Information Distribution System with Distributed Reinforcement Learning for Providing Local Information

Shin'ichiro Murata and Takuji Tachibana

***Abstract*—In this paper, we propose an information distribution system with reinforcement learning so that the information can be distributed to users with a small number of distributions. In this system, the information distributor moves along the same route and distributes the information to users by using the wireless communication technology. Here, we consider a case where the distributor would like to send the information such as advertisement to users who pass a specified area related to the information. As long as the information is distributed at random, however, it is hard to distribute the information to users that pass the specified area effectively. To achieve this goal, in our proposed system, distributed reinforcement learning is utilized. With the distributed reinforcement learning, the distributor learns the optimal positions where the information can be distributed to a larger number of users who pass the specified area. Moreover, by considering road conditions, the information can be distributed to users effectively. We evaluate the performance of our proposed system with simulation. In numerical examples, we investigate the effectiveness of our proposed system by comparing with the conventional method where the information is distributed at random.**

***Index Terms*—Distributed reinforcement learning, Information distribution, Distributed Q-learning, Wireless communication, Optimal action**

I. Introduction

Recently, several kinds of wireless communication systems have been proposed and developed. With such wireless communication systems, the information can be distributed to users easily. [1] has proposed a reliable message transmission in mobile social networks. By using this method, the information can be transmitted to users reliably without the security key management for the mobile social networks. In [2], authors has implemented and experimented publish/subscribe systems as a word-mouth information distribution service for users such as resident, shopper and traveler in Kobe, Japan. The effectiveness of the proposed system for the information distribution has been proposed with simulation. Moreover, [3] has considered an information collection system, which is called ZebraNet system, to support wildlife tracking for biology research. This system utilizes wireless peer-to-peer networking techniques in a mobile sensor network. In this system, custom tracking collars (nodes) are carried by animals across a large and wild area and sensing information is collected via such nodes.

Here, the distribution information system can be used as an advertisement delivery platform. In this case, information distribution charges or advertising expenditures may be required to distribute the information to users [4]. Therefore, it is expected that the information can be efficiently distributed to a larger number of users with a smaller number of distributions. When users receive the information with their own mobile devices, in addition, the number of information distributions should be limited by reducing the power consumption.

In this paper, we propose an information distribution system with reinforcement learning so that the information can be distributed to users with a small number of distributions. In this system, the information distributor moves along the same route and distributes the information to users by using the wireless communication technology such as IEEE 802.11, Bluetooth, and so on. Here, we consider a case where the distributor would like to send the information such as advertisement to users who pass a specified area related to the information. As long as the information is distributed at random [5], however, it is hard to distribute the information to users that pass the specified area effectively.

To achieve this goal, in our proposed system, distributed reinforcement learning is utilized. With the distributed reinforcement learning, the distributor learns the optimal positions where the information can be distributed to a larger number of users who pass the specified area. Moreover, by considering road conditions, the information can be distributed to users effectively. We evaluate the performance of our proposed system with simulation. In numerical examples, we investigate the effectiveness of our proposed system by comparing with the conventional method where the information is distributed at random.

The organization of this paper is as follows. Section II explains reinforcement learning and distributed Q-learning, and then we introduce QLAODV and RLAB. In Sect. III, we explain our proposed system. Section IV denotes numerical results, and finally, conclusions are denoted in Sect. V.

II. Related Work

A. Reinforcement Learning

In reinforcement learning, a learning agent moves from state to state by performing an action [6]. The agent receives a reward at every state after performing an action, and hence the agent tries to perform an action so as to gain more rewards. From these experiences, the agent can learn the optimal action for each state.

Q-learning is one of the reinforcement learning techniques, and a value function Q is utilized in order to decide an optimal action for each state of an environment. Here,

S. Murata is Graduate School of Engineering, University of Fukui, Fukui 910–8507, Japan. Email: shinichiro-m@network.fuis.u-fukui.ac.jp

T. Tachibana is Graduate School of Engineering, University of Fukui, Fukui 910–8507, Japan. Email: takuji-t@u-fukui.ac.jp

ISBN: 978-988-19253-3-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$Q(s,a)$ means a value of action a when a state is s, and action a^* is the optimal action if $a^* = \underset{a}{\operatorname{argmax}} Q(s,a)$ is satisfied. Now, let s_t denote a state of an environment at time t and let a_t denote an action at time t. Moreover, let r_t denote a reward that is gained at time t. When a learning agent moves from s_t to s_{t+1} by performing action a_t, a value function $Q(s_t, a_t)$ is changed into

$$Q(s_t,a_t) = Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right], \quad (1)$$

where $0 < \alpha \leq 1$ is a learning rate and $0 \leq \gamma < 1$ is a discount factor. The learning rate α determines to what extent the latest information will override the old information. The agent does not learn anything when α is equal to zero, while the agent considers only the most new information when α is equal to one. The discount factor γ determines the influence of future rewards. The agent considers only current reward when γ is equal to zero, while the agent tries to find a long-term high reward when γ is equal to one.

From (1), the optimal value function $Q(s_t, a_t)$ is approximated directly without the information about the environment. Therefore, the agent can determine the optimal action a^* that maximizes $Q(s_t, a_t)$. Thus, the agent can perform the optimal action easily, and hence the Q-learning is recently used to control and manage communication networks.

In the learning methods explained in the above, an agent learns the optimal action by considering a state of an environment. However, when the number of agents and/or environments is multiple, those learning methods are not effective. Therefore, in such a case, distributed reinforcement learning is utilized [7].

Distributed Q-learning can be utilized when the number of agents and/or environments is multiple. In the following, we consider a case where the number of agents is one but the number of environments is multiple. In the distributed Q-learning, an agent performs action a_i when the agent is in the ith environment e_i whose state is s_i, the agent gains reward $r_i(s_i, a_i)$ as is the case with the conventional Q-learning. Here, a value of action a_i for state s_i is also denoted as $Q_i(s_i, a_i)$ that is calculated from

$$Q_i(s_i,a_i) = (1-\alpha)Q_i(s_i,a_i) + \alpha\Big\{r_i(s_i,a_i) + \gamma \sum_j f(i,j)V_j(s_j)\Big\} \quad (2)$$

$$V_j(s_j) = \max_a Q_j(s_j,a), \quad (3)$$

where α $(0 < \alpha < 1)$ is a learning rate and γ $(0 \leq \gamma < 1)$ is a discount factor. In (2), $f(i,j)$ is the importance degree of environment e_j for environment e_i, and $\max_a Q_j(s_j,a)$ in (3) is the maximal value of the value function for the environment e_j.

For each environment, the agent selects an optimal action $a_i^* = \underset{a}{\operatorname{argmax}} Q_i(s_i,a)$ and perform the selected action. From the above, the agent can learn the optimal action for each environment while considering other environments with the distributed Q-learning.

B. QLAODV

In Vehicular Ad-Hoc Network (VANET), a network topology changes frequently due to vehicle's movement. It is well known that in VANET, ad hoc routing protocols such as Ad hoc On-Demand Distance Vector (AODV) and Dynamic Source Routing (DSR) cannot work efficiently. In [8], in order to use unicast applications effectively in VANET, a routing protocol called QLAODV has been proposed [8].

QLAODV utilizes a distributed reinforcement learning in order to forward data effectively in a highly dynamic network environment. This method utilizes the information about the availability of paths in the distributed reinforcement learning with unicast control packets. From simulation results, [8] has shown that QLAODV outperforms original AODV significantly in highly dynamic networks.

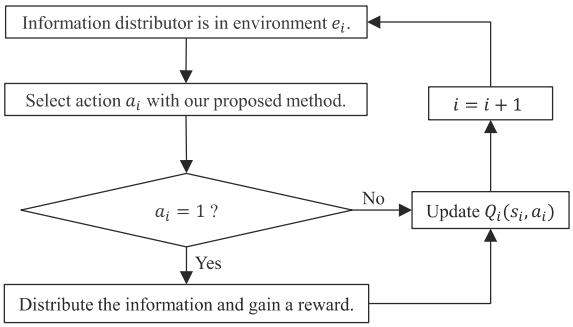

Fig. 1. Flowchart of proposed information distribution.

C. RLAB

In VANET, it is a challenging problem that context-aware broadcasting of information is performed effectively so as to transmit information to the areas of interest (AoI).

In [9], a self-adaptive broadcast scheme has been proposed. This method also utilizes distributed reinforcement learning. In the distributed reinforcement learning, the vehicles are able to collaboratively tune the rate of their broadcast based on the network dynamics. Moreover, in this method, the initial knowledge about geographical distribution of AoI is not required. By using the method, a distributed context-aware broadcasting is implemented easily and used practically. This is because global information is not required and only a partial synchronization is required.

The effectiveness of this method has been evaluated with simulation. From simulation results, it has shown that a significant improvement, in terms of number of useful broadcasts and delay, over the existing approaches, such as gossip-based broadcasting.

III. Proposed System

In this section, we propose an information distribution system with distributed reinforcement learning.

A. Overview

In this system, each road is divided into multiple segments and a segment is used as an environment. In the following, the ith segment is called environment e_i and a state of e_i is denoted as s_i. Moreover, a user moves along the same route as an information distributor and distributes the information to other users. The distributor is an agent for distributed

ISBN: 978-988-19253-3-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Q-learning and learns an optimal action in terms of the information distribution at each environment.

The distributor performs optimal action a_i^*, which maximizes value function $Q_i(s_i, a_i)$, according to state s_i in environment e_i (see Fig. 1). Then, the distributor gains reward $r_i(s_i, a_i)$ from the environment e_i and updates $Q_i(s_i, a_i)$ according to (2). The information distributor moves to the next environment e_{i+1} and continues the above process. Thus, the information is distributed to users by using the distributed Q-learning.

B. Utilization of Distributed Q-Learning

In this subsection, we explain how distributed Q-learning is utilized in our proposed system in detail.

1) State: In this system, state of the ith environment e_i at time t is denoted as $s_i(t)$, and $s_i(t)$ is decided based on the number of users that passed e_i from time $t - L$ to time t. Let $K_i(\tau)$ be the number of users that are passing e_i at time τ. With $K_i(\tau)$, the number of users that passed e_i from time $t - L$ to time t is calculated as $\Sigma_{\tau=t-L}^{t} K_i(\tau)$.

We assume that when $\Sigma_{\tau=t-L}^{t} K_i(\tau)$ is equal to or larger than threshold D, e_i is in a congestion state. Otherwise, we assume that e_i is in a non-congestion state. As a result, $s_i(t)$ is set to zero or one as follows:

$$s_i(t) = \begin{cases} 1 : \text{Congestion state}, & \text{if } \Sigma_{\tau=t-L}^{t} K_i(\tau) \geq D, \\ 0 : \text{Non-congestion state}, & \text{otherwise.} \end{cases} \quad (4)$$

According to state s_i, the information distributor, which is agent, selects an action.

2) Action: The information distributor selects and performs an optimal action at environment e_i. Here, the distributor can select an action at time t among the following actions:

$$a_i(t) = \begin{cases} 1 : \text{Broadcast is performed}, \\ 0 : \text{Broadcast is not performed}. \end{cases} \quad (5)$$

3) Reward: If the information distributor performs action $a_i(t)$ at time t when a state of environment e_i is s_i, the distributor gains reward $r_i(s_i(t), a_i(t))$. In the distributed Q-learning, the agent continues to update value function $Q_i(s_i(t), a_i(t))$, and finally the agent learns the optimal action that maximizes the expected reward. Therefore, by determining a reward function so as to achieve our goal, the information distributor can learn the effective information distribution.

Now, let $N_i(t)$ denote the number of users that receives the information when the distributor performs the broadcast at the environment e_i. In addition, let $G_i(t)$ is the number of users that passed a specified area after receiving the information at e_i. Here, $G_i(t)$ is updated in the specified area and the distributor receives the information about $G_i(t)$ when it passes the area.

With the above parameters, in our proposed system, a reward function is given in the following:

$$r_i(s_i, a_i) = \begin{cases} \zeta N_i(t) + \eta \dfrac{\Sigma_t G_i(t)}{\Sigma_t N_i(t)} + \kappa(t) - 1, & \text{if } a_i(t) = 1, \\ 0, & \text{if } a_i(t) = 0, \end{cases} \quad (6)$$

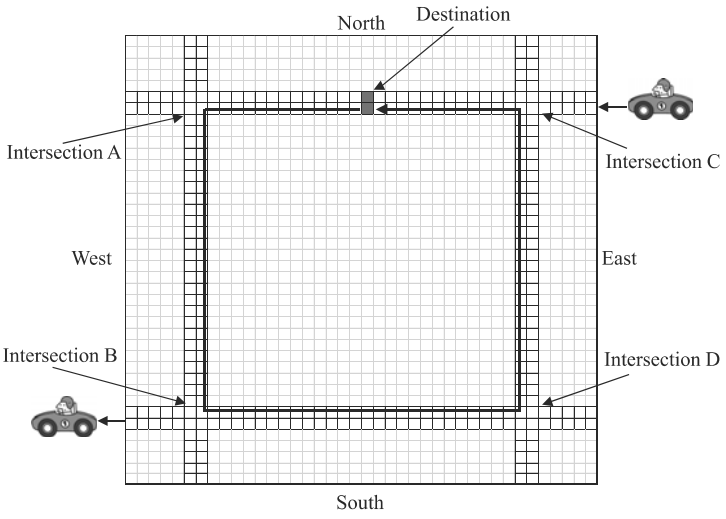

Fig. 2. Simulation scenario.

where ζ and η are setting parameters. In (6), the first term means that a large reward is gained when a large number of users receives the information with a broadcast. The second term denotes that a large reward is gained when the distributor performs the broadcast at a segment where a large number of users that passed a specified area receive the information.

In addition, $\kappa(t)$ in the third term increases if the distributor does not performs the broadcast in the following:

$$\kappa(t) = \begin{cases} \kappa(t-1) + \phi, & \text{if } a_i(t) = 0, \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

where ϕ is a setting parameter. With $\kappa(t)$, the redundant information distribution can be avoided. Finally, the last term is set so that the distributor gains a negative reward when it performs an action that is not appropriate.

4) Update of value function: In this system, $Q_i(s_i(t), a_i(t))$ is updated according to (2). Here, in this update, it is well known that a local optimum action is sometimes selected. To solve this problem, we utilize ϵ-greedy method.

In the ϵ-greedy method, an agent selects and performs an action that maximizes $Q_i(s_i(t), a_i(t))$ with probability $1 - \epsilon$ and otherwise, it selects and performs an action at randam.

Therefore, in our proposed system, with probability $1 - \epsilon$, the distributor compares $Q_i(s_i(t), 0)$ with $Q_i(s_i(t), 1)$ and performs a broadcast if $Q_i(s_i(t), 0) \leq Q_i(s_i(t), 1)$. If $Q_i(s_i(t), 0) > Q_i(s_i(t), 1)$, the distributor does not perform the broadcast. With probability ϵ, on the other hand, the distributor selects an action at random.

IV. Numerical Examples

In this section, we evaluate the performance of our proposed system. In the following, an user that moves along the same route, e.g., a local bus, is an information distributor. The distributor utilizes distributed Q-learning in order to distribute effectively the information to other users.

Figure 2 shows a road map where we evaluate the performance of our proposed system. In this road map, the size of a segment is 20 [m] $\times$ 20 [m], and the distributor moves along a red route at 40 [km/h]. Other users enter the map along

ISBN: 978-988-19253-3-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

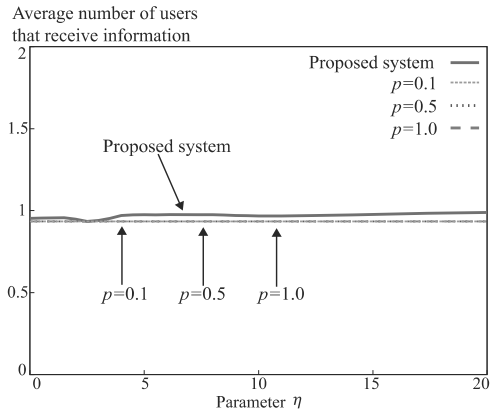

Fig. 3. Impact of η on the average number of users that receives information.

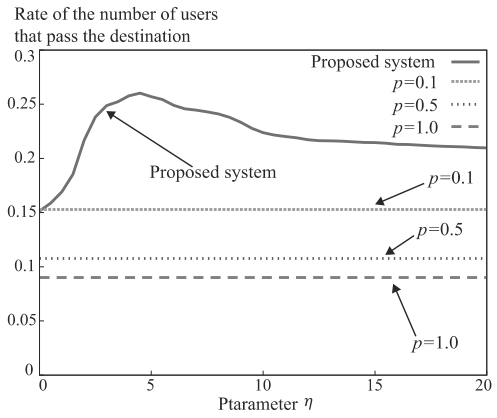

Fig. 4. Impact of η on the average number of users that pass destination.

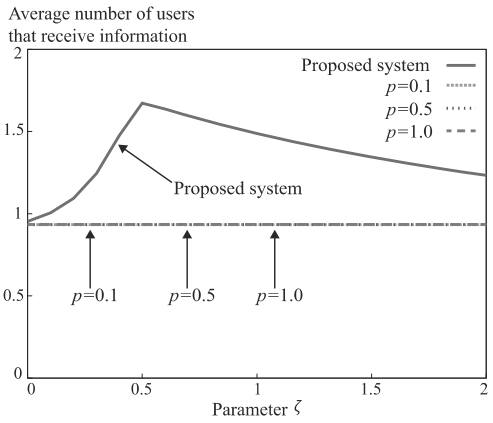

Fig. 5. Impact of ζ on the average number of users that receives information.

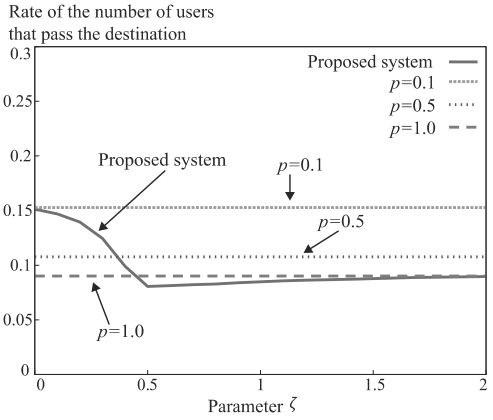

Fig. 6. Impact of ζ on the average number of users that pass destination.

each road with probability P_e and moves at 45 [km/h]. When the users enter intersection B, they go straight down the intersection with probability P_s, turn right with probability P_r, and turn left with probability P_l. At other intersections, they go straight down with probability 0.6, turn right with probability 0.2, and turn left with probability 0.2. The users continue to move until it gets out of the map.

In the system, the distributed information is related to a specified area denoted as *destination* in Fig. 2. Therefore, it is expected that the distributor sends the information to users that pass the destination. In our proposed method, $\alpha = 0.3$, $\gamma = 0.9$, $\epsilon = 0.1$, and $\phi = 0.1$. For the performance comparison, we also evaluate the conventional method where an information distributor performs broadcast with probability p at each segment.

A. Impacts of η and ζ

Figure 3 and Fig. 4 shows the impact of η on the performance of our proposed system. Here, $P_e = 0.3$, $P_s = 0.6$, $P_r = 0.2$, and $P_l = 0.2$. In the following, we evaluate two performance metrics: 1) the ratio of the number of users that pass the destination to the number of users that receive the information and 2) the average number of users that receive the information.

Figure 3 shows the average number of users that receive the information against η in a case of $\zeta = 0$. From this figure, we find that the average number of users that receive the information does not change so much by changing η. Figure 4 shows the ratio of the number of users that pass the destination against η in a case of $\zeta = 0$. From this figure, we find that the ratio of the number of users that pass the destination for our proposed method is larger than that for the conventional method regardless of η and p.

Figure 5 and Fig. 6 show the impact of ζ on the two metrics. Here, η is equal to zero. From these figures, the average number of users that receive the information increases as ζ increases. This is because η affects only the first term in (2). As a result, the number of users that pass the destination decreases by distributing the information redundantly.

B. Impacts of probabilities P_s, P_r, P_l, and P_e

We also investigate the impacts of probabilities with which users move on the effectiveness of our proposed system. In the following, η is equal to eight and ζ is equal to 0.5. Here, we consider three cases: case A, case B, and case C. In case A, $P_s = 0.6$, $P_r = 0.2$, $P_l = 0.2$, and $P_e = 0.3$, and $P_s = 1.0$, $P_r = 0.0$, $P_l = 0.0$, and $P_e = 0.3$ in case B. Moreover, in case C, $P_s = 0.0$, $P_r = 1.0$, $P_l = 0.0$, and $P_e = 0.3$.

Figure 7 and Fig. 8 show the number of information distributions and the average number of users that receive the information, respectively. Figure 9 shows the ratio of the number of users that pass the destination. From these figures, we find that the performance of each system is affected by

ISBN: 978-988-19253-3-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

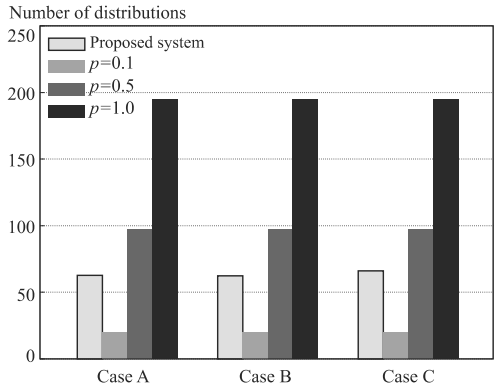

Fig. 7. Number of information distributions for three cases.

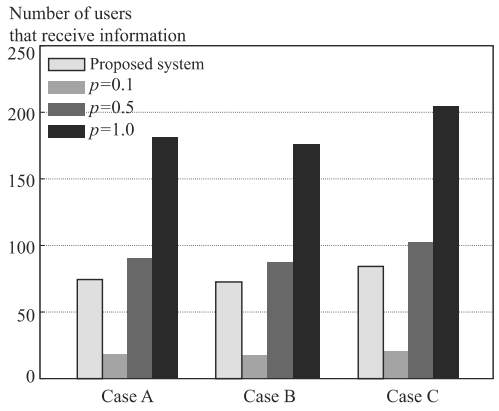

Fig. 8. Number of users that receive information for three cases.

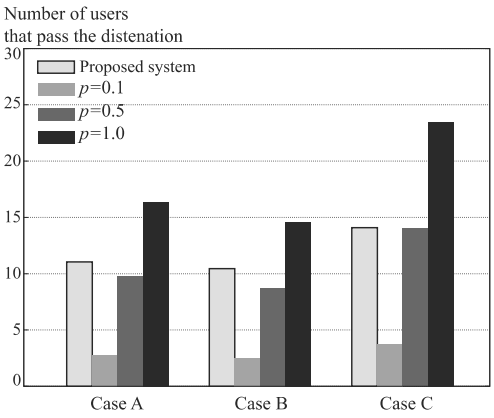

Fig. 9. Number of users that pass destination for three cases.

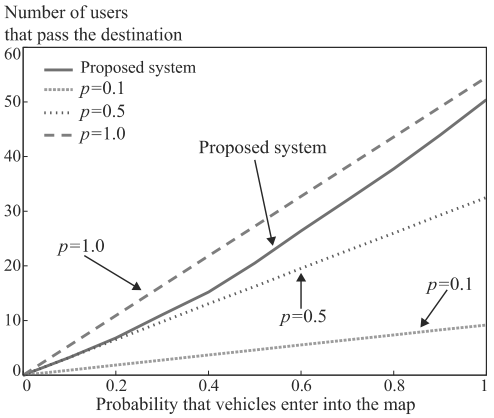

Fig. 10. Number of users that pass the destination vs. P_e.

probability with which users move at intersection. However, our proposed system is effective to distribute the information to the users that pass the destination effectively.

Moreover, Fig. 10 show the impact of probability P_e on the number of users that pass the destination. Here, $P_s = 0.6$, $P_r = 0.2$, and $P_l = 0.2$. From this figure, regardless of P_e, our proposed system can distribute the information effectively. This is because the information distributor can learn the optimal action based on the congestion state.

V. Conclusions

In this paper, we proposed an information distribution system based on distributed Q-learning. We evaluated the performance of our proposed system with simulation. From numerical system, we found that the information can be distributed to users that pass a destination effectively in our proposed system. Moreover, we found that a large number of users can receive the information with a small number of distributions.

ACKNOWLEDGEMENT

This work was partly supported by the Strategic Information and Communications R&D Promotion Programme of the Ministry of Internal Affairs and Communications, JAPAN.

References

[1] T. Yamamoto and T. Tachibana, “Message Transmission with User Grouping for Improving Transmission Efficiency and Reliability in Mobile Social Networks,” submitted to the International MultiConference of Engineers and Computer Scientists 2014 (IMECS 2014).

[2] Y. Kato, Y. Diao, P. Nhy, K. Yagura, T. Kitada, K. Yamaguchi, Y. Takaki, M. Kuwano, and C. Ohta, “Implementation and Experiment of Publish/Subscribe System on Wireless Mesh Networks - Word-mouth Information Distribution Service-,” in *Proc. the Second International Conference on Advances in Mesh Networks*, June 2009, pp. 103–108.

[3] P. Juang, H. Oki, Y. Wang, M. Martonosi, L.S. Peh, and D. Rubenstein, “Energy-Efficient Computing for Wildlife Tracking: Design Tradeoffs and Early Experiences with ZebraNet,” in *Proc. ACM SIGOPS*, July 2002, pp. 96–107.

[4] NEC, “FineChanel,” NEC HP. http://jpn.nec.com/finechannel/index.html.

[5] Z.J. Haas, J.Y. Halpern, and L. Li, “Gossip-based Ad Hoc Routing,” *IEEE/ACM Transactions on Networking*, vol. 14, no. 3, pp. 479–491, June 2006.

[6] R.S. Sutton and A.G Barto, *Reinforcement Learning:An Introduction*, *The MIT Press*, 1998.

[7] J. Schneider, W. Wong, A. Moore, and M. Riedmiller, “Distributed Value Functions,” in *Proc. the Sixteenth International Conference(ICML’99)*, June 1999.

[8] C Wu, K. Kumekawa, and T. Kato,“Destributed Reinforcement Learning Approach for Vehicular Ad Hok Networks,” *IEICE Trans. on Commun.*, vol. E93-B, No. 6, pp. 1431–1442, June 2010.

[9] S. Hosseininezhad, G.N. Shirazi, and V.C.M. Leung, “RLAB: A Reinforcement Learning-based Adaptive Broadcasting for Vehicular Ad-hoc Networks,” *IEEE Vehicular Technology Conference* (*VTC Spring*), May 2011, pp. 1–5.