doi:10.11959/j.issn.1000-436x.2017178

# 基于 Q-learning 的 HTTP 自适应流码率控制方法研究

熊丽荣，雷静之，金鑫
（浙江工业大学计算机科学与技术学院，浙江 杭州 310023）

**摘 要**：基于 HTTP 的自适应流 HAS 已经成为自适应视频流服务的标准。在 HAS 客户端网络状态多变的情况下，硬编码形式的码率决策方法灵活性偏低，对用户体验考虑不足。为了优化用户体验质量（QoE），提出一种基于 Q-Learning 的码率控制算法，结合 HTTP 自适应视频流客户端环境进行建模并定义状态转移规则；量化与用户 QoE 相关的参数，构建新的回报函数；实验表明引入 Q-Learning 进行码率调整的自适应算法在码率切换的稳定性方面表现较好。

**关键词**：HTTP 自适应流；硬编码；Q 学习；码率控制；稳定性
**中图分类号**：TN919.85 **文献标识码**：A

# Research on Q-learning based rate control approach for HTTP adaptive streaming

XIONG Li-rong, LEI Jing-zhi, JIN Xin
(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

**Abstract:** HTTP adaptive streaming (HAS) has become the standard for adaptive video streaming service. In changing network environments, current hardcoded-based rate adaptation algorithm was less flexible, and it is insufficient to consider the quality of experience (QoE). To optimize the QoE of users, a rate control approach based on Q-learning strategy was proposed. the client environments of HTTP adaptive video streaming was modeled and the state transition rule was defined. Three parameters related to QoE were quantified and a novel reward function was constructed. The experiments were employed by the Q-learning rate control approach in two typical HAS algorithms. The experiments show the rate control approach can enhance the stability of rate switching in HAS clients.

**Key words:** HTTP adaptive streaming, hardcoded, Q-learning, rate control, stability

## 1 引言

近年来，各界对多媒体内容传输特别是视频流服务越来越重视。在尽力交付的互联网上支持可靠视频流传输，基于 HTTP 的自适应流（HAS, HTTP adaptive streaming）已经成为视频业务技术的发展趋势。

HAS 采用码流切换技术动态调整码率，整个过程由自适应算法负责。可用带宽估算和码率选择是客户端码率自适应算法的 2 个核心功能。根据媒体片段的 TCP 平均下载吞吐量估算网络带宽[1~4]。码率决策从视频码率集中选择低于估算网络可用带宽的最大码率等级。这种码率选择方法易造成视频码率的频繁切换，给观看者带来不舒适的体验，且设定硬编码的码率选择策略灵活性偏低，无法应对变化多样的网络配置和网络带宽。

将自适应视频流的传输建模成优化控制问题可以提升码率决策的灵活性。增强学习通过早期离线训练，学习最优的控制策略，然后将策略应用在实时自适应控制中，能够提升客户端码率决策机制

**收稿日期**：2016-12-19；**修回日期**：2017-05-22
**基金项目**：浙江省重大科技专项重大工业基金资助项目（No.2012C11026-2）；杭州市重大科技创新专项基金资助项目（No.20132011A16）
**Foundation Items:** The Key Project of Science and Technology Planning Project of Zhejiang Province (No.2012C11026-2), The Special Found for Science and Technology Innovation Program of Hangzhou (No.20132011A16)

的灵活性和自适应性[5~7]。

本文设计基于 Q-learning 的码率决策，设计合适数量的环境状态组成元素，根据可用带宽和实时缓存数据填充量进行环境状态建模；从 HTTP 视频流服务质量的角度考虑，选择与用户体验质量（QoE）相关的 3 个方面[8]：视频质量等级高低、视频播放期间切换带来的损失及缓存区数据溢出危险性来构建新的回报函数。实验结果表明，在 HAS 自适应算法的码率决策部分引入本文提出的 Q-learning 算法，增强了码率决策的灵活性，视频码率切换的稳定性优于未结合机器学习方法的算法。

## 2　相关工作

流媒体的传输必须能够自适应地满足传输网络、客户端和用户需求，才能有效解决不同应用环境下视频的可靠高效传输问题。

HAS 客户端的自适应算法中通常基于 TCP 平均下载吞吐量来估算可用带宽。在此基础上，对估算出的带宽进行平滑处理去除一些异常值[2~4]，较好地应对带宽短期波动。但是，估算出的带宽实时性和准确性在网络状态多变的环境中较差。

混合控制的自适应机制综合考虑多个参数（如网络带宽、缓存等）以准确地反映当前网络状况和客户端情况。Li 等[9]提出的 PANDA 算法采用“探测和自适应”原理去探测实时带宽，基于“dead-zone”量化器决定选择的视频码率。设计调度时间策略，推动缓存趋近均衡级别变化。该算法能够尽可能维持缓存区数据量处于较均衡的范围，保持较高的平均视频质量。Hesse 等[10]提出了一种基于流会话期间媒体数据选择和恢复机制的 BS 算法。设计了一个缓存监控单元，规定缓存上下限，将缓存状态作为平滑函数中的输入参数，能够更好地利用缓存状态和网络带宽之间的关系进行码率选择。但是码率选择时采用最简单选择思想调整码率，即切换至可选码率集中低于当前可用带宽的最高码率等级。上述混合自适应控制机制中，码率选择策略灵活性偏低。在网络带宽多变环境中，视频码率切换的稳定性有待提高。

引入机器学习方法可以增强客户端码率决策机制的灵活性和自适应性。Garcia 等[11]提出了基于随机动态规划（SDP）的解决方案，基于一个估计先验系统模型计算出了离线的最优控制策略。文献[12,13]利用马尔可夫决策过程（MDP）驱动码率调整策略，考虑影响用户 QoE 的关键因子、最大化视频流的质量。

将增强学习应用到 HAS 质量选择的启发式中，可以使 HAS 客户端能够动态学习当前网络环境中的最佳行为。Istepanian 等[14]首次将增强学习方法应用于多媒体传输领域，用来处理无线网络传输医疗视频图像时的码率控制问题。Claeys 等[5]设计了一个 HAS 客户端自学习智能系统，将环境模型划分成 2 500 万个状态，但是该模型收敛遇到困难。后来他们精简了环境状态的组成元素，并重新设计了回报函数，得到性能较好的码率决策方法[6]。Mart 等[7]提出的模型能够在模型的复杂性和动态系统的自学习能力之间合理权衡，选定适合数量的参数构建环境状态，根据用户 QoE 相关的因子（the requested qualities, the quality switches, the freeze）定义回报函数。文献[15]提出(FA)Q-learning HAS 客户端，调整学习速率，增加环境变化后获得新信息的速度。该客户端已在整体性能和收敛速度方面完成了优化并且对多个网络配置进行了评估。Menkovski 等[16]利用增强学习的在线学习 SARSA 算法，构建了一个不需要离线训练的码率自适应码率调节算法，并通过实验证明算法的有效性。Chiariotti 等[17,18]构建了 Q-learning 算法和 SARSA 算法，并通过实验发现在预定带宽内 Q-learning 算法更高效，而在训练的带宽外 SARSA 算法的自适应能力更强。

本文构造基于 Q-learning 的码率决策方法，根据用户观看体验关注的目标，增强决策的灵活性、视频切换的稳定性。根据探测出的可用带宽和缓存状态定义环境状态和客户端行为。从量化 QoE 的 3 个不同方面构建回报函数，基于 Q-learning 算法设计状态转移规则。为了验证本文方法的有效性，将基于 Q-Learning 的码率决策方法与基于“探测和自适应”原理的估算带宽自适应机制[9]结合实现了 PANDA-RL 算法，与构建缓存状态变化因子平滑带宽的自适应机制[10]结合实现了 BS-RL 算法，实验证明在多客户端场景下引入本文的码率决策方法的码率切换稳定性更好。而相比于同类的增强学习算法[6]的工作，本文设计的算法在选择最优的状态时，尽量使缓存趋于均衡状态，保证视频质量的提升，这样能够有效地提高视频的播放质量，而文献[6]更注重视频质量的提升，但在缓存饥饿时，会发出切换最优状态的行为，这会出现较多视频卡顿现象。

## 3 基于 Q-Learning 的码率决策方法

本文将 HAS 系统中服务器和客户端的通信问题抽象成动态系统优化控制问题，应用增强学习的方法优化自适应算法性能。Q-learning 算法是一种与模型无关的增强学习算法[19]，通过与环境的交互获得关于回报的预期和环境状态转换的知识，用于求解马尔可夫决策过程的最大回报和最优策略。在学习过程中，$Q$ 值代表 Agent 已经获得的知识。在探索和利用之间进行折中是 Q-learning 中最具挑战的任务，通常使用的探索策略是 $\varepsilon$-greedy 算法[19]、Softmax 算法[20]或 VDBE-Softmax[21]，VDBE-SOFTMAX 与 Softmax 的探索策略收敛效率更高，同时也需要更复杂的计算。

### 3.1 环境建模

首先对环境状态建模和定义 HAS 客户端行为。

1) 环境状态。根据 2 个参数对 Agent 和网络环境交互的行为进行建模：一个是客户端感知的可用带宽，另一个是当前的客户端缓存数据填充状态，即缓存区内现有片段的总时长。最终确定环境状态的定义为 $s=(Bandwidth, Bufffer filling)$，这 2 个构成状态的元素，取值范围都是连续的区间，本文将它们离散成不同的级别。状态元素的取值范围和离散化后的级别如表 1 所示。

**表 1** 环境状态定义

| 状态元素 | 范围 | 级别 |
|---|---|---|
| *Bandwidth* | $[0; BW_{\max}]\text{bit}\cdot\text{s}^{-1}$ | $N+1$ |
| *Buffer filling* | $[0;\tau B_{\max}]\text{s}$ | $B_{\max}+1$ |

如表 1 所示，$BW_{\max}$ 是物理链路上可能达到的最高带宽，$B_{\max}$ 表示缓存区最多能够存储的片段数目，$N$ 表示码率等级的总数目。本文设定缓存最大容量是 30 s，选用数据集的视频内容的每个片段时长为 2 s。

2）客户端行为。通过感知当前的网络状态，从该状态可达的状态集合中选择将要转换的下一个状态，每个选择视作一种行为。

### 3.2 构建回报函数

回报函数（reward function）是 Agent 用来学习最优策略的基本向导。在视频流传输中，回报函数应该体现为对用户 QoE 的测量。视频流服务质量的评估是一个主观的问题，本文从影响视频观看 QoE 3 个方面来构建回报函数。

1) 转换后的状态映射的视频质量等级 $r_{\text{quality}}$，其定义如式(1)所示，按照等级的高低映射成正的回报，即奖励。

2) 视频播放期间切换带来的损失 $r_{\text{switches}}$，其定义如式(2)所示，根据切换的步长映射成负的回报，即惩罚。

3) 缓存区数据溢出危险性 $r_{\text{bufferstate}}$，其定义如式(3)所示，根据缓存区数据量处于不同区间映射成大小不等的负回报。

$$r_{\text{quality}}=Q_n \tag{1}$$

$$r_{\text{switches}}=-1.0\times\left|Q_n-Q_{n-1}\right| \tag{2}$$

$$r_{\text{bufferstate}}=\begin{cases}A_1\left|B[n]-B_{\text{down}}\right|, & B[n]<B_{\text{down}}\\ A_2\left|B[n]-B_r\right|, & B_{\text{down}}\leqslant B[n]\leqslant B_{\text{up}}\\ A_3\left|B_{\text{up}}-B[n]\right|, & B[n]>B_{\text{up}}\end{cases} \tag{3}$$

其中，$Q_n$ 和 $B[n]$ 分别表示第 $n$ 段请求的码率等级和发出请求前缓存区的数据填充量。$B_{\text{up}}$、$B_{\text{down}}$ 分别是缓存上溢阈值和下溢阈值，将其设定为缓存容量的 80%和 20%。$A_1$、$A_2$ 和 $A_3$ 是对应缓存区数据填充状态计算 $r_{\text{bufferstate}}$ 的不同权重值，可以根据不同的情形来给定它们的具体值。当 $B[n]<B_{\text{down}}$，缓存区数据填充量低于下溢阈值，说明缓存内数据消耗至空的概率很大，此时由 $A_1$ 给定其最大的惩罚值。当 $B_{\text{down}}<B[n]<B_{\text{up}}$，缓存的数据量百分比处于均衡区间内，本文希望缓存区数据量朝均衡级别 $B_r$（设定为缓存容量的 $\frac{2}{3}$）变化，越接近 $B_r$，惩罚值越小。当 $B[n]>B_{\text{up}}$ 时，缓存的数据量高于上溢阈值，说明缓存内数据爆满的概率很大，此时 $A_3$ 决定其惩罚值的大小。

总体回报函数会驱动 Agent 的探索向较高的码率等级、较少的码率切换和较均衡的缓存数据量的行为趋近。为了避免把视频播放停滞的优先级设定为最高，对缓冲区数据溢出危险性 $r_{\text{bufferstate}}$ 给定最强的惩罚权值，其比重必须要明显大于组成回报函数从其他 2 个方面获得的回报值。针对以上 3 个方面考虑回报计算，本文设计的线性总体回报函数为

$$r=C_1r_{\text{quality}}+C_2r_{\text{switches}}+C_3r_{\text{bufferstate}} \tag{4}$$

式(4)将影响 QoE 3 个方面的回报相加得到的

总和定为总体回报，$C_1$、$C_2$、$C_3$分别是对应$r_{\text{quality}}$、$r_{\text{switches}}$、$r_{\text{bufferstate}}$的权重值，根据侧重的 QoE 体现的方面来设定具体的值。

### 3.3　探索策略

在采用探索策略上，由于在实时系统决策中，本文发现采用 Softmax 探测算法需要大量的计算，会影响一定的网速。因此，本文先选用$\varepsilon$-greedy 贪心算法进行探索，探索的过程中以概率$\varepsilon(0\leqslant\varepsilon\leqslant1)$随机地选择下一步行为，以$1-\varepsilon$的概率选择将会带来最大长期回报的行为。$\varepsilon$-greedy 算法的选择策略$\pi(s)$为

$$\pi(s)=\begin{cases} A(s)\text{中的随机丢失行为}, & \xi<\varepsilon \\ \arg\max_{a\in A(s)} Q(s,a), & \text{其他} \end{cases} \tag{5}$$

其中，$0\leqslant\xi\leqslant1$是均匀分布的随机数，每次请求下载下一片段前$\xi$都会被随机赋值。若$\xi<\varepsilon$，Agent 随机从行为集$A(s)$中选一个行为执行。若$\xi\geqslant\varepsilon$，Agent 则遍历行为集$A(s)$中的所有行为，最终执行获得即刻最大$Q$值的某个特定的行为。

### 3.4　Q-learning 算法实现

在可用带宽和客户端缓存数据填充状态成功获取的基础上，本文提出的 Q-learning 算法的思想如算法 1 所示。

**算法 1**　Q-learning

**输入**　参数学习速率$\alpha$，折扣因子$\gamma$，回报函数$r$，环境状态$s$，客户端行为$a$。

**输出**　最终收敛的$Q$值

1) 初始化$Q(s,a)$为任意值；

2) 迭代探索 Repeat (for each episode)；

3) 根据当前$Q$和状态$s$，通过$\varepsilon$-greedy 策略选择动作$a$；

4) 执行动作$a$，通过获得回报值$r$，达到新的状态$\tilde{s}$；

5) $Q(s,a)=(1-\alpha)Q(s,a)+\alpha\left[r+\gamma\max_{\tilde{a}}Q(\tilde{s},\tilde{a})\right]$；

6) 更新当前状态$s:=\tilde{s}$；

7) Until　$Q$　收敛

## 4　实验

### 4.1　实验环境和性能指标

本文在真实网络环境下进行实验，搭建 Apache 服务器，将其结合 Ubuntu 12.04.4 系统提供 HTTP 服务。以 MPEG-DASH 标准的参考平台 libdash[22]为客户端，同时，利用服务器端的带宽控制器 Traffic Control(TC)实现带宽控制。基于 libdash 的客户端与 Web 服务器通信的整体结构如图 1 所示，基于 Q-learning 的决策机制作用于 libdash 库中算法的量化模块。选用实验数据“Big Buck Bunny”，预编码为$R=\{100,200,350,500,700,900,1100,1300\}$ kbit/s，共 300 个片段，每个视频片段持续时间$\tau=2$ s。为了验证实验效果，将本文提出的基于 Q-learning 的码率决策方法分别应用于考虑缓存建模和带宽估算 2 个角度、提高算法自适应性而又未涉及机器学习的混合控制自适应机制 PANDA 和 BS 方法中，并在多客户端竞争带宽的环境中比较新旧算法性能。

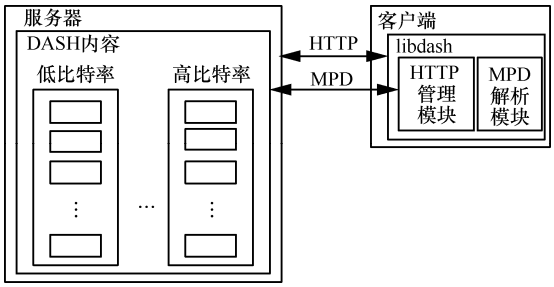



图 1　基于 libdash 的 HTTP 自适应流系统结构

为了搭建带宽波动的实验环境，在一条网络链路上接入 2 个客户端竞争带宽，设计 4 种实验情况如下。

1) 维持链路受限带宽为 1 000 kbit/s。

2) 实验带宽骤降，0~100 s 设置链路带宽为 1 000 kbit/s，100~200 s 链路带宽下降为 500 kbit/s。

3) 实验带宽骤升，0~100 s 设置链路带宽为 500 kbit/s，100~200 s 链路带宽上升为 1 000 kbit/s。

4) 实验带宽短期内多次波动，0~100 s 设置链路带宽为 1 000 kbit/s，100~150 s 后每隔 10 s 降低或升高带宽，直至 150 s 后维持在 1 000 kbit/s。

评价一个自适应算法的性能涉及多个评价标准之间的权衡。本文沿用提出 PANDA 算法的文献[9]定义的 4 个性能评价指标：缓存下溢性（underflow）、低效率性（inefficiency）、不稳定性（instability）、不公平性（unfairness），并且在此基础上添加了一个新的指标：缓存上溢性（overflow）。通过测算缓存区已有片段数目的增加幅度来表明一个算法防止缓存区上溢的能力。将$B_{\text{up}}$作为缓存上限阈值（缓存超过$B_{\text{up}}$有上溢危险），$B_t$是时刻$t$播放器的缓存区现有数据量，$t$的时间单位是 s。定义客户端处于时刻$t$的缓存上溢指数*Overflow*为

$$Overflow=\frac{\max(0,B_t-B_{\text{up}})}{B_{\text{up}}} \tag{6}$$

### 4.2 探索概率的评估

本文设置不同参数进行多次实验，验证得到 $\boldsymbol{Q}$ 矩阵经过多次迭代最后会达到收敛的状态。每次探索都会随机定位到状态集合中的某个状态，按 $\varepsilon$-greedy 贪心算法进行状态转换。根据选用的视频序列的片段个数，训练数据时将 300 次探索记作一次迭代结束。

为了评估探索概率对 Q-learning 探索过程的影响，设计以下实验，设置不同的探索概率和折扣因子，记录在算法收敛时的迭代次数和得到的平均回报值(average reward value)[10]。实验中设定 $\alpha=0.3$，折扣因子 $\gamma(0\leqslant\gamma\leqslant1)$ 的取值集合为 $\{0.1,0.5,0.7,0.9\}$，$\varepsilon\in\{0,0.5,1\}$，一组典型的实验结果如表 2 和表 3 所示。

**表 2 探索概率评估迭代次数实验结果**

| $\varepsilon$ | 收敛时的迭代次数 | | | |
|---|---|---|---|---|
| | $\gamma$=0.1 | $\gamma$=0.5 | $\gamma$=0.7 | $\gamma$=0.9 |
| 0 | 65 | 95 | 169 | 493 |
| 0.5 | 109 | 179 | 312 | 900 |
| 1 | 7 472 | 12 218 | 20 131 | 56 735 |

由表 2 可知，较大的探索概率 $\varepsilon$ 带来大量随机动作选择，随着 $\varepsilon$ 的增大，达到收敛状态时总迭代次数增加。

**表 3 探索概率评估平均回报值实验结果**

| $\varepsilon$ | 收敛时的平均回报值 | | | |
|---|---|---|---|---|
| | $\gamma$=0.1 | $\gamma$=0.5 | $\gamma$=0.7 | $\gamma$=0.9 |
| 0 | −3.805 95 | −0.698 535 | 7.238 54 | 52.5 |
| 0.5 | −3.805 91 | −0.698 533 | 7.238 56 | 52.500 1 |
| 1 | −3.804 86 | −0.693 75 | 7.239 58 | 52.506 3 |

每次训练都设定学习速率 $\alpha=0.3$，采用 $\gamma=0.7$ 的贪心策略进行探索。改变探索概率 $\varepsilon$ 的取值，实验中设置 $\varepsilon$ 的取值范围是 0~1，取值的集合是 $\{0.1,0.3,0.5,0.7,0.9\}$，不同的 $\varepsilon$ 对应的平均回报值变化如图 2 所示。

由图 2 可知，随着探索概率 $\varepsilon$ 增大，收敛至稳定平均回报值的速度变慢。

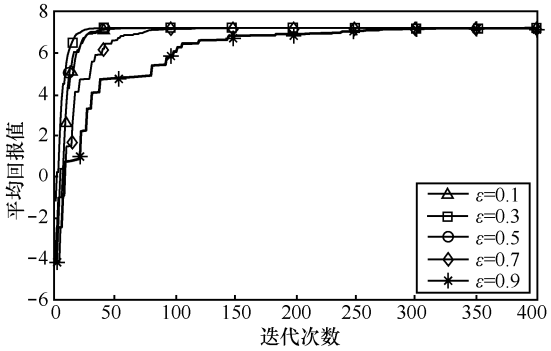



图 2 $\varepsilon$ 对应的平均回报值变化

### 4.3 算法有效性分析

本文设定权重值 $A_1=-5$、$A_2=-0.1$、$A_3=-1$、$C_1=1$、$C_2=2$、$C_3=1$ 构建回报函数，取 $\alpha=0.3$、$\gamma=0.7$、$\varepsilon=0.5$ 离线训练 $\boldsymbol{Q}$ 矩阵的值，用最终得到的 $\boldsymbol{Q}$ 矩阵进行码率决策。结合 PANDA 算法中基于“探测”的原理估算带宽，形成 PANDA-RL 算法。结合 BS 算法中利用随缓存动态变化的平滑因子处理带宽，形成 BS-RL 算法，在 libdash 上分别实现这 4 种算法。通过 TC 调整链路带宽，设置成第 4 种实验情况，根据得出的实验数据，对比 PANDA-RL 算法、PANDA 算法及 BS-RL 算法、BS 算法利用带宽的效果。

由图 3 可知，PANDA-RL 算法对带宽波动反应优于原 PANDA 算法，实验过程中获得的码率更贴近实时变化的带宽。PANDA-RL 算法的码率调整较 PANDA 算法更加平稳，减少了切换的次数。同理，图 4 中 BS-RL 算法相比于原 BS 算法，对于网络带宽的反应也更加合理和灵敏，同时，码率切换也没有 BS 算法频繁。以上实验结果说明，引入 Q-learning 算法进行码率决策能够尽量避免在可用视频码率级别中频繁而显著地改变视频码率，降低切换代价。

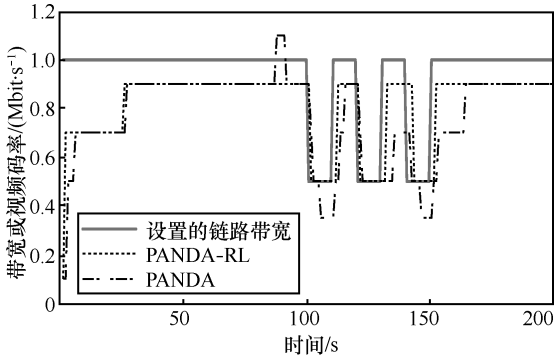



图 3 带宽短期内多次波动时 PANDA-RL 算法和 PANDA 算法的客户端获得的码率变化

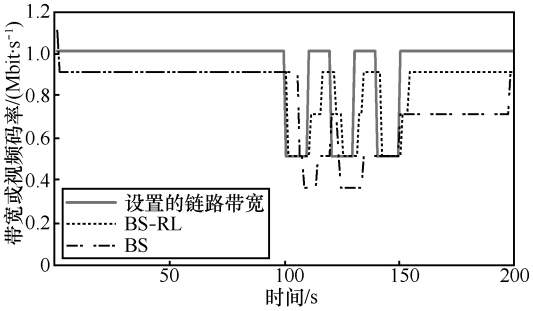



图 4　带宽短期内多次波动时 BS-RL 算法和 BS 算法的客户端获得的码率变化

### 4.4　性能对比

实验过程分别由 PANDA 算法、BS 算法、PANDA-RL 算法和 BS-RL 算法驱动的 libdash 客户端在上述 4 种性能实验情况中持续 200 s 下载视频片段，得到对应性能指标的相关数据。多客户场景中有 2 个客户端同时在网络链路上竞争带宽，可能出现性能差异较大的问题，本文对每种算法的 2 个客户端在每种实验情况下得到的数据求平均值，评估多客户端场景中自适应算法的整体性能。5 个性能评估指标为缓存上溢性、缓存下溢性、低效率性、不稳定性、不公平性，这 5 个指标的具体值越小说明算法性能越佳。

这 5 个性能评估指标在不同的实验情况下具体的值是不同的，波动的特点主要与自适应算法本身的设计有关。如图 5 所示，不同算法在 4 种实验情况下的缓存上溢性波动趋势不同。由图 5~图 7 可知，本文实现的 2 种算法在 4 种实验情况下的指标节点值都比原算法要小，说明本文方法能提升算法在缓存上下溢性和低效率性方面的性能，但是效果不明显。

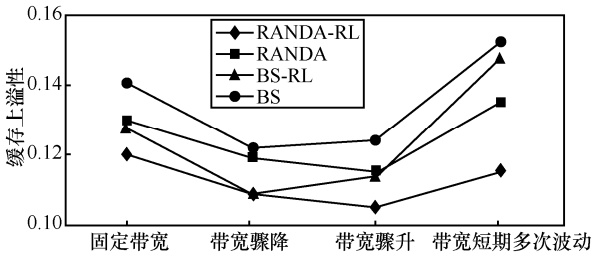



图 5　缓存上溢性

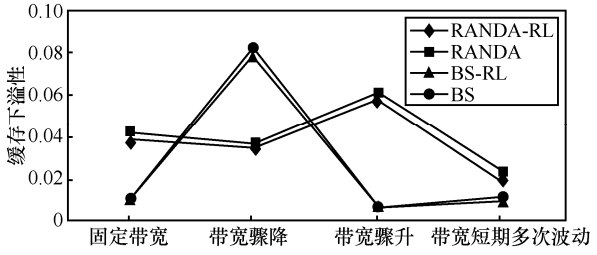



图 6　缓存下溢性

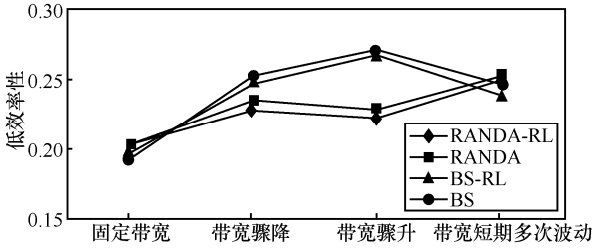



图 7　低效率性

由图 8 可知，当网络带宽出现波动，无论是带宽骤降、骤升还是短期内多次波动，引入 Q-learning 进行码率决策的 PANDA-RL 算法和 BS-RL 算法相比于量化策略采用“Dead-zone”量化器的 PANDA 算法和简单码率选择思想的 BS 算法，不稳定性明显降低，说明增强学习的方法用在量化策略中平滑码率的切换效果好。图 9 表明，对于主动探测带宽的 PANDA 算法来说，引入 Q-learning 算法后多客户共享带宽的公平性的提升比 BS 算法显著。

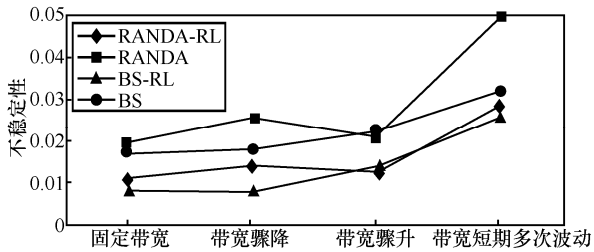



图 8　不稳定性

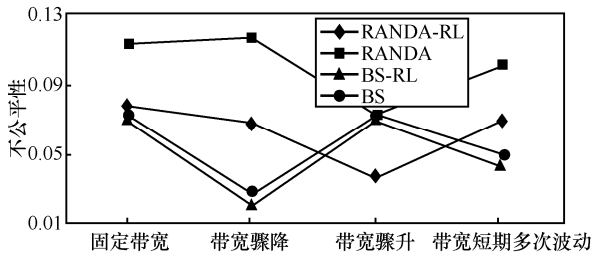



图 9　不公平性

为了分析本文设计的算法性能，将本文的工作和文献[6]的工作进行比较。将 BS 算法与文献[6]算法结合形成 BS-MAY，将 PANDA 算法与文献[6]结合形成 PANDA-MAY。为了尽量使整个模型收敛，本文在线下分别训练了 1 500 个视频片段。同时，也采用文献[6]平均意见得分（MOS）作为主要评价指标，平均意见得分越高，则表示策略越好。

由图 10 可知，在带宽固定的状态下，本文设计的 BS-RL 和 PANDA-RL 算法与融合文献[6]的强化学习方法的 BS-MAY 与 PANDA-MAY 这 2 种算法平均 MOS 提升 5%；而在带宽骤降和带宽骤升中文献[6]

和本文设计的算法都能够有效地应对带宽突变，两者的效率差距不足 1%，没有明显的差别；在带宽的短期多次波动中本文设计的强化学习算法平均 MOS 比文献[6]提升 7%，说明在带宽的短期多次波动中，本文设计的强化学习方法具有更好的稳定性。

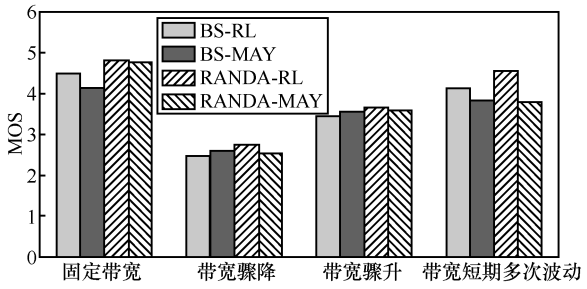



图 10　与其他算法比较平均意见得分

## 5　结束语

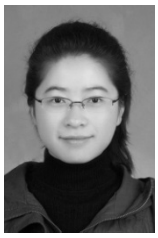
针对现有码率选择策略的局限性，本文提出一种基于 Q-learning 的码率决策方法。通过和现有环境进行交互，将感知到的实时可用带宽和缓存数据填充状态对环境状态和客户端行为建模。根据视频码率等级的改变和缓存数据填充量，建立可调整的回报函数。实验结果表明，引入 Q-learning 的码率自适应算法在带宽波动条件下维持视频码率稳定切换的方面具有显著优势。

## 参考文献：

[1] LIU C, BOUAZIZI I, GABBOUJ M. Rate adaptation for adaptive HTTP streaming[C]//The Second Annual ACM conference on Multimedia systems. 2011: 169-174.
[2] THANG T C, HO Q D, KANG J W, et al. Adaptive streaming of audiovisual content using MPEG DASH[J]. IEEE Transactions on Consumer Electronics, 2012, 58(1): 78-85.
[3] ROMERO L R. A dynamic adaptive http streaming video service for google android[J]. Communication Systems Cos, 2011, 33-51.
[4] LIU C, BOUAZIZI I, HANNUKSELA M M, et al. Rate adaptation for dynamic adaptive streaming over HTTP in content distribution network[J]. Signal Processing Image Communication, 2012, 27(4): 288-311.
[5] CLAEYS M, LATRÉ S, FAMAEY J, et al. Design of a Q-learning-based client quality selection algorithm for HTTP adaptive video streaming[C]//Adaptive and Learning Agents Workshop. 2013: 30-37.
[6] CLAEYS M, LATRÉ S, FAMAEY J, et al. Design and evaluation of a self-learning HTTP adaptive video streaming client[J]. IEEE Communications Letters, 2014, 18(4): 716-719.
[7] MART V, GARC N. Q-learning based control algorithm for HTTP adaptive streaming[C]//2015 Visual Communications and Image Processing (VCIP). 2015: 1-4.
[8] MOK R K P, CHAN E W W, CHANG R K C. Measuring the quality of experience of HTTP video streaming[C]//The 2011 IFIP/IEEE International Symposium on Integrated Network Management. 2011: 485-492.
[9] LI Z, ZHU X, GAHM J, et al. Probe and adapt: rate adaptation for http video streaming at scale[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(4): 719-733.
[10] HESSE S. Design of scheduling and rate-adaptation algorithms for adaptive HTTP streaming[C]//Applications of Digital Image Processing. 2013: 219-224.
[11] GARCIA S, CABRERA J, GARCIA N. Quality-control algorithm for adaptive streaming services over wireless channels[J]. IEEE Journal of Selected Topics in Signal Processing, 2015, 9(1):50-59.
[12] MARTIN V, CABRERA J, GARCIA N. Evaluation of Q-Learning approach for HTTP adaptive streaming[C]//IEEE International Conference on Consumer Electronics. 2016:293-294.
[13] ZHOU C, LIN C W, GUO Z. mDASH: a Markov decision-based rate adaptation approach for dynamic HTTP streaming[J]. IEEE Transactions on Multimedia, 2016, 18(4): 738-751.
[14] ISTEPANIAN R S H, PHILIP N Y, MARTINI M G. Medical QoS provision based on reinforcement learning in ultrasound streaming over 3.5 G wireless systems[J]. IEEE Journal on Selected Areas in Communications, 2009, 27(4): 566-574.
[15] CLAEYS M, LATRÉ S, FAMAEY J, et al. Design and optimisation of a (FA) Q-learning-based HTTP adaptive streaming client[J]. Connection Science, 2014, 26(1): 25-43.
[16] MENKOVSKI V, LIOTTA A. Intelligent control for adaptive video streaming[C]//International Conference on Consumer Electronics. 2013:127-128.
[17] CHIARIOTTI F. Reinforcement learning algorithms for DASH video streaming[D]. Italian: University of Padova. 2015.
[18] CHIARIOTTI F, D'ARONCO S L, et al. Online learning adaptation strategy for DASH clients[C]//International Conference on Multimedia Systems. 2016:8.
[19] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. MA: MIT Press, 1998.
[20] TOKIC M. Adaptive ε-greedy exploration in reinforcement learning based on value differences[C]//Advances in Artificial Intelligence, German Conference on Ai, Karlsruhe, Germany, September 21-24, 2010. Proceedings. DBLP, 2010:203-210.
[21] TOKIC M, PALM G. Value-difference based exploration: adaptive control between epsilon-greedy and softmax[C]//The 34th Annual German Conference on Advances in Artificial Intelligence. 2011: 335-346.
[22] MUELLER C, LEDERER S, POECHER J, et al. Demo paper: libdash-an open source software library for the MPEG-DASH standard[C]//The 2013 IEEE International Conference on Multimedia and Expo Workshops. 2013: 1-2.

## 作者简介：

**熊丽荣**（1973-），女，湖北崇阳人，浙江工业大学副教授，主要研究方向为服务计算和计算机网络。



**雷静之**（1992-），女，江西抚州人，浙江工业大学硕士生，主要研究方向为服务计算、计算机网络。

**金鑫**（1991-），男，浙江台州人，浙江工业大学硕士生，主要研究方向为服务计算。