软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(1):35−58 [doi: 10.13328/j.cnki.jos.005140]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

时空众包数据管理技术研究综述*

童咏昕[1,2], 袁 野[3], 成雨蓉[3], 陈 雷[4], 王国仁[3]

[1](软件开发环境国家重点实验室(北京航空航天大学),北京 100191)

[2](北京航空航天大学 计算机学院,北京 100191)

[3](东北大学 计算机科学与工程学院,辽宁 沈阳 110004)

[4](香港科技大学 计算机科学与工程学系,香港)

通讯作者: 童咏昕, E-mail: yxtong@buaa.edu.cn

摘 要: 近年来,众包为传统数据管理提供了一种通过汇聚群体智慧求解问题的新模式,并成为当前数据库领域的研究热点之一.特别是随着移动互联网技术与共享经济模式的快速发展,众包技术已融入到各类具有时空数据的应用场景中,例如各类 O2O(online-to-offline)应用、实时交通监控与动态物流管理等.简言之,这种应用众包技术处理时空数据的方式称为时空众包数据管理.对近期在时空众包数据管理方面的研究工作进行综述,首先阐述了时空众包的概念,解释了其与传统众包技术的关系,并介绍了各类典型的时空众包应用;随后描述了时空众包应用平台的工作流程及其任务特点;然后讨论了时空众包数据管理的 3 项核心研究问题和 3 类应用技术;最后,总结了时空众包数据管理技术的研究现状并展望了其未来潜在的研究方向,为相关研究人员提供了有价值的参考.

关键词: 时空众包;共享经济;O2O 模式;任务分配;质量控制;隐私保护

中图法分类号: TP311

中文引用格式: 童咏昕,袁野,成雨蓉,陈雷,王国仁.时空众包数据管理技术研究综述.软件学报,2017,28(1):35−58. http://www.jos.org.cn/1000-9825/5140.htm

英文引用格式: Tong YX, Yuan Y, Cheng YR, Chen L, Wang GR. Survey on spatiotemporal crowdsourced data management techniques. Ruan Jian Xue Bao/Journal of Software, 2017,28(1):35−58 (in Chinese). http://www.jos.org.cn/1000-9825/5140.htm

Survey on Spatiotemporal Crowdsourced Data Management Techniques

TONG Yong-Xin[1,2], YUAN Ye[3], CHENG Yu-Rong[3], CHEN Lei[4], WANG Guo-Ren[3]

[1](State Key Laboratory of Software Development Environmnt (Beihang University), Beijing 100191, China)

[2](School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

[3](School of Computer Science and Engineering, Northeastern University, Shenyang 110004, China)

[4](Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, China)

Abstract: In recent years, crowdsourcing, which utilizes the intelligence of crowds to solve problems, provides a novel data processing paradigm for traditional data management challenges and has become one of the hottest research topics. In particular, due to the rapid development of mobile Internet and sharing economy, crowdsourcing not only becomes a new approach for data collection, but is also integrated into all kinds of application scenarios especially spatiotemporal data management such as online-to-offline (O2O) applications,

* 基金项目: 国家重点基础研究发展计划(973)(2014CB340300); 国家自然科学基金(61502021, 61622202, 61572119, U1401256); 北京航空航天大学软件开发环境国家重点实验室开放课题(SKLSDE-2016ZX-13)

Foundation item: National Basic Research Program of China (973) (2014CB340300); National Natural Science Foundation of China (61502021, 61622202, 61572119, U1401256); State Key Laboratory of Software Development Environment (Beihang University) Open Program (SKLSDE-2016ZX-13)

收稿时间: 2016-10-08; 修改时间: 2016-10-24; 采用时间: 2016-11-17; jos 在线出版时间: 2016-11-24

CNKI 网络优先出版: 2016-11-24 13:41:16, http://www.cnki.net/kcms/detail/11.2560.TP.20161124.1341.007.html

real-time traffic monitoring, and logistics management. In this paper, a survey is provided on existing research of spatiotemporal crowdsourcing. First of all, the concept and representative applications of spatiotemporal crowdsourcing is described, and its relationship with traditional crowdsourcing is explained. Then, the workflow of spatiotemporal crowdsourcing is illustrated. Furthermore, three core research problems and three categories of techniques of spatiotemporal crowdsourcing are discussed. Finally, the state-of-the-art studies of spatiotemporal crowdsourcing are summarized and promising future research directions for the research community are presented.

Key words: spatiotemporal crowdsourcing; sharing economy; O2O mode; task assignment; quality control; privacy protection

随着 Web 2.0 技术的兴起,大量在线 Web 应用正在悄然地改变着人类的生活模式,同时也为传统的人本计算(human computation)提供了一种通过群体智慧求解问题的新模式——众包(crowdsourcing)[1].所谓众包,通常是指“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”[2].在过去的 10 余年里,众包技术已与人们的日常生活息息相关.例如,早期的众包平台通常指“问答系统”平台,如维基百科(wikipedia)、雅虎问答(Yahoo! answers)与百度知道等,发展至今已成为现代人们获取知识的必需品.近年来,由于早期众包平台所支持的任务类型单一,其已不能满足当前数据类型多样化与任务复杂化的 Web 应用需求,这也促使新一代“在线众包平台”的诞生,即,大型在线工作招募与任务分包管理平台,例如 Amazon Mechanical Turks(AMT)[3],CrowdFlower[4],oDesk[5]等.该类众包平台不但带来了新的技术革命,更创造了巨大的市场经济价值.根据美国亚马逊公司的年度报告,截至 2010 年,该公司在 AMT 众包平台上的年度盈利已经超过 5.2 亿美元.因此,众包技术为当今互联网时代的技术革命带来了巨大潜能,正如《人民日报》2014 年关于众包的报道所述:“众包模式,大势所趋”[6].与此同时,随着大数据时代的到来,虽然各类数据驱动型应用不断涌现,但由于受到传统数据管理技术自身瓶颈的制约,许多传统数据管理难题(例如实体同一识别问题)在大数据时代将更难解决.然而,众包技术通过汇聚群智可将人类经验融入到求解问题之中,为突破传统数据管理挑战开辟了全新的视角.因此,基于众包的数据管理技术(简称为“众包数据管理技术”)已经引起学术界和产业界的广泛重视[7−10].

作为一项新兴的研究热点,当前的众包数据管理技术主要关注如何将众包策略融入到传统数据库管理系统之中,从而提高数据管理的质量.例如,筛选查询(filtering query)[11,12]、连接查询(join query)[13]、最大值查询(top-1 query)[14,15]、Top-k 查询(top-k query)[16]与聚集查询(aggregation query)[17]等经典查询处理技术都已经被扩展到新型的众包数据管理系统之中.而且,一些众包数据管理原型系统近年来也先后问世.例如,美国加州伯克利大学研发的 CrowdDB 系统[18]、麻省理工学院研制的 Qurk 系统[13]和斯坦福大学开发的 Deco 系统[19].由于现有研究侧重于将众包技术集成到数据库管理系统内从而形成众包数据管理系统,因此可以说现有研究大多是众包数据管理的内涵研究.

另一方面,随着移动互联网技术与共享经济模式的快速发展,移动计算技术为众包数据管理带来了更多的外延需求,其不仅延伸了众包数据管理系统所需管理数据的类型,更延伸了众包数据管理系统可获取数据的方式.所谓延伸所需管理数据的类型是指:由于移动设备自身携带着大量时空数据,且此类时空数据又与众包任务和众包参与者(也称为“众包工人”)的行为密切相关,众包数据管理系统不得不考虑如何有效地处理此类新型数据问题.例如,近年来全球流行的各类实时专车类服务平台,如滴滴出行[20]、神州专车[21]与 Uber[22]等,均采用时空众包方式提供服务,其中,专车用户为众包任务请求者,专车司机即众包参与者.所谓延伸可获取数据的方式是指:移动设备日益强大的功能产生了一类以获取数据为目标的新型众包任务.例如,美国的 Gigwalk 公司[23]组织众包参与者通过智能手机收集不同超市的物品价格,而国内高德地图公司推出的“道路寻宝”服务也旨在组织众包参与者收集国内各大城市的道路周边信息[24].

综上所述,移动互联网与物联网等技术的飞速发展,使得众包数据管理技术从基于在线众包平台的模式转变为一种新型的服务模式,称为“时空众包(spatiotemporal crowdsourcing)”(也称为空间众包或移动众包)[25].简言之,时空众包数据管理技术是指以时空数据管理平台为基础,将具有时空特性的众包任务分配给非特定的众包参与者群体为核心操作,要求众包参与者以主动或被动的方式来完成众包任务并满足任务所指定时空约束条件的一种新型众包计算模式.特别地,当前“互联网+”时代的共享经济模式为时空众包数据管理技术提供了大

量实际应用.具体而言,近年来流行的各类 O2O(online-to-offline)应用、灾情监控、交通管理、公共安全、物流管理和社交媒体等领域,都有意或无意地采用了时空众包技术以提高其服务质量.因此,时空众包数据管理技术已遍及百姓衣食住行等各个领域,并在人们日常生活中扮演着越来越重要的角色.

由于应用驱动研究,时空众包数据管理技术成为国际数据库与数据挖掘领域新近发展起来的一个研究热点[26−49].近年来,已有一些研究人员在国际数据库与数据挖掘的顶级期刊和会议上对众包数据管理技术的相关研究进行了总结[7−10,50−53],并也提及了少量时空众包的相关研究内容[50].例如,Li 等人[50]对众包数据管理技术进行了全面的综述,在其文献[50]中提及了时空众包的静态离线任务分配机制和基于众包最佳路径查询技术,但并未对时空众包数据管理技术的研究现状进行深入而完整的综述.因此,不同于上述众包技术综述类文章,本文仅聚焦于时空众包数据管理技术,首先揭示时空众包与传统众包技术的区别与联系,随后以时空众包平台的工作流程与其任务特点为基础深入讨论了当前时空众包数据管理的 3 项核心研究问题和 3 类应用技术的研究现状,同时也展望了时空众包数据管理领域未来潜在的研究方向,为相关研究人员提供有价值的参考.

本文第 1 节简介时空众包的相关概念与典型应用.第 2 节详细论述时空众包平台的工作流程和任务特征.第 3 节介绍时空众包研究中的 3 项核心研究问题:任务分配、质量控制与隐私保护.第 4 节从时空众包数据管理技术的角度详细综述其在数据集成、数据查询与数据挖掘这 3 类应用中的研究现状,并对各类技术进行完整的对比分析.第 5 节展望时空众包数据管理未来潜在的研究方向.第 6 节是结束语.

1 时空众包概念和应用领域

本节介绍时空众包的概念,阐明其与传统众包技术之间的关系,并介绍时空众包技术的代表性应用领域.

1.1 时空众包概念

定义 1(时空众包任务)[26,45]. 一个时空众包任务被该任务的请求者发布,通常被定义为如下五元组的形式,记为 $t=\langle l_t,a_t,d_t,r_t,u_t\rangle$,其中,$l_t$表示该任务的位置;$a_t$为该任务的发布时间;$d_t$为该任务的截止时间;$r_t$为该任务发布的空间范围,即,在此范围内的众包参与者才有机会接收到该任务;u_t 是完成该任务可以获得的奖励或效用,通常表示为任务的价格或奖金.

对于任意的时空众包任务(下文简称“众包任务”),上述五元组中的前三者应必须被包括以标记此任务的时空属性;是否包含后两者视具体应用而定.例如,某些任务希望众包平台的每位众包参与者都获知,则可去除此空间范围约束.另外,如果某些任务不为参与者带来任何奖励,也可删除该项内容.

此外,时空众包参与者定义如下.

定义 2(时空众包参与者)[26,45]. 一位时空众包参与者也被称为时空众包工人,通常被定义为如下六元组的形式,记为 $w=\langle l_w,a_w,d_w,r_w,c_w,q_w\rangle$,其中,$l_w$表示该参与者当前的空间位置;$a_w$ 为此参与者抵达时空众包平台的时间;d_w 为该参与者预计离开时空众包平台的时间;r_w 为该参与者的空间服务范围,即,对于该范围外的众包任务,此参与者将不能提供服务;c_w 代表该参与者计划承担的众包任务数量;q_w 度量该参与者提供服务的质量,通常表示为历史任务成功率或历史服务满意度等形式.

与时空众包任务的定义相似,对于任意时空众包参与者(下文简称“众包参与者”或“参与者”)的六元组,也是前三者应被包括,而后三者可视具体应用而定.注意:上述时空众包任务与时空众包参与者的定义皆为基础性定义,根据不同的应用需求,可在上述两个定义的基础上进行扩展.因此,基于上述定义,可将时空众包定义如下.

定义 3(时空众包). 时空众包通常是指通过移动互联网设备实时地在时空众包平台上汇聚众包任务与众包参与者,并通过平台对众包任务进行分配调度与质量控制,从而使众包参与者在物理世界完成众包任务并满足任务约束条件的过程.

综上所述,时空众包旨在通过整合移动互联网中线上空闲的大众群体,组织其在线下物理世界完成机器难以解决的问题,从而有效地利用线上与线下的闲置资源.因此,时空众包计算正是各类 O2O 应用的通用计算范式.

1.2 时空众包与传统众包的关系

自 Howe[2]于 2006 年首次提出众包的概念以来,这种通过公开的 Web 平台将任务分配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正在变得日益流行.特别是得益于移动互联网技术与共享经济模式的快速发展,众包技术正在从传统的基于 Web 的应用模式转向各类时空众包应用.因此,一个自然的问题就是时空众包与传统众包之间的区别与联系是怎样的?

一方面,时空众包属于传统众包思想的延伸.从时空众包的名称即可知,其属于众包研究中的一类子问题,更是传统众包在时空维度的新衍生形态.另一方面,时空众包并不是传统众包技术与时空数据的简单组合,其与传统众包有着本质的不同.下文从时空众包的 3 类基本研究对象(即众包任务、众包参与者与众包平台)的视角分别阐述时空众包与传统众包的差异.

- 众包任务的差异.

传统众包任务通常只需参与者在线上完成,所以其一般仅受到在线参与者数量与参与者历史正确率等线上因素的影响.例如,在传统众包平台 AMT 上发布一项图片标注的任务,其完成情况通常依赖于任务发布时的在线参与者数量与参与者的认真态度[10].相较于传统众包任务,时空众包任务不但会受到上述线上因素的影响,且其更多地会受制于各类线下时空属性约束.例如,任务位置附近的人口密度、任务发布时的城市交通情况以及任务自身的实时性需求等.以时下流行的某餐饮派送类时空众包平台上的实时送餐任务为例,若该任务发布于晚高峰时段,且需要将餐饮派送至一个人口稠密的住宅小区,由于该时段交通拥堵而且小区附近订单众多,故该任务通常难以按预计时间送达.因此,不同于传统众包任务一般仅受线上因素的影响,时空众包任务则易受制于各类线下时空属性约束,这也使处理时空众包任务更具有挑战性.

- 众包参与者的差异.

传统众包参与者一般只在线完成众包任务,对其工作地点并无特定要求.而时空众包参与者的工作方式通常需要其在线下移动到指定位置完成任务.以在国内流行的时空众包平台滴滴出行为例,作为众包参与者的专车司机每当被分配任务后,则需要从当前位置驱车至任务指定地点,接到乘客,并将其送达至任务目的地.此外,传统众包与时空众包中众包参与者工作方式的差异也导致了二者的众包参与者男女比例恰好相悖.传统众包中,女性参与者通常多于男性参与者.根据文献[54]对传统众包平台 AMT 的统计分析,该平台众包参与者中女性比例约占 65%,而男性参与者仅为 35%.而时空众包平台中,男性比例远高于女性.根据对美国时空众包平台 Gigwalk[23]的统计分析,该平台中男性参与者高达 71%,而女性参与者仅为 29%[25].文献[25]解释上述差异是因为传统众包的工作方式下参与者无需在物理世界中频繁移动,因此更受到女性参与者的青睐;反之,时空众包通常需要参与者在线下奔赴不同位置来完成任务,从而更适宜男性参与者.因此,时空众包参与者在工作方式与性别比例等方面与传统众包参与者有着本质不同.

- 众包平台的差异.

下文将从众包平台外在的业务类型特征与内在的数据管理技术两个方面进一步比较时空众包平台与传统众包平台的异同.

(1) 服务范围与业务特征.

传统众包平台通常是面向全球众包参与者的,因此该类平台 24 小时内总有在线的众包参与者,并未明显受到具体业务类型与参与者所在地理区域的影响;反之,时空众包平台一般只服务于某一个固定区域范围(例如一个国家或一个城市),而且根据时空众包平台的业务类型,其访问量通常存在明显的周期性规律.例如,出行类的时空众包平台,滴滴出行与 Uber,每日产生的众包任务数量与在线众包参与者数量呈周期性变化.因此,时空众包平台在服务范围与业务类型特征等方面与传统众包平台存在着本质上的不同.

(2) 数据管理技术.

传统众包平台内部的数据管理技术多为传统关系型数据库的查询处理技术,基础操作也多为连接(join)与排序(ranking)等.例如,传统众包平台经常会根据众包参与者的历史数据筛选出最可信的 k 位众包参与者,其本质为执行一个 Top-k 查询.而时空众包平台的数据管理技术则针对于时空数据库,其基础操作通常为最近邻查

询(nearest neighbor query)、最短路径查询(shortest path query)与空间连接查询(spatial join query)等.例如,滴滴出行平台的一项基础操作是为某一个专车请求寻找距其任务位置最近的专车,其本质即为执行一个最近邻查询.因此,时空众包平台与传统众包平台内部的数据管理基础技术大不相同.

综上所述,通过对众包任务、众包参与者与众包平台这 3 类众包研究基础对象的比较可知,传统众包与时空众包存在着本质的差异.表 1 也展示了这两类众包服务在不同维度的区别.因此,针对时空众包上述所独有的特点,其应作为一个单独的主题被提出与研究.下文一方面继续介绍时空众包的代表性应用领域,另一方面也将进一步阐述时空众包研究的核心问题与应用技术.

Table 1 Comparison between traditional crowdsourcing and spatotemporal crowdsourcing

表 1 传统众包与时空众包的对比

众包类型	众包任务		众包参与者		众包平台		
	任务类型	任务约束	工作方式	性别比例	服务范围	业务特征	基础操作
传统众包	线上任务	线上约束	线上完成	女多男少	国际范围	不依赖业务特征	结构化查询
时空众包	线下任务	时空约束	线下移动	男多女少	固定范围	业务特征驱动	时空查询

1.3 时空众包的代表性应用领域

本节将介绍 4 类时空众包数据管理技术的典型应用,从应用实例的视角剖析时空众包技术如何应用于人们的日常生活之中.

- 实时 O2O 应用

共享经济时代典型的互联网+商业模式之一正是O2O商业模式,旨在借助移动互联网技术通过线上招募的方式来整合调度线下的空闲资源,以达到线下空闲资源的高效共享.当前流行的 O2O 应用有实时专车类的滴滴出行[20]、神州专车[21]与 Uber[22];物流派送类的百度外卖[55]、Gigwalk[23]与 TaskRabbit[56]等.以国内互联网巨头百度公司 2014 年提出的“百度外卖”为例,该服务可支持用户发布外卖任务,随后,系统分配某位配送员根据用户任务需求购买餐饮并提供配送.在此类应用中,点餐订单为时空众包任务,而配送员则为时空众包参与者.此外,度量该送餐服务质量的一个重要指标显然是用户的等待时间,而优化配送等待时间这一指标既受制于外卖任务出现的随机性与不确定性,又取决于根据外卖任务的时空分布而采取的配送策略,这也正是时空众包数据管理技术的研究重点.

- 交通管理应用

交通实时路况监控时刻影响着人们的日常出行与生活方式.近年来,随着便携式移动计算设备的普及,基于位置服务的提供商所开发的移动导航类软件,例如国内的百度地图与高德地图,或国外的 Waze[57]等,已可以较为精准地提供实时路况监控信息.该类软件所获得的精准交通监控信息主要源自对其大量用户移动设备中传感器数据的获取与分析.通过获取大量用户在不同时刻的空间分布信息与对应的各类传感器数据,该类软件可分析推测出实时的交通路况.换言之,移动导航类软件在用户使用其软件的同时发布了一项潜在的众包任务,即,分享用户的时空信息与传感器数据,而其用户也被动地成为了众包参与者.此类场景在移动互联网研究中也被称为“参与感知(participatory sensing)”.

- 灾情监控应用

近年来,由于国内外重大自然灾害频发,如海地地震、日本福岛海啸、我国四川雅安地震等,各国政府均高度重视灾后监控,以减轻后续余震等对灾区的影响.因为基础设施破坏严重,灾后的官方监控消息发布与更新通常迟缓与滞后,一些大型社交媒体所开设的灾情信息分享平台反而成为了第一手信息获取的源头.例如,Twitter 在 2010 年 1 月的海地大地震后构建了灾情分享平台,众多灾区人民采用智能手机等移动设备发布自己所在地的第一时间灾情.其实,该分享平台背后隐含着时空众包数据管理技术.换言之,Twitter 分享平台所发布的灾情调查可被视为其发布的一项时空众包任务,任务内容即上报用户所在地灾情,而每位分享灾情的用户充当了众包参与者的角色.若欲对此类应用的监控质量与救援情况作进一步优化,则可根据灾情分享信息的时空特征更有针对性地分配救助小组,这种任务分配技术恰恰就是时空众包数据管理技术的研究重点所在.

- 社交媒体应用

随着 Web 2.0 技术的迅猛发展,各类在线社交媒体(online social medias)如雨后春笋般地不断涌现.特别是随着移动互联网技术的普及,众多社交媒体开始推出一种“基于事件的社交服务[58−62]”.例如,Meetup[63],Plancast[64]和 Whova[65]等.在此类服务中,事件组织者在社交媒体平台上发布社交事件的时空信息(如事件的召集地点与时间等),社交媒体平台会推送不同社交事件给可能的潜在参与者,若参与者规模不足,则事件可能被取消.该类应用也蕴含着时空众包策略的思想,不同社交事件可被视为不同的时空众包任务,而事件的参与者则可被视为时空众包参与者.欲提高该类应用服务质量,如推荐成功率,则需在做推荐决策时不但考虑不同潜在参与者的兴趣偏好,更需考虑社交事件与参与者的时空信息等因素,而这类推送决策的设计正依赖于时空众包数据管理技术的研究.

综上所述,以上 4 类应用的关键挑战都可以被形式化建模为时空众包数据管理中的核心研究问题,因此,时空众包数据管理技术不仅具有重要的理论研究意义,更具有广泛的实际应用价值.

2 时空众包平台

本节首先描述时空众包平台的通用工作流程,随后介绍典型时空众包任务的 3 个基本特性.

2.1 工作流程

时空众包的主要参与者包括众包任务的请求者与众包参与者,他们通过时空众包平台建立联系.如图 1 所示,平台在工作流程中居于中心位置,特将其展开叙述.平台负责对所请求任务和参与者信息进行综合处理.一般地,平台首先将任务/参与者信息进行预处理,然后将其交给任务分配引擎.随后,任务分配引擎基于任务特点和优化目标进行任务分配,并将相应信息反馈给请求者和参与者.根据不同的任务需求,平台既可将任务执行结果直接反馈给请求者,也可对执行结果进行整合汇聚(如图 1 左侧虚线框所示),再反馈给请求者.

下面分别从请求者和参与者视角阐述工作流程.

- 任务请求者工作流程.当请求者打算使用时空众包平台完成任务时,需要依次执行以下步骤:首先,请求者需要设置任务的时空约束,例如派送类任务通常需要设置派送时间和地点等;设置完成后,请求者即可将任务提交到平台;随后,请求者等待平台反馈;
- 众包参与者工作流程.参与者为了完成任务,首先需要提交自己的时空信息.例如,当前所在位置等,以供平台判定其是否满足相关时空约束.在一些平台上,参与者可浏览并自主选择任务,如图 1 右部虚线框所示.随后,参与者等待平台反馈.

以上即为时空众包平台的通用工作流程.

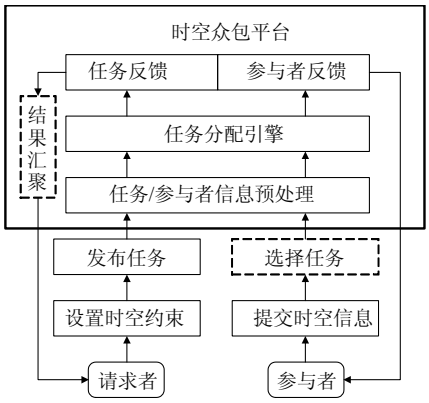


Fig.1 Workflow of spatotemporal crowdsourcing platforms

图 1 时空众包平台的工作流程

2.2 任务特性

本节将从任务的实时性、参与者需求量以及任务选择权这 3 方面介绍不同时空众包应用中任务的独有特性.

2.2.1 任务实时性

根据任务的实时性,可将任务分为静态离线任务和动态在线任务.它们的主要区别在于:动态在线任务中对任务和参与者出现的情况是未知的,参与者与任务随机(或按照某种未知分布)到来,平台只能根据当前参与者和任务的时空分布情况进行分配决策,而无法预知未来参与者和任务的时空分布;对静态离线任务而言,所有的任务和参与者的信息都是已知的,这些信息中没有不确定性,因此往往可以很容易地得到全局最优目标.

以前述物流派送类应用为例,该应用为典型的静态离线任务.通常在一次送货前,派送员(可视为众包参与者)就已经清楚所有的送货地点,因此,平台可为派送员安排出最佳的送货路线.但在实际应用中,通常面对的是动态在线任务,即,任务是实时动态变化的.例如,滴滴出行平台中,出租车司机对于未来可能出现的请求打车服务的用户是无法预测的,因此,平台也只能根据当前时刻已知司机与乘客的信息匹配合理的订单,而当前的匹配策略会对未来订单的匹配结果产生影响,即便当前时刻的匹配策略是最优的,当考虑到下一时刻的订单分布情况时,当前的匹配很可能并不是非常合理的.因此,在各类不同的时空众包应用中,众包任务的实时性特点将在很大程度上影响着众包平台的工作策略.

2.2.2 参与者需求量

根据任务对参与者的需求量,可将任务分为单一参与者需求任务和多参与者需求任务,其区别在于完成任务需要的参与者数量不同.一些任务对众包参与者能力要求较单一,这类任务往往限定仅需一个参与者完成.例如,滴滴出行等专车服务以及最近出现的代驾服务等都是典型的单一参与者需求类任务.另一方面,有一些任务或者由于要求能力种类较多,或者由于任务工作量较大或对工作技能要求较复杂,则需要多个参与者协同完成任务.例如,需要多种能力的参与者共同举办聚会就属于多参与者需求类任务.对于此类多人参与的工作,如何控制协同完成工作的质量并不简单.因此,在各类不同的时空众包应用中,众包任务对参与者需求量的不同将影响到对众包任务完成质量的控制方式.

2.2.3 任务选择权

根据众包参与者是否拥有任务选择权,时空众包任务可分为参与者主动选择任务和平台主动分派任务两种形式.对于参与者主动选择类任务,众包参与者拥有对任务的选择权,其可基于自己的偏好选择适宜的任务.但在很多应用场景中,众包参与者们偏好趋同,这将导致少量性价比高的任务供不应求,反而很多性价比不高的任务无人问津,从而致使时空众包平台的整体效用较低.因此,这也导致了另一类平台主动分配型任务的出现.此类任务仅由时空众包平台根据任务分配算法进行分派,被分配到任务的众包参与者应执行任务,否则会受到平台的相应惩罚.仍以实时专车类应用为例,滴滴出行平台采用了参与者主动选择型任务,其每项专车任务(众包任务)都可由专车司机(众包参与者)选择,当一项任务有多位司机选择时,该平台采用“抢单”策略来保证任务分派的唯一性,但其任务本质为参与者主动选择型任务.与之相反,神州专车与 Uber 平台却采用了平台主动分派型任务,即,每位司机只能被动等待平台分派任务,当任务被分派后,司机需执行此任务,否则将接受惩罚.因此,在各类不同的时空众包应用中,众包参与者对任务的选择权将影响该平台对众包任务的分配机制.

3 时空众包核心研究问题

近年来,关注时空众包数据管理的各类研究大都聚焦于如下 3 个核心问题:任务分配、质量控制与隐私保护.因此,一个很自然的问题就是为什么是这 3 个核心问题?其主要原因来自如下两个方面.

- 一方面,任务分配与质量控制本就是传统众包数据管理中的两个核心问题,其在时空众包环境下更具有研究意义.正如第 1.2 节所述,时空众包与传统众包有着本质的区别,这也导致任务分配与质量控制这两个问题在时空众包环境下的优化目标与约束条件都发生了较大改变.就任务分配问题而言,传统众包研究中的任务分配问题通常只采用离线静态二分图匹配模型加以刻画,但该方法难以适应时空

众包环境下任务实时性的约束和参与者完成多项任务时路径规划的需求.就质量控制问题而言,传统众包研究中质量控制问题通常以最大化单一参与者或参与者群体完成任务的预期正确率作为优化目标,而在一些实时性时空众包应用中质量控制的优化目标变为最小化参与者完成任务所需的时空成本,因此,时空众包环境下的质量控制方式也与传统众包中的质量控制策略存在较大差异;

- 另一方面,时空众包中隐私保护问题是由其真实应用所产生的全新挑战.由于时空众包平台需根据众包任务位置(或众包任务请求者位置)和众包参与者位置信息进行任务分配,因此,任何时空众包平台都存在泄露任务请求者与参与者隐私的潜在风险.仍以实时专车类时空众包平台为例,一旦平台遭受攻击而泄露隐私信息,则每位专车订单请求者与专车司机在过去每日的精准出行信息将被公布于众.但传统众包平台因为无需收集任务请求者与参与者的时空信息而无此类风险.因此,时空众包的隐私保护问题是其独有的核心研究问题,该问题旨在如何设计隐私保护策略,使其既保护任务请求者与众包参与者的时空信息,又可根据保护后的时空信息指导平台进行有效的任务分配.

综上所述,时空众包环境下的任务分配、质量控制与隐私保护这 3 个研究问题彼此间并不独立且相互影响,它们之间的关系如图 2 所示.下文将逐一阐述这 3 个时空众包的核心研究问题.

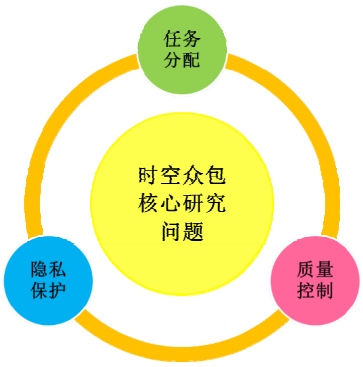


Fig.2 Core research problems of spatial-temporal crowdsourcing

图 2 时空众包的核心研究问题

3.1 任务分配

任务分配问题是指时空众包平台根据众包任务与众包参与者的时空属性和任务特点,为每个任务分配最适合执行的众包参与者.由于不同的时空众包应用对任务分配的需求不同,因此,现存研究通常采用两种不同的算法模型对不同应用场景进行建模:一种建模采用二分图匹配模型,旨在一段时间内为每位众包参与者分配一项众包任务,其典型应用场景为实时专车类服务,如滴滴出行等;另一种建模采取任务规划模型,旨在一段时间内为每位众包参与者分配多项众包任务并规划出执行这些任务的详细顺序与路径,其典型应用场景为物流派送类服务,如百度外卖等.下文将对这两类模型分别展开阐述.

3.1.1 基于匹配的任务分配模型

该类模型通常将实际问题规约为经典的最大化或最小化加权的二分图匹配问题[66].根据不同应用中任务的实时性要求有所不同,又可进一步将此类模型分成静态离线场景的匹配模型与动态在线场景的匹配模型.

- 静态离线匹配模型.此场景下的任务分配问题通常等价于空间匹配(spatial matching)[67,68]问题的一个变种问题,旨在将二分图匹配问题扩展到空间数据中.早期关于时空众包任务分配的研究大都采用这一模型[26,27,37,39],但是由于现实应用中的众包任务与参与者在平台上出现前通常很难提前获知其信息,因此近年来的时空众包任务分配研究大都聚焦于如何建模任务的动态实时特性;
- 动态在线匹配模型.在此应用场景下,每位众包参与者或每项众包任务在分配前均无法获知未来任务与参与者的信息,换言之,仅根据部分二分图信息来进行匹配决策[45,46,69−71].虽然该模型求解不易,但却

很自然地刻画了众包任务的实时性需求.

3.1.2 基于规划的任务分配模型

该类模型适用于对众包参与者在给定时间内要求执行多项众包任务,则众包平台需要为该参与者规划出一份优化的任务执行计划.相似于基于匹配的任务分配模型,现有的基于规划的任务分配模型也可根据任务实时性要求的不同分为静态离线场景的规划模型与动态在线场景的规划模型.

- 静态离线规划模型.此场景下的任务规划问题通常被规约为经典的旅行商问题[72]或者定向问题(orienteering problem)[73],具体而言,即对于某位众包参与者和一个众包任务的集合,在给定众包参与者时空预算成本的条件下,众包平台如何为该参与者规划所执行的任务,从而最大化其所完成任务的数量或者完成任务的效用[28,40,41];
- 动态在线规划模型.有别于静态离线规划模型,在动态在线应用场景下,每当有新任务在众包平台发布时,每位众包参与者需要实时地决定是否将此项新任务加入到其当前的任务规划之中[74].虽然该模型被证明难以提供理论近似保障,但却很自然地刻画了众包任务的实时性需求.

3.2 质量控制

除了上述任务分配问题,质量控制问题也是当前时空众包数据管理研究的一个核心问题[27,75−79].顾名思义,时空众包有两个重要的组成部分:时空与众包.而质量控制的核心问题也由这两个因素引起.

- 一方面,与传统众包数据管理中的质量控制相似,一些时空众包任务也由多位众包参与者重复完成,所以质量控制的关键在于,如何对不同众包参与者反馈进行有效的汇聚从而形成高质量的任务结果.但与传统研究不同的是:时空众包参与者通常还受到空间服务范围的限制,因此在时空众包环境下,基于结果汇聚的质量控制研究通常着重考虑空间服务范围对所汇聚结果的影响;
- 另一方面,许多时空众包任务虽然具有实时性的要求,但是由于众包参与者和众包任务之间存在一定的空间距离,参与者移动到任务所在位置所需时间会直接影响任务的完成质量或任务请求者的满意度.因此,这种由于移动时间引起的延迟也是时空众包质量控制所需研究的新的挑战性问题.

下面分别介绍上述两种时空众包质量控制方法.

3.2.1 基于结果汇聚的质量控制

如上所述,该类研究旨在对不同众包参与者反馈进行有效的汇聚,从而形成高质量的任务结果.所以,任务结果通常会受到如下两类因素的影响:众包参与者的可靠性与众包结果的汇聚机制.由于传统众包数据管理关于质量控制的研究已广泛涉及如何估计众包参与者的可靠性与反馈误差[80−85],因此第 1 类影响因素通常不被视为时空众包环境下质量控制研究的研究热点.此外,在时空众包环境下,众包参与者受到空间服务范围的限制,所以如何有效地集成众包参与者的空间服务范围至众包结果汇聚机制中,成为当前质量控制研究的重点.

具体而言,此类研究通常针对每个众包任务给定一组众包参与者,每位参与者根据其历史表现评估得出其正确完成该任务的可靠性,研究的核心问题旨在为每个众包任务寻找到 n 位众包参与者,并使得至少有一位或$\lceil n/2 \rceil$位参与者能够正确完成该任务的概率满足所给定的概率阈值[27,37].

3.2.2 基于时空约束的质量控制

此外,不同于传统众包数据管理研究通常采用众包参与者集合的可靠性作为质量控制的核心指标,实时性时空众包数据管理通常将完成任务所需的时间开销视为其质量控制的关键指标.例如,对于实时专车服务平台,如滴滴出行,当司机(即众包参与者)接单后,如果需要花费较长时间才能将车驾驶到乘客(即众包任务请求者)所在位置,则该乘客的用户体验通常较差.另外,如果所有的司机所接任务都需要将车开出较远的距离才能接到乘客,那么大量的时间都会被浪费在司机空驶过程中,从而减少司机完成任务的总数量.

文献[46]提出了在线实时任务分配模型,在众包任务随机发布在众包平台的情况下,平台系统需立即决定是将此众包任务分配给某位当前在线的众包参与者还是等待后续抵达平台的众包参与者.一旦平台将众包任务与众包参与者匹配,则匹配关系不可改变.由于在线实时任务匹配的质量主要依赖于众包参与者抵达众包任务位置的时间,即众包任务请求者的等待时间,因此,每项众包任务与其匹配众包参与者之间的时空距离将成为

此类时空众包任务质量控制的最终目标.

综上所述,质量控制与任务分配问题相互影响.质量控制一方面是任务分配问题的重要优化目标,为任务分配提供指导;另一方面,质量控制结果的好坏也依赖于任务分配策略制定得是否合理.因此,两者互相影响且相互依存.

3.3 隐私保护

除了上述任务分配与质量控制两个问题外,隐私保护问题更是时空众包所带来的全新挑战.该问题的目标在于如何既保护众包参与者的时空信息,又可根据众包参与者保护后的时空信息指导其有效地完成任务.

目前,在传统众包数据管理的研究中,隐私保护并未被特别研究.这是因为,在传统众包数据管理中,无论在众包任务的请求者发布任务的过程还是在众包参与者完成任务的环节,人们均无需提供其在真实物理世界中的个人信息,如空间位置信息等.换言之,传统众包数据管理问题并未对现存的隐私保护技术提出新的挑战,因此,隐私保护并非传统众包数据管理研究所关注的问题.然而,时空众包数据管理的研究则不然,由于时空众包平台要求众包任务请求者与参与者提供其在真实物理世界中的时空信息,一旦时空众包平台遭受攻击,时空众包任务请求者与参与者的个人隐私信息将面临泄漏.因此,在时空众包数据管理的研究中,针对新型应用挑战设计合理的隐私保护策略是非常必要的.

当前,面向时空众包数据的隐私保护框架通常如下:首先,众包任务请求者与众包参与者将其敏感的时空信息发送给一个可信的服务器,并在该可信服务器上对其时空信息进行模糊化;随后,该可信服务器将模糊后的信息发送给时空众包平台;最后,该时空众包平台再根据模糊后的信息为任务请求者与参与者提供服务[31,38].

虽然目前已存在一些针对时空数据的隐私保护研究[86],然而现存方法通常采用 k-匿名及 l-多样性等方式以模糊掉用户的位置信息,使用户的位置信息由一个具体的位置点扩大成一个模糊的位置范围,从而实现对用户位置信息的保护.但是,该类方法并不适用于时空众包数据的隐私保护,这是由于,当众包任务与众包参与者的位置被模糊化后,模糊的位置信息会直接影响众包任务与参与者之间的匹配结果,尤其在多人协作和实时任务分配的场景下,位置信息的不准确对任务完成质量的影响非常大.因此,现存研究通常采用差分隐私保护的方法,以参与者在几个不同位置出现的概率来模糊化参与者的精准位置信息[31,38],或在任务分配前需对参与者的空间信息进行隐私保护,而又不可因隐私保护强度过大而导致分配效用过低[87−89].

综上所述,以隐私保护为目的的任务请求者与参与者空间信息模糊化操作会降低任务分配的质量,因此,如何在隐私保护时保障时空众包平台的任务分配质量,是该类问题的研究重点.

4 时空众包数据管理应用技术

本节首先从数据集成、数据查询与数据挖掘这 3 个方面分别介绍各类时空众包数据应用技术,随后对所介绍的应用技术进行深入而完整的对比分析.具体内容如下.

4.1 数据集成

数据集成研究近十几年来一直是数据库与数据挖掘研究领域的核心主题之一,指的是将众多的异构数据源进行有效地清洗、去冗、归并、匹配,且最终将融合后的数据形成统一视图的过程[90].由于数据集成的主要挑战源自异构数据源对同一实体的表述形式不同,而甄别相同语义的不同描述形式正是人类擅长之处,因此,基于众包的数据集成技术成为众包数据管理技术最先涉及的研究主题[91−98].例如,Wang 等人[91,92]研究了基于众包策略的实体同一识别(entity resolution)问题,Zhang 等人[96]讨论了基于众包策略的模式匹配(schema matching)问题,Tong 等人[98]提出了基于众包的数据清洗框架.但是现存的基于众包的数据集成技术通常只针对结构化数据,采用众包数据管理技术对非关系型数据进行集成的研究寥寥无几.特别地,现实生活中有些应用需对时空数据进行集成,例如电子地图数据集成与城市交通数据集成等,但是现有的自动化的时空数据集成方法在数据源获取与集成精准度两个方面都存在较大缺陷.而时空众包数据管理技术一方面可以增加时空数据获取的规模,另一方面也可提高数据集成的精准度,从而为时空数据集成提供了一种高质量的求解新思路.因此,

本节将通过 3 个典型应用进一步阐明时空众包数据管理技术在时空数据集成问题中的作用.

4.1.1 地图数据集成

地图数据是众多基于位置服务(location-based service)的基础.传统的地图数据集成主要通过测绘等手段完成,一方面成本较高,另一方面难以应对真实世界中道路、建筑等变化导致的地图数据更新.Google 公司每年维护地图数据的花费高达 10 亿美金.时空众包数据管理技术为地图数据集成提供了一种新思路,通常称为众包地图(crowdsourced map).开放街道地图(open street map,简称 OSM)[99]是众包地图的典型代表,可视为地图版的维基百科.它通过招募志愿者(众包参与者)对地图进行编辑、标注,实现对地图数据的集成.截止至 2013 年,OSM 已经拥有超过 100 万的注册用户,并收集了 2 100 万英里以上的道路数据和超过 7 800 万条建筑物数据[100].

在地图数据集成过程中,道路类型信息在导航和路径规划方面应用广泛,是不可或缺的重要信息.因此,在众包地图应用中,众包参与者不仅需要标注道路的拓扑结构,还需要对道路类型进行标注.Ding 等人[34]研究利用众包参与者的轨迹数据进行道路类型推断,从而为众包参与者进行道路类型标注提供恰当的候选集,以减轻编辑负担,提高地图数据集成效率.具体而言,文献[34]提出了一种综合考虑路网拓扑结构和轨迹数据的层叠泛化(stacked generalization)推断模型,其集成了基于路网拓扑特征和轨迹特征的逻辑斯蒂回归模型和基于路段联通性的朴素贝叶斯分类器.实验结果表明,该模型能够有效地对道路类型进行推断并提高众包参与者的标注效率.

4.1.2 POI 数据标注

信息点(point of interest,简称 POI)是指人们认为重要或感兴趣的地理坐标,如学校、大型超市、车站等.POI 数据在城市管理、知识发现和基于位置服务等领域都具有广泛的应用价值.例如,Yuan 等人[101]通过综合利用城市不同区域内 POI 信息和人在区域间的移动信息等,推断各区域的主要功能.因此,POI 标注与分类问题也成为时空数据管理与挖掘领域中的重要研究问题.然而,算法自动生成的 POI 标注通常无法保证其质量,易导致较低的标注可用性和较差的用户体验.

为了提高 POI 数据标注的质量,Hu 等人[43]提出了基于时空众包数据管理技术的 POI 数据标注方法.具体而言,该方法将 POI 数据标注问题分解为两个子问题:(1) 如何从不同众包参与者的标注结果中推断出某一 POI 的正确标注;(2) 如何有效地将每个众包任务分配合适的众包参与者以提高推断的准确性.为了解决这两个子问题,文献[43]提出了一个推断模型与一种在线任务分配算法:一方面,推断模型综合众包参与者自身的任务完成质量、众包参与者和 POI 之间的空间距离以及 POI 的影响力等因素,可在众包参与者提交答案后推断出可靠的 POI 标注结果;另一方面,随着众包参与者的动态出现,在线任务分配算法以提高推断效果为目标,将恰当的任务分配给众包参与者.特别地,上述推断模型与任务分配算法交替迭代工作,连续不断地提高 POI 数据标注整体质量.其实验结果表明,结合时空众包数据管理技术的 POI 数据标注方案可明显提高 POI 数据的标注质量.

4.1.3 交通状况监测

除了上述的地图数据集成与 POI 数据标注外,时空众包数据管理技术在数据集成方面的另一个典型应用为交通状况监测,即,对道路交通情况进行实时监测和预估.传统的交通状况监测一般通过在道路上部署传感器来获取车流量与车速等信息,从而监控交通状况.但是由于道路基数大以及道路交通状况复杂等原因,当前所部署的传感器通常无法覆盖全部道路网络,甚至十分稀疏,从而导致当前大多数交通状况监测系统所收集的数据不完备,可用性较差.此外,交通数据具有较强的时效性,也是交通状况监测的另一大挑战.基于上述原因,近年来,时空众包数据管理技术逐渐成为提高交通状况监测质量的一种新策略.

例如,Artikis 等人[33]认为,城市交通状况监测质量较差的根源在于未能从城市中各类异构数据流中获得高质量的数据.如,从位置固定的传感器数据流和位置移动的公交车 GPS 数据流所收集到的数据质量较差,即,存在不准确、不一致、稀疏等问题.因此,文献[33]通过构建时空众包平台来解决数据的不准确和不一致问题.具体而言,时空众包平台通过向距离不准确和不一致数据源较近的多位时空众包参与者进行询问,并对他们的反馈进行汇聚,从而获得较为可靠的结果.尤其是针对数据稀疏问题,文献[33]采用高斯过程回归模型在稀疏数据上对未被传感器覆盖区域的交通状况进行估计,从而构建基于时空众包的人机结合型交通状况监测系统.

除了对城市整体交通状况进行监测,实时车速监测更是交通状况监测的核心内容.然而,现有交通状况监测系统仅能在少数道路上获取粗粒度的车速数据,而无法获得全部道路上细粒度的实时车速信息.为了解决这一问题,Hu 等人[44]采用时空众包数据管理技术对城市各道路的实时车速进行推断估计.具体而言,文献[44]研究了下述问题:若可通过时空众包技术准确获知 k 条道路(称为种子道路)的真实车速,如何利用这些信息推断出整个道路网络中其他道路(称为非种子道路)的实时车速.该问题可进一步划分为两个子问题:(1) 如何准确推断非种子道路的车速;(2) 如何选择种子道路.对于子问题(1),文献[44]提出了有效的种子道路选择算法,并设计了推断模型用来推断非种子道路的车速.具体地,在车速推断方面,由于道路间具有相关性,并且相关联的道路有相似的车速变化趋势,提出了估计车速的两阶段模型.对于子问题(2),文献[44]证明该问题为 NP-难,并提出了具有近似比保障的贪心算法.其实验结果表明,上述采用时空众包数据管理技术的解决方案可将实时车速估计准确度提高约 40%.

4.2 数据查询

4.2.1 空间匹配查询

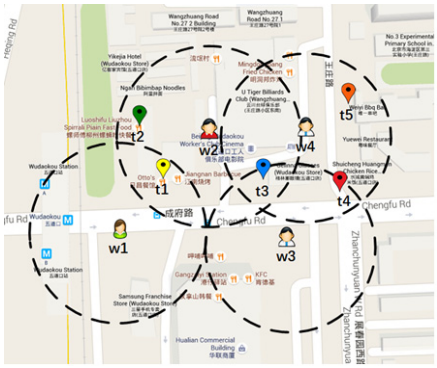
时空众包数据处理以空间匹配查询为核心,首先关注时空众包任务的匹配查询,即:给定一个任务集合与一个众包参与者集合,在满足任务匹配要求的约束条件下优化任务匹配的目标函数.由于不同时空任务匹配查询的优化目标函数与约束条件不同,不同任务匹配问题的建模方式也有所不同.现有的相关研究工作针对不同场景采用了如下 4 种建模方式:最大二分图权值和的匹配模型、最小化二分图权值和的匹配模型、集合覆盖的匹配模型以及基于隐私保护的匹配模型.此外,时空众包数据自身的动态性特征也将产生不同类型的时空众包任务匹配场景,因此在二分图模型中,又可细分为静态离线场景和动态在线场景.

4.2.1.1 最大化二分图权值和匹配模型

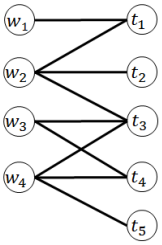
根据不同应用对任务的实时性要求的不同,最大化二分图权值和匹配模型的适用场景又可进一步分为静态离线场景与动态在线场景.

(a) 静态离线场景

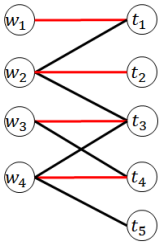
Kazemi 等人[26]率先提出了静态离线场景中的时空众包任务分配查询,即:给定众包任务和众包参与者的空间位置、众包参与者的空间服务范围和计划承担的众包任务数,该查询旨在最大化任务分配总数量.为了求解该查询,文献[26]采用二分图对该查询建模,将众包任务和众包参与者视为二分图中左右两个不相交的点集,将每位众包参与者空间服务范围内的众包任务与该参与者在二分图内构建边,则原查询可规约为最大化二分图匹配规模的问题.如图 3 所示:众包参与者以 w 表示,众包任务以 t 表示.如图 3(a)所示,每位参与者的空间活动范围以虚线圆形所示,对于 w_1,其空间服务范围内的任务只有 t_1,因而在图 3(b)所示二分图中有边(w_1,t_1).如图 3(c)所示,二分图红边代表静态离线场景下的最优解.



(a) 静态离线场景图



(b) 静态二分图



(c) 静态二分图匹配结果

Fig.3 Bipartite matching model

图 3 二分图匹配模型

除静态离线场景外,文献[26]将时间划分为若干时间片,对每个时间片内出现的任务和参与者执行上述静态任务分配,从而对多个时间片支持增量查询.具体而言,文献[26]提出 3 种启发式算法:第 1 种算法以每个时间片下静态离线的二分图为输入,采用传统的最大流算法,累加每个时间片下的最优解;第 2 种算法将众包参与者的位置熵作为二分图的边权,利用传统网络流中最小费用最大流算法得到匹配结果;第 3 种算法选择距离参与者空间位置最近的任务来获得每个时间片下的匹配结果.

基于文献[26],近年来的后续研究通过对众包任务增加“质量约束”“冲突约束”和“时空散度约束”,分别在静态离线场景下提出了如下 3 类最大化二分图权值和匹配查询的变种问题.

- 基于质量约束的任务分配.

文献[26]虽然初步解决了静态离线场景中的时空众包任务分配查询问题,但其忽略了不同参与者对不同任务的完成质量存在差异这一点.因此,Kazemi 等人[27]进一步提出了基于质量约束的任务分配问题以保证任务完成质量.具体而言,文献[27]首先为每位众包参与者定义可靠性的概念以评估其正确完成一个任务的概率;随后,根据少数服从多数的原则继续定义众包任务被正确完成的概率,即:给定 n 个众包参与者执行该任务后,至少 $\lceil n/2 \rceil$ 个参与者能够正确完成该任务的概率.因此,该问题旨在最大化任务分配规模,并同时既满足空间约束又令每个任务被正确完成的概率高于指定的阈值.最终,该问题被证明为 NP-难.文献[27]提出了一系列启发式算法以求解该问题.

- 基于任务冲突约束的任务分配.

考虑到不同众包任务之间可能存在冲突,She 等人[60]提出了基于任务冲突约束的任务分配问题.对于每位众包参与者,若不同时空众包任务间存在冲突,例如同一参与者难以兼顾相距过远的两项任务,因此,众包平台为每位参与者分配的任务不可存在冲突.此外,文献[60]为每位参与者与每项任务标注了一个标签集合,记为一个 d 维向量,故任意一对参与者与任务间的效用值表示为此两个 d 维向量的相似度.该查询旨在满足不同任务间冲突约束条件下最大化全局匹配的效用值,即,将该查询建模为带权二分图匹配模型.由于不同任务间存在冲突,问题被证明为 NP-难.为求解该查询,文献[60]提出了两种具有近似比保障的高效近似算法.

- 基于任务时空散度的任务分配.

在上述基于质量约束的任务分配查询的基础上,Cheng 等人[37]在任务分配过程中不但考虑了众包参与者的可靠性,还融入了完成任务所需众包参与者的时间与空间散度信息.例如,若任务请求者发布了一项拍摄某地标性建筑的时空众包任务,其通常不希望得到许多在相同时段从相同角度对该建筑进行拍摄的照片,而更希望得到不同时段从不同角度拍摄的照片.基于上述考虑,Cheng 等人形式化地建模了基于参与者可靠性和时空散度的任务分配问题,并证明该问题为 NP-难.为解决该问题,Cheng 等人提出了基于贪心、采样和分治的 3 种近似解决方案,并通过建立索引结构提高计算效率.

(b) 动态在线场景

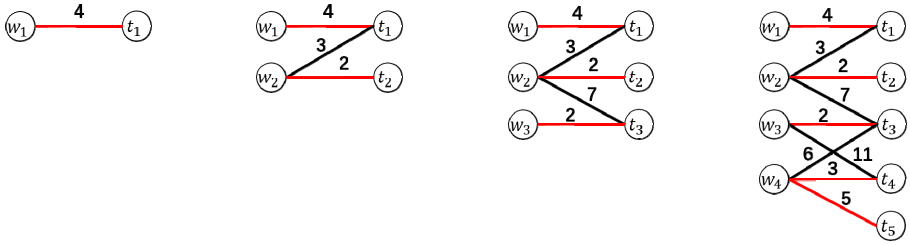
由于许多时空众包的现实应用对众包任务的实时性有较高要求,故众包平台仅可根据当前出现的众包任务与众包参与者执行任务分配,且在分配前无法获知后续众包任务与众包参与者的信息.因此,近年来各类针对动态在线场景空间匹配问题成为研究热点.本节将介绍 3 类动态在线场景中的最大化二分图权值和匹配查询.

- 在线双边动态任务分配.

上述基于静态离线场景任务分配策略需要获知全部任务和参与者信息后方可执行[26],故皆不能满足实时性众包任务的需求.仍以实时专车类服务为例,每项专车订单出现后,系统需立即为其分配专车司机,或告之附近无车可派,而在分配时系统并不知晓后续将出现哪些专车或订单.为了自然且准确地刻画该问题,Tong 等人[45]首次提出了在线双边带权二分图匹配模型.该模型依然将任务与参与者建模为二分图中两个不相交的顶点集,并为每位参与者空间服务范围内的任务与该参与者在二分图内构建带权的边,其权值得自任意函数.特别地:该模型允许任务与参与者以任意顺序动态地出现在二维空间中的任意位置,一旦一项新任务出现,众包平台或立即为该任务分配一个当前未被分配的参与者或令其等待后续的参与者来执行它,直至该任务截止时间为止;反之,对一个新出现的参与者亦然.此外,当一个分配被确定后,则分配结果不可更改.求解此类任务分配问题

的算法被称为在线算法,其算法性能既受制于任务与参与者所构成的二分图结构,又依赖于任务与参与者的抵达顺序.

如图 4 所示,完整的离线二分图结构与图 3 相同,众包参与者 w_1,w_2,w_3 计划承担的任务数量为 1,w_4 计划承担的任务数量为 2.每条边上的数值为边权.



(a) t_1 抵达时匹配结果 (b) w_2 抵达时匹配结果 (c) w_3 抵达时匹配结果 (d) t_5 抵达时匹配结果

Fig.4 Procedure of the greedy algorithm under the first online arrival order

图 4 第 1 种在线抵达顺序的贪心算法匹配过程

表 2 给出两种不同的任务与参与者抵达顺序.若平台采用简单的贪心策略进行分配,即当每项任务或每位参与者出现时为其选取最大权值的边,则在第 1 种顺序下,最终权值和为 16;而在第 2 种顺序下,同样的贪心算法却可获得最优的权值和 27.由此可见,任务与参与者的抵达顺序对算法性能的影响.为了求解该问题,文献[45]分别给出针对最差情况分析与平均情况分析的高效在线算法.

Table 2 Two different online arrival orders

表 2 两种不同的在线抵达顺序

抵达时间	8:00	8:01	8:02	8:07	8:08	8:09	8:09	8:15	8:18
第 1 种顺序	w_1	t_1	t_2	w_2	t_3	w_3	t_4	w_4	t_5
第 2 种顺序	t_1	w_1	t_2	t_3	w_2	t_4	w_3	w_4	t_5

- 基于任务冲突约束的在线任务分配.

针对上述文献[60]在静态离线场景下所提出的基于任务冲突约束的任务分配问题,She 等人[62]扩展该问题模型到动态在线场景之中,进一步提出了基于任务冲突约束的在线任务分配问题.具体而言,该问题假设众包参与者逐一登录众包平台,每当一个新登录的参与者出现在众包平台,该平台将立即为其分配任务,任务一经分配后则不可更改.特别地:与静态离线场景相似,不同众包任务之间存在着一些时空冲突,使得参与者对于某些任务不可能兼得.在满足上述约束条件的基础上,该问题旨在为每位众包参与者分配适当的众包任务,使得众包平台获得最大匹配效用.针对这一问题,文献[62]提出了一种具有竞争比保障的在线算法,并通过大量实验表明,其所提出算法在动态在线场景下具有较好的性能.

- 基于摇臂赌博机的在线动态任务分配.

摇臂赌博机(multi-armed bandit)模型是在线学习(online learning)研究领域中的一个重要模型,该模型假设存在一个多臂赌博机,每摇其中一个臂,就可根据该臂相关的某概率分布获得收益.基于该模型的典型优化问题是最大化有限次摇臂后的总收益.Hassan 等人[36]利用该模型对时空众包中的在线动态任务分配问题进行建模.具体地,众包参与者被视为赌博机的臂,而将某任务分配给某参与者视为一次摇臂.众包参与者接受该任务的概率对应于摇臂收益,该概率受到众包任务和参与者时空属性的影响.每当时空众包任务动态出现时,系统进行一次任务分配,即对应一次摇臂.该问题的优化目标是最大化总任务接受数.为解决该问题,文献[36]扩展了现存的摇臂赌博机求解算法,并提出了适用于时空众包环境的启发式算法.

4.2.1.2 最小化二分图权值和匹配模型

与第 4.2.1.1 节相似,本节也从静态离线场景和动态在线场景分别介绍最小化二分图权值和匹配模型.

- 静态离线场景

Alfarrarjeh 等人[42]采用与文献[26]相同的建模方式,仅将二分图中边权定义为任务与参与者之间的欧式距离,故其优化目标为最小化二分图权值和.其可采用经典的最小费用最大流算法求解,受篇幅所限,在此不赘述.

- 动态在线场景

相较于动态在线最大化二分图权值和匹配模型[45],动态在线最小化二分图权值和匹配模型有着更为广泛的应用空间.仍以实时专车类服务为例,动态在线最小化二分图权值和匹配模型旨在刻画专车平台,令所有乘客在订单被分配后至专车抵达期间的等待时间之和最小,这也正是时空众包研究中的质量控制问题.最近,Tong 等人[46]深入研究了动态在线最小化二分图权值和匹配模型,并发现了一项有悖于该模型过去 25 年研究的新结论.

现存该模型的研究一直认为:贪心算法求解此问题会产生极差的效果,其竞争比为 n 的指数阶,n 为全部任务的数量.这里所谓的贪心算法是指当一位新的参与者出现时,众包平台立即为其分配当前距离该参与者最近且未被分配的任务.Tong 等人证明了基于最差情况分析贪心算法的竞争比虽为 n 的指数阶,但此最差情况发生的概率仅为 $1/n!$.此外,若采用平均情况分析可以发现,最差情况分析理论下的最差实例在平均情况分析下的竞争比仅为 3.195.随后,文献[46]也通过大量真实和仿真实验说明贪心算法实际求解该模型具有很好的执行效果.

4.2.1.3 基于集合覆盖的匹配模型

本节目前所介绍的空间匹配查询均假设针对简单类型的时空众包任务,即,任务仅需要单一众包参与者即可完成.但在实际时空众包应用中存在一些复杂类型的众包任务,需要多位众包参与者组成团队协同完成.具体而言,可在第 1.1 节众包任务与众包参与者的基础定义中分别增添任务技能约束和参与者技能属性,从而令每个众包参与者团队需满足众包任务的技能约束要求方可执行该任务.对于此类由多位众包参与者组队协同完成任务的空间匹配查询,通常采用集合覆盖模型对问题加以刻画,现存的相关研究可分为如下两类.

- 满足集合覆盖约束.

当任务较为复杂、需要多位众包参与者组成团队协同完成时,就自然地涉及到队伍组建(team formation)问题.队伍组建问题来源于集合覆盖问题,最早由 Lappas 等人[102]提出,是指挑选具有不同技能的参与者组成队伍,以满足完成任务的技能需求.文献[102]分别研究了最小化队伍成员组成的社交网络图的直径和最小生成树的权值两种优化问题,并证明该两问题均为 NP-难.对于第 1 个优化问题,提出了近似比为 2 的近似算法;对于第 2 个优化问题,则证明该问题等价于组斯坦纳树(group Steiner tree,简称 GST)问题,并可使用 GST 问题的任一近似算法求解.

不同于上述传统队伍组建问题,在时空众包环境下,众包任务与参与者的时空属性也会对队伍组建产生影响.对该类涉及时空信息的队伍组建问题,Cheng 等人[48]将其定义为多技能时空众包问题:在同时满足众包参与者的空间范围约束、路径长度约束、任务截止时间约束以及覆盖任务所需技能的条件下,将众包任务和参与者团队匹配起来,并最大化总收益.其证明该问题为 NP-难,提出了基于贪心、基于分治和基于成本模型的 3 种启发式算法,并通过实验验证了所提出算法的效率和有效性.

- 最大化覆盖率.

上述队伍组建问题要求在覆盖任务所需全部技能的条件下最小化目标函数,称为完全覆盖问题.该问题的另一变种为在参与者数量不超过某上限的约束下组建队伍,从而最大化覆盖技能的数量,称为最大化覆盖率问题.文献[49]研究了采用时空众包方式进行实时天气预报情境中的最大化覆盖率问题.在该问题中,每位众包参与者可汇报其所在区域的天气情况,众包平台在每个时段选择给定数量的参与者汇报天气状况,并最大化天气情况的汇报范围.文献[49]研究了该问题的两个变种:(1) 给定每个时段的可选择人数上限;(2) 给定全部时段可选人数上限.其证明,上述两个问题在离线静态情形下均为 NP-难.对于第 1 个变种的动态在线情形,给出了 3 种启发式算法;对于第 2 个变种的在线情形,给出了一种具有自适应性的解决策略.人造数据集和真实数据集上的实验,均验证了所提出算法的有效性.

4.2.1.4 基于隐私保护的匹配模型

在传统众包中,众包参与者通过 Web 平台接受和完成任务,并不需要考虑隐私保护问题.而在时空众包中,

众包参与者需将空间位置信息提交至平台,因此存在隐私泄露的风险.位置隐私保护问题在基于位置服务领域已有所研究,一系列方案[86]被提了出来.然而在时空众包中,众包参与者的位置信息同时也是任务分配需考虑的重要因素,因此,如何在保护位置隐私的同时进行有效的任务分配,成为新的挑战课题.下面具体介绍典型的基于隐私保护的匹配模型.

- 基于差分隐私保护模型

时空众包平台要求众包参与者提交其实时位置信息,而众包平台并非一个可完全信任的实体,因此,To 等人[31,38]提出了基于差分隐私的保护框架.众包参与者首先将自身的位置信息提交至可信任的移动服务商,移动服务商将对时空众包平台公布一个索引点.这里,索引点是指位置信息经过了差分隐私转化后,最终存储在索引表上的点.当时空众包平台收到时空众包任务时,根据任务的地理位置信息,在经过转化过的索引上便可以进行范围查询,即,查询时空众包任务附近的众包参与者,最后将合适的众包参与者推送给时空众包任务请求者.

- 基于隐蔽位置保护模型

除了差分隐私保护模型外,同样,为了保护众包参与者的位置信息,Pournajaf 等人[87,88]对所有众包参与者的位置作了假设,认为所有位置均是隐蔽的,确切地说,是只知道众包参与者所在区域及存在于此区域任一点的概率密度函数,但并不清楚具体位置.在这一假设下,尝试最小化所有众包参与者移动至相应时空众包任务的距离加和.该工作提出了一种两阶段的方法:在第 1 阶段,基于隐蔽位置解决一个全局优化问题;在第 2 阶段,允许参与者用自己的精确位置对匹配结果进行调整,以减轻位置的不确定性对第 1 阶段匹配的影响.

4.2.2 最佳路径查询

传统的最佳路径查询通常是指在路网中的最短路径查询(shortest path query),即:在路网中给定起始点和终止点,查找这两点之间距离最短或者时间最短的路径.该问题已经得到广泛研究.根据优化目标的不同,现有研究主要可分为两类:第 1 类为距离最短路径查询,此类研究通常先构建高效的索引,再基于索引设计查询算法[103];第 2 类为时间最短路径查询,由于路网中移动对象的速度依赖于动态变化的交通情况,故此类研究的核心在于如何既实时又准确地预测未来通过每段路径的耗时,其通常使用数据挖掘方法根据历史数据推测未来每段路径的耗时,进而获取时间最短路径.例如,Luo 等人[104]使用频繁模式挖掘,在大量历史数据中寻找到特定时间段内最频繁的路径,以指导时间最短路径查询.

最佳路径查询与最短路径查询相比,除需考虑路径距离与通过时间外,还应考虑其他可能影响交通状况的因素,如路面平坦程度、道路是否施工、车辆拥堵情况、交通信号灯数量、天气情况等.计算机难以处理上述复杂因素,导致很难通过传统建模方法得到高质量的最佳路径查询结果.相比通过计算机算法进行查询,有经验的司机选择的路径可能更为实用.因此,可利用时空众包技术,从司机推荐的路径中选择最佳路径,提高查询质量.例如,Su 等人[30]开发了基于时空众包的路径推荐系统 CrowdPlanner.该系统的主要特点在于:既可快速地将最佳路径查询任务推送给可靠性高且经验丰富的司机(众包参与者),又具有友好且高效的人机交互接口.此外,Zhang 等人[35]同样使用时空众包的方法处理最佳路径查询问题.他们认为:在求解最佳路径选择问题的过程中,由于路网的复杂性,令众包参与者进行两两路径比较的方法并不适用.因此,他们设计了路径查询(routing query)与二分路径查询(binary routing query)这两种问题类型,要求参与者在每个交叉路口选择该去的方向,并提出了一系列有效的算法,动态地管理问题以降低问题的选择难度.

4.2.3 路径规划查询

时空众包环境下,由于众包任务要求参与者满足一定的时空约束,参与者对移动路径和任务完成顺序的选择都将影响任务完成的效率.为解决上述问题,路径规划查询研究如何为众包参与者规划路径和任务完成顺序,以使其在给定时空约束下获取更多收益.Deng 等人[28]为单个众包参与者规划完成多项任务的执行路径,即:在给定一个众包参与者和一系列具有相应空间位置和截止时间约束任务的条件下,为该参与者规划任务完成顺序,最大化该参与者可完成任务的数量.以图 5 为例,给出众包参与者 w_1 及任务 t_1~t_5,任务的空间位置如图所示,并具有截止时间.图中红色路线即为一条可完成任务数最大的路径.Deng 等人证明该问题为 NP-难,并基于动态规划和分支定界两种思路,分别提出两种精确算法.此外,由于精确算法在数据规模较大时运行效率较差,他们

提出了 3 种近似算法以提高运行效率:前两种算法启发式地选择截止时间最早的任务和距离最近的任务;第 3 种算法对基于分支定界思想的精确算法进行修改,通过仅搜索最可能包含最优解的分支以加快算法运行速度.

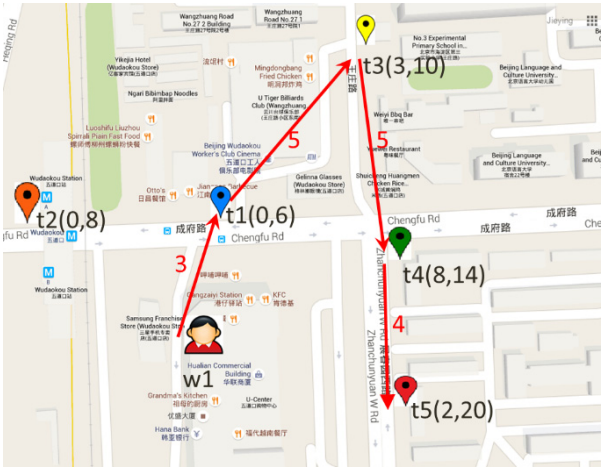


Fig.5 Routing planning example

图 5 路径规划示例

在文献[28]的基础上,Li 等人[74]考虑到实际中任务实时动态出现,研究了单个参与者在线任务规划问题.与文献[28]相比,该问题有两个不同点:(1) 众包任务实时动态出现,只有在其出现后才能被众包参与者获取;(2) 众包参与者完成其任务之后还需按时到达某目的地点.为解决该问题,Li 等人提出了两种启发式算法:GetNextTask 算法贪心地选择众包参与者可到达的任务中使启发式函数 ψ_p 取得最大值的任务;Re-Route 算法在每次任务更新时,基于当前时空信息重新搜索一条最优路径,同时使用剪枝策略缩小路径的搜索空间.

文献[28,74]研究了针对单个众包参与者的路径规划问题;而 She 等人[61]受基于事件社交网络(event-based social network,简称 EBSN)的启发,研究同时为多个众包参与者规划路径,以使得参与者完成时空众包任务的总效用值尽可能地大.She 等人证明该问题为 NP-难,并提出一种两阶段的近似算法.算法第 1 阶段将原问题拆分为若干规模较小的子问题,并利用动态规划予以解决;第 2 阶段归并每个子问题的解,解决其中具有冲突的众包参与者.该算法的近似比为 0.5.Deng 等人[40]研究了相似问题,提出了两个框架,启发式地解决针对多众包参与者的任务分配与路径规划问题.

此外,随着移动设备的普及,利用移动设备进行传感测量正受到广泛关注.移动设备的位置实时地发生变化,而传感测量要求传感器接近测量目标,如何高效地利用移动设备进行传感测量,可视为路径规划查询问题.例如,He 等人[29]提出并形式化定义了众包环境下最大化传感测量平台收益问题,该问题要求在满足移动设备(众包参与者)移动距离和传感任务测量次数约束的条件下,最大化平台收益.其证明该问题为 NP-难,并提出了一种具有常数近似比的近似算法.此外,基于该问题,还提出了一种定价机制,以使平台和移动设备持有者对价格达成一致.

4.3 数据挖掘

一方面,时空众包技术为其他研究领域提供了新思路;另一方面,诸如数据挖掘等其他研究领域的研究成果可用于对时空众包系统进行优化.具体地,通过对时空众包数据进行挖掘,可分析出众包参与者重要的行为模式,用以优化任务分配质量和时空众包系统的用户体验.

对时空众包数据进行挖掘的一个典型实例是实时专车系统中的用户上车地点推荐问题[20,21].在使用滴滴出行等平台呼叫专车时,乘客(任务请求者)与司机(众包参与者)通常难以很快对上车地点达成共识,而需要通过打电话等方式进行协商.并且,若其中一方或双方均对周边环境不熟悉,则可能导致经过长时间沟通依然难以达

成对上车地点的共识,严重影响乘客的用户体验和司机完成任务的效率.解决该问题的一种思路是,利用数据挖掘技术对乘客历史实际上车位置进行挖掘.具体地,可通过聚类或者频繁模式挖掘等方法挖掘出频繁上车地点,并将其作为最佳上车地点推荐的候选集.在任务分配过程中,平台可从候选集中选择一个与用户距离最近或者满足其他优化目标的位置作为最佳上车地点进行推荐.

4.4 对比分析

综上所述,本节从数据集成、数据查询与数据挖掘这 3 方面出发,较详细地介绍了时空众包数据管理的 19 种具体应用技术.为了便于读者区分上述应用技术的异同,本节将从任务实时性、参与者需求量、任务选择权、应用技术的科学问题与优化目标这 5 个维度进一步对其进行深入而完整的对比分析.具体内容见表 3.

Table 3 Comparison on different spatotemporal crowdsourced application techniques

表 3 时空众包应用技术对比

应用技术				文献	任务实时性	参与者需求量	任务选择权	科学问题	优化目标
数据集成	地图数据集成			[34]	动态在线	多位参与者	参与者选择	质量控制	最大化正确率
	POI 数据标注			[43]	动态在线	多位参与者	参与者选择	质量控制	最大化正确率
	交通状况监测			[33,44]	动态在线	多位参与者	平台分派	质量控制	最大化正确率
数据查询	空间匹配查询	最大化二分图权值和匹配模型	基于匹配规模	[26,38]	静态离线	单一参与者	平台分派	任务分配	最大化任务数
			基于质量约束	[27]	静态离线	多位参与者	平台分派	质量控制	最大化匹配数
			基于任务冲突	[60]	静态离线	多位参与者	平台分派	任务分配	最大化效用值
			基于时空散度	[37]	静态离线	多位参与者	平台分派	任务分配	最大化散度值
			基于双边在线	[45]	动态在线	多位参与者	平台分派	任务分配	最大化效用值
			基于动态冲突	[62]	动态在线	多位参与者	平台分派	任务分配	最大化效用值
			基于摇臂赌博机	[36]	动态在线	多位参与者	平台分派	任务分配	最大化接受率
		最小化二分图权值和匹配模型		[42]	静态离线	单一参与者	平台分派	任务分配	最小化等待时间
				[46]	动态在线	单一参与者	平台分派	任务分配	最小化距离和
		基于集合覆盖的匹配模型	集合覆盖约束	[47,48]	静态离线	多位参与者	平台分派	任务分配	最大化效用值
			最大化覆盖率	[49]	静态离线	单一参与者	平台分派	质量控制	最大化覆盖率
		基于隐私保护的匹配模型	基于差分隐私	[31,38]	静态离线	多位参与者	平台分派	隐私保护	最大化效用值
			基于隐蔽位置	[87,88]	静态离线	多位参与者	平台分派	隐私保护	最大化效用值
	最佳路径查询			[30,35]	动态在线	多位参与者	平台分派	质量控制	最大化正确率
	路径规划查询	单一参与者规划		[28,41]	静态离线	单一参与者	平台分派	任务分配	最大化任务数
				[74]	动态在线	单一参与者	平台分派	任务分配	最大化效用值
		全体参与者规划		[29,40]	静态离线	多位参与者	平台分派	任务分配	最大化任务数
挖掘数据	任务最优执行地点推荐			[20,21]	动态在线	多位参与者	平台分派	质量控制	最小化等待时间

5 时空众包未来研究方向

目前,时空众包数据管理技术作为一个新型的研究领域,还有很多研究方向值得学者们深入探究.下面简述其中 4 类潜在的研究方向,供后续研究者们参考.

(1) 时空众包数据的建模问题.

现有工作中,对时空众包的位置信息(众包参与者和请求者的所在地)均采用网格坐标的方式进行建模,且众包参与者在空间的移动方式也只简单地建模为直线移动,这并不符合现实生活中众包参与者真实的应用场景.因此,如何利用路网来建模位置信息及参与者移动方式,是未来建立时空众包数据模型的一个挑战性问题.

(2) 时空众包的存储与索引问题.

由于时空众包数据自身包含动态的时空数据、高维属性数据与时空冲突数据,故传统离线静态场景中的时空数据查询索引技术并不适用于该类问题[105−109].因此,如何对时空众包数据进行有效的存储与索引,进而支持各类时空众包数据查询处理,是未来研究的关键.因此,一个极具潜力的研究问题是设计一种针对时空众包数据自身特性的存储策略与通用性强的索引结构.

(3) 时空众包的激励机制问题.

相似于传统众包数据管理技术的研究,激励机制也将成为未来时空众包数据管理的重要研究问题之一[110−113].对时空众包任务分配的一种潜在影响因素是对不同时空众包任务的最优激励策略,试想:在“百度外卖”类型的送餐应用中,对于居住在较为偏远处的点餐者(即时空众包任务的发布者)通常不得不多付一些费用方可得到派送服务,此应用实例中隐藏着对时空众包任务最优定价对任务分配的影响.因此,如何在任务分配过程中进行最佳定价,是未来非常值得研究的问题.

(4) 时空众包在社交网络中的数据管理问题.

某些数据清洗与集成的时空众包应用,如高德推出的“道路寻宝”等,要求用户到指定地点拍摄符合要求的图片并发送到应用平台中.有些众包参与者在完成这些任务的过程中更喜欢与朋友结伴而行.因此,在进行时空众包任务分配时,若综合考虑不同众包参与者之间的社交关系,将会提升众包参与者使用时空众包平台的满意度和用户体验[114,115].而现有工作中并未对众包参与者之间的社交关系进行系统性研究,这也是时空众包数据管理未来需探索的方向之一.

6 结束语

本文主要阐述了时空众包数据管理技术的研究进展,不但介绍了时空众包的基本概念、典型应用及其工作流程,还从时空众包数据管理的 3 个核心研究问题(任务分配、质量控制和隐私保护)和 3 类应用技术(基于时空众包的数据清洗、数据查询与数据挖掘)这两个视角对该研究领域进行了深入且全面的综述.特别针对 3 类 19 种时空众包应用技术进行了 5 个维度的比较分析,并针对目前的研究状况给出了未来值得探讨的研究方向.具体而言,未来的时空众包数据管理研究在数据模型、存储索引、激励机制与社交关系等方向都值得进一步深入研究.随着移动互联网技术与共享经济模式的快速发展,时空众包数据管理技术作为各类 O2O 应用的一种新型通用计算范式正受到学术界与产业界的双重关注.希望本文所做的工作可以为致力于从事时空众包数据管理的相关研究人员提供参考.

References:

[1] Law E, Ahn L. Human computation. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2011,5(3):1−121.

[2] Howe J. Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business. Crown Business, 2009.

[3] Amazon Mechanical Turks(AMT). https://www.mturk.com/mturk/

[4] CrowdFlower. http://www.crowdflower.com/

[5] oDesk. http://www.odesk.com/

[6] 林琳.众包模式,大势所趋.人民日报,2014-4-4.

[7] Doan A, Franklin MJ, Kossmann D, Kraska T. Crowdsourcing applications and platforms: A data management perspective. Proc. of the VLDB Endowment, 2011,5(11):1495−1506.

[8] Deutch D, Milo T. Mob data sourcing. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. 2012. 581−584. [doi: 10.1145/2213836.2213905]

[9] Chen L, Lee D, Zhang M. Crowdsourcing in information and knowledge management. In: Li JZ, Wang XS, Garofalakis MN, *et al*., eds. Proc. of the 23rd ACM Int'l Conf. on Information and Knowledge Management. Shanghai: ACM, 2014. http://dblp.uni-trier.de/rec/bibtex/conf/cikm/2014

[10] Chen L, Lee D, Milo T. Data-Driven crowdsourcing: Management, mining and applications. In: Proc. of the 31st Int'l Conf. on Data Engineering. 2015. 1527−1529. [doi: 10.1109/ICDE.2015.7113418]

[11] Parameswaran AG, Garcia-Molina H, Park H, Polyzotis N, Ramesh A, Widom J. Crowdscreen: Algorithms for filtering data with humans. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. 2012. 361−372. [doi: 10.1145/2213836.2213878]

[12] Parameswaran AG, Boyd S, Garcia-Molina H, Gupta A, Polyzotis N, Widom J. Optimal crowd-powered rating and filtering algorithms. Proc. of the VLDB Endowment, 2014,7(9):685−696. [doi: 10.14778/2732939.2732942]

[13] Marcus A, Wu E, Karger D, Madden S, Miller R. Human-Powered sorts and joins. Proc. of the VLDB Endowment, 2011,5(1):13−24. [doi: 10.14778/2047485.2047487]

[14] Guo S, Parameswaran AG, Garcia-Molina H. So who won? Dynamic max discovery with the crowd. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. 2012. 385−396. [doi: 10.1145/2213836.2213880]

[15] Venetis P, Garcia-Molina H, Huang K, Polyzotis N. Max algorithms in crowdsourcing environments. In: Proc. of the 21st World Wide Web Conf. 2012. 989−998. [doi: 10.1145/2187836.2187969]

[16] Davidson S, Khanna S, Milo T, Roy S. Using the crowd for top-*k* and group-by queries. In: Proc. of the 16th Int'l Conf. on Database Theory. 2013. 225−236. [doi: 10.1145/2448496.2448524]

[17] Marcus A, Karger D, Madden S, Miller R, Oh S. Counting with the crowd. Proc. of the VLDB Endowment, 2012,6(2):109−120. [doi: 10.14778/2535568.2448944]

[18] Franklin MJ, Kossmann D, Kraska T, Ramesh S, Xin R. CrowdDB: Answering queries with crowdsourcing. In: Proc. of the 2011 ACM SIGMOD Int'l Conf. on Management of Data. 2011. 61−72. [doi: 10.1145/1989323.1989331]

[19] Park H, Pang R, Parameswaran AG, Garcia-Molina H, Polyzotis N, Widom J. Deco: A system for declarative crowdsourcing. Proc. of the VLDB Endowment, 2012,5(12):1990−1993. [doi: 10.14778/2367502.2367555]

[20] DiDi. http://www.xiaojukeji.com/index/index

[21] Shenzhou Taxi. http://www.10101111.com/

[22] Uber. https://www.uber.com/

[23] Gigwalk. http://www.gigwalk.com/

[24] DaoLuXunBao. http://xunbao.amap.com/

[25] Musthag M, Ganesan D. Labor dynamics in a mobile micro-task market. In: Proc. of the 2013 ACM SIGCHI Conf. on Human Factors in Computing Systems. 2013. 641−650. [doi: 10.1145/2470654.2470745]

[26] Kazemi L, Shahabi C. Geocrowd: Enabling query answering with spatial crowdsourcing. In: Proc. of the SIGSPATIAL 2012 Int'l Conf. on Advances in Geographic Information Systems. 2012. 189−198. [doi: 10.1145/2424321.2424346]

[27] Kazemi L, Shahabi C, Chen L. GeoTruCrowd: Trustworthy query answering with spatial crowdsourcing. In: Proc. of the SIGSPATIAL 2013 Int'l Conf. on Advances in Geographic Information Systems. 2013. 304−313. [doi: 10.1145/2525314.2525346]

[28] Deng D, Shahabi C, Demiryurek U. Maximizing the number of worker's self-selected tasks in spatial crowdsourcing. In: Proc. of the SIGSPATIAL 2012 Int'l Conf. on Advances in Geographic Information Systems. 2013. 314−323. [doi: 10.1145/2525314.2525370]

[29] He SB, Shin D, Zhang JS, Chen JM. Toward optimal allocation of location dependent tasks in crowdsensing. In: Proc. of the 2014 IEEE Conf. on Computer Communications. 2014. 745−753. [doi: 10.1109/INFOCOM.2014.6848001]

[30] Su H, Zheng K, Huang JM, Jeung H, Chen L, Zhou XF. CrowdPlanner: A crowd-based route recommendation system. In: Proc. of the 30th Int'l Conf. on Data Engineering. 2014. 1144−1155. [doi: 10.1109/ICDE.2014.6816730]

[31] To H, Ghinita G, Shahabi C. A framework for protecting worker location privacy in spatial crowdsourcing. Proc. of the VLDB Endowment, 2014,7(10):919−930. [doi: 10.14778/2732951.2732966]

[32] Chen Z, Fu R, Zhao ZY, Liu Z, Xia LH, Chen L, Cheng P, Cao CC, Tong YX, Zhang CJ. gMission: A general spatial crowdsourcing platform. Proc. of the VLDB Endowment, 2014,7(13):1629−1632. [doi: 10.14778/2733004.2733047]

[33] Artikis A, Weidlich M, Schnitzler F, Boutsis I, Liebig T, Piatkowski N, Bockermann C, Morik K, Kalogeraki V, Marecek J, Gal A, Mannor S, Gunopulos D, Kinane D. Heterogeneous stream processing and crowdsourcing for urban traffic management. In: Amer-Yahia S, Christophides V, Kementsietsidis A, *et al*., eds. Proc. of the 17th Int'l Conf. on Extending Database Technology. Athens, 2014. 712−723. http://dblp.uni-trier.de/rec/bibtex/conf/edbt/ArtikisWSBLPBMKMGMGK14

[34] Ding Y, Zheng J, Tan H, Luo W, Ni LM. Inferring road type in crowdsourced map services. In: Proc. of the 19th Int'l Conf. on Database Systems for Advanced Applications. 2014. 392−406. [doi: 10.1007/978-3-319-05813-9_26]

[35] Zhang CJ, Tong YX, Chen L. Where to: Crowd-Aided path selection. Proc. of the VLDB Endowment, 2014,7(14):2005−2016. [doi: 10.14778/2733085.2733105]

[36] Hassan U, Curry E. A multi-armed bandit approach to online spatial task assignment. In: Proc. of the 14th Int'l Conf. on Ubiquitous Intelligence and Computing. 2014. 212−219. [doi: 10.1109/UIC-ATC-ScalCom.2014.68]

[37] Cheng P, Lian X, Chen Z, Fu R, Chen L, Han JS, Zhao JZ. Reliable diversity-based spatial crowdsourcing by moving workers. Proc. of the VLDB Endowment, 2015,8(10):1022−1033. [doi: 10.14778/2794367.2794372]

[38] To H, Ghinita G, Shahabi C. PrivGeoCrowd: A toolbox for studying private spatial crowdsourcing. In: Proc. of the 31st Int'l Conf. on Data Engineering. 2015. 1404−1407. [doi: 10.1109/ICDE.2015.7113387]

[39] To H, Shahabi C, Kazemi L. A server-assigned spatial crowdsourcing framework. ACM Trans. on Spatial Algorithms and Systems, 2015,1(1):2. [doi: 10.1145/2729713]

[40] Deng D, Shahabi C, Zhu L. Task matching and scheduling for multiple workers in spatial crowdsourcing. In: Proc. of the SIGSPATIAL 2015 Int'l Conf. on Advances in Geographic Information Systems. 2015. 21. [doi: 10.1145/2820783.2820831]

[41] Deng D, Shahabi C, Demiryurek U, Zhu L. Task selection in spatial crowdsourcing from worker's perspective. GeoInformatica, 2016,20(3):529−568. [doi: 10.1007/s10707-016-0251-4]

[42] Alfarrarjeh A, Emrich T, Shahabi C. Scalable spatial crowdsourcing: A study of distributed algorithms. In: Proc. of the 16th IEEE Int'l Conf. on Mobile Data Management. 2015. 134−144. [doi: 10.1109/MDM.2015.55]

[43] Hu H, Zheng Y, Bao Z, Li G, Feng J, Cheng R. Crowdsourced POI labelling: Location-Aware result inference and task assignment. In: Proc. of the 32nd Int'l Conf. on Data Engineering. 2016. 61−72. [doi: 10.1109/ICDE.2016.7498229]

[44] Hu H, Li G, Bao Z, Cui Y, Feng J. Crowdsourcing-Based real-time urban traffic speed estimation: From trends to speeds. In: Proc. of the 32nd Int'l Conf. on Data Engineering. 2016. 883−894. [doi: 10.1109/ICDE.2016.7498298]

[45] Tong YX, She JY, Ding BL, Wang LB, Chen L. Online mobile micro-task allocation in spatial crowdsourcing. In: Proc. of the 32nd Int'l Conf. on Data Engineering. 2016. 49−60. [doi: 10.1109/ICDE.2016.7498228]

[46] Tong YX, She JY, Ding BL, Chen L, Wo TY, Xu K. Online minimum matching in real-time spatial data: Experiments and analysis. Proc. of the VLDB Endowment, 2016,9(12):1053−1064. [doi: 10.14778/2994509.2994523]

[47] Gao DW, Tong YX, She JY, Song TS, Chen L, Xu K. Top-*k* team recommendation in spatial crowdsourcing. In: Proc. of the 17th Int'l Conf. on Web-Age Information Management. 2016. 191−204. [doi: 10.1007/978-3-319-39937-9_15]

[48] Cheng P, Lian X, Chen L, Han J, Zhao J. Task assignment on multi-skill oriented spatial crowdsourcing. IEEE Trans. on Knowledge and Data Engineering, 2016,28(8):2201−2215. [doi: 10.1109/TKDE.2016.2550041]

[49] To H, Fan L, Tran L, Shahabi C. Real-Time task assignment in hyperlocal spatial crowdsourcing under budget constraints. In: Proc. of the 17th Int'l Conf. on Pervasive Computing and Communications. 2016. 1−8. [doi: 10.1109/PERCOM.2016.7456507]

[50] Li GL, Wang JN, Zheng YD, Franklin JM. Crowdsourced data management: A survey. ACM Trans. on Knowledge and Data Engineering, 2016,28(9):2296−2319. [doi: 10.1109/TKDE.2016.2535242]

[51] Chittilappilly AI, Chen L, Amer-Yahia S. A survey of general-purpose crowdsourcing techniques. ACM Trans. on Knowledge and Data Engineering, 2016,28(9):2246−2266. [doi: 10.1109/TKDE.2016.2555805]

[52] Garcia-Molina H, Joglekar M, Marcus A, Parameswaran AG, Verroios V. Challenges in data crowdsourcing. ACM Trans. on Knowledge and Data Engineering, 2016,28(4):901−911. [doi: 10.1109/TKDE.2016.2518669]

[53] Feng JH, Li GL, Feng JH. A survey on crowdsourcing. Chinese Journal of Computers, 2015,9:1713−1726 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2015.01713]

[54] Ipeirotis P. Demographics of mechanical turk. Social Science Electronic Publishing, 2012,35:119

[55] Baidu Waimai. http://waimai.baidu.com/

[56] TaskRabbit. http://www.taskrabbit.com/

[57] Waze. https://www.waze.com/

[58] Liu XJ, He Q, Tian YY, Lee W, McPherson J, Han JW. Event-Based social networks: Linking the online and offline social worlds. In: Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2012. 1032−1040. [doi: 10.1145/2339530.2339693]

[59] Li KQ, Lu W, Bhagat S, Lakshmanan LVS, Yu C. On social event organization. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2014. 1206−1215. [doi: 10.1145/2623330.2623724]

[60] She JY, Tong YX, Chen L, Cao CC. Conflict-Aware event-participant arrangement. In: Proc. of the 31st Int'l Conf. on Data Engineering. 2015. 735−746. [doi: 10.1109/ICDE.2015.7113329]

[61] She JY, Tong YX, Chen L. Utility-Aware event-participant planning. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. 2015. 1629−1643. [doi: 10.1145/2723372.2749446]

[62] She JY, Tong YX, Chen L, Cao CC. Conflict-Aware event-participant arrangement and its variant for online setting. IEEE Trans. on Knowledge and Data Engineering, 2016,28(9):2281−2295. [doi: 10.1109/TKDE.2016.2565468]

[63] Meetup. http://www.meetup.com/

[64] Plancast. http://plancast.com

[65] Whova. https://whova.com/

[66] Burkard RE, Dell'Amico M, Martello S. Assignment Problems. Philadelphia: Society for Industrial and Applied Mathematics, 2009.

[67] Wong RC, Tao YF, Fu AW, Xiao XK. On efficient spatial matching. In: Koch C, Gehrke J, Garofalakis MN, *et al*., eds. Proc. of the 33rd Int'l Conf. on Very Large Data Bases. ACM Press, 2007. 579−590.

[68] U LH, Yiu ML, Mouratidis K, Mamoulis N. Capacity constrained assignment in spatial databases. In: Proc. of the 27th Int'l Conf. on Management of Data. Vancouver: ACM Press, 2008. 15−28. [doi: 10.1145/1376616.1376621]

[69] Ho C, Vaughan J. Online task assignment in crowdsourcing markets. In: Hoffmann J, Selman B, eds. Proc. of the 26th AAAI Conf. on Artificial Intelligence. Toronto: AAAI Press, 2012. 45−51. http://dblp.uni-trier.de/rec/bibtex/conf/aaai/2012

[70] Ho C, Jabbari S, Vaughan J. Adaptive task assignment for crowdsourced classification. In: Proc. of the 30th Int'l Conf. on Machine Learning. Atlanta, 2013. 534−542. http://dblp.uni-trier.de/rec/bibtex/conf/icml/2013

[71] Karger D, Oh S, Shah D. Budget-Optimal task allocation for reliable crowdsourcing systems. Operations Research, 2014,62(1):1−24. [doi: 10.1287/opre.2013.1235]

[72] Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. New York: W. H. Freeman, 1979.

[73] Vansteenwegen P, Souffriau W, Oudheusden DV. The orienteering problem: A survey. European Journal of Operational Research, 2011,209(1):1−10. [doi: 10.1016/j.ejor.2010.03.045]

[74] Li Y, Yiu ML, Xu W. Oriented online route recommendation for spatial crowdsourcing task workers. In: Proc. of the 14th Int'l Conf. on Advances in Spatial and Temporal Databases. 2015. 137−156. [doi: 10.1007/978-3-319-22363-6_8]

[75] Ipeirotis P, Provost F, Wang J. Quality management on amazon mechanical turk. In: Proc. of the ACM SIGKDD Workshop on Human Computation. 2010. 64−67. [doi: 10.1145/1837885.1837906]

[76] Liu X, Lu M, Ooi B, Shen Y, Wu S, Zhang M. CDAS: A crowdsourcing data analytics system. Proc. of the VLDB Endowment, 2012,5(11):1495−1506. [doi: 10.14778/2350229.2350264]

[77] Cao CC, She JY, Tong YX, Chen L. Whom to ask? Jury selection for decision making tasks on micro-blog services. Proc. of the VLDB Endowment, 2012,5(11):1495−1506. [doi: 10.14778/2350229.2350264]

[78] Cao CC, Tong YX, Chen L, Jagadish H. WiseMarket: A new paradigm for managing wisdom of online social users. In: Proc. of the 19th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. 2013. 455−463. [doi: 10.1145/2487575.2487642]

[79] Gao J, Liu X, Ooi B, Wang H, Chen G. An online cost sensitive decision-making method in crowdsourcing systems. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. 2013. 217−228. [doi: 10.1145/2463676.2465307]

[80] Whitehill J, Ruvolo P, Wu T, Bergsma J, Movellan J. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: Bengio Y, Schuurmans D, Lafferty JD, *et al*., eds. Proc. of the 23rd Annual Conf. on Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2009. 2035−2043. http://dblp.uni-trier.de/rec/bibtex/conf/nips/2009

[81] Raykar V, Yu S. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. Journal of Machine Learning Research, 2012,13:491−518.

[82] Sheng V, Provost F, Ipeirotis P. Get another label? Improving data quality and data mining using multiple, noisy labelers. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2008. 614−622. [doi: 10.1145/1401890.1401965]

[83] Dalvi N, Dasgupta A, Kumar R, Rastogi V. Aggregating crowdsourced binary ratings. In: Proc. of the 22nd Int'l World Wide Web Conf. 2013. 285−294. [doi: 10.1145/2488388.2488414]

[84] Karger D, Oh S, Shah D. Efficient crowdsourcing for multi-class labeling. In: Proc. of the ACM SIGMETRICS Int'l Conf. on Measurement and Modeling of Computer Systems. 2013. 81−92. [doi: 10.1145/2465529.2465761]

[85] Joglekar M, Garcia-Molina H, Parameswaran AG. Evaluating the crowd with confidence. In: Proc. of the 19th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. 2013. 686−694. [doi: 10.1145/2487575.2487595]

[86] Ghinita G, Kalnis P, Khoshgozaran A, Shahabi C, Tan KL. Private queries in location based services: Anonymizers are not necessary. In: Proc. of the 2008 ACM SIGMOD Int'l Conf. on Management of Data. 2008. 121−132. [doi: 10.1145/1376616.1376631]

[87] Pournajaf L, Xiong L, Sunderam VS, Goryczka S. Spatial task assignment for crowd sensing with cloaked locations. In: Proc. of the 15th Int'l Conf. on Pervasive Computing and Communications. 2014. 73−82. [doi: 10.1109/MDM.2014.15]

[88] Pournajaf L, Xiong L, Sunderam VS, Goryczka S, Xu XF. STAC: Spatial task assignment for crowd sensing with cloaked participant locations. In: Proc. of the SIGSPATIAL 2015 Int'l Conf. on Advances in Geographic Information Systems. 2015. 90. [doi: 10.1145/2820783.2820788]

[89] Pournajaf L, Garcia-Ulloa DA, Xiong L, Sunderam VS. Participant privacy in mobile crowd sensing task management: A survey of methods and challenges. SIGMOD Record, 2015,44(4):23−34. [doi: 10.1145/2935694.2935700]

[90] Doan A, Halevy A, Ives Z. Principles of Data Integration. San Francisco: Morgan Kaufmann Publishers, 2012.

[91] Wang J, Kraska T, Franklin MJ, Feng J. CrowdER: Crowdsourcing entity resolution. Proc. of the VLDB Endowment, 2012,5(11): 1483−1494. [doi: 10.14778/2350229.2350263]

[92] Wang J, Li G, Kraska T, Franklin MJ, Feng J. Leveraging transitive relations for crowdsourced joins. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. 2013. 229−240. [doi: 10.1145/2463676.2465280]

[93] Vesdapunt N, Bellare K, Dalvi N. Crowdsourcing algorithms for entity resolution. Proc. of the VLDB Endowment, 2014,7(12): 1071−1082. [doi: 10.14778/2732977.2732982]

[94] Whang S, Lofgren P, Garcia-Molina H. Question selection for crowd entity resolution. Proc. of the VLDB Endowment, 2013,6(6): 349−360. [doi: 10.14778/2536336.2536337]

[95] Gokhale C, Das S, Doan A, Naughton J, Rampalli N, Shavlik J, Zhu X. Corleone: Hands-Off crowdsourcing for entity matching. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. 2014. 601−612. [doi: 10.1145/2588555.2588576]

[96] Zhang C, Chen L, Jagadish H, Cao CC. Reducing uncertainty of schema matching via crowdsourcing. Proc. of the VLDB Endowment, 2013,6(9):757−768. [doi: 10.14778/2536360.2536374]

[97] Fan J, Lu M, Ooi B, Tan W, Zhang M. A hybrid machine-crowdsourcing system for matching Web tables. In: Proc. of the 30th Int'l Conf. on Data Engineering. 2014. 976−987. [doi: 10.1109/ICDE.2014.6816716]

[98] Tong YX, Cao CC, Zhang CJ, Li Y, Chen L. Crowdcleaner: Data cleaning for multi-version data on the Web via crowdsourcing. In: Proc. of the 30th Int'l Conf. on Data Engineering. 2014. 1182−1185. [doi: 10.1109/ICDE.2014.6816736]

[99] Haklay M, Weber P. OpenStreetMap: User-Generated street maps. IEEE Pervasive Computing, 2008,7(4):12−18. [doi: 10.1109/MPRV.2008.80]

[100] 2013 OpenStreetMap Data Report. https://www.mapbox.com/osm-data-report/

[101] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs. In: Proc. of the 18th Int'l Conf. on Knowledge Discovery and Data Mining. 2014. 186−194. [doi: 10.1145/2339530.2339561]

[102] Lappas T, Liu K, Terzi E. Finding a team of experts in social networks. In: Proc. of the 15th Int'l Conf. on Knowledge Discovery and Data Mining. 2009. 467−476. [doi: 10.1145/1557019.1557074]

[103] Wu L, Xiao X, Deng D, Cong G, Zhu A, Zhou S. Shortest path and distance queries on road networks: An experimental evaluation. Proc. of the VLDB Endowment, 2012,5(5):406−417. [doi: 10.14778/2140436.2140438]

[104] Luo W, Tan H, Chen L, Ni LM. Finding time period-based most frequent path in big trajectory data. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. 2013. 713−724. [doi: 10.1145/2463676.2465287]

[105] Zhou AY, Yang S, Jin CQ, Ma Q. Location-Based services: Architecture and progress. Chinese Journal of Computers, 2011,34(7): 1155−1171 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2011.01155]

[106] Hu J, Fan J, Li GL, Chen SS. Top-*k* fuzzy spatial keyword search. Chinese Journal of Computers, 2012,35(11):2237−2246 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2012.02237]

[107] Liu XP, Wang CX, Liu DX, Liao GQ. Survey on spatial keyword search. Ruan Jian Xue Bao/Journal of Software, 2016,27(2): 329−347 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4934.htm [doi: 10.13328/j.cnki.jos.004934]

[108] Wang JB, Gao H, Li JZ, Yang DH. An index supporting spatial approximate keyword search on disks. Journal of Computer Research and Development, 2012,49(10):2142−2152 (in Chinese with English abstract).

[109] Dai J, Xu JJ, Liu KE, Wu B, Ding ZM. DKR-Tree: A dynamic-keyword-R tree. Journal of Computer Research and Development, 2013,50(S1):163−170 (in Chinese with English abstract).

[110] Singer Y, Mittal M. Pricing mechanisms for crowdsourcing markets. In: Proc. of the 22nd Int'l World Wide Web Conf. 2013. 1157−1166. [doi: 10.1145/2488388.2488489]

[111] Singla A, Krause A. Truthful incentives in crowdsourcing tasks using regret minimization mechanisms. In: Proc. of the 22nd Int'l World Wide Web Conf. 2013. 1167−1178. [doi: 10.1145/2488388.2488490]

[112] Gao Y, Parameswaran AG. Finish them!: Pricing algorithms for human computation. Proc. of the VLDB Endowment, 2014,7(14): 1965−1976. [doi: 10.14778/2733085.2733101]

[113] Wu Y, Zeng JR, Peng H, Chen H, Li CP. Survey on incentive mechanisms for crowd sensing. Ruan Jian Xue Bao/Journal of Software, 2016,27(8):2025−2047 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5049.htm [doi: 10.13328/j.cnki.jos.005049]

[114] Tong YX, She JY, Meng R. Bottleneck-Aware arrangement over event-based social networks: The max-min approach. World Wide Web Journal, 2016,19(6):1151−1177. [doi: 10.1007/s11280-015-0377-6]

[115] Tong YX, Cao CC, Chen L. TCS: Efficient topic discovery over crowd-oriented service data. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2014. 861−870. [doi: 10.1145/2623330.2623647]

附中文参考文献:

[53] 冯剑红,李国良,冯建华.众包技术研究综述.计算机学报,2015,9:1713−1726.

[105] 周傲英,杨彬,金澈清,马强.基于位置的服务:架构与进展.计算机学报,2011,34(7):1155−1171.

[106] 胡骏,范举,李国良,陈珊珊.空间数据上 Top-*k* 关键词模糊查询算法.计算机学报,2012,35(11):2237−2246.

[107] 刘喜平,万常选,刘德喜,廖国琼.空间关键词搜索研究综述.软件学报,2016,27(2):329−347. http://www.jos.org.cn/1000-9825/4934.htm [doi: 10.13328/j.cnki.jos.004934]

[108] 王金宝,高宏,李建中,杨东华.RB 树:一种支持空间近似关键字查询的外存索引.计算机研究与发展,2012,49(10):2142−2152.

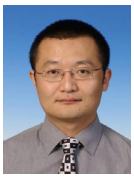
[109] 戴健,许佳捷,刘奎恩,武斌,丁治明.DKR-Tree:一种支持动态关键字的空间对象索引树.计算机研究与发展,2013,50(S1): 163−170.

[113] 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(8):2025−2047. http://www.jos.org.cn/1000-9825/5049.htm [doi: 10.13328/j.cnki.jos.005049]

童咏昕(1982－),男,北京人,博士,副教授,CCF 专业会员,主要研究领域为众包数据管理,不确定数据管理与挖掘,时空数据管理与挖掘,社交网络分析.

袁野(1981－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为图数据管理,众包数据管理,不确定数据管理,云计算.

成雨蓉(1989－),女,学士,CCF 学生会员,主要研究领域为不确定图数据管理与挖掘,众包数据管理,社交网络分析,时空数据管理与挖掘.



陈雷(1972－),男,博士,教授,博士生导师,主要研究领域为众包数据管理,不确定数据管理与挖掘,Web 数据管理,时空数据管理与挖掘.

王国仁(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据密集型计算,众包数据管理,非结构数据管理,云计算,生物信息学.