电子商务中基于Q学习的动态交叉销售方法①

程 岩
(华东理工大学商学院，上海 200237)

摘要：动态交叉销售是电子商务中的一种新型营销手段.在已知关联规则和商品库存水平的情况下，要研究两个决策问题：(1)如何选择交叉销售的商品组合(2)如何为商品组合确定合理的价格，从而使经销商获得最大收益.首先将动态交叉销售映射为事件驱动的马氏决策过程模型，其次结合关联规则理论提出了知识驱动的Q-学习算法，K-Q-learning，用该算法来求解动态交叉销售问题具有较高的效率和效用.

关键词：电子商务；交叉销售；Q-学习；关联规则
中图分类号：TP311 **文献标识码**：A **文章编号**：1007-9807(2008)03-0106-08

0 引 言

交叉销售研究领域的代表人物Kamakura认为："客户与经销商的接触点越多，那么他的转移成本就越高，因此，交叉销售是一种培养稳固的顾客关系的重要工具."[1]交叉销售是电子商务企业的一种重要营销手段，美国E-Tailing Group市场调查公司的研究表明，全球100家最大的在线零售企业中有62家企业采用各种形式的交叉销售策略[2].

在传统的商务模式下，受获取信息能力以及计算能力的限制，营销人员无法在交易过程中实时制定优化的交叉销售商务条款，只能根据对未来市场需求的预测将交叉销售的商务条款事先制定出来，这些商务条款在一段时期里具有相对稳定性，不会随客户个体的特殊需求而动态改变.近年来，Amazon.com、Travel-related.com等电子商务企业利用电子商务系统的实时信息处理能力，实施了一种商务条款可以随客户的交易进程而实时制订的交叉销售策略[2]，在实际运行中，这些商务条款是由智能代理而非营销人员来制定和实施.宾西法尼亚大学的Netessine等人将这一类交叉销售定义为"动态交叉销售"，而将传统的需要事先确定商务条款的交叉销售定义为"静态交叉销售"[2].根据Netessine等人的定义，动态交叉销售是指：当买方提出对一种商品的购买请求时，经销方能够主动建议买方再同时购买另外一些商品，并对由这些商品组成的消费套餐提供一个折扣价格[2].由于动态交叉销售的决策是伴随客户的交易进程而实时制定的，因此，实时决策能力是实施动态交叉销售的关键因素.

目前，有关交叉销售的研究成果大多集中于交叉销售机会的识别问题上，而对策略优化问题的研究非常少[3].交叉销售的策略优化问题是指：在营销资源有限的条件下，如何优化交叉销售的商务条款，从而为企业带来最大收益.对此，Netessine等人在提出动态交叉销售这一概念的同时，重点探讨了动态交叉销售策略的优化模型[2].在他们所提出的随机动态规划模型中，策略的优化需要同时考虑价格折扣水平、价格折扣对客户购物行为的影响以及商品库存资源三者之间的互动关系.尽管Netessine等人从库存资源限制的角度给出了科学的策略优化模型，但是，求解随机动态规划模型需要巨大的计算花销.为了保证决策的实时性要求，Netessine等人在求解过程中引入了启发式信息.然而，启发式信息带有很大

① 收稿日期：2005-11-29；修订日期：2007-11-20；
作者简介：程 岩(1967—)，男，辽宁沈阳人，博士，讲师，Email：chengyan66@163.net.

的经验色彩，一般只有在所允许的计算时间里不可能找到最优解时，才使用这一策略. 因此，Netessine等人的求解策略是在牺牲最优解的前提下，获得实时求解的能力. 另外，Netessine 等人认为，适合与目标商品进行交叉销售的商品候选集合并不是很大，但他们没有给出如何确定这一集合的方法，而只是假设这一集合为已知. Lerzan 等人对银行呼叫中心的动态交叉销售问题进行了专门研究[4]，他们将动态交叉销售的策略优化问题映射为马尔可夫决策模型，模型考虑了通讯能力限制条件下如何优化动态交叉销售的实施，从而使企业获得最大收益. 在他们的模型中，系统的状态仅仅用电话占线数量和空闲电话数量这两个变量来描述，因此系统的状态空间很小，方便了模型的求解. 银行呼叫中心业务的这一特殊性与本文研究的问题有很大差距，因此，Lerzan 的方法不能用来解决本文所要研究的问题. 自从 2004 年动态交叉销售的概念被提出以来，有关这一领域的专门研究还很少. 除上述两篇文献外，本文仅仅检索到 2005 年发表的一篇类似的短文，该文献针对旅游社的动态捆绑销售问题进行了分析[5]，但该文献仅仅从概念上对动态捆绑销售进行了介绍，没有给出具体的实施办法. 通过对上述文献内容的分析，本文发现电子商务环境下实施动态交叉销售的现有理论研究成果中还没有一种特别有效的计算方法，能够既保证决策的实时性又能同时保证获得最优解.

电子商务零售行业中实施动态交叉销售的难度在于，在确定商品组合和价格折扣水平时必须参考商品的库存水平. 例如，尽管某商品组合在一定的价格折扣水平下对买方具有很大的吸引力，但如果这些商品库存水平非常低，正处于供不应求的状态，那么，该商品以非折扣价单独出售可能会给经销商带来更大的收益. 因此，各种商品的库存状态是动态交叉销售策略优化的依据. 但是，在电子商务零售业中，商品的种类往往很多，库存系统的状态空间非常巨大[6]. 例如，假设有 N 种不同的商品，每种商品的最大库存量为 M，用 x_i 表示第 i 种商品的库存量，用向量 $(x_1, x_2, \cdots, x_N)^{\mathrm{T}}$ 来描述库存状态，那么，系统可能出现的状态数量为 N^M. 此外，在动态交叉销售决策中，系统不仅要参考商品的库存状态，还要参考客户首先提出了什么购买申请(即事件)，这样才能做出科学的决策. 可能出现的事件与产品库存共同组成了一个巨大的状态空间，在此状态空间下的马尔可夫决策问题是个大规模马氏决策问题[6]. 一些文献对大规模复杂随机系统的决策问题进行了研究，这些文献大多采用状态集结的方法，但是，状态集结的策略需根据具体问题的特点而不同，目前还没有通用的状态集结策略[6].

动态规划是解马尔可夫决策问题的常用方法，但动态规划方法求解模型的计算量很大，在线计算不能满足决策实时性要求. 最近，机器学习理论以及随机动态规划等领域若干思想的集成产生了一种新的方法学 —— 强化学习，也称神经动态规划，这种方法为求解马尔可夫决策问题提供了一种新的策略. Q- 学习算法是一种基于仿真的强化学习算法，是由 Watkins[7] 针对折扣型马尔可夫决策问题提出的. Q- 学习过程可以在仿真环境下离线进行，其训练结果是一张 LOOK UP 表. 当有决策需要时，系统只需查询 LOOK UP 表就可以立刻采取决策，而无需在线计算，因此，Q- 学习方法能够满足决策的实时性要求. 传统的 Q- 学习方法是对状态空间进行盲目地搜索，通过无数次迭代计算来获得稳定收敛的 LOOK UP 表. 当状态空间非常巨大时，就很难获得稳定收敛的 LOOK UP 表[8,9].

由于动态交叉销售决策中所面临的商品库存状态空间非常巨大，因此，采用传统 Q- 学习方法求解动态交叉销售问题的计算花销十分巨大. 但实际上，大多数商品之间并不适合交叉销售，适合与某一特定商品组合为消费套餐的商品选项并不是很多. 商务智能中的一种重要知识 —— 关联规则，反映的是各种商品在销售中的关联关系，利用这种知识可以发现哪些商品可以和目标商品进行捆绑销售，哪些商品不能够与目标商品组成消费套餐. 为此，本文首先提出一种事件驱动的马氏决策模型来描述动态交叉销售的决策过程，并进而提出一种知识驱动的 Q- 学习方法来求解该模型. 根据本文提出的算法，关联规则被植入 Q- 学习过程，在 Q- 学习的每次迭代计算中，算法不是盲目地选择决策，而是在关联规则指导下，以较高的概率选择比较优良的决策，从而加快学习的收敛速度. 仿真试验的结果表明，本文提出的方法具有较高的效率和精度.

1 动态交叉销售的事件驱动型马氏决策过程模型

本文研究这样一类动态交叉销售问题:1) 商品按单件出售,对同一种商品,买方每次只买一件商品;2) 卖方每次只选择一种商品与买方的第一选择组成消费套餐.

动态交叉销售的决策问题是一类特殊的马氏决策过程问题. 普通型马氏决策过程(MDP) 没有考虑系统中出现的事件对系统状态和决策的影响,而动态交叉销售中的每个决策除了要考虑各种商品的库存状态,还要考虑消费者首先提出购买的是什么商品. 因此,可以将动态交叉销售的决策过程描述为一个由五元组$\{S, E, D, P, f\}$定义的事件驱动型马尔可夫链:$\pi = \{(s_t, e_t) \rightarrow d_t \mid s_t \in S, e_t \in E, d_t \in D, t = 1,2,\cdots,N\}$. 本文对其中的要素分别做如下定义:

有限离散时间序列 t: 假设某营销活动周期的时间长度为 T,将时间 T 划分为足够多的 N 个子时间段,保证每个时间段里最多只有一位买方到达. 令 $t(t = 1,2,\cdots,N)$ 表示某个时间段的序列号.

有限状态空间 S:本文用商品的库存量来描述系统的状态,各种可能的库存状态构成了有限空间 S. 假设在线零售商经营 m 种商品,商品 i $(i = 1,2,\cdots,m)$ 在某时间点 t 的库存量记为 I_i^t,那么此时的状态 s_t 可以描述为向量 $\boldsymbol{s}_t = (I_1^t, I_2^t, \cdots, I_i^t, \cdots, I_m^t)^{\mathrm{T}}$.

有限事件集合 E:如果有一位买方登录电子零售网站,并申请购买某商品 i,那么这一事件记为:e_i,其中 $i = 1,2,\cdots,m$ 表示商品的序号,各种可能事件构成有限事件集合 E. 令 $\lambda_i (0 \leqslant \lambda_i \leqslant 1)$ 表示有客户登陆且他登录时有意购买商品 i 的概率,规定 $\sum_{i=1}^{m} \lambda_i \leqslant 1$,这是因为根据前面的规定,在每个时间段 t 最多只有一个客户登录,而该客户登录网站的目的可能仅仅是浏览网页信息,无意购买任何商品[10]. 做此规定是由于电子商务系统中大量在线客户无任何购买意愿,在线客户由浏览转变为购买的概率远远小于传统业务模式中的客户[10]. 文献[11] 对此进行了研究,并提出了一种利用数据挖掘技术从 web log 数据中发现在线匿名客户购买意愿强度 λ_i 的方法. 令 p_i 表示商品 i 的售价,$F(p_i)$ 表示买方不接受价格 p_i 的概率. 那么,出现事件 e_i 的概率为

$$Pr(e_i) = \lambda_i \times (1 - F_i(p_i)) \tag{1}$$

有限决策集合 D:系统可能做出的所有决策构成决策集合 D. 在某个决策时间点 t,如果发生事件 e_i,那么,智能代理根据目前的库存状态 s 做出的决策 $d(d \in D)$ 由两项子决策 d' 和 d'' 组成,即 $d = \{d', d''\}$,它们分别为

1) d':从商品集合中挑选出商品 $j(j \neq i)$ 与商品 i 捆绑为一个消费套餐;即 $d' = (i, j)$;

2) d'': 给出该消费套餐的折扣率 $\alpha_{ij} (0 < \alpha_{ij} \leqslant 1)$,即 $d'' = \alpha_{ij}$. 令消费套餐 $\{i, j\}$ 的价格为 p_{ij},那么 $\alpha_{ij} = \dfrac{p_{ij}}{p_i + p_j}$ 为了保证决策集合的离散性,本文规定 α_{ij} 只能取离散数值集合中的一个值,即 $\alpha_{ij} \in \Omega, \Omega = \{0.95, 0.9, 0.85, 0.8, \cdots, 0.1\}$.

系统状态转移的概率矩阵 $\boldsymbol{P}$:令 s_t, s_{t+1} 为时间点 t 和 $t+1$ 的两个系统状态. 设在线零售商经营有 m 种商品,用向量 $\boldsymbol{a}_i$ 表示商品 i 被出售给买方,令 $\boldsymbol{a}_i$ 为 m 元向量,$(a_i)_k$ 为向量 $\boldsymbol{a}_i$ 的第 k 个元素,且规定

$$(a_i)_K = \begin{cases} 0 & k \neq i \\ 1 & k = i \end{cases} \tag{2}$$

在时间点 t,系统状态可能发生如下转移:

1) 当出现事件 e_i 且采取决策 $d = \{(i, j), \alpha_{ij}\}$ 时,如果买方拒绝消费套餐 (i, j),那么,买方将仅仅购买商品 i,系统状态 s_t 转变为 $s_{t+1} = s_t - a_i$. 令 $F(p_{ij})$ 表示买方拒绝消费套餐 (i, j) 报价的概率,Pr_1 表示采取决策 d 时系统从状态 s_t 转变为 $s_{t+1} = s_t - a_i$ 的概率,则有

$$\begin{aligned} Pr_1 &= Pr(s_t, d, s_{t+1}) \\ &= Pr(s_t, d, s_{t+1} \mid e_i) \times Pr(e_i) \\ &= F(p_{ij}) \times (\lambda_i \times (1 - F(p_i))) \end{aligned} \tag{3}$$

2) 当出现事件 e_i 且采取决策 $d = \{(i, j), \alpha_{ij}\}$ 时,如果买方接受消费套餐 (i, j) 的报价,那么,系统状态 s_t 转变为 $s_{t+1} = s_t - a_i - a_j$,令 Pr_2 表示其状态转移概率,则有

$$\begin{aligned} Pr_2 &= Pr(s_t, d, s_{t+1}) \\ &= Pr(s_t, d, s_{t+1} \mid e_i) \times Pr(e_i) \end{aligned}$$

$$= (1 - F(p_{ij})) \times (\lambda_i \times (1 - F(p_i))) \tag{4}$$

3）没有任何事件发生，或没有任何买方到达时，系统状态不发生任何变化，即 $s_{t+1} = s_t$，令 Pr_3 表示这种情况出现的概率，则有

$$Pr_3 = Pr(s_t, s_{t+1} \mid s_{t+1} = s_t)$$

$$= 1 - \sum_{i=1}^{m} (\lambda_i \times (1 - F(p_i))) \tag{5}$$

一次状态转移的报酬 f：令 c_i 表示商品 i 的成本，p_i 表示商品 i 的价格，那么，当买方仅仅购买了一件商品 i 时，零售商的收益记为 $f_1 = p_i - c_i$. 如果买方接受消费套餐：(i, j)，那么，零售商的收益记为 $f_2 = p_{ij} - c_i - c_j = \alpha_{ij} \times (p_i + p_j) - c_i - c_j$.

Bellman 递推方程：令 $V_t(s_t)$ 表示在状态 s_t 条件下，系统从时间点 t 开始到整个销售周期结束时的累积收益. 令 C_t 表示在时间点 t 的商品集合，如果某商品 $i \in C_t$，那么，商品 i 至少有一件库存. 随着销售活动的进行，有些商品的库存会耗尽，这时就将该商品从集合 C_t 中删除. 交叉销售策略下收益函数的 Bellman 递推方程可以描述为

$$V_t(s_t) = \sum_{i \in C_t} \max_{j \neq i, j \in C_t} \left(\max_{\alpha_{ij} \in \Omega} (Pr_1 \times (f_1 + V_{t+1}(s_t - a_i)) + Pr_2 \times (f_2 + V_{t+1}(s_t - a_i - a_j)))) + Pr_3 \times V_{t+1}(s_t) \right) \tag{6}$$

在上述模型中，对概率 λ_i，$F_i(p_i)$ 和 $F_{ij}(p_{ij})$ 的估计是一项重要的研究课题. 文献[10,11] 研究了如何利用数据挖掘技术从历史交易数据中获取这些参数.

2　基于聚类分析确定交叉销售的商品集

从动态交叉销售的 MDP 模型可以看出，系统的状态空间十分巨大. 事实上，许多商品之间根本没有交叉销售的可能，可以通过确定适合交叉销售的商品集合来降低问题中状态空间的规模. 利用数据挖掘技术可以从历史交易数据中发现哪些商品有可能被客户同时购买，那么，这些商品就可以聚集在同一个簇中进行交叉销售. 本文利用文献[12] 提出的 ROCK 算法对整个商品集合进行聚类. 在传统的聚类算法中，优化的标准是使每一个点到簇中心点的欧几里德距离之和最小，与此不同，ROCK 算法是以交易记录之间联结（link）量作为簇划分的标准而非采用距离，这与本文问题的要求十分一致. 算法的逻辑大体如下所述：

令数据库中的某个交易记录 T_i 记载了该交易包含的商品项，给定阈值 $0 < \theta \leqslant 1$，如果两个交易 T_i，T_j 满足下面的公式，那么 T_i，T_j 互为邻居.

$$\mathrm{sim}(T_i, T_j) = \frac{\mathrm{card}(T_i \cap T_j)}{\mathrm{card}(T_i \cup T_j)} > \theta \tag{7}$$

公式中 $\mathrm{sim}(T_i, T_j)$ 表示 T_i 与 T_j 之间的相似度，函数 card( ) 表示求集合中商品项的个数.

联结（link）量是指两个交易记录之间共同邻居的数量，如果 $\mathrm{link}(T_i, T_j)$ 越大，那么 T_i，T_j 在同一个簇中的可能性就越高. 由于 ROCK 算法的优化标准是使同一个簇中的交易之间的 link 尽可能大，因此，同一个簇中的商品项具有较高的关联度，适合交叉销售. 本文不给出 ROCK 算法的详细过程，读者可参考文献[12]. 利用 ROCK 算法可以获得规模很小的适合交叉销售的商品集合.

3　知识驱动型 Q- 学习算法

尽管利用 ROCK 算法可以找出一个规模不大的彼此间关联密切的商品集合进行交叉销售，但是在这个商品集合中，有些商品之间的关联度大，有些个别商品之间的关联度很小，甚至是负关联的. 利用数据挖掘技术可以从历史数据中发现商品之间的关联关系，本文利用这些知识来指导 Q- 学习过程，从而提高学习的效率.

3.1　Q- 学习基本原理

传统的 Q- 学习的训练结果是由 $(s, d, Q(s, d))$ 三元组组成的 LOOK UP 表. 在实际应用中，当出现某状态 s 时，系统只需要查询 LOOK UP 表，选择 $Q(s,d)$ 最大的决策. Q- 学习方法的实现如下：

设有普通型 MDP：$\{S, D, P, f\}$，智能系统在每个时间步 t，观察当前状态 s_t，选择和执行决策 d，环境接受该决策后发生一次状态转移 $s_t \rightarrow s_{t+1}(s_t, s_{t+1} \in S)$，并接收到即时报酬 $f(s_t, d, s_{t+1})$，然后根据下式调整 $Q(s,d)$ 值

$$Q(s_t, d) = (1 - \beta) \times Q(s_t, d) + \beta \times$$

$$(f(s_t,d,s_{t+1})+\gamma\times V(s_{t+1})) \quad (8)$$

式中 $V(s_{t+1})=\max_{b\in D}(Q(s_{t+1},b))$，$\beta$ $(0<\beta\leqslant 1)$ 为学习率参数，$\gamma(0\leqslant\gamma\leqslant 1)$ 为收益的时间折扣率，γ 越大表示系统越看重远期收益，γ 越小表示系统越看重近期收益[7]. 经过多次迭代计算，$Q(s,d)$ 将趋于收敛.

3.2 关联规则

利用数据挖掘技术可以从交易数据中挖掘出如下两类关联规则[13,14]：

1）正关联规则：正关联规则是型如：$X\Rightarrow Y$ ($c\%$ ，$s\%$) 的蕴涵式，规则中 X ，Y 为商品项集 ，且 $X\cap Y=\varnothing$，$c\%$ 表示规则的信任度，是指在购买了商品集 X 的交易中有 $c\%$ 的交易还包含商品集 Y，$s\%$ 表示规则的支持度，即总交易数中有 $s\%$ 的交易同时购买了商品集 X 和 Y 的商品. 如果 X 和 Y 中分别只有一个商品项 i 和 j，那么，这样的关联规则被称为简单正关联规则，记为 $i\Rightarrow j$ ($c\%$，$s\%$)，其含义是：如果买方购买商品 i，他还很有可能买商品 j.

2）负关联规则：型为 $X\Rightarrow\neg\ Y$ ($c\%$，$s\%$) 是负关联规则的一种常见形式，它反映的是如果买方购买某些商品后，他很有可能不再买另一些商品. 如果规则 $X\Rightarrow\neg\ Y$ 中 X 和 Y 分别只包含一个商品项，那么，这样的负关联规则被称为简单负关联规则，记为 $i\Rightarrow\neg\ j$ ($c\%$，$s\%$).

令 $c_d\%$ 与 $s_d\%$ 分别表示信任度与支持度的阈值，如果某规则的信任度 $c\%$ 和支持度 $s\%$ 分别大于 $c_d\%$ 和 $s_d\%$，那么该规则成立.

3.3 知识驱动的 Q-学习算法

利用 Q-学习算法求解动态交叉销售问题时，其训练结果是由 $(s,e,d,Q(s,e,d))$ 四元组组成的 LOOK-UP 表. 在进行动态交叉销售决策时，智能代理只需浏览该表，选择当前库存状态 s 和事件 e 所对应的一个 $Q(s,e,d)$ 值最大的决策 d，每项决策由两项子决策 d' 和 d'' 组成. 子决策 $d'=(i,j)$ 表示系统挑选出商品 $j(j\neq i)$ 与事件 e 中的商品 i 捆绑为一个消费套餐，子决策 $d''=\alpha_{ij}$，表示系统给该消费套餐的折扣率为 α_{ij}.

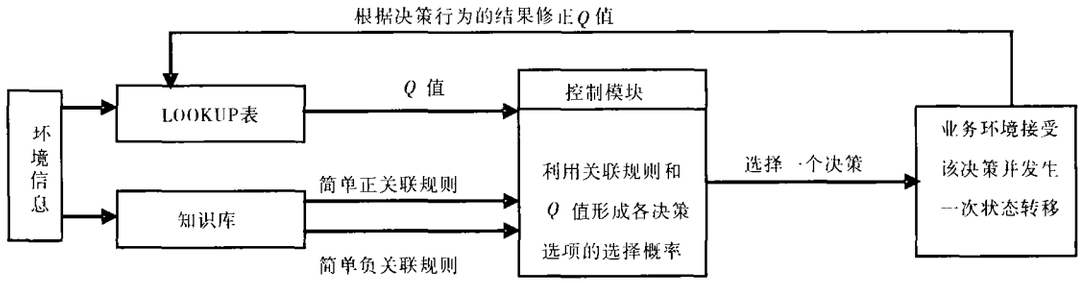

图1 Q-学习过程

Fig. 1 Process of Q-learning

本文将关联规则植入 Q-学习过程，在每次迭代中，不是盲目地选择决策，而是在关联规则指导下，以较高的概率选择比较优良的决策，其计算过程如图 1 所示. 在图 - 1 中，业务环境信息包括目前的库存状态 s_t，以及发生的事件 e_i. 系统将 LOOKUP 表中与环境信息相关的各种决策的 Q 值以及规则前提为商品 i 的简单关联规则提交给控制模块. 控制模块按下述方法调整各种决策的 Q 值效用度，根据各种决策的 Q 值效用度来形成各种决策的选择概率.

为了定量衡量某一决策的 Q 值对选择决策时的指导效用，本文用 $\mu(s_t,e_i,d_t)$ 表示决策的 Q 值效用度，其计算方法如公式(9) 所示，公式中 $H=\{d\mid d\in D,s=s_t,e=e_i\}$.

$$\mu(s_t,e_i,d_t)=\frac{Q(s_t,e_i,d_t)-\min_{d\in H}(Q(s_t,e_i,d))}{\max_{d\in H}(Q(s_t,e_i,d))-\min_{d\in H}(Q(s_t,e_i,d))} \quad (9)$$

设有简单正关联规则为 $i\Rightarrow j$，为了与负关联规则的信任度相区别，用 $c_{ij}^{+}\%$ 表示正关联规则的信任度，负关联规则的信任度记为 $c_{ij}^{-}\%$. 信任度表示规则的正确率，或规则的效用. 公式(10) 是利用正关联规则对决策的 Q 值效用度的修正.

$$\mu(s_t,e_i,d_t)=\frac{1}{1-c_{ij}^{+}\%}\times\mu(s_t,e_i,d_t) \quad (10)$$

如果有负关联规则 $i\Rightarrow\neg\ j$，该规则的信任度为 $c_{ij}^{-}\%$. 公式(11) 是利用该关联规则对 Q 值效用

度的修正.

$$\mu(s_t,e_i,d_t)=(1-c_{ij}^{-}\%)\times\mu(s_t,e_i,d_t) \quad (11)$$

为论述方便，将公式(10)和(11)合并为公式(12). 在公式(12)中如果不存在正关联规则 $i\Rightarrow j$，那么 $c_{ij}^{+}\%=0$，同理，如果不存在负关联规则 $i\Rightarrow\neg\ j$，那么 $c_{ij}^{-}\%=0$.

$$\mu(s_t,e_i,d_t)=\frac{1}{1-c_{ij}^{+}\%}\times\mu(s_t,e_i,d_t)\times(1-c_{ij}^{-}\%),\forall d_t \text{ where} \quad (12)$$

$$d_t=(d_t^{'},d_t^{''}) \text{ and } d_t^{'}=(i,j)$$

从公式(12)中可以发现，如果正关联规则 $i\Rightarrow j$ 的信任度 $c_{ij}^{+}\%$ 越高，那么，采用商品 j 与商品 i 组成消费套餐的决策的 Q 值效用度就越大；如果某负关联规则 $i\Rightarrow\neg\ j$ 的信任度 $c_{ij}^{-}\%$ 越高，那么，采用商品 j 与商品 i 组成消费套餐的决策的 Q 值效用度就越小.

当 Q-学习进行到一定程度后，Q 值趋于稳定，这时，再利用关联规则对 Q 值效用度进行修正反而会破坏 Q 值的效用度，这是因为关联规则也存在着发生错误的风险，其风险为：$1-c\%$. 为了避免这些知识对训练结果的错误影响，按下面的公式对 Q 值效用度进行修正，公式(13)中，η 为经验值，可以从多次实验中获得.

$$\mu(s_t,e_i,d_t)=\begin{cases}\dfrac{1}{1-c_{ij}^{+}\%}\times\mu(s_t,e_i,d_t)\times(1-c_{ij}^{-}\%), & \text{如果}\quad \mu(s_t,e_i,d_t)<\eta\\ \mu(s_t,e_i,d_t), & \text{否则}\end{cases} \quad (13)$$

在 Q-学习的训练过程中，有多种决策探索方法，但最为常用的是 Boltzmann 提出的 Softmax 分布探索[15]. 根据 Softmax 探索策略，在迭代计算中决策 d_t 被选中的概率为

$$Pr(d_t\mid s_x,e_i)=\frac{e^{\frac{\mu(s_t,e_i,d_t)}{\tau}}}{\sum_{d\in H}e^{\frac{\mu(s_t,e_i,d)}{\tau}}} \quad (14)$$

公式(14)中 τ 为“温度”系数，τ 越小，系统越倾向于选择 Q 值效用度大的决策，τ 越大，系统越趋向于以平均的概率选择各种候选决策. 系统将以公式(14)计算出的概率随机挑选一个决策，一旦一个决策被选中，那么就执行该决策，并根据获得的即时报酬和 LOOK UP 表中的信息，利用下面的公式在 LOOK UP 表中修改该决策的 Q 值.

$$Q(s_t,e_i,d_t)=(1-\beta)\times Q(s_t,e_i,d_t)+\beta\times(f(s_t,d_t,s_{t+1})+\gamma\times V(s_{t+1})) \quad (15)$$

式中 $V(s_{t+1})=\max_{d\in D}(Q(s_{t+1},d))$.

根据上述原理，本文设计了一个知识驱动的 Q-学习算法 K-Q-learning，算法描述如下：

1 随机初始化 LOOK UP 表中各个决策的 Q 值

2 repeat

【

a) 令商品集合 C 为交叉销售所考虑的所有商品

b) 令 $t=1$

c) repeat

【 if 发生某事件 e_i，即买方提出购买某商品 i，且有 $i\in C$

【

(1) 观察当前的库存状态 s，从LOOK UP表中找出状态 s 和事件 e_i 所对应的所有决策，形成决策集合 H，如果 H 中某决策所选择的捆绑商品已经不在商品集合 C 中，那么，将该决策从 H 中删除.

(2) 从知识库中找出所有以商品 i 为前提的简单正关联规则和简单负关联规则

(3) 根据公式(13)，利用关联规则对决策集合H中的各个决策的 Q 值效用度进行修正

(4) 利用公式(14)形成集合 H 中的各个决策被选择的概率

(5) 根据集合 H 中的各个决策被选择的概率随机选择一个决策

(6) 执行该决策，计算该决策的即时收益，并观察新状态

(7) 根据公式(15)在LOOK UP 中修改该决策的 Q 值

(8) 更新库存状态，如果某商品的库存量降为0，那么将该商品从商品集合 C 中删除

】

$t=t+1$

当 $t=N$ 时结束本轮循环 】

d) 当 Q 值收敛或达到规定的循环次数，结束循环 】

4　仿真实验

4.1　实验目的

大量的文献已经证明零售商品之间存在各种

关联关系，关联规则的挖掘技术已经日臻成熟. 本文实验的目的不是验证商品间的关联关系，而是验证 *K-Q-* 学习算法的收敛效率，实验的另一个目的是验证动态交叉销售策略的收益能力. 在实验中，需要对同一个商品集合，假设存在不同数量的关联规则的情况下，比较算法的收敛能力以及动态交叉销售策略的收益能利，从而分析本文提出的方法在什么情况下是适用的.

4.2 实验设计

为了分析 *K-Q*-learning 算法在动态交叉销售中的应用效果，本文设计了一个交易数据生成器，并将预设的关联规则存放在知识库中. 在实验中规定有 5 种商品，每种商品的库存量为 10 件. 在每个时间点 t，交易事件发生器根据 λ_i，$F(p_i)$ 和 $F(p_{ij})$ 等随机参数随机生成一笔交易，同时进行 *K-Q*-learning 算法的一次迭代运算. 在实验过程中，先规定商品之间存在的关联关系，即那些商品之间存在关联、关联的强度等，然后再依据上述参数生成交易数据. 为了比较关联规则数量不同时算法的计算能力，本文规定了不同数量的关联规则，并分别生成不同的交易数据. 针对不同的数据集，算法的计算的效率和动态交叉销售的收益能力分别如图 2a 和图 2b 所示.

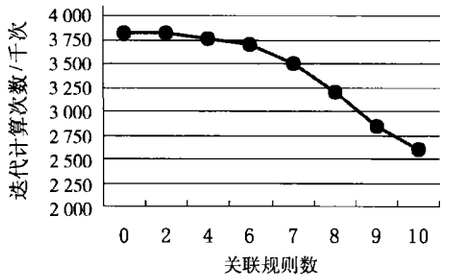

图 2a 计算效率分析

Fig. 2a Analysis on computation efficiency

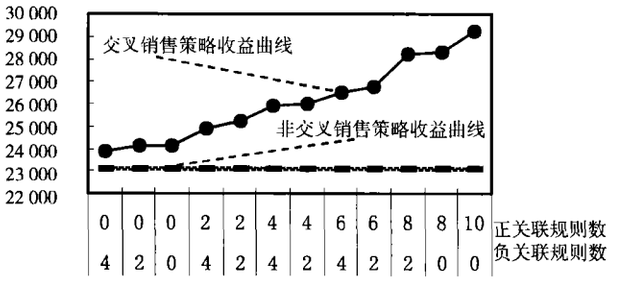

图 2b 收益分析

Fig. 2b Analysis on revenue

为保证所模拟的交易过程体现出商品之间的关联关系，规定：

1）如果存在正关联规则 $i \Rightarrow j$，且规则的信任度为 $c_{ij}^{+}\%$，那么 $0 \leqslant F(p_{ij}) \leqslant 1 - c_{ij}^{+}\%$.

2）如果存在负关联规则 $i \Rightarrow \neg j$，且规则的信任度为 $c_{ij}^{-}\%$，那么 $1 \geqslant F(p_{ij}) \geqslant 1 - c_{ij}^{-}\%$

3）如果商品 i 与商品 j 之间既不存在正关联关系也不存在负关联关系，那么有 $c_d^{-}\% \geqslant F(p_{ij}) \geqslant 1 - c_d^{+}\%$，式中 $c_d^{+}\%$，$c_d^{-}\%$，分别表示正关联规则的信任度阈值和负关联规则的信任度阈值.

4）为避免因为价格折扣率过低（例如低于成本）引起消费者行为理性发生逆转，从而破坏关联规则所体现的购买行为模式，本文规定商品的折扣价格不能低于成本，尽管有时零售商会以低于成本的价格出售商品，但这种策略仅仅适用于特殊的紧急情况，不是本文考虑的范围.

4.3 实验结果分析

从图 2a 中可以发现，随着规则数量的提高，算法的计算效率明显提高. 图 2b 对比了正关联规则和负关联规则数量不同时动态交叉销售的收益与不实行交叉销售策略的收益，从图 2b 会发现，正关联规则越多，动态交叉销售的收益越明显，在只有负关联规则而没有正关联规则的特殊情况下，动态交叉销售的收益与不实行动态交叉销售时的收益几乎相同. 这一仿真试验的结果说明，动态交叉销售并非适用于任何商品集合，而是适用于那些彼此间存在正关联关系的商品集合.

5 结 论

本文提出用 *Q-* 学习方法来解决电子商务环境下动态交叉销售中实时决策问题，针对动态交叉销售的马氏决策模型中的巨量状态空间问题，本文提出用关联规则理论来指导 *Q-* 学习的训练过程，从而保证获得稳定收敛的 LOOK UP 表. 实验结果表明，本文提出的方法适用于那些商品间存在广泛的关联关系的商品集合.

参 考 文 献：

[1] Kamakura W A, Ramaswami S N, Srivastava R K. Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services[J]. International Journal of Research in Marketing, 1991, 4(8): 329—349.

[2] Netessine S, Savin S, Xiao W. Revenue management through dynamic cross selling in e-commerce retailing[J]. Operation Research, 2006, 54(5): 893—913.

[3] 汪　涛，崔　楠. 国外交叉销售研究综述[J]. 外国经济与管理，2005，27(4)：43—50
Wang Tao, Cui Nan. Rerview of foreign research on cross-selling[J]. Foreign Economies & Management, 2005, 27(4): 43—50. (in Chinese)

[4] Lerzan O E, Zeynep A O. Revenue Management through Dynamic Cross-selling in Call Centers[C]. Proceedings of the 4th Annual INFORMS Revenue Management and Pricing Section Conference, MIT, Cambridge, 2004.

[5] Baldwin M. Dynamic packaging[J]. Travel Agent, 2005, 321(5): 76—78

[6] 王利存，郑应平. 一类事件驱动马氏决策过程的 Q- 学习[J]. 系统工程与电子技术，2001，23(4)：80—83.
Wang Licun, Zheng Yingping. Q-learning for a class of event driven Markov decision processes[J]. Systems Engineering and Electronics, 2001, 23(4): 80—83. (in Chinese)

[7] Watkins C, Dayan P. Q-learning[J]. Machine Learning, 1992, 8(3～4): 279—292.

[8] Eyal E D, Yishay M. Learning Rates for Q-learning[J]. Machine Learning Research, 2003, 5(Dec): 1—25.

[9] Gosavi A. Boundedness of iterates in Q-learning[J]. Systems & Control Letters, 2006, 55(4): 347—349.

[10] Moe W W, Fader P S. Dynamic conversion behavior at e-commerce sites[J]. Management Science, 2004, 50(3): 326—335.

[11] Euiho S, Seungjae L, Hyunseok H. A prediction model for the purchase probability of anonymous customers to support real time web marketing: A case study[J]. Expert Systems with Applications, 2004, 27(2): 245—255.

[12] Sudipto G, Rajeev R, Kyuseok S. ROCK: A Robust Clustering Algorithm for Categorical Attributes[C], In Proceedings. of Fifteenth International Conference on Data Engineering, Sydney, Australia, 1999. 512—521.

[13] 董祥军，王淑静，宋瀚涛，等. 负关联规则的研究[J]. 北京理工大学学报，2004，24(11)：978—981.
Dong Xiangjun, Wang Shujing, Song Hantao, *et al*. Study on negative association rules[J]. Transactions of Beijing Institute of Technology, 2004, 24(11): 978—981. (in Chinese)

[14] Chen M S, Han J W, Yu P S. Data mining: An overview from a database perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866—883.

[15] 蒋国飞，吴沧浦. Q 学习算法在库存控制中的应用[J]. 自动化学报，1999，25(2)：236—241.
Jian Guofei, Wu Cangpu. inventory control using Q-learning[J]. ACTA AUTOMATICA SINICA, 1999, 25(2): 236—241. (in Chinese)

Q-learning-based dynamic cross-selling approach in e-commerce

CHENG Yan

School of Business, East China University of Science and Technology, Shanghai 200237, China

Abstract: Dynamic cross-selling is a novel marketing technique in electronic commerce setting. Given association rules and the level of inventory, two issues are analyzed: (1) how to select packaging complements and (2) how to price product packages to maximize profits. First the cross-selling problem was formulated as a event-driven markov decision process model, then a knowledge-driven *Q*-learning algorithm based on association rule, *K*-*Q*-learning, was proposed, which has high efficiency and effectiveness in solving the dynamic cross-selling problem.

Key words: electronic commerce; cross-selling; *Q*-learning; association rule