文章编号:2095-6134(2022)01-0134-10

基于 Q-learning 的飞行自组织网络 QoS 路由方法*

黄鑫陈[1,2†],陈光祖[1],郑敏[1],谭冲[1],刘洪[1]

(1 中国科学院上海微系统与信息技术研究所, 上海 200050; 2 中国科学院大学微电子学院, 北京 100049)

(2020 年 1 月 2 日收稿; 2020 年 4 月 29 日收修改稿)

Huang X C, Chen G Z, Zheng M, et al. Q-learning based QoS routing for high dynamic flying Ad Hoc networks[J]. Journal of University of Chinese Academy of Sciences, 2022, 39(1): 134-143. DOI: 10.7523/j.ucas.2020.0001.

摘 要 针对无人机自组网等高动态飞行自组织网络中,网络拓扑的快速变化导致通信链路断裂和路由重建频繁的问题,研究一种基于 Q-learning 的 QoS(quality of service)路由方法。该方法以 Q-learning 强化学习框架为基础,将邻居节点数量、链路持续时间和链路可用带宽作为路由度量信息,设计一种提供 QoS 保证的 Q-learning 奖励函数。网络节点通过广播 Hello 消息交互各自的本地路由度量信息,邻居节点接收到 Hello 分组或者数据分组,根据奖励函数计算并更新 Q 值,待转发数据分组的节点根据其维护的 Q 值表智能选择下一跳转发节点。EXata 无线网络仿真环境中的仿真结果表明,该方法能为高动态飞行自组织网络中的数据传输提供稳定性好、服务质量高的通信链路。

关键词 飞行自组网;QoS 路由;Q-learning;链路可用带宽;链路持续时间

中图分类号:TN929.5 **文献标志码**:A **DOI**:10.7523/j.ucas.2020.0001

Q-learning based QoS routing for high dynamic flying Ad Hoc networks

HUANG Xinchen[1,2], CHEN Guangzu[1], ZHENG Min[1], TAN Chong[1], LIU Hong[1]

(1 *Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai* 200050, *China*; 2 *School of Microelectronics, University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract In high dynamic flying ad hoc networks (FANETs), such as UAV (unmanned aerial vehicle) ad hoc networks, the rapid change of network topology leads to the breakage of communication links and the frequent reconstruction of routes. To solve this problem, a QoS (quality of service) routing method based on Q-learning is studied. Based on the basic Q-learning framework, this method takes the number of neighbor nodes, link duration and link available bandwidth as routing metrics, and designs a Q-learning reward function to provide QoS guarantee. All nodes exchange local routing metrics information with neighbor nodes by broadcasting Hello messages and forwarding data packets. After receiving Hello packets or data packets, neighbor nodes calculate and update the Q value according to the reward function. Then one of neighbor nodes selects a next hop node to forward data packets intelligently according to the Q value table that

* 中国科学院青年创新促进会(2018269)资助

† 通信作者,E-mail: huangxinchen@csu.edu.cn

it maintains. The simulation results in EXata simulator show that this method can provide stable and high QoS communication links for high dynamic flying ad hoc networks.

Keywords　FANETs; QoS routing; Q-learning; link available bandwidth; link expiration time

随着传感器和通信技术的快速发展,各种有人和无人飞行设备大量应用于军事和民用领域。战斗机、火炮等装备已成为各国强大的战斗力量,国产歼-20 已装备人民空军的王牌部队,在这种趋势下国内外掀起一股无人机的理论和实践研究热潮。飞行自组网(flying ad hoc networks)作为一种新的移动自组网(mobile ad-hoc networks)应运而生[1]。这种网络由具有无线通信功能的节点组成,不依赖任何固定的基础设施,以一种无中心、自组织和多跳传输方式,为战机协同、抢险救灾等应用场景提供应急通信网络。

路由选择作为网络通信的关键技术之一,决定了数据的传输路径,对网络整体性能有着非常重要的影响[2]。然而,在高动态飞行自组织网络中,网络节点频繁入网、退网以及快速移动,网络拓扑变化快,链路容易断裂和路由重建频繁,从而导致数据分组丢失严重,网络性能严重下降[3-5]。传统的 Ad hoc 网络路由协议,例如 AODV(ad hoc on-demand distance vector routing)和 DSR(dynamic source routing),难以适应网络拓扑结构的快速变化,不能保证网络的服务质量(quality of service, QoS)。

链路可用带宽被认为是 QoS 的一个重要指标,它是指由发/收节点组成的链路中未被使用的空闲带宽,表征通信链路还可以提供的数据传输能力,通常定义为在不影响当前网络中现有业务服务质量的前提下,该链路可以为新加入的业务流提供的最大传输速率。在现有的带宽预测方法中,基于被动测量的带宽预测算法利用节点自身的物理载波检测能力,获取带宽利用信息来预测链路的可用带宽。基于被动测量的方法无需向网络中发送探测包,不会对原有业务的服务质量造成额外影响,预测结果更加可靠,得到了较为广泛的应用[6]。

近年来,Q-learning 作为一种离策略、无模型的启发式强化学习方法,受到众多研究人员的关注[7-8],它能够通过与周围环境交互信息动态地调整传输路径,将学习任务分散在每一个网络节点中,通过周期性地与周围邻居节点交换控制信息,可为寻找可靠的传输路径提供依据[9-13]。

基于 Q-learning 强化学习框架,本文研究一种面向高动态飞行自组网的 QoS 路由方法,该方法考虑了转发节点质量、链路稳定性和链路通信质量,分别将邻居节点数量、链路持续时间和链路可用带宽作为路由度量信息,设计一种提供 QoS 保证的 Q-learning 奖励函数。网络节点周期性统计本地路由度量信息并封装在 IP 头部,然后通过广播 Hello 消息和发送数据分组进行交互,当邻居节点接收到 Hello 分组或者数据分组,根据奖励函数计算并更新 Q 值,源节点或者中继节点根据其维护的 Q 值表智能选择下一跳节点进行数据转发。EXata 网络仿真结果表明,该方法在吞吐量和平均端到端时延上具有较好的性能,能为高动态飞行自组织网络中数据传输提供稳定性好、服务质量高的通信链路。

1　Q-learning 与路由重建

强化学习(reinforcement learning, RL)利用智能体(agent)与环境(environment)的交互,通过映射动作(action)和场景进行学习以获得最优策略。它不会告诉 agent 在当前状态(state)下应该采取的最优动作,而是让 agent 与环境进行交互,通过不断地尝试最大化总奖励值进而获得最优策略。Q-learning 作为一种经典的强化学习算法,通过不断与外界交互信息,能够在动态的环境中找出一条到达目的地的最佳路径。

图 1 描述 RL 的基本框架。RL 中的智能体根据系统的当前状态以及从环境中接收到的反馈来选择操作。满足马尔可夫性质的强化学习任务称为马尔可夫决策过程(Markov decision process, MDP),通常用一个 4 元组(s, a, p, r)来描述,该 4 元组分别表示状态、动作、转移概率(transition probabilities)和奖励(reward)。

定义:

1)动作(a):智能体可以采取的所有可能的行动。

2)状态(s):环境返回的当前情况。

3)奖励(r_t):环境的即时反馈值,以评估智能体选择的上一个动作。

4)策略(p):智能体根据当前状态决定下一

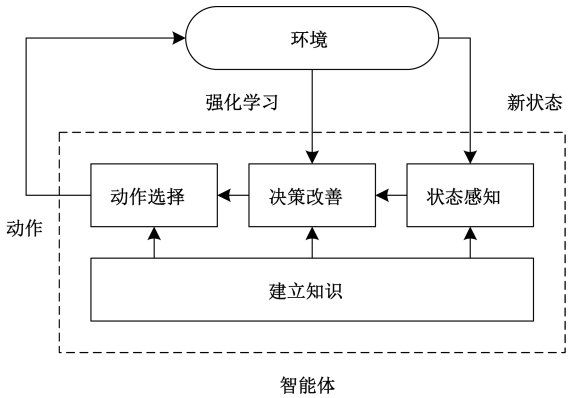


图 1 强化学习的基本框架

Fig. 1 Basic framework of reinforcement learning

步动作的策略。

5)价值(V):折扣(discount)下的长期期望返回值,与 r_t 代表的短期返回相区分。$V^p(s)$ 定义为策略 p 下当前状态 s 长期返回值的期望。

6)Q 值或行动值(Q):与 r_t 相似,但多一个参数 a。$Q^p(s,a)$ 指当前状态 s 在策略 p 下采取动作 a 的长期回报。

Q-learning 是基于贝尔曼方程(Bellman equation)的离策略、无模型强化学习算法。在通信网络中,一个节点代表一个状态,数据分组从一个节点传输到另一个节点称为一个动作。每发送一个数据分组,更新一次平均值,这就是 Q-learning 路由的基本思想。数据分组被转发的次数越多,得到样本就越多,则更新次数越多,Q 的估计值就越接近于真实值,最后依概率收敛于最优值,从而可以找出一条从源节点到目的节点的最佳路径。

标准 Q-learning 的更新公式为

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1},a)], \quad (1)$$

其中:$\alpha \in (0,1]$ 为学习速率,用于控制学习更新的速度;$\gamma \in [0,1)$,用于表示未来奖赏的折扣,意味着相较于以后的回报看重眼前奖励的程度;r_t 为环境的即时反馈值,可根据网络性能需求,将性能参数如跳数、带宽、时延、丢包率、能耗等映射到 r_t 中。

2 基于 Q-learning 的 QoS 路由方法

2.1 奖励函数设计

2.1.1 邻居节点度

邻居节点度即一跳邻居节点数,是衡量节点质量的重要度量指标。如果有待发送数据的节点随机选择邻居节点作为下一跳转发节点,该转发节点的邻居节点度可能较小,则邻居节点稀少甚至没有,容易造成通信链路断裂,从而导致链路的可持续性降低。用 $N(R)$ 表示节点 R 的邻居节点度,$N(R)$ 并非越大越好。假设节点的发送概率为 P_t,在基于竞争接入的移动自组织网络中,节点成功传输数据分组的概率 P_s 为 $1-(1-P_t)^{N(R)-1}$。邻居节点数越多,越有可能产生分组冲突,导致网络性能下降。

2.1.2 链路可用带宽

文献[14]提出一种基于载波检测的链路可用带宽被动测量方法,节点通过载波检测侦听自身的发送和接收可用时长,对链路可用带宽进行初步估计,然后侦听可能导致数据冲突的其他节点发送时长,对初步估计值进行修正。该方法不依赖于邻居节点的数量,考虑了当前网络的业务量,且能获得较为精确的可用带宽预测值。

首先根据数据链路层协议模型确定链路可用带宽的上限值。定义传输周期为链路成功完成一次数据传输所需要的时间,以 IEEE 802.11DCF 协议为例,考虑图 2 所示的 RTS/CTS 4 次握手机制,传输周期包含分布式帧间间隔(distributed interframe space,DIFS)、退避过程(BackOff)所经历的时间、RTS/CTS 控制帧交互过程经历的时间,DATA/ACK(acknowledgement)帧交互过程经历的时间,以及 3 个短帧间间隔(short interframe space,SIFS)。

$$t = t_{\mathrm{DIFS}} + t_{\mathrm{B}} + t_{\mathrm{RTS}} + t_{\mathrm{CTS}} + t_{\mathrm{DATA}} + t_{\mathrm{ACK}} + 3t_{\mathrm{SIFS}}, \quad (2)$$

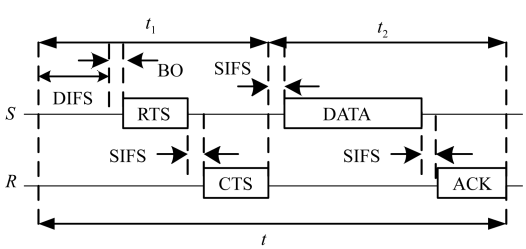


图 2 传输周期示意图

Fig. 2 Transmission cycle diagram

用 L_{DATA} 表示 DATA 帧的大小,考虑到传输周期 t 包含传输一个 DATA 帧的其他协议开销,则网络中一条链路能获得的最大吞吐量 $B_{\max}$ 为

$$B_{\max} = \frac{L_{\mathrm{DATA}}}{t}, \quad (3)$$

最大吞吐量 $B_{\max}$ 可视为链路可用带宽的上限值。

为了获得实际链路可用带宽,节点利用自身的物理载波检测能力侦听周围信道的“忙闲”状态,在一个固定测量周期 T_{mea} 内,统计各自的发送可用时长和接收可用时长。节点的物理层状态有 4 种情况:发送、接收、侦听和空闲,发送可用时长为节点处于空闲状态,且空闲时间大于 DIFS 的时长;接收可用时长为节点处于空闲或侦听状态的时长。用 $T_{\text{s}}(S)$ 表示发送节点的发送可用时长,$T_{\text{r}}(R)$ 表示接收节点的接收可用时长,则由发/收节点对 (S,R) 组成的链路 $L_{S,R}$ 的可用时长 T_{L} 计算为

$$T_{\text{L}} = \min\{[1 - p(S,R)] \cdot T_{\text{s}}(S), [1 - p(R,S)] \cdot T_{\text{r}}(R)\}. \tag{4}$$

其中:$p(S,R)$ 为节点 S 可以发送数据,但节点 R 不能接收的概率;$p(R,S)$ 为节点 R 可以接收数据,但节点 S 不能发送的概率。在每个测量周期 T_{mea} 内,链路 $L_{S,R}$ 的可用带宽初步估计值 B_{pre} 根据链路可用时长的占比计算为

$$B_{\text{pre}} = \frac{T_{\text{L}}}{T_{\text{mea}}} \cdot B_{\max}. \tag{5}$$

在基于竞争接入的多跳 ad hoc 网络中,考虑隐藏节点的信号传输导致节点对 (S,R) 数据分组冲突,以及信道忙而不能应答 CTS,从而造成链路可用带宽损耗的情况,对初步估计值进行修正。在一个测量周期 T_{mea} 内,通过侦听信道统计发送节点 S 的隐藏节点发送信号的总时间为 T_{hid},可以推出隐藏节点导致可用带宽消耗的概率 p_{con} 为

$$p_{\text{con}} = \begin{cases} \dfrac{1}{2}\left(1 + \dfrac{T_{\text{hid}}}{T_{\text{mea}}}\right) - \dfrac{t_{\text{RTS}} \cdot T_{\text{L}}}{t \cdot T_{\text{mea}}}, & \dfrac{T_{\text{hid}}}{T_{\text{mea}}} > \left(1 - \dfrac{t_{\text{RTS}} \cdot T_{\text{L}}}{t \cdot T_{\text{mea}}}\right), \\ \dfrac{1}{2}\left(1 - \dfrac{t_{\text{RTS}} + t_{\text{CTS}}}{t} + \dfrac{3T_{\text{hid}}}{T_{\text{mea}}}\right) - \dfrac{T_{\text{L}}^2 \cdot t_{\text{RTS}}^2}{2 \cdot t^2 \cdot T_{\text{mea}}(T_{\text{mea}} - T_{\text{hid}})} - \dfrac{t_{\text{DATA}} \cdot T_{\text{L}}}{t \cdot T_{\text{mea}}}, & \\ \left(1 - \dfrac{T_{\text{mea}} \cdot (t_{\text{RTS}} + t_{\text{CTS}}) + t_{\text{DATA}} \cdot T_{\text{L}}}{t \cdot T_{\text{mea}}}\right) < \dfrac{T_{\text{hid}}}{T_{\text{mea}}} \leqslant \left(1 - \dfrac{t_{\text{RTS}} \cdot T_{\text{L}}}{t \cdot T_{\text{mea}}}\right), & \\ \dfrac{T_{\text{hid}}}{T_{\text{mea}}} - \dfrac{T_{\text{L}}^2 \cdot t_{\text{DATA}}^2}{2 \cdot t \cdot T_{\text{mea}}^2 \cdot (t - t_{\text{RTS}} - t_{\text{CTS}} - t \cdot T_{\text{hid}})}, & \dfrac{T_{\text{hid}}}{T_{\text{mea}}} \leqslant \left(1 - \dfrac{T_{\text{mea}} \cdot (t_{\text{RTS}} + t_{\text{CTS}}) + t_{\text{DATA}} \cdot T_{\text{L}}}{t \cdot T_{\text{mea}}}\right). \end{cases} \tag{6}$$

则链路 $L_{S,R}$ 最终的可用带宽 $B(S,R)$ 为

$$B(S,R) = (1 - p_{\text{con}}) \cdot B_{\max}. \tag{7}$$

2.1.3　链路持续时间

考虑图 3 所示的平面拓扑,设节点 S 为源节点,D 为目的节点,R 为节点 S 的一个邻居节点,链路持续时间是移动距离 RH 所需的时间 t_{RH}。然而,在基于贪婪和竞争的转发过程中,它是经过距离 RK 所需的时间 t_{RK},而 t_{RK} 明显小于 t_{RH}。

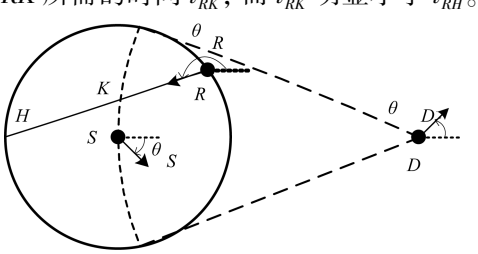


图 3　链路持续时间计算模型

Fig. 3　Calculation model of link duration

设节点 S、R 和 D 的坐标分别为 (x_S,y_S)、(x_R,y_R) 和 (x_D,y_D),节点移动的速度和方向分别为 (V_S,θ_S)、(V_R,θ_R) 和 (V_D,θ_D),参考文献[15],节点 R 处于节点 S 通信范围的时间 $T(S,R)$ 预测为

$$T(S,R) = \frac{-(ab + cd) + \sqrt{(a^2 + c^2)h^2 - (ad - cb)^2}}{a^2 + c^2}, \tag{8}$$

其中 h 为传输距离,且

$$\begin{cases} a = v_S\cos\theta_S - v_R\cos\theta_R, \\ b = x_S - x_R, \\ c = v_S\sin\theta_S - v_R\sin\theta_R, \\ d = y_S - y_R. \end{cases} \tag{9}$$

2.1.4　奖励函数

根据前面对邻居节点度、链路持续时间和链路可用带宽的定义可知,$N(R) \in [0,\infty)$,$T(S,R) \in [0,\infty)$,$B(S,R) \in [0,B_{\max}]$,首先对 3 个参量分别进行归一化,其归一化值 $n(R)$、$t(S,R)$ 和 $b(S,R)$ 为

$$\begin{cases} n(R) = \dfrac{2}{\pi}\arctan[N(R)], \\ t(S,R) = \dfrac{2}{\pi}\arctan[T(S,R)], \\ b(S,R) = \dfrac{B(S,R)}{B_{\max}}. \end{cases} \tag{10}$$

定义节点 S 到 R 的瞬时奖励 $A(S,R)$ 为

$$A(S,R) = -g + [w_N \cdot n(R) + w_B \cdot b(S,R) + w_T \cdot t(S,R)], \quad (11)$$

其中：w_N、w_B 和 w_T 分别为邻居节点度、链路可用带宽和链路持续时间的权重因子，且满足 $w_N + w_B + w_T = 1$。定义 g 为取值是正常数的惩罚因子，则 $-g$ 为负值，因为每次发送数据分组都会消耗节点能量，并且占用一定的信道带宽。基于归一化的 $n(R)$、$t(S,R)$ 和 $b(S,R)$，取 $g=1$，则 $A(S,R) \in [-1,0]$。$A(S,R)$ 表明网络节点发送数据分组之后会获得一个负的奖励，从而迫使源节点最终选择相对跳数较少的转发路径，因为跳数越多，转发节点获得的负奖励越多，Q 值则越小，被选为转发节点的机会越小。对于目的节点 D 的一个邻居节点 X，有 $A(X,D) = -1$。由于 $A(S,R)$ 总是负值，则非目的节点的 V 值也是负的，从而目的节点的 V 值最大，定义为 $V(D,D) = 0$。

根据瞬时奖励对相应邻居节点的 Q 值进行更新，更新当前节点 S 对其邻居节点 R 的质量评估为

$$Q_S(D,R) \leftarrow (1-\alpha)Q_S(D,R) + \alpha \cdot \{A(S,R) + \gamma \max_{X \in N_R} Q_R(D,X)\}, \quad (12)$$

其中：$\alpha \in (0,1]$ 为学习速率，$\gamma \in [0,1)$ 为折扣因子，N_R 为节点 R 的邻居节点集。

2.2 路由方案设计

基于上述的 Q-learning 框架，结合考虑链路可靠性和稳定性保障的奖励函数，提出基于 Q-learning 的 QoS 路由方法（Q-learning based QoS routing，QQR）。

为了计算本节点到相应目的节点的 Q 值，网络节点首先在本地统计路由测度相关信息，然后将其元数据封装在 IP 头部，并通过周期性广播 Hello 分组以及转发数据分组，将本地元数据发送给邻居节点。若其邻居节点正确接收到该分组，就可以从分组头部提取元数据，从而完成后续 Q 值的计算更新。节点周期性广播 Hello 分组的目的是确保所有节点（包括那些没有数据流量的节点）能够更新路由测度信息，以辅助邻居节点做出正确的路由决策，其周期大小应根据网络应用需求进行设置。

IP 头部除包含传统的 IP 版本、协议版本、源地址、目的地址等信息，还需要添加本节点的发送可用时长、节点位置、邻居节点数和 V 值链表，其中 V 值（即 Q_{max}）链表与经过本节点的目的节点数量有关。本节点的邻居节点表中相应地维护邻居节点上一次位置和记录时间、链路持续时间、邻居节点数、可用带宽链表和 V 值链表。为了减少协议开销，计算链路持续时间所需的速度参数由节点及其邻居节点在前、后 2 个时刻的位置进行估算，而不再交互额外的速度矢量信息。此外，每个节点都为其已知邻居维护一个如表 1 所示的 Q 值表（即 Q 矩阵），表中的行代表经过本节点的目的节点 ID，列表示与其相邻的一跳邻居节点 ID。

表 1 Q 值表

Table 1 Q values

Neighbors	Destinations		
	D_1	D_2	…
N_1	$Q_S(D_1, N_1)$	$Q_S(D_2, N_1)$	…
N_2	$Q_S(D_1, N_2)$	$Q_S(D_2, N_2)$	…
…	…	…	…

1）路由发现

图 4 为 QQR 路由方法的路由发现流程框图。当网络节点正确接收到分组（Hello 分组或者数据分组）时，无论该节点是否被指定为下一跳转发节点，它先从分组的报头中提取发送节点携带的信息，即发送可用时长，位置坐标 x、y、z，邻居节点

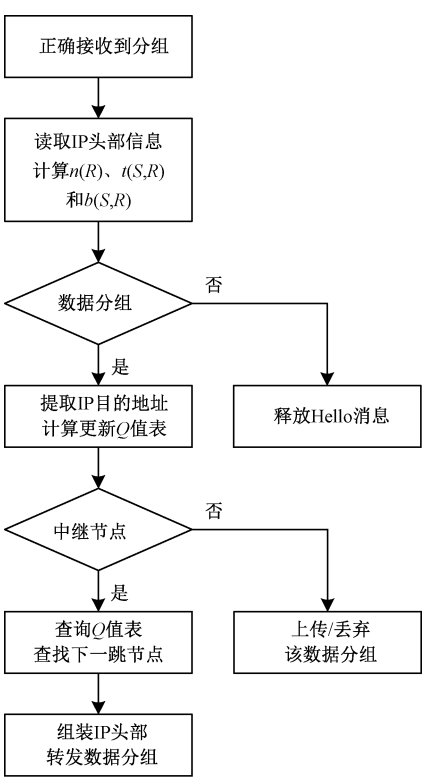


图 4 QQR 路由发现流程

Fig. 4 QQR route discovery process

数和 Q_{max} 链表,计算和更新本节点邻居链表中的邻居节点数、链路持续时间和链路可用带宽。

如果该节点接收到的分组是一个数据包,则提取 IP 头部中的目的地址插入 Q 值表,并通过式(12)计算该目的节点与本节点每个邻居相关联的 Q 值,即更新 Q 值表中该目的节点对应的列。如果本节点接收到的是 Hello 包,将丢弃 Hello 分组释放内存。若节点进一步判断自己是中继节点,即数据包是发给自己但自己不是目的节点,则本节点通过查询 Q 值表选择 Q 值最高的节点作为下一跳转发节点,然后用自己最新的路由测度信息替换 IP 头部中的旧元数据并发送数据分组。如果当前不存在到目的节点的 Q 值,或者存在多个最高 Q 值相同的节点,则从中随机选择一个节点转发本次数据分组。最后,不是发给本节点的数据分组将被丢弃,发给本节点的数据分组将上传给上层。

2)路由维护

当数据分组成功到达目的节点,则与这条路径相邻的部分节点的 Q 值表也得到更新,如果数据分组超过规定跳数或时间没有达到目的节点,则将该分组丢弃而不再进行转发。周期性广播 Hello 分组的主要目的是动态地维护全网节点的 Q 值表,并解决链路断开问题。此外规定了 Q 值表中目的节点的生存时间,如果在生存周期内某一个目的节点相关的 Q 值没有得到更新,则认为此目的节点失效,并删除其对应的一列 Q 值。

2.3　路由方法分析

1)网络开销

路由信息交互需要在 IP 头部添加本节点的发送可用时长(8 字节)、节点位置(24 字节)、邻居节点数(4 字节)和 V 值链表($8\times ND_i$ 字节,其中 ND_i 为有数据包中转至节点 i 的目的节点数)。因此,QQR 路由方法的 Hello 分组交互导致的吞吐量开销 $\mathrm{Thp}_{\mathrm{overhead}}$ 大约为

$$\mathrm{Thp}_{\mathrm{overhead}}=\frac{\sum_{i=1}^{N_{\mathrm{total}}}(36+8\times ND_i)}{T_{\mathrm{hello}}},\qquad(13)$$

式中:N_{total} 为网络总节点数,T_{hello} 为 Hello 分组更新周期,一般将链路可用带宽的估计周期 T_{mea} 设置为 T_{hello}。

2)反馈代价

QQR 算法中节点每接收到一个分组便计算并更新 Q 值,该过程即为一次 Q 学习。假设当前时刻源节点 S 和目的节点 D 之间存在一条最优路径 P,Q 学习的目的就是通过多次迭代学习最后逼近这条最优路径。在网络初始化时采用 ε-greedy 算法进行探索,发现的传输路径与最优路径存在较大的偏差,可能是跳数较大甚至没有找到目的节点,那么吞吐量和时延等网络性能就表现得较差。随着学习次数的增多,Q 值不断更新逼近稳态值,则传输路径越接近最优路径,网络性能逐渐提升。Q 学习发现的传输路径与理想传输路径之间的偏差导致的网络性能降低,就是 Q 学习用于路由时在通信网络中的反馈代价。

3)收敛时间

由于 Q-learning 算法的收敛需要一定的时间,只有网络中所有节点的 Q 值收敛后,建立的路由才会逼近最佳路由。然而对于拓扑运动的网络,可能存在 Q 学习还未收敛时,网络状态已经发生改变的情况,所以要求:①算法收敛时间内网络拓扑不能剧烈变化,不能导致 Q 值总是与稳态值存在很大的偏差;②Hello 包更新时间要小于算法收敛时间,且如果网络拓扑变化较快,更新时间应该取较小的值,从而保证邻居节点能够及时报告网络状态。

3　仿真结果与分析

3.1　收敛分析

本文提出的 QQR 协议已在 EXata 网络仿真环境中实现,设置邻居节点数、链路持续时间和链路可用带宽的权重系数为 0.2、0.3 和 0.5,其他主要仿真参数如表 2 所示。

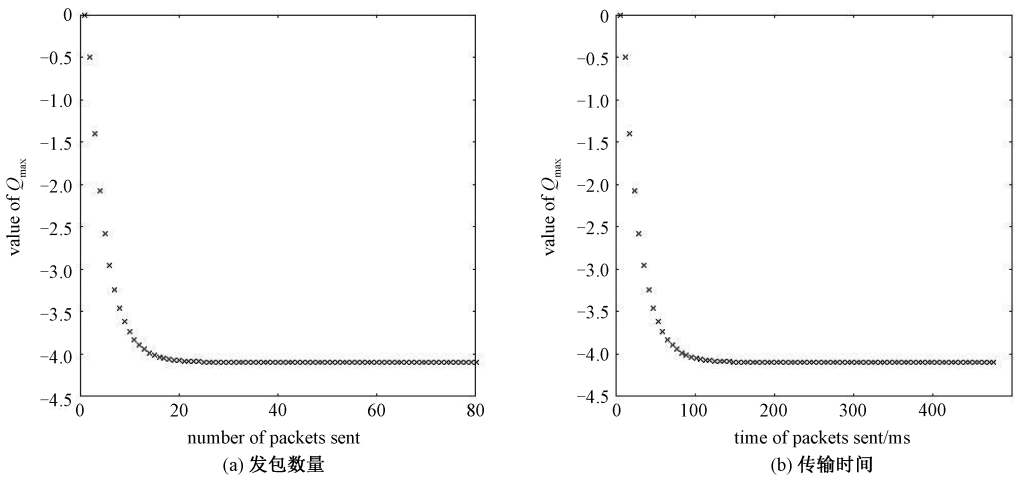
表 2　主要仿真参数

Table 2　Main simulation parameters

参数	设置值	参数	设置值
应用层报文大小/byte	1 000	业务流类型	Poisson 流
信道速率/(Mbit/s)	2	传播层模型	双线地面反射
传输距离/m	250	AODV RREQ/byte	29
物理载波检测距离/m	550	AODV RREP/byte	21
传输功率/dBm	15	QQR Hello/byte	21
信噪比门限值/dB	10	折扣因子	0.5
MAC 协议	802.11b	学习速率	0.5

本文采用的业务流模型为泊松流,数据包的产生时间服从泊松分布。为了评估 QQR 协议的收敛性,建立 4×4 平面网格拓扑,网格边长为 250 m,配置一条业务流使得源节点和目的节点位于整个拓扑主对角线的 2 个端点,源节点发包速

率为 1 pkt/s,Hello 包的广播间隔设置为 1 s。统计仿真运行过程中源节点的 Q_{max} 值随发包数目的变化情况如图 5(a)所示,依据图 5(a)结合根据发包速率、数目和传输时间的关系,可转化出源节点 Q_{max} 值随传输时间的变化情况如图 5(b)所示。



(a) 发包数量 (b) 传输时间

图 5 源节点 Q_{max} 值随发包数目和传输时间的变化情况

Fig. 5 Change of source node Q_{max} value with number of packets sent and transmission time

在该网格拓扑中,因为源节点与目的节点距离最远,所以源节点的 Q_{max} 值收敛时间最长。图 5(a)中曲线开始部分 Q_{max} 值剧烈下降,在发包数为 20 时开始缓慢减少,说明源节点从网络中学习,并获得接近真实网络的信息,包括网络拓扑的运动性、周围邻居节点情况,以及流内竞争程度。当发包数为 40 时曲线趋于稳定,表明 Q_{max} 值基本达到收敛状态。曲线稳定后也可能小幅波动,因为流内竞争导致最佳路径上的可用带宽降低,QQR 路由算法会根据奖励函数选择带宽较高的另一个节点,此时的路径可能不是最短路径。

图 5(a)还反映了源节点的 Q_{max} 值随数据分组传输时间的变化情况。源节点发包速率为 1 pkt/s,考虑“2.1.2 链路可用带宽”中的数据分组传输周期,根据式(2)计算单个数据分组的平均传输时间 $t=5\,952\ \mu s$。则由图 5(a)可知,在没有其他业务流干扰的情况下,源节点业务饱和时 Q_{max} 值收敛的数据分组总传输时间约为 $40t \approx 0.24$ s。在源节点业务非饱和以及存在干扰业务的条件下,节点 Q_{max} 值收敛时间为节点发送最后一个(当前仿真场景中为 40 pkt)使得 Q_{max} 值基本稳定的数据包之前的所有时间,主要包含所有数据包的端到端传输时延以及发包间隔,下面通过静态拓扑仿真对其分析讨论。

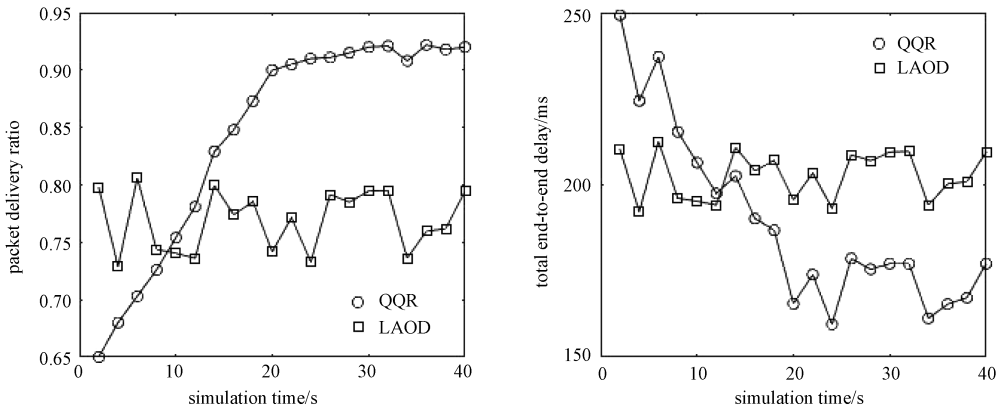
3.2 静态拓扑

在 1 000 m×1 000 m 的方形拓扑中随机均匀分布 25 个静态节点,随机建立 6 条多跳泊松业务流。Hello 包的广播间隔设置为 0.1 s,仿真时间 40 s,统计 6 条业务流总的分组投递率,吞吐量和平均端到端时延,并与 LAOD[16] 和 AODV[17] 比较分析。

1)分组投递率和平均端到端时延随仿真时间的变化

设置每条业务流的业务负载为 50 Kbit/s,在 40 s 的仿真过程中每隔 2 s 统计一次所有业务流的总分组投递率和总平均端到端时延,仿真结果如图 6(a)和图 6(b)所示。

在静态拓扑中,由于网络节点静止不动,LAOD 退化为 AODV。如图 6(a)所示,当业务负载保持不变时,LAOD 协议获得的分组投递率基本保持在一个稳定水平,Poisson 业务流的随机性会导致统计结果有一些小范围波动。因为 LAOD 协议同样通过广播 RREQ 分组和应答 RREP 分组进行路由发现,而且是按需执行该过程,不需要额外的判断和等待,所以能在较短时间内建立路由。然而对于 QQR 协议,在仿真初期需要发送数据包来建立并更新 Q 值表,探索传输路径的过程使得开始阶段业务的分组投递率比较小。随着 Q 值表慢慢收敛,传输路径也逐渐收敛到较优路径,分组投递率逐渐提高并达到稳定水平。静态拓扑中邻居节点数和链路持续时间维持恒定,由于 QQR 协议还考虑了链路可用带宽,减少了网络拥塞,因而较 LAOD 协议提高了分组投递率。



(a) 分组投递率　　(b) 总端到端时延

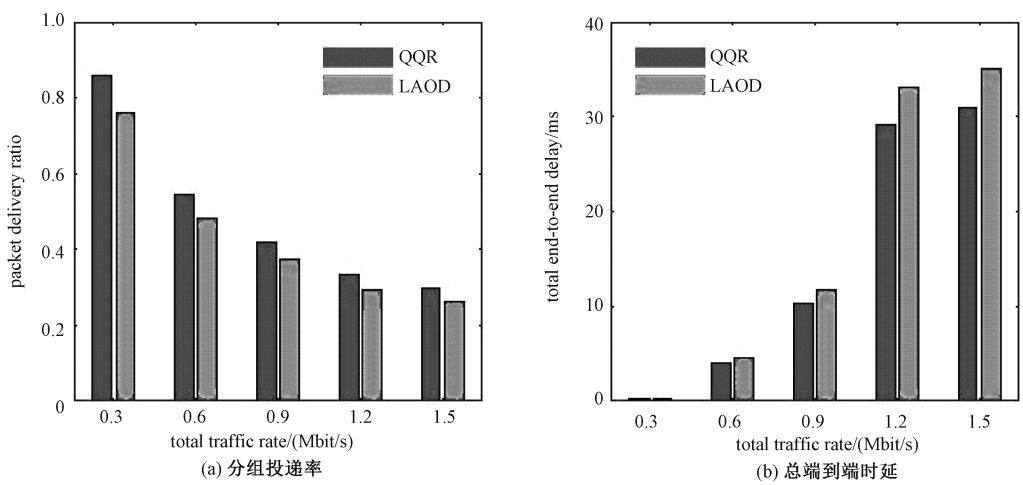
图 6　分组投递率和总平均端到端时延随仿真时间的变化情况

Fig. 6　Change of packet delivery rate and total average end-to-end delay with simulation time

如图 6(b)所示,在静态拓扑条件下,当业务负载保持不变时,经过 2 s 后 LAOD 协议完成路由发现,通过稳定的链路传输数据包,所以平均端到端时延保持在一个平均水平。而 QQR 协议初始化时 Q 值表还未建立,在仿真初期需要发送一定的数据包来建立并更新 Q 值表,Q 值表未收敛时网络中还不存在完整的转发路径,或者转发路径较长,造成大量的数据包积累,所以仿真开始阶段平均端到端时延较 LAOD 更高。仿真后期随着 Q 值表慢慢收敛,逐渐产生稳定且跳数较优的传输路径,故而平均端到端时延逐渐变小,最后趋于稳定。由于 QQR 考虑了网络中链路可用带宽,优先选择可用带宽较大的链路转发数据包,减少了不必要的数据包接入排队,从而提升了时延性能。QQR 协议中稳定状态与未收敛条件下的时延差值可以看做是 Q 学习反馈代价在时延性能上的体现。

2)分组投递率和平均端到端时延随全网业务负载的变化

依次改变单条业务流的负载为 100、150、200 和 250 Kbit/s,统计不同业务负载条件下的分组投递率和平均端到端时延,仿真结果如图 7(a)和图 7(b)所示。小负载条件下 2 种协议均保持较高的分组投递率和较低的平均端到端时延,随着网络总负载的增加,分组投递率下降,平均端到端时延随之增加。由于考虑到链路可用带宽,QQR 协议会轮换使用负载较轻的节点当作中继节点,从而减少分组冲突和网络拥塞,因此整体来看 QQR 的分组投递率要高于 LAOD,同时平均端到端时延比 LAOD 更小。



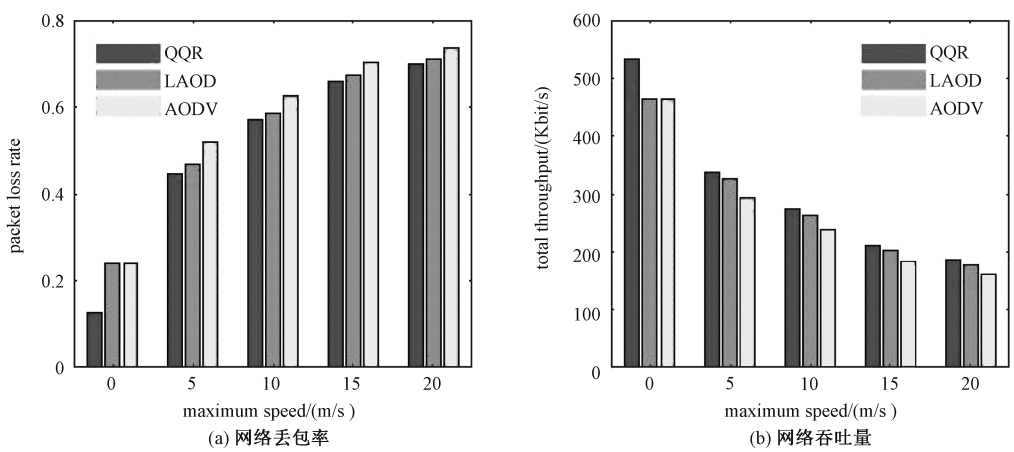
(a) 分组投递率　　(b) 总端到端时延

图 7　不同业务负载下的分组投递率和总平均端到端时延

Fig. 7　Packet delivery rate and total average end-to-end delay under different traffic load

3.3 运动拓扑

在 3.2 节的静态拓扑条件下增加节点的运动性,设置节点运动模型为 random waypoint,停留时间为 0 s,最小速率为 0 m/s,最大速率依次设置为 0、5、10、15 和 20 m/s,统计全网的丢包率和吞吐量如图 8(a)和图 8(b)所示。随着节点运动速率的加快,通信链路断裂变得频繁,3 种协议下的网络丢包率均呈增大趋势,相应的网络吞吐量不断减小。然而,通过周期性交互的 Hello 分组和转发的数据分组,QQR 协议的 Q 值表得以不断地更新,Q 学习的任务被分配到每一个节点中,使得算法能够快速地收敛到最优路径。重要的是 QQR 协议综合考虑了邻居节点数、链路持续时间和链路可用带宽 3 个指标,对网络拓扑的变化能够做出及时的调整,不需要在拓扑变化导致链路断裂时重启路由发现交互控制信息,因此较 AODV 和 LAOD 协议的丢包率更低,吞吐量更大。LAOD 协议考虑了链路持续时间和路由跳数,能够对运动拓扑做出反应,故而网络性能优于 AODV 协议。当拓扑运动速率增大到一定程度,QQR 协议的性能较 LAOD 没有太大优势,因为 Q 值的收敛需要一定的时间,QQR 协议很难适应运动速率很快的网络场景,因此需要改进和提高 Q 值收敛速度以获得更好的网络性能。



(a) 网络丢包率 (b) 网络吞吐量

图 8 不同运动速率下的网络丢包率和网络吞吐量

Fig. 8 Packet loss rate and network throughput at different motion rates

4 结论

Q-learning 作为一种离策略、无模型的启发式强化学习方法,为无线自组织网络路由设计提供了新的思路。本文研究一种基于 Q-learning 的 QoS 路由方法,该方法以 Q-learning 学习框架为基础,将邻居节点数量、链路持续时间和链路可用带宽作为路由测度,设计了一种提供 QoS 保证的奖励函数。网络节点通过广播 Hello 消息和发送数据分组交互路由测度信息,并根据奖励函数计算和更新 Q 值,待转发数据分组的节点根据其维护的 Q 值表智能选择下一跳转发节点。EXata 网络仿真结果证明了该方法的有效性,该方法能为高动态飞行自组网提供可靠的通信链路。后续工作将围绕大规模网络和高业务量场景中 Q 值表的快速收敛和动态维护问题展开,将 Q-learning 与神经网络结合使用也是未来的研究方向之一。此外,可以搭建无人机或无人车硬件系统,将本文实现的 QQR 路由算法移植到移动自组网节点的协议栈中,对路由算法的耗时性和实时性等性能进行实测。

参考文献

[1] Bekmezci İ, Sahingoz O K, Temel Ş. Flying ad-hoc networks (FANETs): a survey[J]. Ad Hoc Networks, 2013, 11(3): 1254-1270.

[2] Moussaoui A, Boukeream A. A survey of routing protocols based on link-stability in mobile ad hoc networks[J]. Journal of Network and Computer Applications, 2015, 47: 1-10.

[3] Fan X R, Cai W L, Lin J Y. A survey of routing protocols for highly dynamic mobile ad hoc networks[C]//2017 IEEE 17th International Conference on Communication Technology (ICCT). October 27-30, 2017, Chengdu, China. IEEE, 2017: 1412-1417.

[4] 陈峥,单剑锋,俞能海. 移动 Ad hoc 网络的分簇后择路由协议[J]. 中国科学技术大学学报, 2008,38(12):1372-1375.

[5] 张福全,吴寅,杨绪兵. 移动性度量的移动时延容忍网络路由策略[J]. 中国科学技术大学学报, 2019, 49(2):

132-137.

[6] Chaudhari S S, Biradar R C. Survey of bandwidth estimation techniques in communication networks[J]. Wireless Personal Communications, 2015, 83(2): 1425-1476.

[7] Chettibi S, Chikhi S. A survey of reinforcement learning based routing protocols for mobile ad-hoc networks[C]// Recent Trends in Wireless and Mobile Networks. CoNeCo 2011, WiMo 2011. Springer, Berlin, Heidelberg, 2011, 162: 1-13.

[8] Alsheikh M A, Lin S W, Niyato D, et al. Machine learning in wireless sensor networks: algorithms, strategies, and applications[J]. IEEE Communications Surveys & Tutorials, 2014, 16(4): 1996-2018.

[9] 张德干, 葛辉, 刘晓欢, 等. 一种基于 Q-Learning 策略的自适应移动物联网路由新算法[J]. 电子学报, 2018, 46(10): 2325-2332.

[10] 李荣, 王芳, 景栋盛, 等. 一种基于 Q 学习的无线传感网络路由方法[J]. 计算技术与自动化, 2017, 36(2): 155-160.

[11] 王月娟, 张苏宁, 吴水明, 等. 基于秩的 Q-路由选择算法[J]. 计算机与现代化, 2018(10): 1-5.

[12] Liang X D, Balasingham I, Byun S S. A reinforcement learning based routing protocol with QoS support for biomedical sensor networks[C]// 2008 First International Symposium on Applied Sciences on Biomedical and Communication Technologies. October 25-28, 2008, Aalborg, Denmark. IEEE, 2008: 1-5.

[13] Hu T S, Fei Y S. QELAR: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks[J]. IEEE Transactions on Mobile Computing, 2010, 9(6): 796-809.

[14] 朱明, 雷磊, 朱钢, 等. 多跳无线 DCF 自组网 MAC 层可用带宽预测[J]. 小型微型计算机系统, 2016, 37(10): 2217-2221.

[15] Chen Y J, Wang G J, Peng S C. Link lifetime-based segment-by-segment routing protocol in MANETs[C]//2008 IEEE International Symposium on Parallel and Distributed Processing with Applications. December 10-12, 2008, Sydney, NSW, Australia. IEEE, 2008: 387-392.

[16] Waheed A, Wahid A, Shah M A. LAOD: link aware on demand routing in flying Ad-Hoc networks[C]//2019 IEEE International Conference on Communications Workshops (ICC Workshops). May 20-24, 2019, Shanghai, China. IEEE, 2019: 1-5.

[17] Wang H P, Cui L. An enhanced AODV for mobile ad hoc network[C]// 2008 International Conference on Machine Learning and Cybernetics. July 12-15, 2008, Kunming, China. IEEE, 2008, 2: 1135-1140.