International Journal of New Computer Architectures and their Applications (IJNCAA) 8(2): 53-60
The Society of Digital Information and Wireless Communications, 2018 ISSN 2220-9085 (Online); ISSN 2412-3587 (Print)

A Study for Dynamically Adjustmentation for Exploitation Rate using Evaluation of Task Achievement

Masashi SUGIMOTO
National Institute of Technology, Kagawa College
Email: sugimoto-m@es.kagawa-nct.ac.jp

ABSTRACT

Until now, in reinforcement learning, a ratio of a random action as known as exploration often has not been adjusted dynamically. However, this ratio will be an index of performance in the reinforcement learning. In this study, agents learn using information from the evaluation of achievement for task of another agent, will be suggested. From this proposed method, the exploration ratio will be adjusted from other agents' behavior, dynamically. In Human Life, an "atmosphere" will be existed as a communication method. For example, empirically, people will be influenced by "serious atmosphere," such as in the situation of working, or take an examination. In this study, this atmosphere as motivation for task achievement of agent will be defined. Moreover, in this study, agent's action decision when another agent will be solved the task, will be focused on. In other words, an agent will be trying to find an optimal solution if other agents have been found an optimal solution. In this paper, we propose the action decision based on other agent's behavior. Moreover, in this study, we discuss effectiveness using the maze problem as an example. In particular, "number of task achievement" and "influence for task achievement," and how to achieve the task quantitative will be focused. As a result, we confirmed that the proposed method is well influenced from other agent's behavior.

KEYWORDS

Reinforcement Learning, Exploration ratio, Action Selection Strategy, Multi Agent, Behavior using Communication, Cooperative Work, Interworking Algorithm, Agricultural Weeding Robot

1 INTRODUCTION

Over the years, many studies have been conducted with the objective of facilitating the working of robots in dynamic environments [1, 2, 3]. Various robots have been developed to assist humans in workspaces, such as a house or factory [4]. In general, robots are required to work effectively and safely in a dynamic environment to achieve their tasks. However, it is not easy to make a robot behave like a human in dynamic environments [5, 6]. When they are working in a certain environment, humans select an appropriate course of action through subconsciously predicting all the changes in the environment and their next state. For achievement these problems, in recent years, various machine learning methods have been suggested. In reinforcement learning, it attracts attention as the technique that often use in the actual robot [7, 8, 9, 10, 11]. However, reinforcement learning has some problems. In one of the problems, a robot does not cope with changing purpose in reinforcement learning. Reinforcement learning has been demanded to achieve various purposes, because what request to robot is diversifying and to achieve various purposes in robot have been wanting, as mentioned above. Therefore, it is important to solve this problem.

Until now, in reinforcement learning, a ratio of a random action as known as exploration [11] often hasn't been adjusted dynamically [12, 13]. However, this ratio will be an index of performance in the reinforcement learning. In this study, agents learn using information from the evaluation of achievement for task of another agent, will be suggested. From this proposed method, the exploration ratio will be adjusted from other agents' behavior, dynamically. In Human Life, an "atmosphere" will be existed as a communication method. For example, empirically, people will be influenced by "serious atmosphere," such as in the situation of working, or take an examination. In this study, this atmosphere as motivation for task achievement of agent will be defined. Moreover, in this study, agent's action decision when another agent will solve the task, will be focused on. In other words, an agent will be trying to find an optimal solution if other agents have been found an optimal solution.

In this paper, we propose the action decision based on other agent's behavior. Moreover, in this study, we discuss effectiveness using a maze problem as an example. In particular, "a number of task achievable" and "influence for task achievement," and how to achieve the task quantitative will be focused. As a result, we confirmed that the proposed method is well influenced from other agent's behavior.

This paper is organized as follows: In section 2, we explain the how to exploration ratio will be adjusted from other agents ' behavior, dynamically. In parallel, we provide details about the proposed method. In Section 3, we explain about the setting for the experiment. Finally, in Section 4, we present the conclusions of this study.

2 A CONCEPT OF ACTION-DECISION BASED ON OTHER AGENT'S BEHAVIOUR

2.1 Basic Idea

In Human life, it seems that there is a kind of information transmission methods called "atmosphere." For example, in behavior such as applause and attitude, interaction and cooperation with the surroundings is performed unconsciously. These are accepted as a kind of "atmosphere." Also, at school examination and some work, "serious atmosphere" propagates, moreover, it is empirically occurring that people are influenced, as unconscious and gradually inflation in the same space (perhaps the readers had been might be confirmed empirically). However, since this concept of "atmosphere" is too abstract. Therefore, in this study will be interpreted it in the form of motivation to accomplish the task. Now, let consider in case of work in human society. When a work is given to someone and work is also given to another person. Think about the pattern that you get. At this time, each one will silently work under "tacit understanding." Of course, at this time, one will quietly and silently attain the work, in the majority of cases, it can be said that it is meaningful to be able to acquire methods to solve work in the shortest time, that is, to finish the work as soon as possible. As mentioned above, behavior selection is done under motivation.

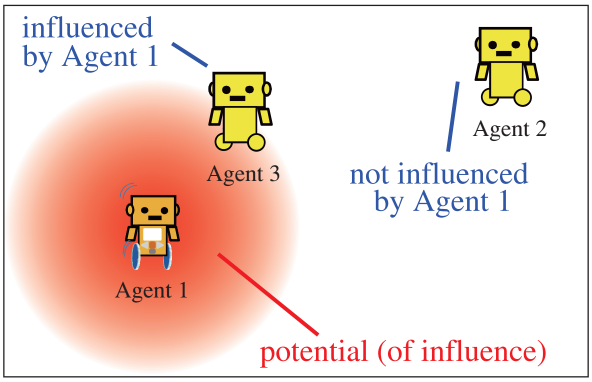

Figure 1. Aim of this study (situation).

Now, in this paper, the above-mentioned "motivation" is greatly thought to propagate like the real environment. This is briefly illustrated in fig. 1. At this time, agent-1 will be contained influence area, and decreasing in a manner of attenuating farther away from himself. When reconsidering mentioned above scenario, if we think that it is working with a maze task, if this influence is within the range, agent-3, which is close to agent-1 in this figure, will be influenced from agent-1 at least. It can be considered harder to receive him influence because agent-2 must be away.

At this time, it can be thought that agent-1 can have some influence in such a way that it decreases as it goes away from itself. If we reconsider previous consideration if operating with a maze task, the number of goals will directly influence the achievement of work, the more seriously it will be accepted as "serious." Moreover, if there is scope for this effect, agent-3 could be thought to be more susceptible than agent-2 and it can be thought that agent-2 won't be affected by the fact that it is away from it. Hence, the impact of this agent by setting the number of achievements of the task can be expressed as a mathematically model, to show this hypothesis; the task given to N_g times:

$$\exp\left|-\frac{p^2}{\frac{\sqrt{2}}{3}rN_g}\right|. \tag{1}$$

In this paper, we express this as the potential possessed by the agent, where p is the position of the agent, r is the range of the potential effect of the agent, and N_g is the number of task achievements. Now, in this case, the exponent part may be indefinite in some cases, especially when $N_g = 0$, is naturally to regard the influence as 1 regardless of the state of the agent.

2.2 How to Influence by Other Agent ?

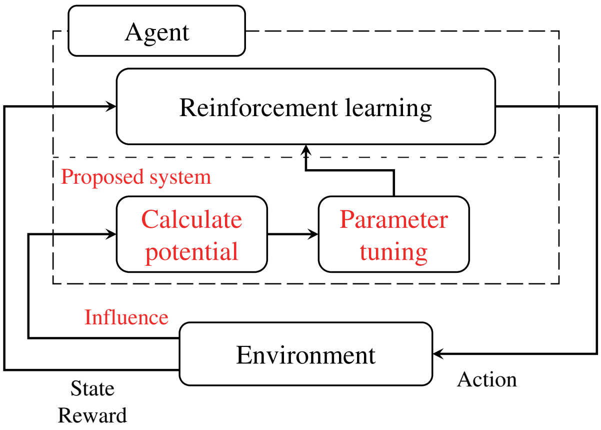

Figure 2. An Outline of the Proposed System.

The figure 2 is the outline of the proposed system. In this method contains two parts; calculate potential part (the part calculating the influence rate of agent) and the parameter tuning part (part calculating the stochastic element from the influence rate). At this time, the agent operating in the proposed system also has the influence of other agents in addition to the state and reward of the environment.

$$\varphi_i = \exp\left|-\frac{(x-p_i)^2}{\frac{\sqrt{2}}{3}r_iN_{g,i}}\right|. \tag{2}$$

On calculate potential part, the influence degree φ_i received from A_i based on the influence received from the potential of another agent A_i will be calculated. This influence degree φ_i is calculated by the following expression:

$$\varphi = \sum_{i\in R}\exp\left|-\frac{(x-p_i)^2}{\frac{\sqrt{2}}{3}r_iN_{g,i}}\right|. \tag{3}$$

Here, x is the position of its own agent, p_i is the position of A_i, r_i is the influence range, and $N_{g,i}$ is the number of achievements of task by A_i. We use the number of achievements as an index. The above expression is based on the expression (2), so the exponent part may be indefinite, so $N_{g,i} = 0, \varphi_i$ is assumed to be 1 irrespective of the state of x, p_i. If R agents affect each other in the same space, as shown in fig. 1.

$$\epsilon = \varphi_i. \tag{4}$$

Next, we describe parameter tuning part, here we decide the probabilistic element of the influence degree φ_i calculated by calculating potential part. In this study, in particular, the ϵ-greedy strategy is used for the behavior selection method. Considering that decision this ϵ dynamically from φ_i.

That is, when another agent discovers an optimal solution, its own agent also selects a behavior of searching for an optimum solution up to the goal according to it.

If R of agents are affecting each other in the same space, they can be summed from the expression (3) as follows:

$$\epsilon = \frac{\varphi_i}{R}. \tag{5}$$

Consider using this φ_i as an *indicator of action selected*, that is, apply it to the random action ϵ in the ϵ-greedy strategy, so the agent A_i in the surrounding r_i. The more seriously the goal is, the more individuals are affected by φ_i, which means that the best action is selected at the local point.

3 VERIFICATION EXPERIMENT – COMPUTATIONAL SIMULATION USING THE PROPOSED METHOD (1)

3.1 Outline of the Experiment

We verify the effectiveness of the proposed method up to the previous section by computer simulation. The effectiveness is evaluated by comparing the difference of the convergence speed of the learning of the proposed reinforcement learning with the proposed method. At this time, the ordinary reinforcement learning method is to learn the route that reaches the goal while avoiding walls and pitfalls through trial and error, and the reinforcement learning to which the proposed method is applied. Behavior will be selected according to behavior facing. Also consider the maze environment with walls and pitfalls consisting of a grid of 4 × 6 shown in fig. 3 as the experimental environment. Moreover, the agent implemented a proposed method will be affected by 2-types agents during task execution.

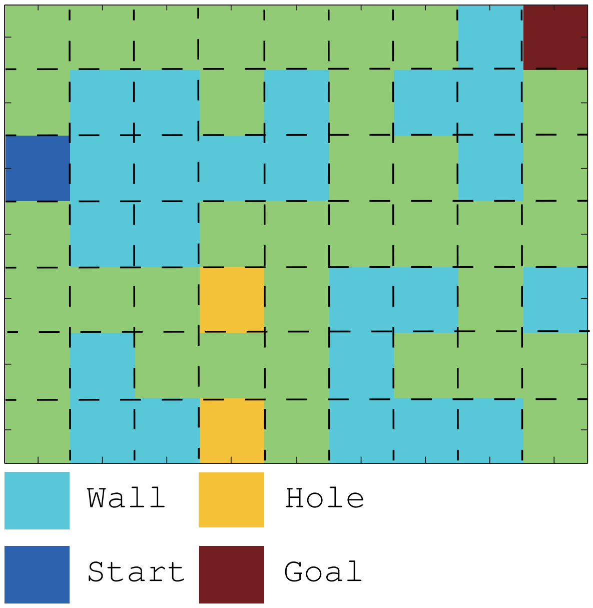

Figure 3. An Experimental Environment (Grid Maze) (1).

In figure 3, the water blue-colored mass is the wall and the orange-colored mass is the catch-point. The two agents are perfect perception and can move up, down, left and right of the grid, among which the pitfall (H). In case of falling to the start point (S) and agent getting a reward -10, or get a reward +1 when agent reaches the goal point (G).

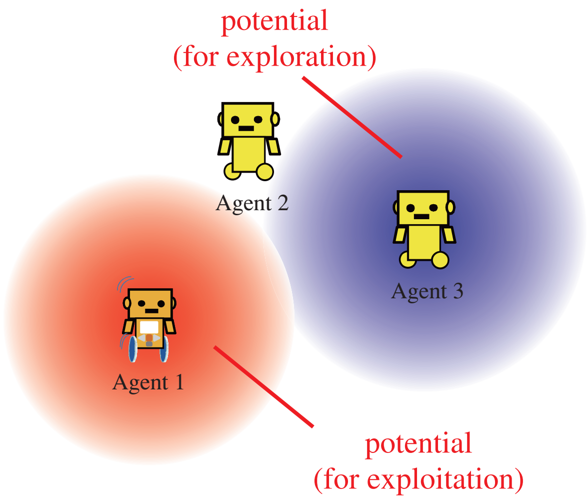

Figure 4. A Structure of Verification Experiment.

Table 1. Experimental parameters for Agents (1).

Agent-1	Agent-2	Agent-3	Property
0	0	0	Initial value of Q values
0.1	0.1	0.1	Learning rate α
0.95	0.95	0.95	Discount value γ
0.1	Eq.(2)	1.0	Exploration rate ϵ
1.0	0.0	2.0	Influences range area r

3.2 Condition of Simulation

In this experiment, we mainly deal with episodic tasks: Agent-1 is an agent that operates with ordinary reinforcement learning, Agent-2 is an agent that combines the proposed method and reinforcement learning. Moreover, Agent-3 is an agent that operates with a completely random action. Agent-1 and Agent-3 are Agent-2’ s learning without being affected, Agent-2 will select actions affected by learning and behavior other agents.

When each agent reaches the goal point (G) from the start point (S), the reward is obtained and the process returns to the start (S) Also, as described above, even when falling into the pitfall (H), it returns to the start point (S). Treat this as one episode In this experiment we will do 400 episodes. Setting of experimental parameters is as shown in the following table 1.

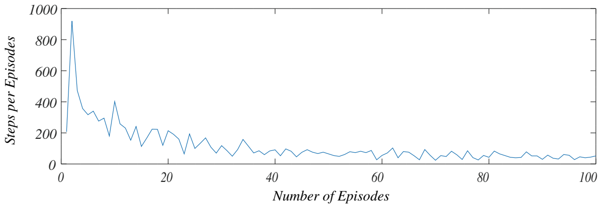

Figure 5. Number of Action per Episodes of Agent-1 (1).

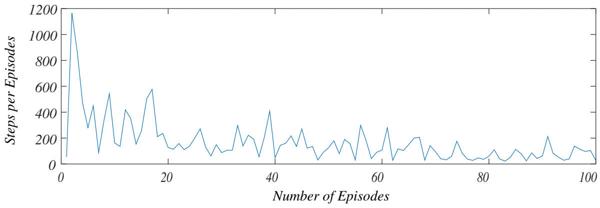

Figure 6. Number of Action per Episodes of Agent-2 (1).

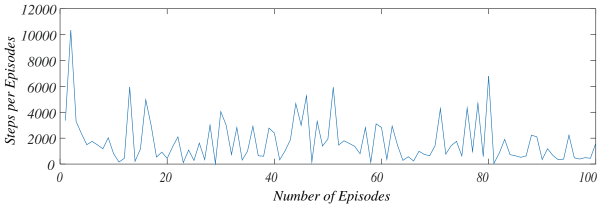

Figure 7. Number of Action per Episodes of Agent-3 (1).

3.3 Discussion on Simulated Results

Figures 5 through 8 shows the results of the experiment. Figures 5 and 7 are the transition of the behavior in each episode by Agent-1 and Agent-3. Figure 6 is the transition of the behavior in each episode of Agent-2 applying the proposed method. The initial value of learning is the number of the behaviors. From these results we can confirm that almost identical to Agent-2, however, as learning progresses, that can be seen that it follows Agent-2 that achieves episode with a fewer number of behaviors than Agent-1.

ϵ is 1 at the beginning of the action and chooses the exploratory behavior, however, as learning progresses, ϵ decreases to 0 as Agent-1's goal number increases, next episode. Moreover, in the next episode, ϵ becomes 1. On the other hands, ϵ had been rising when Agent-1 has faced on Agent-3 was confirmed. We will be considered that the rise resulted from an action strategy of Agent-3. From this result, it could experimentally confirm the fact that actioned will be become while watching the progress of the opponent. Similarly, at the same time. Therefore, the action was realized by searching for an action that finds the optimal solution of the given task (along with it) only when the agent to find the optimal solution of the task that will be confirmed.

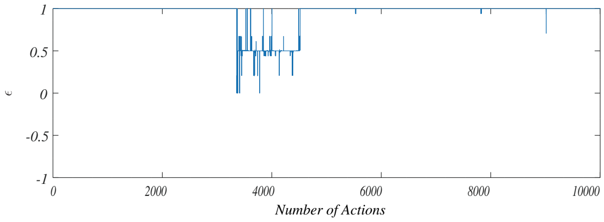

Figure 8. Transition of Exploration Ratio ϵ of Agent-2.

4 VERIFICATION EXPERIMENT – COMPUTATIONAL SIMULATION USING THE PROPOSED METHOD (2)

4.1 Outline of the Experiment

In this section, the potential range will be tried to extend such as circle to ellipse. Therefore, we will extend the potential as below:

$$r_i = \frac{L}{1 + \varepsilon_r \cos \operatorname{arccos}\left(\frac{S_i - S_x}{S_i}\right)} \tag{6}$$

$$L = B^2/A \tag{7}$$

$$\varepsilon_r = C/A \tag{8}$$

$$C = \sqrt{A^2 - B^2} \tag{9}$$

On calculate potential part, the influence degree φ_i received from A_i based on the influence received from the potential of another agent A_i will be calculated. This influence degree φ_i is calculated by the following expression:

$$\varphi = \sum_{i \in R} \exp \left| -\frac{(S_x - p_i)^2}{\frac{\sqrt{2}}{3} r_i N_{g,i}} \right| . \qquad (10)$$

Here, S_x is the position of its own agent, p_i is the position of A_i, r_i is the influence range, and $N_{g,i}$ is the number of achievements of task by A_i. We use the number of achievements as an index. Moreover, A, B are the parameters of ellipse. The above expression is based on the expression (2, 2), so the exponent part may be indefinite, so $N_{g,i} = 0, \varphi_i$ is assumed to be 1 irrespective of the state of x, p_i. If R agents affect each other in the same space, as shown in fig. 1.

$$\epsilon = \varphi_i . \qquad (11)$$

Next, we describe parameter tuning part, here we decide the probabilistic element of the influence degree φ_i calculated by calculating potential part. In this study, in particular, the ϵ-greedy strategy is used for the behavior selection method. Considering that decision this ϵ dynamically from φ_i.

That is, when another agent discovers an optimal solution, its own agent also selects a behavior of searching for an optimum solution up to the goal according to it.

If R of agents are affecting each other in the same space, they can be summed from the expression (3) as follows:

$$\epsilon = \frac{\varphi_i}{R} . \qquad (12)$$

Consider using this φ_i as an *indicator of action selected*, that is, apply it to the random action ϵ in the ϵ-greedy strategy, so the agent A_i in the surrounding r_i. The more seriously the goal is, the more individuals are affected by φ_i, which means that the best action is selected at the local point.

In this experiment, we verify the effectiveness of the proposed method in case of influence potential range will be changed such as a circle to an ellipse. The effectiveness is evaluated by comparing the difference of the convergence speed of the learning of the proposed reinforcement learning with the proposed method. At this time, the ordinary reinforcement learning method is to learn the route that reaches the goal while avoiding walls and pitfalls through trial and error, and the reinforcement learning to which the proposed method is applied. Behavior will be selected according to behavior facing. Also consider the maze environment with walls and pitfalls consisting of a grid of 4 × 6 shown in fig. 9 as the experimental environment. Moreover, the agent implemented a proposed method will be affected by 2-types agents during task execution.

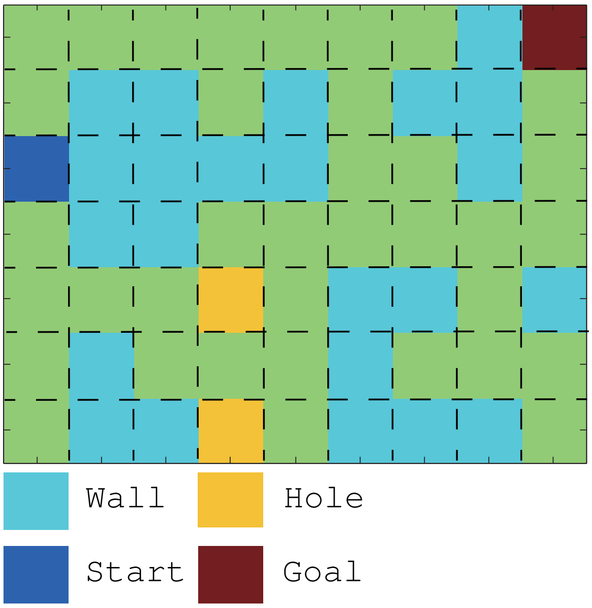

Figure 9. An Experimental Environment (Grid Maze) (2).

In figure 9, the water blue-colored mass is the wall and the orange-colored mass is the catch-point. The two agents are perfect perception and can move up, down, left and right of the grid, among which the pitfall (H). In case of falling to the start point (S) and agent getting a reward -10, or get a reward +1 when agent reaches the goal point (G).

4.2 Condition of Simulation

In this experiment, we mainly deal with episodic tasks: Agent-1 is an agent that operates with ordinary reinforcement learning, Agent-2 is an agent that combines the proposed

Table 2. Experimental parameters for Agents (2).

Agent-1	Agent-2	Agent-3	Property
0	0	0	Initial value of Q values
0.1	0.1	0.1	Learning rate α
0.95	0.95	0.95	Discount value γ
0.1	Eq.(2)	1.0	Exploration rate ϵ
1.0	0.0	Eq.(6)	Influences range area r

method and reinforcement learning. Moreover, Agent-3 is an agent that operates with a completely random action. Agent-1 and Agent-3 are Agent-2's learning without being affected, Agent-2 will select actions affected by learning and behavior other agents.

When each agent reaches the goal point (G) from the start point (S), the reward is obtained and the process returns to the start (S) Also, as described above, even when falling into the pitfall (H), it returns to the start point (S). Treat this as one episode In this experiment we will do 400 episodes. Setting of experimental parameters is as shown in the following table 2.

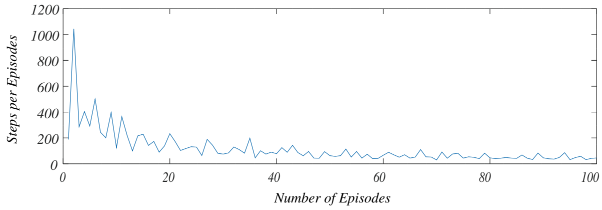

Figure 10. Number of Action per Episodes of Agent-1 (2).

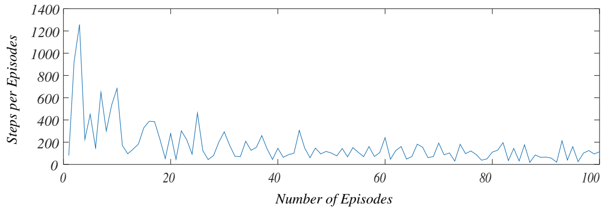

Figure 11. Number of Action per Episodes of Agent-2 (2).

4.3 Discussion on Simulated Results

Figures 5 and 7 are the transition of the behavior in each episode by Agent-1 and Agent-3. Figure 11 is the transition of the behavior in each episode of Agent-2 applying the proposed method. The initial value of learning is the number of the behaviors. From these results we can confirm that almost identical to Agent-2, however, as learning progresses, that can be seen that it follows Agent-2 that achieves episode with a fewer number of behaviors than Agent-1. However, the behavior of Agent-2 seems randomly than former experiment. This symptom is caused that Agent-2 had fallen within the range of Agent-3. Moreover, the range of Agent-3 is extended than former experiment. From the above, also, the action was realized and affected from other agents, will be confirmed.

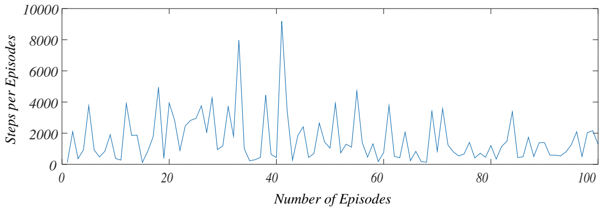

Figure 12. Number of Action per Episodes of Agent-3 (2).

5 CONCLUSION

In this paper, a method to dynamically adjust the action-decision strategy based on other agent's behavioral results, has been proposed. In this method, the evaluation of other agent's behavioral results is its number of agent's task achievement. Moreover, in the proposed method, a number of task achievement and a potential that the agent has unique effective range have been defined. Further, these parameters have been affected the exploration ratio as epsilon. From this method, the simulation results showed the proposed method has been acquired actions to reach the goal more efficiently than conventional method. In other words, the number of trials of proposed method's agent is less, while its agent will not be affected by another agent. Inversely, if another agent will be found the route to reach to goal with the shortest action number, the proposed method's agent will be decided the action that reach to the goal with the shortest action number, repeatedly, and affected by his behavior.

From these results, the proposed method has been confirmed to efficiently accomplish the task, while adjusting itself looking at the other agent's behavior. Therefore, we conclude that the usefulness of the proposed method has been confirmed.

Now, let's consider the living thing, again. Especially, in Human Life, not only "atmosphere," but also "synchronization" has been existed in social life, moreover, this communication method will be important [14]. Further, in other living things, it is often done to specify actions in the form of cautionary or guidance by pheromones [15]. Especially in humans, interaction and cooperation with the surroundings are doing unconsciously within this atmosphere and synchronization [10, 14]. If a robot working in a dynamic environment, such as daily life, will be implemented, these information methods, its will be perform actions and tasks with higher affinity with humans will be predicted.

REFERENCES

[1] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents series)*. The MIT Press, 2005.

[2] S. Asaka and S. Ishikawa. Behavior Control of an Autonomous Mobile Robot in Dynamically Changing Environment. Journal of the Robotics Society of Japan, 12(4):583-589, 1994.

[3] T. Kanda, H. Ishiguro, T. Ono, M. Imai, T. Maeda, and R. Nakatsu. Development of "Robovie" as Platform of Everyday-Robot Research. IEICE Transactions on Information and Systems, Pt.1 (Japanese Edition), J85-D-1(4):380-389, 2002.

[4] International Federation of Robotics. *All-time-high for industrial robots Substantial increase of industrial robot installations is continuing*, 2011.

[5] T. Sogo, K. Kimoto, H. Ishiguro, and T. Ishida. Mobile Robot Navigation by a Distributed Vision System. Journal of the Robotics Society of Japan, 17(7):1-7, 1999.

[6] J. J. Park, C. Johnson, and B. Kuipers. Robot Navigation with MPEPC in Dynamic and Uncertain Environments: From Theory to Practice. IROS 2012 Workshop on Progress, Challenges and Future Perspectives in Navigation and Manipulation Assistance for Robotic Wheelchairs, 2012.

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

[8] N. Sugimoto, K. Samejima, K. Doya, and M. Kawato. Reinforcement Learning and Goal Estimation by Multiple Forward and Reward Models. IEICE Transactions on Information and Systems, Pt.2 (Japanese Edition), J87-D-2(2):683-694, 2004.

[9] Y. Takahashi and M. Asada. Incremental State Space Segmentation for Behavior Learning by Real Robot. Journal of the Robotics Society of Japan, 17(1):118-124, 1999.

[10] A. Agogino and K. Tumer. Reinforcement Learning in Large Multi-agent Systems. In Proc. of AAMAS-05 Workshop on Coordination of Large Scale Multiagent Systems, 2005.

[11] E. A. Guggenheim. *Boltzmann's Distribution Law*. North-Holland Publishing Company, 1955.

[12] N. Shibuya and K. Kurashige. Control of exploration and exploitation using information content. In Proc. of the Nineteenth International Symposium on Artificial Life and Robotics 2014, pp.48-51, 2014.

[13] T. Masaki and K. Kurashige. Decision Making Under Multi Task Based on Priority for Each Task. International Journal of Artificial Life Research, 6(2):88-97, 2016.

[14] D. Li and Y. Du. *Artificial Intelligence with Uncertainty*, Second Edition. CRC Press, 2017.

[15] C. Blum. Ant Colony Optimization: Introduction and recent trends. Elsevier Physics of Life Reviews, 2:353-373, 2005.