DOI:10.13203/j.whugis2012.11.002

文章编号:1671-8860(2012)11-1271-05 文献标志码:A

空间认知导向下利用分层强化学习的最优路径规划

赵卫锋[1] 李清泉[1] 李必军[1]

(1 武汉大学测绘遥感信息工程国家重点实验室,武汉市珞喻路 129 号,430079)

摘 要:针对空间认知导向下模型驱动型路径规划和人们认知偏好多样性之间的矛盾,提出了一种基于分层强化学习的交互学习型路径规划方法。该方法将最优路径标准转换为路口处转向决策的瞬时奖励值,并通过预学习和实时学习两个阶段实现高效地发现总奖励值最大的最优路径策略。其中,预学习阶段自动发现子目标节点,并构建包含局部最优策略的子任务;实时学习阶段利用预定义策略实现高效的 Q 值更新,并根据 Q 值追溯最优路径。实验表明,该方法具有足够好的实时性和最优性。

关键词:空间认知;最优路径规划;分层强化学习;预学习;实时学习

中图法分类号:P208

近年来,一些学者研究了地标、转向方式、路口结构等对导航系统路径规划的影响,提出了一系列面向空间认知的最优路径,如最简单路径[1]、最清楚路径[2]、最可靠路径[3]和最易描述路径[4]等,以方便人们构建认知地图,并降低认知压力[5]。其实,现方法皆为将各种空间认知特征加权变换为路网模型的路段花费或转向延迟,然后采用改进的 Dijkstra 或 A* 算法进行路径计算。然而,由于不同用户通常具有不同的认知偏好,造成其关注的地标或各认知特征的权重系数有较大差异。因此,针对不同用户,通常需要生成不同的路网模型。这种模型驱动型路径规划会给路网数据的共享和维护带来较大麻烦。Cuayahuitl 等尝试了利用 MAXQ 分层强化学习在室内环境下进行路径规划的方法[6],但是其采用的手动分层和值函数分解方法并不适用于具有超大状态空间和不规则动作空间的城市道路环境。为了提高在大范围城市路网内学习的效率,本文提出了一种基于网络 Voronoi 图的分层强化学习方法。

1 分层强化学习原理

作为一种在未知或部分可知环境下通过试错和环境交互获取从状态到动作的最佳映射的机器学习方法,强化学习被广泛应用于最优化控制、机器人导航等领域[7]。作为以状态-动作对的期望奖励(即 Q 值)为学习对象的 Q 学习是最常用的强化学习算法。Q 学习的基本思想是:利用查找表维护各状态-动作对的 Q 值,并在由有限贪心策略(GLIE)决定的各学习周期(episode)中利用后继状态可能采取动作的最大 Q 值更新当前状态-动作对的 Q 值[7]。其 Q 值更新方程为:

$$Q(s_t,a_t)\leftarrow(1-\alpha)Q(s_t,a_t)+\alpha[r+\gamma\max_{a'\in A}Q(s_{t+1},a')] \tag{1}$$

然而,由于缺乏多尺度学习能力,Q 学习存在维数灾难问题,即当状态空间较大时,学习效率低,实时性差。以任务分层和子任务策略复用为指导思想的分层强化学习是克服维数灾难的常见方法。经典分层强化学习方法有时态抽象[8]、分层抽象机[9]和值函数分解[10]。相对于后两种方法,时态抽象法更适合任务自动分层,应用最广。

时态抽象的基本思想是从学习中抽象出若干子任务(常被称为 Option),并将其当作特殊动作加入基本动作集中。每个 Option 表示在状态子空间上按照局部最优策略执行的动作序列,由启动状态集 I、内部策略 π 和终止条件 β 构成。与传统的将包含 Option 的马尔科夫决策过程(MDP)当作半马氏决策过程(SMDP)并进行基于 SMDP

收稿日期:2012-09-15。

项目来源:国家自然科学基金重点资助项目(40830530);国家自然科学基金资助项目(60872132)。

的强化学习不同，Sutton 等提出的 Option 内部 Q 学习能够更新 Option 内部状态的 Q 值，提高了利用多层 Option 的学习效率[8]。其 Q 值更新方程为：

$$Q(s_t,o)\leftarrow Q(s_t,o)+\alpha[r+\gamma U(s_{t+1},o)-Q(s_t,o)] \quad (2)$$

其中，

$$U(s,o)=(1-\beta(s))Q(s,o)+\beta(s)\max_{o'\in O}Q(s,o')$$

由于 Agent 在 Option 内部状态间的转移受到 Option 内部策略的约束，因此，基于 Option 的分层强化学习不能随机访问任意状态-动作对，仅能收敛到局部最优策略(分层最优策略)。然而，合理的任务分层能够使其达到或接近全局最优[8]。

2 最优路径标准

本文定义的面向空间认知的最优路径主要包含如下目标：经过尽可能多的显著地标；涉及尽量少的复杂转向；尽量回避结构复杂的路口；尽量选择等级较高的道路；尽量避免绕行。

作为环境中具有突出特征的空间要素，地标在导航中起到转向标识和路径确认的作用。人们描述转向时，常用不同精度的方向模型，如 4 方向和 8 方向模型[11]。通常，人们会尽量回避可能产生歧义的转向，以减少由于使用复杂方向模型产生的认知压力。而对于一些复杂路口，如环岛，可以利用顺序关系降低转向描述的认知难度[11]。而属于相同方向模型的不同转向也具有不同的执行难度，如直行最容易，掉头最困难，右行制交通中右转易于左转等。此外，路口结构越简单，对转向的描述和执行越容易实现；反之，则需要花费越大的认知心力[3,4]。可以认为，路口的复杂度与在该路口处交汇的道路数量成正比。最后，人们通常更倾向沿着高等级道路前进，并尽量避免由于过多考虑认知因素而造成大幅绕路。

3 基于网络 Voronoi 图的分层强化学习

在基于 Option 的时态抽象方法的基础上，本文提出了一种适用于城市道路网结构的基于网络 Voronoi 图的分层强化学习(简记为 NVD-HRL)方法，以提高 Q 值的更新效率。该方法分为预学习和实时学习两个阶段，如图 1 所示。

3.1 预学习

预学习阶段的任务是利用预处理的方式根据特定的即时奖励方程发现路网中的子目标，并构建子任务，以生成分层的 Option。

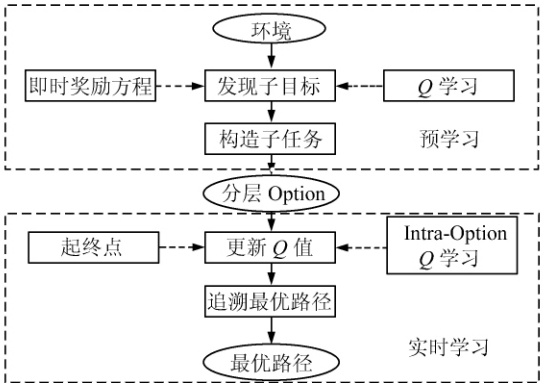

图 1 NVD-HRL 方法的实现流程

Fig. 1 Implementation Process of Our NVD-HRL Approach

3.1.1 即时奖励方程

即时奖励方程将从环境感知到的每个状态-动作对转换为一个表示该状态的内在渴望度的数值化的奖励值。因此，强化学习 Agent 的目标是通过定义当前问题的即时奖励方程，使其累积奖励值达到最大化。

在路径问题中，每个状态-动作对反映一个在路口处的具体转向决策。根据 §2 定义的对应每个转向行为花费认知心力的五条最优路径选择标准，即时奖励方程定量化并衡量了在每个路口处较好的转向方式。因此，期望的即时奖励方程可以表示为如下能够满足认知导向的最优路径目标的线性方程：

$$R=\omega_1\times R_1+\omega_2\times R_2+\omega_3\times R_3+\omega_4\times R_4+\omega_5\times R_5 \quad (3)$$

式中，$R_1\sim R_5$ 表示在每个路口的转向行为考虑五条最优路径选择标准收获的定量化的即时奖励值；$\omega_1\sim\omega_5$ 对应了权衡各指标的权重系数，且可以根据用户的个人偏好进行适当调整。

3.1.2 发现子目标

子目标指状态转移图中具有“瓶颈”特征的状态。在城市路网中，子目标指最优路径经常经过的节点。通过计算路网中各节点的中介中心性(即 BC 值)，发现多层子目标[12]。

BC 值用来度量一个节点位于连接任意相异节点对的最优路径(常为最短路径)上的频率，可以衡量节点在图中的重要性。记 u、s、t 为图 $G=<V,E>$ 中的节点，σ_{st} 为节点 s、t 之间的最优路径，则 BC 值的计算公式可表示为：

$$BC(u)=\sum_{u\neq s\neq t\in V}\frac{\sigma_{st}(u)}{\sigma_{st}} \quad (4)$$

在本文中,用来计算 BC 值的最优路径指符合定义的路径选择标准的最优路径。由于预学习阶段对算法的实时性要求不高,通过 Q 学习发现路网中任意相异两点间的最优路径。Q 学习过程中,状态转移图与道路网拓扑结构一致,每个动作对应转向决策的即时奖励由即时奖励方程(3)确定。计算出各节点的 BC 值后,依据其大小对所有节点进行分层,层次较高的若干层节点被当作子目标。

3.1.3　构造分层 Option

Option 定义了从启动状态到终止状态的局部最优策略,这和 Voronoi 空间分割的思想十分吻合。因此,本文提出在以子目标节点为种子点的网络 Voronoi 图上构建 Option 的方法。其中,作为 Voronoi 图的一种变体,考虑道路通行方向的有向网络 Voronoi 图可以分为入向网络 Voronoi 图和出向网络 Voronoi 图[13]。

根据定义的最优路径选择标准,各对节点之间的往返最优路径通常不重合。因此,以各层子目标为种子点,以各最优路径的总奖励值为网络距离,对所有节点进行有向网络 Voronoi 图划分。在入向或出向网络 Voronoi 图中,同时属于多个网络 Voronoi 区域或者与属于其他网络 Voronoi 区域的节点直接相连的节点被称为桥接点。通过桥接点连接的 Voronoi 区域互为相邻区域。图 2 展示了相邻入向网络 Voronoi 区域{ v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}, v_{11}, v_{12}}和{ v_{10}, v_{11}, v_{12}, v_{12}, v_{14}, v_{15}, v_{16}, v_{17}, v_{18}, v_{19}, v_{20}, v_{21}}的范围,其中星状符号为种子点,黑色填充的节点为它们的桥接点。

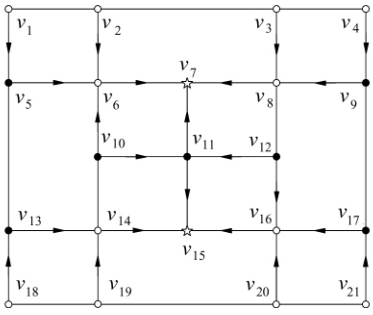

图 2　网络 Voronoi 图示例

Fig. 2　An Example of Network Voronoi Diagram

建立在入向和出向网络 Voronoi 区域上的 Option 分别被称为入向和出向 Option。入向 Option 的启动状态集包括所有非种子点,终止条件为 β(种子点)=1,β(非种子点) = 0,内部策略为非种子点到种子点的最优路径。出向 Option 的启动状态集仅包含种子点,终止条件为 β(种子点)=0,β(非种子点) > 0,内部策略为种子点到非种子点的最优路径。此外,为了方便入向和出向 Option 之间的状态转移,还为 Option 内每个桥接点定义了桥接策略。入向 Option 的桥接策略为从通过当前桥接点与当前 Option 邻接的 Option 的种子点到该桥接点的最优路径;出向 Option 的桥接策略为从当前桥接点到通过该桥接点与当前 Option 邻接的 Option 的种子点的最优路径。

3.2　实时学习

确定路径规划的起终点后,实时学习阶段利用分层 Option 更新转向决策的 Q 值,并根据收敛后的 Q 值追溯最优路径。

3.2.1　更新 Q 值

为了提高 Q 值的更新效率,采用了 Intra-Option Q 学习思想,即利用最高层 Option 生成学习周期,并利用较低层 Option 生成虚拟经历。实时学习过程中,各转向决策的 Q 值记录在查找表中,并被初始化为无效值,Agent 的状态转移由 Option 的内部策略和桥接策略决定,且各动作的即时奖励值记录在 Option 的内部策略或桥接策略中。

在基于最高层 Option 的 Q 学习中,为了减少 Q 值的无效更新,采取从终点到起点的逆序更新策略。生成每个逆序的学习周期为从终点开始搜索一条连接起终点的任意路径的过程,可以分为开始、中间和结束三个阶段,如图 3 所示。开始阶段,Agent 有两种选择:① 从终点移动到其所属出向 Option 的种子点,如终点→$N1$;② 从终点移动到其所属入向 Option 的桥接点,如终点→$N8$。中间阶段是循环过程,Agent 利用入向 Option 内部策略从种子点移动到任意桥接点或利用其桥接策略从桥接点移动到邻接 Option 的种子点,直至抵达起点所在 Option 的种子点,如 $N1$→…→$N7$ 或 $N8$→…→$N7$。终止阶段,Agent 则只有一种移动方式:利用入向 Option 从种子点移动到起点,即 $N7$→起点。在每个学习周期内,利用式(1)更新其经过转向决策的 Q 值。

对于各学习周期实际经过的节点,当其非种子点且所属较低层入向或出向 Option 的种子点非最高层种子点时,在其所属较低层入向和出向 Option 内分别记录的从该节点到种子点以及从种子点到该节点的内部策略可以用来产生虚拟经历。在对虚拟经历包含转向决策的 Q 值进行同步更新后,每个学习周期中被更新了 Q 值的转向决策远多于其实际经历的转向决策。由于相对于

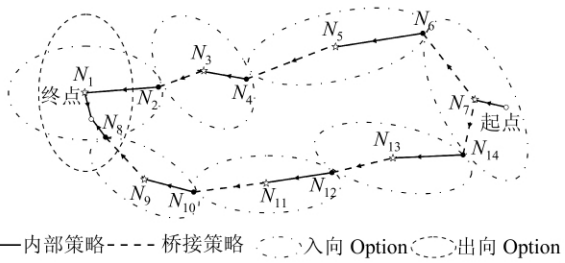

图3 逆序周期生成过程

Fig. 3 Process of Generating Reverse Episodes

产生有效学习周期的花费，生成虚拟经历和更新相关 Q 值需要的额外花费极低，故这种离策略(off-policy)的 Q 值更新方式保证了分层强化学习的高效性。

3.2.2 追溯最优路径

当一定周期内无任何 Q 值发生变化时，可以认为学习达到收敛，并不再生成新周期。此时，有效 Q 值表示在分层 Option 限定的状态-动作空间内从对应转向决策通往终点收获的最大奖励，而不包含有效 Q 值的节点则不可能在起终点间的任意路径上。因此，从起点开始，在每个节点处选择对应最大 Q 值的转向决策并过渡到后继节点，直至抵达终点，就能够发现起终点间的最优路径。

4 实 验

为了验证 NVD-HRL 方法的有效性，选择图4(a)所示的武汉市武昌区域进行路径规划实验。将因上下行路线分离产生的复合路口以及环岛当作复合节点，该区域共包含5 639个节点及40 214个不同的转向方式。另外，采用文献[14]的方法从本区域提取了224个地标。实验计算机采用2.5 GHz Intel Dual-Core CPU和2 GB内存。瞬时奖励值参数设定如下(其中括号内的值为对应的奖励值)。地标：一级地标(−10)、二级地标(−20)、三级地标(−40)、四级地标(−80)、无地标(−100)；转向方式：直行(−10)、右转(−40)、左转(−60)、掉头(−80)、通过环岛(−50)、歧义转向(−100)；道路等级：高速公路(−10)、城市高速(−20)、国道(−30)、城市主干道(−60)、一般道路(−80)、其他道路(−100)；路口结构：路口连接路段数 m(将 m 线性变换到区间(0，−100])；路段长度：路段实际长度 n(将 n 线性变换到区间(0，−100])。

根据§3的定义，将影响转向瞬时奖励值的各项因素转化为参数设定取值范围在区间(0，−100]内的定量参数，以使路径总奖励值最大时满足总花费最小。其中，对 m 和 n 作线性变换时，最大值分别为路网中最复杂路口连接的路段数和最长路段的长度，最小值皆为0。在利用式(3)计算各转向的瞬时奖励值时，将5个变量的权重系数皆设为0.2。此外，强化学习过程中，动作选择策略采用 Boltzmann 分布，即 $e^{Q_t(a)/t}/\sum_{b=1}^{n} e^{Q_t(b)/t}$，其中 τ 为退火系数；学习率 $\alpha=100/(100+T)$，T 为该状态经历的周期数；折扣系数 $\gamma=1$。

图4(b)～4(d)对比了采用以上参数设置的 Q 学习发现的三对起终点间的认知导向下的最优路径和最短路径。可以发现，按照本文定义的路径选择标准发现的最优路径基本符合人们找路或指路的认知习惯。

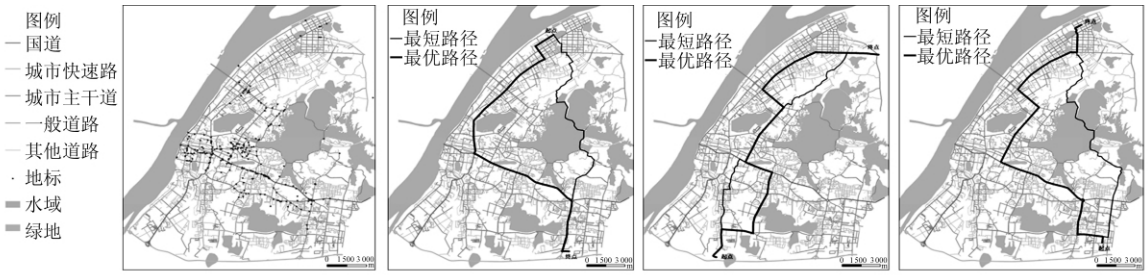

图4 研究区域和三对最优路径、最短路径

Fig. 4 Study Area and Three Pairs of Optimal Routes and Shortest Routes

计算出路网中各节点的 BC 值后，按照 BC 值由高到低依次选取分别占节点总数的5%、10%和25%的282个、564个和1 410个节点作为3层子目标，并构造相应的3层 Option。接下来，采用 NVD-HRL 方法对随机选择的500对起终点进行路径规划，其中每条路径运算10次后取平均值，并设定连续100个周期无 Q 值更新时学习收敛。

从图5(a)可以发现，对于绝大部分路径，该学习方法能够在1 s内达到收敛。这说明该方法能够高效地发现起终点间的最优路径，并接近模

型驱动型最短路径算法(Dijkstra 或 A*)的效率。

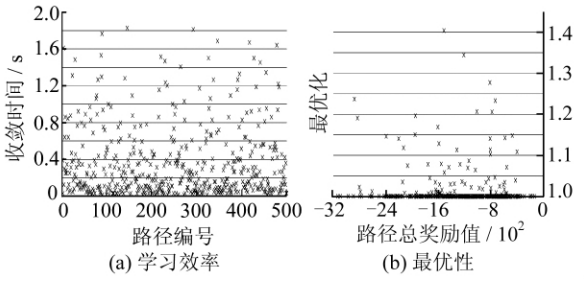

图 5　NVD-HRL 的学习效率和最优性

Fig. 5　Learning Efficency and Optimality of NVD-HRL

由于该方法仅能使规划结果达到分层最优，还将其与保证实现全局最优的 Q 学习进行了对比，如图 5(b)所示。其中，纵轴的最优比表示 NVD-HRL 和 Q 学习发现的各条路径的总奖励值之比，横轴为 NVD-HRL 发现的各条路径的总奖励值。可以发现，本方法规划出来的最优路径超过 80%达到全局最优，且超过 95%的最优比在 1.1 以内。

分层强化学习还使路径规划具有很好的动态性。学习收敛后，对于由追溯各转向决策对应的最大 Q 值获取的已知最优路径，只需在其经过的任意节点处采取与该路径不一致的转向决策，然后在各后继节点选择 Q 值最大的转向决策，就能快速生成一条非最优的派生路径。按照总奖励值将所有可能的派生路径排序，就能发现次最优路径、次次最优路径，依此类推。因此，分层强化学习同时高效地解决了 K 最优路径问题。在实际应用中，总能根据当前交通状况利用已学习的 Q 值快速发现一条避开拥堵或限行路段前提下从当前位置到终点总奖励值最大的路径。另外，当环境要素发生变化时，如增加新地标或扩宽道路，还可以通过对相关 Option 的局部更新使路径规划具有自适应性。在后续研究中，还将探索利用多 Agent 的分层强化学习解决多目标路径优化问题，以同时发现满足不同认知偏好的最优路径。

参　考　文　献

[1] Duckham M, Kklik L. Simplest Paths: Automated Route Selection for Navigation[C]. International Conference on Spatial Information Theory, Berlin, 2003

[2] Caduff D, Timpf S. The Landmark Spider: Representing Landmark Knowledge for Wayfinding Tasks [C]. Reasoning with Mental and External Diagrams: Computational Modeling and Spatial Assistance, Stanford, CA, 2005

[3] Richter K F. Adaptable Path Planning in Regionalized Environments[C]. International Conference on Spatial Information Theory, Berlin, 2009

[4] Richter K F, Duckham M. Simplest Instructions: Finding Easy-to-Describe Routes for Navigation [C]. Lecture Notes in Computer Science, Berlin, 2008

[5] Golledge R G. Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes[M]. Baltimore, MD: Johns Hopkins Press, 1998

[6] Cuayahuitl H, Dethlefs N, Frommberger L, et al. Generating Adaptive Route Instructions Using Hierarchical Reinforcement Learning[C]. Lecture Notes in Computer Science, Berlin, 2010

[7] Sutor R S, Barto A G. Reinforcement Learning: An Introduction[M]. Cambridge, MA: MIT Press, 1998

[8] Precup D. Temporal Abstraction in Reinforcement Learning[D]. Amherst, MA: University of Massachusetts, 2000

[9] Parr R E. Hierarchical Control and Learning for Markov Decision Processes[D]. Berkeley: University of California , 1998

[10] Dietterich T G. Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition [J]. Journal of Artificial Intelligence Research, 2000, 13(1): 227-303

[11] Klippel A, Tenbrink T, Montello D R. The Role of Structure and Function in the Conceptualization of Directions[M]. Oxford: Oxford University Press, 2009

[12] Rad A A, Hasler M, Moradi P. Automatic Skill Acquisition in Reinforcement Learning Using Connection Graph Stability Centrality[C]. 2010 IEEE International Symposium on Circuits and Systems, Canada, 2010

[13] Okabe A, Satoh T, Furuta T, et al. Generalized Network Voronoi Diagrams: Concepts, Computational Methods, and Applications[J]. International Journal of Geographical Information Science, 2008, 22(9): 965-994

[14] 赵卫锋，李清泉，李必军. 利用城市 POI 数据提取分层地标[J]. 遥感学报，2011,15(5): 976-993

第一作者简介：赵卫锋，博士生，现从事智能交通系统研究。

E-mail: wfzhao@whu.edu.cn

(下转第 1320 页)

An Improved SVR Image Fusion Algorithm Base on Low-pass Filter and Histogram Matching

GAO Yonggang[1,2] *XU Hanqiu*[1,2]

(1 College of Environment and Resources, Fuzhou University, 2 Xueyuan Road, Fuzhou 350108, China)

(2 Institute of Remote Sensing Information Engineering, Fuzhou University, 2 Xueyuan Road, Fuzhou 350108, China)

Abstract: To avoid the spectral distortion of SVR(synthetic variable ratio) algorithm, we propose an improved algorithm by using a low-pass filter and histogram matching performance, which is hence named SVR based on low-pass filter and histogram matching (SVRFM) algorithm. Two subsets from the IKONOS image of Fuzhou, representing different land cover types were used as test data. The spectral fidelity and the ability of gaining high frequency information were assessed by using visual and statistical analysis. The fused images were compared with those fused using the SVR, wavelet transform, pansharp, ehlers and Gram-Schmidt algorithms, respectively. The results show that the spectral fidelity of the SVRFM algorithm is generally better than the five algorithms compared.

Key words: image fusion; SVRFM; spectral fidelity

About the first author: GAO Yonggang, lecturer, Ph.D candidate, majors in remote sensing image processing and satellite altimetry. E-mail: yggao@fzu.edu.cn

(**上接第** 1275 **页**)

Spatial Cognition Oriented Optimal Route Planning with Hierarchical Reinforcement Learning

ZHAO Weifeng[1] *LI Qingquan*[1] *LI Bijun*[1]

(1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: Against the contradictions between model-driven route planning and the diversity of human cognitive preferences for spatial cognition oriented optimal routes, we present a kind of interactive route planning approach based on hierarchical reinforcement learning. In this approach, optimal route criterias are translated into immediate rewards of turning decisions at intersections, and optimal route policies with maximal cumulative rewards can be found through a two-stage learning process. The first pre-learning stage automatically identifies some nodes in road network as subgoals and constructs corresponding subtasks containing local optimal route policies for achieving the subgoals. The second real-time learning stage focuses on efficiently updating the Q values of every available state-action pair using predefined policies, and tracing the optimal routes according to Q values. The experimental results show that our proposed approach learns effectively enough and ensures the routes found close to global optimal ones.

Key words: spatial cognition; optimal route planning; hierarchical reinforcement learning; pre-learning; real-time learning

About the first author: ZHAO Weifeng, Ph.D candidate, majors in intelligent transportation system. E-mail: wfzhao@whu.edu.cn