

Traffic Routing Management System with Markov Decision Process (MDP) Using Q-Learning

¹Olatunji M. A. and ²Lawal O. N.

^{1,2}*Yaba College of Technology, Yaba, Lagos, Nigeria.*

Date of Submission: 10-12-2021

Revised: 20-12-2021

Date of Acceptance: 25-12-2021

ABSTRACT: Commuters are very adaptive to any traffic situation, no matter how unfavorable, which is mostly due to the fact that alternative routes are not known. Q-learning is a simple approach for agents to learn how to operate optimally in controlled Markovian domains. It operates by gradually improving its assessments of the quality of specific actions performed at specific states. This research work is aimed at designing a dynamic traffic routing system with traffic weight heuristics in getting the most optimized path from any point in the road network of the University of Lagos in terms of path and weighted costs. One of the reinforcement learning algorithms; Heuristically Motivated Q learning (HQL) was used for a more efficient and dynamic handling of parameters. HQL would help to determine better and optimal route to the same destination. The case study is the road network of the University of Lagos community. The various routes were considered as states in the Markovian domain. It was implemented using C-Sharp (C#) Programming language to derive an optimized path to a destination. This system can help commuters navigate round the school community with minimal delay. Markov Decision Process (MDP) was adopted in modeling the research.

Keywords: Markovian domain, Markov Decision Process, HQL, Q-Learning

I. INTRODUCTION

Traffic congestion has become an everyday occurrence as a result of increasing traffic and the road network's limited capacity. Since delays caused by peak hour traffic congestion are well-predictable and they account for the largest portion of all traffic congestion delays, one strategy to avoid this is to select an alternative route between two nodes at the problematic hours.

II. RELATED WORKS

This research work relates to a problem in the class of time dependent dynamic vehicle routing.

Nambajemariya & Wang (2021) describes the increasing number of private automobiles on the road, transportation has become one of the most prevalent aspects of people's everyday lives, resulting in highly complex traffic in urban areas. Saying, there is energy consumption, environmental degradation, unforeseen accidents, and time is spent as a result of traffic congestion and traffic jams.

Dipak (2019) explains the increasing number of cars on the road, particularly four-wheelers and large vehicles, causes frequent traffic jams and longer commute times, especially in congested areas. Putting restrictions on the usage of vehicles, on the other hand, cannot be the solution to such problems. Instead, in this circumstance, an effective traffic control strategy may be advantageous.

Chavhan & Venkataram (2019) focuses on modern metropolitan regions are the primary markers of a country's economic growth. The quantity and frequency of automobiles in metropolitan areas has expanded dramatically, causing concerns such as traffic congestion, accidents, environmental pollution, economic losses, and excessive fuel waste.

Allan et. al., (2017) describes many areas of modern life are affected by traffic problem, economic development, increased carbon emissions, time spent, traffic accidents and health implications are all factors to consider. In this situation, modern societies can rely on traffic management technologies to decrease traffic congestion and its harmful consequences. Traffic management systems are made up of a combination of application and management technologies that

work together to improve transportation system overall traffic efficiency and safety.

An excellent high-level look at how a learning agent should acquire a sense of purpose – “a cognitive ability to create its own goals and reward itself for its accomplishment” is as described by (Sitkoff, 1996). This is premised upon the existence of layers of cognitive function, a common feature in proposed mental architectures (Sloman & Logan, 2000).

As shown in Figure 1, an agent is connected to its environment through action and

perception in a basic reinforcement learning model. The agent receives some indication of the current state, s , of the environment as input at each phase of interaction, and then picks an action, a , to generate as output. The agent is notified of the value of this state change via a scalar reinforcement signal, r , which influences the state of the environment. The learning agent's behavior, B , should choose activities that increase the learning agent's long-run sum of values or rewards. It can learn to do so over time through systematic trial and error.

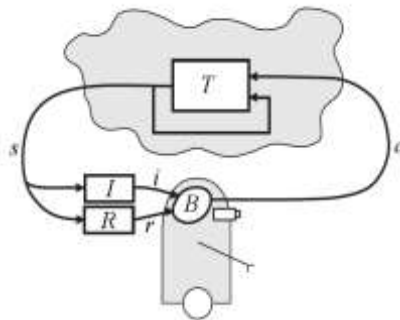


Figure 1.

The figure 1: Input function I determines how the agent perceives the status of the environment.

Formally, the model consists of:

1. A discrete set of environment states, S ;
2. A discrete set of agent actions, A ; and
3. A set of scalar reinforcement signals; typically $\{0, 1\}$, or the real numbers.

The state is a representation of the current circumstances of the learning agent's environment in reinforcement learning terminology. The action is a learning agent output that has the ability to change its surroundings. The policy of the learning system is the set of actions it takes in response to its states (Gaskett, 2002).

III. RESEARCH MATERIALS AND METHOD

Markov Decision Process (MDP) is the underlying model upon which this research work is based. It is one of the many combinatorial design models. It can also be viewed as feedback compliant system that ensures every action is rewarded and the reward helps the learning agent to take smarter decisions as the learning activity progresses. MDP is a 4-tuple (S, A, P, R) model, where: $S = \{s_1, s_2, \dots, s_n\}$ is a finite set of states; $A = \{a_1, a_2, \dots, a_m\}$ is a finite set of actions; P is a Markovian state transition model — $P(s, a, s_')$ is the probability of making a transition to state $s_'$ when taking action a in state s ($s \xrightarrow{a} s_'$); and, R is

a reward (or cost) function. — $R(s, a, s_')$ is the reward for the transition $s \xrightarrow{a} s_'$. The Q Learning method was utilized in particular to make parameter management more efficient and dynamic. The technique was developed as a way to improve Markov Decision Process problems' solutions. Its ability to choose between immediate and delayed rewards is one of its distinguishing characteristics.

The proposed system is modeled as a control theory problem using MDP. It is assumed that the road points are the set of possible states S ; Forward-direct movement, redirection and shut down as the possible action sets A , a road map table indicating the connection among the road points as the Markov transition matrix P , and a reward scheme R . These four MDP components are sufficient to describe the framework and serve as moderating factors for the learning algorithm.

A comparable matrix, "Q" is added to reflect the memory of what our agent has learned through experience. The current state of the agent is represented by the rows of matrix Q , while the possible actions leading to the next state are represented by the columns (the links between the nodes). Because the agent begins with no knowledge, the matrix Q is set to zero.

Hence, the description of the components of the MDP in relation to the proposed system is as follows:

- The road network in the University community constitutes the state space S

S/N	Road Path	Notation	Connection
1	Engineering – Botanical	S ₁	S ₂ , S ₃
2	UBA -Vc's Lodge	S ₂	S ₁ , S ₅
3	Senate -Guest House	S ₃	S ₁ , S ₄
4	Law - Mechanical works Junction	S ₄	S ₃ , S ₆ , S ₁₅
5	UBA - Mechanical works Junction	S ₅	S ₂ , S ₆ , S ₁₅
6	Mechanical works Junction – Science Junction	S ₆	S ₂ , S ₁₅ , S ₈
7	Science Junction – Science Faculty	S ₇	S ₆ , S ₈
8	Science Junction – Medical Junction	S ₈	S ₆ , S ₇ , S ₁₈ , S ₉
9	Medical Junction – DLI Junction	S ₉	S ₈ , S ₁₈
10	DLI Junction – Honours Hostel Road	S ₁₀	S ₉ , S ₁₂ , S ₁₁ , S ₁₃
11	DLI Junction – AP Junction	S ₁₁	S ₁₃ , S ₁₄ , S ₁₇ , S ₉ , S ₁₂ , S ₁₀
12	DLI Junction – Second Gate	S ₁₂	S ₉ , S ₁₀ , S ₁₁ , S ₁₃
13	DLI Junction – Education Axis- First Gate	S ₁₃	S ₁₁ , S ₉ , S ₁₀ , S ₁₂
14	AP Junction – First Gate	S ₁₄	S ₁₇ , S ₁₁
15	Mechanical Works Junction – CITS Junction	S ₁₅	S ₁₆ , S ₅ , S ₆ , S ₁₇
16	CITS Junction – Ozoluwa	S ₁₆	S ₁₅ , S ₁₇
17	CITS Junction – AP Junction	S ₁₇	S ₁₆ , S ₁₅ , S ₁₄ , S ₁₁
18	Medical Junction – High Rise	S ₁₈	S ₈ , S ₉

- The possible actions and their corresponding rewards at any given state:

S/N	Action	Reward	States
1	Direct-Forward Movement	10	Any state in S
2	Redirection	30	Any state in S
3	Shut down	50	Goal state

- State Transition Matrix

S	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	S ₈	S ₉	S ₁₀	S ₁₁	S ₁₂	S ₁₃	S ₁₄	S ₁₅	S ₁₆	S ₁₇	S ₁₈
S ₁	F	T	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
S ₂	T	F	F	F	T	F	F	F	F	F	F	F	F	F	F	F	F	F
S ₃	T	F	F	T	F	F	F	F	F	F	F	F	F	F	F	F	F	F
S ₄	F	F	T	F	F	T	F	F	F	F	F	F	F	T	F	F	F	F
S ₅	F	T	F	F	F	T	F	F	F	F	F	F	F	T	F	F	F	F
S ₆	F	T	F	F	F	F	T	F	F	F	F	F	F	T	F	F	F	F
S ₇	F	F	F	F	F	T	F	T	F	F	F	F	F	F	F	F	F	F
S ₈	F	F	F	F	F	T	T	F	T	F	F	F	F	F	F	F	F	T
S ₉	F	F	F	F	F	F	F	T	F	F	F	F	F	F	F	F	F	T
S ₁₀	F	F	F	F	F	F	F	T	T	F	T	T	T	F	F	F	F	F
S ₁₁	F	F	F	F	F	F	F	F	T	T	F	T	T	T	F	F	T	F
S ₁₂	F	F	F	F	F	F	F	F	T	T	T	F	T	F	F	F	F	F
S ₁₃	F	F	F	F	F	F	F	F	T	T	T	T	F	F	F	F	F	F
S ₁₄	F	F	F	F	F	F	F	F	F	F	T	F	F	F	F	F	T	F
S ₁₅	F	F	F	F	T	T	F	F	F	F	F	F	F	F	F	T	T	F
S ₁₆	F	F	F	F	F	F	F	F	F	F	F	F	F	F	T	F	T	F
S ₁₇	F	F	F	F	F	F	F	F	F	F	T	F	F	T	T	T	F	F
S ₁₈	F	F	F	F	F	F	F	T	T	F	F	F	F	F	F	F	F	F

T indicates link between nodes and F Matrix Q represents the brain of the agent. Because the agent starts with no knowledge, the matrix Q is set to zero. Without an instructor, the virtual agent will learn by experience (unsupervised learning). The agent will explore from state to state until it reaches the goal or destination state.

Policy (π)

The learning policy for the vehicular agent is described in the recurrence formula below. By picking next states using the relevant schemes, the policy prescribes the appropriate conditions that will lead to an optimal value.

$\Pi(s) = (Q = 0 \ \& \ S \neq GS)$ Select state by Random Selection

$\Pi(s) = (Q > 0 \ \& \ S \neq GS)$ Select state by Qmax & Min(TD)

$\Pi(s) = (Q \geq 0 \ \& \ S = GS)$ Goal State(GS), Stop $\alpha = 0.5$

..... 1
 $Q(s, a) = R(s, a) + (\alpha * \text{Max } [Q(s, a)])$
 2

Traffic Density, $K = N/L$
 3

(Delay is the number of vehicles, N on the Road; Weight is length, L of the road)

Therefore equation (2) becomes

$Q(s, a) = R(s, a) + (\alpha * \text{Max } [Q(s, a)]) + TD$
 4

Where:

α = Learning rate of the agent as the activity continues

TD = Traffic density i.e. the delay when N number of vehicles are on a road path whose length is L.

R(s, a) = Reward got for an action a, when in state, S .

Q(s, a) = Q-value of the learning agent at state S for an action a . The equation is a modification to equation 2, the traffic density is added to the basic

indicates no link between nodes. Q-learning formula to acknowledge effect of traffic delay on the learning exercise.

The transition matrix is a table representing the roads network within Unilag community. The roads are represented as numbers beginning from 1 to 10 both horizontally and vertically. The intersection of one number against the other means there is a connection between both roads.

There are always two tables representing the traffic situation at any particular road point as shown in table 5 and table 6

Table 5 is found in other points or nodes in the University road network. It represents the current situation across the road points at that time i.e. the **weight**, the **delay** and the **associated traffic density**. The value for weight (road length in unit of distance) is any number between 1 and 10 while the value for delay (number of vehicles on the particular road) is randomly generated to present a dynamic scenario. The value is any number between 1 and 50.

Table 6 represents the possible states the learning agent or vehicle could move to. The selection of any state is determined by the given policy. The policy is that the road point with the highest Q-value should be selected.

$Q(\text{state, action}) = R(\text{state, action}) + \alpha * \text{Max } [Q(\text{next states, all actions})] + TD$

According to this formula, a value assigned to a specific element of matrix Q, is equal to the sum of the corresponding value in matrix R and the learning parameter Gamma, multiplied by the maximum value of Q for all possible actions in the next state plus the traffic density.

Below is a map out from the program. The agent is starting at state 1 and the destination is state 9. The agent can only traverse through the node with "true" which indicates connection between the nodes.

Analysis of Routing From Location: [Eng-Bot]

Road	Weight	Delay	Speed
Eng-Bot	6	36	6
UBA-VcLodge	5		26 5.2
Senate-GH	1		1 1
Law-Mech	3		18 6
UBA-Mech	3		15 5
Mech-Science	4		22 5.5
Science	5	28	5.6
Sci-Med	4	22	5.5
Med-DLI	5		29 5.8
DLI-Hon	4		20 5

DLI-AP	5	29	5.8
DLI-2ndgate	7	39	5.57
DLI_Edu	3	18	6
AP-1stgate	3	13	4.33
Mech-CITS	4	19	4.75
CITS-Ozo	6	34	5.67
CITS-AP	6	34	5.67
Med-HgRise	6	33	5.5

Table 5

State Selection From : Eng-Bot Analysis above

Position	Traffic Delay	Q-Value
UBA-VcLodge	5.2	10.2
Law-Mech	6	11

Table 6

State Selection From: Law-Mech Analysis above

Position	Traffic Delay	Q-Value
Senate-GH	1	6
Mech-Science	5.5	10.5
Mech-CITS	4.75	9.75

Analysis of Routing From Location: [Mech-Science]

Road	Weight	Delay	Speed
Eng-Bot	6	36	6
UBA-VcLodge	5	26	5.2
Senate-GH	1	1	1
Law-Mech	3	18	6
UBA-Mech	3	15	5
Mech-Science	4	22	5.5

.....

State Selection From : Mech-Science Analysis above

Position	Traffic Delay	Q-Value
UBA-VcLodge	5.2	10.2
Sci-Med	5.5	10.5
Mech-CITS	4.75	9.75

Analysis of Routing From Location: [Sci-Med]

Road	Weight	Delay	Speed
Eng-Bot	6	36	6
UBA-VcLodge	5	26	5.2
Senate-GH	1	1	1
Law-Mech	3	18	6
UBA-Mech	3	15	5
Mech-Science	4	22	5.5

Science 5 28 5.6

.....
State Selection From: Sci-Med Analysis above

Position	Traffic Delay	Q-Value
Mech-Science	5.5	10.5
Science	5.6	10.6
Med-DLI	5.8	10.8
Med-HgRise	5.5	10.5

Analysis of Routing From Location: [Med-DLI]

Road	Weight	Delay	Speed
Eng-Bot	5	28	5.6
UBA-VcLodge	8	46	5.75
Senate-GH	5	25	5
Law-Mech	5	28	5.6
UBA-Mech	1	3	3

.....

State Selection From : Med-DLI Analysis above

Position	Traffic Delay	Q-Value
Sci-Med	6	11
Med-HgRise	5.88	10.88

Analysis of Routing From Location: [Sci-Med]

Road	Weight	Delay	Speed
Eng-Bot	5	28	5.6
UBA-VcLodge	8	46	5.75
Senate-GH	5	25	5
Law-Mech	5	28	5.6
UBA-Mech	1	3	3
Mech-Science	5	27	5.4

.....

State Selection From: Sci-Med Analysis above

Position	Traffic Delay	Q-Value
Mech-Science	5.4	10.4
Science	5	10
Med-DLI	5.5	10.5
Med-HgRise	5.88	10.88

Analysis of Routing From Location: [Med-HgRise]

Road	Weight	Delay	Speed
Eng-Bot	5	28	5.6
UBA-VcLodge	8	46	5.75
Senate-GH	5	25	5
Law-Mech	5	28	5.6

UBA-Mech	1	3	3
Mech-Science	5	27	5.4
Science	3	15	5
Sci-Med	8	48	6
Med-DLI	4	22	5.5

.....

State Selection From : Med-HgRise Analysis above

Position	Traffic Delay	Q-Value
Sci-Med	6	11
Med-DLI	5.5	10.5

Routing Summary

From	To
[Eng-Bot]	[Law-Mech]
[Law-Mech]	[Mech-Science]
[Mech-Science]	[Sci-Med]
[Med-DLI]	[Sci-Med]
[Sci-Med]	[Med-HgRise]
[Med-HgRise]	[Sci-Med]

This is a single traversal of the learning agent. The agent stops here since Sci-Med is the destination still maintaining the highest reward i.e. Q-Value.

IV. CONCLUSION

This research framework can be fully commercialized or made available as accessible software to the College or University community even to Lagos State at large. It will significantly enhance the navigation judgments of commuters when there is heavy traffic load on usually optimal paths to take other paths that are not as optimal but having lower traffic costs to the same destination. It can also be proposed to government and private agencies that may be interested in developing the software into a fully commercial application in managing the teeming national road network. This research work is a veritable solution to traffic delay and congestion; hence a lot is still to be desired in terms of further research to enhance the solution for route optimality in Nigeria.

REFERENCES

[1]. Allan, M. de S., Celso ARL B., Roberto, S. Y, Erick, A. D., Edmundo, R.M.M. & Leandro, A. (2017). Traffic management systems: A classification, review, challenges, and future perspectives. International Journal of Distributed Sensor Networks 2017, Vol. 13(4). The Author(s) 2017 DOI: 10.1177/1550147716683612. journals.sagepub.com/home/ijdsn

[2]. Chapman, D. (1991). The Concrete-Situated Approach, Vision, Instruction and Action. MIT Press.

[3]. Chavhan, S., & Venkataram, P. (2019). Prediction based traffic management in a metropolitan area. Journal of Traffic and Transportation Engineering (English Edition), <https://doi.org/10.1016/j.jtte.2018.05.003>

[4]. Dipak, G. (2019). ICT based Smart Traffic Management System “iSMART” for Smart Cities. International Journal of Recent Technology and Engineering (IJRTE), 8 (3). ISSN: 2277-3878, September 2019. Retrieval Number: C5137098319/2019@BEIESP DOI:10.35940/ijrte.C5137.098319, Published By: Blue Eyes Intelligence Engineering & Sciences Publication.



[5]. Gaskett, C. (2002). Q-Learning for Robot Control. Ph.D. Dissertation, The Australian National University.

[6]. Gerald, T. (1995). Temporal Difference Learning and TD-Gammon. Communications of the ACM, Vol. 38, No. 3, pp. 58 – 68.

[7]. Ian D. Kelly (1997). The Development of Shared Experience Learning in a Group of Mobile Robots. University of Reading, Department of Cybernetics, April.

- [8]. Kok, A. et al. (2010). Vehicle routing under time-dependent travel times: the impact of congestion avoidance, Operational Methods for Production and Logistics. University of Twente.
- [9]. Nambajemariya, F. & Wang, Y.S. (2021). Excavation of the Internet of Things in Urban Areas Based on an Intelligent Transportation Management System. Advances in Internet of Things , 11, 113-122. <https://doi.org/10.4236/ait.2021.113008>
- [10]. Nolfi, S. & Parisi, D. (1996). Learning to Adapt to Changing Environments in Evolving Neural Networks. Tech. Report 95-15, Inst. Psychology, NRC, Rome.
- [11]. Richard, S. S. & Andrew, G. B. (1998). Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA, A Bradford Book.
- [12]. Sitkoff, N. (1996). On Motivation and the Development of Purpose. <http://citeseer.nj.nec.com/384564.html>.
- [13]. Sloman, A. & Logan, B. (2000). Evolvable Architectures For Human-Like Minds. Presented at 13th Toyota Conference on "Affective Minds".
- [14]. Sutton, R. & Barto, A. (2002). Reinforcement Learning: An Introduction. MIT Press, pp. 61-65.
- [15]. Watkins, C. & Dayan, P. (1992). Technical note: Q-Learning. Machine Learning 8, Kluwer Academic Publishers, Boston, pp 279-292.
- [16]. William, D. S. (2002). Making Reinforcement Learning Work on Real Robots. Ph.D. Dissertation, Brown University, Providence, Rhode Island, United States.

Biography of Authors

	<p>OLATUNJI, Michael Abiodun is currently a graduate assistant as system analyst at the University of Lagos and an Adjunct Lecturer at Yaba College of Technology. He received the B.Sc. and M.Sc. degrees from University of Lagos, Nigeria in Computer Sciences in 2014 and 2021, respectively. His research interests are in the areas of Machine Learning, Data Science and Internet of Things, with a current focus on Water work inspection using Ant Colony Optimization. He is a member of the ISOC and NCS.</p>
	<p>LAWAL Olawale Nasiru is currently a PhD student at Morgan State University, Baltimore, U.S.A. He is a Principal Lecturer at Yaba College of Technology, Lagos, Nigeria. He obtained B.Tech (Mathematics & Computer Science) from Federal University of Technology, Minna, Nigeria, and M.Sc. (Computer Science) from University of Lagos, Nigeria, in 1998 and 2003 respectively. His research interest is in the area of Artificial Intelligence, with a focus on Evolutionary Computation Algorithms. He is a member of the Computer Professional Registration Council of Nigeria (CPN) and Nigeria Computer Society (NCS).</p>