Q-Learning Traffic Signal Optimization within Multiple Intersections Traffic Network

Yit Kwong Chin, Wei Yeang Kow, Wei Leong Khong, Min Keng Tan, Kenneth Tze Kin Teo
Modelling, Simulation & Computing Laboratory, Material & Mineral Research Unit
School of Engineering and Information Technology
Universiti Malaysia Sabah
Kota Kinabalu, Malaysia
msclab@ums.edu.my, ktkteo@ieee.org

***Abstract* — Traffic flow optimization within traffic networks has been approached through different kinds of methods. One of the methods is to reconfigure the traffic signal timing plan. However, dynamic characteristic of the traffic flow is not able to be resolved by the conventional traffic signal timing plan management. As a result, traffic congestion still remains as an unsolved problem. Thus, in this study, artificial intelligence algorithm has been introduced in the traffic signal timing plan to enable the traffic management systems' learning ability. Q-Learning algorithm acts as the learning mechanism for traffic light intersections to release itself from traffic congestions situation. Adjacent traffic light intersections will work independently and yet cooperate with each other to a common goal of ensuring the fluency of the traffic flows within traffic network. The experimental results show that the Q-Learning algorithm is able to learn from the dynamic traffic flow and optimized the traffic flow.**

Keywords – Reinforcement learning; Q-Learning; Traffic networks; Traffic signal timing plan management; Multi-agents systems

I. INTRODUCTION

In fully developed urban areas, the transportations are supported by traffic networks with high degree of complexity levels. The traffic networks are fully integrated with different kinds of traffic infrastructures to ensure fluency of the all travelling pedestrians and on-road vehicles within the traffic networks. Unfortunately, within the highly populated urban area, traffic demands towards the traffic networks are extremely high, as the daily operations of the community require high mobility. As the traffic demands rise, the existing traffic networks start to encounter problems like traffic congestions when the traffic infrastructures failed to restrain the saturated traffic flows. The limited landscapes of the urban area have constrained the possibilities of rebuilding the existing traffic networks with infrastructures like fly-over ramps and roundabouts. Thus, improvement of the traffic flow management within the traffic network depends on the traffic light systems.

The performance of traffic light systems also starts to be declined by the traffic network when the traffic demands from the on-road vehicles increased. This is because of the dynamic characteristic of the traffic flows within the networks. Conventional traffic signal timing plan management by using statistic traffic information lacks the ability to rapidly adapt into the dynamic traffic flows. Thus, the necessity of development in intelligent traffic signal timing plan management is need to continuously learning from the dynamic traffic environment for adaptability.

Q-Learning algorithm gathers information from its learning process as its experience and learns from the environments. This learning ability is emphasized in this study of traffic flow optimization system to act as multi-agent systems. The Q-learning algorithm implemented traffic light intersections will able to learn from the traffic environment for increasing its adaptability and making a better decision in the future.

II. REVIEWS OF TRAFFIC SIGNAL TIMING PLAN

In the studies of traffic signal timing plan management, various approaches had been done by researchers to increase the intelligent of the traffic management system. Traffic lights system coordinates the traffic signals among the traffic phases in a traffic intersection, to prevent crashes between traffic phases and ensure the smoothness of the traffic flow within the traffic network.

Traffic signals which are accepted by the world are red, amber and green signals. Red signal represents the restrictions to pass through the intersections as red lights has the lowest frequencies among the three signals, enable it to travel the furthest to warn the arriving vehicles. Amber signal is a warning for vehicles to slow down for stoppage at the intersections. Green signalized a "go" permission for vehicles to pass through the intersections. Every traffic light systems have to complete a sequence of green, amber and red signals for each traffic phase. When the traffic light systems circulate through all the traffic phases, it is called as a complete cycle for the traffic light intersections.

The most conventional traffic signal timing plan is the fixed-time traffic signal plan. In the fixed-time traffic signal plan management, all the duration of the traffic signals are preset in the database and being executed repeatedly. The configurations of the traffic signals durations are done based on the historical traffic statistic gathered for a long period. This method had been eliminated after the increasing traffic demands within the traffic networks start emerge as the method is unable to adapt itself into the dynamic traffic

978-0-7695-4926-2/12 $26.00 © 2012 IEEE
DOI 10.1109/EMS.2012.75

environment. Therefore, researches of developing traffic flow management systems with artificial intelligence has been carried out.

Artificial intelligence techniques with the ability to learn have been proposed in the development of the intelligent traffic light systems history. Artificial neural networks technique is one of the methods that being introduced by researchers into the traffic flow management systems. Extension of green signal duration in a traffic phase is decided by the neural networks algorithm in various traffic situations and the algorithm will continue to learn itself from the environment through reinforcement learning [1]. Communications between vehicles and vehicles to infrastructure are also being proposed as a solution to traffic congestion [2]. Another research is also conducted in a similar way, but using another advance control methods which is fuzzy logic controller. Fuzzy logic is able to interpret the linguistic values into numerical values, thus being used to interpret the traffic congestion situations which are hard to be clearly classified. With the aid of fuzzy logic, the traffic management system able to extend the green signal duration for the continuous incoming traffic flow [3]. There is also research in microscopic traffic control involving only single intersection using fuzzy logic algorithm to comprehend the traffic situation [4]. Fuzzy logic has the advantage of visualizing situation with indistinguishable boundary but still lacking of the ability to learn from the situations.

Evolvable techniques such as genetic algorithm are also being implemented in the traffic management system. Genetic algorithm or evolutionary algorithm is an algorithm that mimics the characteristics of survival in the nature law by letting the populations to evolve themselves throughout the generations. Only the fittest among the populations has the right to survive until the end and the fittest will be the most optimum solution for the designed problem. In traffic management system, populations of genetic algorithm are defined as the green signal durations. So the most surviving chromosomes in the populations are the most optimum green signal durations [5, 6]. In another research, genetic algorithm is use to re-routing the traffic flow within the traffic network to avoid traffic congestions [7].

Reinforcement learning has shown its advantages in the ability of exploring the environment to exploit the most suitable actions in the dynamic situations. Researches involving reinforcement learning in the traffic signal management systems have shown significant results and thus drawn more attentions to traffic management systems' researchers. Reinforcement learning is used to manage the traffic flow within the networks with the longest-queue-first algorithm [8]. Q-Learning as one of reinforcement learning algorithm is applied in the optimization of the traffic flow in single traffic intersection [9, 10]. Besides that, Q-learning algorithm has also been highly valued in the researches of traffic control system as multi-agents systems [11, 12]. In this study, Q-learning algorithm has been proposed to be studied in the traffic signal timing plan management of traffic networks.

III. Q-LEARNING ALGORITHM

Reinforcement learning is an algorithm that can improve and evolve itself from the past experiences. Decisions made from the past is evaluated and stored as experiences data which will provide valuable help in the future. Reinforcement learning is usually presented as Q-learning algorithm which values not only the actions taken but also the states caused by the actions. Therefore, Q-learning has the prospective ability to be implemented into the traffic flow control within the traffic networks.

Q-Learning's concepts can be mirrored by many phenomenons in the real life of nature. The most common metaphor used to illustrate the algorithm is the relationship between a trainer and trainee, for example, a teacher and student, a dog trainer and his dog, or parents and their child. By taking the training of a dog by a trainer as example, the trainer will evaluate each actions of the dog after the commands are given. When a command is given to the dog, the dog will respond to it, and then the trainer will observe and evaluate the performance of the dog. If the dog's action is within the expectations, a reward in the form of food or snacks will be given to the dog; and nothing will be given if the dog did not behave accordingly. In this way, the dog will make a connection between the commands and the actions; it will realize that only the correct command and action pairs will be rewarded. All these experiences encourage the dog to respond according to the command of the trainer, as the dog desires rewards from its trainer. The dog will able to learn every trick that is taught by the trainer after the processes are repeated in times, as it know which command and action pairs will lead it to a reward and which is not. In simple words, Q-learning algorithm is a process of taking actions towards the environment and receiving rewards according to the action taken.

A. Structure of Q-Learning

The flow chart of Q-learning is being illustrated in Fig. 1. The process of Q-learning starts with the initialization of the states and actions in the Q-table. After the initialization, Q-learning will identify its current state in the environment. An action will be chosen from the action lists available by searching for the maximum possible rewards return by the action. Then, the actions chosen will be executed or evaluated. The rewards gained from the actions chosen will be updated in the Q-table. After the actions have been executed, Q-learning will identify the next states in the environment model. Finally, Q-learning will check for the goal accomplishment, the process will start from the beginning again if the goal did not accomplished.

As stated in the previous section, there are many ways to describe about Q-learning algorithm's learning process. In Q-learning algorithm, the role of trainer is played by the environment model, while the Q-learning algorithm itself learns from the environment. Thus, in the development of a Q-learning algorithm, each operation involves in the process must be determined carefully. Q-Learning algorithm is composed with several steps or operations. The

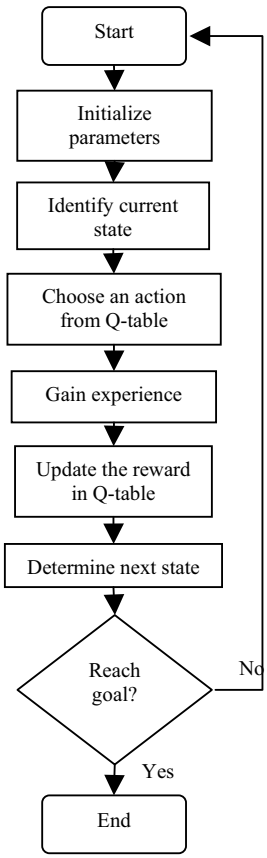

Figure 1. Q-Learning algorithm flow chart.

implementation of Q-learning operation is applied through the evaluation of (1).

$$Q(s,a)_i = (1-\alpha)Q(s,a)_{i-1} + \alpha[R(s,a)_i + \gamma \max_{a'} Q(s',a')] \quad (1)$$

where, s = current state
a = action taken in current state
s' = next state
a' = action taken in next state
i = iteration
α = learning rate
γ = discounting factor

Q-Learning is evaluated from (1), each of evaluated Q-value is the rewards gained from the experiences in the exploration process. Q-table is the memory of the Q-learning algorithm, storing every single state and action pairs along with their rewards.

α, learning rate is an important variable in evaluation of the Q-value. As suggested from the name, α is the factor that will influence the learning rate of the Q-learning algorithm. Learning rate of Q-learning is ranged from 0 to 1, and responsible for the weight of the newly learnt experience. When the learning rate is equal to 1, the Q-learning algorithm will toss away its old experience and treat the newly learnt action as its only experience. This is will set the Q-learning to be opportunist which only care for the immediate rewards. If the learning rate is set to be too low or near to 0, then the Q-learning will suffer from the slow learning rate. For zero learning rates, the experience newly gained is not important, and Q-learning will be in the exploitation mode which only acts based on its past experience, and caused the Q-learning algorithm to stop exploring in the environment.

Discounting factor γ is the variable that decides the importance of the future states. High discounting factor will make Q-learning algorithm to be too speculative, where it will focus more on the possible future rewards and neglect the importance of the current experience. The advantage of having a high discounting factor is enable the Q-learning algorithm converges in a faster rate. Optimum value of the discounting factor is important to let Q-learning algorithm having its speculative characteristic towards the future for long term rewards, and still focus on the short term rewards from the current experience.

In the processing of choosing the actions from the actions list, Q-learning will search for the actions with the maximum rewards benefits. But, there will always be cases, that the actions with current maximum rewards are not producing the real highest rewards return. A mechanism of Q-learning is acting as a support protocol to minimize the possibility of actions being trapped in a local maximum. It is called the ε – greedy selection which is triggered by a greedy probability.

Greedy probability, ε allows the Q-learning to have a chance of choose an action from the actions lists at random which does not return the highest rewards [13]. Greedy probability, ε provides Q-learning to be able to continuously explore itself in the new environment for other possibilities of actions despite of the current highest rewards. However, if the greedy probability is too high, Q-learning will face the difficulty to be converged, as the greedy probability will prompt the Q-learning to continue exploring in the environment.

B. State-Action Pairs

Q-Learning algorithm gains its experience through the exploration in the modelled environment. An accurately defined environment will ease the Q-learning's exploration process. The environment of the Q-learning is formed by the states and actions [14]. Each state which is available in the state space represents the boundary of the environment's map. If the definition of the states is wrongly done, then the Q-learning algorithm might lost itself in the process of exploration or learning.

In this study of traffic signal timing plan management, the states of the designed Q-learning are the level of vehicles in queue at each intersection. There are 4 levels of vehicles in queue defined for this study, where they are categorized from no vehicles in queue to high vehicles in

queue. With each intersection having 4 traffic phases, the total possible states are 256 states combination from the permutation of 4 phases and 4 levels of vehicles in queue.

The actions list that is available for the Q-learning algorithm is also very important parameter, as they act as the navigators for the Q-learning algorithm in the environment. Each action done by the Q- learning will lead it to another state. If the state and action pairs are not set to be correct, then the whole Q-learning will not be able to get the optimum solution in the whole process.

The levels of vehicles in queue at the intersections will be increased by the incoming traffic flow and decreased when they pass through the intersection during the green signal periods. Therefore, green signals are defined as the actions of the proposed Q-learning algorithm. The actions available in the actions lists are 1 second and 5 seconds of green signals distribution. In the process of exploration, the chosen action will be stored in the memory of traffic signal timing plan management system until the goal of the Q-learning is achieved. After that, the green signals will be allocated to each traffic phase in the intersection.

C. Rewards and Penalties Functions

Although states and actions pairs of the Q-learning algorithm are set, rewards and penalties for each chosen actions have to be determined for ensuring the Q-learning algorithm is performing well. In Q-learning algorithm, the basic concept is the best actions will be prized with the highest rewards and the worst actions will be assigned with the least rewards. Besides rewarding each proper action, penalties can be given to those unproductive actions.

The goal of the proposed Q-learning algorithm is to yield the minimum possible vehicle in queue for each intersection. Thus, rewards functions are computed carefully for each appropriate green signal distributions and the actions that yield excessive idle green signals will be penalized. Excessive idle green signals is a period when there are no more waiting vehicles at the intersection, but the green signal is still being allocated and being triggered for that period. Actions that allocate excessive idle green time will be penalized. The proposed Q-learning algorithm will stop after it reach the goal of the system, which is every single traffic link has been allocated with their corresponding traffic signal durations. All of the rewards and penalties returned by the reward and penalties function are stored in the memory of Q-learning as their own experience for their future references.

IV. SIMULATIONS

The study of this paper focused on the implementation of Q-learning algorithm in the traffic signal timing management within traffic network. A traffic network consists of two traffic light intersections has been developed for the simulation of this paper study. The illustration of the traffic network is shown in Fig. 2; intersection A (INTA) is located at the west and intersection B (INTB) at east. The

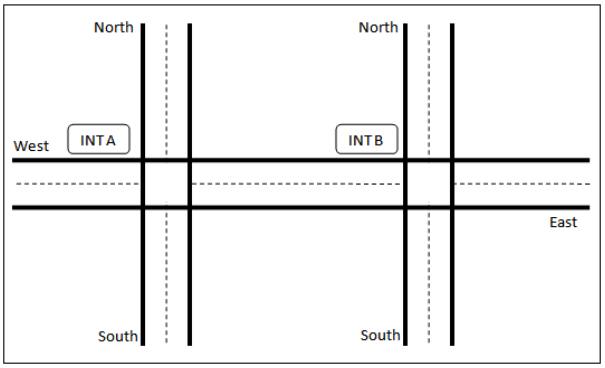

Figure 2. Traffic Network of 2 Intersections.

relationship between INTA and INTB is the main focus of this study.

The two intersections in this traffic networks have the configurations of 4 traffic phases as shown in Fig. 3. Traffic phases are the sequences of the traffic signals activation and only one traffic phase maybe undergo the green signal at a particular time [15]. This is to avoid traffic crashes from different directions at the intersections.

Graphs of simulation results in Fig. 4 are the simulation results of the Q-learning based traffic signal optimizer (QLTSO) at Intersection A for 250 seconds. The 4 graphs in Fig. 4 represented the 4 different traffic phase in Intersection A. The bar graphs at the bottom of each graph show the activation duration of the green signals for each traffic phase. The graphs clearly show that no two traffic phases' green signal is activated at the same moment throughout the whole simulation period, but they are triggered one after another. The decline part of the graphs during green signal is showing that the QLTSO is releasing vehicles to pass the intersection. The QLTSO also performed well in the simulation by distributing the green signal duration effectively to yield minimum vehicles in queue for all the traffic phases in Intersection A.

Different situations within the traffic network are simulated to test the ability of QLTSO in multiple intersections. Each intersection in the network has own QLTSO and work individually but sharing the traffic information together for the global optimization of the traffic flow.

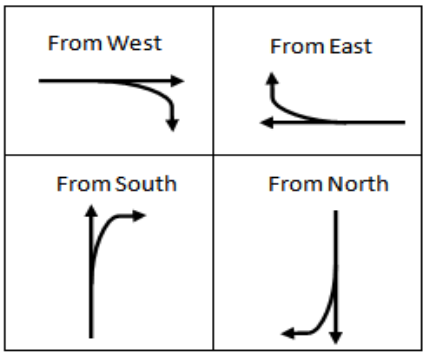

Figure 3. Traffic Phases of Traffic Intersections.

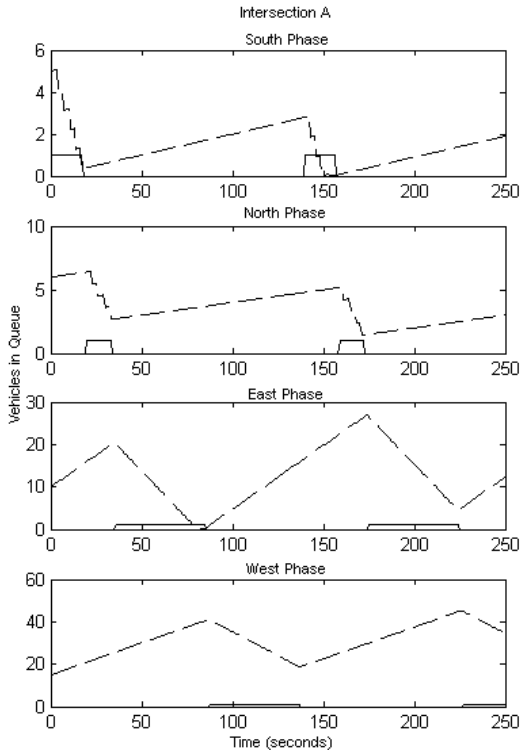

Figure 4. Green Signals Distribution of Traffic Phase.

V. RESULT AND DISCUSSION

The simulation of the developed QLTSO in traffic network with multiple intersections has been carried out for a period of 3600 seconds. Besides QLTSO, a fixed-time traffic signal timing plan has also being simulated in the same environment setting for the analysis purposes. In this simulation, the traffic flows are input through intersection A north phase, south phase and west phase. As for intersection B, north phase, south phase and east phase are being input for the traffic flow.

The east phase of intersection A and west phase of intersection B are linked together and being considered as a closed traffic environment, as the input for both phases are generated from the traffic intersections. The traffic that passes through east link of intersection A will flow into the west link of intersection B and vice versa. The results of the simulation are shown in Table I. In Table I, the number of vehicles that successfully passed through the traffic lights intersections is shown. From Table I, QLTSO has the better performance as compared to the fixed-time traffic signal timing plan.

The difference in the number of vehicles passing through the intersection between the two traffic signal timing management systems can be considered as significant as both of the systems are simulated under the same environment and same kind of data. Besides from north phase and east phase of intersection B, other traffic phases are releasing more vehicles with the developed QLTSO. QLTSO successfully optimized every traffic phases during

TABLE I. NUMBER OF VEHICLES PASS DURING 1HOUR SIMULATION

Traffic Phase	Fixed Time Signal	QLTSO
INT_A_south	55	63
INT_A_north	58	67
INT_A_east	971	1036
INT_A_west	966	1033
INT_B_south	64	64
INT_B_north	71	65
INT_B_east	1067	1024
INT_B_west	1009	1025
Total Pass	4261	4377

the simulation. The results from the table already indicated that QLTSO let more vehicles passing through the intersections compared with the conventional fixed-time traffic signal timing plan. Under the same period of simulation time, QLTSO has released a total of 4377 vehicles within the traffic network compared with the fixed-time traffic management which letting total of 4261 vehicles to travel through the traffic networks.

From the results above, QLTSO has shown the ability of adapting into the traffic environment. Instead of having a fixed duration of green signals during the simulation, QLTSO adapted itself towards the different traffic demands, and calculated the optimum green signal duration for each traffic phase. In south and north phase of intersection A, QLTSO did not neglect the demands from the waiting vehicles and still able to let more vehicles from those two phases to pass through. This shown that, QLTSO has the ability to optimize the traffic flows within the traffic network with the different levels of traffic demands.

Another part of the results from the simulation is shown in Fig. 5. This figure consists of two graphs, the dotted line is representing the fixed-time traffic signal timing plan management and the other solid line is showing the results of QLTSO. In the graphs, the results of intersection A east phase is shown with the QLTSO outperformed the fixed-time traffic signal timing plan. Fixed-time traffic signal started to fail after 1000 seconds and caused the vehicles in queue to rise up throughout the rest simulation.

Observation of the fixed-time traffic signal timing plan during 1500 seconds of simulation time and 2500 seconds of simulation time has revealed the weakness of the fixed-time traffic signal timing plan. The vehicles in queue did not decrease during that green signal duration because of the traffic flow into the traffic phase is larger than the traffic flow out rate. In this situation, the traffic phase need longer green signal as time is needed for the traffic phase to reach its maximum saturation out flow rate. As for QLTSO, the algorithm of Q-learning learnt from the environment and thus allocating longer green signal for that situation to let more vehicles to pass through.

In this graph of Fig. 5, both of the systems also encounter inconsistent traffic flow in. The traffic flow in of this east phase of intersection A is contributed by the vehicles that pass through intersection B from east phase, north and south phase. Improvement of this situation can be done by calibrating the offset between the two adjacent consecutive

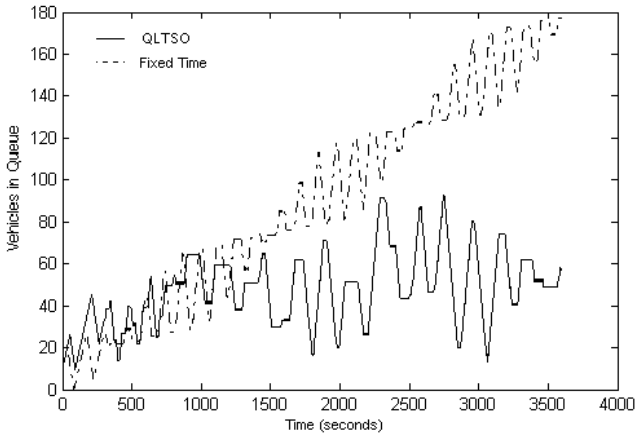

Figure 5. Results of INT A East Phase.

traffic phases. Offset of the traffic phases is important to reduce the waiting time and also reduce the green signal wastage during no-vehicle situation.

VI. CONCLUSION

The developed Q-learning traffic signal optimization system is performing well throughout the simulations. This shows the potentials and the abilities of Q-learning in the traffic signal timing plan management systems. Even without centralized control, the QTSO of each intersections able to work independently and having traffic information sharing with other traffic intersections. Offset between two adjacent traffic phases should be part of the future research of QLTSO in traffic network. Q-learning algorithm has shown its strength in exploration in the dynamic traffic environment and also the adaptability towards the rapid changes of the environment by successfully manages the traffic signals distribution within the traffic networks.

ACKNOWLEDGEMENT

The authors would like to acknowledge the financial assistance from Ministry of Higher Education of Malaysia (MoHE) under Exploratory Research Grant Scheme (ERGS) No. ERGS0021-TK-1/2012, Universiti Malaysia Sabah (UMS) under UMS Research Grant Scheme (SGPUMS) No. SBK0026-TK-1/2012, and the University Postgraduate Research Scholarship Scheme (PGD) by Ministry of Science, Technology and Innovation of Malaysia (MOSTI).

REFERENCES

[1] Y.Dai; J.Hu; D. Zhao; F. Zhu; , "Neural network based online traffic signal controller design with reinforcement training," *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011, pp.1045-1050, doi: 10.1109/ITSC.2011.6083027.

[2] V. Gradinescu, C. Gorgorin, R. Diaconescu, and V. Cristea. "Adaptive Traffic Light using Car-to-Car Communication." *In Proceeding of Vehicular Technology Conference,* 2007, pp.21-25.

[3] K. Khiang Tan, M. Khalid, and R. Yusof. "Intelligent Traffic Lights Control by Fuzzy Logic." *Malaysian Journal of Computer Science*, vol. 9, no. 2, pp. 29-35. 1996.

[4] E. Azimirad, N. Pariz, and M.B.N. Sistani. "A Novel Fuzzy Model and Control of Single Intersection at Urban Traffic Network." *IEEE Systems Journal*, vol. 4, no. 1, pp. 107-111, March 2010.

[5] K.T.K. Teo, W.Y. Kow, and Y.K. Chin. "Optimization of Traffic Flow within an Urban Traffic Light Intersection with Genetic Algorithm." *In proceeding of 2nd International Conference on Computational Intelligence, Modelling and Simulation,* 2010, pp. 172-177. doi: 10.1109/CIMSiM.2010.95

[6] Y.K. Chin, K.C. Yong, N. Bolong, S.S. Yang, and K.T.K. Teo. "Multiple intersections traffic signal timing optimization with genetic algorithm." *In Proceeding of International Conference on Control System, Computing and Engineering,* 2011, pp.454- 459. doi: 10.1109/ICCSCE.2011.6190569

[7] F. Teklu, A. Sumalee, and D. Watling. "A Genetic Algorithm Approach for Optimizing Traffic Control Signals Considering Routing." *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, pp. 31-43, 2007.

[8] I. Arel, C. Liu, T. Urbanik, and A.G. Kohls. "Reinforcement Learning-based Multi-Agent System for Network Traffic Signal Control." *IET Intelligent Transport Systems*, vol. 4, no. 2, pp. 128-135, 2010.

[9] Y.K. Chin, N. Bolong, A. Kiring, S.S. Yang, and K.T.K. Teo. "Q-Learning based Traffic Optimization in Management of Signal Timing Plan." *International Journal of Simulation, Systems, Science & Technology*. ISSN: 1473-804x. vol. 12, no. 3, pp. 29-35. 2011.

[10] Z.Y. Liu, and F.W. Ma, "On-line Reinforcement Learning Control for Urban Traffic Signals." *In Proceedings of the 26th Chinese Control Conference*, 2007, pp. 34 – 37.

[11] P.G. Balaji, X. German, and D. Srinivasan, "Urban Traffic Signal Control using Reinforcement Learning Agents." *IET Intelligent Transport Systems*, vol. 4, no. 3, pp. 177-188, 2010.

[12] B.Abdulhai, R. Pringle, and G.J. Karakoulas. "Reinforcement Learning for True Adaptive Traffic Signal Control." *Journal of Transportation Engineering*, vol. 129, no.3, pp. 278-285, June 2003.

[13] Mi. Tokic, and G. Palm. "Value-Difference based Exploration: Adaptive Control between Epsilon-Greedy and Softmax." *In KI 2011: Advances in Artificial Intelligence*. J. Bach, S.Edelkamp, Ed. Berlin: Springer, 2011, pp. 335-346.

[14] C.J.C.H. Watkins, P. Dayan. "Technical Note: Q-learning." *Machine Learning*, vol. 8, no.3, pp. 279-292, May 1992.

[15] N.J. Garber, and L.A. Hoel. *Traffic and Highway Engineering. 3rd Ed.* Pacific Grove, California: Thomson, 2002.