《交通科学》第 00 卷，第 0 期，0000 年
Xxxxx，页码 000–000ISSN 0041-1655 | EISSN 1526-5447 | 00 | 0000 | 0001

INFORMS

DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

我们鼓励作者使用包含期刊标题的样式文件模板，向 INFORMS 旗下期刊投稿新论文。然而，使用模板并不代表论文已被所列期刊接受发表。INFORMS 期刊模板仅用于向 INFORMS 旗下期刊投稿，不得用于以印刷版或在线形式分发论文，也不得用于向其他出版物投稿。

arXiv:2301.02142v1 [cs.LG] 5 Jan 2023

玩转躲猫猫：应对店内拣选作业，同时提升顾客体验

F´abio Neves-Moreira, PedroA
morimfabio.s.moreira@inesctec.pt, pamorim@fe.up.pt INESC TEC，波尔图大学工程学院，葡萄牙波尔图 4200–465

零售业的发展带来了新的挑战，也引发了一个关键问题：如何重塑商店和供应链，以满足线上渠道日益增长的需求。全渠道零售商近期采取的一项举措，是通过店内拣选应对线上销售的增长，利用现有资源处理在线订单。然而，这也可能损害线下顾客体验。为了制定适应零售商店动态顾客流动的拣选策略，我们将一个新问题形式化，称为动态店内拣货员路径规划问题（diPRP）。在这一重要问题 diPRP 中，拣选员尝试拣选在线订单，同时尽量减少与顾客相遇。我们将该问题建模为马尔可夫决策过程（MDP），并采用结合数学规划与强化学习的混合求解方法进行求解。针对合成实例开展的计算实验表明，该算法能够收敛至高效策略。此外，我们将这一方法应用于一家大型欧洲零售商的实际场景，评估所提出策略在拣选订单数量和遇到顾客数量方面的表现。我们的研究表明，零售商应能够扩大在线订单的店内拣选规模，同时不损害线下顾客的体验。与只专注于拣选订单的策略相比，使用所提出的求解方法学习得到的策略可将与顾客相遇的次数减少 50% 以上。因此，为了实现能够妥善平衡运营效率与顾客体验的全渠道策略，零售商不能依赖现有的简单拣选策略，例如选择可能的最短路径。

Key words：全渠道零售；店内拣选；马尔可夫决策过程；强化学习；实际应用

© 0000 INFORMS

1. 引言

零售的未来充满不确定性。这种不确定性与顾客消费模式、渠道转变，以及人们对速度和便利性的期望变化速度密切相关。新冠肺炎疫情加剧了这种不可预测的行为，但这些变化将持续下去。目前，投资者、企业家、商业专业人士和学者针对零售业提出的几个关键问题仍未得到解答（Caro 等，2020年）。

大型零售商面临的最关键挑战之一，与实体零售店的未来息息相关（Adhi et al. 2022）。有人认为实体零售可能已经“消亡”，但也有许多理由可以为其存在辩护：实体零售能够实现差异化竞争，提供卓越的体验式购物，这依托于优质的客户服务、丰富的商店以及线上与实体渠道的协调融合（Dennis 2018）。如今，许多零售商认识到，专注于全渠道零售是提升顾客价值主张和运营效率的可行之道（Gao and Su 2017a）。尤其是，对于相当一部分食品杂货零售顾客而言，实体渠道仍然十分重要；他们觉得，在线下单，再到附近的零售选址提取杂货，十分便利（Glaeser et al. 2019, Vyt et al. 2022）。然而，这种线上线下二分也带来了诸多挑战。例如，如何向线下顾客提供与线上同等水平的信息（Gao and Su 2017b）、应将大部分库存分配给哪个渠道（Jia et al. 2021），以及线上和线下渠道应采取哪些商品组合决策（Chen et al. 2021），目前都尚无定论。

显而易见，实体零售店在便利性和速度方面具有巨大优势，尤其是在食品零售领域（Barbee 等，2021）。例如，这些优势推动了全渠道零售近期从电子商务转向即时商务（Q-Commerce）的趋势。即时商务侧重于在较短的配送时间内（10 至 30 分钟），提供少量商品，包括杂货、文具和非处方药。这类服务有时

被称为按需配送，需要采用更快速的履约策略，将传统上位于郊区的仓库转变为微型仓库，或利用靠近配送地点的实体零售店。这意味着，未来零售门店内的内部拣选员和第三方拣货员数量应大幅增加，以便零售商应对在线订单和到店自提订单数量的增长，并在较短时间内完成配送。部署全面数字化的商店后，这些全渠道履约策略还将得到进一步推动，使零售商能够利用物联网（IoT）、先进机器人技术和数字孪生等新兴技术（Olsen 和 Tomlin，2020）。

本文以一家欧洲零售商的真实案例为背景，重点关注由店内拣货员负责拣选的线上订单履约。这些订单发布在零售商的网站上，顾客可以选择送货上门，也可以到店自取。这一场景极具挑战性，由于以下四个主要原因，在拣选员的拣选路径中为其提供协助正变得愈发重要。首先，近年来线上业务大幅增长，因为顾客几乎不得不尝试网上购物（Richter 2021）。其次，大多数实体店没有足够的后仓空间，无法设置一个经过优化、能够拣选线上订单中可能出现的所有商品的仓库（Pires et al. 2021）。第三，随着人们近期对即时商务的关注不断升温，越来越多的第三方拣货员（即 Uber、Glovo 和 Delivery Hero 的人员）开始在商店内购物（Nierynck 2020）；因此，当他们不熟悉每种商品的位置时，提供能够帮助他们在零售门店内导航的服务十分重要。第四，也是更重要的一点，随着店内自有拣选员和第三方拣货员大量涌入，店内顾客体验面临风险，这与杂货零售商所采取的线上线下策略相冲突。

在本文中，为了实现能够适应零售商店动态顾客流动并有助于缓解上述挑战的拣选策略，我们引入了一个名为动态店内拣货员路径规划问题（diPRP）的新问题。该动态问题考虑动态零售门店环境，其目标是作出决策，使拣选员在最大化拣选订单数量的同时

尽量减少与顾客相遇的次数。需要注意的是，对商店内不确定顾客流动进行建模是这一挑战的核心要素。基于随机数学规划的传统拣选员辅助方法很可能会产生难以求解的混合整数规划（MIP）模型，并且不适合捕捉序贯决策问题的动态性。因此，diPRP被建模为马尔可夫决策过程（MDP），并通过一种广为人知的强化学习技术求解。通过采用混合 Q 学习算法获得拣选策略，该算法通过提供拣选员应执行的最佳动作（即拣选员移动），最大化即时奖励和未来奖励（即，拣选商品时获得正奖励，而遇到顾客时获得负奖励）。目标是在避免与顾客相遇（购物体验）的同时，拣选尽可能多的订单（运营效率）。

我们展示了如何利用真实世界数据，并统筹运用数学规划、仿真和机器学习技术，解决全渠道零售中的一个相关实际问题——diPRP。本研究的科学贡献有三方面：

1. 我们对零售领域一个新颖且相关的问题进行了形式化：拣选员在零售商店内拣选在线订单和到店自提订单，同时避免与顾客发生身体接触——即 diPRP。

2. 我们使用混合 Q 学习算法求解这一动态问题。该算法包含混合整数规划和马尔可夫决策过程组件，用于确定适应店内拣选作业动态环境的拣选策略。

3. 我们使用合成实例开展了关于算法收敛的计算实验，并针对四种不同的拣选策略总结了管理启示；这些策略已在与一家大型欧洲零售商合作开展的真实应用中得到验证。这些启示旨在帮助零售商更好地制定全渠道策略，为扩大新型履约选项奠定坚实基础。

本文其余部分安排如下。第 2 节回顾与店内拣选员路径规划问题相关的文献。第 3 节介绍问题，并阐述其固有

零售商店仿真环境，该环境将用于我们的研究。第 4 节详细介绍基于 Q-learning 算法的解决方案。第 5 节展示在合成实例上开展的计算实验，结果体现了算法收敛情况。第 6 节介绍其在一家欧洲零售商店中的实际应用。最后，第 7 节总结本文的主要结论，并提出未来的研究方向。

2. 文献综述

在以下各小节中，我们回顾近期针对拣货员路径规划问题（第2.1节）、动态路径规划问题（第2.2节）和动态拣货员路径规划问题（第2.3节）提出的方法。我们认为，这些主题涵盖了处理我们的问题——diPRP——所需的建模和求解方法技术。该问题涉及在动态门店环境中拣选商品，是一种新型拣选作业。

2.1. 拣货员路径规划问题

拣货员路径规划问题或订单路径规划问题由 Ratliff 和 Rosenthal（1983）提出，通常被建模为 Steiner图上的旅行商问题（TSP）（Letchford 等，2013；Cambazard 和 Catusse，2018；Rodr´ıguez-Pereira 等，2019）。文献中已提出多种问题扩展，以适应具有特定业务约束的现实案例（Chabot 等，2017；Quader 和 Castillo-Villar，2018；Ardjmand 等，2018），并结合后续问题，例如存储和批处理（van Gils 等，2018）。在求解方法方面，该问题通常采用启发式算法（Theys 等，2010）进行处理，但研究发现了一些重要的问题特性，可提高基于数学公式的精确方法的效率（Cornu´ejols 等，1985；Pansart 等，2018）。最近，有研究提出了新的公式，以扩大能够精确求解的实例规模（Scholz 等，2016），并将订单拣选问题与供应链中的其他活动相结合（Roodbergen 等，2015），因为该问题仍然是仓库运营管理中的一个重要问题。拣货员路径规划问题通常在窄巷道仓库中求解（Roodbergen 和 de Koster，2001；Roodbergen 和 Koster，2001；de Koster 等，2007；Chabot 等，2018；Masae 等，2020），而不是在内部

商店中。考虑到商店的布局并未针对拣货作业进行优化，并且可能会有顾客拥挤，这就形成了一个订单拣选文献中从未研究过的新情境。在商店中，商品布局可能非常不规则，订单拣选文献中所假设的大多数图属性并不适用。此外，顾客购物路径可能非常混乱，且事先并不知晓事先，这意味着拣选问题具有更 强的随机性和动态性。

2.2. 动态路径规划问题

尽管随机动态路径规划已有30年历史，但该领域发表的大多数论文都是近年来问世的。Ulmer等（2020）的文献综述中，超过半数的论文发表于2010年之后。其他关于这一主题的文献综述由Berbeglia等（2010）、Pillac等（2013）、Ritzinger等（2016）、Psaraftis等（2016）和Ojeda Rios等（2021）发表。值得一提的是，多种趋势（例如电子商务的增长、共享经济和可持续性）以及技术进步（例如数字连接、大数据和自动化），正在推动用于应对动态路径规划问题的新方法不断发展（Savelsbergh和Van Woensel，2016）。为解决这些问题，Powell（2019）提出的随机规划统一框架发挥了决定性作用，因为研究人员一直在对其进行调整，以解决建模为马尔可夫决策过程的动态问题。为了使动态规划适用于大型现实问题，Powell（2007）提出的近似动态规划方法也已成为该领域的重要参考。

近年来，也有一些研究尝试通过强化学习技术求解建模为马尔可夫决策过程的路径规划问题（Nazari等（2018）、Li等（2021）、Kullman等（2022））。然而，尽管这些方法有望应用于新的情境，路径规划领域仍认为它们尚未得到充分探索，正如Hildebrandt等（2021）所指出的那样。

2.3. 动态拣选员路径规划问题

在某些业务情境中，考虑拣货员路径规划问题的动态变体颇有意义。当新信息有助于更好地规划或执行拣货作业时（例如，有新订单到达），便会出现这类问题。关于动态拣货员路径规划问题，

相关文献非常有限。Gong 和 Koster（2008）研究了在一个拣选周期内拣选信息可能动态变化的拣货作业。作者基于随机轮询理论的方法指出，这种做法能够缩短订单处理时间，并提高按时服务完成率。Lu 等人（2016）研究了动态订单拣选策略，该策略允许在拣选路径执行期间更改拣选清单。作者允许在拣货作业期间重新优化最优路径，与问题的静态版本相比，所得解的质量更高。

请注意，这些作者认为，动态信息的到达仅与顾客到达有关。然而，在商店中，有两类实体会执行拣选：线下顾客和线上订单拣选员。拣选员在执行拣选路径时，可以收集顾客位置信息。这由此形成了一个新的动态拣选路径问题。然而，店内拣选员与店内顾客之间的互动尚未得到考虑，也就是说，目前还没有提出解决动态店内拣选员路径规划问题的方法。

尽管已有大量文章探讨拣货员路径规划问题，但尚无文献考虑店内拣选团队所面临的新环境。事实上，2020年的疫情背景带来了需要管理的新运营活动，以及在追求不同目标的同时需要应对的新挑战。例如，针对商店内顾客数量施加的限制，可能会使零售商有兴趣引导顾客沿购物路径行进，以加快顾客的购物过程，而不是让他们在商店内停留更长时间（Hui et al. 2013, Boros et al. 2015, Hirpara and Parikh 2021）。每位店内顾客也在解决自己的拣货员路径规划问题，以确定其购物路径。顾客并不知道商品的确切位置，且部分顾客可能存在拣选顺序偏好（Larson et al. 2005），这可能导致顾客在商店通道中产生混乱流动（Chen et al. 2015, Li et al. 2017）。这意味着，在确定每位拣选员的拣选路径时，商店的状态是未知的先验地，而是随机的。因此，如果我们关注的是不干扰顾客购物体验的拣选路径，

针对拣货员路径规划问题所提出的方法在应对这一问题上仍有不足。这正是我们在本文中要填补的空白。

3. 问题描述

本节通过描述所考虑的零售商店环境中包含的所有相关要素、将在商店内做出决策的拣选员智能体，以及用于建模相关动态问题的马尔可夫决策过程（MDP），对 diPRP 进行形式化。因此，第 3.1 节详细介绍零售商店环境，包括商店图、用于模拟实体顾客购物路径的模型，以及由一组拣货位置构成的线上订单到达。第 3.2 节根据拣选员在每个步骤中可获得的信息（例如，接下来要拣选的商品和附近的实体顾客），阐明拣选员的行为。第 3.3 节介绍用于对我们旨在解决的动态问题进行建模的马尔可夫决策过程，包括决策时点、状态、动作、状态转移函数、奖励和目标函数。

3.1. 零售商店环境

为了对真实的零售商店环境进行建模，我们采用一种仿真环境，同时考虑实体顾客到达和线上订单到达。商店被建模为一个稀疏图 $\mathcal{G}=(\mathcal{V},\mathcal{E})$，其中顶点集合（门店内的位置）被划分为顶点 0（即商店入口，购物路径从此处开始）、顶点 $\{1,...,v\}$（对应商店内的 v 位置）以及顶点 $v+1$（即购物路径结束的位置，例如收银台或订单准备区）。实体顾客按照均值为每个时间段 λ^{Store} 名顾客的泊松分布到达商店。每位实体顾客以 $speed^{Customer}$ 的速度沿商店通道移动，按照给定的购物路径 θ_c 在图 $\mathcal{G}$ 中通行，该路径包含一系列门店内的位置。实体顾客在某个节点拣取商品需要$service\ time^{Customer}$ 个时间段。店内顾客数量不得超过商店的容量 $|\mathcal{C}|$，并且

同一商店节点中的顾客最大数量为 m。线上订单的到达率也是一个随机变量，服从均值为每个时间段 λ^{Online} 个订单的泊松分布。每个线上订单 $o \in \mathcal{O}$ 由一组需要访问的拣货位置 $\mathcal{L}_o \subset \mathcal{V}$ 组成，以便拣取订购的商品。

3.2. 拣选员智能体

每当有线上订单 o 到达时，都需要将其分配给一名空闲的拣选员。此时，根据业务决策规则（例如最短路线、优先拣选重物，以及最后拣选冷冻商品）确定拣选顺序 θ_o。本文首先假设这些拣选顺序旨在最大限度地缩短行进距离，即采用最短路线，以提高运营效率；这也是实践中的常见情况。拣选员按照既定的拣选顺序行动，但为了不干扰顾客的购物体验，会尽量避免与顾客相遇。因此，在每对待拣选商品之间，拣选员都会在避开商店内顾客的同时，求解一个最短路径问题。该问题的动态特性体现在需要连续求解一系列动态最短路径问题。这个问题固有的基本权衡在于：拣选员要遵循零售商店所采用的决策规则（例如最短路线）提供的预设拣选位置顺序，但又需要避开前往两个拣货位置之间路径上的顾客。顾客在店内的位置事先未知，只有当拣选员查看通道时才能发现。一天之中，每名拣选员会接收多个线上订单，因此需要执行多个拣选序列。

3.3. 将动态拣选员路径规划问题建模为马尔可夫决策过程

为了将 diPRP 表述为具有 T 个时段的有限时域马尔可夫决策过程（MDP），考虑决策时点$k \in 0, 1, .., T$，在该时点需要决定接下来要前往商店内的哪个位置节点。每当拣选员到达一个新位置时，就会触发新的决策时点 k。每个决策时点 k 都对应一个决策状态 s_k。在初始状态 s_0 和初始决策时点$k = 0$，拣选员位于起始仓库节点 0。在终止决策时点 $k = T$，

© 0000 INFORMS

拣选员完成当前订单的拣选后，返回终点仓库节点 $v+1$。在状态 s_k下，系统会接收更新信息，包括拣选员所在位置及相邻位置（由弧连接处）的顾客数量。根据系统提供的当前线上订单信息，需要确定一个动作 a_k （即接下来前往哪个位置，以寻找商品或前往仓库）。每到访一个位置，都会增加总行程时间，并可能增加与顾客相遇的次数。拣选员的目标是找到最优路径，既能快速、短程地完成行程，又能尽量减少与顾客相遇的次数。每拣选一件商品，拣选员都会获得正向奖励。总体而言，店内拣货员面临的问题可建模为一个由四个部分组成的马尔可夫决策过程（MDP），即：(1) 状态集合 $\mathcal{S}$；(2) 动作集合 $\mathcal{A}$；(3) 奖励函数 R；以及 (4) 转移概率 P。

3.3.1. 决策时点 决策时点 k 始于拣选员抵达商店内的新位置时。

3.3.2. 状态 决策时点 k 的系统状态由元组 $s_k = (n_k, z_k, t_k)$定义，其中 $n_k \in \mathcal{V}$为拣选员当前所在位置，$z_k = (z_k(1)\ z_k(2)\ \ldots,\ z_k(|\mathcal{L}_o|))$是一个向量，包含在线订单 o在决策时点 k各拣货位置的状态，而 $t_k \in [0,T]$为抵达位置 n_k的时间。对于在线订单的每个拣货位置，$z_k(.)$取值于集合 $\{0,1\}$:

$$z_k(.) = \begin{cases} 0, & \text{if picking location has not been visited by time } t_k \\ 1, & \text{if picking location has been visited by time } t_k \end{cases} \tag{1}$$

在初始状态 $s_0 = (0,\ z_0,\ 0)$ 下，拣选员位于订单准备区，并且对于每个拣货位置 $n \in \mathcal{V} \setminus \{0\}$，$z_0(n)$ 的值为 0 或 1。当拣选员到达集合 $\{(0,\ z_k,\ t_k)$ 中的终止状态 s_K时，即到达最后一个决策时点：$t_k \in [0,T],\ z_k \in \{0,1\}^{|\mathcal{L}_o|}\}$，此时拣选员已在时间 T 前返回备货区，且正在处理的在线订单的拣货位置状态为 0 或 1。状态空间为集合 $\mathcal{S} = \mathcal{V} \times [0,T] \times \{0,1\}^{|\mathcal{L}_o|}$。

3.3.3. 动作 决策时点 k 的一个动作，是将拣选员指派到属于 $\mathcal{V}$ 的一个店内位置。当拣选员处于状态 s_k 时，对应于从 n_k 出发的可能弧的可行行动集合为

$$\mathcal{A}(s_k) = \{a_k \in \{n \in \mathcal{V} : (s_k, n) \in \mathcal{E}\}\} : \tag{2}$$

3.3.4. 状态转移 在从状态 s_k选择动作 a_k 后，系统在决策时点 $k+1$转移到新状态 s_{k+1}。关于拣选员在下一决策时点所处的位置，不存在任何不确定性。然而，关于遇到的顾客数量的未知信息会被揭示。该信息是 Ω_{k+1} 的一次实现，并由 ω_{k+1}给出。令 $R_{k+1}(s_k,\ a_k,\ \omega_{k+1})$ 表示在决策时点 k 从状态 s_k 选择动作 a_k 并观测到随机信息 ω_{k+1}时产生的随机负奖励。

3.3.5. 奖励 当拣选员处于状态 s_k 并执行动作 $a_k \in \mathcal{A}$时，奖励取决于拣选员移动到的店内位置的状况。奖励由四个部分组成：（1）每个拣货员步数对应的固定负奖励；（2）与拣选员位于同一位置的店内顾客数量；（3）从拣选员位置可到达的位置中的店内顾客数量；以及（4）每件已拣选商品对应的正奖励。自上一个决策时点以来的步数由 ϕ_1给出，同一位置的顾客数量由 ϕ_2 给出，附近顾客数量由 ϕ_3给出，已拣选商品数量由 ϕ_4给出。这四个部分分别可通过权重 w_1 至 w_4进行调整。奖励 $R_k(s_k,\ a_k,\ \omega_{k+1})$ 是一个取决于门店环境的随机变量，其定义为

$$R_k(s_k, a_k, \omega_{k+1}) = -w_1\,\phi_1(\omega_{k+1}) - w_2\,\phi_2(\omega_{k+1}) - w_3\,\phi_3(\omega_{k+1}) + w_4\,\phi_4(\omega_{k+1}) \tag{3}$$

3.3.6. 目标 目标是在各决策时点最大化奖励期望总和；也就是说，在权重设置得当的情况下，尽可能多地拣选订单，同时尽量减少在店内与顾客相遇的次数。令 π为一系列动作 a_0，a_1，……，a_T，对应每个决策

© 0000 INFORMS

时点 $k=0$，……，T。目标是找到一个最优策略 π，以最大化期望奖励总和。给定初始状态 s_0 时，目标函数如下

$$\min_{\pi} \mathbb{E}\big[\sum_{k=0}^{T} R_k | s_0\big] \tag{4}$$

4. 求解方法

第3节定义的 diPRP 需要采用适用范围和目的各不相同的求解技术：（1）每当有在线订单到达时，都需要求解一个购物路径规划问题（SRP）模型，即一个混合整数规划，以根据预定义的决策规则获得拣选顺序；（2）对于每个拣选顺序及该顺序上的每一对商品，都要解决一个序贯决策问题，并通过 Q-learning算法递归更新策略。图1概述了所设计的方法，展示了每种技术的应用环节，以及拣货代理如何与门店环境交互。

在第4.1节中，我们将介绍问题、用于获得拣选顺序并将其提供给拣货代理相关流程的数学公式化及求解方法。在第4.2节中，我们将介绍所提出的 Q-learning方法，以学习一种最优策略，从而高效拣选订单，同时避免在店内与顾客相遇。第4.3节将展示采用不同策略得到的拣选路径示例，以及一个通过 Q学习获得的策略示例。

4.1. SRP 定义

为介绍标准 SRP，考虑第 3.1 节所述的零售商店环境。每条边（i，j）$\in \mathcal{E}$ 是一对位置，与一系列网格步相对应，这些网格步构成了拣货位置 i 与 j之间的最短路径。边的权重由因子 w_{ij}确定，通常表示弧的距离或时长。同样，设 $\mathcal{L}_o \subset \mathcal{V}$ 为与为完成给定购物清单 o而必须访问的拣货位置相对应的顶点子集。拣选员的主要目标是在最小化其购物路径距离（或时长）的同时，拣选购物清单中的所有商品。

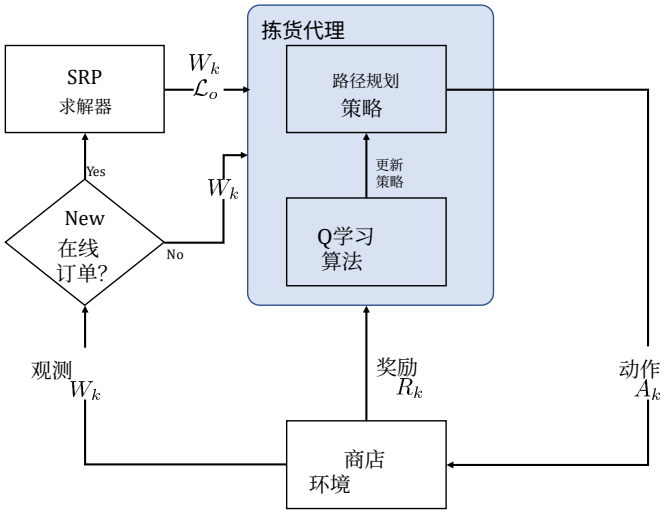


图1 方法概览，包括对拣货代理的详细拆解。

4.1.1. SRP 数学公式化 为对 SRP 进行建模，我们使用二元决策变量x_{ij} 来定义路径决策。设 x_{ij} 为二元变量，用于表示是否经过边（i，j）。所提出的公式如下：

SRP:

$$\text{minimize} \sum_{(i,j)\in\mathcal{E}} w_{ij}\, x_{ij} \tag{5}$$

s.t.

$$\sum_{i\in\mathcal{V}} x_{ji} = 1 \quad \forall\, j \in \mathcal{L}_o \cup \{0\} \tag{6}$$

$$\sum_{i\in\mathcal{V}} x_{ij} = 1 \quad \forall\, j \in \mathcal{L}_o \cup \{v+1\} \tag{7}$$

$$\sum_{i\in\mathcal{O}} \sum_{j\in\mathcal{O}} x_{ij} \le |\mathcal{O}| - 1 \quad \forall\, \mathcal{O} \subset \mathcal{V}, i \ne j \tag{8}$$

$$x \in \{0,1\}. \tag{9}$$

目标函数（5）将购物路径的总距离（或时长）降至最低。约束（6）和（7）确保起始位置、产品拣选位置和结束位置处的流守恒。加入约束（8）以消除子回路。最后，约束（9）定义每个变量的类型和界限。

© 0000 INFORMS

4.1.2. SRP 求解方法在 diPRP 的背景下，每个 SRP 都需要在几分之一秒内求解。此外，在仿真过程中（稍后将用于优化拣选策略），我们需要求解数百甚至数千个 SRP 实例。当拣选清单包含数量相当多的产品时，仅依赖数学求解器是不够的。因此，我们采用一种高效的精确求解方法，以快速求解上一节介绍的数学公式化。我们采用割平面算法，通过加入割平面，从线性松弛的可行域中移除分数解（算法1）。

算法1 割平面

1: **过程** CuttingPlanes($\mathcal{V}$, $\mathcal{L}_o$)
2: $CutsAdded \leftarrow True$; $X \leftarrow \emptyset$; $Y \leftarrow \emptyset$; $Z \leftarrow 0$;
3: **当** $CutsAdded$ **时执行**
4: $X \leftarrow$求解考虑集合 $\mathcal{V}$ 和 $\mathcal{L}_o$的分配模型(5)-(7);
5: $Y \leftarrow$将弧分到连通分量集合中;
6: $Z \leftarrow$计算 Y 中连通分量集合的数量;
7: **如果** $Z = 1$ **则**
8: $CutsAdded \leftarrow False$;
9: **else**
10: 添加割约束 (8) 以消除子回路;
11: $CutsAdded \leftarrow True$;
12: **如果** $CutsAdded$**则**
13: $X \leftarrow$将正变量转换为整数;
14: **else**
15: **如果** 解中的所有边均为整数 **则**
16: $\theta_o \leftarrow$从模型解 X 中获得拣选顺序;
17: **返回** θ_o。

该算法首先求解松弛指派模型（算法1，第4行），并分离连通分量集合（算法1，第5行）。如果连通分量集合不止一个

，则该解包含子回路（算法1，第10行），并添加割以消除这些子回路（算法1，第11行）。添加割时，将取正值的边变量转换为整数。重复这一过程，直至求解松弛指派模型后不再添加割，且所有边变量均为整数，这意味着已找到最优解（算法1，第17行）。

4.2. Q学习

第4.1.2节介绍的切平面方法为在线订单的产品清单提供了一个拣选顺序 θ_o。o。然而，店内拣货员可以通过多种方式从一个拣选位置前往另一个位置。在SRP公式中，我们假设采用最短路径，也就是说， traversing an arc (i， j)时所执行的步骤，是能够使 i 和 j 之间的距离最短的步骤（同样请注意，也可以使用其他决策规则）。因此，为了不打扰店内顾客，从 i 前往 j时，根据店内顾客移动模式的不确定性，采取不同的步骤或绕行路线可能会更有利。

为了获得一种遵循给定拣选顺序、同时尽量避开顾客的路径规划策略，我们设计了一种Q学习（Watkins 1989）方法，用于求解第3.3节中介绍的马尔可夫决策过程。这项技术属于强化学习，是近似动态规划的一个分支。简单来说，这是一种基于价值的方法，通过搜索动作价值函数 Q（s_k， a）来表示在状态 s_k下采取动作 a 所能获得的未来期望奖励。

我们将 $V_\pi(s_k)$ 记为价值函数，该函数表示从状态 s_k 开始并遵循策略 π 时所获得的总期望奖励。该价值包括在决策时点 $k+1$ 获得的即时奖励，以及直至最终决策时点 T 之前未来将获得的期望奖励。我们的目标是找到动作数组，使整个过程中获得的总奖励最大化。根据贝尔曼方程（Powell 2007）， $V^*(s_k) = \max_{a_k \in \mathcal{A}(a_k)}\{R_{k+1} + \mathbb{E}[V(s_{k+1})|s_k, a]\}$，通过求解以下表达式即可执行最优动作 a_k，

$$a_k = \underset{a_k \in \mathcal{A}(s_k)}{\arg\max}\{R_{k+1} + \mathbb{E}[V(s_{k+1})|s_k, a]\} \tag{10}$$

© 0000 INFORMS

在我们所研究的问题中，由于顾客购物路径存在不确定性，智能体无法控制其最终所处状态的特征。因此，根据其当前状态 s_k 和所执行的动作 a_k，它们只能对将要访问的下一个状态产生有限影响。令 $Q_\pi(s_k, a_k)$ 表示一个函数（Q函数），用于表示智能体处于给定状态 s_k 并遵循策略 π 时，动作 a_k 的质量。我们可以将最优Q函数定义为 $Q^*(s_k, a_k)$。$V^*(s_k)$ 与 $Q^*(s_k, a_k)$ 之间的关系由以下表达式给出，

$$V^*(s_k) = \max_{a \in \mathcal{A}(s_k)} Q^*(s_k, a) \tag{11}$$

因此，可以通过以下表达式推导出我们旨在解决的问题的最优策略 π^* ，

$$\pi^*(s_k) = \arg\max_{a \in \mathcal{A}(s_k)} Q^*(s_k, a) \tag{12}$$

如果我们能够确定 Q^*，就能获得一个最优策略，用于控制在门店环境中受训的拣货代理的动作。可以采用文中提到的 Q学习方法确定该策略，并使用贝尔曼方程近似 Q^*。基于 Q函数的贝尔曼方程可递归表示为

$$Q(s, a) = R + \gamma \max_{a'} Q(s', a') \tag{13}$$

用于迭代近似 Q^* 的 Q学习方程如下：

$$Q^{new}(s, a) = Q(s, a) + \alpha\Big(R + \gamma \max_{a} Q(s', a) - Q(s, a)\Big) \tag{14}$$

其中，学习率 α 用于调整前后 Q值之间的差异。随着智能体对环境的探索不断深入，近似得到的 Q^{new} 值将收敛至 Q^*。为提取适用于门店环境的最优拣选策略，我们提出了算法2。

算法以随机的乐观值初始化Q函数（算法2，第2行）作为起点。这些乐观值有助于探索新状态。随后，重复执行一个循环过程若干回合（算法2，第3行）。首先，门店环境从初始状态 s_0 开始初始化（算法2，第4行）；在未达到最终状态时

算法2 Q学习过程

过程 QLEARNING(MDP, α, γ, ϵ)2: 使用随机的乐观值初始化 $Q(s,a)$；3: **对于**每个回合**执行**4: 将环境初始化为初始状态 s_0 （拣选员位于仓库）；5: **当** s_k 尚未达到最终状态（商店尚未关闭）时**执行**6: 遵循 ϵ-贪心策略，从当前状态 s_k 可用的动作中选择一个动作 a_k ；7: 执行动作 a_k，获得奖励 R_k，并观察到新状态 s_{k+1}；8: 通过执行 $Q(s_k,a_k) \leftarrow Q(s_k,a_k) + \alpha\Big(R_k + \gamma \arg\max_a Q(s_{k+1},a) - Q(s_k,a_k)\Big)$来更新Q函数；9: **如果** Q 收敛**则**10: 退出；11: **返回**Q

s_T （算法2，第5行），它会选择一个动作并执行，然后更新Q函数（算法2，第6至8行）。动作的选择采用 ϵ-贪心策略，即以 ϵ的概率随机选择一个动作，以概率$1-\epsilon$选择Q值最佳的动作。完成一个回合后，算法会验证Q函数是否收敛（即检查新旧Q值之间的最大差异是否低于收敛阈值）。如果已收敛，算法就返回Q函数；否则，继续进行下一个回合。

4.3. 示例

本文提出的解决方法会根据所假设的奖励形状收敛到与基于最短路径的策略存在显著差异的最优策略。图2对比了起点和终点位置相近的两条路径。在图中，节点的大小表示曾出现在各位置的顾客平均数量。我们观察到，在模拟回合期间，中央通道和顶部中间通道的商品位置人流更加拥挤。红色的**最短路径**策略（基于最短路径）距离较短，但会遇到更多顾客。另一方面，蓝色的**Q学习**策略（基于Q学习）会选择更长的绕行路线，但平均而言，遇到的顾客更少。

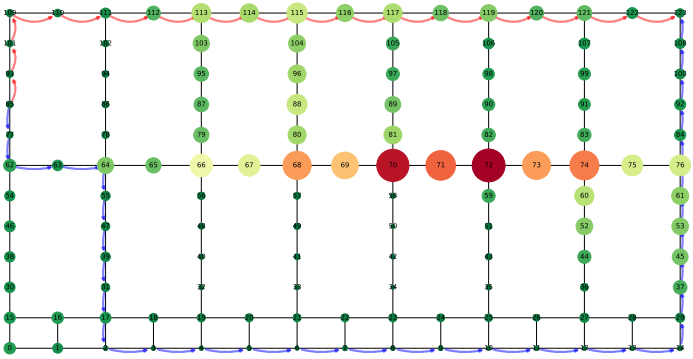


图2 最短路径（红色）和Q学习路径（蓝色）的示例。

蓝色路径是通过直接遵循根据Q值提取的策略得到的。图3展示了一个使用Q表（查找表）执行三个步骤的示例。

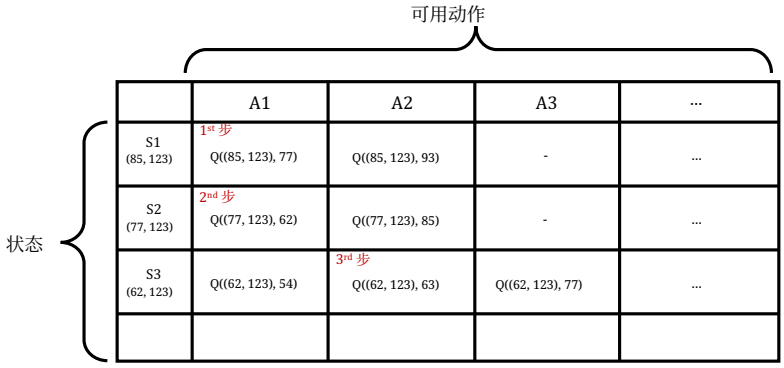


图3 示例中用于获得 Q学习路径的 Q表所提供的三步示例。

每一行代表一个状态，即由当前位置（节点 85）和目标位置（节点 123）组成的元组。每一列对应拣选员根据当前位置可采取的一种动作。例如，在第一步中，拣选员可以走到节点 77 或节点 93。由于节点 77 的 Q 值更大，拣选员会移动到位置 77。

5. 算法收敛

本节通过计算实验，探究所提出方法在合成实例上的有效性和效率。首先，我们介绍合成实例集的生成过程。其次，我们搜索训练参数，以找到

学习率、折扣率和 ϵ-贪婪度的良好组合，并验证所提出方法在不同问题规模下的收敛性。

5.1. 合成实例

为验证所提出的方法，我们通过考虑不同的仿真参数组合生成了合成实例。我们测试了四种不同规模的商店布局与三种不同店内顾客集中度的组合。因此，门店环境可分为极小、小、中和大四种，顾客购物清单中的产品则可能更集中在入口附近、中部或商店后方。顾客购物清单中的产品数量服从均匀分布，最少为1件，最多为10件。这些产品及其拣选顺序均为随机选择，顾客在不同产品位置之间移动时遵循最短路径。实体顾客的到达率设为2，线上订单的到达率设为0。2。在每个回合中，商店营业8小时，店内顾客最大数量 $|\mathcal{C}|$ 设为50。每个节点的顾客最大数量 m 为5，每位顾客需要30秒拣选一件产品。步行速度设为1米/秒。图1汇总了用于构建12个合成实例的参数集。关于合成实例中的奖励和惩罚，当拣选员与顾客处于同一节点时，拣选员会被扣3分。如果顾客能看到拣选员（即顾客位于拣选员可到达的节点），拣选员会被扣1分。拣货员步数每步扣1分，拣选到的产品则奖励100分。

5.2. Q学习训练与结果

尽管强化学习方法潜力巨大，但对其进行调优往往并非易事，而且耗时较长。这主要是因为测试不同参数组合的过程可能成本高昂且结果不稳定。为克服这些困难，我们测试了多种参数配置，以确定适用于每个合成实例的策略。每种配置均使用单线程运行了 5000 个回合。所有计算均在英特尔至强 @ 2.5 GHz 处理器上进行，并且

表1 构建生成实例所用的参数集。

商店布局	{极小、小、中、大}
顾客集中度	{入口、中、后部}
实体顾客到达率	{2}
在线订单到达率	{0.2}
营业时间（小时）	{8}
店内顾客最大数量	{50}
同一位置的顾客最大数量	{5}
商品拣选时间（秒）	{30}
步行速度（米/秒）	{1}
顾客相遇惩罚（p1）	{3}
顾客可见性惩罚 (p2)	{1}
步数惩罚 (p3)	{1}
拣选奖励 (p4)	{100}

表 2 用于构建训练配置的强化学习参数集。

学习率 (α)	$\{0.95, 0.97, 0.99\}$
折扣率 (γ)	$\{0.50, 0.70, 0.90\}$
ϵ-贪婪度（ϵ）	$\{0.01\}$

用于计算拣选员拣选顺序的数学求解器为 Gurobi 9.5。表 2 列出了用于构建各训练配置的参数集。

训练回合获得的累积奖励列于表 3 中。前两列“布局”和“路径”分别表示商店布局以及顾客采用的购物路径类型。第三列“最佳参数”表示获得相应行所列数值的参数组合（$\alpha|\gamma|\epsilon$）。第四至第六列分别列出在所有已测试的参数配置中获得的最低、平均和最高累积奖励。最后一列列出最近 50 个观测值的变异系数。

我们观察到，对于规模较大的商店布局，获得的累积奖励最大值与最小值之间的差距会增大。随着问题规模增加，选择

表3 训练流程中获得的累积奖励（5000 回合）。

布局	路径	最佳参数 $(\alpha\|\gamma\|\epsilon)$	累积奖励 Min.	Avg.	Max.	CV（最近 50 个观测值）
tiny	far	0.97|0.9|0.01	28989.54	29262.92	29440.60	0.11
tiny	中	0.99|0.7|0.01	27240.37	28563.90	29262.10	0.11
tiny	near	0.95|0.9|0.01	28325.40	28810.05	29073.35	0.10
小	far	0.97|0.9|0.01	23471.06	26828.68	28569.88	0.11
小	中	0.97|0.9|0.01	8571.05	21363.16	28362.71	0.10
小	near	0.97|0.9|0.01	11729.02	22830.18	28236.43	0.10
中等	far	0.99|0.9|0.01	-249624.48	-74201.81	26741.11	0.11
中等	中	0.99|0.9|0.01	-280708.50	-136167.61	26352.44	0.12
中等	near	0.97|0.9|0.01	-275092.86	-144711.93	25733.31	0.08
大型	far	0.95|0.9|0.01	-282733.08	-172079.58	22584.38	0.10
大型	中	0.95|0.9|0.01	-286259.32	-152702.76	17371.42	0.18
大型	near	0.97|0.9|0.01	-285686.24	-180264.63	21524.09	0.13

一组合适参数的重要性也随之提高。正如预期，在规模较大的商店中，获得的累积奖励较少，因为拣选员需要花更多时间穿行于更多节点之间，尤其是在需要返回备货区时。此外，我们观察到，对于最后50个观测值，所有实例规模下奖励的变异系数都较低，这进一步说明算法在所有实例中都已收敛。此时，为了更好地理解算法收敛情况，我们展示图4。

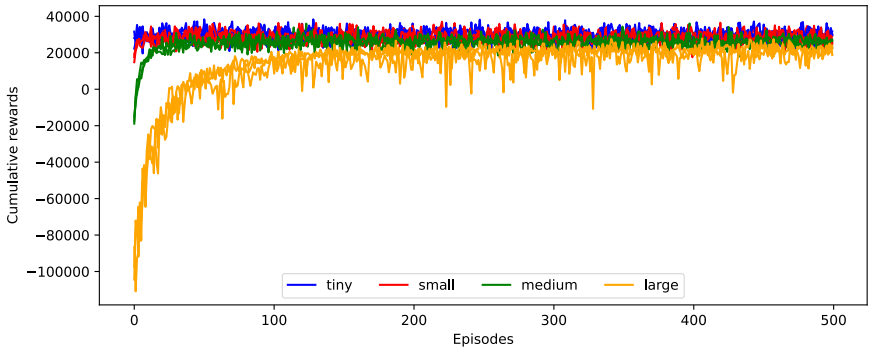


图4 合成实例中使用的最佳参数配置所获得的累积奖励（训练过程的前500个回合）。

累积奖励表明，算法只需经过几个回合便会收敛。正如预期，每个实例所考虑的商店规模与累积奖励的收敛时间有关。在中型布局的实例中，我们观察到算法经过几个回合后便收敛；而在大型布局的实例中，累积奖励则要在200多个回合后才趋于稳定。尽管如此，所提出的方法仍能成功得出无需大量计算能力的拣选策略。

6. 实际应用

本节详细介绍了一个大型欧洲零售商的实际应用案例。我们将所设计的方法应用于一家真实商店的布局，并使用顾客通过移动应用程序购物时收集的真实购物路径数据；该应用程序可帮助顾客记录他们在店内选取的商品。

所设计方法的一个非常重要的特点是，它不需要大量计算能力。Q学习方法的输出是一种以Q表（查找表）形式呈现的策略，可用于简单的移动设备（即手机和个人数字助理）。此外，还可以随着时间推移，通过离线或在线方式，利用从顾客和拣选员购物路径中收集的新数据来更新Q表，从而改进该策略。我们甚至可以计算针对一天中不同时段或一周中不同日期定制的Q表，以适应这些时段更有可能出现的顾客流动。

6.1. 现实实例

将所提出的方法应用于所研究的实际情境，第一步是对零售商店进行详尽的表示。由于实际中缺乏商品位置相关数据，这一过程尤其具有挑战性。如今，大型零售商正对实体店进行数字化，因为新收集的数据所带来的机遇十分可观。然而，本研究所考察的零售商在这方面仍处于起步阶段。在零售业中，由于各种原因（例如开展促销活动、在商品组合中引入新商品），许多产品可能每天都会调整位置，这使得可靠地维护

商品位置信息更加困难。尽管如此，我们仍将产品映射到通道，从而对实际商店进行了逼真的表示。图 5 展示了实际应用中所考察的商店布局和图。

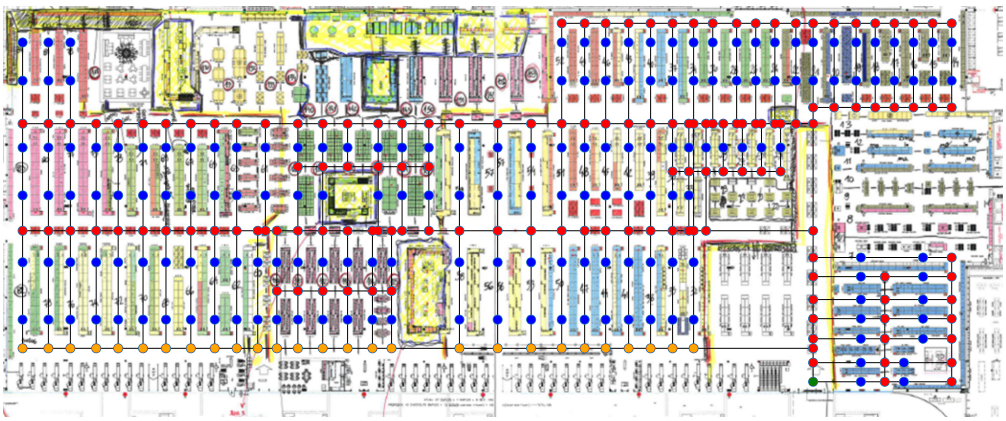
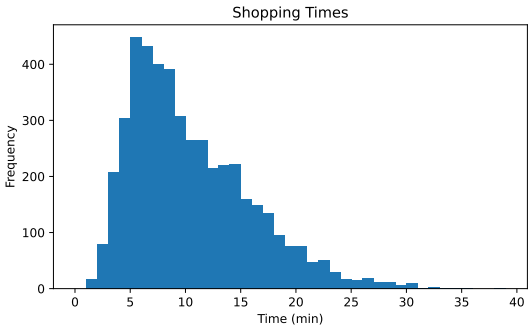


图5 实际应用中所考虑的商店布局及图示。

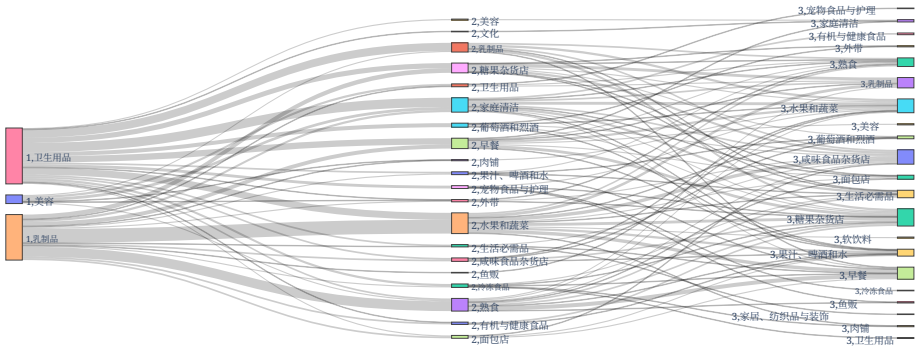
在此图中，以红色表示的节点是通道交叉口，以蓝色表示的节点是商品位置。我们选择在每条通道内最多设置两个商品节点，因为这样已经能够很好地指示每件商品的位置。在所考虑的图中，顾客可以通过绿色表示的节点进入商店，并在 traversing their shopping path 后，随机选择橙色表示的出口节点之一离开。实体顾客拣选的商品序列基于零售商通过移动应用程序收集的真实世界数据。为了计算顾客购物路径中每对商品之间的最短路径，我们采用了 Dijkstra 算法的一个实现。其余参数根据与零售店经理的讨论进行设置。实体顾客的到达率 λ^{Store} 设置为 10，线上订单的到达率 λ^{Online} 设置为 0。3。商店内允许的顾客最大数量设置为 300，在实际应用中我们没有设置每个节点的顾客最大数量。每位顾客花费 30 秒拣选一件商品，并以 1 米/秒的速度行走。商店营业 8 小时。拣货代理每拣选一件商品可获得 50 分；每当在同一节点遇到一名顾客时扣 3 分；每当在相邻节点看到一名顾客时扣 1 分。



(a) 购物时长直方图。



(b) 最拥挤的节点。



(c) 从卫生用品、乳制品和美容类商品开始挑选 3 件商品的顾客所到访的品类。

图 6 真实商店仿真环境指标。

为了进一步描述用于训练拣选员智能体的真实商店仿真环境，我们展示了图 6，其中包括顾客购物时长的分布、商店内最拥挤区域的表示（每个决策时点的平均顾客数量），以及顾客仅拣选 3 件商品时执行的品类路径示例（以便更好地进行可视化），这些顾客分别从卫生用品、美容和乳制品品类开始。

我们观察到，大多数顾客在商店内停留的时间为5至15分钟。最拥挤的节点位于图像左侧，主要对应食品，以及图像底部、靠近收银员的商店前方过道。拣选员出发的订单准备区位于图片右上方。该区域通常不太拥挤，因为其中包含工作工具和各种DIY活动所需的材料。我们还观察到，即使顾客的购物清单中共享少量节点/商品，他们所遵循的购物路径也各不相同。

6.2. 在现实情境中训练

现实案例中采用的训练过程与第5.2节所介绍的过程类似。首先，我们设置了多个训练环节，考虑强化学习参数的不同组合，即学习率 α、折扣率 γ和探索概率 ϵ。所使用的强化学习参数集如图4所示。

表 4　　rei 套装　强化学习参数，以便为现实实例构建训练配置。

学习率（α）	$\{0.1, 0.5, 0.9, 0.95, 0.97, 0.99\}$
折扣率 (γ)	$\{0.1, 0.5, 0.9, 0.95, 0.97, 0.99\}$
ϵ-贪婪度 (ϵ)	$\{0.01, 0.03, 0.05, 0.07, 0.09\}$

每种配置共运行了 30000 个回合。每 1000 个回合保存一次策略和平均累积奖励。图 7 展示了在真实零售商店环境相关训练结果中获得的 10 个最佳策略（按平均奖励计算）。每

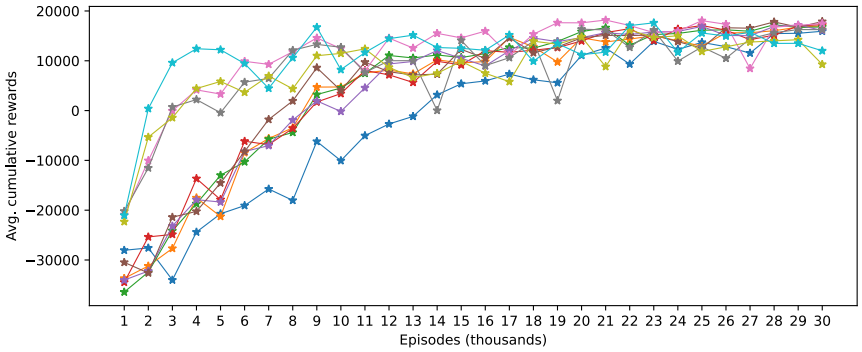


图 7 真实门店环境中所有参数配置下排名前10 的策略的平均累积奖励。

1000 回合后，我们重新开始计算平均累积奖励。请注意，在最初的 1000 回合后，一些策略的表现已经有所提升，但尚未收敛。总体而言，在训练流程中，最优策略在经过 20000 回合后收敛。少数策略即使经过 30000 回合，也未能达到最优策略的表现。

© 0000 INFORMS

在某些强化学习问题中，训练阶段获得的平均累积奖励（考虑随机动作）可能与测试阶段获得的奖励不同（测试阶段始终根据 Q表选择最优动作）。关于如何评估这些方法，以及在生产环境或测试阶段（此时不会对拣选策略进行更新）是否应保留 ϵ-贪婪探索，存在很大争议。在某些问题中，保留这种探索机制可能是合理的，因为智能体可能会因遇到未见状态而陷入停滞。Mnih 等人（2015）在游玩 Atari 游戏时成功采用了这种做法。在求解路径问题时，如果不考虑随机性，非最优策略可能会受到循环的影响。因此，我们认为在实际使用策略时应保留 ϵ 参数。

6.3. 管理启示

前几节展示的结果表明，该方法能够学习可应用于真实零售商店环境的拣选策略。事实上，我们可以观察到，随着拣选员获得越来越高的累积奖励，其学习能力也在不断提升。尽管如此，我们不禁要问，为了拣选一组产品，是否有必要做这么多工作，还是只需采用更简单的策略就足够了，例如仅基于距离度量的最短路径策略。

促使我们着手解决这一问题的几个核心问题仍未得到解答。我们能否改善顾客体验？这对拣选订单数量有何影响?

为回答这些问题，我们衡量了四种不同拣选策略在业务相关指标方面的表现，例如拣选的订单／产品数量以及与顾客相遇的次数。**Q-learning（QL）**策略：采用所设计的Q-learning进行训练，旨在寻找一种良好的折中方案，在拣选订单的同时避免与顾客相遇。

最短路径（SP）策略基于最短路径策略（采用 Dijkstra 算法），旨在拣选尽可能多的订单，并忽略与顾客相遇。每一步都通过求解最短路径问题获得，其中弧的权重按其长度设置。

短视策略（MP）基于短视行为的策略，旨在沿着顾客最少的节点前进，同时不远离目标拣选位置。

拥挤节点（CN）基于各节点顾客平均数量的静态指标，尽量减少与顾客相遇次数的策略。每一步都通过求解最短路径问题来确定，其中弧的权重为经过这些弧的顾客平均数量。

此外，对于每种策略，我们考虑了两种路径成本，用于确定订单中商品的拣选顺序。因此，每当有在线订单到达时，均可使用考虑距离或拥挤度指标（起点与终点节点之间的顾客平均数量）的成本矩阵来求解 SRP。这两种路径规划依据分别称为**弧距离**和**弧拥挤度**。这会影响订单中商品的拣选顺序。

表 5 列出了采用所考虑的各路径规划依据时，四种策略对应的指标。

表5 最短路径（SP）、拥挤节点（CN）、短视策略（MP）和 Q-learning（QL）策略在 500 个回合中得到的平均指标。

路径规划依据	策略	Avg. 奖励	Avg. #订单	Avg. #产品	Avg. #遭遇
弧距离	QL	16582.12	41.03	487.01	1025.39
弧距离	SP	15388.14	46.56	553.79	2072.22
弧距离	CN	16085.83	38.51	456.64	745.71
弧距离	MP	15294.34	47.65	564.76	2134.58
	Avg.	15838.69	43.45	515.55	1494.40
弧拥挤度	QL	16997.61	41.49	491.58	1018.92
弧拥挤度	最短路径	14991.66	44.45	527.50	1868.61
弧拥挤度	CN	16845.25	39.84	470.69	768.99
弧拥挤度	MP	15244.25	45.66	542.28	1891.15
	Avg.	16019.69	42.86	508.01	1386.92

© 0000 INFORMS

分析使用弧距离生成路线序列的情况可以发现，**SP**和**MP** 是运营效率最高的策略，平均分别拣选 46.56 和 47.65 个订单。**QL** 和**CN** 在拣选订单数量方面效率较低，平均分别拣选 41.03 和 38.51 个订单。这些策略显然侧重于改善顾客体验，与**SP** 和**MP**策略相比，将与顾客相遇的平均次数降低了 50% 以上。**CN** 策略实现了最少的与顾客相遇次数，但其拣选的订单数量也是所有策略中最低的。**QL** 策略大幅减少了与顾客相遇的次数，同时没有严重影响拣选的订单数量。

观察使用弧拥挤度指标生成路线序列后得到的结果，可以发现，各策略在运营效率和顾客体验方面的相对表现相似。然而，这些路线序列的平均拣选订单数量有所减少，降幅为1.35%，从平均43.45个订单降至42.86个。与顾客相遇的次数略有改善，减少了7.19%，从平均1494.40次降至1386.92次。因此，以弧拥挤度作为路径规划依据，有助于所有策略在顾客体验方面取得小幅改善，代价是拣选订单数量略有减少。

我们发现，运营效率最高的策略是基于遵循最短距离路径的策略（**最短路径** 和**MP**）。就提供最佳顾客体验而言，**CN** 策略的平均顾客相遇次数最低。不过，这些策略都只专注于一个目标。**最短路径** 和**MP** 策略无法避开顾客，而**CN**策略则没有意识到最终目标是拣选订单。**QL**策略则在拣选较多订单的同时，与较少顾客相遇，实现了良好的折中。

这些结果中一个有趣的细节是，当路径规划依据从距离改为拥挤度时，**QL** 必须重新训练。请注意，我们处理的问题会受到状态数量方面的“维数灾难”影响。鉴于路线序列

发生了变化，不同的当前位置与目标位置组合（状态）开始更频繁地出现。由于使用弧距离训练的**QL** 策略并未充分访问这些状态，因此必须在路线序列考虑弧拥挤度的环境中训练该策略。

从策略中得到的若干起讫点对，可以观察到拣货代理学到的有趣行为。图8展示了一个示例，说明**最短路径**、**CN** 和**QL** 策略所遵循的路径。

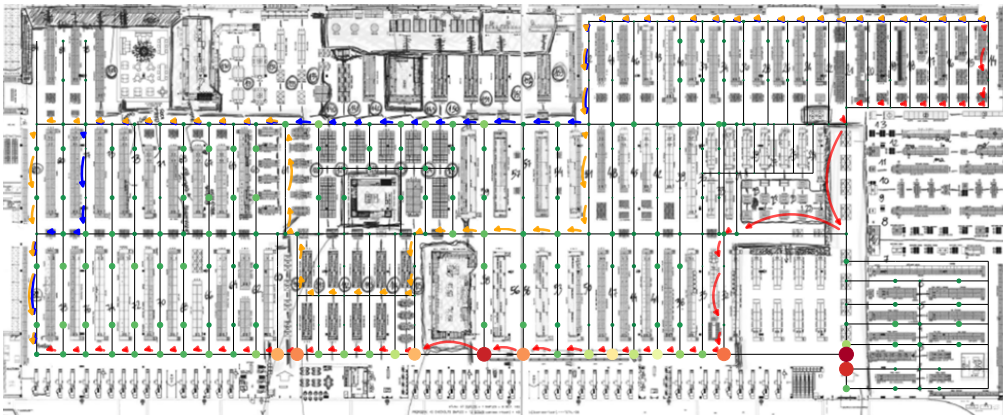


图8 最短路径（红色）、CN路径（橙色）和QL路径（蓝色）的示例。

尽管**最短路径**策略执行的步数较少（行进距离较短），但最终会经过最拥挤的节点。另一方面，**QL** 策略遵循的路径较长，但会尽量避开拥挤节点，因此与顾客相遇的次数较少。**CN** 策略走的路径最长，因为它试图不惜一切代价避开顾客。

7. 结论

本文提出了在线零售中的一个新问题 diPRP，并采用强化学习方法加以解决。在 2020 年的疫情背景下，这一问题尤为突出：大多数提供网上购物服务的零售商都转而采用新的店内拣货作业。这些新的拣货作业不会只是昙花一现，而将成为全渠道零售商的重要履约方式。

© 0000 INFORMS

我们提出的解决方法综合运用了多种技术，融合了数学规划、仿真和机器学习等领域，以获得一种在已拣选产品数量方面具有竞争力、同时能尽量减少与顾客相遇次数的店内拣选策略。我们的研究表明，强化学习可以成为有效解决新型管理问题的重要方法。

我们设计的解决方案通过训练拣选员智能体执行购物路径，以减少拣选作业对店内顾客购物体验的干扰。在合成实例和一家大型欧洲零售商的现实实例上进行的计算实验表明，拣选员智能体能够在合理数量的回合内学会高效的拣选策略。我们的方法有一项重要特点：所需的计算能力较低，且做出决策所需的计算时间极短。

此外，基于现实案例研究，我们提供了有趣的数值结果和解决方案可视化，以支持若干管理启示。不出所料，顾客的购物顺序往往会导致店内区域拥挤。由于顾客在店内的移动动态且难以预测，这些拥挤的店内区域也难以预判。不过，我们观察到，设计的拣选策略会避开这些区域，以减少与顾客相遇的次数。这与基于最短路径的策略不同；后者按逻辑可能多拣选一些商品，但在我们的仿真中，受到干扰的顾客数量几乎翻倍。我们基于强化学习的方法能够更好地控制运营效率与顾客体验之间的权衡。这一结果或可助力零售商进一步推进其全渠道履约策略。

在这个问题中，还可以进一步引入许多改进。例如，制定协作式拣选策略，以协调这一问题中的拣选团队，将是一项很有意思的工作，尤其是在能够评估顾客对与拣选员相遇次数的感知时。此外，还可以改进仿真环境，例如纳入购物车容量以及在拥挤的通道中遇到的行进减速。此外，

探索更丰富的状态变量或许能提升拣选策略的质量。例如，每条通道中遇到的顾客数量，也可以进一步影响拣选员的在线决策（在这种情况下，状态空间可能会受到所谓的“维数灾难”更严重的影响）。最后，我们想强调，本文提出的方法有多种应用场景。未来，各类活动将实现自动化，例如使用机器人清洁公共空间、在仓库和商店拣选产品，以及使用自动驾驶车辆。这些技术很可能会部署在与人类共用的狭窄空间中。由于这类自动化设备通常不太美观，我们的方法未来或许可以成为一种有趣的替代方案。

致谢

本研究成果得到了欧盟地平线2020计划——欧盟2014–2020年研究与创新框架计划的资助，资助协议编号为952060。

参考文献

Adhi P、Hough G、Calais S、Lange T、LenzenC（2022）重新构想实体店在全渠道配送网络中的作用。网址：https://www.mckinsey.com/industries/retail/our-insights/重新构想实体店在全渠道配送网络中的作用。

Ardjmand E、Shakeri H、Singh M、Sanei Bajgiran O（2018）在波次拣选仓库中通过多名拣选员将订单拣选完工时间降至最低。《国际生产经济学》206：169–183，ISSN 0925-5273。

Barbee J, Jayakumar J, Touse S, Venkataraman K (2021) 零售业对速度的需求：释放全渠道配送价值。URL https://www.mckinsey.com/industries/retail/our-insights/retails-need-for-speed-unlocking-value-in-omnichannel-delivery.

Berbeglia G, Cordeau JF, Laporte G (2010) 动态取送问题。*European Journal ofOperationalResearch* 202(1):8–15.

Boros P, Feh´er O, Lakner Z, Niroomand S, Vizv´ari B（2015）从所有者视角对超市重新布局进行建模。《运筹学年鉴》238(1-2):27–40。

Cambazard H, Catusse N（2018）平面直角斯坦纳树和直角旅行商问题的固定参数算法。《欧洲运筹学杂志》270(2):419–429，ISSN 0377-2217。

Caro F, K¨ok AG, Mart´ınez-de Alb´eniz V（2020）零售运营的未来。《制造与服务运营管理》22(1):47–58。

Chabot T, Coelho LC, Renaud J, Cˆot´e JF（2018）窄通道中订单拣选的数学模型、启发式算法和精确方法。《运筹学会杂志》69(8):1242–1253。

Chabot T, Lahyani R, Coelho LC, Renaud J（2017）考虑重量、易碎性和类别约束的订单拣选问题。《国际生产研究杂志》55(21):6361–6379。

Chen J, Liang Y, ShenH, Shen ZJM, Xue M（2021）智能全渠道零售中的线下渠道规划。《制造与服务运营管理》。

Chen X、Li Y、Hu T（2015）基于遗传算法解决超市购物路径规划问题。*2015 IEEE/ACIS* 第*14*届计算机与信息科学国际会议（*ICIS*），529–533。

Cornu´ejols G、Fonlupt J、Naddef D（1985）图上的旅行商问题及一些相关的整数多面体。数学规划33：1–27。

de Koster R、Le-Duc T、Roodbergen KJ（2007）仓库订单拣选的设计与控制：一篇文献综述。《欧洲运筹学杂志》182(2)：481–501。

Dennis S（2018）实体零售并未消亡，消亡的是乏味零售。网址 https://www.《福布斯》.com/sites/stevendennis/2018/03/19/physical-retail-is-not-dead-boring-retail-is-understanding-retails-great-bifurcation。

Gao F, Su X (2017a) 采用线上购买、门店自提模式的全渠道零售运营。管理科学63(8):2478–2492。

Gao F, Su X (2017b) 全渠道零售的线上与线下信息。制造与服务运营管理19(1):84–98。

Glaeser CK, Fisher M, Su X (2019) 最优零售选址：实证方法与实践应用。《制造与服务运营管理》21(1):86–102.

Gong Y, Koster RD (2008) 面向在线零售商的基于轮询的动态订单拣选系统。*IIE Transac-tions* 40(11):1070–1082.

Hildebrandt FD, Thomas B, Ulmer MW (2021) 行动所在之处：让强化学习应用于随机动态车辆路径问题!

Hirpara S, Parikh PJ (2021) 考虑购物者路径的零售设施布局。《计算机与工业工程》154:106919.

Hui SK, Inman JJ, Huang Y, Suher J (2013) 店内行走距离对非计划消费的影响：在移动促销策略中的应用。《营销学杂志》77(2):1–16.

Jia S, Karp J,Ravi R,Tayur S (2021) 全渠道履约的有效在线订单接收策略。《制造与服务运营管理》.

Kullman ND, Cousineau M, Goodson JC, Mendoza JE (2022) 使用电动汽车提供动态网约车服务。《交通科学》56(3):775–794.

Larson JS, Bradlow ET, Fader PS（2005）对超市购物路径的探索性考察。《国际营销研究杂志》22(4):395–414，ISSN 0167-8116。

Letchford AN, Nasiri SD, Theis DO（2013）斯坦纳旅行商问题及相关问题的紧凑公式。《欧洲运筹学杂志》228(1):83–92，ISSN 0377-2217。

Li J, Ma Y, Gao R, Cao Z, Lim A, Song W, Zhang J（2021）用于求解异构容量车辆路径问题的深度强化学习。《*IEEE*控制论汇刊》1–14，ISSN 2168-2275。

Li YM, LinLF,Ho CC（2017）一种用于店内购物支持的社交路径推荐机制。《决策支持系统》94:97–108，ISSN 0167-9236。

Lu W, McFarlane D, Giannikas V, Zhang Q（2016）仓库运营中的动态订单拣选算法。《欧洲运筹学杂志》248(1):107–122，ISSN 0377-2217。

Masae M, Glock CH, Grosse EH（2020年）仓库中的订单拣选员路径规划：系统性文献综述。《国际生产经济学》224:107564。

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, SadikA, AntonoglouI, KingH, Kumaran D, WierstraD, Legg

S, Hassabis D（2015）通过深度强化学习实现人类水平的控制。《自然》 518(7540):529– 533, ISSN 00280836。

Nazari M, Oroojlooy A, Snyder LV, Tak´ac M（2018）运用强化学习求解车辆路径问题。*NeurIPS*。

Nierynck R（2020）即时零售：开创下一代配送：Delivery Hero。网址：https: // www.deliveryhero.com/blog/quick-commerce/。

Ojeda Rios BH, Xavier EC,Miyazawa FK, Amorim P, Curcio E, Santos MJ（2021）近期动态车辆路径问题综述。《计算机与工业工程》160:107604，ISSN 0360-8352。

Olsen TL, Tomlin B（2020）工业4.0：运营管理的机遇与挑战。制造与服务运营管理22(1):113–122。

Pansart L, Catusse N, Cambazard H（2018）订单拣选问题的精确算法。《计算机与运筹学》100:117–127，ISSN 0305-0548。

Pillac V, Gendreau M, Gu´eret C, Medaglia AL (2013) 动态车辆路径问题综述。《欧洲运筹学杂志》225(1):1–11。

Pires M, Silva E, Amorim P(2021) 杂货店后场布局问题求解。《国际生产研究杂志》59(3):772–797。

Powell WB (2007)近似动态规划：解决维数灾难（威利概率与统计丛书）（美国：Wiley-Interscience），ISBN 0470171553。

Powell WB (2019) 随机优化的统一框架。《欧洲运筹学杂志》275(3):795–821。

Psaraftis HN, Wen M, Kontovas CA (2016) 动态车辆路径问题：三十年发展历程。《网络》67(1):3–31。

Quader S, Castillo-Villar KK (2018) 考虑存储分配和路径规划启发式算法的增强型多巷道订单拣选系统设计。《机器人与计算机集成制造》50:13–29。

Ratliff HD, Rosenthal AS（1983）矩形仓库中的订单拣选：旅行商问题的一个可求解情形。*Operations Research* 31(3):507–521。

Richter F（2021）疫情加速了向在线零售的转变。网址 https://www.statista.com/chart/14011/ 电子商务占零售总额的比例/。

Ritzinger U、Puchinger J、Hartl RF（2016）关于动态与随机车辆路径问题的综述。《国际生产研究杂志》54(1):215–231。

Rodr´ıguez-Pereira J、Fern´andez E、Laporte G、Benavent E、Mart´ınez-Sykora A（2019）斯坦纳旅行商问题及其扩展。《欧洲运筹学杂志》278(2):615–628，ISSN 0377-2217。

Roodbergen KJ、de Koster R（2001）在设有中间通道的仓库中为订单拣选员规划路径。《欧洲运筹学杂志》133(1):32–43。

Roodbergen KJ、Koster R（2001）设有多条横向通道的仓库路径规划方法。《国际生产研究杂志》39(9):1865–1883。

Roodbergen KJ、Vis IF、Jr GDT（2015）同时确定仓库布局和控制策略。《国际生产研究杂志》53(11)：3306–3326。

Savelsbergh M、Van Woensel T（2016）50周年特邀文章——城市物流：挑战与机遇。《交通科学》50(2)：579–590。

Scholz A、Henn S、Stuhlmann M、W¨ascher G（2016）单订单拣选员路径规划问题的新数学规划公式。《欧洲运筹学杂志》253(1)：68–84，ISSN 0377-2217。

Theys C、Br¨aysy O、Dullaert W、Raa B（2010）使用旅行商问题启发式算法为仓库中的订单拣选员进行路径规划。《欧洲运筹学杂志》200(3)：755–763，ISSN 0377-2217。

Ulmer MW, Goodson JC, Mattfeld DC, Thomas BW (2020年) 关于随机动态车辆路径问题的建模。《*EURO*交通与物流杂志》9(2):100008。

van Gils T, Ramaekers K, Braekers K, Depaire B, Caris A (2018) 通过整合存储、批处理、分区拣选和路径规划策略决策，提高订单拣选效率。《*International Journal of Pro- duction Economics*》197:243–261，ISSN 0925-5273。

Vyt D, Jara M, Mevel O, Morvan T, Morvan N（2022）线上下单、门店自提零售场景中便利性的影响：基于消费者的方法。《国际生产经济学》248:108491，ISSN 0925-5273。

© 0000 INFORMS

Watkins CJCH（1989）从延迟奖励中学习。博士论文，国王学院，牛津。