ROK MARSETIČ, M.Sc.
E-mail: rok.marsetic@fgg.uni-lj.si
DARJA ŠEMROV, B.Sc.
E-mail: darja.semrov@fgg.uni-lj.si
MARIJAN ŽURA, Ph.D.
E-mail: marijan.zura@fgg.uni-lj.si
Faculty of Civil and Geodetic Engineering,
University of Ljubljana
Jamova 2, SI-1000 Ljubljana, Slovenia

Intelligent Transportation Systems (ITS)
Original Scientific Paper
Accepted: June 14, 2013
Approved: Apr. 8, 2014

ROAD ARTERY TRAFFIC LIGHT OPTIMIZATION WITH USE OF REINFORCEMENT LEARNING

ABSTRACT

The basic principle of optimal traffic control is the appropriate real-time response to dynamic traffic flow changes. Signal plan efficiency depends on a large number of input parameters. An actuated signal system can adjust very well to traffic conditions, but cannot fully adjust to stochastic traffic volume oscillation. Due to the complexity of the problem analytical methods are not applicable for use in real time, therefore the purpose of this paper is to introduce heuristic method suitable for traffic light optimization in real time. With the evolution of artificial intelligence new possibilities for solving complex problems have been introduced. The goal of this paper is to demonstrate that the use of the Q learning algorithm for traffic lights optimization is suitable. The Q learning algorithm was verified on a road artery with three intersections. For estimation of the effectiveness and efficiency of the proposed algorithm comparison with an actuated signal plan was carried out. The results (average delay per vehicle and the number of vehicles that left road network) show that Q learning algorithm outperforms the actuated signal controllers. The proposed algorithm converges to the minimal delay per vehicle regardless of the stochastic nature of traffic. In this research the impact of the model parameters (learning rate, exploration rate, influence of communication between agents and reward type) on algorithm effectiveness were analysed as well.

KEY WORDS

reinforcement learning; Q learning; road artery; traffic control; traffic lights

1. INTRODUCTION

Due to the continuous growth of population and motorization and due to changes in travel behaviour we are facing increase of traffic volumes on the existing road system. Consequently, problems related to high traffic volumes (traffic jams, time loss, increased pollution...) occur. An initial solution could be road extensions, but in many cases there is no available space for additional road lanes; therefore, the solution should be found within the existing road infrastructure, e.g. the optimization of the traffic flow. The most common way of traffic control in various types of intersections are traffic lights. Traffic light controllers are designed to coordinate the time between crossing traffic flows that use the same space at an intersection. In the past, for traffic light coordination most commonly the pre-timed signal controllers were used. Signal plans for pre-timed traffic light controllers were defined on the basis of historical traffic volume data. One could say the traffic flow is a living organism which is changing continuously. In this context, the question arises as to whether a system based only on historical data is sufficiently effective. This suggests that traffic light controllers sensitive to traffic changes should be developed. Advanced control systems are known as actuated traffic light controllers. A signal plan for this type of signal controllers continuously checks traffic flow and adjusts itself to the current traffic volume. Unlike pre-timed traffic light controllers with a fixed program, actuated traffic light controllers adjust the program according to the detected traffic volume. The duration and sequence of phases are calculated in a way that all vehicles on the road network have minimal loss time and at the same time the capacity utilization of an intersection or group of intersections is the highest. Despite the flexibility of the system, considerable work for system calibration at major traffic volume change is required [1] .

A few commercial systems with integrated actuated traffic light controllers are in use. The most known methods are TRANSYT [2], SCOOT [3] and SCATS [4].

The SCOOT and SCATS systems coordinate traffic flow on a group of intersections, where all the intersections have the same signal plan. Systems adopt phase duration and sequences to traffic volume detected in real-time on all lanes of an intersection. The next system for signal plan optimization in real time is the OPAC system [5]. The first generation of the OPAC system was applicable only for isolated intersections, and the last generation has been upgraded to control road arteries and networks. The next step ahead for the afore mentioned systems are systems based on artificial intelligence. Using the reinforcement learning principle, new, learning and adaptive, systems are being developed. Reinforcement learning as one of the branches of artificial intelligence has been proven to be effective in various areas in traffic engineering, e.g. ramp metering [6] and traffic light control.

Sen and Head [7] proposed dynamic programming, where a model for traffic forecast is needed, for signal plan optimization of each cycle. The neuro-fuzzy approach for signal plan optimization [8] has had limited success due to the lack of responsiveness to traffic changes. Reinforcement learning was used for adaptive algorithm for signal plan implemented in a Dutch simulation tool [9]. This algorithm improves the level of service, but has not proved promising in the case of oversaturated traffic flow and in significant fluctuations in traffic volume. Thorpe [10] has presented results of signal plan optimization with reinforcement learning. Using the SARSA method Thorpe has improved the level of service compared to conventional pre-timed traffic light controllers, but this approach is not applicable for real-time use. Abdulhai et al. [11] reported improved level of service using Q learning for signal plan optimization. This approach is not suitable for moderate volume of traffic network nor for adding lanes and intersections, due to an unmanageable number of states. The Q learning algorithm for signal plan optimization was also used only for a straight traffic flow [12], for an individual intersection [13], for a group of intersections with a multi-agent [14] and with a single-agent [15] approach. Reinforcement learning algorithms can be easily implemented into practice and can operate in real time. Q learning is a reinforcement learning algorithm which converges to the optimal strategy.

Literature review identifies that artificial intelligence can be used for signal plan optimization more efficiently than an existing conventional approach. In this paper optimal signal plan on a road artery with a sequence of three intersections was tested with Q learning algorithm. The effectiveness of the algorithm is represented with delay per vehicle and the number of vehicles that have left the network compared to an actuated signal plan optimization approach. Since the primary goal of signal plan optimization on a road artery is fluent traffic on the main road, the priority to minimize vehicle delay and maximize the number of vehicles that have left the network on the main road is set up. In one case all agents were independent and in the second case the agents interact (the leader agent has additional information about the phase and phase duration of the neighbour agent). The impact of local and global reward was analysed, and the most appropriate level of information between neighbouring agents in different traffic conditions is proposed.

2. THEORETICAL APPROACH

With the reinforcement learning approach, which summarizes different fields of machine learning, medicine, psychology, computer and mathematical science, useful engineering applications were developed [16]. Reinforcement learning problems can be solved by using one of three main methods, namely, dynamic programming (DP) in correlation with the Markov decision process and the Bellman equation, temporal difference learning (TD) and the Monte Carlo methods (MC).

The basic principle of reinforcement learning is a learning agent which interacts with the environment. By taking actions, an agent changes the state of the environment through which the agent wins a reward. Based on this feedback the agent learns and adopts decisions to maximize the utility function. Reinforcement tasks are usually treated in discrete time steps. In each time step t the system gets information of environment state s_t. Based on this information an agent performs an action a_t, and in the next state gets a reward r_t. Through the reward, an agent is told the adequacy of the previously chosen action. In the next time step $t+1$ the environment responds to the agent with the change of the state (s_{t+1}) [17].

The groundwork of reinforcement learning is temporal difference (TD) learning. In general TD methods are learning algorithms for the long-term forecast of dynamic systems [18]. TD methods are incremental learning procedures developed specially to forecast systems where the reward is assigned based on the difference between successive steps [19]. In TD learning methods the principles of DP and MC methods are combined. Like in MC methods, in TD methods an agent learns directly from its experiences, without knowing the environment model. The similarity of TD and DP methods are in continuous utility function updates, without waiting for the final outcome. In MC methods, a utility function $V(s_t)$ is updated at the end of the process, while in TD methods a state-value function $V(s_{t+1})$ is updated in every time-step based on the reward r_{t+1} [16]. The basic form of TD method, designated with the TD(0) is as follows:

$$V(s_t) = V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)], \qquad (1)$$

where α is the learning rate and γ the future reward discount factor. The advantage of TD over DP methods is the ability to find optimal strategy only by experience, without knowing the environment model, therefore only the reward and the probability of the next state are needed.

Q Learning

Q Learning is an off-policy algorithm of the TD method. The matrix of the value-action function $Q(s_t,a_t)$ is updated for each transition between states [16]. In the matrix values $Q(s_t,a_t)$ for all state-action pairs are written. At the transition from state s_t into state s_{t+1}, where action a_t was chosen and reward r_{t+1} was assigned, the algorithm makes the following update:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \\ +\alpha\left[r_{t+1} + \gamma \max_{a+1} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right], \qquad (2)$$

where α is the learning rate and γ the future reward discount factor.

To come closer to real traffic situations on real road network the Stochastic Q learning algorithm was introduced. The expected value of the value-action function in the next step is:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \\ +\alpha[r_{t+1} + \gamma \textit{ expected } Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)], \qquad (3)$$

where:

$$\textit{expected } Q(s_{t+1},a_{t+1}) = \\ = \frac{\sum_j \left[n(s_t,s_{t+1,j},a_t) \times \max_{a_{t+1}} Q(s_{t+1},a_{t+1})\right]}{n(s_t,a_t)} \qquad (4)$$

where n is the number of agent transitions from state s_t into state $s_{t+1,j}$ at action a_t and $n(s_t,a_t)$ is the number of all previously accomplished actions in state s_t. At the beginning of the algorithm matrix Q was initialized at zero.

One of the conditions for an algorithm in a stochastic environment to converge to the optimal strategy is learning rate reduction. Parameter α (learning rate parameter) defines to what extent the newly acquired information influences the experiences; factor $\alpha \approx 0$ means that an agent will not learn, while $\alpha = 1$ means that the agent will consider only the most recent information. In our study we examined factor α in interval [0, 1]. The discount factor γ defines the importance of (potential) future reward; factor $\gamma = 0$ means that the agent will consider only the most recent reward, while $\gamma \approx 1$ means the agent will aim for long-term high reward. From the literature [15] and [20] it is evident that the researchers got better results with parameter γ closer to 1. In our study $\gamma = 0.8$ was used.

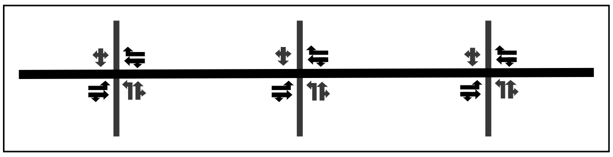

Figure 1 - Road network

3. METHODS AND RESEARCH RESULTS

3.1 Input data and simulation start-points

The efficiency and effectiveness of the Q learning algorithm was analysed on three intersections. An example of the road network was used for verification of the proposed algorithm applicability in a stochastic environment and for testing parameters that influence the algorithm. The road network consisted of three four-leg intersections with left turning lanes at east, south and west legs. All intersections are controlled with a two-stage signal plan. In the first phase green is in the west-east direction, and in the second phase green is in the south-north direction. At all intersections on all legs all turning movements are allowed (right, straight, left). Road artery alignment and intersections geometry are presented in *Figure 1*.

Traffic volumes are provided in vehicle per hour, the ratio main road vs. side road was from 100 : 10 to 100 : 15, which corresponds to the ratio of the traffic flow on a road artery in a real traffic situation. A vehicle enters the road network following the Poisson distribution.

The efficiency of the proposed algorithm was tested in both, peak hours when the traffic flow is oversaturated (v/c = 1) and in off-peak hours when the traffic flow is saturated (v/c = 0.9). Verification of input parameters was simulated in iterations, where one iteration presented one hour in real life. In one test 80 runs were performed, which is approximately 3,000 traffic light cycles. In each run random vehicle arrivals were changed by changing the initial random speed of arrivals for the precise evaluation of the stochastic nature of arrivals. The effectiveness of the algorithm was determined through total delay time and the number of vehicles that have left the network by comparing with the results of traffic flow optimization with an actuated signal plan. In both cases the same traffic volume and parameters were taken into account. The comparison with the pre-timed signal plans is not suitable, due to the inability of the pre-timed signal plans to adjust to the changed traffic conditions (traffic volumes). The micro-simulations were carried out with the software VISSIM, where traffic situations can be simulated in detail. VISSIM can be run with various external applications and can serve as a tool for verifying different algorithms. The algorithm was written in Visual Basic and we accessed the model information and ran the simulation through the COM interface. The optimization and coordination of signal plan for the examined road network and for chosen traffic volume was made with available commercial state-of-the-art software. The optimized signal plan was later used as input for simulation of traffic flow coordinated with actuated traffic light controllers in VISSIM. VISSIM was also used for the simulation of traffic light optimization with the proposed Q learning algorithm. In both cases the same traffic volume and parameters were taken into account.

A state was defined as a 4-dimensional vector of queue length on the main and side lanes, current signal phase and the duration of green signal. In the case of a multi-agent approach the influence of additional information on the leader agent (the leader agent is the neighbouring agent on the left side) was examined. Agents get the information about the leader agent phase ($f_{va} = \{0,1\}$) and the information about the leader agent's green phase duration (t_{va}), and in this case the state is defined as a 6-dimensional vector. Since a state is a vector of all possible queues on the main and side roads, all possible durations of green phases, all possible leader agent green phase duration and information of the leader agent phase, the number for all possible states increase dramatically and the learning process of the agent is threatened. For this reason three classes of queues (short, middle and long), twelve classes of duration of green phase (class size 8 seconds) and four classes of duration of leader agent green phase (class size 24 seconds; $t_{va} = \{1 \text{ to } 4\}$) were introduced. The number of all possible states is 216, and if the leader agent information is known to other agents, the number of all possible states is $216 * |f_{va}| * |t_{va}|$.

An agent can choose one of the possible actions, whether it extends or changes phase (switching the green signal to the other direction). An agent makes a decision every 4 seconds, but not before minimum duration of the green signal (10 seconds). Variable signal cycles were taken into account.

The reward function depends on the queue lengths on all intersections legs. The queue length was chosen for reward function since queues can be easily measured in the field. And since queues longer than the distances between two neighbouring intersections are not desirable in real life situations they can be penalized. In other words, the use of queue lengths for the reward function prevents situations where the queue would extend to the neighbouring intersection. The value of the reward function is negative. In the case of global reward, the reward function value is the sum of all queues on all intersections, and in case of local rewards, the value of the reward function is the sum of all queues on the intersection controlled by the agent. The goal of the reward function is the optimization of traffic lights signal plans in the way the sum of queues on all lines l is minimal. In our case the traffic flow on the main road has higher priority. Reward function can be written as follows:

$$R_{t+1} = \sum_i q_i(t+1) w_i \tag{5}$$

where q is the value of queue length, w is the weight of individual lane defined as

$$w_i = \begin{cases} 0.6, i \in L_m \\ 0.4, i \in L_s \end{cases},$$

where L_m is a set of lanes on the main road and L_s is a set of lanes on a side road. The queue lengths values of the individual lane are:

$$q_i = \begin{cases} 0, & l_i \leq 30 \\ 0.2, & 30 < l_i \leq 60 \\ 0.4, & 60 < l_i \end{cases} \tag{6}$$

where l_i is the length of the queue on lane i.

3.2 Parametric study

Q learning algorithm converges to the optimal result if adequate level of exploration in order to achieve a sufficient number of state visits is ensured. One of the methods is ε-greedy strategy [11]. In our study the ratio between exploration vs. exploitation ε was tested for different traffic volumes. Different combinations of learning rate, exploration vs. exploitation and required number of visits when an agent can start to exploit the knowledge and the different combinations and level of information between agents were tested. The efficiency and effectiveness of the proposed algorithm was evaluated based on two parameters, namely the average delay per vehicle and the speed of algorithm convergence as a function of the number of state visits when the agents start to exploit the knowledge. The number of state visits n_ε, when parameter ε was set to zero (agents do not explore), was tested for the values between 100 and 800, with step 100. The goal was to determine the combination of parameters when the algorithm converges to the optimal value the fastest and with the lowest oscillations.

3.3 Results

The proposed Q learning algorithm for the signal plan optimization was tested for the different combination of parameters mentioned above for two different traffic volumes, namely for oversaturated and saturated traffic flow. In all cases, regardless of the traffic volume, the results were the most promising in case of $\varepsilon = 10$ (an agent in 1 of 10 trials explores and does not exploit its knowledge).

3.3.1 Oversaturated traffic flow

Analysis of the result for the oversaturated traffic flow has shown that the proposed Q learning algorithm achieves the best results with parameters $\alpha = 0.2$ and $\varepsilon = 10$. The information about the leader agent's phase duration is not needed. The results are better if agents know only the phase of the neighbouring agent (of the leader agent) or even no additional information is needed. It can be summarized that in the case of oversaturated traffic flow information about the leader agent, which would have an influence on traffic lights off-set between signal lights, does not improve the results. The use of a local reward serves better mainly due to traffic behaviour in oversaturated traffic flow and the fact that the distance between intersections enables that information about queue length contains all information each agent needs to know about the neighbour agent, in our case about the leader agent. In all parameter combinations, the results were more promising in case of local reward, which is the opposite of [11]. Global reward in the case of oversaturated traffic flow blurs the action efficiency of an individual agent. This in the long term leads to a less solid experience and leads to less efficient decisions. The analysis of the parameter ε impact on the convergence speed was carried out. *Figure 2* shows the results where the convergence speed was the highest. Lines present the results of 3 tests, where all parameters were the same, only the number of state visits when the agent stops to explore the environment and starts to exploit the acquired knowledge differ. As the optimal result we choose the parameter combination with results, having the lowest data dispersion. In case of oversaturated traffic flow we get the most promising results when we set $n_\varepsilon = 400$.

Effectiveness and efficiency of proposed algorithm was estimated by comparison with an actuated signal system. Two parameters of effectiveness and efficiency were tested, namely the average delay per vehicle and the number of vehicles that left the road network in one simulation hour. Comparison was made with the same software and with the same input data. As shown in *Figure 3*, the results are better in case of traffic light optimization with proposed Q-learning algorithm, since more vehicles left the road network and at the same time delays were shorter.

3.3.2 Saturated traffic flow

In the case of saturated traffic flow vehicle delays are shorter if agents interact, if agents know the information about the leader agent phase and leader agent phase duration. Additional information configures an adequate off-set between signal lights, and convoys of vehicles travelling together form. Also in the case of saturated traffic flow the results are better with the use of a local reward. The impact of parameter ε on the speed of convergence was tested. From the results, it is evident, that delays converge fastest if an agent stops exploring after 200, 400 or 600 visits of each state. The lowest data dispersion is when agent starts to exploit its knowledge after 600 visits (*Figure 4*).

The proposed algorithm is effective also for the traffic light optimization in saturated traffic flow conditions. The proposed algorithm improves delays and increases the number of vehicles that left the network in comparison with the actuated signal plan (*Figure 5*).

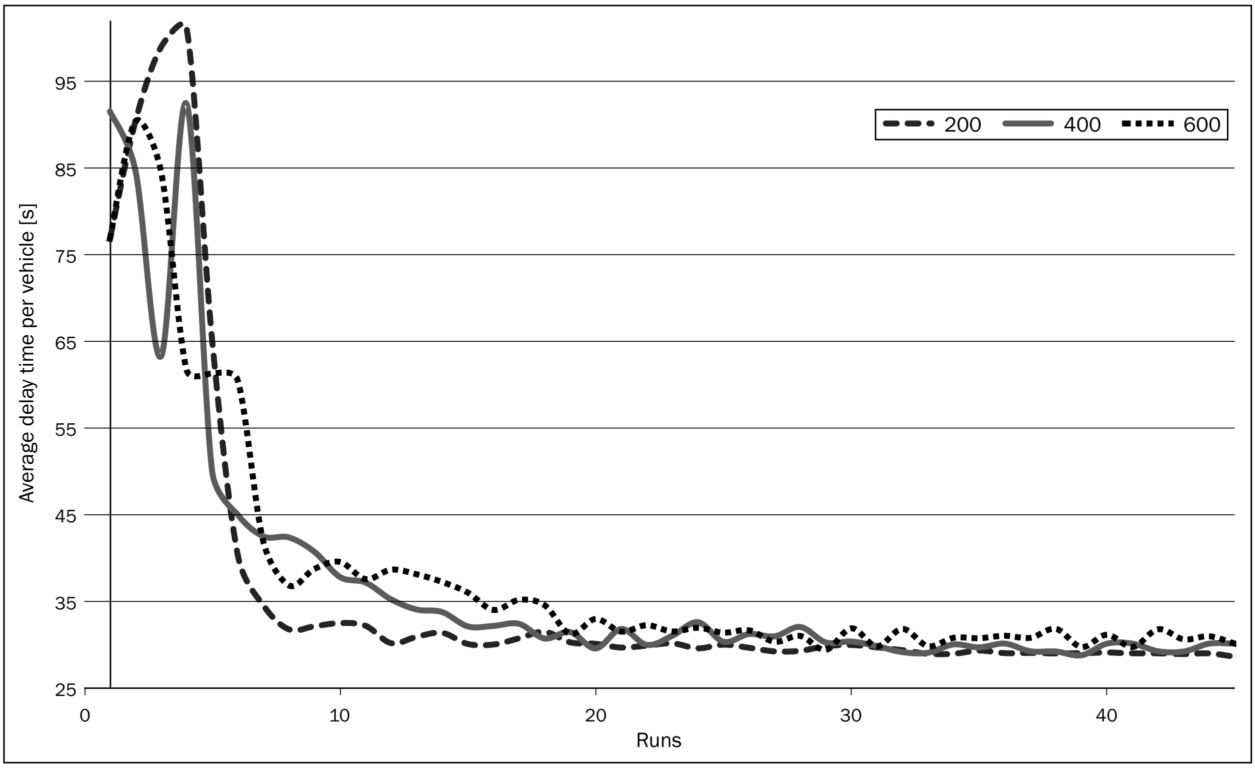

Figure 2 - Behaviour of proposed algorithm for different n_ε (oversaturated traffic flow)

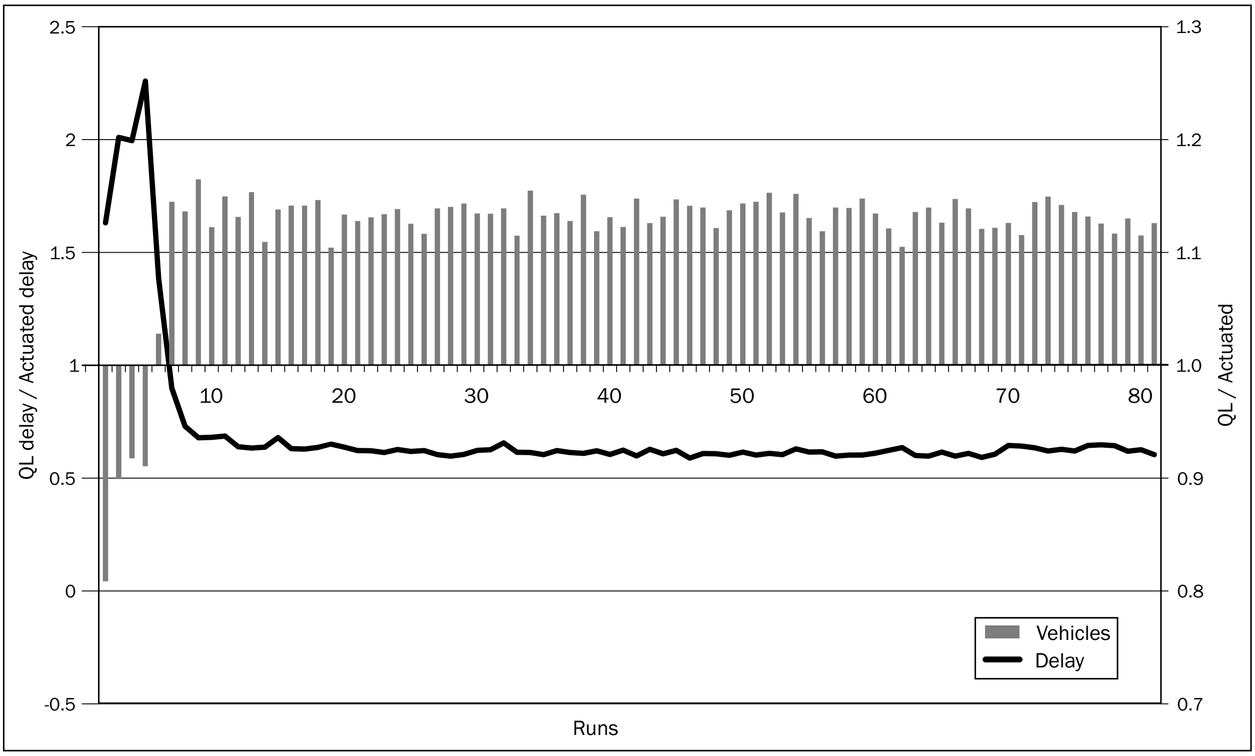

Figure 3 - Q learning vs. actuated traffic light optimization (oversaturated traffic flow)

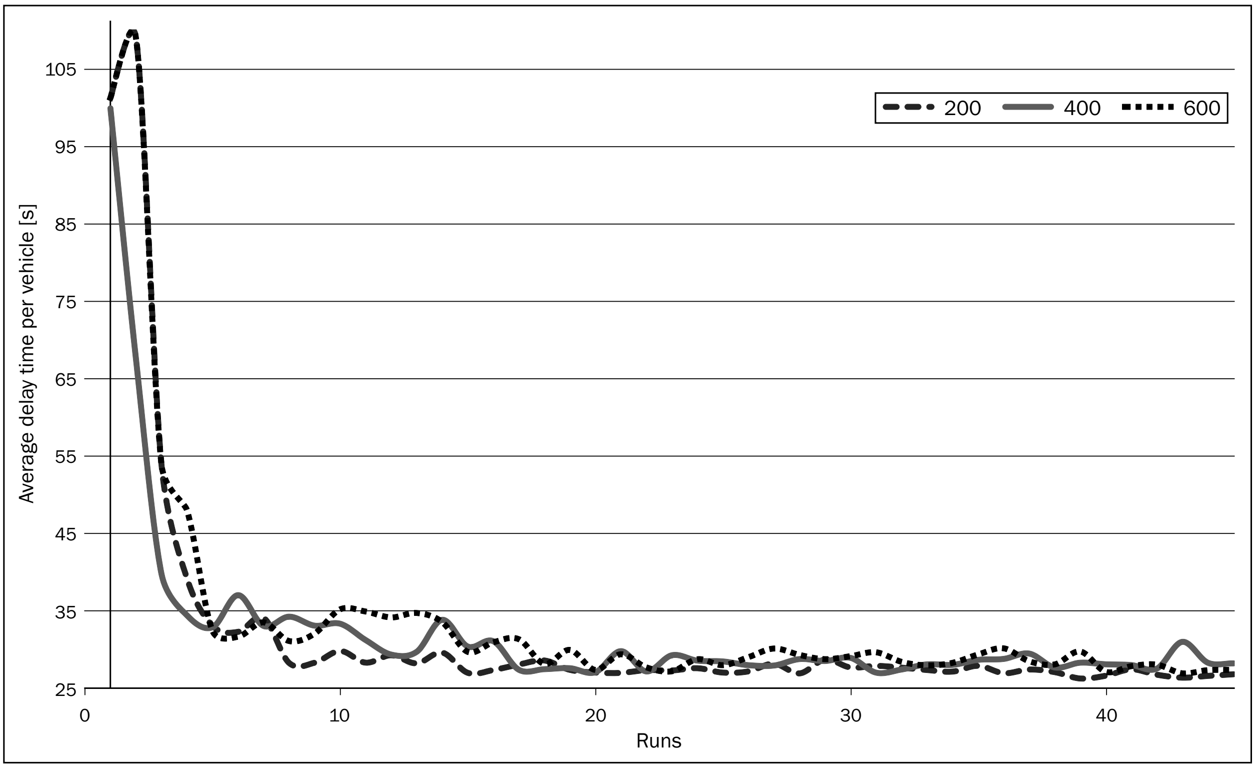

Figure 4 - Behaviour of proposed algorithm for different n_ε (saturated traffic flow)

4. DISCUSSION

From the cited works it can be concluded that reinforcement learning can be successfully used for signal plan optimization. In previous works different simplifications were done, e.g. limited number of intersections taken into account [13], not taking into account turning movements [12] and simplified traffic flow model [9]. Signal plan optimization is a very complex problem and simplifications are understandable and

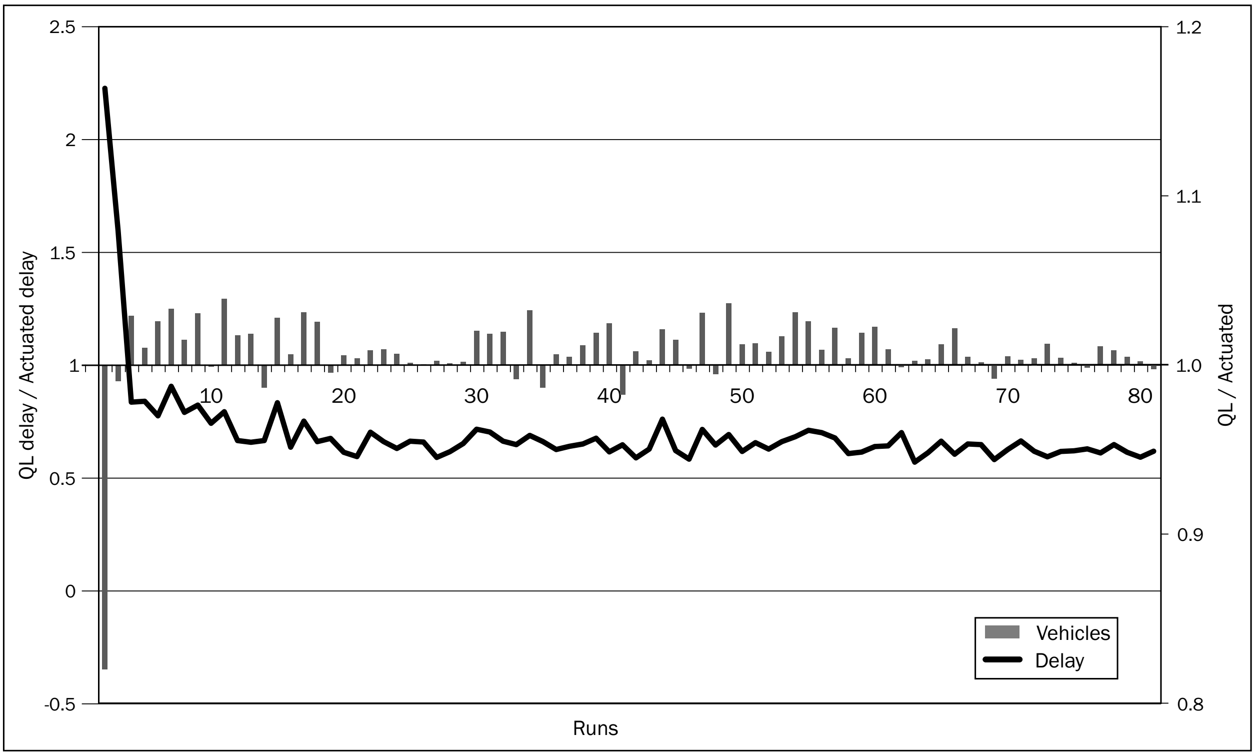

Figure 5 - Q learning vs. actuated traffic light optimization (saturated traffic flow)

necessary. In our research, although this is our first attempt to optimize a signal plan with reinforcement learning, we made only few simplifications. Q learning algorithm with a multi-agent approach, where agents independently take actions with or without knowing additional information about a leader agent's state is proposed. In our research traffic flows on all intersection legs were taken into account, variable traffic light cycles and micro-simulation tool with very accurate real traffic flow simulation was used. The results are more promising than with an actuated signal plan, in both saturated and oversaturated traffic flows. Significant improvement of delays and increased number of vehicles that left the network in oversaturated traffic flow is very important for improving traffic conditions in peak hours.

5. CONCLUSION

The research was carried out with the aim to explore the efficiency of the proposed Q learning algorithm for traffic light optimization on a road artery in real time. The proposed algorithm contains two stages. It begins with the learning and training phase. In *Figure 2* and *Figure 4* one can see that the system learns over 30 or 20 simulation hours and after that the average total delay curve oscillates near the optimal average total delay. The agent has learned and can respond in every step (in our study every 4 seconds) to the moderate changes in traffic volume and provides optimal signal plan for a given situation in real time. In the case of completely different traffic volume, e.g. in case of traffic deviation due to accident on the main road the traffic could increase on side roads, the agent would react instantly, but would need some time to optimize the signal plan to new traffic conditions. Due to the selected state parameters the intersections are considered to be almost independent and consequently the computation complexity is linearly dependent on the number of intersections. Although computation complexity rises with the number of intersections, it uses very small amount of processor time for real-time decision-making even with hundreds of intersections due to the basic nature of exploitation in Q learning algorithms.

The average delay per vehicle and the number of vehicles that left the road network were considered and compared with actuated traffic light controllers. The parameters were examined in saturated and oversaturated traffic flows. The proposed Q learning algorithm decreases the average delay per vehicle and increases the number of vehicles that left road network compared to the actuated signal approach, since agents learn and adopt decisions to traffic stochastic nature. Due to these results we conclude that the multi-agent Q learning algorithm for signal plan optimization could be used in traffic engineering for traffic light optimization. In the case of implementation of the proposed algorithm in a real environment the agent would have to adopt the policy to traffic changes continuously.

During our research we got several ideas for upgrading the algorithm. The number of states will be optimized and afterwards the algorithm will be tested on an artery with more intersections and for different traffic volumes on all legs. The algorithm response to different traffic volumes will be optimized during the learning phase, where traffic volumes will be changed constantly and agents will learn for new experiences.

Mag. **ROK MARSETIČ**
E-mail: rok.marsetic@fgg.uni-lj.si
DARJA ŠEMROV, univ.dipl.inž.grad.
E-mail: darja.semrov@fgg.uni-lj.si
Dr. **MARIJAN ŽURA**
E-mail: marijan.zura@fgg.uni-lj.si
Fakulteta za gradbeništvo in geodezijo,
Univerza v Ljubljani
Jamova 2, SI-1000 Ljubljana, Slovenija

POVZETEK

OPTIMIZACIJA KRMILJENJA CESTNE ARTERIJE Z UPORABO SPODBUJEVANEGA UČENJA

Temeljno načelo optimalnega krmiljenja prometa je ustrezen odziv na dinamične spremembe prometa v realnem času. Učinkovitost krmilnega programa je odvisno od velikega števila vhodnih parametrov. Prometno odvisni signalni programi se dobro prilagajajo prometnim razmeram, vendar se ne morejo popolnoma prilagoditi stohastični naravi prometa. Zaradi kompleksnosti problema analitične metode niso uporabne, zato v tem članku predstavljamo uporabo hevristične metode za optimizacijo svetlobno signalnih naprav (SSN) v realnem času. Z razvojem področja umetne inteligence so se pojavile nove možnosti reševanja kompleksnih problemov. Namen članka je prikazati uporabnost algoritma učenja Q za krmiljenje SSN. Algoritem smo preverjali na arteriji s tremi križišči in njegovo uspešnost in učinkovitost primerjali s prometno odvisnimi SSN. Analiza rezultatov povprečnih zamud na vozilo in število prepeljanih vozil skozi prometno mrežo je pokazala, da je predlagan algoritem učenja Q uspešnejši od krmiljenja SSN s prometno odvisnim krmiljenjem. V predstavljeni raziskavi smo analizirali tudi vpliv parametrov modela (faktor stopnje učenja, faktor stopnje raziskovanja, vpliv poznavanja informacij med agenti, vrsta nagrade) na uspešnost algoritma.

KLJUČNE BESEDE

spodbujevano učenje; učenje Q; cestna arterija; krmiljenje prometa; svetlobno signalne naprave

REFERENCES

[1] **Anžek M**, **Kavran Z**, **Badanjak D**. *Adaptive Traffic Control as Function of Safety*. 12th World Congress on Intelligent Transport Systems and Services. San Francisco, 2005.

[2] **Robertson DI**. *Research on the TRANSYT and SCOOT Methods of Signal Coordination*. ITE Journal. 1986;56(1):36-40.

[3] **Hunt PB**, **Roberetson DI**. **Bretherton RD**, **Winton RI**. *A traffic responsive method of coordinating signals*. Crowthorne, Berkshire: Transport and Road Research Laboratory; 1981.

[4] **Lowrie PR**. *Scats, Sydney co-ordinated adaptive traffic system: a traffic responsive method of controlling urban traffic*. Darlinghurst, NSW Australia: Roads and Traffic Authority NSW; 1990.

[5] **Gartner N. Opac**: *A demand-responsive strategy for traffic signal control*. Transportation Research Board; 1983.

[6] **Veljanovska K**, **Bombol K**, **Maher T**. *Reinforcement Learning Technique in Multiple Motorway Access Control Strategy Design*. PROMET - Traffic&Transportation. 2010;22(2):117-123.

[7] **Sen S**, **Head K**. *Controlled optimization of phases at an intersection*. Transportation science. 1997;31(1):5-17.

[8] **Bingham E**. *Neurofuzzy Traffic Signal Control* [Master's thesis]. Helsinki, Finland: Dept. of Engineering Physics and Mathematics, Helsinki Univ. of Technology; 1998.

[9] **Wiering M**. *Multi-agent reinforcement learning for traffic light control*. ICML '00 Proceedings of the 17th International Conference on Machine Learning; 2000.

[10] **Thorpe TL**, **Anderson CW**. *Traffic Light Control Using SARSA with Three State Representations*. IBM Corporation; 1996.

[11] **Abdulhai B**, **Pringle R**, **Karakoulas GJ**. *Reinforcement Learning for True Adaptive Traffic Signal Control*. Journal of Transportation Engineering. 2003;129(3):278–285.

[12] **Gregoire PL**, **Desjardins C**, **Laumonier J**, **Chaib-draa B**. *Urban Traffic Control Based on Learning Agents*. IEEE Intelligent Transportation Systems Conference. IEEE; 2007. p. 916-921.

[13] **Lu S**, **Liu X**, **Dai S**. *Q-Learning for Adaptive Traffic Signal Control Based on Delay Minimization Strategy*. 2008 IEEE International Conference on Networking, Sensing and Control; 2008. p. 687-91.

[14] **Arel I**, **Liu C**, **Urbanik T**, **Kohls AG**. *Reinforcement learning-based multi-agent system for network traffic signal control*. IET Intelligent Transport Systems. 2010:4(2):128-135.

[15] **Prashanth, L.A.**, **Bhatnagar, S.**: *Reinforcement Learning With Function Approximation for Traffic Signal Control*. IEEE Transactions on Intelligent Transportation Systems. 2011;12(2):412-421.

[16] **Sutton RS**, **Barto A**. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press; 1998.

[17] **Abdulhai B**, **Kattan L**. *Reinforcement learning: Introduction to theory and potential for transport applications*. Canadian Journal of Civil Engineering. 2003;30(6):981–991.

[18] **Perez-Uribe A**. *Introduction to reinforcement learning* [Internet]; 1998. Available from: http://ape.iict.ch/teaching/AIGS/AIGS_Labo/Labo5-RL/sarsa.html

[19] **Sutton RS**. *Learning to predict by the methods of temporal differences*. Machine Learning. 1988;3:9–44

[20] **Wiering M**, **van Veenen J**, **Vreeken J**, **Koopman A**. *Intelligent traffic light control*. Technical Report. Utrecht University, Institute of Information and Computing Sciences; 2004. 31 p.