北京航空航天大學
BEIHANG UNIVERSITY

第二十六届“冯如杯”学生学术科技作品竞赛

项目论文

基于时空众包的实时调度系统

摘要

众包特指通过 web 平台以群体智慧解决问题的新模式。伴随着移动互联网、物联网等技术的发展以及智能移动设备的大范围普及，在移动设备自身携带的大量时空数据对众包问题形式的影响下，传统众包根据新的外延需求被赋予了新的概念——时空众包。人群通过移动设备实时向平台发布任务。当伴随着地理位置等信息的任务实时大量涌现至平台时，如何高效地处理这些任务，即将任务及时调度给特定的工人，成为了核心问题。这类问题常见于现实应用，如神州专车、滴滴打车、gigwalk 等。我们基于时空众包背景，嵌入合理的调度算法以设计实时的调度系统，以此解决这类问题。

关键字：时空众包，调度算法，系统设计

Abstract

Crowdsourcing, using the wisdom and sources from crowds, has become a new pattern in web platform. With the development of the Mobile Internet, the Internet of Things and the widespread using of the smart mobile devices, a new concept of crowdsourcing, spatialtemporal crodsourcing, which comes from the influence of the huge spatial data in mobile devices, is proposed to adapt to new demands. The crowds publish tasks by the mobile devices. When the tasks with large information such as spatial information poured into the platform in real time, the way to dispatch the tasks has become the key problem. The attempt in this kind of problem is universal in reality, such as Shenzhou car, Drops taxi and Gigwalk etc. To give a thorough and reasonable solution, a dispatching algorithm is built by the concept of spatialtemporal crowdsourcing. A dispatching system in real time based on the algorithm is designed to solve the problem.

Keywords: Spatialtemporal Crowdsourcing, Dispatching Algorithm, System Design

目录

图表清单

第一章 项目简介

1.1 项目背景

随着 Web 2.0 技术的兴起，大量在线 Web 应用正悄然改变着人类的生活模式，同时也为传统的人本计算提供了一种通过群体智慧求解问题的新模式——众包。早期的众包平台如维基百科、百度知道、Amazon Mechanical Turks (AMT)， CrowdFlower 及 oDesk 等，已深入我们的生活。根据美国亚马逊公司的报告，公司已经在 AMT 众包平台上的年度盈利超过 5.2 亿美元。因此，众包技术为当今互联网时代提供巨大机遇与挑战，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”。

现今，伴随着移动互联网、物联网与移动设备技术的快速发展，众包数据有了新的变化。众包数据更多地包含了移动端用户的时空信息，使得早期众包平台不能满足当前数据类型的多样化与任务的复杂化，新一代时空众包因此诞生。

时空众包的例子包括如美国的 Gigwalk 公司，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”服务也有异曲同工之处，其组织众包参与者收集了国内各大城市的道路周边信息。又如国内百度外卖服务，美国的 TaskRabbit 公司提供的“社区物品派送”服务，以及最近火热的打车应用如神州专车、滴滴打车、uber 等。

这些时空众包平台应用的背后都有着能够处理时空信息的一套系统。在时空众包的概念下，我们以其典型应用打车应用为例设计了一套实时调度系统，同时对现有的系统提出问题并做出改进。

1.2 项目特点分析

时空众包背景下，需求方向平台发布包含着地理位置信息的需求（如打车客户的位置）。这些需求会实时大量地涌入到平台上，如何高效地处理这些信息，即把需求及时分配给最合适的人（即把订单分给合适的司机），成为了问题的关键。这里以打车应用为例明确两类实体（即订单和专车司机）的特点和我们的问题着眼点。首先，系统的实时性强。多个订单会同时出现在平台上，且订单一端的客户通常希望尽快得到平台的应答，即收到平台选择的某个专车司机的响应。专车司机同样如此，当他们送完一个客户恢复空车后，会重新出现在平台上，此时司机心外尽快有新的工作。其次，系统的规模

性大。神州专车一天订单数约为四万单，平均每秒有十单以上。而司机在平台上的多次出现，从数据方面来讲也可看作是新的数据。最后，数据都包含地理位置信息。站在平台的角度来讲，希望一个时间单位内（如一天）可以尽可能多地完成订单。这就要求给司机分配订单时候要采取更优的策略（衡量策略优劣后文详细说明，这里可以理解为让司机少跑冤枉路），这个策略的制定当然与地理位置信息相关。这也是一个系统成败的关键。因而我们的问题着眼于策略制定，最终通过可视化方法描述系统的成功。

这里再对策略进一步说明。常识来讲知道得越多自然可以做得更好，然而有些时候为了一个目标近优去知道一些信息是要付出一定代价的，这会导致另一个指标的下滑，而满足另一个指标会导致这个目标不足。因此，权衡二者的策略才是更优的策略。当加之以理论及实验保证时，可以说明系统的成功。

1.3 相关工作

国外方面，美国南加州大学的 Shahabi 教授团队率先提出了一系列空间众包静态场景中任务分配问题，即在当前时刻给定众包任务与众包参与者当前的空间位置，以及众包参与者的活动范围和最大所能承接的空间众包任务数量。但由于时空众包任务和众包参与者均是随机动态出现，其采用的静态任务分配机制无法解决任务分配动态性与实时性的需求。此外，香港科技大学陈雷教授团队研发了国际首项基于时空众包的最优路径选择系统，即 gMission 系统，但该系统并不能解决实时任务调度问题。此外，本项目的主要特色在于首次采用预测技术来提高实时调度的有效性。

众包数据处理技术的研究在国内也开始浮现。清华大学的李国良老师团队首先研发了基于 Web 的众包任务分配算法。由于此类研究不涉及用户的时空信息，故不涉及与本系统相同的应用背景。此外，该团队最近也采用时空众包技术来预测实时交通情况，但其不涉及时空众包中的实时任务分配的问题。

第二章 具体设计

2.1 总体设计概述

如前文所说，打车应用是时空众包的典型应用。由于我们具有神州专车的数据，所以仅以打车应用为例设计系统。实际系统可以灵活变迁至其余时空众包应用，具有普适性。

整体系统包括前端和后台。前端即下图中数据源，指移动用户以及专车司机使用的手机 app。通过前端与数据库的交互，将时空数据迁移到后台。后台分两大模块，预测模块以及调度模块。预测模块一方面根据数据库中的历史信息进行预测，另一方面反馈给调度模块具体信息，涉及机器学习相关技术。调度模块收到预测模块的信息后，通过与数据库交互得到实际平台实时数据，又根据系统预先设定的优化目标制定策略，涉及组合优化相关理论。调度模块同时对可控实体（这里即专车司机）进行实时调度，将司机与客户的匹配通知信息通过数据库又返还给前端。

其系统结构图如下图。

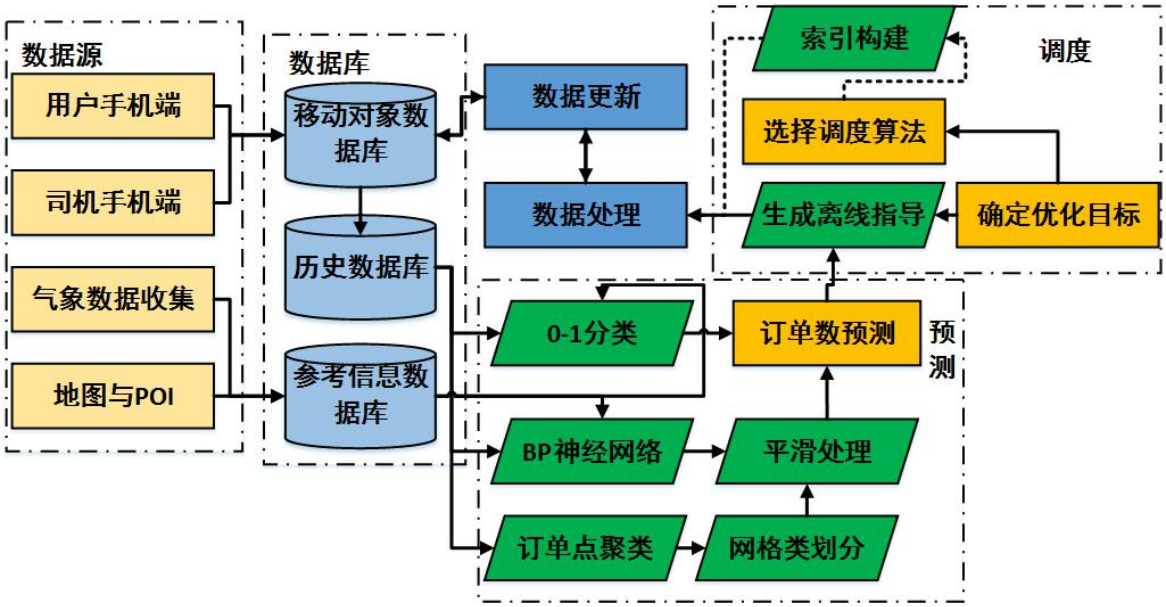

图 1 系统结构图

2.2 预测模块概述

预测模块根据已拿到的历史数据，建立起机器学习中的 BP 神经网络，对未来某一个时间窗口的某一地区内将会出现的订单数和专车司机数进行预测。

首先，对历史数据进行清洗。这里提前对毛刺数据进行删去，合理选定一些特征量，最终可以在记录条数上百万条时达到百分之八十以上正确率。前多一半进行学习，后少一半真实验证学习优劣，且创新性提出最后的平滑方法加强记录之间的地理位置相关性以进一步提升准确率。

2.3 调度模块概述

调度模块在实时更新了现有订单和专车司机信息后，结合对订单和专车司机的下一

时刻的预测信息，根据系统预先设定的策略目标，以高效的算法实时给出具体调度方案。

下图中空白的代表订单，含车的代表专车，当专车选择一个订单时，会有轨迹显示，右上方是实时的统计信息包括整个平台现有的等待客户数已经已完成订单数，右下方是实时的客户与专车司机来到平台时的信息。

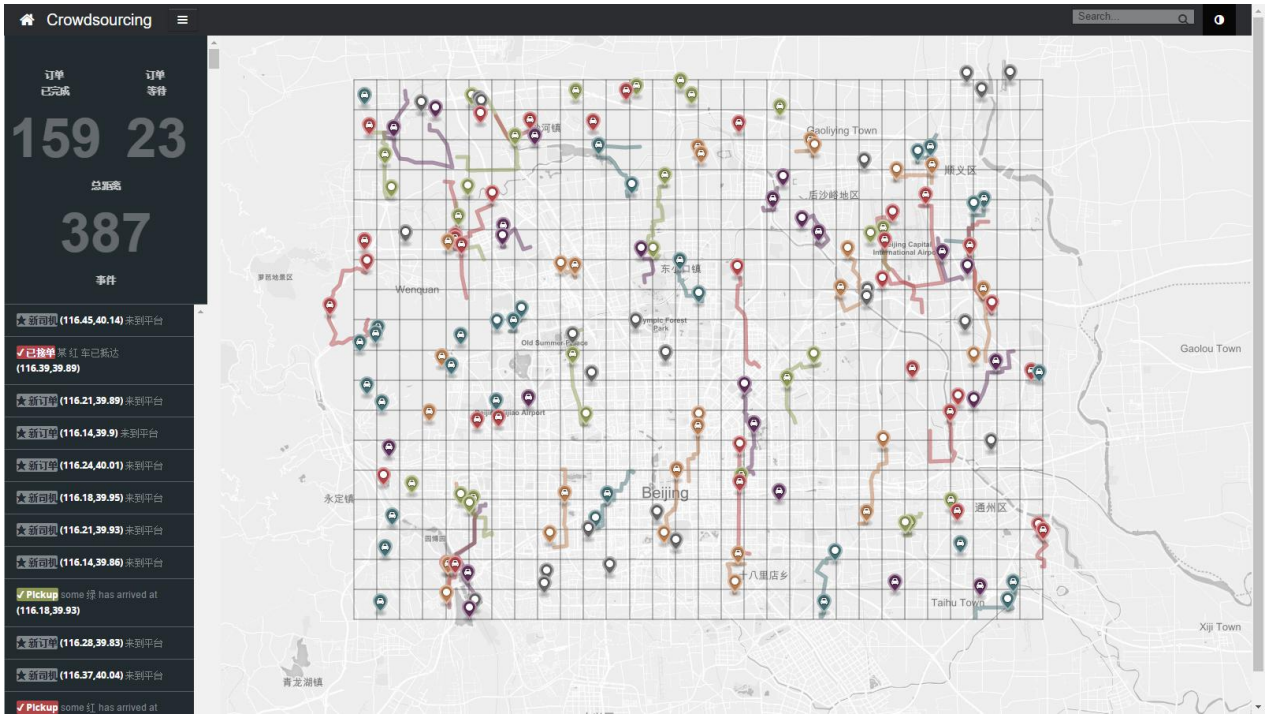

图 2 调度模块图

2.4 前端展示

前端 app 仅实现了 android 版，为系统结构图中数据源端中用户端，展示图如下图。用户通过输入自己的上车地点进行叫单。

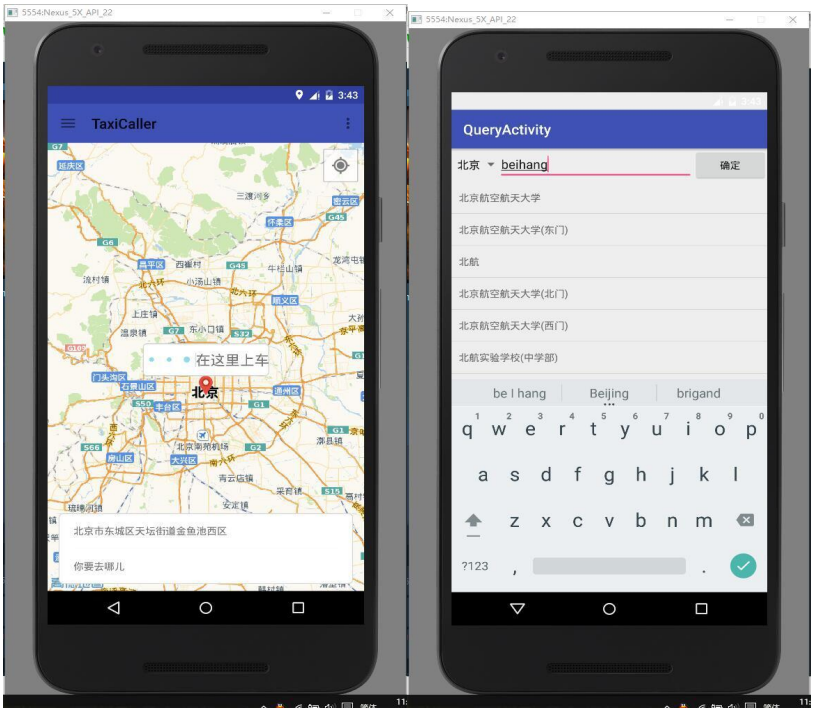

图 3 app 图

第三章 应用技术分析

3.1 传统调度方法

如相关工作中所提到，一部分工作用来满足实时性。如我们了解到的神州专车的做法，当一个订单出现时，采取最近邻的原则，即以订单为中心选取十个最近的空闲专车司机，采取询问的方式寻找合适的专车司机进行调度。这种调度方法显然可以满足实时性，然而这样的贪心启发式思想并没有任何理论保证，可以证明存在极端差的情况，使得这种方案下的多个司机的一次调度总行驶距离大大增加，却存在其他更优策略。

另一部分工作用来满足成本尽可能低。如前文提到的 Shahabi 教授团队所提出的策略，他们将时间划分为时间窗口，比如一个时间窗口代表二十分钟。他们同上边最近邻的原则不同，不在每个订单出现的时候都调用一次算法做匹配结果寻找司机，而是选择在一个时间窗口结束的时刻，这时将会攒下一些没有处理的订单和司机数据，然后进行处理。这样做的区别在于允许了数据的滞留，而不是尽可能的在可以处理数据的情况下处理，带来的坏处在于无法满足订单和司机所需的短响应时间，破坏了实时性。好处在于这样知道的信息更多了，可以更进一步的站在全局的角度考虑问题，而不会陷入局部最优。这里我们认为更进一步站在全局也只是五分钟的全局，即延迟五分钟做出的决策并不比立刻做决策有过多的优势。而如果能有调度方案以任一时间长为单位作为全局考

虑，且满足实时性的特点，则这一方案将是合理且健壮的调度方案。

之前的调度方案仍有可以借鉴的地方。他们均采用划格子的思想，在每格子内或是相邻一个格子进行调度。其优点在于可以做索引，大大提高处理数据的速度。在实际应用中，新加坡的出租车派遣系统就是如此。

3.2 基于 BP 神经网络的预测

BP 神经网络也称后向传播神经网络。通过对大量历史训练数据在神经网络上构成的映射关系不断加以学习（学习时利用梯度下降法），能够在不事先知道自变量与因变量之间隐含的数学函数的条件下，建立起一整套映射机制，对给出的新的自变量的取值，通过对神经网络模型的数学公式推导，得出以隐含的规律反映出的因变量的取值。

神经网络模仿的是人的大脑。不同于人脑，神经网络因为传输数据是比特流，其运行速度远远高于人脑，能够迅速的对大量历史数据加以学习，以建立完整的数学模型。

整个神经网络学习流程为前向传播信息，后向传播误差，并加以调整参数。每一层将本层的特征结合一个偏置量进行线性组合传递至下一层，传递意思就是得出下一层各个特征量的值，并同大脑类似经过神经元的一个突触时做一次激活函数的映射，直至最后一层因变量的取值，求得与因变量的误差后进行后向传播，调整每一层次的参数。

神经网络的具体拓扑结构由特征的选取开始。我们预先猜测我们需要预测的某个时间窗口内某个格子出现的订单数受哪些特征的影响，选定本格子的 POI，天气情况，前多周这一时间这个格子作为特征，并采取多层传递，到最后一层时采取新的平滑操作，具体利用了格子与格子间的地理相关性，对格子进行聚类后同处一类的格子一并做了新的一层网络，即最后因变量真实值取值为一类格子的订单数和。

先进行的是点聚类：

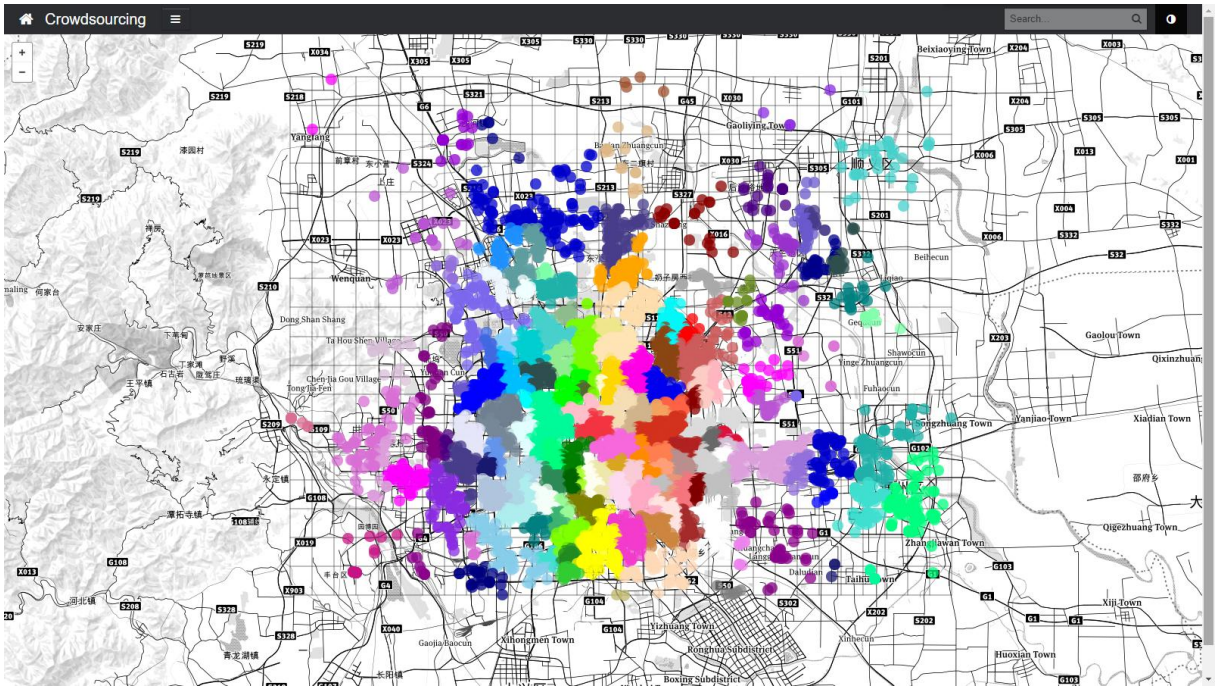

图 4 订单点聚类图

点聚类如系统结构图所示是为了后边的格子聚类，格子聚类的结果由点聚类结果得知，在某一网格内哪一类的点数最多这个格子就属于哪一类，格子聚类结果根据结构图我们可以看到是为了神经网络的平滑处理，这是我们预测技术的创新点。因为处于一类的格子是地理上相关联的，即这类地区的订单数是有关系的，所以在每个格子的神经网络模型之后再加上了整个格子类的大的神经网络，这部分建模即为平滑处理，可以有效降低预测与真实差值。格子的聚类结果如下：

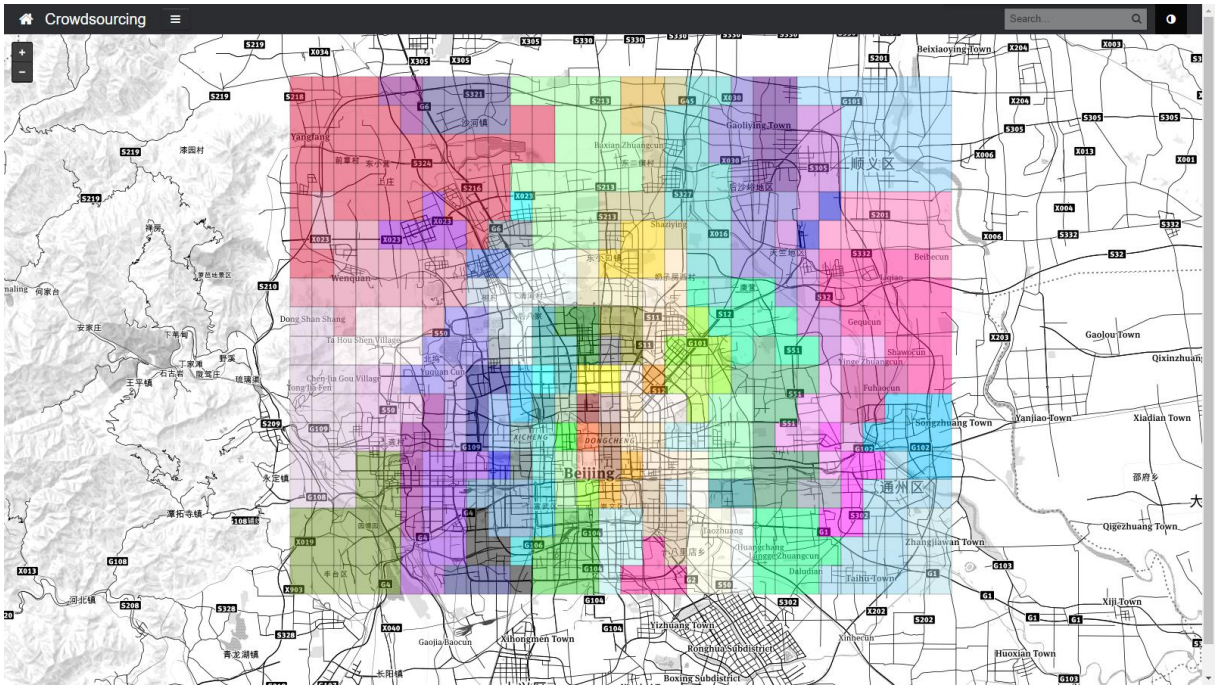

图 5 格子聚类图

我们根据神州专车在北京市 15 年的每月订单和车数据，将前多半进行训练，后多半进行测试，得到至少百分之八十正确率的预测结果。

预测结果可用两张不同时间的热力图代表。

图 6 密集订单预测热力图

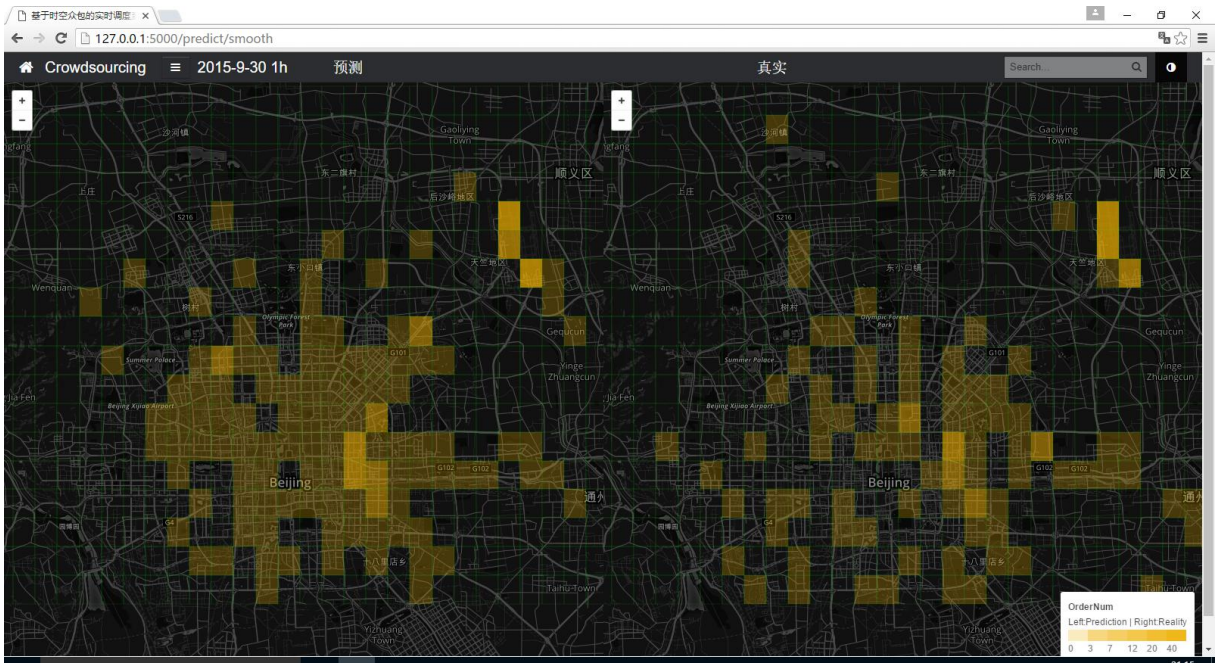

图 7 稀疏订单预测热力图

3.3 基于预测的调度

3.3.1 以最大化订单数为目标调度方案

当系统指定最大化订单数时，我们主要考虑的问题就是订单的实时性问题。客户的等待时间是有限的，当超过客户的忍受程度后，用户将会采取别的手段打车，这个订单将消失在平台上。而系统平台选择司机去接单的时间是一部分时间，司机发车去订单所在位置又是一部分时间，这两部分时间加和需小于客户预计的期限，否则将出现司机未接到客户时，客户已放弃订单的矛盾情况。

我们所要做的就是使第一部分系统选择司机的时间尽可能得短，体现在每当一个订单出现时就为它寻找司机。这里结合了预测信息，即若在下一两个时间窗口内有司机会出现也算作这个订单完成的候选者里。除非平台上现在没有司机且预测将来订单附近也没有司机，我们才选择滞留它，其余情况结合预测信息分配出去。

我们的调度方案同时对第二部分司机前往订单所在位置的时间进行了优化。如果可以提前预知未来，我们自然可以提前让司机前往订单所在区域，大大减少这部分时间。这里的问题从高层次来看，可以说将每个方格看作一个工作的机器，工作能力即司机数目的多少，等同于组合优化中经典的负载均衡问题。基于有效的预测预测，我们便可以从一个全局静态的角度合理处理这个经典问题。

3.3.2 以最小化成本为目标调度方案

在满足完成订单数尽可能多的情况下，我们使司机接单的空驶距离尽可能小。这里采取了一种理论研究已经很广泛的树结构——层次分割树。

层次分割树（HST），树上的每一个结点都代表一个簇，由一个结点到它的所有孩子结点的距离均相等。同一层到下一层的距离为固定值，每三层中两两层之间的距离成固定系数的比例。

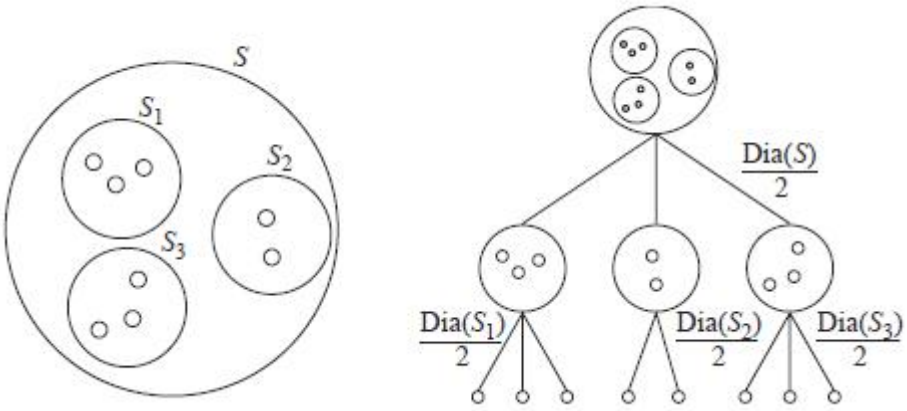

图 8 层次分割树

层次分割树是一种度量，类似欧式空间也是度量，一种度量给出任意两点之间的距离。我们希望寻找一种度量，能够将原欧式空间中的点投影到这种度量方式上，且保证在原来二维空间中两点距离能够小于这种度量下的距离，以及这种度量方式下的距离小于一个系数倍的原距离。这样的转换度量方式的好处在于，我们将原问题迁移到了新的度量方式下，而新的度量方式具有很好的性质，如我们这里采用的层次分割树具有树上的对称性和递归性，可以进一步从理论上保证在新度量方式下原问题的解不超过某系数倍的最优解，再乘以投影所需的损失的比例，即可得到实际的理论保证。

我们将所有格子中心点投影到层次分割树上，当每个订单和专车司机出现在平台时，结合本时间窗口和下一时间窗口的预测结果，将所有这些实体代表的点投影到树上，在树上完成匹配。最后对匹配结果进行通知，安排合适的车去接合适的订单。

第四章 项目分析

4.1 项目难点

项目的调度方案中运用了近似的手段。无论是时间窗口还是网格，我们认为一个时间窗口或是一个网格的距离通常用户都是可以忍耐的，即用户截止时间过一个时间窗口是可以接受的。这里强调虽然运用了近似，然而理论上仍然有保证，且这里的近似是为了一个全局角度的调度。

我们无论是求算网格间的距离，或是在建立 HST 时都是采用的欧式距离，和路网上的距离并不相同，但因路网上问题讨论的复杂性，没有采用路网距。

项目的可视化方面无法呈现真实数据，仅在几分钟之内单数就上以千计，我们在试

图可视化后发现所有的实体均在北京地图密密麻麻地排布，无法肉眼直观感受出系统的优点。只能通过实验，求得不同方法的在同一衡量指标下的值的大小来说明我们所采取的方案的优势。

层次分割树的理论比是基于构建算法中的某几步的随机选择，而实际中我们为了可视化，仅仅建立了 2-HST，且认为每次随机选择恰好构成我们所看到的完美四叉树，对于其他的树结构不利于展示。

项目尚未投入到实际使用，手机 app 暂时只能模拟人为操作，只能看到手机 app 在少量数据下的执行情况。

4.2 项目创新

项目的调度方案所采取的预测的方法能够将机器学习中的理论嵌入进设计的调度方案中，且调度方案采取组合优化的方法允许之前的预测有一定的偏差，综合两大方面理论提出的基于预测的调度方案能够实际对现实中的实时调度系统加以改进，以一种全局的角度审视及解决问题，且很好的保证了实时性和成本的需求。

项目对所设计的系统加以可视化呈现，帮助深入理解系统运行过程。

项目所设计的系统具有一定的普适性，能够迁移至时空众包的同类应用中。

结论

在真实生活中群智计算代表的一类应用已经逐渐普及，在计算机科学界中数据库国际会议上已将众包作为新的论文接受方向，而时空众包又以其全新的概念流行起来。本文便是以当下火热的打车应用为例，设计了一种基于时空众包的实时调度系统，同时在系统内嵌入相关各类模块，模块间彼此交互以保证系统稳定且高效运行。我们以可视化的方式呈现了系统的运行情况，同时对系统的创新处加以说明，从多个角度展现出系统的各方面优点。最后对目前项目的难点进行总结，期望在后续的工作中能够加以完善。

参考文献

[1] E. Law and L. Ahn. Human computation [B]. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers, 2011.

[2] J. Howe. Crowdsourcing: Why the power of the crowd is driving the future of business [B]. Crown Business, 2009.

[3] L. Kazemi and C. Shahabi. Geocrowd: enabling query answering with spatial crowdsourcing [C]. In Proceedings of the SIGSPATIAL 2012 International Conference on Advances in Geographic Information Systems (GIS 2012), Redondo Beach, USA, 189-198.

[4] L. Kazemi, C. Shahabi, and L. Chen. GeoTruCrowd: trustworthy query answering with spatial crowdsourcing [C]. In Proceedings of the SIGSPATIAL 2013 International Conference on Advances in Geographic Information Systems (GIS 2013), Orlando, USA, 304-313.

[5] H. To, G. Ghinita, and C. Shahabi. A framework for protecting worker location privacy in spatial crowdsourcing [J]. Proceedings of the VLDB Endowment, 2014, 7(10): 919-930.

[6] S. He, D. Shin, J. Zhang, and J. Chen. Toward optimal allocation of location dependent tasks in crowdsensing [C]. In Proceedings of the 2014 IEEE Conference on Computer Communications (INFOCOM 2014), Toronto, Canada, 745-753.

[7] C. Zhang, Y. Tong, L. Chen. Where To: crowd-aided path selection [J]. Proceedings of the VLDB Endowment (PVLDB), 2014, 7(14): 2005-2016.

[8] Z. Chen, R. Fu, Z. Zhao, Z. Liu, L. Xia, L. Chen, P. Cheng, C. Chen Cao, Y. Tong, and C. Zhang. gMission: a general spatial crowdsourcing platform [J]. Proceedings of the VLDB Endowment (PVLDB), 2014, 7(13): 1629-1632.

[9] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, Jianhua Feng. QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications [C]. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD 2015), Melbourne, Victoria, Australia, 1031-1046.

[10] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, Jianhua Feng. iCrowd: An Adaptive Crowdsourcing Framework [C]. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD 2015), Melbourne, Victoria,

Australia,1015-1030.

[11] Huiqi Hu, Guoliang Li, Zhifeng Bao and Jianhua Feng. Crowdsourcing-Based Real-Time Urban Traffic Speed Estimation: From Speed to Trend [C]. In Proceedings of the 32nd International Conference on Data Engineering (ICDE 2016), Helsinki, Finland.

[12] Jittat Fakcharoenphol, Satish Rao, and Kunal Talwarc. A tight bound on approximating arbitrary metrics by tree metrics. Journal of Computer and System Sciences 69 (2004) 485–497.