基于 Q 学习的自主联合无线资源管理算法

张永靖 冯志勇 张 平
(北京邮电大学电信工程学院 北京 100876)

摘 要：该文提出了一种基于 Q 学习的联合无线资源管理(JRRM)算法，用于异构无线接入技术条件下 B3G 系统的自主资源优化。JRRM 控制器通过与无线环境的“试错”交互，学会为每个会话分配合适的接入技术和业务带宽。为降低存储需求，算法引入了反向传播神经网络用于泛化其输入状态空间。仿真结果表明，该算法不仅通过在线学习实现了 JRRM 的自主化，且在频谱效用和阻塞率之间获得了很好的性能折衷。

关键词：无线接入技术(RAT)；联合接纳控制；带宽分配；Q 学习；神经网络
中图分类号：TN915.65 **文献标识码**：A **文章编号**：1009-5896(2008)03-0676-05

A Q-learning Based Autonomic Joint Radio Resource Management Algorithm

Zhang Yong-jing Feng Zhi-yong Zhang Ping
(*Telecommunication Engineering School*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)

Abstract: A Q-learning based Joint Radio Resource Management (JRRM) algorithm is proposed for the autonomic resource optimization in a B3G system with heterogeneous Radio Access Technologies (RAT). Through the “trial-and-error” interactions with the radio environment, the JRRM controller learns to allocate the proper RAT and the service bandwidth for each session. A backpropagation neural network is adopted to generalize the large input state space to reduce memory requirement. Simulation results show that the proposed algorithm not only realizes the autonomy of JRRM through the online learning process, but also achieves well trade-off between the spectrum utility and the blocking probability.

Key words: Radio Access Technology (RAT); Joint admission control; Bandwidth allocation; Q-learning; Neural network

1 引言

多种异构无线接入技术(RAT)共存将成为未来 B3G 环境的一个重要的特征。重叠的网络覆盖、多样的业务需求以及互补的技术特性使得异构 RAT 间的协同和资源共享成为必须。为此，人们提出了很多联合无线资源管理(JRRM)的方法(如异构网络选择[1]、负载均衡[2]等)以获得更好的系统性能、频谱效率和用户体验。端到端重配置(E^2R)技术[3]的出现，为终端和相关网元设备提供了动态选择、配置 RAT 及工作频率的能力，使得对无线资源(接入权、时隙、码字、信道、功率等)的联合管理更加灵活和可行。相对而言，接入权以负载为表现形式，对异构 RAT 间的资源分配的作用更加直接和有效，因此成为很多 JRRM 算法[4−6]的研究对象。但是，现有的算法均未涉及 JRRM 的自主性问题，而这一点对于在 B3G 环境中所要面对的复杂系统来说尤为重要。考虑到业务需求在空间和时间上的动态变化及其不规则性，一个同时运营多个 RAT 的网络运营商将很难为其大量的基站和接入点配置最佳的 JRRM 策略。为实现网络对资源的自主管理以减少人力参与的规划和维护的成本，需要网络具有能根据实际运行情况不断修正其控制策略的自主学习能力。

强化学习(RL)[7]为我们提供了一种“试错”的在线学习技术。学习者通过与环境不断交互获得学习经验，能够逐步改进其行为策略。RL 以其灵活性和自适应性，广泛应用于机器人和自动控制领域[7]，并被引入无线蜂窝网络的动态信道分配问题中[8]。本文用 RL 中的 Q 学习[9]算法来实现异构 RAT 间的自主的联合接纳控制和带宽分配。算法设计中考虑了不同 RAT 的业务能力差异和负载的均衡性要求，以提高系统的频谱效用并降低呼叫阻塞率。针对实际系统参数的连续性特点，本文还采用神经网络的方法泛化状态空间以节约存储空间。

2 算法模型

2.1 Q 学习

基本的 RL 模型由有限、离散的环境状态的集合 $S=\{s_1, s_2,\cdots,s_n\}$，有限、离散的学习者动作的集合 $A=\{a_1,a_2,\cdots, a_m\}$，标量的强化信号 r 和学习者的策略 π. $S\rightarrow A$ 等基本要

2006-09-11 收到，2007-04-27 改回
欧盟 FP6 端到端重配置(IST-2005-027714)，国家自然科学基金重点项目(60502035)，国家 863 计划(2006AA01Z276)和科技部中欧盟科技合作项目(0516)资助课题

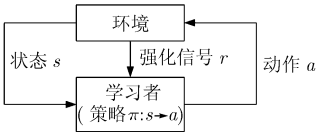

图 1 强化学习模型

素组成。它们的关系如图 1 所示：每一轮的迭代中，学习者感知环境状态 $s\in S$，并根据当前策略 π 选择动作 $a\in A$ 作用于环境；环境状态由此变化为 $s'\in S$，同时产生一个强化信号(称为“回报”)$r(s,a)$反馈给学习者；学习者据此更新其策略，并进入下一轮迭代。通过不断的“试错”，学习的最终目标是找到每个状态的最佳策略 $\pi^*(s)\in A$ 以最大化期望的长期累积回报(即状态的“值”)：

$$V^{\pi}(s)=E\left\{\sum_{t=0}^{\infty}\gamma^t r(s_t,\pi(s_t))\mid s_0=s\right\} \tag{1}$$

其中 $\gamma\in(0,1)$为常数时间折现因子，它体现了未来回报相对当前回报的重要性。根据 Bellman 最优准则，式(1)的最大值为

$$\begin{aligned}V^*(s)&=V^{\pi^*}(s)=\max_{\pi}V^{\pi}(s)\\&=\max_{a\in A}\left[R(s,a)+\gamma\sum_{s'\in S}P_{s,s'}(a)V^*(s')\right]\end{aligned} \tag{2}$$

其中 $R(s,a)$为 $r(s_t,a_t)$ 的数学期望，$P_{s,s'}(a)$ 为状态 s 在动作 a 的作用下达到状态 s'的转移概率。

Q 学习的好处是能够在未知 $R(s,a)$和 $P_{s,s'}(a)$ 的情况下，通过简单的 Q 值迭代找到最优的策略 π^* 满足式(2)。将策略 π 下的每一对状态和动作(s,a)与一个“Q 值”相关联：

$$Q^{\pi}(s,a)=R(x,a)+\gamma\sum_{s'\in S}P_{s,s'}(a)V^{\pi}(s') \tag{3}$$

由式(2)和式(3)可以得到

$$V^*(s)=\max_{a\in A}Q^*(s,a) \tag{4}$$

$$\pi^*(s)=\arg\max_{a}Q^*(s,a) \tag{5}$$

Q 学习通过以下迭代规则来获得 $Q^*(s,a)$:

$$Q_{t+1}(s,a)=(1-\alpha)Q_t(s,a)+\alpha(r_t+\gamma\max_{a'}Q_t(s',a')) \tag{6}$$

其中，$\alpha\in[0,1)$为学习率。随着 $t\to\infty$，若每对(s,a)的 Q 值能够经历无穷多次更新，且 α 递减至 0，则 $Q_t(s,a)$将以概率 1 收敛到最优值 $Q^*(s,a)$[9]。此时，最优策略 π^* 可以由式(5)得到。

作为一种离策略(off-policy)算法，Q 学习中式(6)的收敛并不依赖于动作空间的探索方法[7]。为了使所有(s,a)都能被充分地尝试同时兼顾效率，在迭代过程中本文采用 ε 贪婪算法来选择动作(详见第 5 节)。

2.2 神经网泛化

Q 值的迭代更新离不开对存储空间的需求，而实际问题中的状态或动作空间往往非常庞大甚至连续化，这使一些简单的存储方法(如查找表)难以实现。本文研究的 JRRM 问题中涉及小区负载(见第 4 节)这样非离散的状态参数，因此需要一种泛化方法[7]来处理。本文采用基于神经网函数近似的方法来记忆和表达 Q 值，并用最小化均方误差的反向传播(BP)算法[10]进行调整。

如图 2 所示，一个多层前馈神经网络(MFNN)被集成在学习者(JRRM 控制器)内部，其输入为构成环境状态的参数向量，输出为对应当前状态下对应所有动作的 Q 值向量。基于获得的 Q 值向量，动作选择模块采用 ε 贪婪算法决定所要采取的动作。根据动作执行后反馈强化信号，Q 值得到更新，而 MFNN 的权值则以更新的 Q 值为训练集用 BP 算法进行调整。为提高算法稳定性，更新的 Q 值先被缓存到训练队列里，再以批处理的方式送给 MFNN。

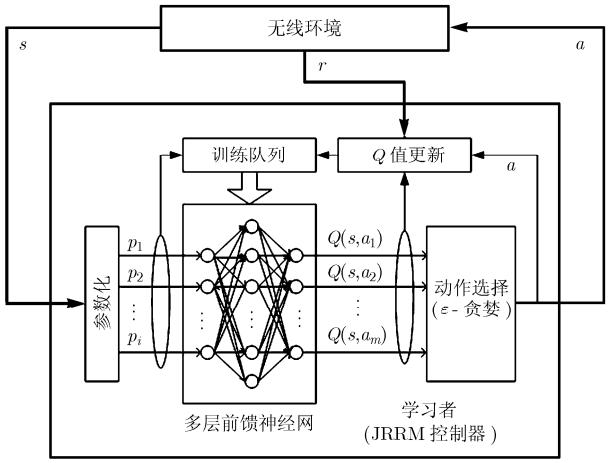

图 2　基于神经网泛化的 Q 学习模型

3　问题映射

设给定服务区内存在 K 个异构 RAT，它们属于同一运营商并由公共的 JRRM 控制器管理。各 RAT 的覆盖范围、业务能力、小区容量各不相同。在重叠覆盖的服务区内，终端可通过重配置工作在任意 RAT 下，并可能发起 V 种不同业务类型的会话请求。需要解决的问题是 JRRM 控制器如何自主地学习为请求的会话分配合适的 RAT 和业务带宽，以获得最佳的频谱效用同时降低呼叫阻塞率。本文的方法是将问题映射到上述 Q 学习模型中，并将 JRRM 控制器看作学习者，将网络条件和会话业务看作环境。相应地，定义如下几个要素：

(1) 状态　会话到达和结束的一系列离散时间事件构成了影响网络状态的主要因素，但由于会话的结束不会引起 JRRM 操作，所以仅将状态与会话的到达相关联以简化模型，并定义为

$$s=(c,v,\boldsymbol{L}) \tag{7}$$

其中，$c\in\{1,2,\cdots,K\}$表示发起会话的终端的网络覆盖条件(即可见 RAT 的数量)，$v\in\{1,2,\cdots,V\}$表示请求的业务类型；$\boldsymbol{L}$ 为所有 RAT 中各种业务的负载分布，可表示为

$$\boldsymbol{L}=\begin{bmatrix} l_{11} & \cdots & l_{1V} \\ \vdots & \ddots & \vdots \\ l_{K1} & \cdots & l_{KV} \end{bmatrix}$$

其中 l_{kv} 表示 RAT k 中业务 v 的负载量，一般情况下为连续值变量。

(2) 动作　JRRM 控制器对于到达会话的处理有两步：首先是分配 RAT 或者拒绝接入，然后是分配一定的业务带宽给接纳的会话。为了处理方便，将它们合为一步：

$$A=\{0,1,2,\cdots,m-1\} \tag{8}$$

其中，m=BK+1 表示可能动作的数量，而 B 为网络所支持的业务带宽等级数。式(8)中，动作 0 表示拒绝接入，而其它非零值表示不同的 RAT 和业务带宽组合的索引。

(3) 回报(强化信号)　即时回报是驱使 JRRM 控制器合理选择动作的直接信号。本文将合理性解释为所能获得的频谱效用(SU)：

$$u^{(k)}=\sum\nolimits_j b_j\eta_j(v,k) \tag{9}$$

其中 b_j 是为会话 j 所分配的业务带宽，$\eta_j(v,k)$为匹配系数，体现了 RAT k 对于会话 j 请求的业务类型 v 的适应度。频谱效用的定义体现了频谱资源利用的程度和效率，也可以理解为网络通过提供其频谱资源来服务用户所获得的收益。

单纯最大化网络的总频谱效用可能会带来资源利用的极度不平衡，特别是当适应某特定 RAT 的会话到达占据大部分业务量的时候。为了兼顾效率与公平，本文借鉴文献[11]中比例公平的概念，定义比例公平的频谱效用(PFSU)如下：

$$U_{pf}=\prod\nolimits_k u^{(k)} \tag{10}$$

此处，将原定义中的对数和运算替换为连乘，以避免某 RAT 的频谱效用接近 0 时 PFSU 趋于负无穷的问题。在此基础上，定义即时回报为状态 s 下动作 a 所获得的 PFSU 增量，即

$$r(s,a)=U_{pf}(s,a)-U_{pf}(s,0) \tag{11}$$

其中 $U_{pf}(s,0)$为动作 a 执行前的 PFSU(它在数值上也等于采取拒绝动作后的 PFSU)。

4　算法实现

基于图 2 中的算法模型以及上述问题映射，算法实现的详细工作流程如下：

(1) 初始化　开始时，随机设定 MFNN 的初始权值，设定式(6)中的折现因子 γ 和初始学习率 α_0，以及动作选择算法中的初始探索概率 ε_0，初始化训练队列(大小为 q)。

(2) 状态构建　当新会话到达时，JRRM 控制器需要搜集各 RAT 的负载信息以及到达会话的业务特征，根据式(7)构造出当前状态 s，并将结果转换为一维参数向量 $\boldsymbol{p}$=(p_1,p_2,$\cdots$,p_i) 送给 MFNN 的输入层，同时缓存在训练队列里。

(3) Q 值获得　向量 $\boldsymbol{p}$ 经过 MFNN 的运算，在输出层得到当前状态 s 下所有可能动作所对应的 Q 值 $\boldsymbol{Q}_t(s)$=($Q(s,a_1),Q(s,a_2),\cdots,Q(s,a_m)$)。该向量在被送往动作选择模块的同时被缓存起来以备更新。

(4) 动作选择和执行　根据输入的 $\boldsymbol{Q}_t(s)$，动作选择模块采用 ε 贪婪算法，从 A={0,1,$\cdots$,m–1}中选择一个动作 a，由 JRRM 控制器根据式(8)中的定义执行。具体地，算法将以概率(1–ε)选择 Q 值最大的动作，而以概率 ε 选择其它任一动作。被执行的动作会被记录，以备用于 Q 值更新。

(5) 获得回报　根据式(9)–式(11)，计算动作执行前后网络的 PFSU 的差，获得即时回报。

(6) Q 值更新　一旦新的会话到达，下一状态 s'及其所有的 Q 值 $\boldsymbol{Q}_t(s')$就能够由步骤(2)和(3)得到。结合记录的动作 a 以及相应的即时回报，缓存的 $\boldsymbol{Q}_t(s)$可由式(6)更新为 $\boldsymbol{Q}_{t+1}(s)$，并被送入训练队列。

(7) 参数更新　每轮迭代结束时，学习率 α 以及探索概率 ε 都需要更新。本文设置它们以负指数规律随着学习的过程逐渐减小为 0，以满足 Q 学习的收敛性要求。

(8) MFNN 更新　每次迭代的过程(步骤(2)–步骤(7))将产生一对 $\boldsymbol{p}$ 和 $\boldsymbol{Q}_t(s)$，分别作为输入向量和相应的目标输出向量被缓存到训练队列里。一旦队列被填满，所有缓存的向量将被一起送入 MFNN 进行 BP 运算以调整 MFNN 的权值，从而获得更加精确的 Q 值函数近似。这里采用的 BP 算法是带动量项的梯度下降法[10]。

5　仿真评估

考察一个由单小区重叠覆盖的 GERAN, UMTS 和 WLAN 构成的异构环境，如图 3 所示。各 RAT 具有不同的覆盖范围和小区容量，但均支持 16kbps 和 64kbps 的业务带宽。设该服务区内仅有话音和数据两种业务，且具有相同的到达率和最小带宽需求 16kbps。假定 GERAN 和 WLAN 分别适合话音和数据业务，而 UMTS 对二者同样适合。总的会话到达率服从均值为 λ 的泊松分布，会话时长服从均值为 120s 的负指数分布。WLAN 覆盖的中心区域 C 为热点地区，会话到达率较高为 0.8λ；而区域 A 和 B 较低，各为 0.1λ。MFNN 采用 3 层结构，其中输入层为 8 节点，隐藏层为 10 节点并采用双曲正切函数，输出层为 7 节点并采用线形函数。其它仿真参数见表 1。

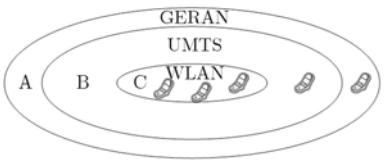

图 3 单小区重叠覆盖场景

仿真评估了基于 Q 学习的 JRRM 算法(QL)的阻塞率及频谱效用性能。作为参考，本文还仿真了基于负载均衡的算法(LB)和效用最大化的算法(UM)。前者以最小化阻塞率为目标，为会话分配负载最轻的 RAT 和最小的业务带宽；后者以最大化总频谱效用为目标，为会话分配匹配系数最大的 RAT 和最大的业务带宽。

表 1 仿真参数

		GERAN (4carriers)	UMTS(FDD) (single carrier)	WLAN
小区容量 (kbps)		640	800	2,000
$\eta(v,k)$	话音	5	3	1
	数据	1	3	5
到达率 (calls/h)		$\lambda = 150, 300, 450, 600, 750, 900$		
$\gamma = 0.5$		$\alpha_0 = 0.5$	$q = 30$	$\varepsilon_0 = 0.5$
$V = 2$		$B = 2$	$K = 3$	$m = 7$
MFNN=(8,10,7)			迭代次数: 20,000	

图 4 和图 5 分别给出了不同会话到达率情况下各算法的阻塞率和总频谱效用。图 6–图 8 则以 λ =600 呼叫/小时为例，进一步分析了各 RAT 中的带宽、负载和频谱效用分布。

如图 4 所示，LB 算法拥有最低的呼叫阻塞率，因为它总是分配最小业务带宽且均衡网络负载(见图 6，7)，因此各 RAT 均不易饱和。然而，最小带宽分配导致了资源利用不充分，且由于 LB 没有考虑业务和 RAT 的适配性，导致资源利用效率不高(图 7 中话音和数据业务在各 RAT 中都均匀分布)，因此总频谱效用最低(见图 5)。相反地，UM 算法用最差的阻塞率性能换来了最高的总频谱效用(见图 4，图 5)。原因与 LB 相反：最大带宽分配(见图 6)和最佳的业务/RAT 匹配(图 7 中话音和数据业务分别占 GERAN 和 WLAN 负载的绝大部分)带来了最充分和有效的资源利用，却也使得网络的剩余资源变得非常有限(如图 7 中的 GERAN)，因此导致新会话无法被接入(特别是在仅被 GERAN 覆盖的边缘地区)。

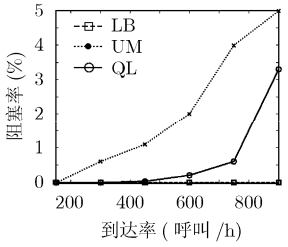

图 4 不同到达率下的阻塞性能

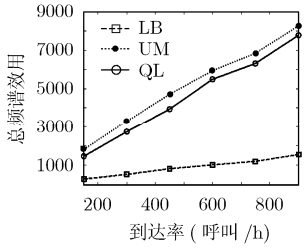

图 5 不同到达率下的总频谱效用

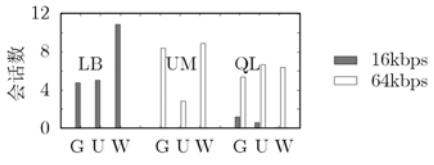

图 6　不同 RAT 中的业务带宽分布

相比而言，QL 算法在网络可用性和资源利用率方面获得了很好的折衷。原因有两方面：首先，基于 PFSU 的即时回报本身既体现了 RAT 的业务适配性，又照顾到各 RAT 间资源利用的公平性。如图 7，GERAN 和 WLAN 的业务分布符合其技术特性,而同时 UMTS 的资源也得到了充分利用，从而缓解了 GERAN 的负载压力。这一点在频谱效用的分布中体现得更加明显(见图 8)。因此，QL 既能够达到很高的频谱效用，又能够在很大程度上避免由于个别 RAT 过载而导致的呼叫阻塞。其次，对于长期累积回报的追求避免了对当前网络资源的过度利用。越是高负载的 RAT 被分配小带宽的会话越多(见图 6)，这样在获得较高资源利用率的同时也保证了网络的可用性。

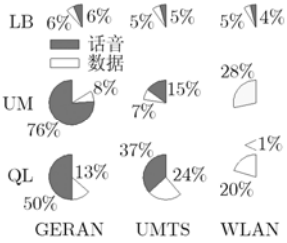

图 7 不同 RAT 中各类业务的负载分布图

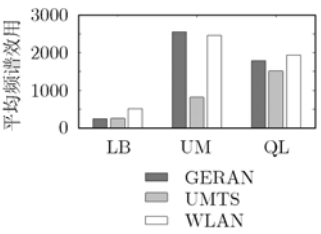

图 8 各 RAT 中的频谱效率分布

最后，图 10 给出了 QL 算法在图 9 所示的动态业务量下的累积阻塞率和累积平均总频谱效用。容易看出，在连续两天同样的会话到达率情况下，第二天的阻塞率和总频谱效用性能均明显好于第一天，这说明 QL 算法通过在线学习，能够有效地将已有经验用于后续的策略选择中，从而获得性能的提升。

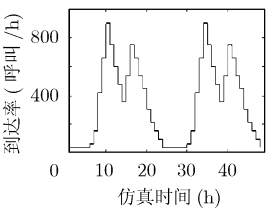

图 9 动态业务到达率

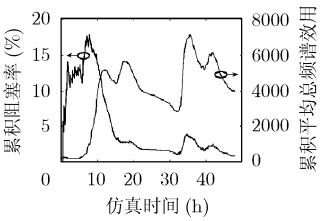

图 10 Q 学习算法的在线性能

6　结束语

本文提出了一种基于 Q 学习的 JRRM 算法，以解决未来异构 RAT 环境下自主的联合接纳控制和带宽分配问题。同时，采用神经网泛化的方法解决了实际情况中的连续状态空间问题，节省了存储开销。通过基于 PFSU 的强化信号，JRRM 控制器能以“试错”迭代的在线学习方式找到优化的 JRRM 策略，在阻塞率和频谱效用之间获得了很好的折衷。

参 考 文 献

[1] Song Q and Jamalipour A. Network selection in an integrated wireless LAN and UMTS environment using mathematical

modeling and computing techniques[J]. *IEEE Wireless Commun.*, 2005, 12(3): 42–48.

[2] 3GPP TR 25.881 v5.0.0. Improvement of RRM across RNS and RNS/BSS (Release 5) [OL]. http://www.3gpp.org, Dec. 2001.

[3] IST-2003-507995 Project E²R (End-to-End Reconfigurability) [OL]. http://e2r.motlabs.com, Jan. 2004.

[4] Agusti R, Sallent O, and Perez-Romero J, *et al*.. A fuzzy-neural based approach for joint radio resource management in a beyond 3G framework[C]. First Int. Conf. on Quality of Service in Heterogeneous Wired/Wireless Networks, Barcelona, Mar. 2004: 216–224.

[5] Luo J, Mohyeldin E, and Dillinger M, *et al*.. Performance analysis of joint radio resource management for reconfigurable terminals with multi-class circuit-switched services[C]. Wireless World Research Forum 12th Meeting, Toronto, Nov. 2004: 138–150.

[6] Zhang Y, Zhang K, and Ji Y, *et al*.. Adaptive threshold joint load control in an end-to-end reconfigurable system[C]. IST Mobile and Wireless Summit 2006, Mykonos, Jun. 2006: 332–337.

[7] Kaelbling L P, Littman M L, and Wang X, *et al*.. Reinforcement learning: a survey[J]. *Journal of Artificial Intelligence Research*, 1996, 4(2): 237–285.

[8] Nie J and Haykin S. A Q-learning-based dynamic channel assignment technique for mobile communication systems[J]. *IEEE Trans. on Vehicular Technology*, 1999, 48(5): 1676–1687.

[9] Watkins C J C H and Dayan P. Q-learning[J]. *Machine Learning*, 1992, 8(3): 279–292.

[10] 张乃尧，阎平凡. 神经网络与模糊控制. 第一版，北京：清华大学出版社，1998 年: 12–18.
Zhang X N and Yan P F. Nueral Networks and Fuzzy Control, 1st ed., Beijing: Tsinghua University Press, 1998: 12–18

[11] Radunovic B, Le Boudec J Y. Rate performance objectives of multihop wireless networks[J]. *IEEE Trans. on Mobile Computing*, 2004, 3(4): 334–349.

张永靖： 男，1981 年生，博士，研究方向为端到端可重配置系统中的无线资源管理.

冯志勇： 女，1971 年生，副教授，主要研究方向为异构无线网络.

张 平： 男，1959 年生，博士生导师，主要研究方向为移动通信.