摘 要

激励学习因具有较强的在线自适应性和对复杂系统的自学习能力，备受机器人导航研究者的关注。但其在连续状态和动作空间的泛化，局部环境的反应式控制，大状态空间和部分可观测环境定性导航等都存在着亟待解决的问题，且用传统的算法很难满意地解决这些问题。本文利用人工势场和激励学习的优点针对机器人在较大状态空间和部分可观测环境下的导航问题进行了研究。

本文首先对激励学习研究现状，课题研究的背景和现实意义进行了综述性介绍，并分析了当前激励学习中两种比较成熟的方法，瞬时差分法和 Q 学习方法。

其次，研究了人工势场中斥力势函数和引力势函数的选取，人工势场法的优缺点。然后重点研究了如何将激励学习模型转换成人工势场模型，即利用激励学习和人工势场的优点应用虚拟水流法如何构建一个具有记忆学习功能的激励势场模型。

最后，用三个著名的网格世界问题对激励势场模型进行了测试，同时在较大状态空间中用 Q 学习和 HQ 学习等方法做了对比实验。实验结果表明：对较大状态空间和部分可观测环境新方法都能简洁有效地给出理想的解；与 Q 学习和 HQ 学习等方法相比激励势场模型更稳定有效。

关键词：激励学习，人工势场，路径规划，移动机器人导航，虚拟水流法

ABSTRACT

Reinforcement learning (RL) has attracted most researchers in the area of robotics, because of its strong on-line adaptability and self-learning ability for complex system. But with the development of robot, more challenges come up, such as environment perception, generalization of RL, reactive control in local environment, large scale and partially observable environments, etc. It is difficult for common algorithms to obtain a satisfied solution. In this thesis, we try to research the mobile robot navigation in the large scale and partially observable environments, with the artificial potential field (APF) and reinforcement learning.

Firstly, the introduction reviews the research on RL, its relative aspects in the world, the background and practical significance. Then temporal difference learning and Q-learning, two relatively mature kinds of algorithms are analyzed.

Secondly, the repulsion force function and the gravitation force function of the potential field are introduced. And the excellent and shortcoming of the artificial potential field method have analyzed at the same time. A reinforcement learning problem is transferred to a path planning problem by using artificial potential field is the main contents of this thesis. That is, an efficient reinforcement potential field model (RPFM) is presented, with a virtual water-flow concept.

Finally, the performance of RPFM is tested by the three well-known gridworld problems, and also the experiment with HQ and Q-learning for comparison had been done. Experimental results show that the RPFM is simple and effective to give an optimal solution for observable and partially observable RL. Compared with HQ and Q learning, our model is more stable and effective.

Key words: Reinforcement Learning; Artificial Potential Field; Path Planning; mobile robot navigation; Virtual Water-flow.

长沙理工大学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：刘泽文　　日期：2008年5月19日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权长沙理工大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1、保密□，在_____年解密后适用本授权书。

2、不保密☑。

（请在以上相应方框内打“√”）

作者签名：刘泽文　日期：2008年5月19日

导师签名：　　日期：2008年5月19日

第一章 引言

1.1 本文研究的背景

根据反馈的不同，机器学习可以分为：有监督学习(Supervised Learning)、无监督学习(Unsupervised Learning)和激励学习(Reinforcement Learning,RL)三大类。其中激励学习是智能体（Agent）从环境到行为映射的学习，通过“试错”方式与环境进行交互作用以期许奖赏（激励）信号达到最大（或最小）的机器学习方法[1]。从20世纪80年代末开始，随着对激励学习的数学基础研究取得突破性进展后，对激励学习的研究和应用日益开展起来，并成为机器学习领域的研究热点之一。激励学习方法能够通过获得与环境的交互过程中的评价性反馈信号来实现行为决策的优化，因此在求解复杂的优化控制问题中具有广泛的应用价值。但激励学习对许多现实问题，如在连续状态和动作空间的泛化，局部环境的反应式控制，大状态空间和部分可观测环境定性导航等都存在着亟待解决的问题，且用传统的算法很难满意地解决这些问题。

1.2 激励学习理论与应用综述

激励学习(RL)又称为强化学习、增强学习或再励学习，是不同于监督学习和无监督学习的另一大类机器学习方法。激励学习的基本思想与动物学习心理学“试错法”学习[2]的研究密切相关，即强调在与环境中的交互中学习，通过环境对不同行为的评价性反馈信号来改变行为选择策略以实现学习目标。来自环境的评价性反馈信号通常称为奖赏((Reward)或激励信号(Reinforcement Signal)，激励学习系统的目标就是期望激励信号最大化。虽然监督学习方法如神经网络反向传播算法[3]、决策树学习算法[4]等的研究取得了大量成果，并在许多领域得到了成功的应用，但由于监督学习需要给出不同环境状态下的教师信号，因此限制了监督学习在复杂的优化控制问题中的应用。无监督学习虽然不需要教师信号，但仅能完成模式分类等有限的功能。由于激励学习方法能够通过获得与环境的交互过程中的评价性反馈信号来实现行为决策的优化，因此在求解复杂的优化控制问题中具有更为广泛的应用价值。

基于激励学习的上述特点，在早期的人工智能研究中曾一度将激励学习作为一个重要的研究方向，如Minsky有关激励学习的博士论文[5]，Samuel的跳棋学习程序[6]等，但后来由于各种因素特别是求解激励学习问题的困难性，在二十世纪七、八十年代人工智能和机器学习的研究主要面向监督学习和无监督学习方法。进入

二十世纪九十年代，激励学习在理论和算法上通过与其他学科如运筹学、控制理论的交叉综合，取得了若干突破性的研究成果，并且在机器人控制、优化调度等许多复杂优化决策问题中取得了成功的应用。

目前激励学习不但成为人工智能和机器学习领域的重要研究方向，而且吸引了许多其他学科研究人员的注意。

1.2.1 激励学习研究的背景

激励学习在算法和理论研究方面的一个重要特点是多学科的交叉综合性。激励学习的研究与动物学习心理学、运筹学、进化计算、自适应控制和神经网络等学科都具有密切的联系。

1. 动物学习心理学

有关动物学习心理学的研究为激励学习算法和理论的研究提供了思想基础。在动物学习心理学的研究中，关于动物“试错”(Trial)学习的思想最早由 Edward Thorndike 于 1911 年提出[7]，该思想的实质是强调行为的结果有优劣之分并为后继行为选择提供依据。Thorndike 称这种规律为“效应定律”(Law of Effect)，并指出效应定律描述了激励性事件对于动物行为选择趋势的影响，即能够导致正的回报的行为选择概率将增大，而导致负回报的行为选择概率将减小。文献[2]指出，效应定律包括了“试错”型学习的两个主要方面，即选择性和联想性。进化学习中的自然选择具有选择性，但不具有联想性;监督学习则仅具有联想性而不具有选择性。另外，“效应定律”反映了激励学习的另两个重要特性，即搜索和记忆。

在动物学习心理学中与激励学习密切相关的另一个研究内容是瞬时差分(Temporal Difference, TD)(或称为时间差分)理论[8][9]。所谓瞬时差分是指对同一个事件或变量在连续两个时刻观测的差值，这一概念来自于学习心理学中有关“次要激励器”(Secondary reinforcers)的研究。在动物学习心理学中，次要激励器是伴随主要激励信号如食物等的刺激，并且产生类似于主要激励信号的行为激励作用[2]。在早期的激励学习研究中，瞬时差分学习方法成为一个重要研究内容，如 A. Samuel 的跳棋程序中就采用了瞬时差分学习的思想[6]。在近十年来激励学习算法和理论的研究中，瞬时差分学习理论和算法同样具有基础性的地位。

2. 运筹学

运筹学是与激励学习研究紧密联系的另一个学科。运筹学中有关 Markov 决策过程(MDP)[10][11]和动态规划的算法和理论为激励学习的研究提供了数学模型和算法理论基础。其中主要包括 Bellman 的最优性原理和 Bellman 方程、值迭代、策略迭代等动态规划算法[10][12]。动态规划和激励学习方法的联系由 Minsky[5]在分析 Sameul 的跳棋程序时首先提出，并逐渐得到普遍重视。许多激励学习算法如 Q 学习算法[13]等都可以看作与模型无关的自适应动态规划算法。激励学习和动态规划

两个学科的交叉综合成为推动激励学习算法和理论研究的重要因素。近年来，求解大规模状态空间的动态规划方法,如值函数逼近方法[14]等在激励学习领域也得到了广泛的研究和应用。

3. 进化计算

进化计算(Evolutionary computation)是基于自然世界生物的自然选择和基因遗传原理实现的一类优化算法，并被广泛的用于求解机器学习问题。目前，进化计算在算法和理论上已取得了大量的研究成果，形成了遗传算法[15]、进化策略[16]和进化规划[17]三个主要的分支，并且在组合优化、自动程序设计、机器学习等领域获得了成功的应用[18]。虽然早期的进化计算与激励学习的研究相互独立，但随着研究的深入，进化计算方法在求解激励学习问题中的应用逐步得到重视。对于利用评价性反馈的激励学习问题，进化计算方法能够通过将回报信号映射为个体的适应度进行求解。在应用进化计算方法求解激励学习问题时，一个关键课题是如何对延迟回报进行时间信用分配(Temporal Credit Assignment)。 J.Holland的分类器学习算法(Classifier System) [19]对上述问题进行了开拓性的研究，在该算法中体现了瞬时差分学习的思想。近年来，求解激励学习问题的进化激励学习方法成为一个重要的研究课题。文献[20]对进化激励学习的研究进行了综述。如何综合利用两种方法的优点实现多策略的高效激励学习系统是一个值得研究的课题。

4. 自适应控制

自适应控制是控制理论的一个重要分支，研究模型未知或不确定的对象的控制问题。在自适应控制中，按照是否对对象模型进行在线估计可以分为直接自适应控制方法和间接自适应控制方法两类。其中，直接自适应控制不建立对象的显式估计模型，而直接通过调节控制器参数实现闭环自适应控制；间接自适应控制则在对对象模型进行在线辨识的基础上，调节控制器的参数。文献[21]对激励学习作为一类直接自适应最优控制方法的特性进行了分析和研究，指出了激励学习与自适应控制理论的联系。与动态规划不同，激励学习不需要 Markov(马氏)决策过程的状态转移模型，而直接根据与环境的交互信息实现马氏决策过程的优化控制。在自适应控制中的辨识与控制的关系类似于激励学习中行为探索(Exploration)和利用((Exploitation)的关系。激励学习的行为探索是指不采用当前策略的随机化行为搜索，与自适应控制的辨识信号输入相对应；行为利用是指采用当前策略的控制行为选择，对应自适应控制的控制器优化设计。

5. 神经网络

神经网络的研究起源于对人类大脑的神经生理学和神经心理学的研究，目前已取得了丰富的研究成果，包括多种神经网络的结构模型和学习算法。在神经网络的学习算法中，针对监督学习和无监督学习已开展了大量的研究工作，近年来

神经网络的激励学习算法以及激励学习与监督学习的混合算法也取得了许多成果。神经网络作为一种通用的函数逼近器，对于解决激励学习在大规模和连续状态空间问题中的泛化(Generalization)具有重要的意义。研究神经网络在激励学习值函数逼近和策略逼近中的应用，克服激励学习和动态规划的“维数灾难”(Curse of Dimensionality)，是实现激励学习方法在实际工程中广泛应用的关键。目前，利用神经网络求解激励学习和动态规划问题的有关算法和理论(又称为神经动态规划(Neuron-dynamic Programming)方法[22])的研究是一个重要的研究方向。

1.2.2 激励学习算法的研究进展

按照学习系统与环境交互的类型，已提出的激励学习算法可以分为非联想激励学习（Non-associative RL)方法和联想激励学习(Associative RL)方法两大类。非联想激励学习系统仅从环境获得回报，而不区分环境的状态;联想激励学习系统则在获得回报的同时，具有环境的状态信息反馈，其结构类似于反馈控制系统。

在非联想激励学习研究方面，主要研究成果包括：Thathachar 等[8]针对 N 臂赌机问题（N-armed bandit)提出的基于行为值函数的估计器方法和追赶方法(Pursuit method)，R.Sutton 提出的激励信号比较方法(Reinforcement comparison)等。由于非联想激励学习系统没有环境的状态反馈，因此主要用于一些理论问题的求解，如多臂赌机等。

联想激励学习按照获得的回报是否具有延迟可以分为即时回报联想激励学习和序贯决策(Sequential decision)激励学习两种类型。即时回报联想激励学习的回报没有延迟特性，学习系统以极大(或极小)化期望的即时回报为目标，已提出的算法包括联想搜索(Associative search)方法、可选自助方法等[8]。

由于大量的实际问题都具有延迟回报的特点，因此用于求解延迟回报问题的序贯决策激励学习算法和理论成为激励学习领域研究的重点。在序贯决策激励学习算法研究中，采用了运筹学中的 Markov 决策过程(Markov Decision Processes: MDPs)模型，激励学习系统也类似于动态规划将学习目标分为折扣型回报指标和平均回报指标两种。同时根据 MDP 行为选择策略的平稳性，激励学习算法可以分为求解平稳策略 MDP 值函数的学习预测(Learning Prediction)方法和求解 MDP 最优值函数和最优策略的学习控制(Learning Control)方法。下面按照优化指标的不同，分别对折扣型回报指标激励学习和平均回报激励学习在算法和理论方面的研究概况进行介绍。

1.折扣型回报指标激励学习算法

(1) TD(λ)学习算法。瞬时差分学习方法在早期的激励学习和人工智能中占有重要的地位，并取得了一些成功的应用(如著名的跳棋学习程序等)，但一直没有建立统一的形式化体系和理论基础。R.Sutton 首次提出了求解平稳 MDP 策略评

价问题的瞬时差分学习算法(TD(λ)算法)，并给出了瞬时差分学习的形式化描述，证明了 TD(λ)学习算法在一定条件下的收敛性，从而为瞬时差分学习奠定了理论基础[23]。TD(λ)学习算法是一种学习预测算法，即在 MDP 的模型未知时实现对平稳策略的值函数估计。在 TD(λ)学习算法中利用了一种称为适合度轨迹(Eligibility Traces)的机制来实现对历史数据的充分利用。并且通常采用称为增量式(Accumulating)适合度轨迹。 Singh 和 Sutton 提出了一种新的替代式(Replacing)适合度轨迹，并验证了该方法的有效性[24]。

为提高 TD 学习算法的收敛速度，同时克服学习步长的设计困难，文献[25]提出了一种基于线性值函数逼近的最小二乘 TD(0)学习算法，该算法直接以 MDP 值函数逼近的均方误差为性能指标，没有采用适合度轨迹机制，因此称为 LS_TD(0)和 RLS_TD(0)学习算法。文献[26]中 J.Boyan 提出了一种 LS_TD(λ)学习算法，该算法能够获得优于 TD 学习算法的收敛速度，但存在计算量大和矩阵求逆的数值计算病态问题，难以实现在线学习。

(2) Q 学习算法。针对优化折扣回报指标的学习控制问题，Watkins 提出了表格型的 Q 学习算法，用于求解 MDP 的最优值函数和最优策略[27]；Peng 与 Williams 提出了 Q(λ)算法[28]，在该算法中结合了 Q 学习算法和 TD 学习算法中的适合度轨迹(Eligibility traces)，以进一步提高算法的收敛速度。为进一步提高激励学习算法的学习效率，基于自适应控制中模型辨识的思想，Sutton 提出了具有在线模型估计的 Dyna_Q 学习算法[8]，Peng 等提出了优先遍历方法(Prioritized sweeping) [29]。上述方法都在学习过程中对 MDP 的模型进行在线估计，虽然能够显著提高效率，但必须以较大的计算和存储量为代价。

(3) Sarsa 学习算法。 Rummery 等提出了一种在线策略(On-policy)的 Q 学习算法，称为 Sarsa 学习算法[30]。在 Q 学习算法中，学习系统的行为选择策略和值函数的迭代是相互独立的，而 Sarsa 学习算法则以严格的 TD 学习形式实现行为值函数的迭代，即行为选择策略与值函数迭代是一致的。Sarsa 学习算法在一些学习控制问题的应用中被验证具有优于 Q 学习算法的性能。在 Sarsa 学习算法中，行为探索策略的选择对算法的收敛性具有关键作用，文献[31]提出了两类行为探索策略，即渐近贪心无限探索(Greedy in the Limit and Infinitely Exploration, GLIE)策略和 RRR 策略(Restricted Rank-based Randomlized Policy)，以实现对 MDP 最优值函数的逼近。

(4) Actor-Critic 学习算法。学习控制算法具有的一个共同特点是仅对 MDP 的值函数进行估计，行为选择策略则由值函数的估计完全确定。A.Barto 和 R.Sutton 提出的 Actor-Critic 学习算法[32]则同时对值函数和策略进行估计，其中 Actor 用于进行策略估计，而 Critic 用于值函数估计。在 Actor-Critic 学习算法中，Critic 采用 TD 学习算法实现值函数的估计，Actor 则利用一种策略梯度估计

方法进行梯度下降学习。在文献[32]提出的 Actor-Critic 算法仅针对离散行为空间，文献[33]进一步研究了求解连续行为空间 MDP 最优策略的 Actor-Critic 算法。

(5)直接策略梯度估计算法。激励学习控制算法的另一种类型是不对 MDP 的值函数进行估计，而只进行最优策略估计的算法。早期的研究如 Williams 的 REINFORCE 算法[8]。与前面两类激励学习算法相比，这一类算法存在策略梯度估计困难、学习效率低的缺点。

2. 平均回报指标激励学习算法

随着对折扣型回报指标的激励学习方法研究的不断深入，平均回报指标的激励学习方法也逐渐得到重视。这是由于在某些工程问题中，优化目标更适合用平均回报指标来描述，如果采用折扣回报指标，则要求折扣因子接近 1。目前平均回报指标的激励学习已提出了多种算法，主要包括：

(1)基于平均回报指标的瞬时差分学习算法。在文献[34]中，Bertsekas 等将求解平均回报指标 MDP 策略评价问题的动态规划理论和方法应用于瞬时差分学习，提出了基于平均回报指标的瞬时差分学习算法。在该算法中，通过引入动态规划中相对值函数(Relative Value Function)的概念，实现了在 MDP 模型未知时对平稳策略 MDP 的值函数估计。文献[35]提出了类似的平均回报 TD 学习算法。

(2) R_学习算法。与求解基于平均回报指标 MDP 的学习控制问题，类似于 Q 学习算法，文献[36]提出了 R_学习算法。在 R_学习算法中，通过对相对值函数的迭代和贪婪的行为选择策略实现广义策略迭代过程。在文献[36]的仿真研究中，R_学习算法在某些情况下可以获得优于Q学习算法等折扣型激励学习算法的性能。

(3) H_学习算法。H_学习算法[37]可以看作是一种基于在线模型估计的 R_学习算法。为验证 H_学习算法的有效性，文献[37]对 H_学习算法在一个仿真的 AGV 调度问题中的应用进行了研究，获得了较好的学习性能。

需要说明的是，由于折扣型指标的激励学习算法在折扣因子接近 1 时的性能与平均回报指标的性能类似，而在理论分析方面，折扣指标算法要远比平均回报指标算法简易。

1.2.3 激励学习的泛化方法研究概况

上述介绍的激励学习算法基本都是针对离散状态和行为空间 MDP 的，即状态的值函数或行为值函数采用表格的形式存储和迭代计算。但实际工程中的许多优化决策问题都具有大规模或连续的状态或行为空间，因此表格型激励学习算法也存在类似于动态规划的“维数灾难”。为克服“维数灾难”，实现对连续状态或空间MDP最优值函数和最优策略的逼近，必须研究激励学习的泛化(Generalization)或推广问题，即利用有限的学习经验和记忆实现对一个大范围空间的有效知识获取和表示。由于激励学习的泛化方法的研究是影响其广泛应用的关键，因此对该

问题的研究成为当前的研究热点。目前提出的激励学习泛化方法主要包括以下几个方面：

1. 值函数逼近方法的研究

虽然值函数逼近在动态规划中的研究中开展得较早，但激励学习中的值函数逼近方法研究则与神经网络研究的重新兴起密切相关。随着神经网络的监督学习方法如反向传播算法的广泛研究和应用，将神经网络的函数逼近能力用于激励学习的值函数逼近逐渐开始得到学术界的重视。在瞬时差分学习的研究中，线性值函数逼近器得到了普遍的研究和注意。Sutton 在首次提出 TD(λ)学习算法时，就给出了线性值函数逼近的 TD(λ)算法(以下简称线性 TD(λ)算法或 TD(λ)算法)[9]。在线性 TD(λ)算法的基础上，Brartke 等利用递推最小二乘方法提出了 LS-TD(0)算法和 RLS-TD(0)算法[25]。Boyan 给出了直接求解 TD(λ)算法稳态方程的 LS-TD(入)算法[26]。在神经网络作为值函数逼近器的研究中，小脑模型关节控制器(Cerebellar Model Articulation Controller: CMAC)是应用得较为广泛的一种。Watkins 首次将 CMAC 用于 Q 学习算法的值函数逼近中，Sutton 在文献[9]和文献[38]中分别将 CMAC 成功地用于连续状态空间 MDP 的瞬时差分预测学习和学习控制问题中。基于一般的前向多层神经网络的值函数逼近方法也得到了广泛研究，如文献[39]利用神经网络的瞬时差分学习实现了西洋棋的学习程序 TD-Gammon。在上述研究中，神经网络的学习算法都采用了与线性 TD(入)学习算法相同的直接梯度(Direct Gradient)下降形式。L.Biard 指出上述直接梯度下降学习在使用非线性值函数逼近器求解 MDP 的学习预测和控制问题时可能出现发散的情况，提出了一种基于 Bellman 残差指标的梯度下降算法，称为残差梯度学习((Residual Gradient) [40]。残差梯度学习可以保证非线性值函数逼近器在求解平稳 MDP 的学习预测问题时的收敛性，但无法保证求解学习控制问题时的神经网络权值收敛性。

2. 策略空间逼近方法

与值函数逼近方法不同，策略空间逼近方法通过神经网络等函数逼近器直接在 MDP 的策略空间搜索，但存在如何估计策略梯度的困难。早期的 REINFORCE 算法只针对二值回报信号，文献[41]提出了一种离散行为空间的策略梯度估计方法。

3. 同时进行值函数和策略空间逼近的泛化方法

在同时进行值函数和策略空间逼近的泛化方法中，基本都采用了 Actor-Critic 的结构，即 Actor 网络实现对连续策略空间的逼近，Critic 网络实现对值函数的逼近。文献[42]研究了基于模糊系统的 AHC 学习算法，文献[33]研究了基于模糊神经网络的 AHC 学习算法，提出了网络结构和参数同时在线调整的算法。在上述研究中，Critic 通常采用基于神经网络的瞬时差分 TD(λ)学习算法，而 Actor 网络则基于一种高斯分布的随机行为探索机制对策略梯度进行在线估计。

4. 基于解释的神经网络激励学习方法

基于解释的学习是一种结合归纳学习和演绎推理的混合策略机器学习方法，在符号学习领域中已得到了广泛的研究和应用。S. Thrun 提出了一种基于解释的神经网络学习方法，并应用于激励学习的值函数逼近中，通过对神经网络的梯度信息的归纳解释，有效地加速了激励学习值函数逼近的收敛[36]。

1.2.4 激励学习的理论与应用研究进展

1. 在理论方面，类似于自适应控制中对闭环系统稳定性的研究，算法的收敛性研究成为激励学习理论的主要研究内容。同时对于激励学习泛化的有关基础理论，如值函数逼近方法的权值学习收敛性和性能误差上界分析等也取得了初步的成果。

(1)瞬时差分学习理论和 TD(λ)预测学习算法的收敛性

瞬时差分学习理论的建立以 Sutton 首次给出瞬时差分学习的形式化描述[9]和 TD(λ)学习算法为标志，已取得了许多研究成果，并成为其他激励学习算法如 AHC 学习算法、Q 学习算法的基础。针对 TD(λ)学习算法在求解平稳策略 MDP 值函数预测时的收敛性，文献[43]证明了任意 $0<\lambda<1$ 的表格型折扣回报 TD(λ)学习算法的概率收敛性；对于采用线性值函数逼近的 TD(λ)学习算法(又称为线性 TD(λ)学习算法)，文献[44]证明了平均意义下的收敛性；Tsitsiklis 等[41]证明了线性 TD(λ)算法在概率 1 意义下的收敛性并给出了收敛解的逼近误差上界。针对 TD(λ)学习算法中 λ 的选取对学习性能的影响，文献[45]研究了 TD(λ)学习算法均方误差与 λ 的函数关系，给出了一定假设下的表达式，并通过实验进行了验证。

(2)表格型激励学习控制算法的收敛性

用于求解 MDP 的学习控制问题的激励学习方法主要包括 Q 学习算法、Sarsa 学习算法和 AHC 学习算法等。Watkins 等在 1992 年证明了在学习因子满足随机逼近迭代算法条件并且 MDP 状态空间被充分遍历时，表格型 Q 学习算法以概率 1 收敛到 MDP 的最优值函数和最优策略[13]。文献[46]进一步基于异步动态规划和随机逼近理论证明了 Q 学习算法的收敛性。Singh 等研究了表格型 Sarsa(0)学习算法的收敛性，证明了在两类学习策略条件下 Sarsa 学习算法的收敛性[31]。

(3)有关激励学习泛化的理论研究

对于采用值函数逼近器的激励学习控制算法，目前在收敛性分析理论方面还比较缺乏。Baird 提出的残差梯度算法[40]仅能保证在平稳学习策略条件下的局部收敛性，无法实现对马氏决策过程最优值函数的求解。VAPS 算法虽然能够保证权值的收敛性，但无法保证策略的局部最优性[47]。Heger 研究了值函数逼近误差上界与策略性能误差上界的关系，指出当值函数逼近误差上界较小时，获得的近似最优策略具有性能保证，从而为基于值函数逼近的激励学习泛化方法提供了理论分析基础[48]。

2.在应用方面，随着激励学习在算法和理论方面研究的深入，激励学习方法在实际的工程优化和控制问题中得到了广泛的应用。目前激励学习方法已在非线性控制、机器人规划和控制、人工智能问题求解、组合优化和调度、通讯和数字信号处理、多智能体系统、模式识别和交通信号控制等领域取得了若干成功的应用。

(1)激励学习在非线性控制中的应用

在激励学习的研究中，小车倒摆系统作为一种典型的非线性控制对象，也成为激励学习应用和研究的目标之一。张平等采用伪熵来改进 Q 学习算法并对小车倒摆系统进行了学习控制仿真[49]。Bart 等研究了 AHC 算法在倒立摆学习控制中的应用[32]。蒋国飞等研究了基于神经网络 Q 学习的倒立摆学习控制[50]。

(2)激励学习在机器人规划和控制中的应用

在机器人学中，基于行为的机器人体系结构由 R.Brooks 于 80 年代提出[51]，近十年来已取得了大量的研究成果，该体系结构与早期提出的基于功能分解的体系结构逐渐开始相互结合，成为实现智能机器人系统的重要指导性方法。在基于行为的智能机器人控制系统中，机器人能否根据环境的变化进行有效地行为选择是提高机器人的自主性的关键问题。要实现机器人的灵活和有效的行为选择能力，仅依靠设计者的经验和知识是很难获得对复杂和不确定环境的良好适应性的。为此，必须在机器人的规划与控制系统引入学习机制，使机器人能够在与环境的交互中不断激励行为选择能力。机器人的学习系统研究是近年来机器人学界的研究热点之一。一些著名大学都建立了学习机器人实验室。C.K.Tham 等采用模块化 Q 学习算法实现了机器人手臂的任务分解和控制，在每个 Q 学习模块中采用了 CMAC 逼近值函数[52]；L.J.Lin 提出了结构化 Q 学习方法用于移动机器人的控制和导航[53]；S.Singh 采用复合 Q 学习算法用于机器人的任务规划和协调[54]；在 R.Jacobs 等的工作中也采用了模块化神经网络进行机器人手臂逆动力学的学习[55]。

(3)激励学习在优化与调度中的应用

基于 MDP 的激励学习算法将随机动态规划与动物学习心理学的“试错”和瞬时差分原理相结合，利用学习来计算状态的评价函数，因而能够求解模型未知的优化和调度问题。采用基于函数逼近的激励学习算法来求解大规模的优化和调度问题是激励学习应用的一个重要方面。

J.Boyan 提出了一种基于值函数学习和逼近的全局优化算法一 STAGE，在一系列大规模优化问题的求解中，STAGE 算法的性能都超过了模拟退火算法((SA)。采用乐观的 TD(λ)算法和神经网络逼近器，Crites 和 Bart 等进行了电梯调度的优化[56]，Zhang 和 Dietterich 等进行了生产中的 Job-Shop 问题的优化[57]，上述应用都取得了令人满意的结果，显示了激励学习在优化和调度中广泛应用前景。

激励学习在优化调度中的其它应用还包括:基于线性函数逼近Q学习算法的多

处理机系统的负载平衡调度[58]等。

(4)人工智能中的复杂问题求解

各种复杂问题求解一直是人工智能研究的重要领域，早期的各种启发式搜索方法和基于符号表示的产生式系统在求解一定规模的复杂问题中取得了成功。但这些方法在实现过程中都存在知识获取和表示的困难，如 IBM 的 Deep Blue 有大量参数和知识数据库，必须通过有关专家进行手工调整才能获得好的性能。

激励学习算法与理论的研究为人工智能的复杂问题求解开辟了一条新的途径，激励学习的基于多步序列决策的知识表示和基于“试错”的学习机制能够有效地解决知识的表示和获取的问题。在早期的 Samuel 的跳棋程序中就应用了一定的激励学习思想。

目前，激励学习在人工智能的复杂问题求解中已取得了若干研究成果，其中有代表性的是 G. J. Tesauro 的 TD-Gammon 程序[39]，该程序采用前馈神经网络作为值函数逼近器，基于 TD(λ)算法通过自我学习对弈实现了专家级的 Back-Gammon 下棋程序。其它的相关工作包括:S. Thrun 研究了基于激励学习的国际象棋程序[59]，并取得了一定的进展。

(5)激励学习在其他领域的应用

激励学习除了在上述领域得到了广泛的应用外，在其他领域也取得了初步的应用成果，包括:在 MAS(多智能体系统)中的应用[55] [60]和交通信号的控制[61] [62]。

1.2.5 存在的问题和本文的研究重点

尽管激励学习的研究在国外已广泛开展，但在国内还没有得到应有的普遍关注。近年来，国内若干高校和研究所已开展了有关激励学习算法和理论的研究工作，从相关文献来看，目前的研究工作还不够深入和广泛，激励学习方法的工程应用还有待进一步拓展。

目前，虽然关于激励学习的算法和理论的研究已经取得了大量的研究成果，但仍然有许多关键问题有待解决。在算法和理论方面，已提出了多种表格型的激励学习算法，并建立了较为完善的收敛性理论，但对于连续、高维空间的马氏决策问题将面临类似动态规划的“维数灾难”。目前已提出的激励学习泛化方法如基于神经网络的激励学习方法等仍然存在学习效率不高,在理论上的收敛性难以保证等缺点，还有待进一步深入研究，以扩大激励学习在实际工程问题中的应用。在应用方面，实现激励学习在复杂、不确定系统优化和控制问题中的应用对于推动工业、航天、军事等各个领域的发展都具有显著的意义和工程价值，特别是对于移动机器人系统来说，激励学习是实现具有自适应、自学习能力的智能移动机器人系统的重要途经，为解决智能系统的知识获取瓶颈问题提供了一条可行之路。本文为解决大状态空间问题和部分可观测问题，重点研究了如何将激励学习问题

转换成人工势场问题，即应用虚拟水流法给出激励势模型。

1.3 本文内容组织结构

首先，引言部分对课题研究的背景和现实意义及激励学习研究的现状进行了综述性介绍。

第二章，先要介绍了激励学习的一些基本知识。然后着重分析了当前激励学习中研究比较成熟的两种方法，瞬时差分法和Q学习方法。

第三章，介绍人工势场中斥力势函数，引力势函数的选取及人工势场的生成，同时分析了人工势方法的优缺点。

第四章，重点研究了如何将激励学习问题转换成人工势场问题，即应用虚拟水流法给出激励势模型。其思想是首先将激励学习模型转换成人工势场模型。然后应用虚拟水流法的思想解决激励势场模型中存在的局部最小问题。在每次试验中智能体能记住局部最小位置的最新的势能，以使下一次试验时智能体能获得更优的解。

最后，第五章用三个著名的网格世界问题对第四章中提出的激励势场模型进行了测试，同时也做了与Q学习、HQ学习等方法的对比实验。实验结果表明：对可观测和部分可观测激励学习新方法都能简洁有效地给出理想的解；与Q学习和HQ学习等方法相比激励势场模型更稳定有效。

第二章 激励学习

激励学习作为智能体的一种自主学习的方法，就是研究从交互中学习的计算方法。它源于动物心理学研究，最初可以上溯到巴普洛夫的狗的反射实验。近十年来，激励学习已经发展成为多学科的交叉领域。它包括人工智能、统计学、心理学、控制工程、运筹学、神经科学、人工神经网络和遗传算法等方面的研究工作。激励学习日益引人瞩目，已经广泛地应用于智能控制、机器人和决策分析等领域。这一章首先详细给出了激励学习的理论基础及基本概念；然后介绍了两种常用的激励学习算法：瞬时差分法（Temporal Difference Method TD）和Q学习算法（Q-learning Algorithm）；最后分析了Q学习的优缺点。

2.1 激励学习的理论基础及基本概念

激励学习是通过奖惩来激励Agent，并且无需说明任务是如何成功的。激励学习中要解决的关键问题，包括通过马尔可夫决策过程（Markov Decision Processes, MDPs）理论建立该领域的基础、从延迟的激励中学习、建立经验模型来加速学习、探索和利用的折衷（Trade-off Of Exploration And Exploitation）、一般化和层次化的利用等。解决激励学习问题主要有两种策略：第一种是在行为空间内搜索以找到在环境中执行得很好动作。这种方法和其它新颖的搜索技术一样，已经被应用于遗传算法和遗传设计。第二种方法是使用统计学技术和动态规划方法来估计在给定环境状态下采取的行为的效用。到底哪一种技术更好更为有效则取决于具体的应用研究。激励学习方法大多数都是建立在马尔可夫决策过程理论框架之上的，虽然激励学习方法并不局限于马尔可夫决策过程，但离散、有限状态的马尔可夫决策过程框架是激励学习算法的基础，这里首先简单介绍马尔可夫决策过程，然后激励学习标准模型说明其基本概念和原理。

2.1.1 马尔可夫决策过程

马尔可夫过程是具有无后效性的随机过程[63]。所谓无后效性是指：当过程在t_n时刻所处的状态为已知时，过程在大于t_n的时刻t所处状态的概率特性只与过程在t_n时刻所处的状态有关，而与过程在t_n时刻以前的状态无关。马尔可夫决策过程的基本概念来源于Bellman关于动态规划的研究工作以及Shapley的有关随机对策的论文，其理论基础则由Howard和Blackwell在其论著中奠定。马尔可夫决策过程（MDP）[1]按照决策时间的特性和模型中的其他因素的不同，可以有多种分类方法，如平稳和非平稳MDP、离散时间和连续时间MDP、离散状态空间和连续状态空间MDP

等。马尔可夫决策过程和动态规划的理论和应用研究已取得了大量的成果，并成为解决随机动态最优化问题的重要工具。在本文中将以离散时间平稳MDP为研究对象。离散时间平稳MDP的定义如下：

定义2.1（马尔可夫决策过程）离散时间马尔可夫决策过程(简称马尔可夫决策过程)由如下五元组定义：

$$\{S, A_{(i)}, P_{ij}^{a}, r_{ij}^{a}, V \mid i, j \in S, a \in A_{(i)}\}$$

其中，S 是环境状态集合；$A_{(i)}$是智能体在状态i的动作集合；P_{ij}^{a}是智能体在状态i采取动作a后转移到状态j的概率(状态转移概率函数)；r_{ij}^{a}是智能体在状态i采取动作a后转移到状态j时获得的瞬时奖赏；V 是评价函数或目标函数；奖赏函数为：$r_{ij}^{a}:\Gamma=\{(i,a)\mid i\in S, a\in A_{(i)}\}\rightarrow[-\infty,+\infty]$。

在马尔可夫决策过程模型中，智能体在与环境进行交互作用的每一步都感知环境的状态s_t($s_t\in S$, t = 0, 1, 2, …,)，然后选择一个动作a_t($a_t\in A_{(i)}$)，在执行动作a_t之后，环境反馈给智能体奖赏r_{t+1}和状态S_{t+1}。即马尔可夫决策过程的实质是当前状态向下一状态转移的概率和获得的奖赏值只取决于当前状态和所选择的动作，与历史状态和历史动作无关。

马尔可夫决策过程的历史是由相继的状态和决策组成，形式为：

$$h_n=(i_0,a_0,i_1,a_1,\ldots,i_{n-1},a_{n-1},i_n),\ n\geq 0 \qquad (2.1)$$

考虑单步的状态转移概率和期望奖赏：

$$P_{ss'}^{a}=P_r\{s_{t+1}=s' \mid s_t=s, a_t=a\} \qquad (2.2)$$

$$r_s^a=E\{r_{t+1}\mid s_t=s, a_t=a\} \qquad (2.3)$$

智能体的目标就是要学习这样一个最优策略$\pi:S\times A\rightarrow[0,1]$，使得目标函数V满足某一设定的要求。

因此根据马尔可夫决策模型，一般来说在状态转移概率函数和奖赏函数已知的环境模型中，可以采用动态规划技术求解最优策略；而激励学习则着重于研究状态转移概率函数和奖赏函数未知的情况下，系统如何学习到最优行为策略。

2.1.2 激励学习的几个基本概念

1.策略(Policy)

策略定义了Agent在给定时刻的行为方式，直接决定了Agent的动作，是激励学习的核心。策略的定义如下：

定义2.2 策略$\pi:S\rightarrow A$是一个从状态集合到动作集合的映射，描述了针对状态集合S中的每一个状态s，Agent应完成的一个动作a。

关于任意状态所能选择的策略组成的集合F，称为允许策略集合，$\pi\in F$。在允许策略集合中存在的使问题具有最优效果的策略π^*称为最优策略。策略与心理学中的刺激-反射(Stimulus-response)规则相对应，在某些情况下策略可能是一

个简单的函数或者查找表(Lookup Table)；而在另一些情况下则可能需要大量的计算，例如搜索过程等。激励学习方法确定了Agent怎样根据经验改变其策略。

2.报酬函数(Reward Function)

报酬函数定义了一个激励学习问题的目标，它将感知的环境状态(或状态一动作对)映射到一个报酬信号r，对产生的动作的好坏作一种评价。报酬信号通常是一个标量，例如用一个正数表示奖赏，而用负数表示惩罚.激励学习的目的就是使Agent最终得到的总的报酬值R达到最大。报酬函数往往是确定的、客观的，可以作为改变策略的标准。

3.值函数(Value Function)

报酬函数是对一个状态(或状态一动作对)的即时评价，而在选择行为策略过程中，要考虑到环境模型的不确定性和目标的长远性，因此在策略和瞬时报酬之间构造值函数(即状态的效用函数，又称评价函数)。

定义2.3(状态值函数) 状态s_t的值定义为Agent在状态s_t，执行动作a_t，及后续策略π所得到的累计报酬的期望，记为$V(s_t)$。

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = r_{t+1} + \gamma R_{t+1}, \quad 0<\gamma \le 1 \tag{2.4}$$

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\{R_t \mid s_t = s\} = E_{\pi}\{r_{t+1} + \gamma V(s_{t+1}) \mid s_{t+1} = s'\} \\ &= \sum_{a} \pi(s,a) \sum_{s'} P^{a}_{ss'} [R^{a}_{ss'} + \gamma V^{\pi}(s')] \end{aligned} \tag{2.5}$$

式(2.5)给出的是在π策略下，系统在状态s的值函数，即学习系统遵循π策略的情况下所能获得的期望累计折扣报酬总和。这是无限折扣模型，此外常用的值函数度量模型还有:

有限范围模型: $V^{\pi}(s_t) = \sum_{i=0}^{h-1} r_i$

平均回报模型: $V^{\pi}(s_t) = \lim_{h\to\infty}(\frac{1}{h}\sum_{i=0}^{h} r_i)$

在实际应用中由于带有折扣的模型在数学上较易控制。有时，记录状态-动作对的值比记录状态的值要方便，Watkins 把状态-动作对的值称为Q值，因此可定义另一种评价函数:Q函数。

定义2.3(Q值函数) Q值函数表示在状态s执行动作a及后续策略的报酬折扣和的期望，记为Q(s,a)。

$$Q^{\pi}(s,a) = E_{\pi}\{R_t \mid s_t = s, a_t = a\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a\} \tag{2.6}$$

状态值函数$V(s_t)$和Q值Q(s,a)函数都是对报酬的一种预测，值函数是激励学习方法的关键，绝大部分强化学习算法的研究就是针对如何快速有效地估计值函数.

4.环境的模型(Model of Environment)

环境的模型是某些激励学习系统的一个可选的组成部分。环境的模型就是模拟环境的行为方式，例如给定一个状态和动作，模型可以预测下一个状态和报酬。利用环境的模型，Agent在作决策的同时可以考虑未来可能发生但尚未实际经历的情形，进行规划((Planning)。将模型和规划加入到激励学习系统是一个比较新的发展，联系了强化学习与动态规划等其它基于模型或部分模型的方法，形成了一些较实用的新方法。

2.1.3 激励学习的模型

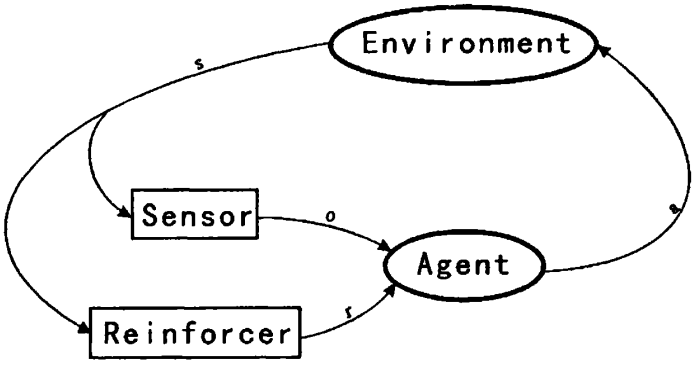


图 2.1 激励学习的标准模型

在标准的激励学习模型中，智能体通过观测和动作与它所处的环境关联在一起，如图 1 所示。智能体（Agent）与环境（Environment）的每步交互作用通过传感器都得到一个在一定程度上能反映当前状态的观测（o）。该观测作为输入。然后智能体选择一个动作（a）即产生输出。这一动作将改变智能体在环境中所处的状态，该状态的值转化成智能体的激励通过分级激励信号（r）传递给智能体。智能体选择的动作应该使激励信号最终值的和趋向增长。

形式上，激励学习模型可以用四个变量清楚地描述，它们是 S，A，O 和 R。这最初是由 Kearns 和 Coller 提出的[23]。

S，环境状态的离散集。S 可以由随机变量集 $S=\{S_1, S_2, \cdots, S_{|S|}\}$来描述，这里每个 S_i 的值在 $Val(S_i)$ 的每一有限范围内。$Val(S_i)$ 表示集合中的可能实例 S_i，$|S|$ 表示 S 的维数。状态 s 为：$s \in Val(S)$，也就是说 $s=\{s_1, s_2, \cdots, s_{|s|}\}, s_i \in Val(S_i), i=1, \cdots, |S|$。s(t)表示 t 时刻的状态，这里的 t 是离散的时间引索。

A，智能体的动作离散集。A 可以用随机变量集 $A=\{A_1, A_2, \cdots, A_{|A|}\}$，这里的 A_i 取值在 $Val(A_i)$ 的每一有限范围内，$Val(A_i)$ 表示集合中的可能实例 A_i，$|A|$ 表示 A 的维数。对每个动作 a 存在 $a \in Val(A)$，也就是说对 $a=\{a_1, a_2, \cdots, a_{|A|}\}$ 有 $a_i \in Val(A_i), i=1, \cdots, |A|$。a(t)表示 t 时刻的动作，这里的 t 是离散的时间引索。

O，智能体观测的离散集。O 可以用随机变量集合 $O=\{O_1, O_2, \cdots, O_{|O|}\}$ 来描述，这里的每一个 O_i 取值都在 $Val(O_i)$ 的每一有限集内，$Val(O_i)$ 表示集合中可能实例

O_i，$|O|$表示O的维数。对每个观测O存在$O \in Val(O)$，也就是说对$O=\{O_1, O_2, \cdots, O_{|O|}\}$有$O_i \in Val(O_i)$，i=1，…，$|o|$。O(t)表示t时刻的观测，这里的t是离散的时间引索。

R，智能体获得的分层激励信号集。R是奖赏函数，$R: S \rightarrow [-R_{max}, R_{max}]$（$R_{max}$是一实数，$R_{max} \in R$表示最大的可能奖赏），即$r=R(S')$表示智能体在状态S'获得的奖赏，S'是智能体在状态S选择动作a后的状态。有时奖赏函数与其说与状态有关不如说是与（状态，动作）对有关。为了简化表示我们假设奖赏仅依赖于状态，这并不影响结果。

2.1.4 激励学习的目标函数或优化标准

给定一个策略π从S到A的映射，记为$\pi: S \rightarrow A$，每个状态对应一个具体的动作。给定任意策略π，就可以定义目标函数或优化标准

$$\pi^* = \arg\max_{\pi} \left[E\left(\sum_{t=1}^{\infty} \gamma^{t-1} r(t) \right) \right] \qquad (2.7)$$

等式（2.7）说明模型的目标是随着时间的推移智能体的奖赏值达到最大或找到智能体的一个优化策略。这里的π^*是优化策略，r(t)是t时刻的奖赏，E()是数学期望，γ $(0 \le \gamma < 1)$为折扣因子以确保等式(2.7)中的和是有限值。

定义激励学习模型的四要素为状态(S)、动作(A)、观测(O)和奖赏(R)。激励学习中的“任务”或“问题”用奖赏函数能很清楚地描述。给定奖赏函数和领域模型<S,A,O>,优化策略就确定了。若智能体能准确地观测环境的状态，即S=O时，模型为可观测激励学习。不过智能体在许多现实世界环境中不可能对环境的状态有理想完整的观测，即$S \neq O$，此时模型为部分可观测激励学习。

2.2 激励学习的基本算法

激励学习问题通常是采用迭代技术更新值函数的估计值来获取最优策略。根据在学习过程中是否需要学习马尔可夫决策模型知识，激励学习方法可以分为基于模型（Model-based)和模型无关(Model-free)的两类。如，Dyna_Q算法属于基于模型的，瞬时差分法(Temporal Difference, TD)和Q学习算法为与模型无关的。一般地说，由于模型无关法没有充分利用学习中获取的经验知识，相比较于基于模型法收敛速度要慢；但模型无关法每次迭代计算量较小，且对动态未知环境适应性好，是激励学习中智能体最主要的学习方法之一。本文所研究的激励学习方法大都是模型无关法，下面主要介绍两种应用最广泛的基本算法：TD算法和Q学习算法。

2.2.1 瞬时差分方法

1988年Sutton提出的瞬时差分(Temporal Difference, TD)[23]学习是激励学习研究中一个非常重要的进展。TD方法是Monte Carto方法和动态规划技术的结合[64]，在动态规划中，状态转移概率函数和奖赏函数已知，从任意设定的策略π_0出发，可采用策略迭代的方法（如（2.8）式和（2.9）式）逼近最优的V*和π^*。

$$\pi_k(s)=\arg\max_a \sum_{s'} P_{ss'}^a[R_{ss'}^a+\gamma V^{\pi_{k-1}}(s')] \qquad (2.8)$$

$$V^{\pi_k}(s)\leftarrow \sum_a \pi_{k-1}(s,a)\sum_{s'} P_{ss'}^a[R_{ss'}^a+\gamma V^{\pi_{k-1}}(s')] \qquad (2.9)$$

TD方法直接从智能体经验中学习，同时利用估计的值函数（（2.10）式）进行迭代。

$$V(s_t)\leftarrow V(s_t)+\alpha[r_{t+1}+\gamma V(s_{t+1})-V(s_t)] \qquad (2.10)$$

最简单的TD算法是一种自适应策略迭代的一步TD算法，即TD(0)算法。所谓一步TD算法，是指智能体获取的瞬时奖赏值仅向后退一步，也就是只迭代修改相邻状态的估计值，其迭代公式如下：

$$V(s_t)=V(s_t)+\alpha[r_{t+1}+\gamma V(s_{t+1})-V(s_t)] \qquad (2.11)$$

其中$\alpha\in(0,1]$学习因子，$V(s_t)$指智能体t在t时刻的状态s_t时估计的状态值函数，$V(s_{t+1})$指智能体在t+1时刻的状态s_{t+1}时估计的状态值函数，r_{t+1}指智能体从状态s_t向s_{t+1}转移时获得的瞬时奖赏值，TD(0)具体算法如下：

TD(0)算法[8]

Initialize V(s) arbitrarily, π to the policy to be evaluated
Repeat (for each episode):
Initialize s
Repeat (for each step of episode):
a ←action given by π for s
Take action a: observe reward, r, and next state, s'
V(s) ← V(s)+α[r+γV(s')-V(s)]
$s\leftarrow s'$
until s is terminal

由于TD(0)算法中智能体获得瞬时奖赏只修改相邻状态的值函数估计，收敛速度较慢。因此为提高收敛速度，使智能体获得的瞬时奖赏可以后退任意步，即TD（λ））算法。其迭代公式（2.12）如下：

$$V(s_t)=V(s_t)+\alpha(r_{t+1}+\gamma V(s_{t+1})-V(s_t))e(s) \qquad (2.12)$$

其中e(s)定义为状态s的选举度(Eligibility Trace)，实际应用时可用下面公式

(2.13)计算。

$$e(s)=\begin{cases}\gamma\lambda e(s)+1, & \text{如果}s\text{是当前状态}\\ \gamma\lambda e(s), & \text{其它情况}\end{cases} \tag{2.13}$$

其中$0\le\lambda\le1$。当λ=0时，即为TD(0)算法的形式，参考文献[1]对e(s)作了进一步的讨论。TD(λ)的算法如下：

TD(λ)算法[8]

Initialize V(s) arbitrarily and e(s)=0, for all s∈S

Repeat (for each episode):

Initialize s

Repeat (for each step of episode):

a ← action given by π for s

Take action a : observe reward, r, and next state, s'

$\delta \leftarrow r+\gamma V(s')-V(s)$

$e(s)\leftarrow e(s)+1$

For all s :

$V(s)\leftarrow V(s)+\alpha\delta e(s)$

$e(s)\leftarrow\gamma\lambda e(s)$

$s\leftarrow s'$

Until s is terminal

2.2.2 Q学习算法

Q学习是由Watkins提出的一种模型无关的激励学习算法[6]，也可以被看作是一种异步动态规划方法。是目前使用最普遍的激励学习算法之一，尤其是在机器人控制中得到了广泛的应用。Q学习让Agent可以通过执行行为序列，而不是建立关于环境的映射，来学习在马尔可夫环境下应该如何优化地动作。也就是说Q学习的思想是不去估计环境模型，而是直接优化一个可迭代计算的Q函数[13]。Watkins定义此Q函数为在状态s_t时执行动作a_t，且此后按（2.14）式将最优动作序列执行时的折扣累计奖赏值。

$$Q_{t+1}(s_t,a_t)=r_t+\gamma\max_{a_t}\{Q(s_{t+1},a_t)\,|\,a_t\in A\} \tag{2.14}$$

Q学习通过在不同的时刻估计值函数，然后通过多次迭代学习逼近真实的值函数，因此可以认为Q学习算法是TD算法的一个特例;然而不同于一般的TD算法，Q

学习迭代时采用（状态，动作）对的奖赏和Q* (s, a)作为估计函数，在智能体每一次学习迭代时需要考察每一个行为，以确保学习过程的收敛。即按策略（2.15）在状态s下的最优动作为使Q(s, a)取得最大值的动作。

$$\pi^*(s) = \arg\max_a Q(s,a) \qquad (2.15)$$

类似TD学习，Q学习首先初始化Q（s，a）值；然后智能体在状态s，根据设定的探索策略，从动作集中选择一个动作a，到达下一个状态s'同时获得奖赏值r，并根据更新规则修改Q值；若智能体到达了目标状态，则结束本次迭代循环，智能体又从初始状态开始新一轮的迭代循环直至学习结束。

Q学习每一个阶段学习的步骤如下：

（1） 观察现在的状态s_t

（2） 选择并执行一个动作a_1

（3） 观察下一个状态s_{t+1}；

（4） 收到一个即时激励信号r_t；

（5） 按式(2. 16)，(2. 17)调整Q值：

$$Q_t(s_t,a_t) = \begin{cases} (1-\alpha_t)Q_{t-1}(s_t,a_t) + \alpha_t[r_t + \gamma V(s_{t+1})] & s=s_t;\ a=a_t \\ Q(s_t,a_t) & \text{其它情况} \end{cases} \qquad (2.16)$$

$$V(s_{t+1}) = \max_{a\in A}\{Q_{t-1}(S_{t+1},a)\} \qquad (2.17)$$

其中α_t是学习因子，γ是对$V(s_{t+1})$的折扣因子。

Q学习算法[13]

```
Initialize Q(s,a) arbitrarily
Repeat (for each episode):
    Initialize s
    Repeat (for each step of episode):
        Choose a from s using policy derived from Q (e.g., ε-greedy)
            Take action a, observe r  s'
            Q(s, a) ← Q(s, a) + α[r + γ max_a' Q(s', a') - Q(s, a)]
            s ← s'
    until s is terminal
```

针对确定性(Deterministic)和非确定性(Nondeterministic)，马尔可夫决策过程Q学习算法收敛性分别有如下定理[65]。

1. 对确定性Markov决策过程Q学习算法收敛性有：

定理2.1　Q学习在确定性Markov决策过程环境下，如果满足条件：

(1) 奖赏值有界，$(\forall s,a)|r(s,a)|\le c$

(2) 用lookup表来表示Q函数

(3) 所用折扣因子γ，$0\le\gamma<1$

(4) 每个状态一动作对都可以被无限多次访问

$$Q(s,a)\leftarrow r+\gamma\max_{a'}Q(s',a') \tag{2.18}$$

则根据上式(2.15)进行训练，设第n次更新Q值为$Q_n(s,a)$，则对于所有s，a，当$n\to\infty$时，$Q_n(s,a)$将以概率1收敛于最优$Q^*(s,a)$。

证明：定义Δ_n为$Q_n(s,a)$与$Q^*(s,a)$之间的最大误差

$$\Delta_n\equiv\max_{s,a}|Q_n(s,a)-Q^*(s,a)|$$

做第n+1次迭代更新之后

$$\begin{aligned}|Q_{n+1}(s,a)-Q^*(s,a)|&=|(r+\gamma\max_{a'}Q_n(s',a'))-(r+\gamma\max_{a'}Q^*(s',a'))|\\&=\gamma|\max_{a'}Q_n(s',a')-\max_{a'}Q^*(s',a')|\\&\le\gamma\max_{a'}|Q_n(s',a')-Q^*(s',a')|\\&\le\gamma\max_{s',a'}|Q_n(s'',a')-Q^*(s'',a')|\end{aligned}$$

$$|Q_{n+1}(s,a)-Q^*(s,a)\le\gamma\Delta_n$$

由于$(\forall s,a)$ $Q_0(s,a)$和$Q^*(s,a)$有界，则Δ_0有界$\Delta_n\le\gamma^n\Delta_0$，$0\le\gamma<1$，所以对所有s，a，当$n\to\infty$时，$\Delta_n\to0$。

2. 对非确定性Markov决策过程Q学习算法收敛性有：

定理2.2　Q学习在非确定性Markov决策过程环境下，如果除了满足定理2.1中的四个条件外，其学习因子$\alpha_n(0\le\alpha_n<1)$对所有s，a，还满足：

(1) $\sum_{n=1}^{\infty}a_n=\infty$

(2) $\sum_{n=1}^{\infty}[a_n]^2<\infty$

则$n\to\infty$当时，$Q_n(s,a)$以概率1收敛于最优$Q^*(s,a)$。

2.2.3 Q学习存在的问题

（1）收敛速度慢。标准Q学习的实现主要采用Lookup表格方法，每次只更新一个Q函数，由于没有明确的教师信号，仅仅依靠外部的评价来调整自己的行为，这样势必要经过一个漫长的学习过程。

（2）利用表格表示Q函数，当环境状态集合S，Agent可能的动作集合A较大时需要占用大量的内存空间。

（3）不具有“泛化”能力。即不能利用有限的学习经验和记忆对一个大范围空间的有效知识获取和表示。

（4）对大规模或连续的状态或行为空间，表格型的激励学习算法存在类似动态规划的“维数灾难”。

（5）存在搜索（Exploration）和利用(Exploitation)均衡的问题。选择最高Q值的动作会导致Agent总是沿着相同的路径而不可能探索到其它更好的值。

第三章 人工势场

机器人(智能体)路径规划是实现机器人智能的一关键技术。它的任务是在具有障碍物的环境中，按照一定的评价标准，寻找一条从起始状态(包括机器人的位置和姿态)到达目标状态(包括机器人的位置和姿态)的无碰路径。目前常用到的路径规划方法主要有人工势场法、栅格法和神经网络等，其中人工势法以其数学计算上的简单明了被广泛应用。这一章首先介绍了人工势场的基本思想；然后着重分析了人工势场中斥力势函数，引力势函数的选取及人工势场的生成；最后分析了人工势方法的优缺点。

3.1 人工势场

人工势场（Artificial potential field APF）的基本方法是梯度下降搜索法，即指向最小的势函数。为规避障碍物，在障碍物周围环绕斥力势场；在目标点环绕引力势场，形状通常为碗状，如果环境中没有障碍物则其能量能很好驱动机器人到达目标的中心位置。在有障碍物的环境中，在障碍物的位置排斥机器人的斥力势垒将叠加到引力势场中。机器人所受力的方向为势场的负梯度方向。这个力驱使机器人向着势能下降的方向运动直到势能最小的位置。

人工势场最早是 1986 年由 Khatib 提出来的[66]，应用于路径规划。基本思想是：构造目标位置引力势场和障碍物周围斥力势场共同作用的人工势场，通过搜索势函数的下降方向来寻找无碰撞路经。人工势场的方法在机器人的运动空间中创建了一个势场，如图 3.1 所示，箭头方向为势函数的下降方向。该势场由两部分组成：一个是引力势场，随着机器人与目标的距离的增加而单调递增，且方向指向目标；另一个是斥力势场，在机器人处于障碍物位置的时候有一个极大值，并随与障碍物距离的增大而单调递减，方向为背离障碍物的方向；整个势场是其引力势场和斥力势场两部分的叠加。机器人在引力势场的引力 F_{att} 和斥力势场的斥力 F_{rep} 共同作用下向目标运动，如图 3.2 所示。其中 Obstacle 为障碍物，Robot 为智能体（机器人），Goat 为目标；Repulsive Force 为障碍物对智能体的斥力，Attractive Force 为目标对智能体的引力，Resultant Force 为智能体所受斥力和引力的合力。数学上表示为搜索势函数的下降方向来寻找从初始位置到达目标位置的无碰撞路径。

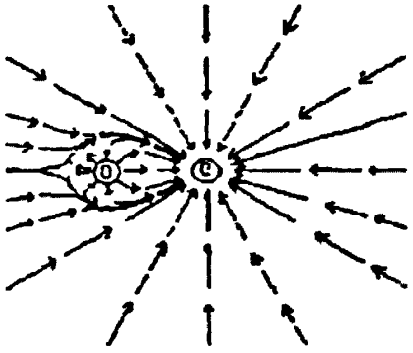


图 3.1 人工势场示意图

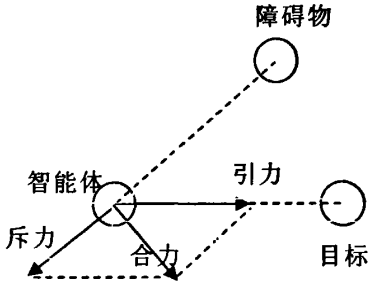


图 3.2 传统人工势场中智能体(机器人)受力示意图

3.2 势函数的选取

势场是对机器人运行环境的人为的抽象描述，障碍物和目标点按某种规则(势函数)产生一定的势场，机器人在势场中具有一定的势能。势函数分为斥力势函数和引力势函数。势函数应该满足连续，可导等一般势场所具有的性质。

3.2.1 斥力势函数的选取

在势场中，由障碍物产生的势场对机器人产生排斥作用，且距离越近，排斥作用应越大，即机器人与障碍物越近，说明机器人具有的势能越大，反之就越小。这种势场与电势场相似，势能与距离成反比，因此斥力势函数可取为：

$$U_{rep}=\begin{cases}k_1/\rho & \rho \le \rho_{\max}(\rho_{\max}\neq 0)\\ 0 & \text{其它情况}\end{cases} \tag{3.1}$$

其中 ρ 为机器人测量点与障碍物的距离，$\rho_{\max}$ 为势场作用的最大距离，k_1 为加权系数。

机器人受的斥力为：

$$F_{rep} = -\text{grad}\ (U_{rep}) = \begin{cases} k_1 / \rho^2 & \rho \le \rho_{\max}(\rho_{\max} \ne 0) \\ 0 & \text{其它情况} \end{cases} \tag{3.2}$$

斥力的角度为：

$$\theta = \pi + \text{arctg}(\frac{y_{obst} - y_{robot}}{x_{obst} - x_{robot}}) \tag{3.3}$$

即斥力的方向为背离障碍物。当 $\rho \to 0$ 时， $F_{rep} \to \infty$，即机器人与障碍物相碰时，受到的排斥力为无穷大。若要避免机器人与障碍物相碰，可设一个最小的安全距离 ρ_0，当 $\rho \to \rho_0$ 时， $F_{rep} \to \infty$。同时为使 F_{rep} 连续， F_{rep} 可修改为：

$$F_{rep} = \begin{cases} \frac{k_1}{(\rho - \rho_0)} - \frac{k_1}{(\rho_{\max} - \rho_0)} & \rho \le \rho_{\max} \\ 0 & \rho > \rho_{\max} \end{cases} \tag{3.4}$$

ρ_0， $\rho_{\max}$ 的大小取决于机器人的大小、速度、加速度以及环境中障碍物的稀疏程度等因素。

当障碍物在机器人与全局引力势场的运动路径上时，斥力势场排斥机器人以使其远离障碍物。斥力势场函数如下，即由Khatib提出的FIRAS函数：

$$U_{rep}(x) = \begin{cases} \frac{1}{2} k_r \left(\frac{1}{\rho(x)} - \frac{1}{\rho_{\max}} \right)^2 & \text{if } \rho(x) \le \rho_{\max} \\ 0 & \text{if } \rho(x) > \rho_{\max} \end{cases} \tag{3.5}$$

同样这里的 $\rho_{\max}$ 表示势场影响到的有限距离，$\rho(x)$ 是从 X 到障碍物 X_{rep} 的最短距离，$\rho(x) = \|x - x_{rep}\|$，$k_r$ 是比例增益函数。$\rho_{\max}$ 的选择要依据机器人的最大速度和控制周期[66]。斥力计算公式如下：

$$F_{rep}(x) = -\nabla U_{rep}(x) = \begin{cases} k_r \left(\frac{1}{\rho(x)} - \frac{1}{\rho_{\max}} \right) \frac{1}{\rho^2(x)} \frac{x - x_{rep}}{\rho(x)} & \text{if } \rho(x) \le \rho_{\max} \\ 0 & \text{if } \rho(x) > \rho_{\max} \end{cases} \tag{3.6}$$

3.2.2 引力势函数的选取

目标位置对机器人的势函数同样是基于距离的概念。目标位置距机器人距离越远，吸引作用越大，即机器人距目标位置越远，所具有的势能就越大；反之，距离越近，势能越小。当距离为零时，机器人的势能为零，此时机器人到达目标位置(终点)。这种性质与重力势能和弹性势能相似。而重力势能与距离成正比，弹性势能与距离的平方成正比。这样就有两类势函数可以用来类比。

(1)采用重力势函数以及由重力势函数产生的力：

$$U_{att}(X) = K_a \cdot \rho(X, X_{goal}) \tag{3.7}$$

$$F_{att} = K_a \tag{3.8}$$

(2)采用弹性势函数以及由弹性势函数产生的力：

$$U_{att}(X) = \frac{1}{2} K_a \cdot \rho(X, X_{goal})^2 \tag{3.9}$$

$$F_{att} = K_a \cdot \rho(X, X_{goal}) \tag{3.10}$$

式中：K_a为引力势场常量，X_{goal}为目标位置向量,引力的方向指向目标点。两种势函数得到的引力是不同的，采用弹性势函数在距离$\rho(X, X_{goal})$很大或很小时，产生的引力差别很大，如果障碍物产生的斥力不大而引力较大时，会是机器人与障碍物相撞。而采用重力势函数，产生的引力是常值。不过Andrews提出了通常称为二次抛物曲线拟合方法的引力势函数，即引力势场U_{att}随机器人远离目标而增加（像我们远离地球表面势能增加一样）。用等式（3.11）描述如下：

$$U_{att}(x) = \begin{cases} \frac{1}{2} k_a \rho_{goal}^2(x) & \text{if } \rho_{goal}(x) \le d \\ d k_a \rho_{goal}(x) & \text{if } \rho_{goal}(x) > d \end{cases} \tag{3.11}$$

这里的X表示机器人的位置矢量，$\rho_{goal}(x) = \|x - x_{goal}\|$表示从$X$到$X_{goal}$位置矢量的欧氏距离，d是方形范围的半径，另外$k_a$是比例增益函数。引力$F_{att}$可以从这个引力势场的负梯度方向得到，即：

$$F_{att}(x) = -\nabla U_{att}(x) = \begin{cases} -k_a (x - x_{goal}) & \text{if } \rho_{goal}(x) \le d \\ -d k_a \dfrac{x - x_{goal}}{\rho_{goal}(x)} & \text{if } \rho_{goal}(x) > d \end{cases} \tag{3.12}$$

3.2.3 全局势场的生成

全局势场可由引力势场和斥力势场的和得到。应用叠加原理可得到U(x)如下：

$$U(x) = U_{att}(x) + U_{rep}(x) \tag{3.13}$$

合力的计算公式如下：

$$F(x) = -\nabla U(x) = -\nabla U_{att}(x) - \nabla U_{rep}(x) = F_{att}(x) + F_{rep}(x) \tag{3.14}$$

3.3 应用人工势场法的优缺点

在机器人到达目标位置规避未知障碍物的过程中，势场法提供了一个简单的易于计算的方法。该方法有一个众所周知的经常被提及的缺陷就是机器人可能陷于势场的局部最小位置[67]。当机器人进入死胡同（如进入U形障碍物内）就会产生局部最小问题。多种不同类型的障碍物在一起时也可能会产生这样的问题。为了让机器人找到跳出局部最小位置的路径，提出了包括启发式和全局恢复等许多方法[68-70]。但这些方法都可能得不到最优路径同时增加了计算的复杂性。在局部路径规划中使用全局导航函数（局部极小处释放势能）的方法也被提出过[71-73]。不

幸的是，努力构造的全局导航函数只能适应于在目标Xgoal处存在一个局部极小的情况。

3.3.1 人工势场法的优点

(1)结构简单，使用方便，便于底层的实时控制。

(2)在进行运动规划时不仅考虑了机器人的避障和轨迹规划，而且也考虑了机器人的动态运动特性，使机器人运动更加自然和具有柔顺性，规划出来的路径一般比较平滑、安全。

3.3.2 人工势场法的缺点

(1)是一种局部寻优方法，只着眼于得到一条能够避障的可行路径，对路径是否最优没有加以考虑。

(2)陷阱区域。对简单的环境有效，但是对于复杂的多障碍物的环境容易产生局部极值点。势场合力为0 时，使得机器人在某点停止不能到达目标点。

(3)在障碍物前振荡。势场角度不容易控制，会因为机器人或者障碍物等位置的稍微改变而发生较大的变化，出现在障碍物前震荡的现象。

(4)相近的障碍物之间不能发现路径。

(5)在狭窄通道中摆动。

第四章 激励势场模型

激励势场(Reinforcement Potential Field RPF)模型是结合激励学习和人工势场的优点应用虚拟水流法构建的一个具有记忆学习功能的模型。激励势场相对 Q 学习等方法在计算的时间复杂度和所需的空间复杂度方面都有所降低，同时虚拟水流法的应用使势场中可能出现的局部极小问题得以巧妙解决。激励势场模型对较大状态空间和部分可观测环境能有效地给出理想的解。这一章首先详细介绍了激励势场模型的有关知识，然后给出了激励势模型的算法。

4.1 激励势场模型

激励势场模型是通过比较激励学习模型与人工势场模型在结构上存在许多相似，利用它们各自的优点将其结合而成的一个具有记忆学习功能的新模型。激励势场是在势场中引入激励，与人工势场相似全局激励势场由引力势场和斥力势场构成。下面是对激励势场中引力势场、斥力势场和全局激励势的描述。

4.1.1 引力源与斥力源集合的定义

给定 s 为状态变量，r 为奖赏(激励值)，R 为激励集合，S 为状态集合则有：

定义1：引力源集合 S^+。如果智能体在状态s获得的奖赏为r，$r\in R$, $s\in Val(S)$,

且r>0,定义状态S为一引力源，则引力源集合可以表示为：

$$S^+=\{s:r=R(s)>0,\ s\in Val(S)\}$$

定义2：斥力源集合 S^-。如果智能体在状态s获得的奖赏为r，$r\in R$, $s\in Val(S)$,

且r<0,定义状态S为一斥力源，则斥力源集合可以表示为：

$$S^-=\{s:r=R(s)<0,\ s\in Val(S)\}$$

4.1.2 引力势场的描述

当前状态S的引力势场 $U_{att}(s)$ 按等式（4.1），根据等式(3.11)让 k_a 取在引力源状态 s_{att} 时的奖赏 r_{att} 的绝对值，记为：$|R(s_{att})|$。在当前状态的引力势场可描述如下：

$$U_{att}(s)=\sum_{s_{att}\in S^+}U(s_{att})=\begin{cases}\sum_{s_{att}\in S^+}\frac{1}{2}|R(s_{att})|\rho_{att}^2(s) & \text{if } \rho_{att}(s)\le d\\ \sum_{s_{att}\in S^+}d|R(s_{att})|\rho_{att}(s) & \text{if } \rho_{att}(s)>d\end{cases} \tag{4.1}$$

这里的s表示智能体所处的当前状态，s_{att}表示引力源的状态（如定义1所示），$\rho_{att}(s)=\|s-s_{att}\|$表示从状态s到引力源状态$s_{att}$的欧氏距离，d 为方形范围的半径。为了简单起见，这里设定d为正无穷，即$d=+\infty$。

4.1.3 斥力势场的描述

当前状态s的斥力势场$U_{rep}(s)$按等式（4.2），根据等式(3.5)让k_r取在斥力源状态s_{rep}时的奖赏r_{rep}的绝对值，记为：$|R(s_{rep})|$。在当前状态的斥力势场可描述如下：

$$U_{rep}(s)=\sum_{s_{rep}\in S^-}U(s_{rep})=\begin{cases}\sum_{s_{rep}\in S^-}\frac{1}{2}|R(s_{rep})|\left(\frac{1}{\rho_{rep}(s)}-\frac{1}{\rho_0}\right)^2 & \text{if } \rho_{rep}(s)\le\rho_0\\ 0 & \text{if } \rho_{rep}(s)>\rho_0\end{cases} \quad (4.2)$$

这里的s表示智能体所处的当前状态，s_{rep}表示斥力源的状态（如定义2所示），$\rho_{rep}(s)=\|s-s_{rep}\|$表示从状态s到斥力源状态$s_{rep}$的欧氏距离，$\rho_0$势场影响的有限距离。

4.1.4 全局激励势场的生成

状态s的全局激励势场$U(s)$可以通过求引力势场和斥力势场的和得到，如等式（4.3）所示。

$$U(s)=U_{att}(s)+U_{rep}(s) \quad (4.3)$$

通过上述变换，一个激励势场模型就被构建出来了。等式(4.1)和(4.2)清楚地表明激励势场模型的要素，欧氏距离$\rho(\ )$的计算需要<S,O,A>要素的陈述。激励势场模型既可以用于可观测的激励学习，也可以用于部分可观测的激励学习。S表示可观测激励学习和部分可观测激励学习的环境状态，很多研究者已将环境状态S转换成预测状态智能体的内状态[1]。为简化计算和提高程序的应用效率，一个理想的选择就是对内状态的表示采用无记忆的方法，这里简单地将观测集（O）当作状态集（S）。激励势场模型的目标是让智能体获得的奖赏在一段时间内达到最大或找到一个优化策略。这与激励学习模型的目标相似，除非等式（2.7）中的折扣因子γ是多余的并取值为1。

4.2 虚拟水流法

这部分为参考同一课题组提出的虚拟水流法。这是我们为解决激励势场（RPF）中的局部最小问题而提出的。水的流动具有两个特征：①从高处往低处流；②流

到凹处注水。下面的步骤就是根据这两个特征来执行的。

第一步：判断机器人是否处于局部最小位置。如果 $U(s)<\min_{s'\in neighbor(s)}U(s')$ 并且当前状态S不是目标状态，则智能体陷于局部最小位置。这里的 $neighbor(s)$ 表示当前状态的邻域集。如果在当前状态S执行动作 $a\in Val(A)$ 后到达新状态 s'，则 $s'\in neighbor(s)$。这里的A是智能体动作离散集，如2.1.3部分所描述。

第二步：增加局部最小位置S状态处的势能。如果智能体处在局部最小位置S状态处，则按如下规则增加势能 $U(s)$：

$$U(s)=\begin{cases}\min_{s'\in neighbor(s)}U(s')+(\nu-1)|U(s)| & \text{if } \min_{s'\in neighbor(s)}U(s')=0\\ \min_{s'\in neighbor(s)}U(s')+(\nu-1)\left|\min_{s'\in neighbor(s)}U(s')\right| & \text{otherwise}\end{cases} \tag{4.4}$$

这里的 ν 表示注水速度，如果 $\nu\geq 1$ 定义等式（4.4）表示“注水”过程，否则表示“释水”过程。等式(4.4)利用水流的第二个特征(流到凹处注水)来解决局部最小问题，智能体为进一步导航将记住新调整的势能。让智能体导航的过程中记住新调整的势能，即使智能体具有记忆和学习的能力。

第三步：检查并选择最佳邻域。如果当前状态S不是局部最小状态，智能体将选择邻域S中的最小势(最佳位置)作为下一个状态 s'。这里利用了水流的第一个特征，“水从高处往低处流”。

这里的虚拟水流法是受流水特征的启示为了解决标准人工势场模型中的局部极小问题而提出的。方法的步骤如上面所描述那样分三步。假定机器人导航的环境的全局势场形状如图4.1。机器人运动要么向前要么退后，也就是说机器人的动作集中只包括“向前”和“退后”两个动作。

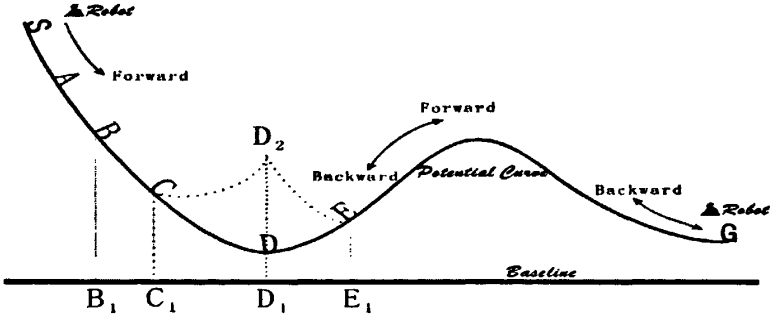


图4.1 机器人使用虚拟水流法如何解决局部最小问题

假设C为机器人所在的当前位置，势能$U(C)=CC_1$，B和D是C的邻域，它们的势能分别为$U(B)=BB_1$，$U(D)=DD_1$，因为$U(D)<U(C)<U(B)$，所以C不是局部最小点，因此机器人直接从高势能C向低势能D移动。当机器人到达位置D时，C和E是D的邻域。由$U(C)=CC_1$，$U(D)=DD_1$，$U(E)=EE_1$,并且$DD_1<EE_1<CC_1$，根据第一步可知D是一个局部最小位置。在这种情况下运用等式(4.4)调整D处的势能。根据等式(4.4)将D处的势能调整为$U(D)=D_1D_2=vEE_1$。若$v=2$，则$U(D)=D_1D_2=2EE_1$。如图4.1，用点划线表示将额外的势能加入原来的全局势能中。这样D点的势能就比它周围的势能高了，机器人就可跳出局部最小位置D并运动到新的位置E。势能可能要继续修改直到机器人到达目标位置G。这正像自然界中水流的过程。

4.3 激励势场的算法

1. 设置全局变量 RESTWATER。//其列向量由状态变量和该状态的势构成。
   给动作集赋值；
   设置机器人初始位置；
   设置引力源(目标位置，该状态的瞬时奖赏值）和斥力源(各障碍物的位置，该状态的瞬时奖赏值);
   设置引力源状态集和斥力源状态集；
   设置最大影响距离 DM（DM=3），涨水速度 RS（RS=2），最大步数(maxsteps=10000);
2. 如果没有达到目标且循环的步数小于最大步数 maxsteps, 则进行下列循环：
   (1)将当前状态的势和执行动作集中的每一个动作后的下一个状态的势进行比较，以选择最佳动作，确定势最小的为下一个最佳状态。
   导入全局变量 RESTWATER;
   ①计算当前状态的势：
   如果当前状态曾涨过水，则返回当前状态的势为上次调整后的势；
   否则：如果引力源非空,则计算所有引力源在该处的引力势并求和；
   如果斥力源非空,则计算所有斥力源在该处的斥力势并求和；

返回引力势和斥力势的和为当前的势。

②初始化当前状态的邻接状态及其势；

③获取当前状态的邻接状态及计算其势，即对动作集中的每一个动作进行如下操作：

获取执行动作后的状态，并记录下来。

计算该状态的势，并记录下来。

④从邻接状态中选取势最小的(minP)；

⑤如果没有引力源，则随机选取一动作为最佳动作，执行该动作到达的状态为下一最佳状态，并返回。

如果邻接状态中最小的势不大于当前状态的势，则最小势的状态为最佳状态，到达该状态的动作为最佳动作。

如果邻接状态的最小的势大于当前状态的势，则使用虚拟水流法：

若 minP=0, 则置当前势(currP)为：

currP = minP+(Rising-1.0)*asb(currP)；

否则：currP = minP+(Rising-1.0)*abs(currP)。

将 currState 和 currP 记录在 rwater 中；置：cs=currState, rw=RESTWATER。

(2)记录下一个最佳状态和选择的最佳动作；

(3)判断此状态：

若为目标状态，则输出“达到目标状态！”字样，并终止循环。

若不是障碍物，则将其设为当前状态。转第 2 步。

第五章 实验仿真与结果分析

著名的网格世界问题是激励学习中用来测试新算法经典基准，它易于理解并能很好地展现问题。因此这里应用它来测试所提出的激励势场模型。

第一个实验使用的是完全可观测四房间网格世界环境[74]，其最优解为20步。为了使问题更富有挑战性，我们使用了两条线路。一个使原始问题仅为部分可观测；另一个是使原始问题变为较大状空间下的一个复杂问题。即第二个实验使用的是部分可观测四房间网格世界环境[74]，其最优解也是20步。第三个实验为Wiering的钥匙与门迷宫问题的环境[75]，这是复杂环境下的一个相当复杂的任务，里面智能体先要找到钥匙打开门后才能规划出从起始位置到目标位置的路径。它的最优路径需要83步。

5.1 完全可观测四房间网格环境

5.1.1 问题描述

如图5.1，这里选用的四房间网格世界是由13x13个网格组成。里面包括各种障碍物，在底部右下角用字母G表示的一个目标和用字母S表示的起始位置。智能体在这个世界里能够从向北，向南，向东和向西移动四个动作中选择动作。智能体运动的方向如果不是目标所在方向得到的奖赏为0，如果是向着目标所在方向运动则获得的奖赏为1。状态的转变是确定的，并且每一次向适当的方向前进为一个方格。如果智能体选择一个动作，执行时遇到障碍物或墙，那么智能体的位置将保持不变并给以-1为奖赏；若智能体到达了目标状态则重置到用字母S表示的起始状态；直至得到最优路径。

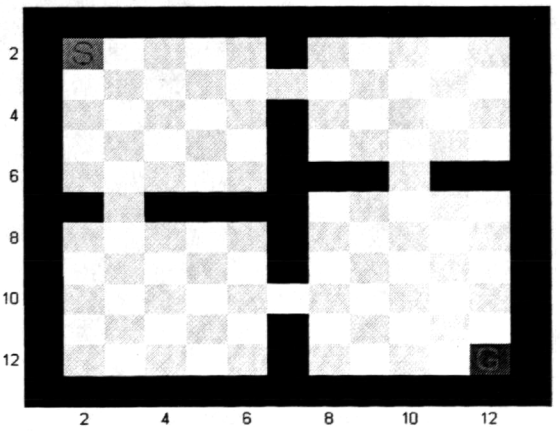


图5.1 13×13四房间网格世界问题

5.1.2 模型描述

RL模型. 环境状态：$S=\{S_1,S_2\}=\{Row,Column\}$，$|S|=2, Val(S_1)=\{1,\cdots,13\}$，$Val\{S_2\}=\{1,\cdots,13\}$；智能体的动作：$A=\{A_1\}$，$|A|=1$，$Val(A_1)=\{down,up,right,left\}=\{1,2,3,4\}$；智能体的观测：S=O，模型为完全可观测的激励学习模型；奖赏函数：R={1，-1，0}，当智能体到达目标位置时取1，当智能体撞上障碍物时取-1，其它情况取0。

激励势场模型. 引力源：$S^+=\{s:r=R(s)=1\}=\{(12,12)\}$，$S^+$包括1个状态，如图5.1所示；斥力源：$S^-=\{s:r=R(s)=-1\}=\{(1,1),(2,1),\cdots,(13,13)\}$，$S^-$包括69个状态；

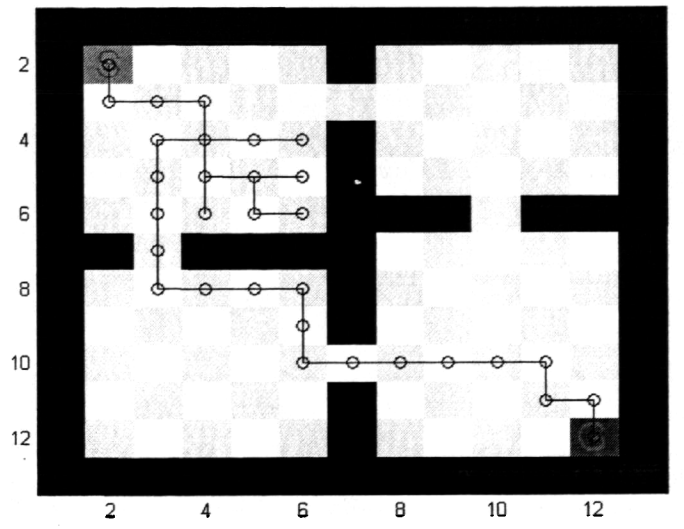
引力势：$U_{att}(s)=\sum_{s_{att}\in S^+}\frac{1}{2}|R(s_{att})|\rho_{att}^2(s)=\frac{1}{2}\sum_{s_{att}\in S^+}\rho_{att}^2(s)$；

斥力势：$U_{rep}(s)=\begin{cases}\frac{1}{2}\sum_{s_{rep}\in S^-}\left(\frac{1}{\rho_{rep}(s)}-\frac{1}{\rho_{\max}}\right)^2 & \text{if } \rho_{rep}(s)\le\rho_{\max}\\ 0 & \text{otherwise}\end{cases}$，

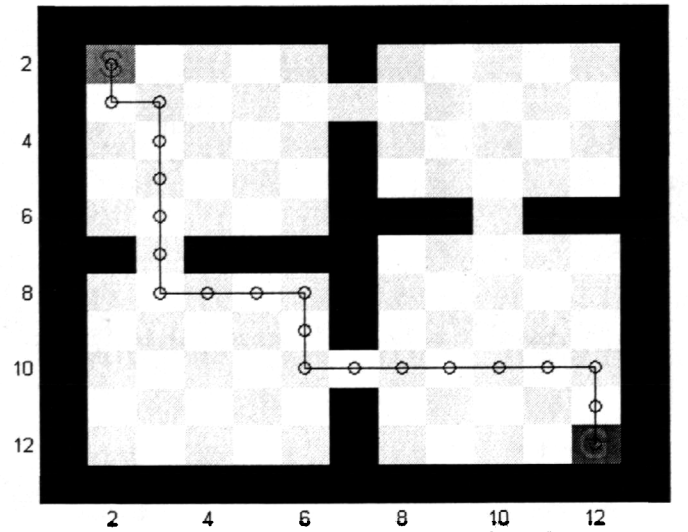
这里设定$\rho_{\max}=3$；实验中注水速度(如4.2部分与等式(4.4)所描述的)设定$\nu=2.0$。

5.1.3 应用激励势场模型进行实验的结果

将提出的激励势场模型应用到完全可观测的激励学习模型中，图5.2描述了两次实验中得到的路径。在第二次实验中智能体能得到稳定的优化路径。在第一次智能体从开始位置到目标位置需要38步，如图5.2(A)。很明显这不是最优的因为智能体在“注水”过程中用去了多出的步数。图5.2(B)画出了第二次实验中得到的路径。智能体到达目标仅用了20步，这确实是一条最优路径。测试的结果表明，我们提出的激励势场模型对完全可观测的激励学习问题是非常有效的。



(A) Path generated from [2,2] to [12 12], Steps: 38

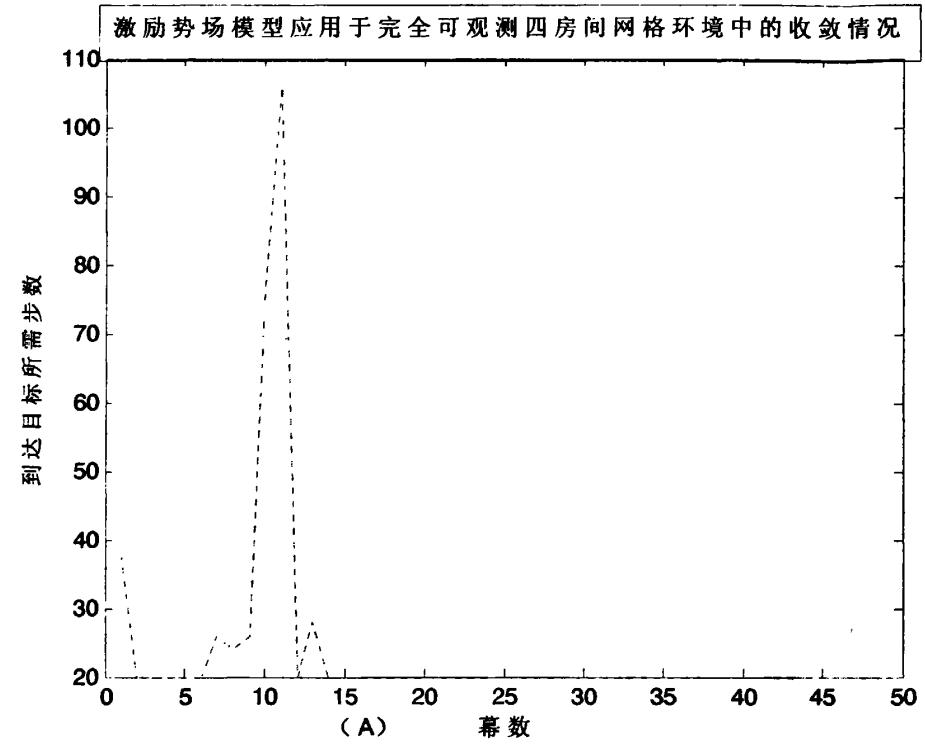


(B) Path generated from [2,2] to [12 12], Steps: 20

图5.2(A-B)激励势场模型应用到完全可观测四房间网格世界环境中的仿真结果

5.1.4 与Q学习进行比较的实验结果

图5.3(A—B) 为在完全可观测四房间网格世界环境中激励势场模型与Q学习的收敛情况。其中图（A）为用激励势场模型进行实验的结果；图（B）为用Q学习方法进行实验的结果。实验结果表明应用激励势场模型通过14幕学习就得到最优路径并收敛，同样的环境下用Q学习方法需要通过21幕学习才能得到最优路径并收敛；另外实验中可以看到同样的环境下每一幕的学习所需时间激励势场模型更是明显优于Q学习。



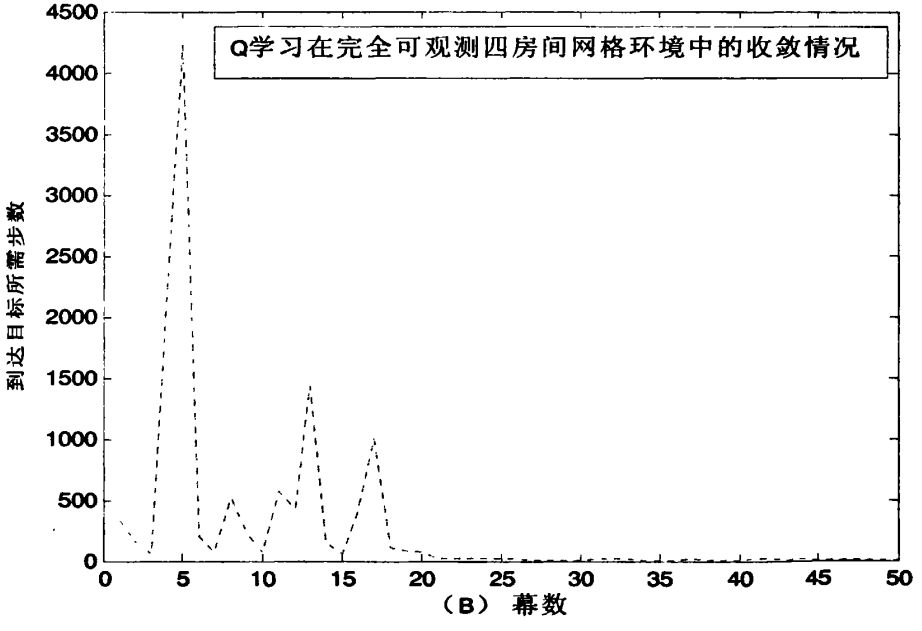


图5.3(A-B)在完全可观测四房间网格世界环境中激励势场模型与Q学习的收敛情况

5.2 部分可观测四房间网格世界环境

5.2.1 问题描述

像Littman所做的那样[74]，如图5.4，四房间网格世界通过引入智能体加以修改成四方位四房间网格世界。智能体通过观测四个方位距离当前状态最近的障碍物形成状态表示。这同用笛卡儿坐标表示智能体位置不同。它更能反映机器人在建筑物中所面对的问题。一个四方位四房间网格世界就是多场景的别称。环境由169(13x13)个状态组成，其中有92个清楚的观测。模糊状态的比率(the ratio of aliased states)为：

$$Alias\ Ratio = 1 - \frac{amount\ of\ observations}{amount\ of\ states} = 1 - \frac{92}{169} \approx 0.544 .$$

根据我们的经验，这个值越大，问题就越难。这里对可观测的激励学习有：$Alias\ Ratio = 0$；对部分可观测的激励学习有：$0 < Alias\ Ratio < 1$；对不可观测的激励学习有：$Alias\ Ratio = 1$。

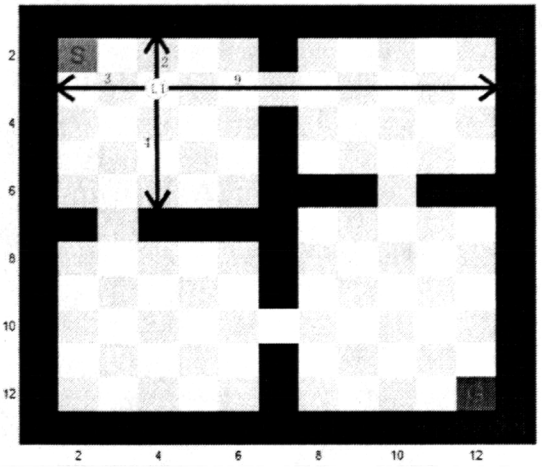


图5.4 部分可观测四房间网格世界环境

5.2.2 模型描述

RL模型. 环境状态：$S = \{S_1, S_2\} = \{Row, Column\}$，$|S| = 2$，$Val(S_1) = \{1, \cdots, 13\}$，$Val\{S_2\} = \{1, \cdots, 13\}$；智能体的动作：$A = \{A_1\}$，$|A| = 1$，$Val(A_1) = \{down, up, right, left\} = \{1,2,3,4\}$；智能体的观测(如图5.4所描述)：

$O = \{O_1, O_2, O_3, O_4\}$，$|O| = 4$，这里$O_1$表示当前状态右方离最近障碍物的距离，$O_2$表示当前状态左方离最近障碍物的距离，$O_3$表示当前状态后方离最近障碍物的距离，

O_4表示当前状态前方离最近障碍物的距离。$S \neq O$为部分可观测激励学习模型；奖赏函数：R={1，-1，0}，当智能体到达目标位置时取1，当智能体撞上障碍物时取-1，其它情况取0。

激励势场模型. 引力源：$S^+ = \{s : o_{att}\}$，S^+包括一个目标状态，o_{att}为智能体从当前状态到目标状态的观测距离；斥力源：$S^- = \{s : o_1, o_2, o_3, o_4\}$，$S^-$包括当前观测(O)四个元素相应的四个障碍物的状态。

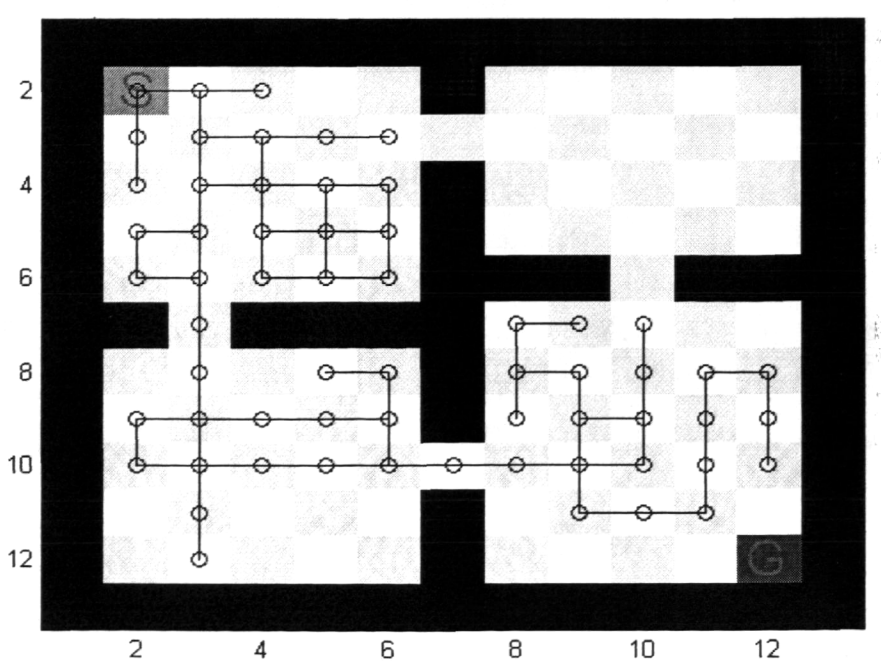
引力势场：$U_{att}(o) = \frac{1}{2} o_{att}^2$

斥力势场：$U_{rep}(o) = \frac{1}{2} \sum_{i=1}^{|Val(S^-)|} \frac{1}{o_i^2}$

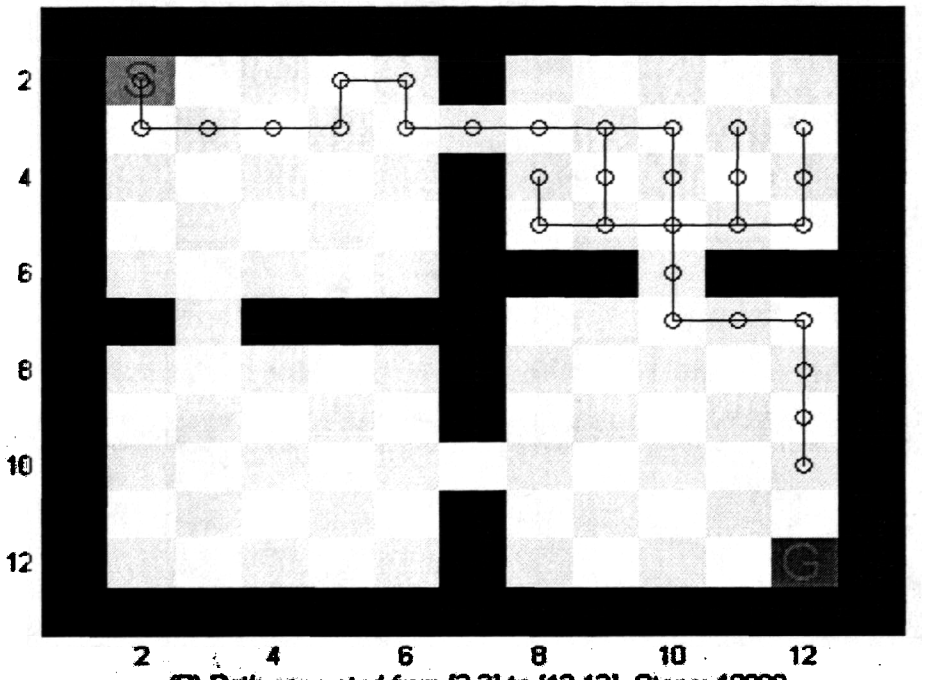
这里$|Val(S^-)|$表示S^-包括元素的个数；实验中注水速度(如4.2部分与等式(4.4)所描述的)设定$\nu = 2.0$。

5.2.3 应用激励势场模型进行实验的结果

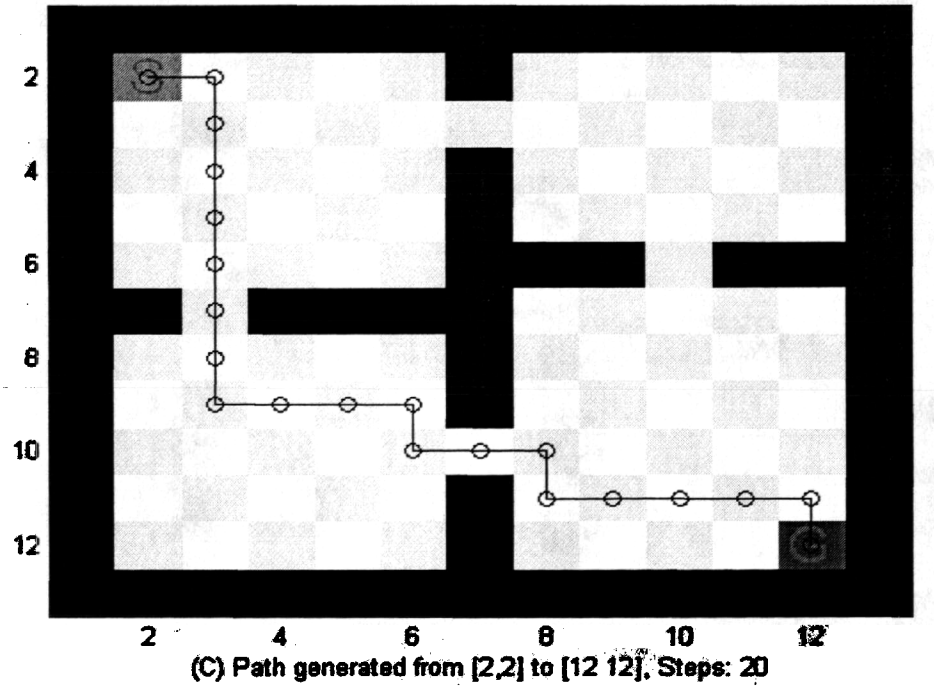
图5.5为用新方法进行八次实验得到的路径。智能体在第八次实验中获得优化路径(从开始位置S到目标位置G仅需20步)。与图5.2相比，在部分可观测环境中需要更多的“注水”过程和更多的试验次数。在第一次和第二次试验中智能体在有限的10000步内不能到达目标位置，如图5.5(A)和(B)所示。等式(4.4)表明智能体对局部最小处势能的调整具有记忆功能。因此在获得最优路径之前智能体的导航性能会越来越好，图5.5(C-H)也表明了这一点。



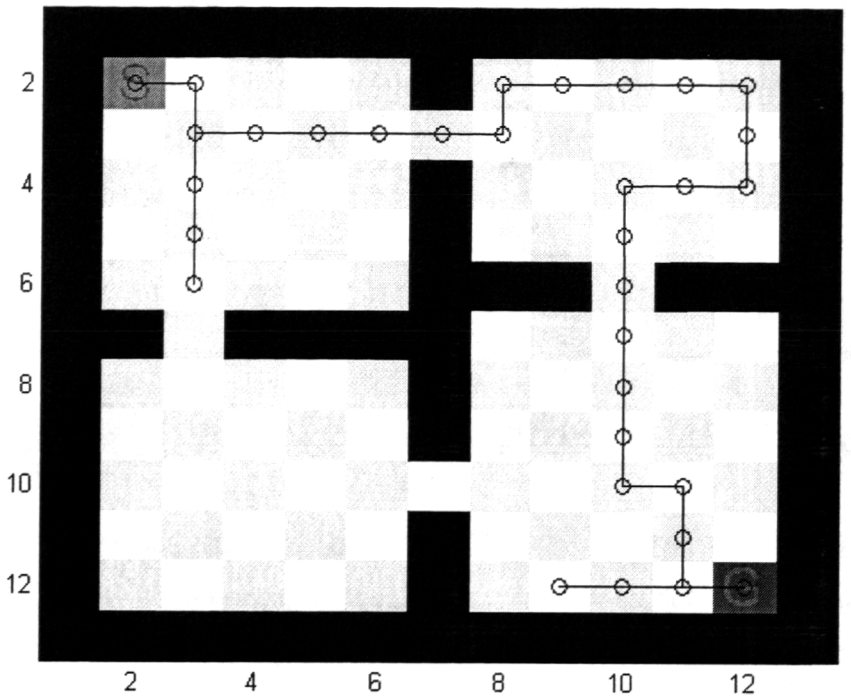
(A) Path generated from [2,2] to [12 12], Steps: 10000



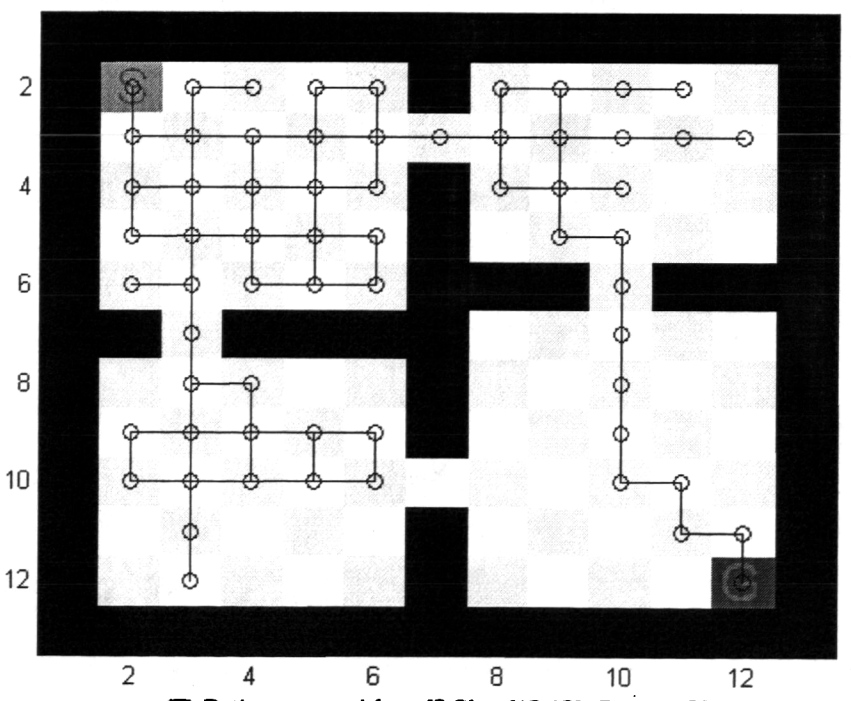
(B) Path generated from [2,2] to [12 12], Steps: 10000



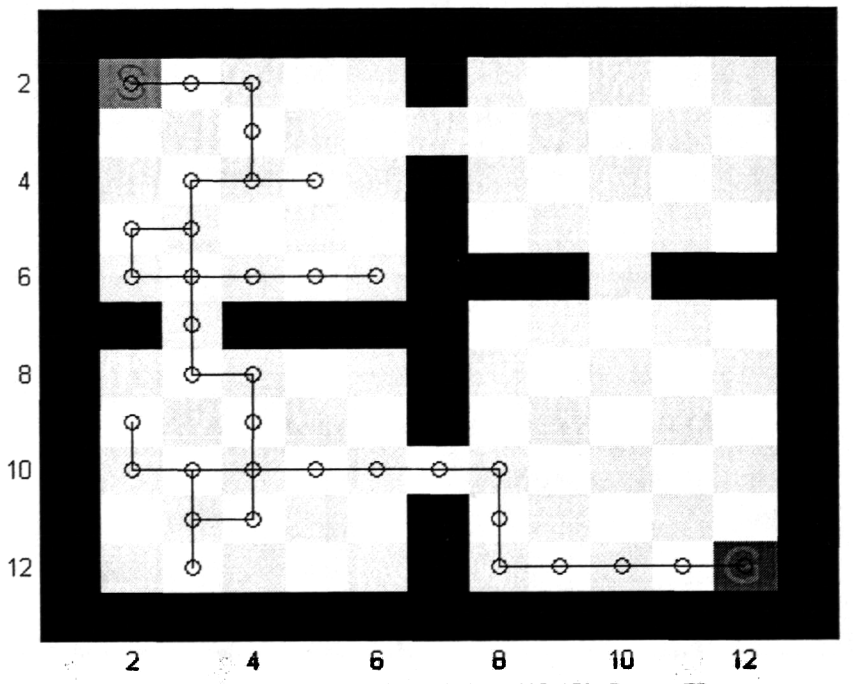
(C) Path generated from [2,2] to [12 12], Steps: 20



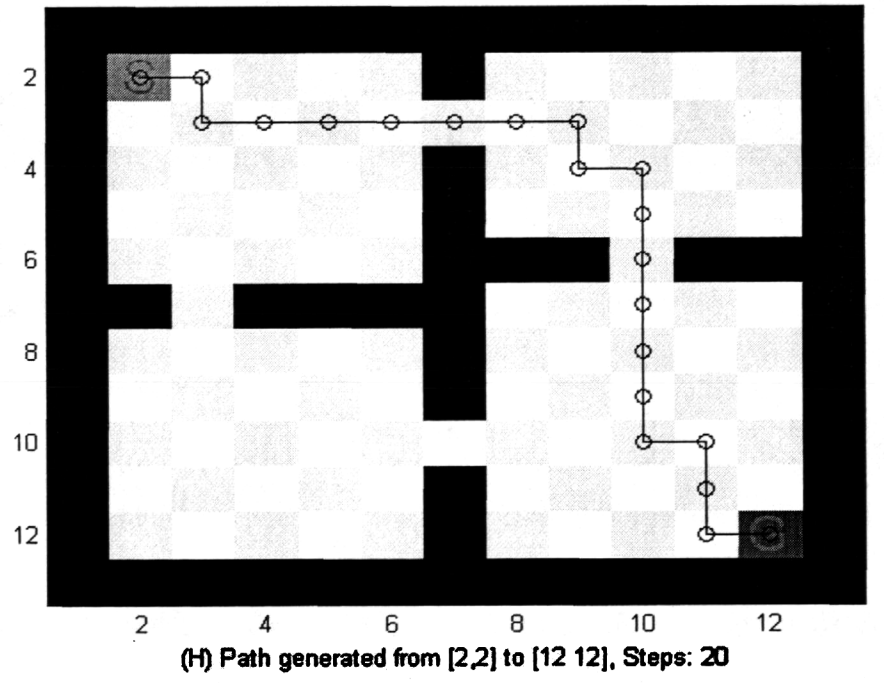
(D) Path generated from [2,2] to [12 12], Steps: 36



(E) Path generated from [2,2] to [12 12], Steps: 180



(F) Path generated from [2,2] to [12 12], Steps: 56



(H) Path generated from [2,2] to [12 12], Steps: 20

图5.5(A-H) 激励势场模型应用到部分可观测四房间网格世界环境中的仿真结果

5.3 钥匙与门迷宫问题

5.3.1 问题描述

这个实验环境是28×25的迷宫，如图5.6。智能体从位置S处开始走要先去位置

K取回钥匙，再走向“门”(Door)的位置(门的位置是通常情况下看似一堵墙的阴影区域，只有当智能体拥有钥匙后“门”才会打开，阴影区域自动消失)打开门继续向目标位置G前进。一旦智能体到达目标位置，它就会获得值为10的奖赏。如果智能体撞着了障碍物(黑区)，它就会获得值为-1的奖赏。在其它情况下智能体获得的奖赏值为0。对智能体取得钥匙或走过门都没有额外的中间奖赏。

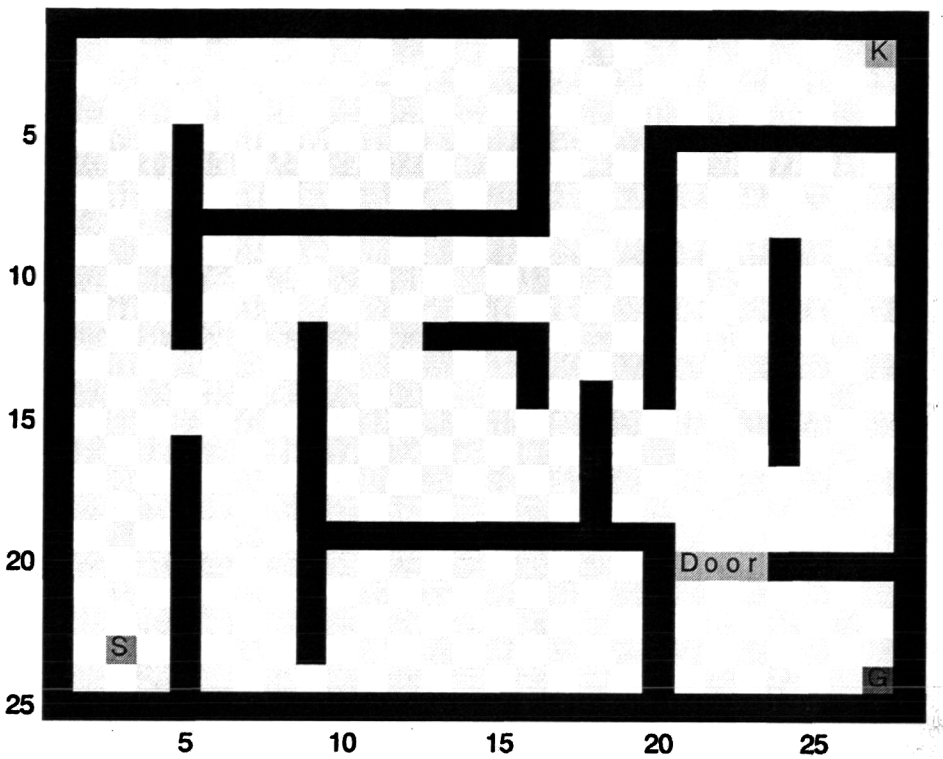


图5.6 钥匙与门28×25的迷宫环境

5.3.2 模型描述

RL模型. 环境状态：$S=\{S_1,S_2,S_3\}=\{Row,Column,HaveKey\}$，$|S|=3$，$Val(S_1)=\{1,\cdots,25\}$，$Val\{S_2\}=\{1,\cdots,28\}$，$Val\{S_3\}=\{0,1\}$，因此状态空间包括25×28×2 = 1400个状态；智能体的动作：$A=\{A_1\}$，$|A|=1$，$Val(A_1)=\{down,up,right,left\}=\{1,2,3,4\}$；智能体的观测：S=O，为完全可观测的激励学习模型；奖赏函数：R={1，-1，0}，当智能体到达目标位置时取1，当智能体到达的是障碍物位置时取-1，其它情况取0。

激励势场模型. 引力源：$S^+=\{s:r=R(s)=10\}=\{(24,27,1)\}$，$S^+$包括一个状态如图7所示；斥力源：$S^-=\{s:r=R(s)=-1\}=\{(1,1,0),(1,2,0),\cdots,(25,28,1)\}$，$S^-$包括411个状态，其中有钥匙的状态为204个，没有钥匙的状态为207个。

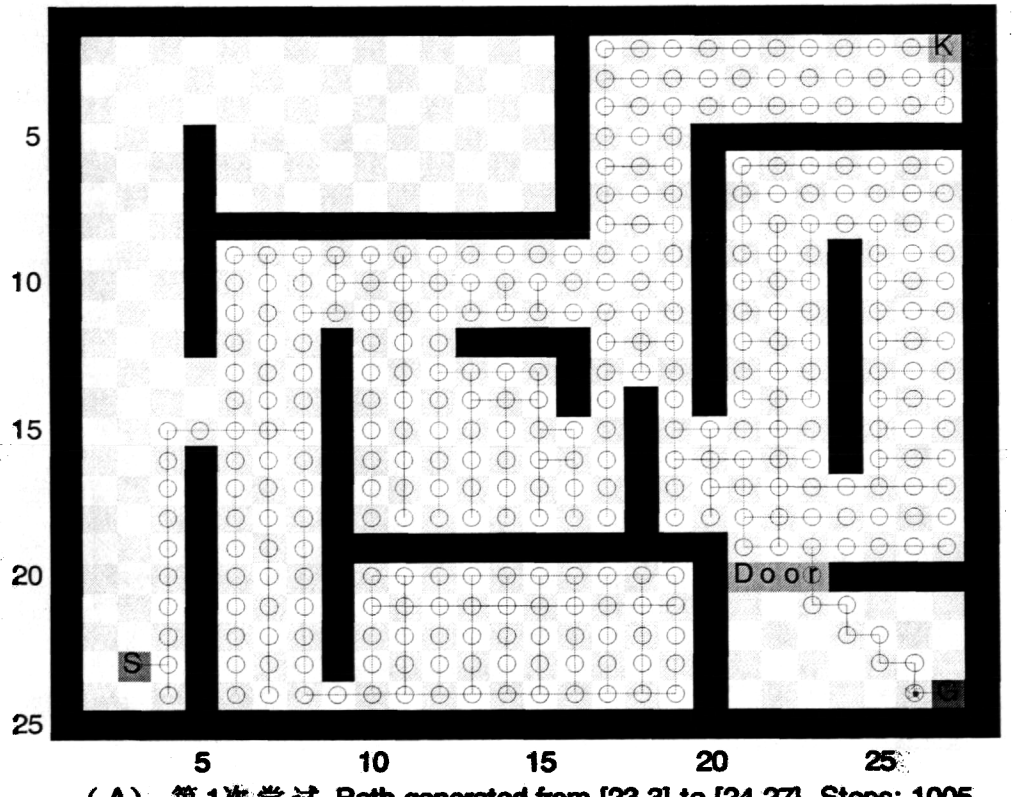
引力势场：$U_{att}(s)=\sum_{s_{att}\in S^+}\frac{1}{2}|R(s_{att})|\rho_{att}^2(s)=\frac{1}{2}\times 10\sum_{s_{att}\in S^+}\rho_{att}^2(s)$；

斥力势场：$$U_{rep}(s)=\begin{cases}\frac{1}{2}\sum_{s_{rep}\in S^-}\left(\frac{1}{\rho_{rep}(s)}-\frac{1}{\rho_0}\right)^2 & \text{if } \rho_{rep}(s)\le\rho_0\\ 0 & \text{otherwise}\end{cases}$$

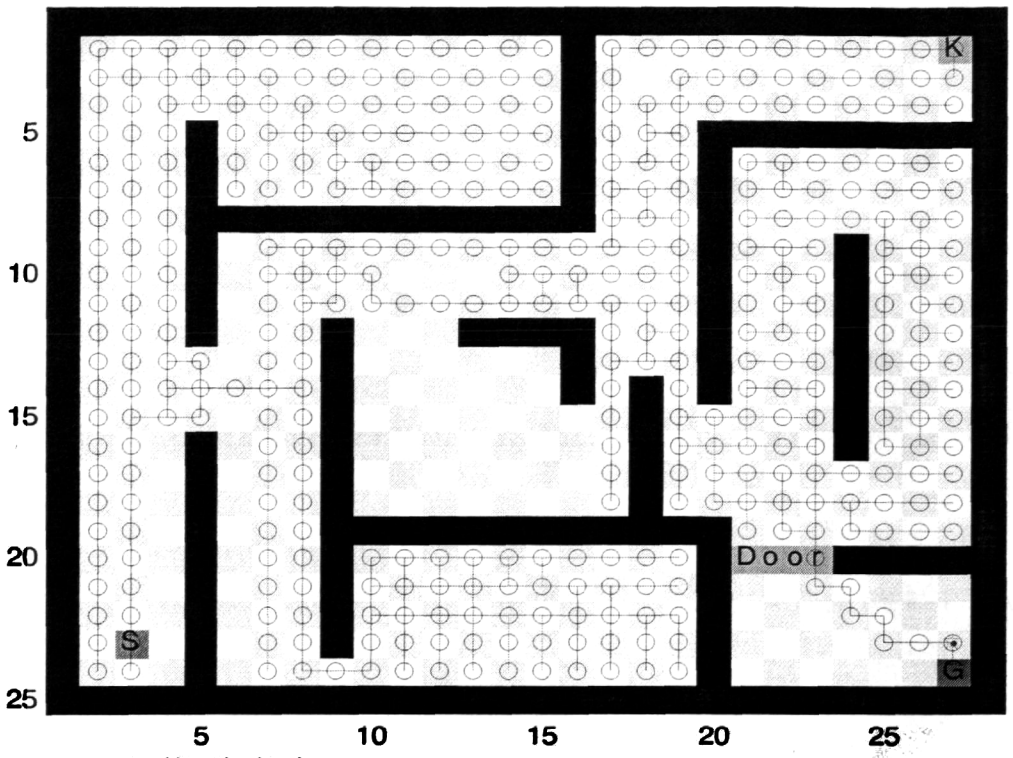
这里设定 $\rho_0=3$；实验中注水速度(如4.2部分和等式(4.4)所描述的)设定 $v=2.0$。

5.3.3 应用激励势场模型进行实验的结果

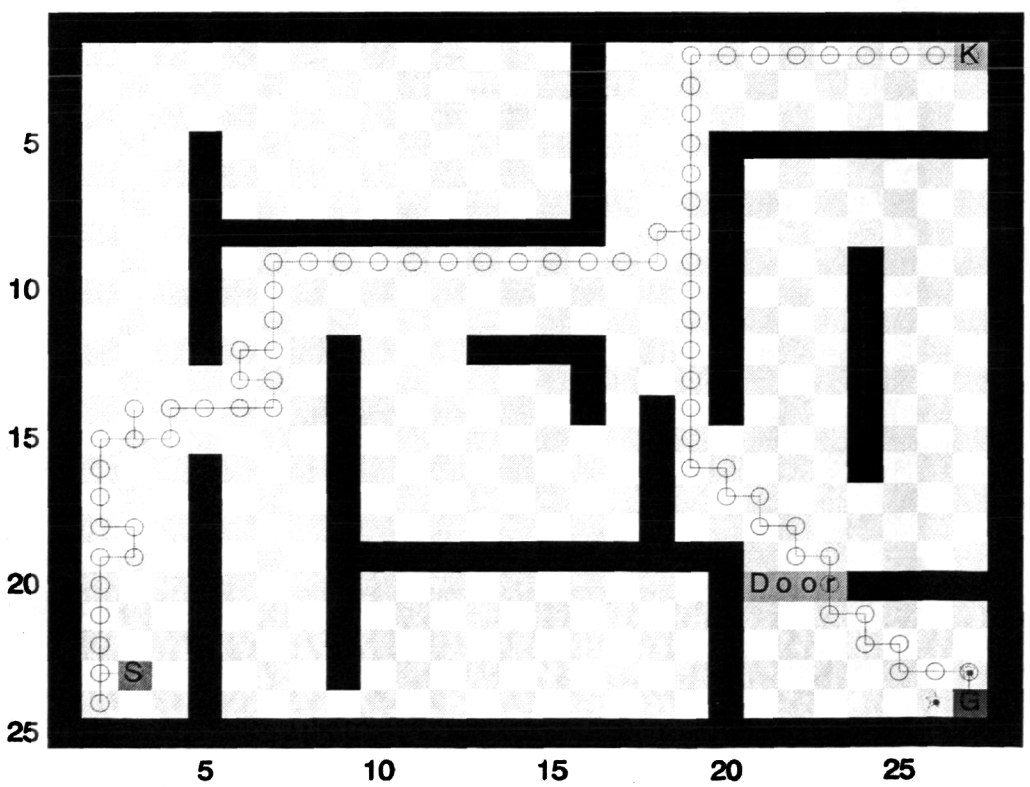
将激励势场模型用于大尺度激励学习问题的路径时，实验结果如图5.7（A—E）。第一次试验智能体用1005步找到目标，如图5.7(A)。在第34次试验试验中找到一条83步的优化路径，如图5.7(D)。从图5.7(E)可知最长的是第4次试验花了1400步。图7(E)表明几乎所有的仿真都能在45次试验内找到好的决策性策略。与Wiering的HQ学习系统(获取稳定的解需要20000次试验)相比，我们的新方法能更快地获得好而稳定的策略。



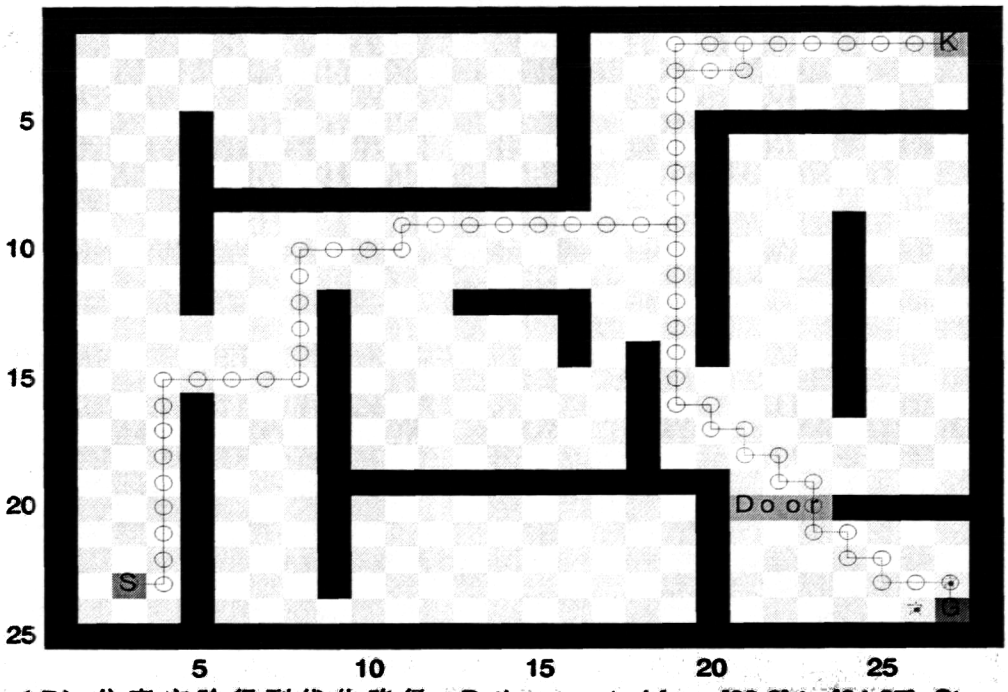
（A） 第1次尝试 Path generated from [23,3] to [24 27], Steps: 1005



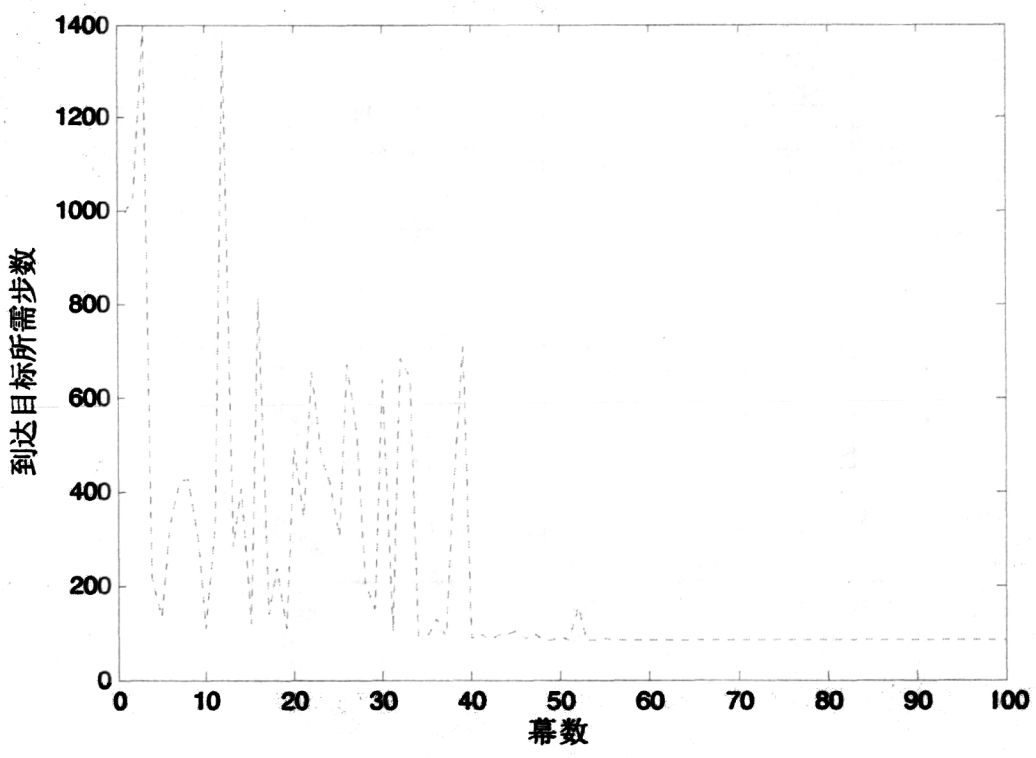
（B）第2次尝试 Path generated from [23,3] to [24 27], Steps: 901



(C)第36次尝试 Path generated from [23,3] to [24 27], Steps: 93



（D）仿真实验得到优化路径 Path generated from [23,3] to [24 27], Steps: 83



（E）激励势场模型应用于钥匙与门问题中的收敛情况

图5.7 (A—E)为激励势场模型应用到钥匙与门问题中的仿真结果

5.3.4 与其它多种学习方法进行比较的实验结果

用三种不同的方法对“钥匙与门”迷宫问题分别作100次仿真实验测试的最终结果如表5.1。从左到右，其中第2列为每次实验的平均步数；第3列为智能体到达目标的次数与其总实验次数的百分比；第4列为实验中智能体从起点到达目标的平均路径长；第5列是智能体在找到最优路径后的后继实验中到达目标次数与其实验次数的百分比。实验结果表明：在每一次仿真试验中，我们的方法都能找到最优解；用我们的方法每次试验的平均步数为83步，到达目标的百分比率为100%；平均路径长(Av.Sol)等于平均步数，获得最优解百分比率为100%。与HQ学习相比，实验结果表明我们的方法更快更稳定。

表5.1 用不同方法对“钥匙与门”迷宫问题分别进行100次仿真实验的结果

结果 / 方法		Av. Steps	(%) Goal	Av. Sol.	(%) Optimal
Random		9310	19	6370	0
HQ-学习	3 Agents	224	85	87	8
	4 Agents	126	96	90	9
	6 Agents	127	96	91	8
	8 Agents	101	99	92	6
Our Method		83	100	83	100

结论与展望

结论

本文在前人研究的基础上，首先在绪论部分介绍了课题研究的背景和激励学习研究的现状。第二章先要介绍了激励学习的一些基本知识。然后着重介绍了当前激励学习中研究比较成熟的两种方法，它们是瞬时差分法和Q学习方法。第三章介绍人工势场中斥力势函数，引力势函数的选取及人工势场的生成，同时分析了人工势方法的优缺点。第四章是激励势场模型的系统介绍，包括激励势的斥力源和引力源定义，激励势场生成的详细描述。即提出了激励势场模型。其思想是首先将激励学习模型转换成人工势场模型。将每一个状态看成一个引力源或斥力源，这样智能体在每一个状态都获得一个瞬时奖赏。同时应用虚拟水流法的思想解决激励势场模型中存在的局部最小问题。最重要的是在每次试验中智能体能记住局部最小位置的最新的势能，这样在下一次试验中智能体就能获得更优的解。最后在第五章对第四章中提出的激励势场模型在不同环境下进行模拟仿真。实验结果表明：对可观测和部分可观测激励学习新方法能简洁有效地给出理想的解。与Q学习和HQ学习等方法相比激励势场模型更稳定有效。

研究展望

激励势场模型吸取激励学习和人工势场方法两者的优点。模型简洁明了，使用方便，在进行运动规划时不仅考虑了智能体的避障和轨迹规划，而且也考虑了智能体的动态运动特性，使智能体运动更加自然和具有柔顺性，且具有自主学习能力。该模型的不足之处是智能体必须知道确切的状态或它的邻域是可观测的，这样其邻域的势能就才能准确被估测出来。在以后研究中需考虑的问题有：

(1) 该方法需在理论上证明算法的收敛性和进行计算复杂性的分析；

(2) 估测邻域状态或观测的准确性和有效性；

(3) 用处于大尺度的未知的动态环境下的机器人来测试该方法的性能。

参考文献

[1] L. P. Kaelbling, M. L. Littman and A. W. Moore. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 1996, vol.4, 237-285

[2] 张奇, 林崇德主编. 学习理论. 武汉: 湖北教育出版社, 1999.5

[3] Rumelhart D.E., G.E.Hinton and R.J.Williams. Learning Internal Representations By Error Propagation. in Parallel Distributed Processing. Vol.l, Cambridge, MA: MIT Press, 1986, 318-362

[4] 洪家荣. 归纳学习. 科学出版社，1999

[5] Minsky M.L. Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain-Model Problem. Ph.D. Thesis, Princeton University, 1954

[6] Samuel A.L. Some studies in machine learning using game of checkers. IBM Joual on Research and Development, 1959. 3, 211-229

[7] Thorndike, E.L. Educational Psychology. Brifer Course, 1914

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

[9] Sutton R. Learning to Predict by the method of temporal differences. Machine Learning, 1988, 3(I),9-44

[10] Howard R.A. Dynamic Programming and Markov Processes. M.LT. Press, Cambridge, 1960.

[11] 胡奇英. 马尔可夫决策过程引论. 西安: 西安电子科技大学出版社, 2000.7

[12] Bertsekas D.P. Dynamic Programming and Optimal Control. Vol.2, Athena Scientific, Belmont Mass, 1995.

[13] Watkins J.C.H. Dayan P. Q-Learning. Machine Learning, 1992.8,279-292

[14] Philips Semiconductors. SJA1000-Stand-alone CAN Controller-data Sheet, 1997

[15] Holland J.H. Adaptation in Natural and Artificial Systems. MIT Press, 1991

[16] Back T. Evolutionary Algorithms in Theory and Practice. Oxford University Press, New York, 1996

[17] Fogel L.J. Artificial Intelligence Through Simulated Evolution. New York: John Wiley, 1966

[18] Goldberg D.E. Genetic Algorithms in Search. Optimization and Machine Learning, Addison- Wesley Publishing, 1989

[19] Holland J.H. Genetic Algorithms and Classifier Systems, Foundations and Future

Directions. Genetic Algorithms and Their Applications: Proc. Of the Second International Conf. On GA.1987, 82-89

[20] Moriarty, D.E. Efficient Reinforcement Learning Through Symbiotic Evolution. Machine Learning,1996.22, 11-32

[21] Sutton, R.S., Barto, A.G., Williams, R. Reinforcement learning is direct adaptive optimal control. Proceedings of the American Control Conference, 1991, 2143-2146.

[22] Boyan J. Learning Evaluation Function for Global Optimization. PhD Thesis, Carnegie Mellon University, 1998

[23] Sutton R. Learning to Predict by the method of temporal differences. Machine Learning, 1988, 3(I),9-44

[24] Singh,S.P., R.Sutton. Reinforcement Learning with Replacing Eligiblity Traces. Machine Learning, 1996, Vol.22, 23-158

[25] Brartke. S.J, A.Barto. Linear Least-Squares Algorithms for Temporal Difference Learning. Machine Learning, 1996, Vo1.22, 33-57

[26] Boyan. J. Least-Squares Temporal Difference Learning. Machine Learning: Proceedings of the Sixteenth International Conference (ICML),1999

[27] Watkins C. Learning from delayed rewards, PhD. Thesis, King's College, University Of Cambridge, 1989

[28] Peng J. and R.J.Williams. Incremental multi-step Q-learning. Machine Learning, 1996,Vol.11, 283-290

[29] Peng J. and R.J.Williams. Efficient Dynamic Programming-Based Learning for Control. PhD Thesis, Northeastern University, 1993

[30] Rummery, G.A. & Niranjan,M. On-line Q-learning using connectionist systems. Technique Report, CUED/ F-INFENG/TR-166, Cambridge University Engineering Department.1994

[31] Singh,S.P. Convergence Results for Single-step On-policy Reinforcement learning Algorithms. Machine Learning, 2000, Vo1.38, 287-308

[32] Barto, A.G., R.S. Sutton, C.W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Transactions on System, Man, and Cybernetics, SMC-13,1983,834-846

[33]Lin C.T, C.S.G. Lee. Reinforcement Structure/Parameter Learning for Neural Network-Based Fuzzy Logic Control System. IEEE Trans. Fuzzy System, 1994, Vol.2, No. 1, 46-63

[34] Bertsekas D.P., J.N.Tsitsiklis. Neuro-Dynamic Programming. Athena

Scientific ,Belmont, Mass.,1996

[35] Hu G.H.,N.Q.Liu and C.P Wu. TD(λ) Algorithms for Average Reward Problems. Proc. Of the 14th IFAC World Congress,Vol.Q, Beijing,1999.7,375-380

[36] Schwartz A. A Reinforcement Learning Method for Maximizing Undiscounted Rewards. In Proceedings of the Tenth International Conference on Machine Learning, Morgan Kaufmann, San Mateo,CA,1993

[37] Tadepalli P. H-learning: A Reinforcement Learning Method to Optimize Undiscounted Average Reward. Technical Report 94-30-1, Oregon State University, Department of Computer Science, 1994

[38] Sutton, R. Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding. Advances in Neural Information Processing Systems, MIT Press,1996, 1038-1044

[39] Tesauro G.J. Temporal Difference Learning and TD-Gammon. Communications of ACM, 1995,Vol.38, 58-68

[40] Baird L.C. Residual Algorithms: Reinforcement Learning with Function Approximation. In Proc. of the12th Int. Conf. on Machine Learning, July, San Francisco,CA.,1995

[41] Sutton R. Policy Gradient Methods for Reinforcement Learning with Function Approximation. Proc. Of Neural Information Processing Systems. Cambridge: M.LT. Press, 1999

[42] Berenji H.R. and P.Khedkar. Learning and tuning fuzzy logic controllers through reinforcements. IEEE Trans. Neural Networks, 1992, Vol.3,no.5, 724-740

[43] Dayan P. and T.J. Sejnowski. TD(λ) Converges with Probability. Machine Learning, 1994, Vol.14. 295-301

[44] Dayan P. The Convergence of TD(λ) for General λ. Machine Learning,1992, Vol.8, 341-362

[45] Singh S., P.Dayan. Analytical Mean Squared Error Curves for Temporal Difference Learning. Machine Learning, 1998, 32. 5-40

[46] Tsitsiklis J. Asynchronous Stochastic Approximation and Q-Learning. Machine Learning,1994, Vol.l6,185-202

[47] Baird L.C., A. Moore. Gradient Descent for General Reinforcement Learning. Proc. Of NIPS-11. Cambridge: MLT Press,1999

[48] Heger M., The Loss From Imperfect Value Functions in Expectation-Based and Minimax -Based Tasks. Machine Learning, 1996, Vol.22, 197-225

[49] Zhang P. and Stephane C. Uncertainty Estimate with Pseudo-Entropy in

Reinforcement Learning. Control Theory and Apllications,1998, Vo1.15, No.1, 100-104

[50] 蒋国飞,吴沧浦. 基于Q学习算法和BP神经网络的倒立摆控制. 自动化学报, 1998, Vo1.24, No.5, 662-666

[51] Brooks R. A Robust Layered Control System for a Mobile Robot. IEEE Trans. Robotics and Automation.RA-2(1986), 14-23

[52] Tham C.-K. and R.W.Prager. A Modular Q-Learning Architecture for Manipulator Task Decomposition. In Proc. Of the 11th Int. Conf. On Machine Learning. San Franciso,1994

[53] Jacobs R.and M. Jordan. Learning Piecewise Control Strategies in a Modular Neural Network Architecture. IEEE Trans on Sys. Man and Cybernetics.1993, 23(2), 337-345

[54] Singh S. Transfer of Learning by Composing Solutions of Elemental Sequential Tasks. Machine Learning, 1992,8(3/4), 323-339

[55] Stone P. Layed Learning in Multiple Agents Systems, Ph.D. Thesis,CMU,1998

[56] Crites,R.H. & Barto A.G. Elevator Group Control Using Multiple Reinforcement Learning Agents. Machine Learning, Vo1.33,Issue 2/3, 1998, 235-262.

[57] Zhang W. and Thomas G. Dietterich. Value Function Approximation and Job-Shop Scheduling. Proceedings of the Workshop on value Function Approximation. Carnegie-Melton University, School of Computer Science, Report Number CMU-CS-95-206, 1995

[58] Haykin S. Adaptive Filter Theory, 3rd edition, Englewood Cliffs, NJ: Prentice-Hall,1996

[59] Thrun S. Learning to Play the Game of Chess. In Advances in Neural Information Processing Systems 7.Cambridge: MLT Press, 1995

[60] Crites R.H. Multi-Agent Reinforcement Learning. Ph.D. Thesis. Univ. Of Mass.,1999

[61) Xu Xin, He Han-gen. Multi-phase Intelligent Traffic Controller Based on Reinforcement Learning. Advances in System Science and Applications, Vo1.2 77-82

[62] Spong,M.W. The swing up control problem for the acrobot. IEEE Control Systems Magazine,1995,15(1), 49-55

[63] 汪荣鑫 随机过程 西安: 西安交通大学出版社（第二版），2006.8

[64] 高阳，陈世福，陆鑫. 强化学习研究综述[J].自动化学报，2004, 30(1),86-100

[65] Mitchell T.M. Machine Learning. 北京:机械工业出版社，2003

[66] O. Khatib. Real-time Obstacle Avoidance for Manipulators and Mobile Robots. International Journal of Robotics Research, 1986, Vol. 5, No. 1, 90-98

[67] Y .Koren and J. Borenstein. Potential field methods and their inherent limitations for mobile robot navigation. Proceedings of the IEEE Conference on Robotics and Automation, Sacramento, California, April, 1991, 1394-1404

[68] W. H. Huang, B. R. Fajen. Visual navigation and obstacle avoidance using a steering potential function. Robotics and Autonomous Systems,2006, Vol.54, 288-299

[69] M. G. Park and M. C. Lee. Artificial potential field based path planning for mobile robots using a virtual obstacle concept. Proceedings of IEEE/ASME International Conference on Advanced Intelligent Mechatronics, 2003,735-740

[70] C. Q. Liu, M. H. A. Jr, H. Krishnan and L. S. Yong. Virtual Obstacle Concept for Local-minimum-recovery in Potential-field Based Navigation. Proceedings of the IEEE International Conference on Robotics & Automation San Francisco, CA,2000, 983-988

[71] O. Brock and O. Khatib. High-speed navigation using the global dynamic window approach. Proceedings of the IEEE International Conference on Robotics and Automation, 1999, 341–346

[72] K. Konolige. A gradient method for realtime robot control. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2000, Vol.1, 639–646

[73] E. Rimon and D. Koditschek. Exact robot navigation using artificial potential functions. IEEE Transactions on Robotics and Automation, 1992, Vol.8, No.5, 501–518

[74] M. L. Littman. Memoryless policies: Theoretical limitations and practical results. Proceedings of the Third International Conference on Simulation of Adaptive Behavior (SAB'94), Cambridge : MIT Press, 1994, 238-245

[75] M. Wiering and J. Schmidhuber. HQ-Learning. Adaptive Behavior,1998, Vol.6, No.2, 219-246

致 谢

在本文完成之际，谨向我的导师陈焕文教授致以深深的谢意。本文的研究工作是在陈老师的启发鼓励和悉心指导下完成的。陈老师对学科前沿与研究方向的敏锐洞察和正确把握，使作者在课题研究中受益非浅。正是由于他的严格要求、精心教诲和指导，本文的研究工作才得以顺利完成。陈老师严谨的治学态度、高深的学术造诣和朴实的作风永远是作者学习的楷模。

感谢长沙理工大学的罗可教授、李峰教授、殷苌茗教授、蒋加伏教授、徐蔚鸿教授、陈书开教授、谢丽娟教授和张明媚老师，他们在百忙之余，始终对论文的研究给予了关心和指导。

感谢湖南信息职业技术学院机器人实验室为论文研究的顺利开展提供良好条件，另外湖南信息职业技术学院还给作者提供了做论文期间的基本生活费。

感谢各级领导对作者学习、工作的指导和关心。

感谢三年年来学习、生活在一起的各位同窗好友，特别是李立云、陈一栋、薛丽华、杨健雄、胡明辉、王曦和付超红等等。

感谢各位师兄弟和师姐妹对作者论文研究提供的帮助，他们是:唐中勇、付强、卓佳、易良、蔡琼、陈鹏慧、鲁新军、万杰、邓慈云、陈哲平和张煌辉等。

感谢多年来我的亲人给予我的无私关爱和无限支持，尤其是我的母亲对我的牵挂和期望，成为我努力拚搏、刻苦钻研的动力。

感谢我的父亲给予我的无私的爱和最坚强的支持。父亲永远是为我人生导航的不灭塔灯。

再次感谢所有关心和帮助过我的人！

刘泽文

2007-12-26

附录（在学习期间完成的学术论文和参加的科研项目）

发表的论文：

[1] 刘泽文，谢丽娟，陈焕文．比较治疗专家系统在临床决策中的应用．医学与哲学（临床决策版），2007.11, 55-58

[2] 鲁新军，陈焕文，谢丽娟，刘泽文．机器人导航中势场局部最小的水流解决法．微计算机信息，已录用

参加的科研项目：

[1] 国家自然科学基金资助项目：智能体在部分可观测 Markov 条件下的激励学习研究（60075019）