基于Q-learning的RFID多阅读器防碰撞算法

袁 源 郑嘉利 石 静 王 哲 李 丽

(广西大学计算机与电子信息学院 南宁 530004) (广西多媒体通信与网络技术重点实验室 南宁 530004)

摘 要 为了解决无线射频识别(RFID)系统中多阅读器与标签通信的碰撞问题,文中将此问题建模为马尔可夫决策过程,并提出了一种基于Q-learning的防碰撞算法。该算法通过智能体agent不断与周围环境进行交互和学习,从而产生Q值函数,得到最佳信道分配策略;取消了HiQ算法中复杂的分层结构,简化了系统模型,引入ε贪婪策略以得到全局最优解,改进奖赏函数以得到最优状态。仿真结果表明,与HiQ算法和EHiQ算法相比,该智能算法能够自适应地为阅读器分配不同的信道来进行数据传输,从而有效降低碰撞率,提高信道利用率和吞吐率。

关键词 无线射频识别,Q-learning,阅读器防碰撞,Q值

中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.018

Anti-collision Algorithm Based on Q-learning for RFID Multiple Readers

YUAN Yuan ZHENG Jia-li SHI Jing WANG Zhe LI Li

(School of Computer and Electronics Information, Guangxi University, Nanning 530004, China)

(Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning 530004, China)

Abstract Due to the collision problem between multiple readers and tags communication in RFID system, this paper modeled the problem as a Markov decision process, and proposed an anti-collision algorithm based on Q-learning. By continuously interacting with the environment, the Q-value function is generated, as well as the optimal channel resources allocation. The complex hierarchical structure in HiQ algorithm is eliminated for simplifying the system model. The algorithm not only imports the concept of ε-greedy strategy to obtain the global optimal solution, but also improves the reward function to get the best state. Simulation results show that compared with HiQ and EHiQ, this intelligent algorithm can adaptively assign different channels to the reader for data transmission, therefore reduces the collision rate and improves the channel utilization and throughput rate.

Keywords Radio frequency identification, Q-learning, Reader anti-collision, Q-value

1 引言

随着物联网技术的普及,RFID技术被越来越广泛地应用到各行各业中,如身份证系统、车牌识别、公交卡系统等。RFID技术,即射频识别技术,是一种无线通信技术,即一种非接触的自动识别技术,其基本原理是通过无线电信号识别特定目标并读写相关数据。射频识别是目前国际上发展最迅速的技术之一,它具有非接触性、识别距离远、抗干扰、环境适应性强、保密性高等优点。

在RFID系统中,受到外部环境和自身系统的影响,会存在两种形式的碰撞——多标签碰撞和多阅读器碰撞。多阅读器防碰撞算法主要分为调度方式和功率调整方式,其中功率调整方式的算法通过动态调整阅读器信道功率的范围来减少碰撞,但是此类算法比较复杂、效率较低,而且需要在无线射频识别系统中加入新的中心控制设备,系统成本高[1-2],因此本文的研究属于调度方式的一种。

文献[3]提出的基于分布式的TDMA的防碰撞算法——Colorwave算法中,每个阅读器都被分配了对应的时隙,即分配了对应的颜色,阅读器网络中任意两个相邻阅读器分配到相同颜色的概率最小则达到防碰撞的效果。文献[4]提出的PULSE算法采用了双信道机制,控制信道和数据信道分别适用于阅读器之间的通信和与标签之间的相互通信。但是,工作中的阅读器每进行一次通信都要竞争一次信道,增加了系统能耗;控制信道的范围远大于数据信道的范围,会造成系统效率降低。从而得到最优时隙资源和频率资源。文献[6]提出了一种基于退火策略的混沌神经网络防碰撞算法(RAC-ACNN),其为任何可能存在碰撞的阅读器都分配了不同的时隙。这种算法运行在控制中心上,属于集中式算法,不适合阅读器网络拓扑结构频繁变化的场景。文献[7]提出的HiQ算法是较为经典的Q学习算法,通过学习阅读器的冲突模式并有效地分配频率,采用分层结构——阅读器、R-Servers和Q-Servers,以确定频率和时间分配,通过与系统的反复交互找到最佳频率分配;但该算法的分层结构增加了阅读器设计的难度,而且算法比较复杂。文献[8]通过使用CSMA技术

到稿日期:2018-05-14 返修日期:2018-08-13 本文受国家自然科学基金项目(61761004)资助。

袁 源(1995-),女,硕士生,主要研究方向为多媒体通信网络理论与技术;郑嘉利(1979-),男,教授,主要研究方向为多媒体通信、物联网技术,E-mail:zhengjiali@vip.163.com(通信作者);石 静(1992-),女,硕士生,主要研究方向为多媒体通信网络理论与技术;王 哲(1993-),男,硕士生,主要研究方向为多媒体通信及其网络工程;李 丽(1994-),女,硕士生,主要研究方向多媒体通信网络理论与技术。

并测量接收信号功率,提出了一种基于距离的 RFID 阅读器避险协议,该防冲突协议在密集读卡器环境中提供了比其他协议更高的吞吐量。文献[9] 基于 TDMA 和 SDMA 算法的创新组合,提出了一种有效的快速检测和冲突避免(FD-CA)算法来解决阅读器碰撞问题,该算法有效地增加了吞吐量,减少了平均等待时间和平均速率冲突。杨健等[10] 在文献[7]的基础上提出了一种增强型的 HiQ 读写器 MAC 协议——EHiQ 算法。

针对上述调度算法比较单一,且当阅读器数量庞大时不能有效降低碰撞率的问题,本文提出一种基于 Q-learning 的多阅读器防碰撞算法,不仅取消了 HiQ 中的分层结构,而且采用 ε 贪婪策略,即随机探索的概率从大到小变化,使系统尽可能地选择最优动作;同时改进了奖赏函数的定义,通过最优策略得到最优状态,使阅读器之间碰撞的可能性最小,从而分配最佳的信道资源。

2　Q-learning 防碰撞算法

2.1　Q-learning 算法

Q-learning 算法是强化学习中非常著名的一种算法[10]。它的目的是学习特定状态 s 下特定动作 a 的价值,使奖赏值达到最大,然后得到最优策略。Q-learning 包含一个学习智能体 agent,即决策过程中的行为主体,智能体 agent 根据 Q 值选择最佳的动作,得出执行动作的奖赏值 c。如图 1 所示。

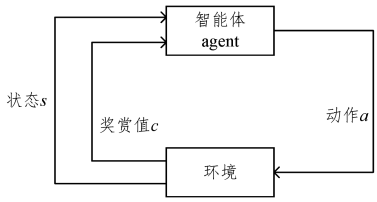

图 1　智能体与环境交互示意图

Fig. 1　Agent interaction with environment

在多阅读器防碰撞问题中,将环境看作马尔可夫过程(Markoff Decision Process, MDP)[11],该过程可以表示为一个五元组(S,A,P,γ,C)。其中,S 是有限状态集合;A 是有限动作集;P 表示转移概率,即 $P_a(s,s')$ 表示 t 时刻系统处于状态 s 下,采取动作 a 后系统状态在 $t+1$ 时刻变为 s' 的概率,也就是马尔可夫的性质,即一个随机过程的未来状态的条件概率分布仅仅依赖于当前状态与该状态下的动作;γ 为折扣因子;C 为奖赏值,表示系统处于状态 s 时执行动作 a 后得到的奖赏值。

马尔可夫过程的核心是找到最优策略 π,它是状态集到行为集的映射。为了评价系统的好坏,还需定义一个值函数 V,它是执行一个方案所能获得的奖赏期望值,定义如下:

$$V^{\pi}(s,a)=E(\sum_{t=0}^{\infty}\gamma^{t}c_t\mid s_0=s) \tag{1}$$

其中,E 表示数学期望,$\gamma\in[0,1)$ 表示折扣因子,c_t 表示在 t 时刻产生的即时奖赏值。将上式用递归的形式表示为:

$$V^{\pi}(s,a)=C(s,a)+\sum_{s'\in S}P_{ss'}(a)V^{\pi}(s') \tag{2}$$

其中,$C(s,a)=E[c(s,a)]$ 是 c 的平均值,$P_{ss'}(a)$ 是从状态 s 转移到状态 s' 的概率,s' 表示下一状态。最优策略 π^* 满足 Bellman 准则,即:

$$V^*(s)=V^{\pi^*}(s)=\min_{a\in A}[C(s,a)+\gamma\sum_{s'\in S}P_{ss'}(a)V^*(s')] \tag{3}$$

由策略 π,定义 Q 函数如下:

$$Q^{\pi}(s,a)=C(s,a)+\sum_{s'\in S}P_{ss'}(a)V^{\pi}(s') \tag{4}$$

其中,$Q^{\pi}(s,a)$ 是根据策略 π,状态 s 执行动作 a 后的期望奖赏值。令:

$$Q^*(s,a)=Q^{\pi^*}(s)=C(s,a)+\gamma\{\sum_{s'\in S}P_{ss'}(a)V^*(s')\} \tag{5}$$

其中:

$$V^*(s,a)=\min_{a\in A}[Q^*(s,a)] \tag{6}$$

那么 Q^* 表示为:

$$Q^*(s,a)=Q^{\pi^*}(s)=C(s,a)+\gamma\{\sum_{s'\in S}P_{ss'}(a)\min_{a'\in A}V^*(s',a')\} \tag{7}$$

然后以递归方式寻求最优值 $Q^*(s,a)$,Q 表的初始值为 0,那么 Q-learning 的准则如下:

$$Q_{t+1}(s,a)=\begin{cases}Q_t(s,a)+\alpha\{[c(s,a)+\gamma\min\limits_{a'\in A}Q_t(s',a')]-Q_t(s,a)\}, \\ \qquad \text{if } s=s_t \text{ 且 } a=a_t \\ Q_t(s,a), \text{if } s\neq s_t \text{ 或 } a\neq a_t\end{cases} \tag{8}$$

其中,γ 为折扣因子,满足 $0\leqslant\gamma<1$;α 为学习效率:

$$\alpha=\frac{1}{1+\text{当前状态 } s \text{ 被访问次数}} \tag{9}$$

如果每个可访问的(s,a)对的 Q 值是可以无限访问的,并且学习速率 α 以合适的方式减小到 0,那么 Q 值将以概率 1 收敛。根据 Q 值的学习准则,通过不断更新迭代 Q 表,可以得出最佳信道资源[12-13]的分配方案。

2.2　ε 贪婪策略

在 HiQ 算法中使用低概率随机化来确定是否接受或拒绝潜在的移动,即每一时刻均选取最小值,这样容易陷入局部最优。因此,我们在 Q-learning 算法的学习过程中运用 ε 贪婪策略,执行状态 s 后需要进行动作的选择。ε 贪婪策略基于一个概率来对探索和利用进行折中,为了最大化每一次执行完动作后的奖赏值,需考虑两个方面:首先是每个动作带来的奖赏值,其次是要执行的奖赏值最大的动作。如果每个动作 s 对应的奖赏是一个确定值 c,那么把所有可能的动作都执行一遍就能找出奖赏值最大的动作。但是一个动作的奖赏值 c 是来源于一个概率分布,仅通过遍历一次动作不能准确地得到平均奖赏值。

每次尝试时,以 ε 的概率进行探索,即以均匀概率随机选取一个动作;以 $1-\varepsilon$ 的概率进行利用,即选择最大 Q 值所对应的动作。通常,ε 取一个较小的常数,如 0.1,若尝试次数足够多,则在一段时间后,动作的奖赏值都能很好地近似出来,最大化地探索了未知状态,而且尽可能地选择了最优动作,即选择了最优信道。

2.3　奖赏函数的定义

式(8)中的 $c(s,a)$ 为奖赏函数,$c(s,a)$ 评估了在某一时刻的 s 状态下执行动作 a 的奖赏值。更具体地说,它是当前阅读器 i 选择信道 k 时的奖赏,若执行动作后产生的效果是积极的,比如成功识别标签,则奖赏值低;若执行动作后产生的效果是不积极的,则奖赏值高。我们的目的是通过最优策略得到最优状态,即分配最佳信道资源,使阅读器之间碰撞的可能性降到最低。因此,我们对文献[7]中的奖赏函数进行了如下改进:

$$c(s,a)=n_1(s,a)c_1+n_2(s,a)c_2+n_3(s,a)c_3+n_4(s,a)c_4 \tag{10}$$

列出以下已知条件：有 N 个阅读器，R 个可用频率资源数，定义阅读器 i 的信道状态是 R 维向量，即 $\boldsymbol{U}(i)=\{u_1(i),u_2(i),\cdots,u_R(i)\}$，$i\in\{1,2,\cdots,N\}$，且

$$u_k(i)=\begin{cases}1, & \text{阅读器 } i \text{ 被信道 } k \text{ 占用}\\ 0, & \text{其他}\end{cases},k\in\{1,2,\cdots,R\}$$

其中，$n_1(s,a)$表示占用信道 k 且在阅读器 i 的干扰范围内的阅读器数量，$n_2(s,a)$表示不占用信道 k 且在阅读器 i 的干扰范围内的阅读器数量，$n_3(s,a)$表示占用信道 k 但不在阅读器 i 的干扰范围内的阅读器数量，$n_4(s,a)$表示不占用信道 k 且不在阅读器 i 的干扰范围内的阅读器数量，系数 c_1,c_2,c_3,c_4 通常满足 $c_1>c_2>c_3>c_4$，一般地，c_1 取正数，c_4 取负数。

3 算法流程

通过以上分析，可以得到算法的流程描述，如图 2 所示。

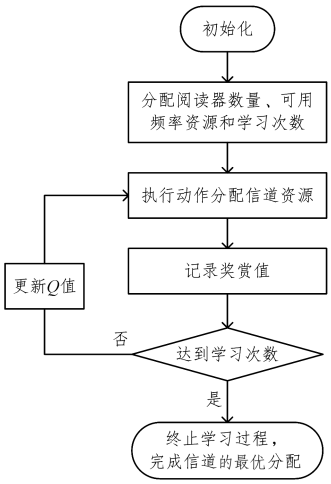

图 2 基于 Q-learning 算法的具体流程图

Fig. 2 Flowchart based on Q-learning algorithm

基于 Q-learning 算法的具体流程如下。

步骤 1 参数约定阅读器个数为 N，系统可用频率资源数为 R，学习次数为 T，贪婪策略概率为 ε，折扣因子为 γ，学习效率为 α，初始时刻 t 为 1。

步骤 2 初始化。根据 RFID 网络中阅读器和频率资源的个数，初始化状态空间 S 和动作空间 A，初始化 Q 表中所有的 $Q(s,a)=0$，其中 $s\in S,a\in A$；同时定义阅读器 i 的信道状态是 R 维向量 $\boldsymbol{U}(i)=\{u_1(i),u_2(i),\cdots,u_R(i)\}$，$i\in\{1,2,\cdots,N\}$，且

$$u_k(i)=\begin{cases}1, & \text{阅读器 } i \text{ 被信道 } k \text{ 占用}\\ 0, & \text{其他}\end{cases},k\in\{1,2,\cdots,R\}$$

步骤 3 执行动作。在当前状态 s 下产生一个随机数，若随机数大于 ε，或者对应的 Q 值都为 0，即第一次访问该状态，则随机选择动作；若随机数小于 ε，则选择 Q 表中最大值所对应的动作，通过该动作 a 分配一个信道资源给阅读器，$a\in\{1,2,\cdots,R\}$。某一时刻的状态由 i 和 $A(i)$ 确定，表示为 $(i,A(i))$，其中 i 表示当前阅读器，$i\in\{1,2,\cdots,N\}$，$A(i)$表示当前阅读器的可用信道资源数，$A(i)\in\{1,2,\cdots,R\}$。所述折扣因子 γ 满足 $0\leqslant\gamma<1$。

步骤 4 记录奖赏值。奖赏值 $c(s,a)$评价动作产生后的效果，其值由公式 $c(s,a)=n_1(s,a)c_1+n_2(s,a)c_2+n_3(s,a)c_3+n_4(s,a)c_4$ 确定，执行步骤 3 中的动作 a 之后，记录下奖赏值 c 和下一状态 s'。

步骤 5 更新 Q 表。根据奖赏值 $c(x,k)$和 Q 值更新函数(式(8))，对应更新 $Q_{t+1}(s,a)$的值。当 $t<T$ 时，转到步骤 3；当 $t>T$ 时，终止学习过程，完成信道的最优分配。

4 仿真测试与实验结果分析

本文使用 Python 3.0 来模拟仿真算法，使用 Matlab 2016a 对得到的实验数据进行分析并做图。仿真次数设为 12×10^5，为了排除数据的异常性，使实验结果更准确，仿真中每隔 24000 次取一个平均值作为数据点标注在图中。本文采用类似于 HiQ 算法的方形阅读器拓扑结构，这里举例说明有 25 个阅读器的拓扑图，如图 3 所示，图中圆表示阅读器，圆之间的连线表示频率约束，16 个阅读器和 9 个阅读器分别为 4 * 4 和 3 * 3 的网络拓扑结构。

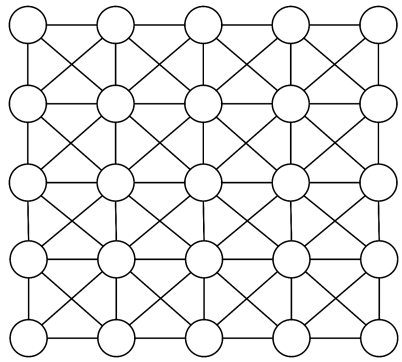

图 3 方形阅读器的拓扑结构图

Fig. 3 Topology of square reader

为了验证算法，首先对没有 Q-learning 的防碰撞算法进行仿真，如图 4 所示。在这种情况下，分别对 9 个阅读器、16 个阅读器和 25 个阅读器的系统给定信道数量 5，9 和 16。由仿真结果可知，在没有 Q-learning 学习机制时，结果呈一条平稳的直线，表示没有学习的过程。本文考虑一个频率对应一个信道，频率碰撞率是指发生碰撞的信道数与系统可用信道资源数的比值。

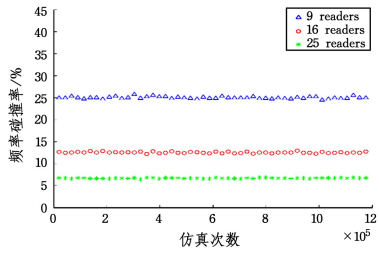

图 4 无 Q-learning 的频率碰撞率

Fig. 4 Reader collision rate without Q-learning

仿真 Q-learning 算法图，分别对 9 个阅读器、16 个阅读器和 25 个阅读器的系统分配 5 个、9 个和 16 个信道，加权系数设置为 $c_1=5,c_2=1,c_3=-1,c_4=-3$，折扣因子 γ 设为 0.9，学习效率 α 设为 0.1，ε 贪婪策略中的 ε 设为 0.1，即以 0.1 的概率随机选择动作，并以 0.9 的概率选择最大 Q 值对应的动作。结果如图 5 所示，由结果可见，当阅读器数量为 25 时，由于系统分配了 16 个信道资源，因此频率碰撞率最

低;当阅读器数量为 9 时,由于系统只分配了 5 个信道资源,因此频率碰撞率最高,且开始时碰撞率比较高,经过一段时间的学习后,最终保持在一个稳定的值,这说明在运用了 Q-learning 算法后,在开始的学习过程中系统不稳定,经过多次学习后,系统已学习到该如何分配信道资源,以使得碰撞率保持在较低水平。比较图 4 和图 5,当有 5 个信道资源、9 个阅读器时,使用本算法达到稳定后碰撞率降低了51.61%;当有 9 个信道资源、16 个阅读器时,碰撞率降低了 39.04 %;当有 16 个信道资源、25 个阅读器时,碰撞率降低了 40.69%。

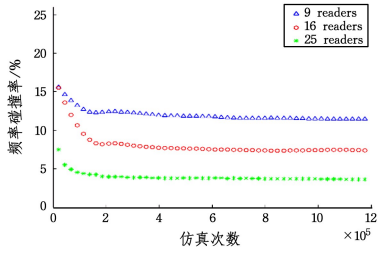

图 5　应用本算法后的频率碰撞率

Fig. 5　Reader collision rate by applying proposed algorithm

图 6 是在 16 个信道数量和 25 个阅读器的条件下,本算法与 HiQ 算法、EHiQ 算法的频率碰撞率比较图。由图可见,相比于 HiQ 算法和 EHiQ 算法,本文提出的算法的频率碰撞率显著降低,达到稳定时,相比 EHiQ 算法降低了67.03%,相比 HiQ 算法降低了 54.94%,这是由于引入 ε 贪婪策略和改进奖赏函数后系统尽可能地选择最优动作,从而降低了碰撞率;而且本算法的收敛速度比 HiQ 算法和 EHiQ 算法快,这说明本算法能在较短时间内达到稳定,这是由于本算法采用的结构简单,没有采用 HiQ 算法的分层结构,而仅是将阅读器与其对应的可用信道数定义为系统的状态,将信道资源定义为系统的动作,只需要维护 Q 表并从 Q 表中选择动作,从而使系统效率得到提高。

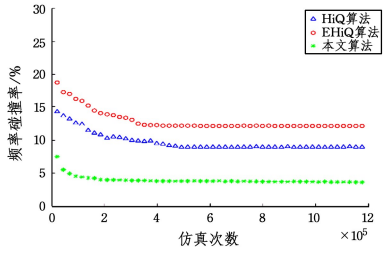

图 6　不同算法频率碰撞率的对比

Fig. 6　Comparison of reader collision rates by applying different algorithms

结束语　本文提出了一种基于 Q-learning 阅读器的防碰撞算法。整个过程包括建立马尔可夫模型、定义状态与动作空间、迭代学习、更新参数以及寻找最佳策略,最后根据强化学习得到的 Q 表判断如何最有效地分配信道资源。相比于经典的 HiQ 算法和 EHiQ 算法,本文算法在引入 ε 贪婪策略和改进奖赏函数后,频率碰撞率得到了显著降低,效率得到了提高。本算法在有效降低阅读器碰撞率后,能够有效提高标签的识别率,对物联网等应用具有相当高的实用价值。后续工作主要是考虑当系统作用范围较大时,如何通过深度学习方法来有效减少碰撞,从而提高系统吞吐率,达到资源利用的最大化。

参 考 文 献

[1] WEI D X,ZHENG J L,LI L L,et al. Study of Novel Adaptive Multi-tree Anti-collision Search Algorithm[J]. Computer Science,2013,40(10):52-55. (in Chinese)
韦冬雪,郑嘉利,李亮亮,等. 一种新颖的自适应多叉树防碰撞算法的研究[J]. 计算机科学,2013,40(10):52-55.

[2] XIE L,YIN Y F,CHEN X,et al. RFID Data Management:Algorithms,Protocols and Performance Evaluation[J]. Chinese Journal of Computers,2013,36(3):457-470. (in Chinese)
谢磊,殷亚凤,陈曦,等. RFID 数据管理:算法、协议与性能评测[J]. 计算机学报,2013,36(3):457-470.

[3] WALDROP J,ENGELS D W,SARMA S E. Colorwave:An Anticollision Algorithm for the Reader Collision Problem[C]// IEEE International Conference on Communications. IEEE,2003:1206-1210.

[4] BIRARI S M,IYER S. PULSE:A MAC Protocol for RFID Networks[C]// International Conference on Embedded and Ubiquitous Computing. Springer-Verlag,2005:1036-1046.

[5] SEO H,LEE C. A New GA-Based Resource Allocation Scheme for a Reader-to-Reader Interference Problem in RFID Systems [C] // IEEE International Conference on Communications. IEEE,2010:1-5.

[6] TIAN J,FAN Y,ZHU Y,et al. RFID Reader Anti-collision Using Chaos Neural network Based on Annealing Strategy[C]// World Congress on Intelligent Control and Automation,2008 (WCICA 2008). IEEE,2008:6128-6132.

[7] HO J,ENGELS D W,SARMA S E. HiQ:a Hierarchical Q-learning Algorithm to Solve the Reader Collision Problem[C]// 2006 International Symposium on Applications and the Internet Workshops. IEEE,2006:88-91.

[8] GOLSORKHTABARAMIRI M,ISSAZADEHKOJIDI N. A Distance Based RFID Reader Collision Avoidance Protocol for Dense Reader Environments[J]. Wireless Personal Communications,2017,95(2):1-18.

[9] SAADI H,TOUHAMI R,YAGOUB M C E,et al. TDMA-SDMA based RFID algorithm for fast detection and efficient collision avoidance[J]. International Journal of Communication Systems,2018,31(3).

[10] YANG J,WANG Y H,CAI Q L,et al. EHiQ:A RFID Reader MAC Protocol Based on Enhanced HiQ[J]. Computer Science,2011,38(7):85-87. (in Chinese)
杨健,王永华,蔡庆玲,等. EHiQ:一种基于增强型 HiQ 的 RFID 读写器 MAC 协议[J]. 计算机科学,2011,38(7):85-87.

[11] LIU Q,ZHAI J W,ZHANG Z Z,et al. A Survey on Deep Reinforcement Learning[J]. Chinese Journal of Computers,2017,40(1):1-28. (in Chinese)
刘全,翟建伟,章宗长,等. 深度强化学习综述[J]. 计算机学报,2017,40(1):1-28.

[12] GU J Y,ZHANG G A,BAO Z H. Joint multi-path routing and channel assignment strategy for cognitive wireless mesh networks[J]. Computer Science,2011,38(5):45-48. (in Chinese)
顾金媛,章国安,包志华. 认知无线 Mesh 网络联合多路径路由和信道分配策略[J]. 计算机科学,2011,38(5):45-48.

[13] AVALLONE S,BANCHS A. A Channel Assignment and Routing Algorithm for Energy Harvesting Multiradio Wireless Mesh Networks[J]. IEEE Journal on Selected Areas in Communications,2016,34(5):1463-1476.