随着机器人技术发展的愈发成熟，机器人在我们生活中扮演着越来越重要的角色。在机器人相关技术领域中，路径规划是非常重要的一项技术。在理论中，我们可以将地图数据传送给机器人进行分析，从而得到最佳路径，但是在实际应用中，机器人是不可能知道一个未知环境中的地理信息的，因此，对于在陌生环境中的，不具有先验知识的机器人，研究其探索和规划路径功能是具有非常重要的意义的。

本文以机器人在未知环境中的路径规划为背景，学习强化学习算法实现机器人的路径规划。在原有的强化学习算法Q-learning基础之上，提出改进，提高性能，完成规避障碍，求得最优路径功能。

With the development of increasingly robotics, robots play an increasingly important role in our lives. In the related field of robotics, the path planning is a very important technology. In theory, we can transport map data to the robot for analysis to obtain the best path. However, in practice, it is impossible to know the map data in an unknown environment, geographic information. Therefore, the study which a robotic without prior knowledge explore and find the best way in an unfamiliar environment is a very important significance.

In this paper, robot path planning in unknown environment as the background, we learn reinforcement learning algorithm to achieve robot path planning. Based on the original reinforcement learning Q-learning algorithm, we propose improvements to enhance performance, complete obstacle avoidance, get the optimal path function.

Q-learning，移动机器人，路径规划，避障，强化学习，最优路径

移动机器人是一种在复杂环境下工作，具有自行组织、自主运行、自主规划的智能机器人，它集中了传感器技术、信息处理、电子工程、计算机工程、自动化控制工程以及人工智能等多学科的研究成果，代表机电一体化的最高成就，是目前科学技术发展最活跃的领域之一。随着机器人性能不断地完善，移动机器人的应用范围大为扩展，不仅在工业、农业、医疗、服务等行业中得到广泛的应用，而且在城市安全、国防和空间探测领域等有害与危险场合得到很好的应用。因此，移动机器人技术已经得到世界各国的普遍关注。

移动机器人根据移动方式来分，可分为：轮式移动机器人、步行移动机器人单腿式、双腿式和多腿式、履带式移动机器人、爬行机器人、蠕动式机器人和游动式机器人等类型；按工作环境来分，可分为：室内移动机器人和室外移动机器人；按控制体系结构来分，可分为：功能式水平式结构机器人、行为式垂直式结构机器人和混合式机器人；按功能和用途来分，可分为：医疗机器人、军用机器人、助残机器人、清洁机器人等。

移动机器人的研究始于60年代末期。斯坦福研究院的和Charles Rosen等人，在1966年至1972年中研发出了取名为Shakey的自主移动机器人，目的是研究应用人工智能技术，在复杂环境下机器人系统的自主推理、规划和控制。我国机器人研究开始于世纪年代，至今已有多年。年，北京机械工业自动化研究所曹祥康在江苏嘉兴召开了全国性机械手技术交流大会，是我国第一个以机器人为主题的大型会议。比较有代表性的有：清华大学研制了智能移动THMR-III V型机器人；中科院自动化研究所研制出智能保安机器人，全方位移动式机器人视觉导航系统，智能移动机器人通用平台AIMR;香港城市大学研发了自动导航车及服务机器人；哈尔滨工业大学研制了导游机器人；国防科技大学研发成功了双足机器人等；“海龙2号”是我国自主研制的水下机器人，是我国目前仅有能在3500米水深、海底高温和复杂地形的特殊环境下幵展海洋调查和作业的最高精技术装备，它除了在潜水深度上的优势之外，还在国际上首次采用了一些自主研发的先进技术，包括虚拟控制系统和动力定位系统。

移动机器人导航移动机器人学的关键技术，是指按照预先给出的任务，根据已知的地图信息作出全局路径规划，并在行进过程中，不断感知周围的局部环境信息，自主地作出各种决策，随时调整位姿，引导自身安全行驶或跟踪已知路径，达到目标位姿，是移动机器人各项研究应用的基础和前提。自主导航是移动机器人导航的一项重要的基本能力，它主要解决的问题可以归纳为三个方面：“1.现在何处？2.要往哪去？3.如何去？”也就是移动机器人的地图构建、定位和路径规划题。地图构建是移动机器人在自主主导航的过程中，通过传感器感知环境并建立环境模型，这是述立环境模型最常用且有效的方法；定位移动机器人在工作环境中探索所处位置的过程，也就是机器人的全局坐标和姿态；路径规划技术就是移动机器人在具有障碍物的环境上按照一定的评价标准（如作代价最小、行走路径最短、行走时间最短等）、寻找一条从给起点到达目标终点的无碰路径。

但是不管采取何种导航方式，移动机器人上主要是完成路校规划、定位和避障等任务。路径规划问题是移动机器人自助式导航技术研究中的关键技术之，它是指移动机器人按照某一性能指标，搜索一条从起始位置到目标位置的最优或次优的无碰路径

随着计算机、传感器及控制技术的发展，特别是各种新算法的不断涌现，移动机器人路径规划技术已经取得了丰硕的研究成果，特别是在周围环境未知或部分未知的局部路径规划（人工势场法、模糊逻辑算法、神经网络法、进化算法、强化学习等），国内外学者已经作了大量的研究和改进，下面综述中将对与本文有关的一些方法进行讨论。

1986年，Khatib首次提出用人工势场法解决机器人避障问题，在此之后许多学

者不仅发展了这一方法并将其应用于机器人的实时路径规划和足球机器人的运动

控制。人工势场法可以实现快速的控制，所以人工势场法丨被广泛的应用于实

时的运动控制。

E.Shi等进一步对传统的人工势能场方法进行改进，他提出的方法虽然能够使移动机器人避幵障碍物，但其实施性并不强，虽然克服了零势能场（在这种情况下机器人无法到达目标位置）的情况，但又出现了很多问题，比如障碍物的影响范围、引力增益和斥力增益会影响到机器人运动过程中振动的程度、运动路径的平滑度、路径的长短及其运动过程中是否会碰撞到障碍物等，所以必须同时选择这三个系数的最优值，而在实验过程中是很难选取的，必须通过反复的实验才能选取较为合适的值；王奇志等提出了一种改进的人工势场法，通过排除一个距离机器人最远的障碍物，同时加一个同等大小反方向力来消除零势能点，从而实现消除零势能域，达到多障碍物情况下机器人运动规划的快速、实时、避障的效果，结果表明，该算法对多个障碍物和非静态的障碍物同样适用。陈华志等采用一种具有速度负反馈的基于模糊控制的移动机器人避障算法进行研究和仿真对移动机器人在实际环境中进行了实验，并证明了方法的可行性；Lee等使用模糊控制来决定移动机器人可能的十三个移动方向，使机器人在这种导航下避免机器人内部碰撞；Pradhan等在完全未知的环境中使用模糊逻辑对

多达1000个机器人进行了导航实验，在试验中，对四个输入变和两个输出变贤进行模糊化，最后证明了高斯隶属函数在移动机器人导航中最有效；Yang等设计了移动机器人的动态模型，根据多机器人的局部信息，建立了引力、斥力函数的模糊控制器，基于最优控制理论，来衡量系统的状态性能指标；刘利等针对海域不同的障碍环境，从运动学特性出发，提出了一种多主水航行器（AUV）编队避障策略。该策略是在领航跟随法的基础上，利用编队队形的几何关系，引入虚拟AUV，再采用模糊控制避障策略，使编队能顺利通过障碍物，到达目的地；Rigatos等研究了在包含固定障碍物和移动物体的部分未知的环境下，根据移动机器人的运动问题和控制规律，结合滑膜与模糊逻辑控制的基本原则，提出使机器人达到期望的运动效的控制器，即降低滑动模式模糊逻辑控制器的复杂度，具有稳定性和简易性；Montaner等为移动机器人导航设计了个模糊逻辑控制器，该控制器可以接收确定性和模糊的信息，并在一个由七个超声波传感器来获信息的环境中对移动机器人进行了实验研究了移动机器人在未知群集环境屮反应式导航的问题，中文把反应式导航定义为感知数据和命令的关系，建立一个反应式导航相当于对机器人提供了一个运动的地图，在此基础上提出了另一种“普遍”的逼近方法—模糊逻辑，并揭示了如何利用行为分解的方法来选择模糊规则；Tan等提出了一种可加速减速的模糊控制器模型，使移动机器人在动态环境中自主导航，包括寻找目标、躲避障碍和寻找最优路径。

Yang和Meng提出了一种离散拓扑结构的祌经网络模型，移动机器人使这种并联的祌经网络模型在离散化的完全已知环境中进行路径规划，并假定机器人的速度恒定，这在真实的环境中对移动机器人进行控制是没意义的；Zhu等提出了种基于模糊神经网络将传感器信息与机器人的移动结合起来的导航方法，这种力法可以使移动机器人充分感受周围环境、自主避开静态和动态障物，并在不同情况下避开“死循环”产生到达目标的合理的轨迹，通过仿真证明了该方法的实用性和有效性；Nishida等基于神经元系统提出了种感知预测机器人伙伴的控制方法，该预测方法对于减少计算量、提取感知信息是非常重要的；Kala等在文中提出了遗传算法在移动机器人航巾的应川，使机器人可用于业界派调查，数据采集，移动机器人可以在动态环境中避免内部碰撞证明了这种算法的有效性。

其主要方法有：可视图法，自由空间法，最优控制法，栅格法，拓扑法，神经网络法等。

可视图法

可视图法视移动机器人为一点，将机器人、目标点和多边形障碍物的各顶点进行组合连接，并保证这些直线均不与障碍物相交，这就形成了一张图，称为可视图。由于任意两直线的顶点都是可见的，从起点沿着这些直线到达目标点的所有路径均是运动物体的无碰路径。搜索最优路径的问题就转化为从起点到目标点经过这些可视直线的最短距离问题。运用优化算法，可删除一些不必要的连线以简化可视图，缩短搜索时间。该法能够求得最短路径，但假设忽略移动机器人的尺寸大小，使得机器人通过障碍物顶点时离障碍物太近甚至接触，并且搜索时间长。切线图法和Voronoi图法对可视图法进行了改造。切线图用障碍物的切线表示弧，因此是从起始点到目标点的最短路径的图，即移动机器人必须几乎接近障碍物行走。其缺点是如果控制过程中产生位置误差，移动机器人碰撞的可能性会很高。Voronoi图法用尽可能远离障碍物和墙壁的路径表示弧。由此，从起始节点到目标节点的路径将会增长，但采用这种控制方式时，即使产生位置误差，移动机器人也不会碰到障碍物。

拓扑法

将规划空间分割成具有拓扑特征子空间，根据彼此连通性建立拓扑网络，在网络上寻找起始点到目标点的拓扑路径，最终由拓扑路径求出几何路径。拓扑法基本思想是降维法，即将在高维几何空间中求路径的问题转化为低维拓扑空间中判别连通性的问题。优点在于利用拓扑特征大大缩小了搜索空间。算法复杂性仅依赖于障碍物数目，理论上是完备的。而且拓扑法通常不需要机器人的准确位置，对于位置误差也就有了更好的鲁棒性；缺点是建立拓扑网络的过程相当复杂，特别在增加障碍物时如何有效地修正已经存在的拓扑网是有待解决的问题。

栅格法

将移动机器人工作环境分解成一系列具有二值信息的网格单元，多采用四叉树或八叉树表示，并通过优化算法完成路径搜索。该法以栅格为单位记录环境信息，有障碍物的地方累积值比较高，移动机器人就会采用优化算法避开。环境被量化成具有一定分辨率的栅格，栅格大小直接影响环境信息存储量大小和规划时间长短。栅格划分大了，环境信息存储量小，规划时间短，但分辨率下降，在密集环境下发现路径的能力减弱；栅格划分小了，环境分辨率高，在密集环境下发现路径的能力强，但环境信息存储量大，规划时间长。

栅格法是由w．E．Howden在1968年提出的。栅格法将机器人工作环境分解成一系列具有二值信息的网格单元，工作空间中障碍物的位置和大小一致，并且在机器人运动过程中，障碍物的位置和大小不发生变化。用尺寸相同的栅格对机器人的二维工作空间进行划分，栅格的大小以机器人自身的尺寸为准。若某个栅格范围内不含任何障碍物，则称此栅格为自由栅格；反之，称为障碍栅格。自由空间和障碍物均可表示为栅格块的集成。栅格的标识方法有两种：直角坐标法和序号法。多采用四叉树或八叉树表示工作环境，并通过优化算法完成路径搜索。该方法以栅格为单位记录环境信息，栅格粒度越小，障碍物的表示越精确，但同时会占用大量的存储空问，算法的搜索范围将按指数增加。栅格的粒度太大，规划的路径会很不精确。所以栅格粒度的大小的确定，是栅格法的主要问题。

自由空间

自由空间应用于移动机器人路径规划，采用预先定义的如广义锥形和凸多边形等基本形状构造自由空间，并将自由空间表示为连通图，通过搜索连通图来进行路径规划。自由空间的构造方法是：从障碍物的一个顶点开始，依次作其它顶点的链接线，删除不必要的链接线，使得链接线与障碍物边界所围成的每一个自由空间都是面积最大的凸多边形：连接各链接线的中点形成的网络图即为机器人可自由运动的路线。其优点是比较灵活，起始点和目标点的改变不会造成连通图的重构，缺点是复杂程度与障碍物的多少成正比，且有时无法获得最短路径用栅格法建模受到了空间分辨率和内存容量的矛盾限制。而自由空间法建模，解决了这一矛盾。但自由空间法的分割需构造想象边界，想象边界本身具有任意性，于是导致路径的不确定性。

最优控制法

在确定的空间里，二维平面上的一条边界可由方程来描述。那么，机器人在运动过程中，从起点到终点的众多路径里，有障碍物的路径是不允许机器人通过的。这些路径可以作为约束条件，由数学表达式表示。非完整移动机器人通过适当的变换，可将其转化为链式形式。因此，通过选择适当的控制量就可以驱使机器人从一个位置运动到另一个位置。

神经网络法

可视图法缺乏灵活性，且不适用于圆形障碍物的路径规划问题。神经网络法用于全局路径规划可以解决以上问题。

人工势场法

人工势场法是Khatib提出的一种虚拟力法。其基本思想是将移动机器人在环境中的运动视为一种虚拟人工受力场中的运动。障碍物对移动机器人产生斥力，目标点产生引力，引力和斥力周围由一定的算法产生相应的势，机器人在势场中受到抽象力作用，抽象力使得机器人绕过障碍物。该法结构简单，便于低层的实时控制，在实时避障和平滑的轨迹控制方面，得到了广泛应用，其不足在于存在局部最优解，容易产生死锁现象，因而可能使移动机器人在到达目标点之前就停留在局部最优点。

主要的缺陷：

1)陷阱区域

2)在相近的障碍物之间不能发现路径。

3)在障碍物前振荡。

4)在狭窄通道中摆动。

针对这些缺陷，提出了一些改进办法。针对人工势场法存在“机器人在到达目标位置前由于陷入局部极小点而无法到达目标位置”的问题，解决的方法有：重新定义势函数，使之没有或有更少的局部极小点；利用搜索算法跳出局部极小点。还可以利用模拟退火算法使势函数跳出局部极小点，到达机器人的目标位置。

模糊逻辑算法

模糊逻辑算法基于对驾驶员的工作过程观察研究得出。驾驶员避碰动作并非对环境信息精确计算完成的，而是根据模糊的环境信息，通过查表得到规划出的信息，完成局部路径规划。优点是克服了势场法易产生的局部极小问题，对处理未知环境下的规划问题显示出很大优越性，对于解决用通常的定量方法来说是很复杂的问题或当外界只能提供定性近似的、不确定信息数据时非常有效。假设检测的是障碍物与机器人的距离和障碍物的运动信息，输出机器人速度变化和转角变化。

模糊控制算法有诸多优点，但也有固有缺陷：人的经验不一定完备；输入量增多时，推理规则或模糊表会急剧膨胀。由于模糊隶属度函数的设计、模糊控制规则的制定主要靠人的经验和试凑，总结模糊控制规则时比较困难，而且，控制规则一旦确定，在线调整困难，无法很好地适应情况的变化。因此，如何得到最优的隶属度函数、控制规则以及对控制规则进行在线调整是该方法最大的问题。

模拟退火算法

模拟退火算法由Kirkpatrick S于1983年提出，源于物理退火过程。基本思想是利用随机优化问题求解过程与统计力学中热平衡问题的相似性，通过设定初温、初态和降温率控制温度的不断下降，结合概率突跳特性，利用解空间的邻域结构进行随机搜索。模拟退火算法用于路径规划可避免局部极值，但其理论收敛条件过于苛刻，在实际应用中往往无法满足。在有限计算量条件下的收敛性能依赖于自身参数，这使得参数设定成为算法应用过程中的一个关键环节。

智能仿生算法

神经网络法

路径规划是感知空间到行为空间的一种映射。映射关系可用不同方法实现，很难用精确数学方程表示，但采用神经网络易于表示。将传感器数据作为网络输入，由人给定相应场合下期望运动方向角增量作为网络输出，由多个选定位姿下的一组数据构成原始样本集，经过剔除重复或冲突样本等加工处理，得到最终样本集。将神经网络和模糊数学结合也可实现移动机器人局部路径规划。先对机器人传感器信息进行模糊处理，总结人的经验形成模糊规则。再把模糊规则作用于样本，对神经网络进行训练。通过学习典型样本，把规则融会贯通，整体体现出一定智能。实际中允许输入值偏离学习样本，只要输入接近一个学习样本的输入模式，则输出也就接近该样本输出模式。该性质使得神经网络可以模仿人脑在丢失部分信息时仍具有对事物正确的辨识力。

遗传算法

遗传算法以自然遗传机制和自然选择等生物进化理论为基础，构造了一类随机化搜索算法。利用选择、交叉和变异编制控制机构的计算程序，在某种程度上对生物进化过程作数学方式的模拟。只要求适应度函数为正，不要求可导或连续，同时作为并行算法，其隐并行性适用于全局搜索。多数优化算法都是单点搜索，易于陷入局部最优，而遗传算法却是一种多点搜索算法，故更有可能搜索到全局最优解。遗传算法的整体搜索策略和优化计算不依赖于梯度信息，解决了一些其它优化算法无法解决的问题。遗传算法运算速度不快，进化众多的规划要占据较大存储空间和运算时间，优点是克服了势场法的局部极小值问题，计算量不大，易做到边规划边跟踪，适用于时变未知环境的路径规划，实时性较好。遗传算法运用于移动机器人路径规划的研究近来取得了许多成果，其基本思想：将路径个体表达为路径中一系列中途点，并转换为二进制串。首先初始化路径群体，然后进行遗传操作，如选择、交叉、复制、变异。经过若干代进化以后，停止进化，输出当前最优个体，其过程如以下算法所示。

开始

随机初始化群体P(0)

计算群体P(0)中个体的适应度

t： =0

while(不满足终止准则)do

{

由P(t)通过遗传操作形成新的种群P(t+1)；

计算P(t+1)中个体的适应度，t：=t+1；

}

蚁群算法

蚁群优化算法是由意大利学者Dorigo等人在2O世纪9O年代从蚁群觅食行为受到启发，通过模拟自然界蚂蚁寻径的行为，提出的一种全新的模拟进化算法。蚁群优化算法在并行运行环境(如网格环境)下可以同步寻优，加快了寻优速度。另外，它是一种通用性强的算法，稍加修改便可用于其他优化问题。但计算量较大，搜索时间较长，易于陷入局部最优解。

粒子群算法

粒子群算法是由Kennedy博士和Eberhart博士于1995年从鸟类的捕食行为中受到启发提出的一种基于群体的智能随机优化算法。粒子群算法具有收敛速度快、算法简单、容易编程实现和鲁棒性强等特点，但是，粒子群算法也有一些缺陷，一是容易陷入局部极值点，导致得不到全局最优解；二是粒子群算法本身的参数设置，若参数选择不当，会导致寻优过程中粒子的多样性迅速消失，造成算法“早熟收敛”。

启发式搜索方法

启发式方法的最初代表是A\*算法 ，其新发展是D\*和Focussed D\*。这2种由Stentz A提出的增量式图搜索算法的产生。D\*算法可以理解为动态的最短路径算法，而Focussed D\*算法则利用A\*算法的主要优点即使用启发式估价函数，2种方法都能根据机器人在移动中探测到的环境信息快速修正和规划出最优路径，减少了局部规划的时间，对于在线的实时路径规划有很好的效果。此外，还出现了一些基于A\*的改进算法，它们一般都是通过修改A\*算法中的估价函数和图搜索方向来实现的，可以较大地提高路径规划的速度，具有一定的复杂环境自适应能力。

第1步训练

第2步训练

第3步训练

….

第n步训练

近年来，由于单步的Q-learning算法结构简单且易于扩展，已成为强化学习中的一个研究热点，国内外学者结合Q-learning算法，对强化学习算法展开了深入的研究。通过对Q-learning算法数据流动特性的分析，可以看出要提高Q-learning算法的收敛速度，其一可以通过打破数据传递的滞后特性，让后续状态的动作决策可以快速的影响到前面状态的动作选择，其二可以通过加快目标状态附近状态对应的状态动作对的收敛，为前面状态动作对的收敛提供必要条件，其三可以通过估计，利用相似的先验知识，估计状态-动作对。

Peng等通过对Q学习算法进行改进,提出了多步的算法，算法是借鉴算法而产生的，通过回溯的思想，使得后续产生的数据能够及时反馈回来。其实际意义为：我们通过数据的不断传递，使得某一状态的动作决策受到其后续状态的影响。如果未来某一决策*a*，是一个失败的决策，那么当前的决策也要承担相应的责任，也会把这种影响追加到当前决策上来；如果未来某一决策*a*，是一个正确的决策，那么当前的决策也获得相应的奖励，同样也要把这种影响追加到当前决策上来。

算法能够提高算法的收敛速度，满足在线学习的实用性，关键在于引入了跟踪迹，跟踪迹最初是由Klopf从认知科学的知度出发提出的一种记忆机制，Sutton和Singh把跟踪迹用于强化学习，使其成为一种重要的强化学习基本机制。实际上跟踪迹的思想比较简单，每当一个状态被访问时，该状态就会被短暂的记忆下来，随着时问的推移，状态的跟踪将会衰减，这个迹标志这个状态对学习足有资格的，该过程称为渐进遗忘。跟踪迹不为零时，若一个状态发生了，那么该状态就会随机赋予一定的信度，所以跟踪迹可以用于延时强化学习的信度分配问题。

通常，渐进遗忘的过程里现指数衰减状态。般令遗忘哀减系数和

折扣系数：的乘积指数衰减，山此来展现跟踪状态的渐进遗忘过积。利

用有效的跟踪以解决强化学习中的效用分配问题。

跟踪迹般可以分为累积迹和科换迹两种，传统计弹跟踪迹的方法累枳法，

当个状态被访时，其迹就巳经建立起来了，累积迹的计；法丨以通过公式

得到：

第1步训练

第2步训练

第3步训练

第n步训练



都是基于回溯的思想。该算法的设计初衷是为了提高强化学习算法的收敛速度，根据不同状态具有不同收敛优先级的特点，让距离目标状态近的状态区域提前收敛，为后续状态的收敛提供必要条件，并且通过回溯的思想改善Q-learning学习中数据传递的滞后性，使当前状态的动作决策能够快速的受到后续动作决策的影响。云南大学的金剑博士于2010年提出了基于链串的Q学习，其方法与此类似。

把一次路径规划尝试（机器人从初始状态到目标状态的一次循环叫做一次尝试）看作是一系列数据的传递，我们用一条链来形象比喻这一系列的数据传递。该循环中每一步搜索的状态可以是相同状态，也可以是不同状态，每一步状态中依次排列，前一少状态指向下步状态，形成一个从起始状态始逐级相连，最后指向当前状态的单条状态链。

本文通过公式递归回溯，使得后续产生的数据能够及时反馈回来。其实际意义为：通过数据的不断传递，使得某一状态的动作决策受到其后续状态的影响。如果未来某一决策a，是一个失败的决策，那么当前的决策也要承担相应的责任，要把这种影响追加到当前决策上来；如果未来某一决策a，是一个正确的决策，那么当前的决策也获得相应的奖励，也要把这种影响追加到当前决策上来。

改进的Q-learning算法就是利用单链和回溯的思想，把一次路径规划尝试（机器人从初始状态到达目标状态的一次循环叫做一次尝试）看作是一系列数据的传递，我们用一条链来形象比喻这一系列的数据传递，该循环中经历过的每个状态中都封装在一起，形成一个状态链，每一个状态形成链中的环。训练循环中的每一步，由当前状态跳到下一个状态，通过公式(3-3)更新当前状态对应的值的同时，通过重复循环，使链中环与环之问的数据直接地简单迭代计算，一环接一环，由现在向过去，迭代更新前面经历过状态的Q值。

这种数据传递方式一方面使得靠近目标状态的区域更快收敛，为其他区域状态的收敛提供必要条件；另一方面使后续产生的评估他能够及时反馈上来，从而使某一状态的动作决策受其后续状态的影响，及早地识别错误动作决策，避免无意义的搜索，通过回溯的思想改善Q-learning学习中数据传递的滞后性，使收敛速度加快。

第1步训练

第2步训练

第3步训练

第n步训练 …

MATLAB（矩阵实验室）是MATrix LABoratory的缩写，是一款由美国The MathWorks公司出品的商业数学软件。MATLAB是一种用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境。除了矩阵运算、绘制函数/数据图像等常用功能外，MATLAB还可以用来创建用户界面及与调用其它语言（包括C,C++,Java,Python和FORTRAN）编写的程序。

我们通过建立Q值表，并让机器人遍乂遍的从起点出发到达目标点，不停的完成训练循环，并不停的修改Q值表，当智能体找到了一条最优路径后结束机器人探索。每次的训练循环中，机器人从起始点出发，不停的训练，当智能体在动作集合中选择某动作后，环境接收该动作并发出状态转移，同吋给出奖赏。如果其动作获得环境正的奖赏(+r)，则移动机器人此后产该动作的趋势就会增强；反之，如果获得环境负的奖赏(-r)，则会减弱。最终通过足够多次的训练次数，Q值表会出现相对收敛的情况，机器人将会在Q值表中找到一条最优的路径。