文章编号:1009-6744(2020)05-0072-07 中图分类号:U231 文献标志码:A
DOI:10.16097/j.cnki.1009-6744.2020.05.011

基于强化学习的城轨信息发布策略研究

贾飞凡[1a,1b],蒋 熙[1a],李海鹰*[1a],于雪峤[2]
(1.北京交通大学 a. 轨道交通控制与安全国家重点实验室,b. 交通运输学院,北京 100044;
2.中国铁道科学研究院集团有限公司 运输与经济研究所,北京 100081)

摘 要: 通过信息发布影响乘客选择行为进而改变路网客流分布,是从需求侧缓解拥堵问题的重要手段之一.本文提出基于强化学习的城市轨道交通信息发布策略生成方法,根据路网各区间客流满载率提取系统状态,再根据系统状态在学习器生成由各OD推荐路径组成的信息发布动作,对乘客进行信息发布;通过发布信息后路网系统状态变化,评估获得实施信息发布动作的奖励值.依托城市轨道交通客流分布动态仿真系统,使用 Q-learning 算法进行训练,获得最优信息发布策略.以实际路网为例进行算例验证,通过对比有无信息发布情景得到,在有信息发布情景下路网客流拥堵情况得到了较大缓解.

关键词: 智能交通;信息发布;强化学习;客流诱导; Q-learning

Information Release Strategy of Urban Rail Transit Based on Reinforcement Learning

JIA Fei-fan[1a, 1b], JIANG Xi[1a], LI Hai-ying[1a], YU Xue-qiao[2]
(1a. State Key Lab of Rail Traffic Control & Safety, 1b. School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China; 2. Institute of Transportation and Economy, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: Guidance information can change the choice behavior of passengers and thus network passenger flow distribution. Information release is one of the key measures to alleviate the congestion problem from demand ideas. A method is proposed to generate an information release strategy based on reinforcement learning. The system state is extracted based on the load rate of passenger flow in each section in the network. The information release action is composed of the recommended paths of each OD. The reward value of implementing an information release action is evaluated according to system state change. By an urban rail transit dynamic passenger flow simulation system, the Q-learning algorithm is employed to obtain optimal information release strategy. A practical network is taken as an example to verify the proposed method. It was found that network congestion can be alleviated by using the proposed information release strategy.

Keywords: intelligent transportation; information release; reinforcement learning; passenger flow guidance; Q-learning

0 引 言

随着城市轨道交通客流需求激增,拥堵问题日趋严重.运营者开始采用各种措施以缓解路网中客流拥堵.从供给侧角度,可通过压缩发车间隔、增加列车编组等措施提高输送能力;从需求侧角度,通过信息发布影响乘客行为是缓解拥堵的有效措施之一.

交通信息发布是指运营者在出行者出行前或

收稿日期:2020-05-22 **修回日期:**2020-08-02 **录用日期:**2020-08-03
项目基金:中央高校基本科研业务费专项基金/ Fundamental Research Funds for the Central Universities of Ministry of Education of China(2018YJS194).
作者简介:贾飞凡(1993-),男,安徽宣城人,博士生. ***通信作者:**hyli@bjtu.edu.cn

出行过程中提供出行信息的服务.交通领域关于信息发布策略的研究刚刚起步.陈芳等[1]基于双层规划模型原理建立了可变信息标识信息发布策略优化模型.孙智源等[2]建立了诱导与控制协同的双层规划模型.曹守华等[3]研究了在交通流诱导条件下的路网随机平衡问题,提出两种不同条件下建设交通流诱导系统的双层规划模型.杨泳等[4]通过仿真方法模拟信息影响下的驾驶员路径选择、匝道排队等行为,研究不同信息发布策略的效果.目前关于城轨信息发布问题的研究较少.Yin等[5]同样建立双层规划模型,使用仿真进行客流分配作为下层模型,并使用遗传算法求解上层模型.

随着大数据时代的到来,运营者积累了海量乘客出行数据.数据驱动的机器学习方法因其自学习、高效率等优势在众多领域得到广泛应用.本文针对城轨信息发布问题,提出基于强化学习的信息发布策略算法,为决策者对乘客制定个性化的信息服务提供辅助决策支持.

1　问题分析

随着数据存储挖掘等技术的飞速发展,运营者积累了海量的乘客出行信息,能够挖掘乘客个体出行规律,借助定位技术可以进一步获得出行路径.粗略无针对性的信息于个体的价值比较随机,乘客更想获取于自己有针对性的信息,因此,定制信息是未来信息发布的一个方向.本文针对OD级的推荐路径发布问题进行研究,为定制信息服务提供决策支持.

信息发布的目的是通过信息影响乘客路径选择行为,改变路网客流分布,从而缓解拥堵.区间断面客流是表征路网客流分布的关键状态,任意时段 T 区间 m 断面客流量 $L_m(T)$ 计算公式为

$$L_m(T)=\sum_{w=1}^{W}\sum_{k=1}^{K_w}Q_w(t\leqslant T)\cdot p_k\cdot\delta_k^m(T) \tag{1}$$

式中:W 为所有OD集合;K_w 为第 w 个OD对的有效路径个数;$Q_w(t\leqslant T)$ 为第 w 个OD对在 T 时段及其以前进入路网的客流量;p_k 为选择第 w 个OD对的第 k 条路径的乘客占第 w 个OD对客流量的比例;$\delta_k^m(T)$ 为选择第 w 个OD对的第 k 条路径的乘客在 T 时段是否经过区间 m 的0-1变量.影响断面客流量的主要因素是OD客流和各OD路径选择比例.OD客流的时空分布动态变化通过乘客个体选择,出行行为与列车运行交互影响断面客流在时间上的变化.

向乘客推荐出行路径可以影响乘客路径选择,但不能强制改变选择结果.乘客个体选择具有自主性、随机性和动态性,因此,区间断面客流与信息诱导存在复杂的非线性关系.准确描述信息影响下乘客选择行为是信息发布策略的基础,把握信息与路网各区间断面客流分布变化之间的关系是制定策略的关键.针对这两个问题,本文采用仿真模拟列车运行过程与信息影响下乘客选择行为与出行过程以获得路网客流分布,利用强化学习方法寻找何种客流分布状态下应该发布何种信息的映射关系,以获得信息发布策略.

Q-learning算法[6]其因免模型、无监督等特征适用于信息发布问题.本文设计基于Q-learning的信息发布策略生成算法,诱导客流从而缓解客流拥堵.学习器根据路网各区间断面客流提取系统状态产生相应的信息发布动作并输入仿真系统,模拟信息影响下乘客决策,出行过程和列车运行过程,得到区间断面客流变化;学习器根据断面客流分布变化情况评估信息发布动作的奖励值,进行Q-table更新,并根据新的客流分布情况产生相应的信息发布内容,多次训练迭代后,即可获得最优信息发布策略,如图1所示.

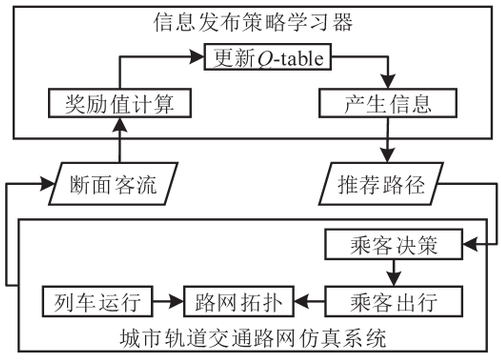


图1　基于强化学习的信息发布策略框架

Fig. 1　Framework of information release based on reinforcement learning

2　基于Q-learning的信息发布策略算法

2.1　要素构建

强化学习包含3个关键要素:状态S、动作A、奖励R,在信息发布问题中路网客流分布,即为系

统状态，用路网各区间单位时段内的平均满载率描述路网客流分布.假设 l_m^t 为第 m 个区间在第 t 个时段的平均满载率，$l_{\max}$ 为运营者设定的满载率控制上限.满载率是一个连续变量，若直接使用满载率进行状态描述，则状态集合是无限的.为缩小状态规模，系统状态使用各区间是否高于控制上限的0-1变量组成的向量表示，即

$$\boldsymbol{s}(t)=\left[C(l_1^t),C(l_2^t),\cdots,C(l_m^t),\cdots\right] \tag{2}$$

$$C(l_m^t)=\begin{cases}0, & l_m^t\leqslant l_{\max}\\ 1, & l_m^t>l_{\max}\end{cases} \tag{3}$$

式中：$\boldsymbol{s}(t)$ 为第 t 个时段的路网状态；$C(l_m^t)$ 为第 m 个区间第 t 个时段的状态值，当满载率大于控制上限则取1，否则取0.

信息发布中，动作即为运营者向不同OD乘客推荐的出行路径.在第 t 个时段内，对各OD乘客推荐的相应路径索引组成的向量，即为第 t 时段的信息发布动作.

$$\boldsymbol{a}(t)=\left[I_1^t,I_2^t,\cdots,I_w^t,\cdots\right] \tag{4}$$

$$0\leqslant I_w^t\leqslant K_w \tag{5}$$

式中：$\boldsymbol{a}(t)$ 为第 t 个时段发布的信息；I_w^t 为在第 t 个时段给第 w 个OD对推荐的路径索引，若不推荐任何路径则为0.

信息发布后，若区间满载率变低，说明执行该动作获得效果较好，值得被“奖励”，被“奖励”的动作被选择的概率会增加.信息发布需要在调控功能和服务功能间进行权衡，采用分段函数设定奖励值，在兼顾乘客出行效率的基础上，尽量调节路网客流分布达到均衡状态.此外，路网中各区间在拓扑网络和客流网络中的重要程度不同，借鉴复杂网络中介数的概念，设计权重函数评估各区间的重要程度以计算总奖励.综上，设立奖励函数为

$$\omega_m=\frac{\sum_{w=1}^{W}\sum_{k=1}^{K_w}d_k^w\cdot Q_w\cdot\delta_k^m}{\sum_{m=1}^{M}\sum_{w=1}^{W}\sum_{k=1}^{K_w}d_k^w\cdot Q_w\cdot\delta_k^m} \tag{6}$$

$$r_m^t=\begin{cases}100, & l_m^t\leqslant l_{\max}\\ \left(l_{\max}-l_m^t\right)\times 100, & l_m^t>l_{\max}\end{cases} \tag{7}$$

$$r(t)=\sum_{m=1}^{M}r_m^t\cdot\omega_m \tag{8}$$

式中：ω_m 为第 m 个区间的权重；d_k^w 为第 w 个OD对的第 k 条路径所经过的区间数；Q_w 为第 w 个OD对的历史平均客流量；δ_k^m 为0-1变量，若第 w 个OD对的第 k 条路径经过区间 m ，则为1，否则为0；M 为路网内总区间数；r_m^t 为第 t 个时段发布信息后第 m 个区间获得的奖励值；$r(t)$ 为第 t 个时段发布信息后获得的总奖励值.

2.2 信息影响下的乘客选择

使用经典随机效用模型(RUM)描述信息发布前的乘客选择行为.进行路径选择时，根据路径特性评估不同路径效用，选择效用最大的路径作为本次出行路径.

$$U_{w,k}=T_{w,k}^{\text{inout}}+T_{w,k}^{\text{wait}}+\zeta\cdot T_{w,k}^{\text{invehicle}}+T_{w,k}^{\text{transfer}}+\tau\cdot n_{w,k}+\sigma_{w,k} \tag{9}$$

$$p_{w,k}=\frac{\exp\left(-\theta U_{w,k}\right)}{\sum_{k=1}^{K_w}\exp\left(-\theta U_{w,k}\right)} \tag{10}$$

式中：$U_{w,k}$ 为OD对 w 第 k 条路径的效用；$T_{w,k}^{\text{inout}}$ 为进出站走行时间；$T_{w,k}^{\text{wait}}$ 为候车时间；$T_{w,k}^{\text{invehicle}}$ 为随车运行时间；$T_{w,k}^{\text{transfer}}$ 为换乘走行时间；$n_{w,k}$ 为换乘次数；ζ 为拥挤度惩罚系数；τ 为换乘惩罚系数；$\sigma_{w,k}$ 为随机项；$p_{w,k}$ 为第 w 个OD对的第 k 条路径被选择的概率；θ 为效用系数.

接收到信息后，乘客首先判断当前路径和推荐路径是否一致.若不一致，需根据自身位置判断是否具备改变路径的条件，若不具备则保持当前路径；否则，需进一步评估是否接受推荐路径.由于乘客的异质性，不同乘客面对信息的态度也不同.已有通过SP调查、实验室实验等研究出行者对信息的态度[7].整合已有研究成果，使用高斯分布 $N(0.7,1)$ 表示乘客对信息接受率分布.乘客接受信息的流程如图2所示.

本文依托城市轨道交通客流动态仿真系统[8]进行二次开发，增加信息发布和乘客接受信息模块，作为强化学习算法的平台.

2.3 算法流程

Q-learning算法的核心是状态—动作值 $Q(\boldsymbol{s},\boldsymbol{a})$ 的更新，基本思想是根据路网客流分布状态，通过尝试不同信息发布动作获取不同奖励值更新状态—动作值，多次训练可以得到在不同状态下，执行不同信息发布动作的价值，选取价值最高的动作执行.

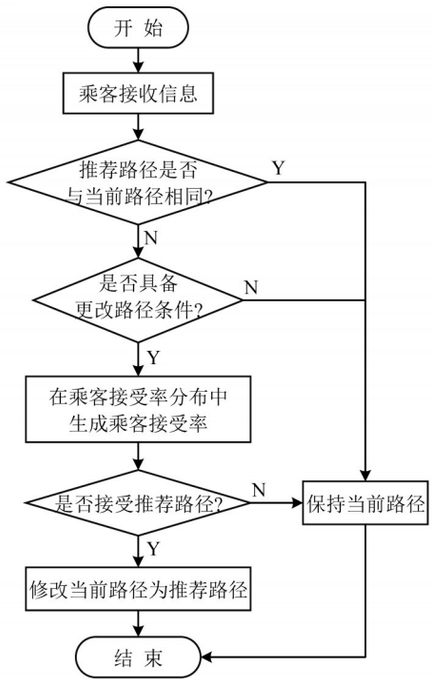


图2　乘客接收信息过程

Fig. 2　Passenger receive information process

将仿真实验与 Q-learning 算法结合，利用仿真系统推演信息发布后，路网客流分布动态变化，为 Q-learning 算法提供训练样本. 假设进行 K 次仿真实验，每次仿真开始时间为 t_s，结束时间为 t_e，每仿真 Δt 时间进行一次信息发布及算法学习. 算法流程如图3所示，具体如下.

Step 1　设 $k=1$，初始化 Q-table，初始的 Q-table为空.

Step 2　进行第 k 次仿真实验，设当前时间 t 为 t_s.

Step 3　提取路网客流分布状态 $\boldsymbol{s}(t)$，选取信息发布动作 $\boldsymbol{a}(t)$ 进行信息发布. 使用贪婪策略进行动作选取，指定探索速率为 $\varepsilon\in(0,1]$. 生成随机数 x，若 x 小于探索速率或现有 Q-table内没有当前状态 $\boldsymbol{s}(t)$，则在每个OD对中随机选取路径作为推荐路径，共同组成动作 $\boldsymbol{a}(t)$；否则在 Q-table内选取状态为 $s(t)$ 的最大 Q 值所对应的动作 $\boldsymbol{a}(t)$. 选定动作 $\boldsymbol{a}(t)$ 后，在 Δt 时间内按照 $\boldsymbol{a}(t)$ 中各OD的推荐路径向仿真系统内的乘客发布信息.

Step 4　仿真系统推演 Δt 时间，利用仿真系统输出的各区间满载率数据，提取执行信息发布动作 $\boldsymbol{a}(t)$ 后路网客流分布状态 $\boldsymbol{s}(t+\Delta t)$，并计算执行动作 $\boldsymbol{a}(t)$ 后的奖励值 $r(t)$.

Step 5　更新 Q-table. 根据系统状态 $\boldsymbol{s}(t)$，信息发布动作 $\boldsymbol{a}(t)$，执行动作后的系统状态 $\boldsymbol{s}(t+\Delta t)$ 和奖励 $r(t)$，更新 Q-table的公式为

$$Q^{\text{new}}\left[\boldsymbol{s}(t),\boldsymbol{a}(t)\right]=(1-\alpha)Q\left[\boldsymbol{s}(t),\boldsymbol{a}(t)\right]+\alpha\left\{r(t)+\gamma\max_{a'}Q\left[\boldsymbol{s}\left(t+\Delta t\right),a'\right]-Q\left[\boldsymbol{s}(t),\boldsymbol{a}(t)\right]\right\} \tag{11}$$

式中：$Q^{\text{new}}\left[\boldsymbol{s}(t),\boldsymbol{a}(t)\right]$ 为更新后的状态 $\boldsymbol{s}(t)$ 下采取动作 $\boldsymbol{a}(t)$ 的 Q 值；α 为学习率（$0<\alpha\leqslant 1$）；$Q\left[\boldsymbol{s}(t),\boldsymbol{a}(t)\right]$ 为更新前的状态 $\boldsymbol{s}(t)$ 下采取动作 $\boldsymbol{a}(t)$ 的 Q 值；γ 为衰减系数（$0<\gamma\leqslant 1$），γ 数值越大，学习过程中越重视未来长期奖励，反之则更重视当前短期的奖励；$\max_{a'}Q\left[\boldsymbol{s}\left(t+\Delta t\right),a'\right]$ 为状态 $\boldsymbol{s}\left(t+\Delta t\right)$ 下采取各动作能够获得的最大 Q 值.

Step 6　判断是否达到仿真结束时间 t_e，若达到，执行Step 7；否则，$t=t+\Delta t$，执行Step 3.

Step 7　判断是否达到仿真实验次数，若达到，学习结束；否则，$k=k+1$，执行Step 2.

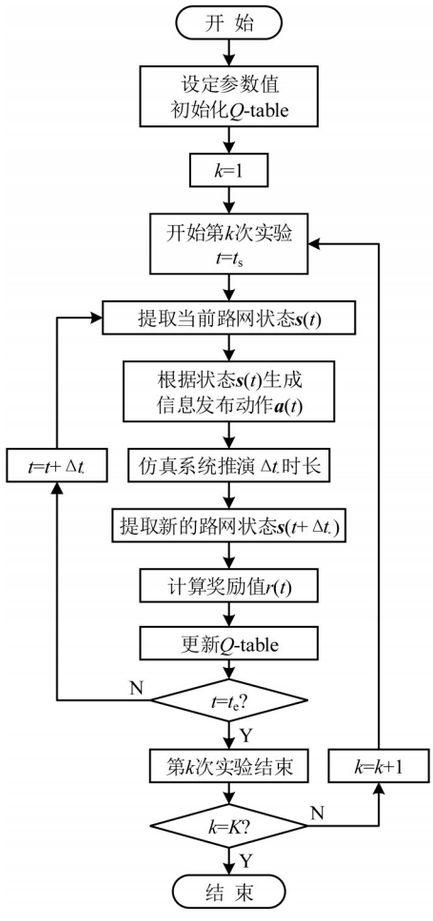


图3　算法流程

Fig. 3　Algorithm process

3 案 例

选择2018年广州城市轨道交通路网进行算例实验，路网中共有14条线路，205个车站.使用某工作日07:00-10:00的客流进行1 000次仿真实验，每15 min进行模型训练并更新信息内容.参数设置如下：满载率控制上限 $l_{max}=0.8$，探索速率 $\varepsilon=0.3$，学习率 $\alpha=0.1$，衰减系数 $\gamma=0.9$.训练过程中平均奖励值变化如图4所示，随着训练次数的不断增加，平均奖励不断增大直至趋于稳定.训练初期，Q-table中状态数较少，算法不断探索新的动作，平均奖励波动较大；随着训练次数增加，Q-table中涵盖的状态数增多，算法探索新动作的概率变小，多数情况下，在已有动作集中选择Q值最大的动作.

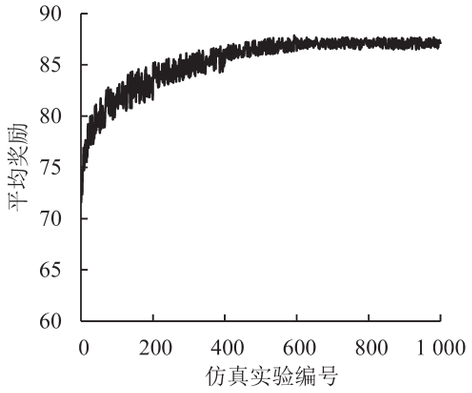


图4 训练过程平均奖励值变化

Fig. 4 Average reward in train process

利用仿真系统输出数据，从个体平均旅行时间和断面满载率两个角度评估信息发布效果.无信息发布情况下，全网乘客平均旅行时间为27.62 min；信息发布后，平均旅行时间降至26.31 min.断面满载率方面：无信息发布情况下，路网中有6条线路存在满载率大于控制上限的区间时段；信息诱导后，有4条线路超过控制上限的区间时段数下降，如图5所示.5号线下降数量最多，下降12个，降幅24.49%；2号线下降数量最少，下降6个，降幅13.95%.有2条线路没有变化，分析客流成份后发现，经过1号线和8号线拥堵区间的客流OD主要是单路径OD，客流无法进行诱导.若想进一步疏解，需要从供给侧调整运输组织方案，提高能力供给.这一结果说明，只有结合供给侧和需求侧进行运输组织优化才能系统性地疏解客流拥堵.从时间维度分析区间满载率变化，5号线杨箕—五羊邨区间分时段满载率变化情况如图6所示.无信息发布情况下，有2个时段满载率超过控制上限；信息发布后，整个早高峰时段满载率都低于控制上限.

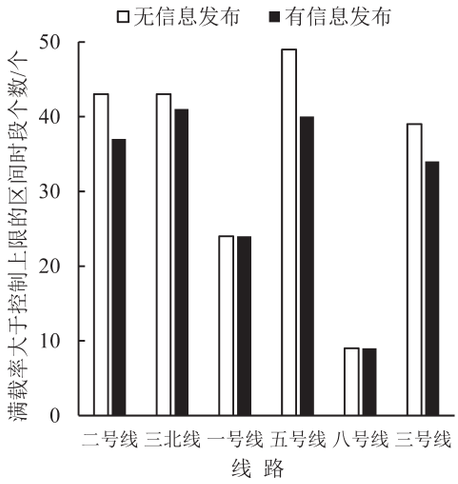


图5 满载率大于控制上限的区段时段数对比

Fig. 5 Comparison of section-time number over load rate control up bound

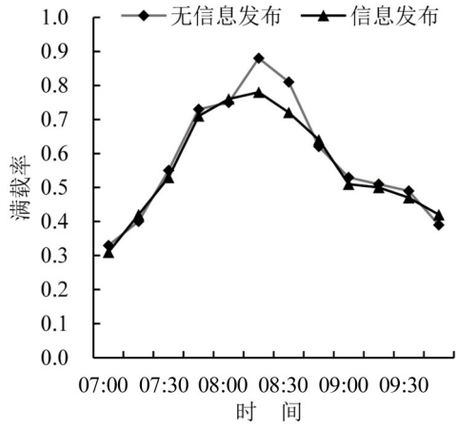
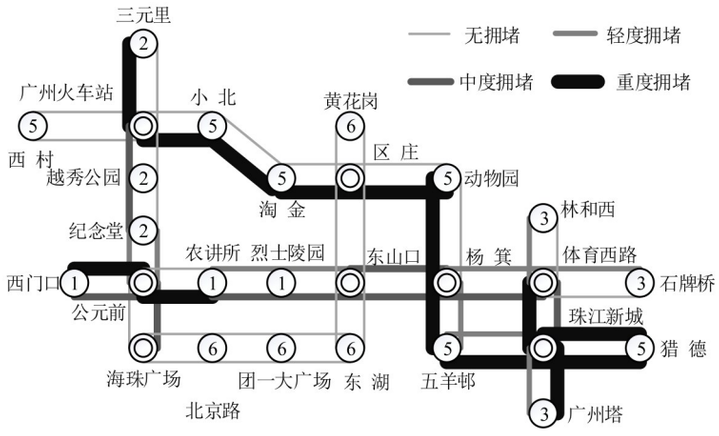


图6 杨箕—五羊邨分时段满载率变化

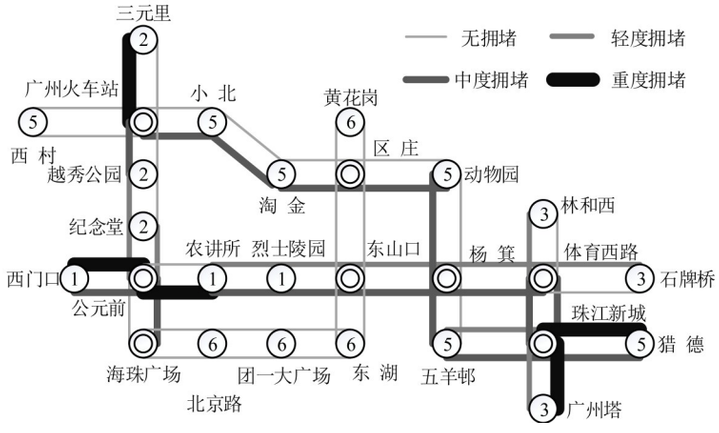
Fig. 6 Load rate change in time series of section Yangji-Wuyangcun

高峰时段局部路网平均满载率分布变化如图7所示.由图7可知：无信息发布情况下，重度拥堵区间集中在5号线下行方向；信息发布后，5号线客流拥堵情况得到缓解，重度拥堵区间消失，一部分客流被疏解到其他线路.利用仿真系统输出数据分析OD各路径客流占比，部分典型OD路径客流占比变化情况如表1所示.以江夏—番禺广场为例，该OD共有3条有效路径，在没有信息发布时，各路径客流比例为0.42∶0.22∶0.36；信息发布后，各路径客流比例变为0.11∶0.28∶0.61，经过5号线拥

堵区段的路径1的客流比例降低，客流被引导至经过8号线的路径3.经过拥堵区段的路径客流比例降低，说明信息发布策略为乘客推荐了不经过拥堵区间的路径.乘客对推荐路径存在一定接受概率，所以客流不会全部转移到某条路径，且算法根据路网客流分布动态变换推荐路径，避免拥堵转移.



(a) 无信息发布



(b) 有信息发布

图7　局部路网平均满载率对比

Fig. 7　Comparison of average load rate in partial network

表1　OD路径分配比例变化

Table 1　Path proportion change of OD

OD	路径编号	经过线路	客流比例	
			无信息发布	信息发布
江夏—番禺广场	1	2号线、5号线、3号线	0.42	0.11
	2	2号线、三北线、3号线	0.22	0.28
	3	2号线、8号线、3号线	0.36	0.61
车陂—客村	1	4号线、8号线	0.85	0.75
	2	4号线、5号线、3号线	0.15	0.25
广州火车站—番禺广场	1	5号线、3号线	0.84	0.7
	2	2号线、8号线、3号线	0.16	0.3

4　结　论

本文提出基于*Q*-learning的城轨信息发布策略生成算法.通过识别路网客流分布状态，在*Q*-table中选择相应的信息发布动作，仿真推演获得发布信息后的路网客流分布状态，计算动作的奖励值并更新*Q*-table.利用实际路网进行案例分析，

对比有无信息发布情况下路网的客流分布.本文提出的信息发布策略算法能够诱导乘客选择非拥堵路径,有效降低路网中拥堵的区间时段个数.

参考文献:

[1] 陈芳, 张卫华, 丁恒, 等. 基于出行者路径选择行为的VMS诱导策略研究[J]. 系统工程理论与实践, 2018, 38(5): 1263-1276. [CHEN F, ZHANG W H, DING H, et al. Research on VMS inducing strategy based on the route selection behavior of travels[J]. System Engineering- Theory & Practice, 2018, 38(5): 1263-1276.]

[2] 孙智源, 陆化普, 张晓利, 等. 城市交通控制与诱导协同的双层规划模型[J]. 东南大学学报(自然科学版), 2016, 46(2): 450-456. [SUN Z Y, LU H P, ZHANG X L, et al. Bi-level programming model for cooperation of urban traffic control and traffic flow guidance[J]. Journal of Southeast University (Natural Science Edition), 2016, 46(2): 450-456.]

[3] 曹守华, 袁振洲, 李艳红, 等. 交通流诱导条件下的路网随机平衡双层规划模型[J]. 交通运输系统工程与信息, 2007, 7(4): 36-42. [CAO S H, YUAN Z Z, LI Y H, et al. A model for road network stochastic user equilibrium based on bilevel programming under the action of traffic flow guidance sytem[J]. Journal of Transportation System Engineering and Information Technology, 2007, 7(4): 36-42.]

[4] 杨泳, 户佐安. 城市路网交通管制下的最优路径诱导仿真研究[J]. 计算机仿真, 2013(7): 155-158. [YANG Y, HU Z A. Optimal path routing simulation research on urban traffic network with traffic management and control strategies[J]. Computer Simulation, 2013(7): 155-158.

[5] YIN H, WU J, LIU Z, et al. Optimizing the release of passenger flow guidance information in urban rail transit network via agent-based simulation[J]. Applied Mathematical Modelling, 2019, 72: 337-355.

[6] WATKINS, CHRISTOPHER J C H, DAYANPETER. *Q*-learning[J]. Machine Learning, 1992, 8(3/4): 279-292.

[7] VAN ESSEN M, THOMAS T, VAN BERKUM E, et al. Travelers' compliance with social routing advice: Evidence from SP and RP experiments[J]. Transportation, 2020, 47(3): 1047-1070.

[8] 蒋熙, 冯佳平, 贾飞凡, 等. 大规模城轨路网客流分布推演的建模与仿真方法[J]. 铁道学报, 2018, 40(11): 13-22. [JIANG X, FENG J P, JIA F F, et al. Modeling and simulation of passenger flow distribution in large-scale urban rail transit network[J]. Journal of the China Railway Society, 2018, 40(11): 13-22.]