} 具有 T_i, dir_w, x_c, y_c 四项特征。

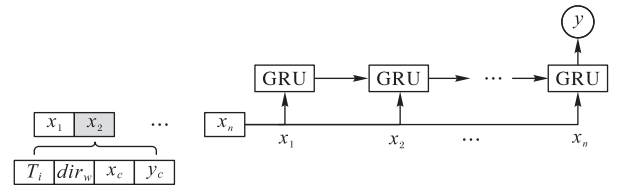


图 2 GRU 循环神经网络时序预测展开图

Fig. 2 GRU recurrent neural network time series prediction expansion diagram

对于样本输入中任意时刻 T_i 的轨迹点 p_{T_i} , GRU 处理过程如图 3 所示, GRU 输入包括当前轨迹点输入 x_i ($x_i = p_{T_i}$) 和上一时刻的隐藏状态 h_{i-1} ($h_0 = 0$), GRU 首先基于 h_{i-1} 和 x_i 通过更新门 z_i 确定上一时刻隐藏状态中信息的保留程度:

$$z_i = \sigma(x_i \cdot W_{xz} + h_{i-1} \cdot W_{hz} + b_z) \quad (16)$$

其中: σ 表示 sigmoid 函数, W_{xz}, W_{hz} 和 b_z 对应 z_i 计算时权重和偏置。同时根据重置门 r_i 将当前输入 x_i 和上一时刻隐藏状态 h_{i-1} 中信息结合:

$$r_i = \sigma(x_i \cdot W_{xr} + h_{i-1} \cdot W_{hr} + b_r) \quad (17)$$

其中 W_{xr}, W_{hr} 和 b_r 对应 r_i 计算时权重和偏置。进而基于重置门 r_i 和当前输入 x_i 确定候选隐藏状态 \tilde{h}_i :

$$\tilde{h}_i = \tanh(x_i \cdot W_{xh} + r_i \cdot h_{i-1} \cdot W_{hh} + b_h) \quad (18)$$

其中 W_{xh}, W_{hh} 和 b_h 对应 \tilde{h}_i 计算时权重和偏置。最后基于上一时间戳隐藏层状态 h_{i-1} 和当前时间戳候选隐藏状态 \tilde{h}_i 确定当前隐藏层状态 h_i :

$$h_i = z_i \cdot h_{i-1} + (1 - z_i) \cdot \tilde{h}_i \quad (19)$$

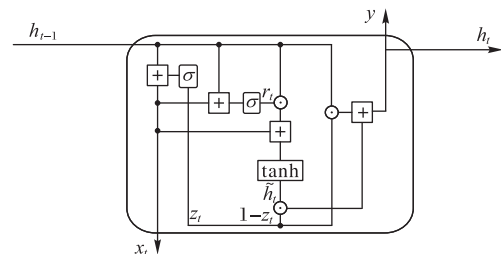


图 3 GRU 神经元内部结构

Fig. 3 GRU neuron internal structure

如果本次样本输入的是最后一个时刻 x_n 的数据,则预测模型在本次处理得到隐藏状态 h_i 的基础上添加全连接层输出 y , 否则 h_i 作为下一时刻的隐藏状态输入。

在训练过程中选用均方差函数作为损失函数,即预测值与真实值之差的平凡的期望值,根据 BPTT(Back Propagation



Through Time)算法计算梯度,进而更新模型的参数,完成模型训练。

由此,根据上述训练完成的RNN-GRU模型预测出当前时间戳内已有众包工人 w 在下一时间戳时所处空间位置是 $l_{nw}(x_{c+1}, y_{c+1})$ 的轨迹点 $p_{T_{c+1}}$ 。完成已有众包工人在下一时间戳内分布预测。

3.3 基于KM算法在线任务分配

在时空众包任务分配中,将众包任务和工人分别看作是二分图的两个点集,可分配任务工人匹配对集合 MP 看作边集,对应任务工人匹配对综合效益 CB 作为边的权重,执行二分图最优匹配算法(KM算法),以完成三方综合效益全局优化的任务分配。

在本文提出的基于预测分析的全局优化在线任务分配算法GOMA中,首先构建基于预测分析的二分图模型,将当前待分配任务集合 M_t 和DPRF预测新出现的任务集合 M_m 看作是二分图的一个点集 V_m 。当前待分配工人集合 W_t 、DPRF预测新出现的工人集合 W_m 以及基于WMTP算法预测出下一时间戳众包工人集合 W_{mt} 看作是二分图的另一点集 V_w 。其中: W_{mt} 是将WMTP算法预测出工人 w 在下一时间戳所在空间轨迹点 $p_{T_{c+1}}$ 看作是处于 l_{nw} 位置,属性参数相同的众包工人 w_m 。工人点集中每一个工人 w 与其最适接单范围内的任务进行预匹配,构成可分配任务工人匹配对 mp ,进而根据任务工人匹配对集合 MP 连接二分图。并假设当前时间 CT 为分配时间,计算每个任务工人匹配对的 U_p 、 AT'_m 、 CO'_w 三项指标,进而计算综合匹配效益 CB ,以 CB 作为边的权重。

最后基于上述搭建的二分图模型执行KM算法,在算法匹配结果 S_p 中,将众包任务和工人均处于当前时间戳内的匹配对 mp 加入分配结果集 R 中。对于含有预测任务或工人的匹配对,则表明这些众包对象在后期分配,会使得全局分配效果更优,故当前时间戳不予分配。

3.4 算法具体实现

GOMA伪代码如下:

Input: 众包任务集合 M ,众包工人集合 W ,综合收益指标 CB 的计算函数、已训练完成RNN-GRU预测模型;

Output: 任务分配的结果集 R 。

- 1) for (每个时间戳 T_i):
- 2) 将当前时间戳内的待分配的众包对象加入对应的待分配集合 M_t 或 M_m 中
- 3) 基于历史数据样本集构造随机森林预测模型
- 4) for ($n \times n$ 网格图中的每一个小室 $cell_{ij}$)
- 5) 根据当前小室 $cell_{ij}$ 的时空属性基于随机森林预测模型预测初始值
- 6) 在初始预测值的基础上基于滑动窗口预测众包对象数目
- 7) 基于当前时间戳小室 $cell_{ij}$ 内众包对象的属性参数,采用正态分布的策略预测动态出现对象的属性参数
- 8) 将预测对象均匀分布在小室的四周并将加入相应的预测集合 M_m 或 W_m
- 9) end for
- 10) for(当前时间戳在线工人 $w \in W_t$)
- 11) 将工人 w_i 的连续 n 个时间戳移动轨迹序列输入RNN-GRU预测网络中,预测出工人 w_i 在下一时间戳的空间位置 $l_{nw}(x_{nc}, y_{nc})$
- 12) 假设工人 w'_i 以 $l_{nw}(x_{nc}, y_{nc})$ 为空间位置,具有与工人 w_i 相同的属性参数,将工人 w'_i 加入集合 W_{mt}
- 13) end for
- 14) 任务集合 M_t 、 M_m 组成二分图点集 V_m ,工人集合 W_t 、 W_m 、

W_{mt} 组成二分图点集 V_w ,构建可分配匹配对集合 MP ,计算匹配对相应的匹配系数 mf ,搭建带权二分图

- 15) 执行带权二分图最优匹配算法,得到匹配集合 S_p
- 16) for($mp \in S_p$)
- 17) if(mp 中 m 和 w 均属于当前时间戳)
- 18) mp 加入分配结果集 R 中
- 19) else
- 20) 等待下一时间戳分配
- 21) end for
- 22) 根据本轮分配结果更新在线随机森林的历史数据样本集
- 23) end for
- 24) return R

GOMA每一时间戳内的算法时间复杂度分析如下:

DPRF的时间复杂度是 $O(num_{SF} \times num_{REC})$,其中 num_{SF} 是随机森林中决策树的特征数量, num_{REC} 是决策树的深度。基于GRU的工人移动轨迹预测算法(WMTP算法)的时间复杂度为 $O(num_{PW} \times step)$,其中 num_{PW} 是当前待预测的工人数目, $step$ 是GRU网络中输入轨迹序列中轨迹点的数目。进而执行带权二分图最优匹配算法的时间复杂度是 $O(|V_m|^2 \times |V_w|)$,其中 $|V_m|$ 是二分图边集的容量, $|V_w|$ 是二分图点集的容量。所以GOMA一轮任务分配的时间复杂度是: $\max(O(num_{SF} \times num_{REC}), O(num_{PW} \times step), O(|V_m|^2 \times |V_w|))$ 。

3.5 算法举例

假设当前时间戳 $T_i = 15$ 内待分配任务有 m_1, m_2 ,待分配工人有 w_1, w_2 ,平台工人的接单范围半径均为 $r_w = 1.5$,时间戳时间间隔 $|CT| = 3$,工人移动速度 $speed = 0.5$ 单位长度/单位时间,工人的移动成本 $per = 1$ 单位成本/单位长度,任务工人匹配对综合效益 $CB = 0.5U_p + 0.15AT'_m + 0.2/MT'_m + 0.15CO'_w$, $AT'_m = AT_m/9$, $MT'_m = MT_m/4$, $CO'_w = CO_w/4.5$ 。工人属性参数见表1,任务属性参数见表2。

表1 工人信息

Tab. 1 Information of workers

工人	位置 l_w	移动方向 dir_w	上线时间 p_w	成功率 s_w
w_1	(4, 5, 6)	W	9	0.75
w_2	(6, 5, 5)	S	12	0.80
w_{n1}	(5, 3, 5)	N	18	0.70
w_{m1}	(4, 5)	E	9	0.75
w_{m2}	(6, 4)	S	12	0.80

表2 实验数据参数

Tab. 2 Parameters of experimental data

参数	取值
工人的数量 $ W $	5 000, 7 000, 9 000, 11 000, 13 000
任务数量 $ M $	5 000, 7 000, 9 000, 11 000, 13 000
任务和工人坐标	100×100网格图的网格线上
网格图中小室单位长度	1.0
时间戳长度	2
任务和工人出现的时间	在 $[0, 240]$ 区间均匀分布
任务的有效时间上限 ET_m	8, 9, 10, 11, 12
工人的接单范围半径 r_w	3, 4, 5, 6, 7
工人的任务成功率均值 \bar{s}_w	0.7
工人的任务成功率标准差	0.1
任务的回报值均值	65
任务的回报值标准差	10
工人的移动速度 $speed$	1.5单位长度/单位时间
工人单位路程成本 per	3



按照GOMA的执行步骤:

1) 基于DPRF预测下一时间戳 T_{i+1} 内该平台上新发布任务 m_{n1} 、 m_{n2} 和新上线工人 w_{n1} 。

2) 基于WMP算法预测当前时间戳的待分配工人 w_1 、 w_2 在下一时间戳空间分布点 w_{m1} 、 w_{m2} 。

3) 当前对象和预测对象的分布情况如图4所示,构建二分图模型并计算权重 T'_m 和 CO'_w 。权重计算以 $mp_1 = \langle w_1, m_2 \rangle$ 为例, $U_p = 0.4875$, $AT'_m = 3/9 = 1/3$, $MT'_m = 3/4$, $CO'_w = 3/4.5 = 2/3$, $CB = 0.5U_p + 0.15AT'_m + 0.2/MT'_m + 0.15CO'_w = 0.658$ 。

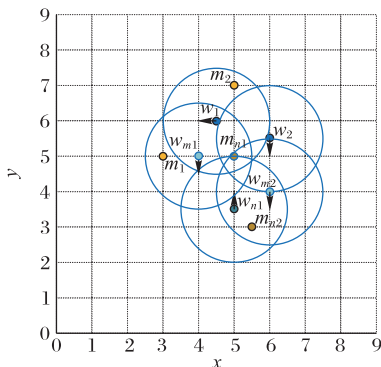


图4 众包任务和众包工人分布示意图

Fig. 4 Crowdsourcing tasks and distribution of crowdsourcing workers

二分图模型如图5所示,然后执行带权二分图最优匹配算法(KM算法),得到最优匹配集:

$$S_p = \{ \langle w_1, m_2 \rangle, \langle w_2, m_{n1} \rangle, \langle w_{n1}, m_{n2} \rangle, \langle w_{m1}, m_1 \rangle \}$$

选择任务和工人均属于当前时间戳内的匹配对 $\langle w, m \rangle$ 加入结果集 R 。剩余众包对象等待下一轮任务分配。本轮GOMA结束。

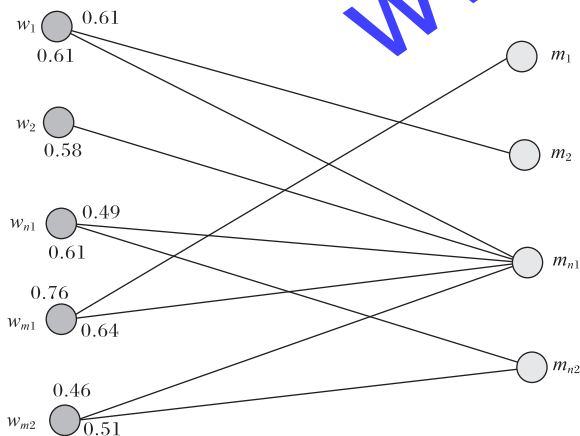


图5 众包任务和众包工人匹配示意图

Fig. 5 Matching of crowdsourcing tasks and crowdsourcing workers

4 实验分析

4.1 实验数据与环境

为验证本文算法的有效性,从微软开源数据集T-Drive项目中获得出租车的行驶轨迹点数据作为众包工人的初始数据集,通过相应的预处理得到15 174组数据;从时空众包平台gMission^[23]上爬取了17 296条记录,进行相应处理取得众包任

务的相应数据。将任务和工人的位置映射在100×100的网格图中,得到任务和工人的二维坐标。从处理后的数据集中基于均匀分布的原则选取13 000组众包工人和任务进行仿真实验。

本实验在处理器为2.3 GHz Inter Core i5-6300HQ,内存为8 GB的计算机上运行,操作系统为Windows 10。基于Tensorflow的上层框架Keras进行模型的搭建。实验使用的编程语言为Python,使用的集成开发环境是PyCharm。

4.2 预测模型参数取值

本文WMP算法所设计的工人轨迹预测模型为3层,分别为输入层、GRU隐藏层和输出层。本文设置GRU隐藏层神经元节点数CELL_SIZE、输入层步数step作为模型可调节参数,均方差(Mean Squared Error, MSE)作为模型的评估指标,以确定最佳的模型参数取值。

设置输入层步数step从3到7,隐层神经元节点数CELL_SIZE为同一step下的最优值,不同step对应的模型均方差如图6所示。

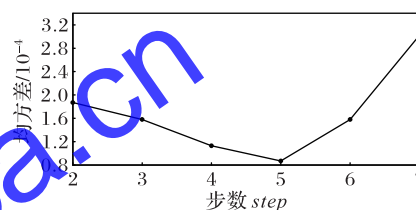


图6 GRU输入时序step的均方差分布

Fig. 6 MSE distribution of step for GRU input time series

随着输入的维度越大,模型的复杂程度越高,预测效果越好,但当输入维度过高时模型会发生过拟合。根据实验结果选取step = 5作为输入层序列 Seq_w 中的轨迹点数量,此时模型的预测效果最好。

将连续5组工人(其中每组包含工人数100名)的平均移动轨迹点组成轨迹序列 Seq_w 作为网络输入,设置5~15不同的GRU隐层神经元节点数,评估模型的预测效果,实验结果如图7所示。由图7可知,当隐层神经元数目为11时模型均方差最小。表明当输入层step = 5时,隐层神经元个数为11,模型的预测效果最佳。

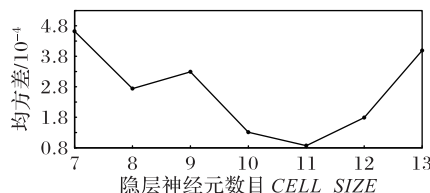


图7 GRU神经元均方差分布

Fig. 7 MSE distribution of GRU neurons

4.3 GOMA有效性实验

本实验从任务分配成功率、分配的平均综合效益、工人平均机会成本三个方面对贪心算法、随机阈值算法和GOMA进行比较分析,实验数据的参数设置如表2,其中任务回报值 g_m 和工人的任务成功率 s_w 满足正态分布。



1) 在任务分配成功率方面,具体实验结果如图8所示。GOMA 始终优于贪心算法和随机阈值算法。当工人数量 $|W|$ 增加时,三种算法的任务分配成功率均由增长逐渐趋于稳定;当任务数量 $|M|$ 增加时,由于工人数量相对逐渐减少,三种算法的任务分配成功率均有一定程度的降低;当任务的有效时间上限 ET_m 增加时,待分配任务可参与多轮任务分配,GOMA 和贪心算法均有小幅度增长,但随即阈值算法的波动较大;当工人的接单范围半径 r_w 增加时,GOMA 的任务分配成功率增速明显高于贪心算法和随机阈值算法。

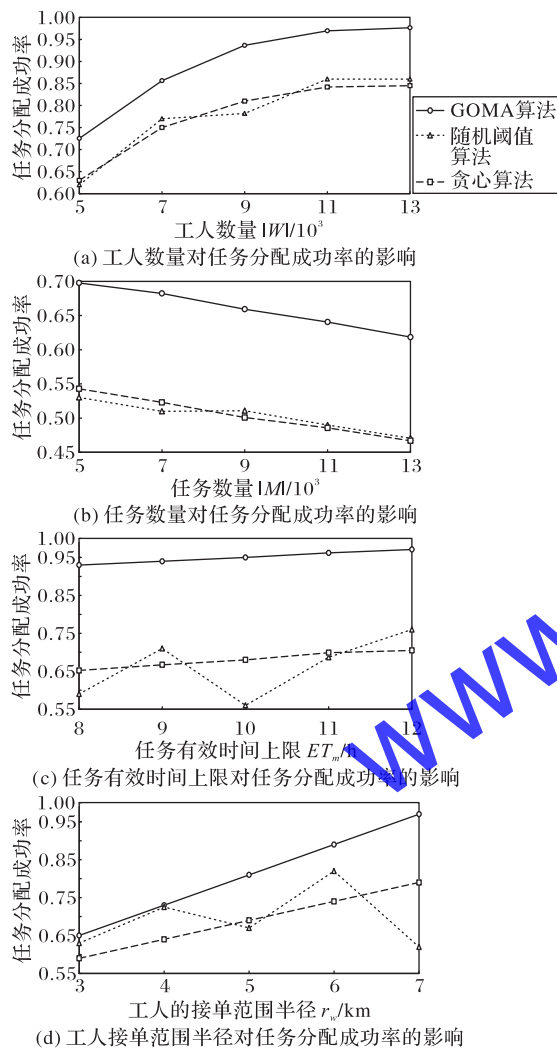


图8 三种算法在任务分配成功率方面的实验结果

Fig. 8 Experimental results of three algorithms in success rate of task allocation

2) 在平均综合效益方面,具体实验结果如图9所示。GOMA 整体相对稳定,对比贪心算法和随机阈值算法,平均综合效益有大幅度的提升。工人数量 $|W|$ 与任务数量 $|M|$ 对GOMA 和贪心算法的影响不大,随机阈值算法由于随机选取阈值,平均综合效益变化较大;当任务的有效时间上限 ET_m 和工人的接单范围半径 r_w 增加时,可分配的任务工人匹配对数量增加,GOMA 和贪心算法的平均综合效益均有小幅度增长。

3) 在工人平均机会成本方面,具体实验结果如图10所示。当工人数量 $|W|$ 和任务数量 $|M|$ 增加时,GOMA 的工人平

均机会成本有小幅度变化,但一直明显优于贪心算法和随机阈值算法;当任务的有效时间上限 ET_m 增加时,GOMA 处于较为稳定状态,贪心算法和随即将阈值算法的工人平均机会成本逐渐增大;当工人的接单范围半径 r_w 增加时,贪心算法和随机阈值算法的工人平均机会成本增速明显高于GOMA。

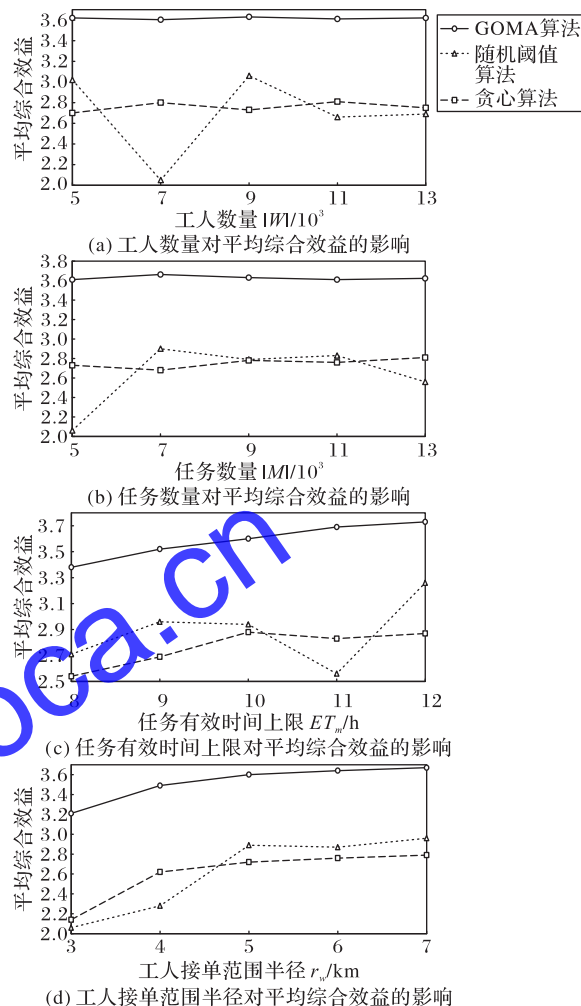


图9 三种算法在平均综合效益方面的实验结果

Fig. 9 Experimental results of three algorithms in average comprehensive benefit

从以上实验不难发现:

1) 在任务分配成功率方面,GOMA 始终优于贪心算法和随机阈值算法。

2) 在分配的平均综合效益方面,工人数量 $|W|$ 和任务数量 $|M|$ 的变化对GOMA 影响不大,当任务的有效时间上限 ET_m 和工人的接单范围半径 r_w 增加时,GOMA 分配的平均综合效益逐渐增长并趋于稳定,对比贪心算法和随机阈值算法具有明显的优势。

3) 在工人的平均机会成本方面,当工人数量 $|W|$ 、工人的接单范围半径 r_w 增加时,三种算法工人的平均机会成本均逐渐增加,但贪心算法和随机阈值算法的增速明显高于GOMA。当任务数量 $|M|$ 增加时,三种算法的工人平均机会成本均逐渐降低,但GOMA 工人的平均机会成本低于另外两种算法。当任务的有限时间上限 ET_m 增加时,GOMA 较为稳定,但随机



阈值算法的波动比较大。

综上,可以看出本文提出的GOMA具有较好的实际应用价值,能有效地解决本文研究的CMA-MCB问题。

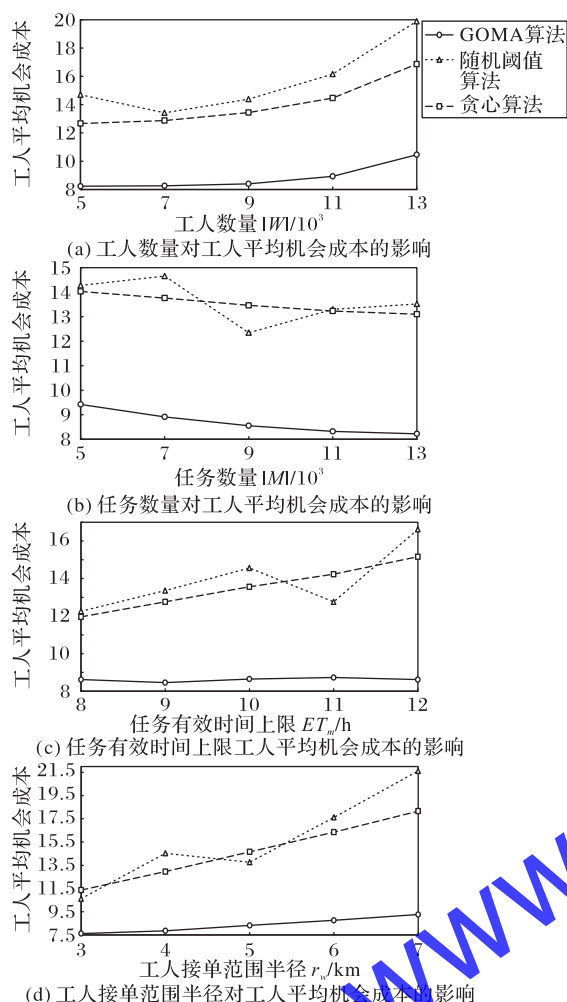


图10 三种算法在工人平均机会成本方面的实验结果

Fig. 10 Experimental results of three algorithms in average opportunity cost of workers

5 结语

本文研究了时空众包环境下面向全局优化的在线任务分配问题。首先采用基于在线随机森林的众包对象动态预测算法(DPRF)预测下一时间戳内众包对象动态出现的情况,然后利用基于GRU的工人移动轨迹预测算法(WMTP)预测众包工人的移动轨迹,最后结合当前众包对象的分布情况,基于带权二分图最优匹配算法进行任务分配,能够有效地提高任务分配的综合效益。通过实验证明了本文所提算法在任务分配率、分配的平均综合效益、工人的平均机会成本方面具有较好的性能表现,能够有效地解决时空众包中全局优化的在线任务分配问题。未来的时空众包研究,可从以下两个方面展开:1)针对众包对象的分布情况预测,进行多步时间戳预测,进一步优化任务分配的效果;2)引入强化学习、预训练等深度学习方法优化预测模型,进一步提高众包对象分布情况预测的准确率。

参考文献 (References)

- [1] HOWE J. The rise of crowdsourcing[J]. Wired Magazine, 2016, 14(6): 1-5.
- [2] 芮兰兰,张攀,黄豪球,等. 一种面向众包的基于信誉值的激励机制[J]. 电子与信息学报, 2016, 38(7): 1808-1815. (RUI L L, ZHANG P, HUANG H Q, et al. Reputation-based incentive mechanisms in crowdsourcing[J]. Journal of Electronics and Information Technology, 2016, 38(7): 1808-1815.)
- [3] 施战,辛煜,孙玉娥,等. 基于用户可靠性的众包系统任务分配机制[J]. 计算机应用, 2017, 37(9): 2449-2453. (SHI Z, XIN Y, SUN Y E, et al. Task allocation mechanism for crowdsourcing system based on reliability of users[J]. Journal of Computer Applications, 2017, 37(9): 2449-2453.)
- [4] SCHEE B A V. Crowdsourcing: why the power of the crowd is driving the future of business[J]. American Journal of Health-System Pharmacy, 2010, 67(18): 1565-1566.
- [5] LI Y, YIU M L, XU W. Oriented online route recommendation for spatial crowdsourcing task workers[C]// Proceedings of the 2015 International Conference on Advances in Spatial and Temporal Database, LNCS 9239. Cham: Springer, 2015: 137-156.
- [6] ALT F, SHIRAZI A S, SCHMIDT A, et al. Location-based crowdsourcing: extending crowdsourcing to the real world[C]// Proceedings of the 10th Nordic Conference on Human-Computer Interaction: Extending Boundaries. New York: ACM, 2010: 13-22.
- [7] MUSTHAG M, GANESAN D. Labor dynamics in a mobile micro-task market[C]// Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 2013: 641-650.
- [8] TO H, GHINITA G, SHAHABI C. A framework for protecting worker location privacy in spatial crowdsourcing[J]. Proceedings of VLDB Endowment, 2014, 7(10): 919-930.
- [9] CHENG P, LIAN X, CHEN Z, et al. Reliable diversity-based spatial crowdsourcing by moving workers[J]. Proceedings of the VLDB Endowment, 2015, 8(10): 1022-1033.
- [10] DENG D, SHAHABI C, DEMIRYUREK U. Maximizing the number of worker's self-selected tasks in spatial crowdsourcing[C]// Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2013: 314-323.
- [11] 童咏昕,袁野,成雨蓉,等. 时空众包数据管理技术研究综述[J]. 软件学报, 2017, 28(1): 35-58. (TONG Y X, YUAN Y, CHENG Y R, et al. Survey on spatiotemporal crowdsourced data management techniques[J]. Journal of Software, 2017, 28(1): 35-58.)
- [12] 宋天舒,童咏昕,王立斌,等. 空间众包环境下的3类对象在线任务分配[J]. 软件学报, 2017, 28(3): 611-630. (SONG T S, TONG Y X, WANG L B, et al. Online task assignment for three types of objects under spatial crowdsourcing environment[J]. Journal of Software, 2017, 28(3): 611-630.)
- [13] TONG Y, SHE J, DING B, et al. Online mobile micro-task alloca-



- tion in spatial crowdsourcing [C]// Proceedings of the IEEE 32nd International Conference on Data Engineering. Piscataway: IEEE, 2016: 49-60.
- [14] UL HASSAN U, CURRY E. A multi-armed bandit approach to on-line spatial task assignment [C]// Proceedings of the IEEE 11th International Conference on Ubiquitous Intelligence and Computing and IEEE 11th International Conference on Autonomic and Trusted Computing and IEEE 14th International Conference on Scalable Computing and Communications and Its Associated Workshops. Piscataway: IEEE, 2014: 212-219.
- [15] 余敦辉, 张灵莉, 付聪. 基于动态效用的时空众包在线任务分配[J]. 电子与信息学报, 2018, 40(7): 1699-1706. (YU D H, ZHANG L L, FU C. Online task allocation of spatial crowdsourcing based on dynamic utility [J]. Journal of Electronics and Information Technology, 2018, 40(7): 1699-1706.)
- [16] CHENG P, LIAN X, CHEN L, et al. Prediction-based task assignment in spatial crowdsourcing [C]// Proceedings of the IEEE 33rd International Conference on Data Engineering. Piscataway: IEEE, 2017: 997-1008.
- [17] 张晨, 郭玉超, 林培光, 等. 空间众包中基于位置预测的任务分配算法[J]. 南京大学学报(自然科学), 2018, 54(2): 471-480. (ZHANG C, GUO Y C, LIN P G, et al. Location prediction-based task assignment in spatial crowdsourcing [J]. Journal of Nanjing University (Natural Science), 2018, 54(2): 471-480.)
- [18] LONG C, WONG R C W, YU P S, et al. On optimal worst-case matching [C]// Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2013: 845-856.
- [19] TONG Y, SHE J, DING B, et al. Online minimum matching in real-time spatial data: experiments and analysis [J]. Proceedings of the VLDB Endowment, 2016, 9(12): 1053-1064.
- [20] 张兴盛, 余敦辉, 聂茜婵, 等. 基于预测分析的时空众包在线任务分配[J]. 计算机工程, 2019, 45(6): 67-74. (ZHANG X S, YU D H, NIE X C, et al. Spatiotemporal crowdsourcing online task allocation based on predictive analysis [J]. Computer Engineering, 2019, 45(6): 67-74.)
- [21] DENG D, SHAHABI C, ZHU L. Task matching and scheduling for multiple workers in spatial crowdsourcing [C]// Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM, 2015: No. 21.
- [22] 张兴盛, 余敦辉, 张万山, 等. 时空众包环境下时效均衡的在线任务分配算法[J]. 计算机应用, 2019, 39(5): 1357-1363. (ZHANG X S, YU D H, ZHANG W S, et al. Time utility balanced online task assignment algorithm under spatial crowdsourcing environment [J]. Journal of Computer Applications, 2019, 39(5): 1357-1363.)
- [23] CHEN Z, FU R, ZHAO Z, et al. gMission: a general spatial crowdsourcing platform [J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1629-1632.

This work is partially supported by the National Key Research and Development Program of China (2017YFB1400602), the National Natural Science Foundation of China (61572371, 61832014).

NIE Xichan, born in 1999. Her research interests include spatial crowdsourcing, mapping knowledge domain.

ZHANG Yang, born in 1999. His research interests include data mining.

YU Dunhui, born in 1974, Ph. D., professor. His research interests include service computing, big data.

ZHANG Xingsheng, born in 1998. His research interests include spatial crowdsourcing.