Adaptively Finding Optimal Routes under Principles of Spatial Cognition

A Hierarchical Reinforcement Learning Approach

Weifeng Zhao[1, 2] and Qin Zhang[1]
[1]College of Geology Engineering and Geomatics, Chang'an University, Xi'an, China
[2]State Key Laboratory of Geo-information Engineering, Xi'an, China

***Abstract*—Way finding research has paid much attention to the selection of optimal routes under principles of spatial cognition. However, the commonly employed implemental approaches suffer inevitably from the contradictions between personalized network modelling and network data sharing. This paper presents one kind of interactive route selection approach based on hierarchical reinforcement learning. In this approach, a complete network model is unnecessary, but the environmental states are automatically perceived by the agent and then mapped into the reward function defining the goal of cognitively optimal routes. The optimal routes corresponding to the policies with maximal cumulative rewards can be found through a two-stage learning process including a pre-learning stage and a real-time learning one. Our experimental results show that the proposed approach learns effectively enough for real-time route selection and ensures found routes close to global optimal ones.**

Keywords-spatial cognition; route selection; hierarchical reinforcement learning; pre-learning; real-time learning

I. INTRODUCTION

In recent years, several approaches have been presented that cover the selection of optimal routes under principles of spatial cognition, such as simplest routes [1], clearest routes [2], most reliable routes [3, 4] and easy-to-describe routes [5, 6], which respect for human principles of route wayfinding and direction giving and the cognitive complexity of traveling through a road network [7].

In those approaches, landmarks, intersection structures and turning styles are usually taken into account for constructing cognitive map of environment and reducing human cognitive load of route following. Their implementations all follow the same process that firstly translating all kinds of weighted cognitive criteria into edge costs and turn penalties of a network model and then adopting modified Dijkstra or A* algorithms to search the routes of least cost from origin to destination. However, as the diversity of user cognitive preferences leads to different concerned landmarks and different weight coefficients of cognitive criteria, personalized network models need be generated to meet different user. Therefore, such model-driven route selection approach could bring big trouble to the sharing and maintenance of road network data between different users.

This paper presents a model-independent interactive route selection approach using hierarchical reinforcement learning (HRL). In this learning process, public road network model defines the topological structure of streets, landmark sets contain personalized landmarks with regard to user cognitive preferences, the agents of HRL find the optimal routes satisfying user-defined reward conditions during interaction with the environment including public road network model and personalized landmark sets. Cuayahuitl et al. [8] tried to generate adaptive route instructions with MAXQ based hierarchical reinforcement learning in the indoor environment. Nevertheless, the applied manual task hierarchization and value function decomposition are unsuitable to urban road environments with large state space and irregular action space.

In order to promote the learning efficiency in large-scale urban road network, we propose a two-stage approach to adaptively learn optimal routes employing network Voronoi diagram based hierarchical reinforcement learning (NVD-HRL). In the first pre-learning stage, we automatically find multilayer subgoals in the road network, and construct hierarchical subtasks (generally called options) on the basis of network Voronoi diagrams generated by the multilayer subgoals. In the second real-time learning stage, off-policy intra-option Q-learning is adopted to update the estimated Q-values of available state-action pairs, and then the optimal route is traced according to Q-values after convergence. After experiments on Wuchang district of Wuhan city, we demonstrate that the NVD-HRL is efficient enough at finding the optimal routes, and ensures that most of the selected routes achieve or close to global optimality.

II. PROPOSED METHOD

On the basis of option framework, we presents a network Voronoi diagram based hierarchical reinforcement learning (NVD-HRL) approach which is divided into pre-learning stage and real-time learning stage. In this approach, a reward function is firstly defined to generate immediate rewards responded to turns taken by the agent at every intersection, and then the two learning stages are introduced in detail successively for finding optimal route.

A. Reward Function

A reward function defines the goal in a reinforcement learning problem [9]. Roughly speaking, it maps each perceived state-action pair of the environment to a single number, a reward, indicating the intrinsic desirability of that state. As such, a reinforcement learning agent's sole objective is to maximize the total reward it receives in the long run, while the reward function defines the immediate features of the problem faced by the agent.

In this routing application, every state-action pair corresponds to a turn taken at an intersection. As route selection criteria need reflect the cognitive efforts of taking a turn, they can be quantified and then traded off to indicate which turn is good in an immediate sense. In the literature of factors based on cognitive and perceptual aspects influencing human route choice, referable landmarks, turning styles, intersection structures, road grades and route length are most commonly used route selection criteria under principles of spatial cognition [1-7]. Therefore, the expected reward function can be expressed as the following linear equation to make a system learn to satisfy the goal of finding optimal routes under principles of spatial cognition.

$$R = \omega 1 \times R1 + \omega 2 \times R2 + \omega 3 \times R3 + \omega 4 \times R4 + \omega 5 \times R5 \quad (1)$$

Where R1 to R5 represent the quantified rewards respectively obtained from every turn in consideration of the above five route selection criteria, and ω1 to ω5 are their weight coefficients when the trade-off is made.

B. Pre-Learning Process

The tasks of pre-learning include subgoal identification and option construction. Subgoals are states that are believed to process some “bottleneck” importance and are worthwhile reaching. Options are treated as subtasks that efficiently take the agent to reach these subgoals.

We adopt betweenness centrality measure in this paper for identifying bottleneck nodes in road network crucial to develop useful options [10, 11]. Betweenness centrality of a node u, denoted by BC (u), is defined as the frequency that a node lies on an optimal route connecting two distinct nodes:

$$BC(U) = \sum_{u \neq s \neq t \in V} \frac{\sigma_{st}(u)}{\sigma_{st}} \quad (2)$$

Where σst is the number of optimal route from node s to node t, σst(u) is the number of such routes that pass through node u. As algorithm efficiency is not critical to pre-learning, Q-learning could be applied to find the optimal route in line with above route selection criteria under principles of spatial cognition between any pair of nodes. In Q-learning process, the transitions among states are constrained with the topological relations among nodes in road network, and the immediate reward of every state-action pair is obtained in accordance with equation (1). When every node in road network is scored based on the betweenness centrality measure, the top most scored nodes are ranked into several levels considered as hierarchical subgoals.

After the identification of subgoals, we utilize network Voronoi diagrams (NVD) [12] to construct hierarchical options. Directed NVDs would be hierarchically generated, with the hierarchical subgoals as generators at different levels and cumulative rewards of the routes between arbitrary nodes and generators as inward/outward distances defined on the road network. In an inward or outward NVD, we call the nodes belonging to more than one Voronoi subnetworks or directly connecting to nodes in other Voronoi subnetworks as bridge nodes. Figure 1 shows an example of inward NVD with two Voronoi subnetworks, which are {v1, v2, v3, v4, v5, v6, v7, v8, v9, v10, v11, v12} and {v10, v11, v12, v12, v14, v15, v16, v17, v18, v19, v20, v21}.

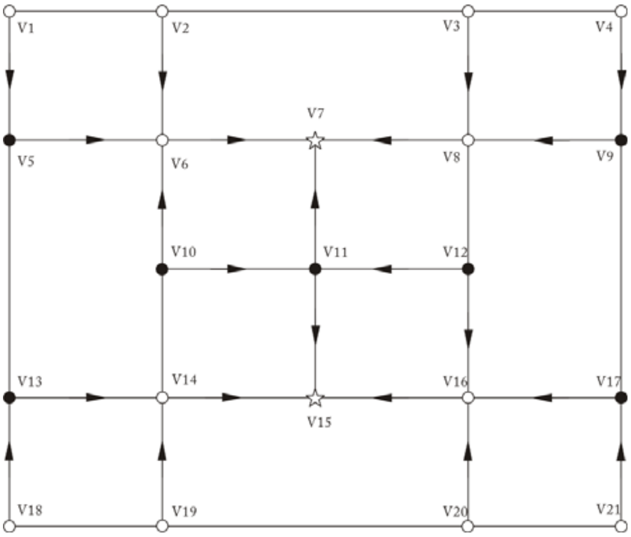

FIGURE I. AN EXAMPLE OF NETWORK VORONOI DIAGRAM

Options built upon inward and outward NVDs are respectively called inward options and outward options, which are constructed as following rules. The initiation sets of inward and outward options include the non-generator and generator nodes in the corresponding Voronoi subnetwork respectively. The inward options terminate with probability one when the generators of their corresponding Voronoi subnetworks are reached and with probability zero at other nodes, while the outward options may terminate at any nodes in their corresponding Voronoi subnetworks except the generators. We define two kinds of policies for each option, called internal policy and bridge policy, for efficiently executing other options when an option terminated. The internal policies of inward and outward options are defined as sequences of turning decisions from initial states to terminal states. The bridge policy of an inward or outward option defines a sequence of turning decisions from one Voronoi subnetwork adjacent to another through a bridge node.

C. Real-Time Learning Process

The main task of real-time learning is updating the estimated Q-values of state-action pairs constrained by hierarchical options with a reinforcement learning algorithm. We apply the same update rule from Q-learning equation (as in [9]) for each transition considered, but perform these updates from goal state to initial state. In this case, after every full episode the agent will have updated its Q-value estimate for every transition along the route it took to the goal.

As the internal policies always start and terminate at subgoal nodes, the subgoal nodes passed in any episode need not be considered for executing other options. Consequently, the available lower level inward options are executed to generate training examples from the calling nodes to their subgoal nodes under corresponding internal policies, while the available lower level outward options are executed to generate training examples from their subgoals to the calling nodes under corresponding internal policies. In this process, the estimated Q-values of all the training examples are updated according to Intra Option Q-Learning [13, 14].

When there is not any Q-value updated in a predefined number of successive episodes, it can be thought that convergence has been reached in this learning process. Thus, the expected optimal route can be traced by way of taking the turns with maximal Q-values at any following node from the origin node and until the destination node reached.

III. EXPERIMENTS

We carried out route selection experiments on the Wuchang district of the city of Wuhan. In the employed navigation electronic map, the road network for this district comprises 5639 nodes and 40214 turns at intersections. In addition, 224 landmarks, being ranked into 4 levels according to their significance, were extracted from the POI data of this area according to the approach proposed in literature [15].

TABLE I. PARAMETER SETTING FOR IMMEDIATE REWARDS

Criteria	Situations	Rewards
Landmark	Existing any level 1 landmark at an intersection	-10
	Existing any level 2 landmark at an intersection	-20
	Existing any level 3 landmark at an intersection	-40
	Existing any level 4 landmark at an intersection	-80
	No landmark at an intersection	-100
Turning style	Turning distinctly straight	-10
	Turning distinctly right	-40
	Turning distinctly left	-60
	Turning distinctly back	-80
	Passing through a roundabout	-50
	Taking an ambiguous turn	-100
Road grade	Highway	-10
	Urban highway	-20
	National road	-30
	Main prefecture road	-60
	General prefecture road	-80
	Other road	-100
Intersection structure	The number of competing out ways	Normalized into range (0, -100]
Segment length	Actual length of the road segment	Normalized into range (0, -100]

A. Experimental Setting

The five aspects of immediate rewards that a user can obtain from every turn at an intersection, as defined in equation (1), were all quantified into normalized values in range (0, -100], to ensure the cost of the route receiving maximal cumulative rewards is minimal. The set values shown in Table 1 mean that the larger a value is, the less cognitive effort need to be made. In addition, the five weight coefficients in equation (1) were all set as 0.2 in current experiments.

The learning parameters used by the algorithms were the same for both Q-learning and our NVD-HRL approaches. The learning rate parameter α decays from 1 to 0 according to $\alpha = 100/(100 + \tau)$, where τ represents the past episodes in current state. The discount factor $\gamma = 1$ makes future rewards as valuable as immediate rewards. The action selection strategy employed ε-Greedy with $\varepsilon = 0.1$, and initial Q-values of 0.

B. Experimental Results

After the computation of between's centrality of every node, we treated the top scored 282, 564 and 1410 nodes as three levels of subgoals, respectively constituting about 5%, 10% and 25% of the total nodes in the network. Then, there levels of inward and outward options were constructed upon the inward and outward NVDs generated from these levels of subgoals.

We randomly selected 500 different origin-destination pairs for finding the optimal routes under principles of spatial cognition with Q-learning and our NVD-HRL approaches respectively. The proposed approaches are both coded with Visual C++ 2010 and ran on a 2.5 GHz Dual-Core Pentium with 2-GB RAM. The learning of each OD was executed over five independent trials with different approaches, and related computational results were averaged for the comparison between the two approaches. In addition, we thought the convergence of learning could be reached if no Q-value was updated in 100 successive episodes.

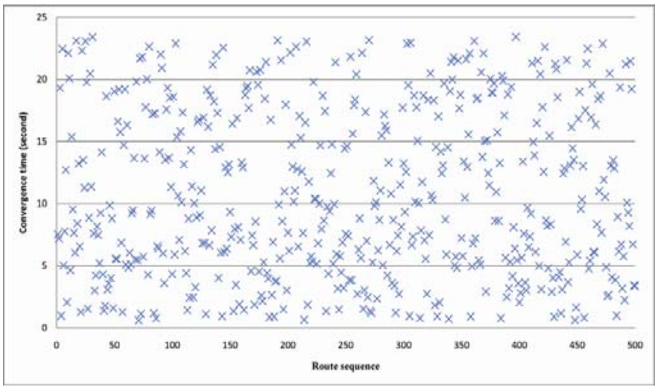

FIGURE II. CONVERGENCE TIME OF Q-LEARNING

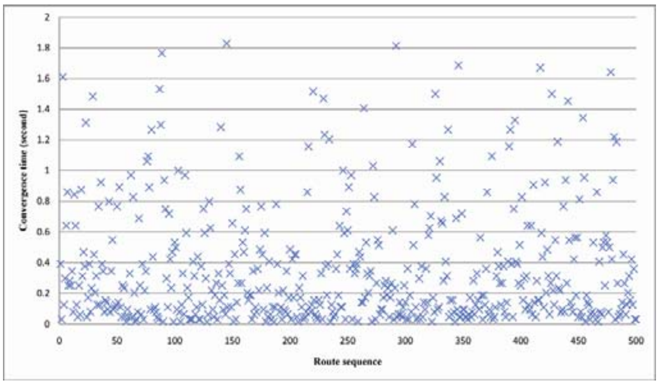

FIGURE III. CONVERGENCE TIME OF OUR NVD-HRL

Figure 2 and Figure 3 respectively show the convergence time for learning the optimal route between any pair of origin and destination with the approaches of Q-learning and NVD-HRL. Obviously, our NVD-HRL performs much more efficiently than the flat Q-learning in this route selection case, and basically meets the efficiency requirement of real-time route selection.

As the optimal routes learned by Q-learning are always the globally optimal ones, we can define the optimal degree of any optimal route learned by the NVD-HRL as the ratio of cumulative rewards of NVD-HRL to Q-learning learned optimal route. It can be observed from Figure 4 that more than 80% of the 500 optimal routes learned by our NVD-HRL approach are globally optimal, and the cumulative rewards of more than 95% of these routes are 10% larger than the corresponding globally optimal route.

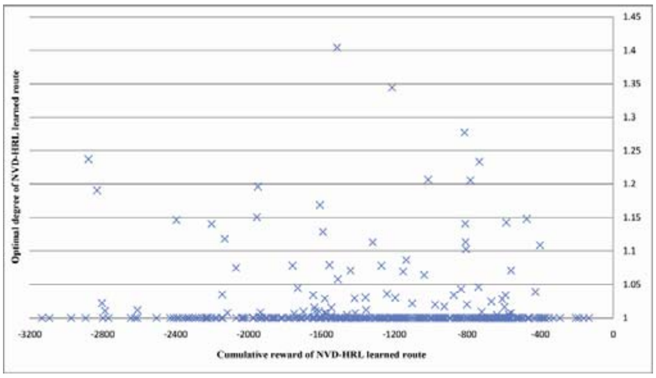

FIGURE IV. OPTIMALITY ANALYSIS of NVD-HRL

IV. Conclusions

We have proposed a hierarchical reinforcement learning approach to find the optima route considering human route selection criteria. This approach automatically identifies hierarchical subgoals and constructs corresponding hierarchical options in advance, and then learns the optimal route from origin to destination with these options at real-time. It can easily overcome the contradictions between model-driven route selection and the diversity of human cognitive preferences for optimal routes under principles of spatial cognition.

In this paper, the environment is assumed to be unfamiliar with the users, causing the action rewards are only judged by the spatial features at intersections. However, users may differ in their amount of prior knowledge and hence in their informational needs or preferences. In future work, we will consider integrating the user's a-priori spatial knowledge of the environment into the state representation and the learning process.

Acknowledgment

The work described in this paper was fully supported by the National Natural Science Foundation of China (No. 41301513), the Open Research Fund Program of Key Laboratory of Geo-information Engineering (No. SKLGIE 2014-M-4-2), and the Fundamental Research Funds for the Central Universities (No. 2014G1261056).

References

[1] M. Duckham and L. Kulik, Simplest paths: automated route selection for navigation, COSIT 2003, W. Kuhn, M. Worboys and S. Timpf, Berlin: Springer, 2003, pp. 169–185.

[2] D. Caduff and S. Timpf, The Landmark Spider: Representing Landmark Knowledge for Wayfinding Tasks, the 2005 AAAI spring symposium, Menlo Park, CA, 2005, pp. 30–35.

[3] S. Haque, L. Kulik and A. Klippel, Algorithms for reliable navigation and wayfinding, Spatial Cognition V, T. Barkowsky, M. Knauff, G. Ligozat and D. R. Montello, Berlin: Springer, 2007, pp. 308–326.

[4] K. F. Richter, Adaptable Path Planning in Regionalized Environments, COSIT 2009. K. S. Hornsby, C. Claramunt, M. Denis and G. Ligozat, Berlin: Springer, 2009, pp.453–470.

[5] D. M. Mark,. "Automated route selection for navigation." IEEE Aerosp Electron Syst Mag. Vol.1, pp. 2–5, 1986.

[6] K. F. Richter and M. Duckham, Simplest Instructions: Finding Easy-to-Describe Routes for Navigation, Geographic Information Science - 5th International Conference. T. J. Cova, H. J. Miller, K. Beard, A. U. Frank and M. F. Goodchild, Berlin: Springer, 2008, pp.274–289.

[7] R. G. Golledge, Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes. Baltimore, MD: Johns Hopkins Press, 1999.

[8] H. Cuayahuitl, N. Dethlefs, L. Frommberger, K. F. Richter and J. Bateman, Generating Adaptive Route Instructions Using Hierarchical Reinforcement Learning. Spatial Cognition VII, C. Holscher, Berlin: Springer, 2010, pp. 319–334.

[9] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press, 1998.

[10] P. Moradi, M. E. Shiri and N. Entezari, Automatic Skill Acquisition in Reinforcement Learning Agents Using Connection Bridge Centrality, Communications in Computer and Information Science, Berlin: Springer, 2010, pp. 51–62.

[11] A. A. Rad, M. Hasler and P. Moradi, Automatic Skill Acquisition in Reinforcement Learning using Connection Graph Stability Centrality, the IEEE International Symposium on Circuits and Systems (ISCAS), Paris, France, May 30 2010-June 2 2010, pp. 697–700.

[12] A. Okabe, T. Satoh, T. Furuta, A. Suzuki and K. Okano, "Generalized network Voronoi diagrams: Concepts, computational methods, and applications." International Journal of Geographical Information Science, vol. 22, pp. 965–994, 2008.

[13] R. S. Sutton, D. Precup and S. Singh, "Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning." Artificial Intelligence, vol. 112, pp. 181–211, 1999.

[14] D. Precup, Temporal Abstraction in Reinforcement Learning. University of Massachusetts Amherst, PhD Thesis, 2000.

[15] W. Zhao, Q. Li and B. Li, "Extracting Hierarchical Landmarks from urban POI Data." Journal of Remote Sensing, vol. 15, pp. 976–993, 2011.