Reinforcement Learning Based Network Selection for Hybrid VLC and RF Systems

Chunxi Wang[1,*], *Guofeng* Wu[1], *Zhiyong* Du[2], and *Bin* jiang[2]

[1]National Digital Switching System Engineering and Technological Research Center, 450001 Zhengzhou, China
[2]National University of Defense Technology, 91944 Changsha, China

Abstract. For hybrid indoor network scenario with LTE, WLAN and Visible Light Communication (VLC), selecting network intelligently based on user service requirement is essential for ensuring high user quality of experience. In order to tackle the challenge due to dynamic environment and complicated service requirement, we propose a reinforcement learning solution for indoor network selection. In particular, a transfer learning based network selection algorithm, i.e., reinforcement learning with knowledge transfer, is proposed by revealing and exploiting the context information about the features of traffic, networks and network load distribution. The simulations show that the proposed algorithm has an efficient online learning ability and could achieve much better performance with faster convergence speed than the traditional reinforcement learning algorithm.

1 Introduction

Visible light communication (VLC), as an emerging wireless access technology has been regarded as a promising member in the 5G era that possesses tremendous value and potential [1]. It exhibits multi-fold advantages such as high data rate, huge bandwidth, no electromagnetic interference and high security. In a heterogeneous wireless access environment for indoor communication, where LTE, WLAN and VLC are available, simultaneously. Selecting network intelligently based on user service requirement is essential for ensuring high user Quality of Experience (QoE). However, different types of wireless technologies show varieties in the aspect of coverage, data transmission rate and other features. Meanwhile, end users are no longer satisfied with the basic data communication and emerging virtual reality, ultra-high definition video etc., pose higher requirements on both uplink and downlink performances. Due to these two concerns, selecting the optimal access network is always a challenging task.

In this paper, a reinforcement learning solution is proposed for indoor network selection. Specifically, context information is leveraged to tackle the network selection on two aspects. On one hand, the feature of asymmetric downlink and uplink performance requirements of traffic are explicitly revealed and modeled. On the other hand, some distinguishing features of network as well as the stationary distribution law of network load are used to assist the algorithm design. In particular, such information enables us to present knowledge transfer for reinforcement learning, providing an effective algorithm for the network selection in dynamic and unknown environment.

Our main contributions are two-folds. First, we proposed a fine-grained network selection model that takes the diverse traffic requirements and network performance of uplink and downlink into account. This vision is important since many newly emerging traffic types such as virtual reality need customized performance requirements. Although there is extensive research on network selection, e.g.,[2][3], the utility design differentiating uplink and downlink requirements of different traffic types proposed in this paper seems to be absent. Second, the idea of transfer learning [4] based algorithm is used in network selection. Even though some works such as [5] have studied the context-aware network selection, they worked in different ways from the knowledge transfer. Compared with some existing work using reinforcement learning [6][7], the introduction of transfer learning could significantly enhance the algorithm performance. This method may provide a new perspective on endowing context awareness in solutions for related problems [8].

2 System Model

We consider an indoor heterogeneous wireless access environment which consists of N networks of $\mathcal{N}=\{1,2,\cdots,N\}$ LTE, WLAN and VLC. For simplicity, we use the term "network" to represent a base station (BS) in LTE or an access point (AP) in WLAN and VLC. We assume that a user locates in the overlapping area of N wireless networks and equipped with multi-homing capability. In a slotted system with epoch duration l

* Corresponding author: firefiy211@126.com

seconds, the user can dynamically change its access network but only one access network can be selected at any given slot.

We use throughput as the main performance metric of the networks. The max instantaneous rate of a user that is determined by SNR (signal to noise ratio) according to Shannon formula constitutes the upper bound of its throughput. Meanwhile, the multi-user access behavior determines the network load distribution and thus affects the achieved throughput of each user in the network. Therefore, the achieved instant throughput $\Theta(i,n)$ of user i in network is a function of the instantaneous rate and the network load K_n (the total number of users in network n) as $\Theta(i,n) = f(R, K_n)$ for a given slot. The function $f()$ could be modeled depending on specific network. In the following, the uplink and downlink throughput models of LTE, WLAN, and VLC are given.

1. LTE: OFDMA is the downlink multiple access technology of LTE. According to the model in [2], the throughput under weighted-proportional fairness can be expressed as

$$\Theta_{\mathrm{DL}}(i,n) = \frac{\omega_i R_{n\to i}}{W_k} \tag{1}$$

Where ω_i is user i 's weight, $W_k = \sum_{i\in\mathcal{K}_n} \omega_i$ is the total users' weight, $\mathcal{K}_n$ is the set of users in network n, i.e., $K_n = |\mathcal{K}_n|$, $R_{n\to i}$ is the instantaneous downlink rate of user i.

In the uplink, LTE uses the SC-OFDMA based MAC protocols with fair subcarrier sharing. Hence, the throughput of a user i is roughly dependent on the total number of users sharing the same network,

$$\Theta_{\mathrm{UL}}(i,n) = \frac{R_{n\leftarrow i}}{K_n} \tag{2}$$

2. WLAN: In 802.11 WLAN MAC protocols, the distributed coordination function (DCF) leads to a fair access opportunity to uplink users. Hence, the low rate user capturing the channel will use it for a long time thus penalizes high rate users. The uplink throughput of a WiFi user can be expressed as

$$\Theta_{\mathrm{UL}}(i,n) = \frac{L}{\sum_{j\in\mathcal{K}_n} \frac{L}{R_{n\leftarrow i}}} \tag{3}$$

Here, L is the packet size. The throughput a user can obtain on the downlink is related to the schedule mechanism of the access point. According to [10], when a round-robin (RR) scheme is used, then the uplink can also be derived via replacing $R_{n\leftarrow i}$ by $R_{n\to i}$ in formula(3).

3. VLC: We consider an all-optical VLC network. Downstream data transmission and illumination are combined. Currently, there is no common view on the MAC protocol specified for VLC. In most existing works, it is assumed that the system uses TDMA with RR scheduling. Thus, if user i is assigned to the n-th VLC AP, the achieved throughput becomes [11]

$$\Theta_{\mathrm{DL}}(i,n) = \frac{R_{n\to i}}{2\cdot K_n} \tag{4}$$

Note that the intensity modulation with direct detection (IM/DD) is used in VLC and only real-valued signals can be transmitted to receivers. Thus, at least half of the sub-carriers must be used to realize the Hermitian conjugate of the complex-valued symbol after modulation. Consequently, the formula is divided by 2.

Using visible light in uplink may not be practical, as it would constrain equipment power and user's psychological feelings. Referring to [12], we use infrared in uplink. The main limitation of infrared link is determined by its low power transmission, thus it often leads to a low rate data transmission (up to 4Mbps or 1.152Mbps in [9]). As visible light and IR light exhibit very similar qualitative behavior, the uplink throughput model could also be derived by replacing $R_{n\to i}$ by $R_{n\leftarrow i}$ in formula (4).

3 Reinforcement learning based network selection framework

3.1 Problem formulation

Considering the diverse features of various traffic types, we propose a general utility model differentiating uplink and downlink performance requirements, which is still absent in current network selection research to our knowledge. Note that we mainly focus on the throughput, but this model can be easily extended to incorporate many other performance metrics. The achieved utility $u(\Theta_{\mathrm{UL}}, \Theta_{\mathrm{DL}})$ is designed from a novel perspective.

1. Uplink dominant traffic: For traffics such as sending files or backing up files on the cloud, the uplink throughput is the main factor affecting the performance, but the downlink throughput is negligible since it is just for transmitting some control and feedback messages (no less than a small threshold, e.g., Θ_0). As an example, it can be defined using a similar utility representing file transfer.

$$u(\Theta_{\mathrm{UL}}, \Theta_{\mathrm{DL}}) = \mathrm{I}\{\Theta_{\mathrm{DL}} \ge \Theta_0\}\lambda \log(\beta\cdot\Theta_{\mathrm{UL}}) \tag{5}$$

Where $\mathrm{I}(x) = 1$ when $x = 1$; otherwise, $\mathrm{I}\{x\} = 0$. $\mathrm{I}\{\Theta_{\mathrm{DL}} \ge \Theta_0\}$ is the minimal downlink throughput requirement and $\lambda\log(\beta\cdot\Theta_{\mathrm{UL}})$ models the utility-throughput function [2], where λ and β are parameters dependent on special maximal and minimal throughput demand of the user.

2. Downlink dominant traffic: On the contrary, downloading files and watching online video mainly care the downlink throughput and can be classified as downlink dominant traffic. Since most of existing works focus on such traffic type, the utility $u(\Theta_{\mathrm{DL}})$ can be easily derived by explicitly indicating the downlink throughput Θ_{DL} in existing utility models. For instances, the file download utility can using the above model by replacing Θ_{UL} with Θ_{DL}. Video traffic shows threshold

effect on throughput, then a piecewise function of the downlink throughput plus the basic uplink throughput requirement is

$$u(\Theta_{UL,}\Theta_{\mathrm{DL}})=\begin{cases}0 & \Theta_{\mathrm{DL}}\le\Theta_1\\ \dfrac{c(\Theta_{\mathrm{DL}}-\Theta_1)}{\Theta_2\text{-}\Theta_1}\mathrm{I}\{\Theta_{DL}\ge\Theta_0\} & \Theta_1<\Theta_{\mathrm{DL}}<\Theta_2\\ cI\{\Theta_{\mathrm{UL}}\ge\Theta_0\} & \Theta_{\mathrm{DL}}\ge\Theta_2\end{cases} \quad (6)$$

Where c is a constant.

3. Uplink-downlink symmetric traffic: For video call and video conference traffics, they have high requirements on both the downlink and uplink throughput. Either uplink or downlink throughput can be the bottleneck. We can replace Θ_{DL} by $\Theta_{\min}=\min(\Theta_{\mathrm{UL}},\Theta_{\mathrm{DL}})$ in formula (6) to get a utility function.

Due to channel fading and shadowing effect, the instantaneous rates $R_{n\leftarrow i}(t)$ and $R_{n\rightarrow i}(t)$ are time-varying. Moreover, the network load K_n is a random variable since the active users in a network is dynamic. Consequently, the achieved throughput $\Theta(i,n)$ and the resulting $u(\Theta_{\mathrm{UL}},\Theta_{\mathrm{DL}})$ are dynamic and random variables. Hence, it is reasonable to select the network providing the best average performance. However, since we have no prior knowledge on the average performance of the available networks, we have to learn the optimal selection from the interaction with the environment. Mathematically, this learning problem can be formed to select a network selection policy π^* maximizing the long term average reward, that is, a series of actions $\{a(1),a(2),\cdots\}$ that can maximize the total expected return as

$$V^*=\max E\{\sum_{t=0}^{\infty}\gamma^t u[\Theta_{\mathrm{UL}}(\mathrm{t}),\Theta_{\mathrm{DL}}(\mathrm{t})]\} \quad (7)$$

Where $\gamma\in(0,1)$ represents the discount factor, which reflects the future returns relative to the current level of importance. $u\left[\Theta_{\mathrm{UL}}(t),\Theta_{\mathrm{DL}}(t)\right]$ is the instant reward received at time t, $\Theta_{\mathrm{UL}}(t)$ and $\Theta_{\mathrm{UL}}(t)$ are the instant uplink and downlink throughput, respectively.

3.2 Algorithm design

Q learning is the most commonly used reinforcement learning algorithm for above problem in femto and small cells net [13]. In Q learning algorithm, the controller (learner) to learn how to optimize its decision through historical experience. However, the standard Q learning algorithm may show slow convergence speed and poor performance due to the exploration. Especially, when the available strategy set is relatively large, there will be significant random exploration costs on bad strategies. Nevertheless, the idea of transfer learning [4] provides a feasible way to enhance the Q learning algorithm. More specifically, the transfer learning enables us to speed up the algorithm convergence by using some knowledge or context information. Fortunately, we notice that the following observations may be useful

Observation 1: Not all networks are inherently suitable for all traffic types. There could be mismatch between the downlink/uplink features of networks and traffic requirements. For instance, VLC itself has poor uplink throughput due to the inherent limitation as we have mentioned, thus, it is not suitable for the traffic with strict requirement on uplink performance. Some other prior rules could also be applied, such as fee and privacy considerations.

Observation 2: Network load distribution is space-time dependent. The recent literature [14] has revealed that the traffic/load shows spatial and temporary distribution law, which means that the information about the load dynamics of networks may be used. For example, the load dynamics of a specific location and a fixed duration of weekdays are generally the same. With these observations, we propose the Q learning algorithm with knowledge transfer as shown in algorithm 1.

Algorithm 1 Q learning with knowledge transfer

1: **if** $(s,\mathcal{N}^*,i)$ has learning record in the database **then**
2: Initialize Q table with previous learn value $\boldsymbol{Q}=\boldsymbol{Q}_{(s,\mathcal{N}^*,i)}$ and set $\varepsilon=\varepsilon'$.
3: **else**
4: Initialize Q table with $\boldsymbol{Q}=\boldsymbol{0}$ and set $\varepsilon=\varepsilon''$.
5: **end if**
6: **loop**
7: For each slot t,based on the traffic type, select network $a(t)$ from the refined action set $\mathcal{N}_s\subseteq\mathcal{N}^*$ as follows
8: ● With probability ε, choose an action at random;
9: ● Else, choose $a(t)=\max_{n\in\mathcal{N}_s}Q(n)$.
10: Receive the reward $u(t)$.

$$Q\left[a(t)\right]=(1-\alpha)Q\left[a(t)\right]+\alpha\left[u(t)+\gamma\max_n Q(n)\right]$$

11: Update Parameters: In each iteration, the learning rate and the exploration probability need to be gradually reduced in order to meet the convergence requirements.
12: Update $(s,\mathcal{N}^*,i)$.
13: **if** $(s,\mathcal{N}^*,i)$ has changed **then**
14: Go to 1.
15: **end if**
16: **end loop**

To this end, we introduce a vector $(s,\mathcal{N}^*,i)$ to represent the traffic type-location-time context information, where $s\in\mathcal{S}$, $\mathcal{N}^*\subseteq\mathcal{N}$ and $i\in\mathcal{I}$ are the current traffic type, available network set and time period index, respectively. $\mathcal{S}$ is the set of traffic types, e.g., the three types defined in Section 3, and $\mathcal{N}$ is the maximal available network set as introduced in Section 2. Note that since the available networks may changes in different location, we use the set of available network set to indicate the "location" instead of exact coordinates.

One day is divided into several time periods. For example, the daytime of weekdays from 8:00 pm to 17:00 am could be divided into 9 periods each corresponding to 1 hour duration. The load distribution law is assumed to stay unchanged in each time period.

Specifically, observation 1 enables us to decrease the size of action set according to the traffic types. That is, some network choices are removed in the Q learning action set considering they are not suitable. This is realized by selecting the traffic type-dependent action set, i.e., the refined action set $\mathcal{N}_s \subseteq \mathcal{N}^*$ and Q vector $\mathbf{Q}=\left[Q(1),Q(2),...,Q\left(|\mathcal{N}_s|\right)\right]$ as shown in the 7th line of the algorithm. Observation 2 actually indicates that the load distribution laws of the same time period across different weekdays are approximately the same, thus, we can reuse the learned experience in the past. In the algorithm, the context-specific learning experience in terms of Q tables $\mathbf{Q}_{(s,\mathcal{N}^*,i)}$ are stored in a data base. Once it is found that there is already some learning record for the current $(s,\mathcal{N}^*,i)$, the learn Q table will be used; otherwise, the Q table is initiated with 0 vector, as shown in the 1st to 5th line of the algorithm. Accordingly, the initial exploration probability $\varepsilon' < \varepsilon''$.

4 Performance Evaluatiion

We consider an indoor scenario composed of LTE femtocells, WLAN and VLC with single cell overlap. In LTE, WLAN and VLC standards, the user achieved instantaneous rate is discrete, which is determined by the user's location and varies with the fading effect over time. Similar to the lecture [15], we make a set of discrete achievable peak rates $R_{1,k} < R_{2,k} < \cdots < R_{M_k,k}$ referencing to some measured data by the "Speedtest" app. Specifically, the dynamic ranges of uplink date rates of the LTE, WLAN and VLC are [4000kbps, 7000kbps], [3000kbps, 10000kbps] and [8000kbps, 13000kbps], and their dynamic ranges of downlink date rates are [500kbps, 6000kbps], [3000kbps, 9000kbps], and [80kbps, 120kbps], respectively. M_k is the network k maximum number of achievable rates. Furthermore, we set 6 as the maximum number of users accessing to the same network and adjust the number of access users at the beginning of each epoch, which further characterizes the dynamics of the network load. Some other parameters are listed in the Tab. 1. The simulation listed below is based on the Monte-Carlo simulation method averaged by 500 times.

Table 1. Parameter set

Parameter	Value	Parameter	Value
l	30s	c	10
λ	2	α	0.3
β	3	γ	0.3
Θ_1	200kbps	ε'	0.4
Θ_2	5000kbps	ε''	0.6

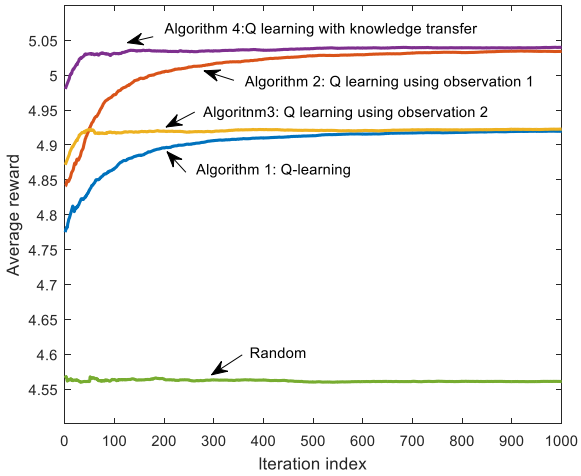

Fig. 1. Convergence comparison of different algorithms.

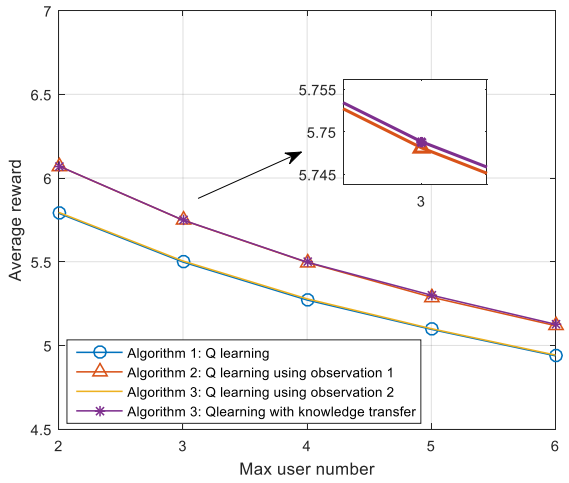

Fig. 2. Performance comparison with different user numbers.

We use the uplink dominant traffic to examine the convergence performance of the proposed algorithm. Since observation 1 has revealed that the uplink of VLC could hardly support the high uplink performance requirement, we can remove VLC to reduce the action space. This reduced action set combined with the standard Q learning is denoted by algorithm 2. The reusing of learning experience that revealed by observation 2 combined with the standard Q learning is denoted by algorithm 3. Finally, the proposed Q learning with knowledge transfer algorithm (i.e., Q learning using observations 1 and 2) in algorithm table is denoted by algorithm 4. In addition, the random access in each slot and the standard Q learning are added for comparison. As we can see in Fig.1, the random selection obtains the lowest and constant average reward. The other four Q learning based algorithms could converge after a certain number of iterations. We can observe that: i) algorithm 3 converges much faster than algorithm 1(the standard Q learning); ii) algorithm 2 achieves a significant gain in the average reward compared with algorithm 1(the standard Q learning) and iii) the proposed algorithm performs the best in terms of both the convergence speed and average reward. These results indicate that the considerations of observation 1 and observation 2 could improve the algorithm convergence speed and the

achieved performance, respectively. Note that the fast convergence speed (less than 50 iterations) of the proposed algorithm is important for practical applications. The performance comparison with different maximal user number of each network in Fig. 2 further shows that the proposed algorithm is the best. Moreover, its performance gain grows as the maximal user number increases.

5 Conclusion

In this letter, we studied the indoor network selection problem under dynamic environment, taking into account both the uplink and downlink performance requirements of traffics. We first formulated the network selection differentiating the network performance and traffic requirements of uplink and downlink as a learning problem. On this basis, we exploited the context information by resorting to transfer learning to propose a reinforcement learning with knowledge transfer based algorithm. The simulation results revealed that the introduction of transfer learning could significantly improve both the convergence speed and achieved performance of reinforcement learning based network algorithm.

Acknowledgments

The work is supported by the NSF of China No. 61671477 and No. 61601490..

References

1. Elgala, H., Elgala, H., Jungnickel, V., Jungnickel, V., Little, T., & Shao, S., et al. 2016. Coexistence of wifi and lifi toward 5g: concepts, opportunities, and challenges. *IEEE Communications Magazine*, 54(2), 64-71.
2. Du, Z., Wu, Q., Yang, P., Xu, Y., Wang, J., & Yao, Y. D. 2015. Exploiting user demand diversity in heterogeneous wireless networks. *IEEE Transactions on Wireless Communications*, 14(8), 4142-4155.
3. Du, Z., Wu, Q., Yang, P., & Xu, Y. 2014. User-demand-aware wireless network selection: a localized cooperation approach. *IEEE Transactions on Vehicular Technology*, 63(9), 4492-4507.
4. Olivas, E. S., Guerrero, J. D. M., Sober, M. M., Benedito, J. R. M., & Lopez, A. J. S. 2009. *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques*. Information Science Reference - Imprint of: IGI Publishing.
5. Boran, M., Gönen, F., & Cetin, S. 2013. Matching with Externalities for Context-Aware User-Cell Association in Small Cell Networks. *IEEE Global Communications Conference* (Vol.113, pp.4483-4488). IEEE.
6. Du, Z., Wu, Q., & Yang, P. 2014. Dynamic user demand driven online network selection. *IEEE Communications Letters*, 18(3), 419-422.
7. Wu, Q., Du, Z., Yang, P., Yao, Y. D., & Wang, J. 2016. Traffic-aware online network selection in heterogeneous wireless networks. *IEEE Transactions on Vehicular Technology*, 65(1), 381-397.
8. Xu, Y., Wang, J., Wu, Q., & Du, Z. 2015. A game-theoretic perspective on self-organizing optimization for cognitive small cells. *IEEE Communications Magazine*, 53(7), 100-108.
9. IrDA Standards. http://irda.org.
10. Bianchi, G. 2000. Performance analysis of the ieee 802.11 distributed coordination function. *IEEE J. Sel. Areas Commun*, vol. 18(3), 535-547.
11. Basnayaka, D. A., & Haas, H. 2015. Hybrid RF and VLC Systems: Improving User Data Rate Performance of VLC Systems. *IEEE, Vehicular Technology Conference* (pp.1-5). IEEE.
12. Kavehrad, M. 2010. Sustainable energy-efficient wireless applications using light. *Communications Magazine IEEE*, 48(12), 66-73.
13. Kaelbling, L. P., Littman, M. L., & Moore, A. W. 1996. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 4(1), 237--285.
14. Lee, D., Zhou, S., Zhong, X., Niu, Z., Zhou, X., & Zhang, H. 2014. Spatial modeling of the traffic density in cellular networks. *IEEE Wireless Communications*, 21(1), 80-88.
15. Ibrahim, M., Khawam, K., & Tohme, S. 2010. Congestion Games for Distributed Radio Access Selection in Broadband Networks. *Global Telecommunications Conference* (Vol.45, pp.1-5). *IEEE*.