基于CTBN的移动对象不确定轨迹预测算法

乔少杰[1]，彭 京[2]，李天瑞[1]，朱 焱[1]，刘良旭[3]

(1. 西南交通大学信息科学与技术学院 成都 610031; 2. 成都市公安科学技术研究所 成都 610017;
3. 宁波工程学院电子与信息工程学院 浙江 宁波 315016)

【摘要】为了高效准确地预测移动对象动态运动轨迹，提出了一种基于轨迹时间连续贝叶斯网络(CTBN)的不确定性轨迹预测算法，充分考虑了移动速度和方向对移动对象动态运动行为的影响，包含3个主要步骤：热点区域挖掘将轨迹数据集划分为不同的热点聚簇；轨迹时间连续贝叶斯网络的构建，其由3个变量(街区号、移动速度、移动方向)构成的状态组合；利用该网络预测移动对象动态运动行为计算可能运动轨迹。不同数据集上的实验结果表明该算法的预测精度优于朴素预测算法，并证明了热点区域挖掘的作用在于能够在保证较高预测准确性的前提下提高预测时间性能近60%。

关 键 词 热点区域； 移动对象数据库； 轨迹时间连续贝叶斯网络； 轨迹预测； 不确定性

中图分类号 TP311.13 **文献标识码** A doi:10.3969/j.issn.1001-0548.2012.05.022

Uncertain Trajectory Prediction of Moving Objects Based on CTBN

QIAO Shao-jie[1], PENG Jing[2], LI Tian-rui[1], ZHU Yan[1], and LIU Liang-xu[3]

(1. School of Information Science and Technology, Southwest Jiaotong University Chengdu 610031;
2. Department of Science and Technology, Chengdu Municipal Public Security Chengdu 610017;
3. School of Electronic and Information Engineering, Ningbo University of Technology Ningbo Zhejiang 315016)

Abstract In order to predict the uncertain trajectories in an efficient and accurate fashion, this paper introduces an uncertain trajectory prediction algorithm based on trajectory continuous time Bayesian networks (CTBN). It contains three essential phases: mining hotspot regions by partitioning trajectories into distinct hotspot clusters; constructing trajectory CTBN which is a states combination of three important variables including street identifier, moving speed, and moving direction; predicting the motion behavior of moving objects in order to obtain possible trajectories. Experimental results demonstrate that the proposed method can accurately predict the possible motion curves of moving objects in different trajectory data sets when compared with the naive prediction algorithm. In addition, experiments verify the essential role of hotspot region mining, which can help save prediction time at about 60% with a guarantee of high prediction accuracy.

Key words hotspot region; moving objects databases; trajectory continuous time Bayesian networks; trajectory prediction; uncertainty

准确预测移动对象轨迹存在4个特殊困难：

1) 不确定性是移动对象的固有属性，由于连续移动或网络延迟，很难实时精确地描述对象真实位置[1]。2) 轨迹预测算法必须保证在不产生大量位置或者距离计算代价的前提下准确地计算对象的位置。3) 预测方法性能不会随移动对象数量的增加剧烈下降。4) 必须保证在有限时间内准确地获得对象的位置[2]。

现有移动对象轨迹预测算法[2-6]取得了较好的预测结果，但是随着移动对象数目的增加，仍然可能存在不足，如计算代价较高，算法不具备可伸缩性。本文主要解决带有时间戳的轨迹预测问题。

1 相关工作

挖掘移动对象不确定性轨迹问题最近得到国内外学者的广泛关注。现有关于轨迹预测的工作主要集中于发现频繁轨迹模式[2,7-8]及轨迹索引和查询[1,3-4]。这些工作通常假设精确的轨迹可以在某一给定时间间隔内获得。然而，这一假设没有考虑移动对象数据库中轨迹数据的不确定性和不完整性。

收稿日期：2010 – 12 – 25； 修回日期：2012 – 01 – 04

基金项目： 国家自然科学基金(61100045，61165013)；中国博士后科学基金特别资助项目(201104697)；教育部人文社会科学研究青年基金(10YJCZH117)；高等学校博士学科点专项科研基金(20110184120008)；中国博士后科学基金(20090461346)；中央高校基本科研业务费专项资金专题研究项目(SWJTU11ZT08)；四川省科技支撑计划(09ZC0516)

作者简介：乔少杰(1981 –)，男，博士，副教授，主要从事移动对象数据库、移动数据管理、数据挖掘等方面的研究.

文献[9]提出了一种两阶段自上而下的挖掘算法STPMiner2，用于从时空数据库中发现周期性运动模式。文献[10]将轨迹地理坐标的每一维描述成以时间为参数均匀分布的随机过程，实验结果表明利用均匀分布描述不确定性轨迹的合理性。上述关于移动对象不确定性管理的方法一般假设对象运动具有线性变化趋势，这一假设限制了其应用范围，因为对象的运动通常是未知并且符合动态变换模式。时间连续贝叶斯网络(CTBN)[11]用于描述符合连续性时间变化、具有有限状态的结构化随机过程，适宜描述不确定性移动对象的轨迹。可以借助CTBN实时高效地近似计算对象连续状态变化的概率，尤其适合预测对象的连续运动轨迹。

2 问题描述

移动对象的位置通常表示在笛卡尔坐标系中，一条运动轨迹被认为由不同时间戳标识的结点构成的一个序列，可以用来描述移动对象的轨迹和实时位置信息[8]，其形式化定义如下。

定义 1(轨迹) 移动对象的轨迹是由一系列三元组构成的序列：

$$S=<(x_1, y_1, t_1),\cdots,(x_i, y_i, t_i),\cdots,(x_n, y_n, t_n)> \quad (1)$$

式中，t_i表示时间戳，$\forall i\in[1,n)$，$t_i<t_{i+1}$；(x_i,y_i)表示移动对象的二维空间坐标。

因为在真实场景中很难精确定位移动对象的位置，本文使用包含在轨迹圆柱中的线段构成的可能运动曲线逼进真实轨迹，具体方法参见文献[1]。

轨迹可以使用以时间为参数均匀分布的马尔科夫随机过程表示。对象在某一时刻位置满足马尔科夫性质。轨迹通常表示在电子地图中，定义如下[6]。

定义 2(电子地图) 电子地图由若干街区构成，是三元组(ID,Polyline,Length)的集合。其中，ID为街区标识符，用数字表示；Polyline是折线，表示从起点到终点内不同街区组成的折线，用两个字母表示；Length表示街区的长度，用实数表示。

移动对象状态的改变不仅受街区的影响，此外受其他两个重要因素影响，即：移动速度和移动方向。这两个因素的作用主要体现在：1) 移动速度的变化可以导致移动对象从某一运动状态转换到其他状态，此外可以影响状态转换概率；2) 移动方向。在地图中方位信息通常利用两对不同的方向表示，即东西和南北。对象通常沿直线移动，在特殊情况下会违反常规运动行为模式选择后退，因此必须考虑移动方向可能产生的影响。

3 轨迹热点区域挖掘

热点区域挖掘的本质是轨迹聚类，其工作原理为：1) 热点区域构建。通过统计历史数据获得电子地图中各街区的访问频率，找出构成轨迹的高频边并对其进行标识。其中，由高频边聚合而成的街区构成一个热点区域。2) 轨迹划分。扫描整个历史数据集，对于每条轨迹选取其落在热点区域内的片段并对其进行存储，算法如下。

算法 1 轨迹热点区域挖掘算法。

输入：电子地图E中边的集合，概率阈值θ。

输出：热点区域集合$O=\{O_1,O_2,\cdots,O_n\}$。

```
1) for (每条边e∈E) do
2)     {e.freq←FreqCal(e);}//统计边e访问频率
3)  对E中所有边按频率降序排列;
4) f←E.size*θ;
5) for (每条边e∈E) do
6)       if (e.freq<f) then
7)               {删除边e;}
8) n←0;
9) for (每条边e∈E) do
10)       {创建一个新的热点区域Ok;
11)        合并热点区域O和Ok;
12)        n++;}
13) while (O中对象个数发生改变)
14)        for (<Oi,Oj>包含在O中) do
15)              if(Oi和Oj之间有边相连) then
16)                   {合并热点区域Oi和Oj;
17)                    删除Oj;}
18)  输出O。
```

算法主要步骤为：

1) 统计E中所有边的访问频率，按降序排列(1～3行)。2) 计算第(E.size*θ)条边的访问频率f(第4行)，并删除E中所有访问频率低于f的边(第5～7行)。3) 对于剩下的所有边建立热点区域，方法如下：① 将每条边标记为一个热点(第8～12行)；② 反复扫描热点集合，当两个热点之间连通时，合并这两个热点，直到热点集合中所有热点数量不再变化为止(第13～17行)。4) 输出所有热点区域的集合O(第18行)。

其中，参数θ的作用在于限定热点区域所包含边的数量占所有边总数的最小比例。

算法复杂性分析：假设一共有N条轨迹，所有轨迹的平均长度为l，街区的总数为m。

步骤1)中，对所有N条平均长度为l的轨迹片段的访问频率进行统计，每次查询并记录街区访问频率的复杂度为$O(\log_2 m)$，所以该步骤复杂度为$O(N*l*\log_2 m)$。对所有街区排序的复杂度为$O(m*\log_2 m)$。因此步骤1)的复杂度为$O(N*l*\log_2 m+ m*\log_2 m)$。步骤2)只需要对排序后的所有街区进行一次不完全遍历，所以复杂度为$O(m)$。步骤3)中，最坏情况下需要对每条街区和其他街区的连通性进行记录，复杂度为$O(m^2)$。因此，算法理论复杂度为$O(N*l*\log_2 m+ m*\log_2 m+m^2)$。由于真实应用中，轨迹中包含边的数量远远大于街区本身的数量，因此算法复杂度为$O(N*l*\log_2 m)$。

4　轨迹时间连续贝叶斯网络

4.1　基本概念

本文提出的模型定义在CTBN基础上，具体描述如下[6]。

定义 3(轨迹时间连续贝叶斯网络—TCTBN) 令三元组的集合$X=\{(x_1,x_2,x_3)|x_1,x_2,x_3$是3个随机变量$\}$，分别表示街区标识符、移动速度和方向。每个变量具有有限的值域Dom(x_i)。TCTBN由两部分构成：1) 关于变量的初始分布ψ_X^0，用关于X的贝叶斯网络N表示。2) 满足如下两个条件的模型：① 由(x_1,x_2,x_3)3个变量作为结点构成的有向图G，Par(x_i)表示x_i的前一状态；② 一个轨迹强度矩阵$\boldsymbol{M}_{x_i|\mathrm{Par}(x_i)}$，其中$x_i\in X$。

下面给出轨迹条件强度矩阵(trajectory conditional intensity matrix，TCIM)的定义[6]。

定义 4(轨迹条件强度矩阵—TCIM) 令L表示由3个随机变量x_1,x_2,x_3(分别表示街区标识符、移动速度、移动方向)构成的组合状态，其值域为$f(L)=\{l_1, l_2,\cdots,l_n\}$。假设$L$满足TCTBN的基本条件要求，$L(t)$表示的状态转换依赖于其前一个状态$L'(t)$的变化，条件强度矩阵定义为：

$$\boldsymbol{M}_{L|L'}=\begin{bmatrix} -p_1^l(L') & p_{12}^l(L') & \cdots & p_{1n}^l(L') \\ p_{21}^l(L') & -p_2^l(L') & \cdots & p_{2n}^l(L') \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1}^l(L') & p_{n2}^l(L') & \cdots & -p_n^l(L') \end{bmatrix} \tag{2}$$

TCIM中的元素$p_{ij}^l(L')$表示从状态l_i变化到状态l_j的概率，定义为$p_{ij}^l(L')=f_{ij}^l(L')/\sum_{i\neq j} f_{ij}^l(L')$。其中，$f_{ij}^l(L')$表示轨迹中经由一条街区由状态$l_i$变化到状态$l_j$的频率，其取值可以通过统计历史数据获得。TCIM中对角线上元素之和$p_i^l(L')=\sum_{i\neq j} p_{ij}^l(L')=1$，表示对象脱离状态$l_i$的概率。该值设置为1是合理的，因为对象往往动态运动，不可能始终保持一种运动状态不变，势必发生状态转换。

4.2　高阶马尔科夫链

移动方向实质上是马尔科夫链中考虑了特殊情况的二阶信息。如果已知对象先前所在街区，可以粗略地估计其运动方向。表1给出了一个关于方向的二阶概率矩阵，其中E、W、S、N分别表示东西南北4个不同的方向。

表1　二阶概率矩阵

方向	W	E	N	S
WE	0.5	0.1	0.2	0.2
WW	0.1	0.4	0.2	0.3
WS	0.1	0.1	0.7	0.1
WN	0.2	0.15	0.35	0.4
EE	0.45	0.05	0.25	0.25
EW	0	0.7	0.15	0.15
ES	0.15	0.15	0.7	0
EN	0.15	0.15	0.1	0.6

以表1第3行元素为例，先前和当前状态用“WS”表示，对象具有70%的概率向北移动，具有10%的概率向南移动。对于正西和正东方向，其具有相等的状态转换概率，即10%。可以直观地发现每行所有概率之和为1。

5　轨迹预测

轨迹预测算法如下[6]。

算法 2　基于TCTBN的轨迹预测算法。

输入：轨迹条件强度矩阵$\boldsymbol{M}$，对象初始状态s_0，预测概率阈值ε，轨迹圆柱中圆盘半径r。

输出：可能轨迹集合$T=\{T_1,T_2,\cdots,T_n\}$。

1) $N\leftarrow\varnothing$；$C_0\leftarrow\varnothing$；

2) C.put(s_0)；C_0.prob=1；N.put(C_0)；

3) for each $C_p\in N$ do

4) {$i\leftarrow C_p$.laststate();}//定位到C_p最后一个状态

5) for(each state j in the ith row of $\boldsymbol{M}$ & $M(i,j)\neq 0$)

6) if (C_p.prob$*M(i,j)\geqslant\varepsilon$) then

7) {$C'_p = C_p$; C'_p.put(j);

8) C'_p.prob$\leftarrow C'_p$.prob$*M(i,j)$;

9) N.put(C'_p);}

10) for each $C_p\in N$ do

11) {T_p=CalTraj(C_p); //计算C_p对应的轨迹

12) for each state s in C_p do

13) {$\tau\leftarrow$CalTime(t_p); //计算经过状态s对应

轨迹点t_p的时间

14) T_p.put(t_s,t_e,τ);}

15) Output(T_p);} //输出T_p

轨迹预测算法包含以下主要步骤：

1) 创建一个空的轨迹序列集合N和一个空状态链C_0(第1行)，然后将初始状态s_0加入C_0中并将C_0加入N中(第2行)。对于N中的每一状态链，首先找到C_p的最后一个状态(第3～4行)。2) 对于$\boldsymbol{M}$中第i行的每个状态j，如果$M(i,j)\neq 0$并且当前的转换状态乘以先前的转换概率乘积大于事先指定的预测概率阈值ε，则将j加入到轨迹序列中，更新其转换概率值(第5～8行)，将经过扩展的状态链C'_p加入到N中，同时删除原状态链C_p(第9行)。3) 对于N中的每个状态链生成相应的轨迹(第10～11行)。利用运动学公式计算经过一条街区起点t_s和终点t_e的时间间隔τ并与端点信息一同添加到结果轨迹数据中(第13～14行)，输出轨迹信息(第15行)。

通过分析可以知道算法2的时间复杂度和空间复杂度分别为$O(m*n)$和$O(n)$，其中m表示状态链的数目，n表示每条状态链上状态的个数。

6 实 验

6.1 实验环境及数据集

本节通过比较朴素轨迹预测算法(简称Naive)、PutMode[6]和TPMO算法，证明TPMO的性能优势。TPMO在进行轨迹聚类时利用热点区域挖掘算法将轨迹集合划分为不同聚簇；PutMode借助CTBN网络进行轨迹预测并采用改进的DBSCAN算法对轨迹聚类，但是没有考虑轨迹热点区域的挖掘；朴素算法在构建TCTBN和轨迹预测时没有考虑速度和方向的影响，状态转换概率仅由不同街区访问频率决定。

实验硬件平台为：AMD Athlon 5000+，2.6 GHz CPU，2 GB内存。实验数据集利用Brinkhoff提出的基于交通网络的轨迹生成器生成，具体描述如下：Oldenburg数据集(用Oldenburg表示)包含由6 105个结点和7 035条边组成的约1万条轨迹；New York数据集(用NY表示)，该数据集由1 448个结点和1 573条边组成，地图信息来源于美国地图局。

6.2 参数调节

轨迹预测中需调节轨迹圆柱中圆盘半径r和预测概率阈值ε这两个参数。r用于确定一条轨迹是否为可能运动曲线，可以影响预测结果准确性。实验中首先指定一个足够大的预测时间值t(选取的原则是在t时间内可以预测大多数轨迹)，通过渐进的方式增加r值以期获得足够大的预测精度。借助τ-约束[8]的概念，本文采用的预测精度度量标准定义为：

$$\text{Accuracy} = n / N \tag{3}$$

式中，n表示轨迹预测命中的次数；N表示预测总次数。其中“命中”的含义参见文献[6]。

本节实验数据集具体参数设置如表2所示。

表2 参数调节实验设置

参数	Oldenburg	NY
地图宽度	23 572	563 287
地图长度	26 915	435 186
移动对象数量	10 005	15 005
ε的取值	0.02	0.02
时间间隔t	7	12
θ的范围	0.1~0.7	0.1~0.7
r的范围	200~1 200	1 600~3 600

以Oldenburg数据集为例，先设置r为200个像素，在23 572×26 915的地图上，以100个像素的间隔逐渐增大到1 200个像素。θ表示建立热点区域的概率阈值，它的取值决定了热点区域的数量及规模。实验观察θ值从0.1变化到0.7的预测结果。随着状态数目的增加，一条状态链上的状态转换概率乘积会急剧下降，因此需要选取一个较小的预测概率阈值ε。大量实验表明ε=0.02可确保所预测轨迹的长度。其中，t设置为7个时间单元(对Oldenburg)，12个时间单元(对NY)，这两个时间间隔足以获得多数完整轨迹。

6.3 预测精度分析

本节实验使用预测精度评价算法性能，以Oldenburg数据集为例，结果如图1所示。其中，x轴表示移动对象初始状态数量，y轴表示预测精度。图1表明TPMO在所有实验中预测精度略优于PutMode算法，平均高出2.6%；对于NY数据，平均高于4.9%。原因在于TPMO算法中热点区域建立相比PutMode具有优势，因为它能够通过挖掘频繁轨迹选取具有普遍运动规律的轨迹。实验表明TPMO在Oldenburg和NY数据集上的平均预测精度分别为79%和74.3%，这一结果经领域专家证实是比较满意的。

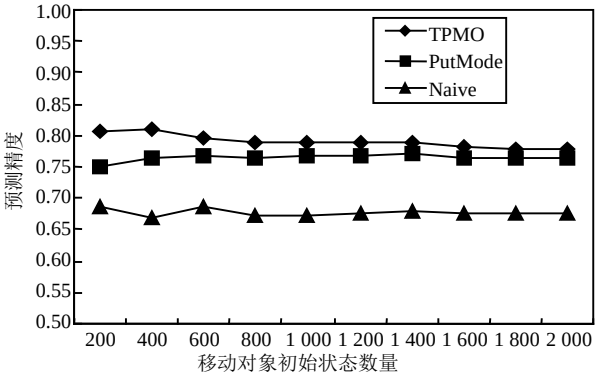

图1 Oldenburg数据集上算法预测精度比较

可以发现TPMO和PutMode的预测精度明显高于朴素预测算法Naive。因为它们在TCTBN构建和轨迹预测过程中考虑了更加复杂的因素，即移动速度和方向。此外，Naive仅考虑一阶概率强度矩阵，而TPMO和PutMode考虑了更为复杂的n阶强度矩阵。实验表明在Oldenburg数据集上TPMO预测精度平均优于Naive算法11.3%，在NY数据集上高出15.7%。

6.4　预测时间对比

本节将比较随着移动对象初始状态数量的增加3种轨迹预测算法的时间性能。

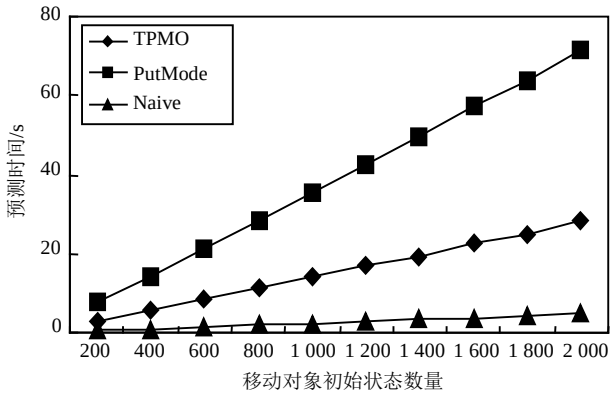

图2　NY数据集上算法预测时间比较

Naive和PutMode算法没有采用轨迹热点区域挖掘算法，因此对历史数据集不敏感，而TPMO算法可以过滤掉不频繁轨迹产生的影响，节省预测时间。以NY数据集为例，通过图2可以发现：TPMO的预测时间均低于PutMode，对于NY数据集平均减少61.5%，在Oldenburg数据集上的预测时间平均缩减62.5%。TPMO对历史数据具有一定敏感性，轨迹数据分布越集中，热点区域挖掘算法的效果越明显，TPMO相对于PutMode预测的优势就越大。但是，Naive算法在所有情况下均是最优的。因为其仅使用从一个街区到另外一个街区的变化表示一次状态转换。对于TPMO，其借助3个随机变量决定一次状态转换过程，并且其候选状态数目多于Naive算法。

7　结　论

本文提出了一种兼顾有效性及高效性的轨迹预测方法，其主要特点为：1) 热点区域的建立用于去除异常轨迹并对轨迹进行聚类处理；2) 利用TCTBN网络刻画不确定性轨迹。此外，利用轨迹生成器对真实地图生成的数据进行实验，证明了TPMO在保证时间性能前提下可以实现准确预测。

参　考　文　献

[1] TRAJCEVSKI G, WOLFSON O, HINRICHS K, et al. Managing uncertainty in moving objects databases[J]. ACM Transactions on Database Systems, 2004, 29(3): 463-507.

[2] MORZY M. Mining frequent trajectories of moving objects for location prediction[C]//Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition. Berling: Springer, LNCS 4571, 2007: 667-680.

[3] TRAJCEVSKI G, TAMASSIA R, DING H, et al. Continuous probabilistic nearest-neighbor queries for uncertain trajectories [C]//Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. New York: ACM, 2009: 874-885.

[4] 丁治明, 李肖南, 余波. 网络受限移动对象过去、现在及将来位置的索[J]. 软件学报, 2009, 20(12): 3193-3204.
DING Zhi-ming, LI Xiao-nan, YU Bo. Indexing the historical, current, and future locations of network-constrained moving objects[J]. Journal of Software, 2009, 20(12): 3193-3204.

[5] 郭黎敏, 丁治明, 胡泽林, 等. 基于路网的不确定性轨迹预测[J]. 计算机研究与发展, 2010, 47(1): 104-112.
GUO Li-ming, DING Zhi-ming, HU Ze-lin, et al. Uncertain path prediction of moving objects on road networks[J]. Journal of Computer Research and Development, 2010, 47(1): 104-112.

[6] QIAO Shao-jie, TANG Chang-jie, JIN Hui-dong, et al. PutMode: Prediction of uncertain trajectories in moving objects databases[J]. Applied Intelligence, 2010, 33(3): 370-386.

[7] GIANNOTTI F, NANNI M, PEDRESCHI D. Efficient mining of temporally annotated sequences[C]//Proceedings of the 6th SIAM International Conference on Data Mining. Philadelphia: SIAM, 2006: 346-357.

[8] GIANNOTTI F, NANNI M, PINELLI F, et al. Trajectory pattern mining[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007: 330-339.

[9] MAMOULIS N, CAO H, KOLLIOS G, et al. Mining, indexing, and querying historical spatiotemporal data[C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 236-245.

[10] MOKHTAR H M O, SU J. Universal trajectory queries for moving object databases[C]//Proceedings of the 2004 IEEE International Conference on Mobile Data Management. Washington: IEEE Computer Society, 2004: 133-144.

[11] NODELMAN U, SHELTON C R, KOLLER D. Learning continuous time Bayesian networks[C]//Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 2003: 451-458.

编　辑　漆　蓉