Proceedings of the 25th Chinese Control Conference
7-11 August, 2006, Harbin, Heilongjiang

The Research of Elevator Group Dynamic Scheduling During Up-peak Traffic Based on POMDP

Zong Qun, Sun Zheng-ya

Academy of Electrical Engineering & Automation ,Tianjin University, Tianjin 300072
E-mail: zongqun@tju.edu.cn, zhengya1982@163.com

Abstract: Due to the existence of the uncertainty of the traffic flows in the problem of elevator group control scheduling, markov decision processes can't deal with the model of the elevator group control scheduling well. The model based on partially observable markov decision processes is studied to solve this problem. With application of the feedforward neural network, which is integrated into the linear Q-learning to construct the whole algorithm for elevator group scheduling, the optimal or nearly optimal policies are obtained gradually. Compared with previous algorithms, this algorithm greatly improves the adaptability to the up-peak traffic flow.

Key Words: reinforcement learning, POMDP, linear Q-learning, elevator group control scheduling

1. INTRODUCTION

The problem of elevator group scheduling has been researched extensively due to its high practical significance. Some stochastic models like Markov decision process (MDP) are used to model the elevator group system [1] [2]. In the MDP framework, it is assumed that there is never any uncertainty about the system's current state—it has complete and perfect perceptual abilities. But the existence of the uncertainty about traffic flow makes this representation of the elevator system model not quite well. The MDP-based elevator system performance degrades fast when the uncertainty about traffic flow increases. Because the hypothesis that the system is fully observable is hardly tenable in the practical world [3], we propose a novel elevator group control scheduling model based on the partially observable Markov decision process (POMDP). Our POMDP-based model can handle the uncertainty in the traffic flow. We use hidden system states as the state set, i.e. the number of the passengers in the elevator system. We approach the model using linear Q-learning algorithm to find an approximation to the optimal policy in the thesis. The simulation experiments are done in the virtual environment for elevator group control during the up-peak traffic flow. The experimental results demonstrate such a programming model that reduces the passenger staying time has good adaptability to the up-peak traffic flow.

1.1. Instructions for Authors

Zong Qun (1961-), a professor at Tianjin University and a member of the International Association of Elevator Engineers (IAEE). His research interests include computer and intelligent control system, intelligent elevator group control system.

Sun Zheng-ya (1982-), a postgraduate student at Tianjin University. Her research interests are reinforcement learning and optimization scheduling algorithms.

IEEE Catalog Number: 06EX1310

This work is supported by National Natural Science Foundation of P.R. China under Grant 60574055 and Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20050056037.

2. PRELIMINARIES

2.1. POMDP

A partially observable Markov decision process can be described as a 6-tuple: $\{S, A, T, R, \Omega, O\}$[4], where S is a finite set of states of the world; A is a finite set of actions; $R: S \times A \to R$ is the reward function; $T: S \times A \to \prod(S)$ is the state-transition function; Ω is a finite set of observations; $O: S \times A \to \prod(\Omega)$ is the observation function, which gives, for each action and resulting state, a probability distribution over possible observations.

The agent's goal is to choose a policy π in order to maximize the expected discounted sum of future reward. In the finite-horizon case, we use $V_t^*(b)$ to represent the optimal expected discounted sum of future reward for starting in belief state b and executing a non-stationary t-step policy. Clearly,

$$V_{\pi,1}^*(b) = \max_a \left[\rho(b,a)\right] = \max_a \left[\sum_i b(s_i) R(s_i,a)\right] \quad (1)$$

where $\rho(b,a)$ denotes the expected immediate reward gained by the agent for taking action a in belief state b. Then $V_{\pi,t}^*(b)$ is defined inductively as

$$V_{\pi,t}^*(b) = \max_a \left[\rho(b,a) + \gamma \sum_{o \in \Omega} \Pr(o|b,a) V_{\pi,t-1}^*(SE(b,a,o)) \right] \quad (2)$$

where $SE(b,a,o)$ is the state estimation function, which has its output the new belief state b'.

$$SE(b,a,o)(s') = O(s',a,o) \sum_{s \in S} T(s,a,s') b(s) / \Pr(o|a,b) \quad (3)$$

The denominator $\Pr(o|a,b)$ can be treated as a normalizing factor, independent of s', which causes

$\sum_{s\in S} b(s) = 1$.

In the infinite-horizon discounted case, we write $V_\pi^*(b)$ for the optimal expected discounted sum of future reward for stating in belief state b and executing a stationary policy π. It is recursively defined by

$$V_\pi^*(b) = \max_a \left[\rho(b,a) + \gamma \sum_{o\in\Omega} \Pr(o|b,a) V_\pi^*(SE(b,a,o)) \right] \quad (4)$$

Value iteration serves as the basis for computing the optimal expected discounted sum of future reward for POMDP [5]. Reinforcement learning, as a method of learning optimal policy from interaction with the environment, is suitable for dynamic optimization problems [6] such as elevator group control scheduling. We use an extension of Q-learning, which is linear Q-learning, to learn approximate Q functions for the POMDP model.

2.2. Elevator Group Control Scheduling during Up-peak Traffic

During the up-peak traffic flow pattern, the directions of passenger flows are mostly or entirely upward, i.e. most or all of the passengers arrive at the building's hall and go upwards. Hundreds of simulation experiments are done with multi-agent coordination non-zoning algorithm, half-zoning algorithm and static zoning algorithm in the virtual environment for the elevator group system with 16 floors and 4 elevators. The conclusions are [7]:

1) While the passenger arrival rate is low, non-zoning algorithm is preferable.

2) While the passenger arrival rate is high, static zoning algorithm is preferable.

3) While the passenger arrival rate is medium, half-zoning algorithm is preferable.

When choosing the corresponding method according to the passenger arrival rate, we can attain better scheduling results. In the next section, the framework of POMDP formulates the problem of elevator group control scheduling and then the elements in the model are identified.

3. THE POMDP MODEL OF ELEVATOR GROUP CONTROL SCHEDULING

The elevator group control scheduling system can be considered as a discrete-time event system. To utilize the algorithm for solving POMDP, it is necessary to establish the POMDP model for elevator group scheduling. Here, we suppose that there is a finite state set, a finite set of actions and a finite set of observations. We obtain the state transition probabilities, observation probabilities and immediate rewards through a number of simulation tests.

1) States: We define s as the number of passengers in the elevator system (namely the sum of the number of passengers being served and waiting to be served) at the beginning of every minute. For the sake of analyzing and calculating conveniently, we divide the number of passengers in the elevator system into 5 classes of states according to passenger arrival rate [7]. The meaning of every state symbol is shown in Table1.

Table1. The State Symbol and the Corresponding Number of Passengers in the Elevator System

State symbol	The number of passengers
0	0～9
1	10～19
2	20～26
3	27～35
4	36～∞

Clearly, the set of states can be described as $S = \{0,1,2,3,4\}$.

2) Actions: We apply a to denote the elevator scheduling method chosen every other minute. The meaning of every action symbol is shown in Table2.

Table2. The Action Symbol and the Corresponding Elevator Scheduling Method

Action symbol	Scheduling method
0	Non-zoning method
1	Half-zoning method
2	Static zoning method

Obviously, the set of actions can be described as $A = \{0,1,2\}$. The scheduling method is determined at the beginning of every minute. The elevator running chooses the last minute scheduling method until it returns to the hall.

3) Reward: In this paper, we take the number of passengers that the elevator system serves in this minute into account.

The number of served passengers in the minute:

$$R_s = \sum_p r_p \quad (5)$$

where p denotes the index number of the passenger and we apply r_p as the variable for judging whether passenger p is served in this minute.

$$r_p = \begin{cases} 1 & t-1 < t_{pr} \le t \\ 0 & otherwise \end{cases} \quad (6)$$

where t_{pr} denotes the riding time of passenger p .

The expected number of served passengers over belief state b in the minute:

$$\rho_t(b) = \sum_{s\in S} R_s \cdot b(s) \quad (7)$$

4) Value function: In this paper, we formulate the elevator group scheduling into infinite-horizon discounted POMDP and define the value function according to the cost function.

$$V_\pi(b) = E(\rho_{t+1} + \gamma\rho_{t+2} + \gamma^2\rho_{t+3} + \ldots | b_t = b, \pi)$$

$$= E(\sum_{i=0}^{\infty} \gamma^i \rho_{t+i+1} \mid b_t = b, \pi) \quad (8)$$

5) Observations and observation function: The observations derive mainly from the recognition results of the passenger number in the elevator system. The set of observations can be described as $\Omega = \{0,1,2,3,4\}$. The meaning of each observation symbol is the same as that of the state symbol.

The most complicated part of the POMDP model for the elevator group scheduling system is determination of observation function. Even if the system executes the same action for the same state, it may still lead to different observations. For example, the system dispatches the elevators in the same way for the same state while the traffic flow statistical recognition system may give us a different observation result. Therefore, the designer should acquaint with the performance of the traffic flow statistical recognition system when designing the observation probability function.

The variable at time $t+1$ is dependent only on the variable at time t on the basis of the Markov property. As a result, for every action, we use 2-stage temporal Bayes network to determine the relationship between states and observations. According to the networks we can get the observation probability function finally.

Based on the definition of the system POMDP model, it is obvious that: $|S| = 5$, $|A| = 3$, $|\Omega| = 5$.

And then we can view elevator group scheduling system as an extremely small POMDP model. Linear Q-learning can produce nearly optimal policies for the extremely small POMDP model [8].

4. ELEVATOR GROUP SCHEDULING SCHEME BASED ON POMDP

In order to illustrate the effectiveness of the proposed scheduling approach, we should first utilize the virtual elevator environment of our lab to produce two up-peak traffic flows (TR) whose intensities are different. The traffic flow lasts for 15 minutes within which 150 passengers ask for service. TR1 is the pure up-peak traffic mode and TR2 is less intensive up-peak traffic mode, which are shown in figure 1.

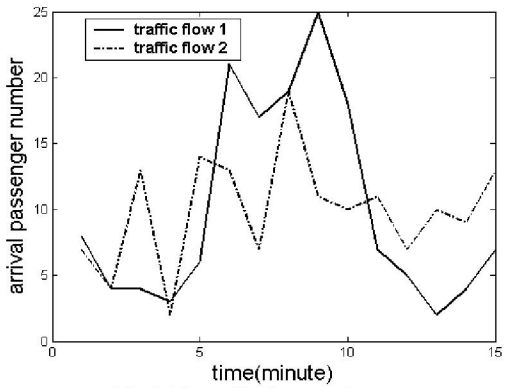

Fig 1. The traffic flow graph.

Next we combine linear Q-learning with the neural network to approach the POMDP model of elevator group control scheduling system. The update rule for linear Q-learning is as follows:

$$\Delta q_a(s) = \alpha b(s)(r + \gamma \max_{a'} Q_{a'}(b') - q_a(s) \cdot b) \quad (9)$$

where $q_a(s)$ denotes the state-action value function, α is a learning rate, $b(s)$ is the probability in state s, r denotes the reward received, and γ is the discount factor.

We approximate the belief state-action value function $Q_a(b)$ for each action according to formula (10).

$$Q_a(b) = q_a \cdot b \quad (10)$$

The update rule is evaluated for every $s \in S$ each time the system makes a state transition. Linear Q-learning adjusts the components of q_a to match the coefficients of the linear function that predicts the Q values.

The complete algorithm for the elevator group control scheduling based on POMDP is as follows:

① Initialize $q_0(s,a), \forall s \in S, \forall a \in A$;

② At time t, obtain the belief state $b_t(s)$ from the state estimator, $\forall s \in S$;

③ At time t, Get the $q_a(s)$ with the equation:

$$\Delta q_a(s) = \alpha b_{t-1}(s)(r(s,a) + \gamma \max_{a'} Q_{a'}(b_t) - q_a(s) \cdot b_{t-1}(s)),$$
$$\forall s \in S, \forall a \in A;$$

④ Calculate the $Q_a(b)$ based on the updated $q_a(s)$ with the equation:

$$Q_a(b) = \sum_{s \in S} q_a(s) \cdot b_t(s) \qquad \forall a \in A;$$

⑤ Define $\hat{Q}(b) = \max_{a \in A} Q_a(b)$;

⑥ Choose the dispatch method a_t, according to $\hat{Q}$ for belief state b_t;

⑦ $b_{t-1} \leftarrow b_t, a_{t-1} \leftarrow a_t$;

⑧ return to ② until the task is finished.

5. SIMULATIONS AND ANALYSIS

In this section, we carry out simulation tests of the algorithm mentioned above. The parameters of the elevators and the building are chosen in the virtual elevator environment.

1) The algorithm parameters:

$$\gamma = 0.9, \quad \eta = 0.02, \quad a = 1.$$

where γ is the discount factor, η is the learning constant in the delta rule, a is the parameter of the active function in the hidden layer of the neural network.

The learning rate based on linear Q-learning is attenuating with the time. The attenuation rule is as follows:

$$\alpha_t = k^t \alpha_{t-1} = k^t \alpha_0 \quad (11)$$

where α_0 is the initialized learning rate, and k is the attenuation factor.

In the simulation, we have $\alpha_0 = 0.99$, $k = 0.999$.

2) The scheduling algorithms:

LQ: linear Q-learning algorithm based on POMDP

SZ: static zoning algorithm
MA: multi-agent coordination non-zoning algorithm
3) The performance:
AWT: average waiting time of passengers
RWLT: rate of waiting longer time of passengers
ATT: average travel time of passengers
NSS: number of start-up and stop
ACD: average crowding degree of passengers

Table3. Results of Simulation

Performance: Algorithm: Traffic flow:		AWT	RWLT	ATT	NSS	ACD
TR1	LQ	**40.35**	**23.33**	**44.20**	**146**	**5.37**
	SZ	46.63	30.00	43.21	143	5.21
	MA	31.65	18.67	57.88	151	6.87
TR2	LQ	**28.39**	**14.67**	**38.49**	**173**	**4.05**
	SZ	34.11	16.00	35.54	168	3.55
	MA	14.05	3.33	48.75	179	5.14

It is shown from the results of simulation in Table 3 that the comprehensive performance of the linear Q-learning algorithm based on POMDP is excellent. Formulating the problem of elevator group control scheduling into the framework of POMDP is closer to the real world.

The static zoning algorithm regulates the floors that every elevator should respond, which distributes effectively the passengers to different elevators so as to decrease the number of start-up and stop clearly, and at the same time reduce the average travel time of passengers. With the passenger arrival rate cutting down during up-peak time, the algorithm, on the other hand, will increase the average waiting time of passengers as a result of prolonging the average travel time of passengers due to one elevator serves one zone, which is shown in Fig. 2.

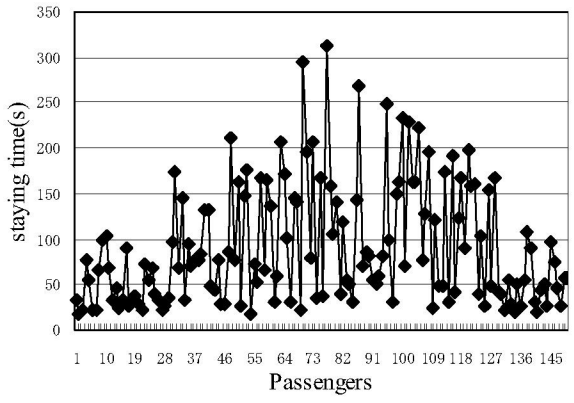

Fig 2. The staying time at SZ in the case of the pattern of TR1.

Multi-agent coordination non-zoning algorithm which makes the service for passengers more balance can ensure that the hall call be responded by the most suitable elevator via coordination among of agents. In consequence, the algorithm makes for decreasing the average waiting time of passengers. On the other hand, it increases the number of start-up and stop and prolongs the average travel time of passengers simultaneity due to every elevator serving all the floors of the building. With the passenger arrival rate cutting down during up-peak time, the algorithm, on the other hand, will decrease the average waiting time of passengers remarkably as a result of shortening the average travel time of passengers, which is shown in Fig. 3.

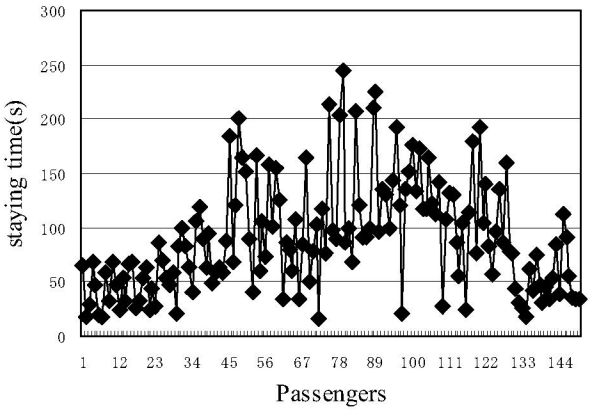

Fig 3. The staying time at MA in the case of the pattern of TR1.

Linear Q-learning algorithm based on POMDP model optimize the dispatching method step by step through learning in respect to the number of passengers in the elevator system every minute. Although the performances of AWT and ATT are not optimum, we can obtain the shortest average staying time of passengers (AST=AWT+ATT), namely the comprehensive performance is optimum, which is shown in Fig. 4, namely the comprehensive performance is optimum, which is shown in Fig. 4.

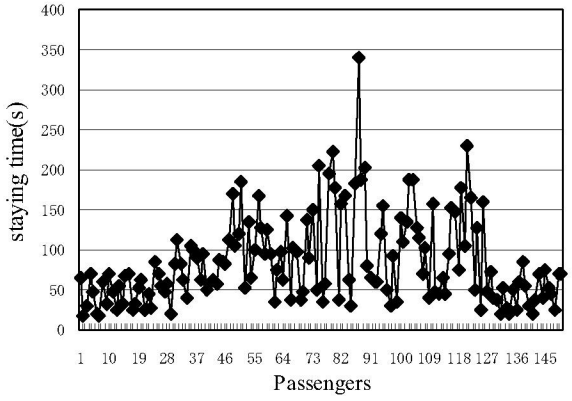

Fig 4. The staying time at LQ in the case of the pattern of TR1.

6. CONCLUSION

In this paper, a new elevator control group scheduling approach with respect to stochastic up-peak traffic flow is presented. We formulate the problem into the POMDP

model in which we consider not only the uncertainty of actions, but also the uncertainty of states. When linear Q-learning is applied to solve POMDP, feed-forward neural network is used to handle the generalization of value function, whose goal is to find Q functions that can be used to produce a good policy. Finally, the algorithm is simulated and compared with other scheduling algorithms. The results demonstrate good learning ability, good performance of staying time and good adaptability during up-peak time.

REFERENCES

[1] Zong Qun, Song Chao-feng, Xing Guan-sheng. A Study of Elevator Dynamic Scheduling Policy Based on Reinforcement Learning. Elevator World, January 2006.

[2] Crites R. Barto A. Elevator group control using multiple reinforcement learning agents. Machine Learning, 33, pp: 235-262, 1998.

[3] Zhang Bo. Planning and Acting under uncertainty: theory and application. Ph.D. Thesis. University of Science and Technology of China, 2001.

[4] Leslie Pack Kaelbling, Michael L. Littman, Anthony R. D. Cassandra. Planning and acting in partially observable stochastic domains. Artificial Intelligence, Volume 101, pp. 99-134, 1998.

[5] Anthony R. Cassandra. Exact and Approximate Algorithms for Partially Observable Markov Decision Processes. Ph.D. Thesis. Brown University, Department of Computer Science, Providence, RI, 1998.

[6] Gao Yang, Chen Shi-fu, Lu Xin. Research on Reinforcement Learning Technology: A Review. Acta automatica Sinica, Vol. 30, No.1 2004.

[7] Zong Qun, Ya Shu-hong, Wang Zhen-shi. Elevator Group Control Dispatching Method During Up-Peak Traffic Based on Queuing Theory. Systems Engineering And Electronics, Vol.25, No.6, 2003.

[8] Michael Littman, Anthony Cassandra, Leslie Kaelbling. Learning policies for partially observable environments: Scaling up. Proceedings of the Twelfth International Conference on Machine Learning, pp. 362--370, 1995.