一个基于增强学习算法的路由模型

张志坚　刘惟一
（云南大学信息学院　昆明 650091）

摘　要　由于 Internet 的不断发展，现有的路由算法为适应不同的网络要求，从一开始的 RIP、OSPF、BGP 等几种，衍生出很多新的适用于特殊网络的路由协议。本文提出一种基于增强学习算法的路由模型。将每个路由节点看作一个 Agent，利用增强学习算法的思想使得每个节点在不了解网络拓扑结构的情况下从向邻居转发的概率获得网络的信息，这样路由节点可以选择一个较优的转发方向。同时，节点能对网络的拥塞等情况作出调整。该模型为一些具体网络的路由协议，特别是 QoS 类路由算法提出了一个新的路由思想。

关键词　增强学习，路由模型，策略搜索，QoS 路由

A Routing Model Based on Reinforcement Learning Algorithm

ZHANG Zhi-Jian　LIU Wei-Yi
(College of Information, Yunnan University, Kunming 650091)

Abstract　With the development of internet, the routing algorithms have been improved for some special demands. Based on some former algorithms such as RIP, OSPF, BGP, some new algorithms have been given. This paper approaches a routing model which is based on reinforcement learning algorithms. Every routing note is treated as an Agent. Using the idea of reinforcement learning algorithms, the notes can gain some information of the net from the probability of forwarding, though it does not know the topology of the net. So the note can choose a better forwarding path. Meanwhile it can make some changes, when the block of net is occurred. This model also gives a new idea for routing algorithms which used in some special field such as QoS Routing.

Keywords　Reinforcement learning, Routing model, Policy search, QoS routing

1　引言

网络已经成为当今发展最快的技术之一。也对网络提出了许多新的具体要求，其中如何保证服务质量(QoS)[1~9]已经成为下一代互联网络的核心问题之一。而多约束的服务质量路由(QoSR)则是其中的核心和热点问题。QoSR 解决如何寻找一条满足多个约束条件的可行路径的问题。然而 QoSR 是一个 NP 完全问题，所以研究人员设计了许多启发式和非启发式算法[2]，这些算法针对每个 QoS 请求使用不同的算法，试图找到最小化线性能量函数的路径。然而，它们普遍存在计算复杂度过高、没有普遍使用性等问题，使得这部分算法很少能在实际中应用。

本文提出一个基于增强学习算法的路由模型。该算法将每个路由节点看作一个 Agent。它并不需要存储网络的拓扑结构，只需要建立和维护一张转发表，根据转发表的概率来选择转发对象。并且利用一个耗费函数 C_j（在 QoSR 中可以理解为满足约束条件的程度）来评估路径的优劣，从而更新转发表的概率分布，使路由节点在不消耗大量资源的同时得到较优路径，并且对网络状况的变化作出及时的反应。

2　相关工作

近年来，随着越来越多的网络应用，很多新的路由算法受到人们的关注。而 QoS 类路由算法已经成为网络技术研究和发展的焦点。

研究人员已经提出了很多 QoS 类路由算法。单播 QoSR 算法中出现了很多将问题转换到多项式可解的算法。Wang 和 Crowcroft 利用多项式非启发类思想使用 Dijkstra 最短路径树算法实现了带宽延迟受限的源路由求解[10]。这些算法在每个节点需要保存网络的全局状态，消耗了大量的路由器资源，同时对网络情况的变化反应不够及时。有的算法是从源节点开始，逐个询问其他节点，逐步逼近并到达目的节点，需要耗费网络资源来保证寻找到较优路径。Shin 和 Chou 提出了一个延迟受限分布式算法[11]，每个节点不保存网络全局状态，通过试验的方法探测邻居节点的延迟，从而选择一个邻居转发。但为了保证没有回路，需要记录大量的探测信息。

Cui 提出了一个性能可调的启发式多约束路由算法 BFS_MCP[4]，用一个能量函数来对每个节点编号。基于一个广度优先搜索的 Dijkstra 算法，设计了一个启发式的路由算法来寻找可行路径。该算法对网络规模和约束个数有良好的可扩展性。由于使用了能量函数对节点进行编号，算法的复杂度由于函数的非线性将大大增加。

在时延受限多播路由算法中，BSMA(bounded shortest multicast algorithm)被公认是最好的，但由于很多路由算法的复杂度太大而没有得到广泛的使用。很多新的路由算法借鉴了其他学科的思想，例如参数可调的克隆多播路由算法[2]，利用了生物技术中的克隆思想，首先对目标节点进行编码，建

张志坚　硕士研究生，主要研究博弈论、增强型学习算法和服务质量路由算法；刘惟一　教授、博士生导师，主要研究领域为数据挖掘和知识发现、网络环境下的信息检索额 Bayesian 网。

立亲和度函数，通过克隆、变异和克隆选择得到路径，从而避免了利用遗传算法(GA)所带来的“早熟”现象。然而在变异和克隆时，算法仍然有很高的复杂度，不利于实际应用。

3 背景知识

3.1 增强学习算法模型[5,6]

增强学习的历史要追溯到计算机科学和控制论的早期。增强学习就是利用奖励和惩罚机制为编程代理提供一个执行某任务的方法，而不需要指定任务如何完成。标准的增强学习模型包含两个主要内容：一个学习代理和一个动态环境。代理和环境通过感知和行为信号相连，如图1。

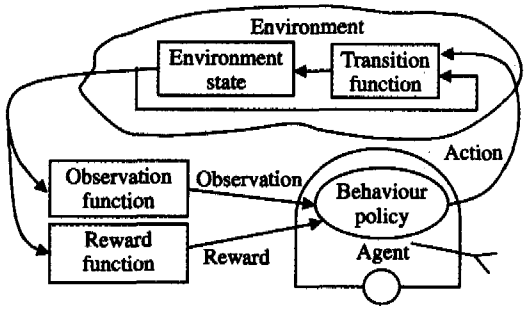

图1 增强学习模型

以代理的观点，存在两种环境：完全可观察环境和部分可观察环境。在完全可观察环境中，代理在每一步都可信地观察到环境所处的确切状态。而在部分可观察环境中，代理依据它的观察无法区别环境的某些状态。正式描述部分可观察环境的模型称为部分可观察 MDP(PoMDP)。

按照观察环境的不同和是否基于模型，现已有一些较好的增强学习算法，如 Q-Learning，Policy search 等。

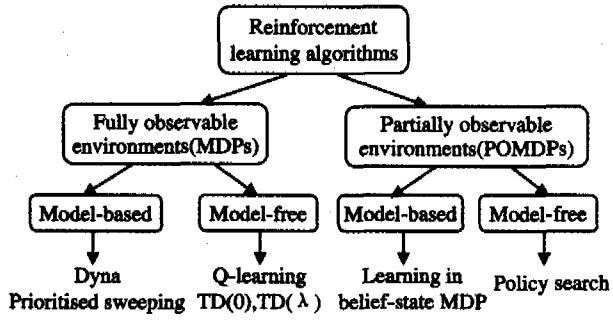

图2 增强学习算法分类

对于网络来说，网络的情况是变化的。作为单个路由要及时地掌握整个网络的情况也会消耗大量资源，所以将网络描述为一个不基于模型的部分可观察模型(POMDPs)是恰当的。每个路由根据历史的数据利用一个 Policy search 算法思想寻找一条较优的路径。

3.2 不基于模型的 PoMDPs 下的策略搜索爬山算法的思想

代理的环境可以用 MDP(Markov Decision Process)建模。

定义1 一个 MDP 是一个四元组$\langle S,A,R,T\rangle$

- S 是环境的状态集；
- A 是代理在每个状态下可能的行为集；
- $R:S\times A\to R$ 是收益函数；
- $T:S\times A\times S\to[0,1]$是状态转换函数、当前状态 s 和行为 a 在不同的下一状态的概率分布。

在部分可观察模型中，处理部分可观察模型的最简单方法就是忽略它。这种方法中，观察被看作是环境的状态，代理用在完全可观察 MDP 中同样的方法学习最优策略。为了按照观察来区分环境状态，代理需要利用以前的观察结果。于是 PoMDP 中最佳策略可能依赖于长期无限的观察序列。因此，求这样的策略的问题，总的说来是很困难的。结果，许多方法都集中于较优策略的学习上。

一个策略搜索的爬山算法的思想是这样的：对每个状态 s 有状态行为对的概率 $\Lambda(s,a)\in[0,1]$，并且有 $\sum_a\Lambda(s,a)=1$。算法每一步在当前状态 s 按概率选取行为 a，得到立即收益 r，并且观察下一状态 s'。按照收益评估函数 $Q(s,a)$，并根据评估函数更新状态行为对的概率。这样，较优的策略概率会高于其他的策略，从而实现寻找较优策略的目标。

4 基于增强学习算法的路由模型

4.1 模型表示

用有向图 $G\langle V,E\rangle$来表示一个网络[4]，其中 V 为路由节点集(路由器集合)，$V=\langle 1,2,3,\cdots,n\rangle$。$E$ 为弧集，元素 $e_{ij}\in E$ 称为图 G 中 v_i 到 v_j 的一条边(弧)，代表网络中的一条链路。在 QoSR 中，给每个链路 e 关联一组相互无关的权值$(w_1(e),w_2(e),\cdots,w_k(e))$，称为链路 e 的 QoS 度量，简写为 $w(e)$。这里我们规定每条边(弧)e_{ij}上有权重 $w(e_{ij})$(表示网络传输的耗费)，并满足可加性(这里我们只考虑满足可加性的情况[1])，即对于路径 p 为 $v_0\to v_1\to\cdots\to v_n$，有 $w(p)=\sum_{i=0}^{n}w(e_{i,i+1})$。这样，可以定义源节点 v_i 到目标节点 v_j 的最佳路径 P^* 为：$p^*=\arg\min_{\text{所有}i\text{到}j\text{的路径}p}w(p)$。算法希望逼近最近路径，尽量降低网络消耗。

在模型中，为方便描述，我们加入以下几个定义。

定义2 节点 v_i 的邻居节点集 $N_i=\langle 1,2,3,\cdots,m\rangle$。如果有 $v_j\in N_i$，则必有 $e_{ij}\in E$。

定义3 节点 v_i 的混和策略$\sigma_i=\langle\sigma_{i1},\sigma_{i2},\cdots,\sigma_{in}\rangle$，为节点 v_i 到各个节点的混和策略序。有 $\sigma_{ij}=\langle\delta_{ij}^1,\delta_{ij}^2,\cdots,\delta_{ij}^m\rangle$，其中 δ_{ij}^k 中的 k 表示在节点 v_i，对于目的地为 v_j 的包节点 v_i 转发给它的邻居 k 的概率。有 $\sum_{k\in N_i}\delta_{ij}^k=1$。

定义4 耗费函数 C_j 表示目的地为 v_j 的包传送到以后的收益($C_{j_\min}$表示目的地为 v_j 的包传送的最小耗费)；其中 $C_j=w(p)=\sum_{i=0}^{n}w(e_{i,i+1})$($p$ 为传送路径)。

定义5 效益对(C_j,k)为将包从源节点传送到目的节点后分发给的传送路径上节点 v_i 的效益对。其中 C_j 为传送包后的耗费。k 表示传送到节点 v_i 时，v_i 选择的传送方向为邻居 v_k。

为讨论方便，我们加入一个控制机构，它负责观察包的传递过程，记录包的传送路径，并且在包到达目的节点后，为每个除目的地节点外的所经过的节点 v_i 分发效益对(C_j,k)，其中 k 为 v_p 节点所选择的转发对象。

同时，我们作如下约定：

- 每个节点每次只能发出一个包。
- 时间分片，每个时间片内包从一个节点传送到下一个节点。

4.2 路由算法

将环境视为部分可观察环境，节点并不需要了解整个网

络具体结构和状况。路由节点通过建立一张路由表，在需要把包转发到目的节点 v_j 时，节点利用历史数据在邻居中选择一个转发概率最高的方向，把包转发给这个邻居，这样一直转发到目标节点。

节点 v_i 的路由算法：

初始化：所有 $C_{j\_\min}=\infty$，$\delta_{ij}^k=1/|N_i|$
Loop 收到邻居 v_p 发来的包
If 该包不是收益对 than
  Begin
    目的地为 v_j 的包，$(j\neq i)$放入等待转发队列
    从等待转发队列首出一个包，设目的地为节点 v_j
    在 σ_{ij} 中按概率 δ_{ij}^k 找一个 k，其中 k 不等于 p
    将包转发给 k
  End
Else
  Begin
  收到效益对(C_j,k)(发给目的地为 v_j 的包，经 v_i 节点。节点 v_i 当初选择转发给了 k)
  做如下更新：
    当 $C_j\leqslant C_{j_\min}$时　$\delta_{ij}^k=\delta_{ij}^k+\Delta$　令 $C_{j\_\min}=C_j$
    否则　$\delta_{ij}^k=\delta_{ij}^k-\Delta/|N_i-1|$($\Delta$ 为初始化参数)
    调整其他概率，使之满足 $\sum_{k\in N_i}\delta_{ij}^k=1$ 条件。
  End
End Loop

控制机制端算法如下：

Begin
  观察每个包经过的节点序 $p=\langle i,\cdots,p,k,\cdots,j\rangle$，发送端为 v_i，接收端为 v_j。
  包到达后计算 $C_j=w(p)$
  给每一个经过的节点 v_p 传送收益对(C_j,k)，(k 为 v_p 节点所选择的转发对象)
End

5 示例

v_i 为不同的节点。它们的连接关系网络结构如图 3。

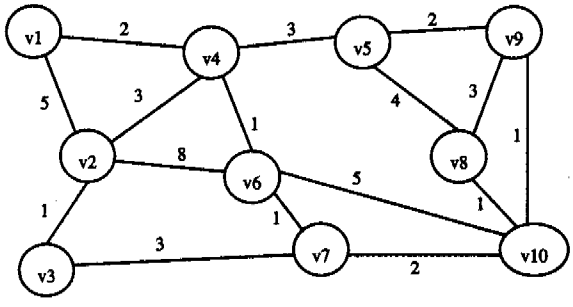

图 3　网络情况

开始时，每个节点建立一张表。以节点 $v4$ 为例：

列为节点 $v4$ 的邻居(现有 1,2,5,6 共 4 个)。行为包可能的目标节点。$C_{j\_\min}$ 表示到目标节点 v_j 的最小收益。$\delta_{ij}^k=1/|N_i|=0.25$。初始化后如表 1。

表 1　节点 v_i 开始时的路由表

	1	2	5	6	$C_{j\_\min}$
1	$\delta_{41}^1=0.25$	0.25	$\delta_{41}^5=0.25$	0.25	$r_{1\min}=\infty$
2	0.25	0.25	0.25	0.25	∞
3	0.25	0.25	0.25	0.25	∞
5	0.25	0.25	0.25	0.25	∞
6	0.25	0.25	0.25	0.25	∞
7	0.25	0.25	0.25	0.25	∞
8	0.25	0.25	0.25	0.25	∞
9	0.25	0.25	0.25	0.25	∞
10	0.25	0.25	0.25	0.25	∞

假设此时有一个包要从源节点 N1 发送到目标节点 $v10$。$v1$ 将包发给了 $v4$，$v4$ 收到包后查看路由表，在 δ_{41}^2，δ_{41}^5，δ_{41}^6 中按概率选取一个，设选到 5。将包转发给 $v5$。假设最后该包的传送路径为$\langle v1,v4,v5,v8,v10\rangle$。到达后控制端计算 $C_1=2+3+4+1=10$。并将收益对发给$\langle v1,v4,v5,v8,v10\rangle$中的 $v1,v4,v5,v8$。发给 $v4$ 的收益对为$(r_{10}=10,v5)$。

N4 在收到收益对时进行更新。$C_1=10<C_{1\_\min}=\infty$，所以 $C_{1\_\min}=0.8$，$\delta_{41}^5=\delta_{41}^5+0.03=0.28$(设 $\Delta=0.03$)。并调整其他概率，得到 $\delta_{41}^2=0.25-0.01=0.24$，$\delta_{41}^1=0.25-0.01=0.24$，$\delta_{41}^6=0.25-0.01=0.24$。表 1 中的第一行更新为表 2。

表 2　节点 v_i 更新过的路由表(部分)

	1	2	5	6	$C_{j\_\min}$
1	$\delta_{41}^1=0.24$	0.24	$\delta_{41}^5=0.28$	0.24	$r_{1\min}=10$

每次需要转发时，都根据路由表来选择转发方向，收到收益对时对路由表进行更新。这样，路由表中的概率能很好地反映网络情况(概率大的方向网络情况较好)，同时也保证路由节点多条路径的选择。

结论和后续工作　本文给出了一个基于增强学习算法的路由模型。该模型在不增加路由复杂度的同时能很好地发现、避开网络拥塞等。和现有的一些路由算法相比，虽不能保证每个包都走最佳路径，但其复杂度大大降低，并能利用历史数据调整路由策略，及时反映网络情况。下一步我们会进一步改进模型设计，让模型更接近实际应用。

参 考 文 献

1　崔勇，吴建平，徐恪，等. 互联网络服务质量路由算法研究综述. 软件学报，2002,13(11)：2065～2075
2　刘芳，杨海潮. 参数可调的克隆多播路由算法 . 2005 Journal of Software
3　胡进锋，黎明，郑纬民，等 . 带宽自适应的 P2P 网络路由协议 . 2005 Journal of Software
4　崔勇，徐恪，吴建平. 性能可调的启发式多约束路由算法. 电子学报，2002，30(12A)：1968～1973
5　Khoussainov R. Economics of Distributed Web Search：A Machine Learning Approach：[PhD thesis]. Dublin：Department of Computer Science National University of Ireland，2004
6　Peshkin L. Reinforcement Learning by Policy Search：[PhD thesis]. Weizmann Institute of Science，1995
7　Khoussainov R，Kushmerick N. Distributed Web Search as a Stochastic Game. Distributed Multimedia Information Retrieval 2003. 58～69
8　Peshkin L，Savova V. Reinforcement learning for adaptive routing. 2003
9　Wang Bin，Hou J C. Multicast Routing and Its QoS Extension：Problems，Algorithms and Protocols. IEEE Network，Jan/Feb 2000
10　Wang Z，Crowcroft J. QoS Routing for Supporting Resource Reservation. IEEE Journal on Selected Areas in Communications，September 1996
11　Shin K G，Chou C C. A distributed route-selection scheme for establishing real-time channel. In：Puigjaner R. ed. Precedings of IFIP Sixth International Conference on High Performance Networking (HPN'95). Palma，Spain：Chapman & Hall，1995. 319～29
12　Nurmi P. Modelling routing in wireless ad hoc networks with dynamic Bayesian games. 2004
13　MacKenzie A B，Wicker S B. Game theory and the design of self-configuring，adaptive wireless networks. Cornell University，IEEE Communication Magazine，2001
14　Khoussainov R，Kushmerick N. Learning to Compete in Heterogeneous Web Search Environments. IJCAI 2003. 1429～1431