基于用户偏好的智能业务选取研究

黄海清[1],张 平[1],张曦文[2]
(1. 北京邮电大学,北京 100876;2. 航天科工集团二院,北京 100854)

摘 要: 将马尔可夫判决过程和强化学习算法相结合,给出了异构无线网络环境下用户业务偏好评估模型的技术框架. 为动态环境下用户需求的感知、量化和适配特征的研究提供了基本的数学描述,对解决用户体验的评价问题和业务与业务环境的适配问题提供了新的研究思路. 仿真结果表明构建的模型能够在满足用户偏好的前提下智能选择业务.

关键词: 效用理论; 用户偏好; 马尔可夫判决过程; 强化学习
中图分类号: TN91 **文献标识码**: A **文章编号**: 0372-2112 (2006) 12A-2537-04

Modeling of User Preference Based on Agent for Service Selection

HUANG Hai-qing[1], ZHANG Ping[1], ZHANG Xi-wen[2]
(1. *Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
2. *The 2th Institute of China Aerospace Science and Industry CORP, Beijing* 100854, *China*)

Abstract: We present a technical architecture for user preference model, and show the nature of the problem represented within a Markov Decision Process combined with adaptive reinforcement learning algorithm. We provide a possible candidate solution for user modeling dynamically to satisfy the users expected preference based on minimal or missing information, it is also a exploration for the evaluation of the user experience when selecting service providers. Simulations of the representative user models show that the adaptive reinforcement learning solutions are effective.

Key words: utility theory; user preference; Markov decision process; reinforcement learning

1 引言

互联网络和无线移动网络的发展趋势,是建立以无线接入技术为基础的多网融合、网络集成与网间协作环境,并向用户提供最佳的业务体验. 未来移动通信研究的出发点和根本目的就是不断发现和满足用户的通信需求,在业务开发和业务应用过程中充分考虑用户的主客观感受,研究和创建新的用户需求模型和行为机制. 随着无线通信网络和计算机网络的不断融合,业务环境的多样性和移动业务逻辑的复杂性推动着业务理论的研究向深层次发展,如何将用户的需求抽象为可表示、可量化、可感知、可适配的特征,是探索未来移动业务研究必须要解决的关键问题.

本文基于用户偏好学习理论,通过将 MDP(Markov Decision Process)建模方法与智能强化学习算法 RL(Reinforcement Learning)相结合,构建了异构无线网络环境下移动用户业务偏好评估模型的基本技术框架. 仿真结果表明,所建模型对用户偏好具有较强的在线学习能力,能够以用户体验为目标智能地选择业务.

2 用户偏好提取技术

未来用户将处在一个复杂的通讯网络环境中,业务的内容和提供方式将面临巨大的变化. 各种业务提供商和运营商将为用户提供丰富多彩的选择,服务的价格、速率、QoS(Quality of Service)等都会各有不同,但一切的服务都将以用户的需求为中心,为用户提供最佳的服务体验. 通过设计智能代理系统[1],可以根据用户以往的使用经验,自动评估并接入服务. 用户的偏好并不是一成不变的,在特定的目标和场合下具有不确定性,因此用户偏好评估模型的构建实际上是研究基于不确定信息的自适应反馈判决问题[2,3].

偏好提取的目标是构建精确的用户模型,从而判决系统可以通过用户模型协助用户完成任务. 构成这些模型的理论基础是采用判决和多属性效用理论[4],对判决问题或场景的输出和选择作出评估.

输出由一系列属性变量的值来定义,$X=\{X_1,\cdots,X_n\}$. 判决问题的输出集合 O 包含于输出空间 Ω, $O\subseteq\Omega$, $\Omega=\{X_1\times X_2\times\cdots\times X_N\}$. 为了基于输出空间 O 作出判决,判决系统经常

收稿日期:2006-02-04;修回日期:2006-06-11
基金项目:863 高科技项目(No. 2003AA12331004)和三星电子中国通信研究院联合资助

要根据用户偏好决定输出的次序,称为偏好关联,用符号 $\succeq$ 表示. 假定 $o_i, o_j \in O$,如果 $o_i \succeq o_j$,表明 o_i, o_j 相比,用户更偏好于 o_i. 偏好关联一般是由值函数推导而来的, $v(o): O \to R$. 值函数能够在输出集和属性变量上进行运算,有下式:

任意 $a, b \in X_i$, $a \succeq b \Leftrightarrow v(a) \geq v(b)$.

由于效用函数必须考虑用户对风险的态度,因此效用函数在输出空间上推导出的偏好关联是基于概率分布的,如果 P_{r_i}, P_{r_j} 分别对应动作 a_i, a_j,则有:

$$a_i \succeq a_j \Leftrightarrow \sum_{o \in O} \Pr_i(o)\, u(o) \geq \sum_{o \in O} \Pr_j(o)\, u(o) \tag{1}$$

上述关联式暗示着系统实际上在寻求具有最大期望效用(MEU ——Maximum Expected Utility)的动作. 由于在系统交互的初始阶段,用户的效用函数经常是未知的,因此偏好提取的主要目标就是构建精确的效用函数和用户偏好关联表示.

为了构建简单而易于管理的效用函数,引入条件偏好的概念. 对一个条件偏好关联, $Y \subset X$,令 α, α' 分别为 Y 中的两个属性, $Z \subseteq X - Y$, γ 为 Z 中的分配值. 则当 $\alpha \succeq \alpha' \Leftrightarrow (\alpha, \gamma) \succeq (\alpha', \gamma)$ 时,表明系统相对 γ 而言,对 α 的偏好强于 α'. 在 X 上引入一个概率分布 Pr,则存在唯一的分布 $\Pr^{\gamma}$, $\Pr^{\gamma}$ 在 Y 上的边缘分布是 Pr, $\Pr^{\gamma}$ 的值在 1 和 γ 之间. 假定一个效用函数具有一个偏好关联次序 $\succeq$,那么在 X 上的条件偏好可以定义为

$$\Pr_a \succeq \Pr_b \Leftrightarrow \Pr_a^{\gamma} \succeq \Pr_b^{\gamma} \tag{2}$$

定义了具有概率分布的条件偏好,就可以具体描述效用独立的表达式. 如果 $X = (Y, Z)$,且 Y 效用独立于 Z,可知[5] 偏好结构的效用函数具有下列形式:

$$u(X) = f(X - Y) + g(X - Y)\, h(Y) \tag{3}$$

其中 g 是个确定的函数[5].

如果下面条件满足,偏好结构可以定义更强的独立性. 令 X 分区为 $X_1, \cdots, X_k$,让 p_1, p_2 为两个任意的概率分布,对所有的 X_i,共享同样的边缘分布. 如果 p_1, p_2 在偏好结构中彼此不相关,则 $X_1, \cdots, X_k$ 是加性独立的. 对于加性独立的输出空间,变量集合具有下述形式:

$$u(X) = \sum_{i=1}^{k} f_i(X_i) \tag{4}$$

f 是针对属性 X_i 的子效用函数.

3 MDP 用户偏好模型框架

用 *MDP* 建模框架可以比较准确的描述上述的用户建模问题. 离散时间 *MDP* 模型[8,9] 应由五重组组成: $\{S, A(i), p_{ij}(a), r(i,a), V, i, j \in S, a \in A(i)\}$. 各元素分析如下: (1) S 是系统所有可能的状态所组成的非空的状态集. 由业务场景分析可知系统状态包含的主要因素有:用户状态集合、业务特征集合及用户偏好集合. 其中用户状态集合用 C 表示,元素 $c \in C$ 为有限维的随机序列,包含如用户当前的位置、服务请求期限、应用等内容. 集合 C 中的元素根据用户的不同目标分成各个子集,用 c^g 表示. 所有可能的业务特征集合用 P 表示,集合中的元素 P 由 n 个具体特征 f_i 构成, $P = (f_1, \cdots, f_n)$,具体的业务特征会随着时间、地点、服务种类及用户漫游的状态而有所不同. 用户的偏好可以用集合 U 表示,其中的元素 U 代表了用户在 c^g 和业务特征 P 的条件下,基于一定概率分布对业务的选择排序. 综上分析,在某个地点(用 loc 表示), t 时刻,系统状态变量 $S \in S$ 可以表示为:

$$S^t = (c^g, t, loc, P, U) \tag{5}$$

(2) MDP 的另一个重要元素 $A(i)$ 是在状态 i 处可用的有限决策集. 它应由用户和代理共同实现,使得系统状态发生转移. 在 MDP 过程中,这种状态的转移表示了系统状态的配置发生了变化. 在本文讨论的问题中,用户决策的变化体现在用户位置的改变、对服务质量或价格的要求、目标应用的改变、用户偏好的改变等因素. 针对某个用户 u,可以简化表示为 $A^u = \{\Delta loc, \Delta app, \Delta U, \phi\}$,分别代表位置、应用、偏好的变化, ϕ 表示当前状态下无动作.

(3) 当系统在决策点时刻 n 处于状态 i,采取决策 $a \in A(i)$ 时,系统在下一个决策点 $n+1$ 时处于状态 j 的概率为 $p_{ij}(a)$,与决策时刻 n 无关.

(4) 当系统在决策时刻点 n 处于状态 i,采取决策 $a \in A(i)$ 时,系统在本阶段获得的报酬函数为 $r(i, a)$.

(5) V 为获得最佳报酬设定的准则函数.

为了表示在当前系统状态和用户的偏好下,用户对服务满意度,可以定义函数

$$f^{c^g}(P_i) = \sum_{1}^{n} w(i, j, c^g)\, v(P_{ij}) \tag{6}$$

进行量化分析,其中 w 是加权系数, $v(P_{ij})$ 是针对业务的特征进行评价的函数,函数形式应根据具体情况有所调整.

MDP 问题的一个重要特征是系统通过定义策略序列 $\pi = (\pi_0, \pi_1, \cdots)$,当系统在时刻 n 时的历史 $h_n = (i_0, a_0, i_1, a_1, \cdots, i_{n-1}, a_{n-1}, i_n)$ $n \geq 0$ 时 $(i_k \in S, a_k \in A(i_k))$ 分别表示系统在第 k 个时刻点所处的状态和采取的决策),策略则按 $A(i_n)$ 上的概率分布 $\pi_n(\cdot \mid h_n)$ 采取决策. 对应策略 π,在 MDP 中与之相关的随机序列 $R = (R_0, R_1, R_2, \cdots)$ 为报酬过程,其中 $R_n = r(S_n, a_n)$ 是系统在时刻 n 采取决策 a_n 时获得的报酬. 针对本文讨论的问题最优策略[6] 的选择应保证在有限维空间内获得最佳的期望总报酬,即

$$V_N(\pi, i) = \sum_{n=0}^{N} E\{ r(i_n, a_n) \}, \quad i \in S \tag{7}$$

式中的报酬 r 就是代理收到的用户对所选业务的反馈,所以最佳策略的选择就是要体现用户对所选业务的最大满意度,策略 π 又可由下式表示:

$$\pi = \arg\max E\{ \sum_{n=0}^{N} f_n \} \tag{8}$$

从式(7)、(8)的分析可知,当系统越来越复杂,难以获得用户精确模型,且动态特性为时变的时候,常规的控制方法难以解决问题. 对用户偏好的学习强调的就是对变化环境的适应,强化学习方法 RL 应该是可选的重要方法之一. 通过 RL 方法能够较好解决移动代理和随机环境交互获取信息后,如何获得最优策略的问题.

4 基于 RL 方法的最佳策略选择

解决最佳策略问题有很多种方法,包括 dynamic program-

ming、Gittins allocation indices 及 learning automata[7],但这些方法对文中描述的模型并不合适. 我们的目的是要逐步增加 MDP 过程的复杂性,因此希望采用启发式的、易于计算的、接近最佳的动作选择技术. 将用户对代理选择不同动作的偏好,用一个均值为 $Q^*(a_i)$ 的奖赏概率分布来表示,如果这个概率在一段时间内为常量,则认为用户的偏好是静态的,反之则称为非静态的. 由于代理缺少概率分布信息,必须反复估计和更新动作值,因此动作选择前需要一个智能学习的过程.

在动作 a_i 经过 k 次选取后对 $Q^*(a_i)$ 更新可以通过下式的指数加权平均来完成:

$$Q_{k+1}(a_i) = Q_k(a_i) + \alpha_k(a_i)[r_{k+1}(a_i) - Q_k(a_i)] \quad (9)$$

其中 $0 \leq \alpha_k(a_i) \leq 1$,表示对动作 a_i k 次选择后的学习速率. 如果 $\alpha_k(a_i) = 1/k$ 则表示学习速率在每次更新时都是变化的. 如何根据动作的估值来选择合适的动作,我们研究了三种不同的策略:greedy,ε-greedy 及 softmax[6,7]. 第一种策略通过选择具有最高估值的动作 $a_i^* = \arg\max Q_t(a_i)$ 来开发当前代理的知识. 然而用户与代理交互的增加会有利于探索动作空间,长期的动作奖赏将优于短期的动作奖赏值. 探索过程中,动作选择的概率应基于一个常数速率 ε(即 εgreedy 策略)或由 Gibbs 或 Boltzman 分布给出:

$$P_t(a_i) = \frac{e^{Q_t(a)/T_t}}{\sum_{a \in A}^{n} e^{Q_t(a_j)/T_t}}$$

$$T_t = T_0(1 - \mu)^t \quad (10)$$

即是当 ε 为变值时的 softmax 策略.

学习最佳的动作也可以通过保持参考奖赏值的办法作为判决准则,采用这种方法的技术有 RC(Reinforcement Comparison). RC 方法对每个动作的偏好是独立衡量的,同样依据 softmax 的准则,但由 $\rho_t(a)$ 决定动作选择概率. 偏好的更新见下式,其中 α 是正值的步长参数,$\bar{\gamma}$ 为参考奖赏值:

$$\rho_{t+1}(a_t) = \rho_t(a_t) + \alpha[\gamma_t - \bar{\gamma}_t] \quad (11)$$

参考奖赏值是近期得到的所有奖赏的增量平均:

$$\bar{\gamma}_{t+1} = \bar{\gamma}_t + \alpha[\gamma_t - \bar{\gamma}] \quad (12)$$

$0 \leq \alpha \leq 1$ 同样是学习速率.

PM(Pursuit method)方法兼顾了 εgreedy 和 RC 的特点,同时保持了动作值和动作偏好.

如果用 $\pi_t(a)$ 表示依据 softmax 决定的 t 时刻动作 a 的选择概率,且 $t+1$ 时刻贪婪动作 $a_{t+1}^* = \arg\max_a Q_{t+1}(a)$,则 $a_{t+1} = a_{t+1}^*$ 概率由下式给出

$$\pi_{t+1}(a_{t+1}^*) = \pi(a_{t+1}^*) + \alpha[1 - \pi_t(a_{t+1}^*)] \quad (13)$$

而选择其他动作的概率为:

$$\pi_{t+1}(a) = \pi_t(a) + \alpha[0 - \pi_t(a)] : a \neq a_{t+1}^* \quad (14)$$

5 仿真结果

为了验证上文 MDP 模型对用户偏好的学习能力. 选用了 εgreedy、Gibbs softmax、PM、RC 方法. 其中 softmax 方法中 $T_0 = 10$, $u = \{0.05, 0.01\}$. 上述所有算法中的指数平均加权参数 $\alpha = 0.1$, softmax 算法当 $u = 0.01$ 时, $\alpha = 1/k$.

图 1 给出用户偏好在静态条件下,模型通过对用户偏好的学习而对业务作出最佳选择的过程,用每次动作学习后获得的平均最佳奖赏值来表示. 共执行 10000 次任务,每个任务包含 1000 次的动作学习,奖赏 $Q^*(a_i)$ 的初始值是 - 1 和 + 1 间的任意值. 图 1 表明了模型采用不同的学习方法,对最佳业务选择及长期平均奖赏的影响程度. εgreedy 和 PM 算法在学习初始阶段改进很快,但长期奖赏值并不是最优的; $u = 0.01$ 时 softmax 在初始探索阶段表现一般,但长期奖赏值却是最优的,表明代理采用 softmax 能够更大限度地满足用户对服务的需求.

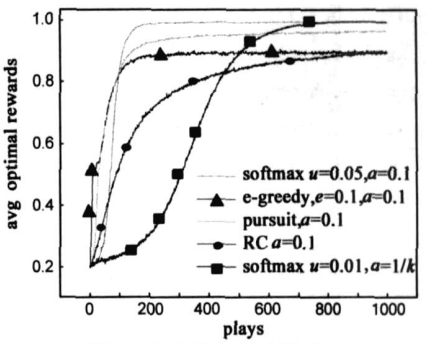


图 1　用户偏好静态模型

图 2 给出用户偏好在非静态条件下,模型通过对用户偏好的学习而对业务作出最佳选择的过程,同样用每次动作学习后获得的平均最佳奖赏值来表示. 在仿真环境中假设用户的偏好在每 150 次动作选择后随机增加或减少 0.4. 偏好的这种规律性变化将有效验证模型学习用户偏好的能力. 由图 2 可知 εgreedy 方法具有最佳性能,在每次偏好变化后都能快速满足新用户的需求. 而 $\alpha = 1/k$ 的 Gibbs softmax 方法性能最差,不能快速捕捉用户需求的变化.

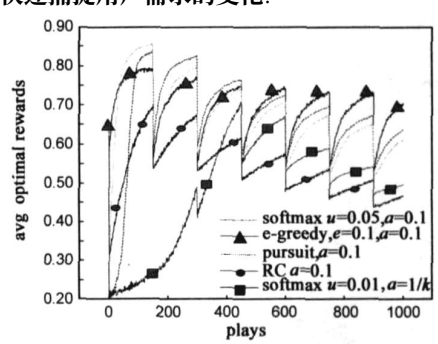


图 2　用户偏好非静态模型

6 结论

本文将 MDP 建模方法和 RL 技术相结合,给出了无线异构网络环境下用户业务偏好评估模型的基本技术框架,为量化、感知用户需求和智能业务的选择与适配问题提供了新的研究思路. 仿真结果表明构建的模型能够在满足用户偏好的前提下智能的选择业务. 为进一步提高偏好学习的能力,一方面要深入研究更加复杂 MDP 模型下智能学习算法的评估,还要将目前采用的单一代理判决机制扩展为多代理系统的协商机制,多代理系统协商机制的研究将为用户偏好提取建模开拓一个新领域.

参考文献:

[1] B Liver, J Altmann. Social carrier recommendation for selecting services in electronic telecommunication markets: A preliminary report[R]. Technical Report TR-97-033, ICSI, Berkeley, CA, USA, 1997.

[2] U Chajewska, D Koller, R Parr. Making rational decisions during adaptive utility elicitation[A]. In Proceedings of the Seventeenth National Conference on Artificial Intelligence [C]. Austin, TX, USA, 2000. 363 - 369.

[3] Craig Boutilier. A POMDP formulation of preference elicitation problems[A]. In Proceedings of American Association of Artificial Intelligence[C]. Edmonton, Alberta, Canada, 2002. 239 - 246.

[4] Simon French. Decision Theory: An Introduction to the Mathematics of Rationality [M]. New York, USA, Halsted Press, 1986.

[5] Daniel P Boulet, Niall M Fraser. Improving preference elicitation for decision support systems[J]. IEEE Trans, 1995. 1574 - 1579.

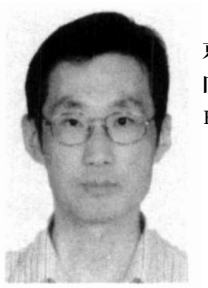
[6] Leslie Pack Kaelbling, Andrew W Moore. Reinforcement learning: a survey[J]. Journal of Artificial Intelligence Research, 1996, 237 - 285.

[7] R S Sutton, A G Barto. Reinforcement Learning[M]. MIT Press, Cambridge, MA, 1998.

[8] 胡奇英, 刘建庸. 马尔可夫决策过程引论[M]. 西安: 西安电子科技大学出版社, 2000.

[9] 刘克. 实用马尔可夫决策过程[M]. 北京: 清华大学出版社, 2004.

作者简介:



黄海清 男, 1971 生于辽宁沈阳, 现为北京邮电大学 2003 级博士研究生. 主要研究方向为无线通信.

E-mail: haiqingmail2003 @yahoo. com. cn

张 平 男, 1960 年出生, 北京邮电大学教授, 博士生导师. 研究方向为无线通信.

张曦文 女, 1971 年出生, 航天二院中心军代室高级工程师. 研究方向为通信与电子系统.