普适环境下陌生交互实体的行为分析与预测 *

刘广宾, 满君丰, 李长云, 温向兵
(湖南工业大学,湖南 株洲 412008)

摘　要: 基于普适环境下陌生交互实体之间没有先验交互经历,为了该环境的计算安全,必须确保其交互的行为可信,提出一种基于嵌套的狄利克雷过程和无限隐马尔可夫模型的行为分析与态势预测模型。该模型能有效地预防陌生实体交互过程中反常、欺诈等行为,对不可信的行为提前干预,避免产生严重的后果。最后举出模型在智能商场中应用并仿真分析了模型的可行性。

关键词: 普适环境; 嵌套的狄利克雷过程; 无限隐马尔可夫模型; 转移概率矩阵
中图分类号: TP393; TP31　　**文献标志码**: A　　**文章编号**: 1001-3695(2009)12-4562-03
doi:10.3969/j.issn.1001-3695.2009.12.045

Recognizing and predicting of strange interactive entity's behavior in pervasive environment

LIU Guang-bin, MAN Jun-feng, LI Chang-yun, WEN Xiang-bing
(*Hunan University of Technology*, *Zhuzhou Hunan* 412008, *China*)

Abstract: There is no priori interactive experience between the strange interactive entities in the pervasive environment, so must make sure that the interactive behavior between the two entities is reliable for the computing safety in this environment. Based on this, this paper proposed a behavior-analyzing and situation-predicting model based on the nested Dirichlet process and the infinite hidden Markov model. It could predict the behavior of abnormality and cheat effectively in the entity's strange interactive process and intervene the unreliable behaviors in advance so that serious consequences could be avoided. Took some examples about the model's application in the intelligent market and analysed the feasibility of model by simulated.

Key words: pervasive environment; nested Dirichlet process; infinite hidden Markov model; probability matrix of transition

0　引言

普适计算的思想最早是 1991 年 Mark Weiser 在《Scientific American》的"The Computer for the 21st Century"中提出的,强调把计算机嵌入到环境或日常工具中去,让计算机本身从人们的视线中消失,让人们注意的中心回归到要完成任务的本身。他的思想在 20 世纪 90 年代后期开始在国际上得到广泛关注和接受,许多相关的研究计划纷纷启动,目前已经成为一个极具活力和影响力的研究领域。关于普适计算的准确定义,不同的研究者有不同的理解,其目标都是"要建立一个充满计算和通信能力的环境,同时使这个环境与人们逐渐地融合在一起"。清华大学徐光祐教授给出的定义是:普适计算是信息空间与物理空间的融合,在这个融合的空间中人们可以随时随地、透明地获得数字化服务[1,2]。

普适计算环境的特点是:a)环境对用户不熟悉,用户没有与环境的拥有者存在信任关系;b)数据常常动态地产生;c)用户的访问权限动态变化。因此双方必须要有足够级的信任。然而在开放、动态的普适计算环境中,仅凭身份可信来判断交互实体的可信性是不够的,即存在用户的身份是可信的,但用户的行为不一定可信的情况。

为了普适计算环境的安全可靠,实体的行为分析与态势预测是普适计算中关键的一个环节。在陌生实体交互的过程中,对交互行为及其效应进行动态分析,并根据分析结果对后续行为发展态势进行主动预测,解决这个环节将是普适计算软件发展过程中重要的突破。这样它能有效地预防实体交互过程中反常、欺诈等行为,对不可信的行为提前干预,避免产生严重的后果。例如智能会议室服务系统、智能家居、智能商场等,如能有效实现交互行为的主动预测和提前风险干预,可以有效提高系统的可靠性和安全性,降低采取各种安全措施的费用。

本文例举出了一个智能商场购物情景,通过仿真实验证实:本文行为分析与态势预测模型能有效分析普适计算环境下陌生实体交互行为,并对行为的后续发展态势进行有效的预测。这样可以保证该环境下实体之间正常交互。

1　相关工作

在陌生实体交互的行为预测方面,国内外研究者作了一些研究,文献[3]中利用 Beta 概率密度函数更新建立贝叶斯声誉系统,能过滤不公平评估和对不公平的评估进行折扣可以提高模型可信预测的性能。文献[4]提出了合格声誉,一种克服了当前信任模型基于自身的直接交互经验和第三方见证实体报

收稿日期: 2009-04-21; **修回日期**: 2009-06-04　　**基金项目**: 国家自然科学基金资助项目(60773110);中国博士后科学基金资助项目(20080440216);湖南省自然科学基金资助项目(09JJ6087);湖南省教育厅科研资助项目(08C286)

作者简介:刘广宾(1983-),男,湖南娄底人,硕士研究生,主要研究方向为普适计算、可信软件(gb_liu24@tom.com);满君丰(1976-),男,黑龙江海伦人,副教授,主要研究方向为可信软件、普适计算;李长云(1971-),男,湖南衡阳人,教授,主要研究方向为可信软件;温向兵(1982-),男,甘肃天水人,硕士研究生,主要研究方向为可信软件.

道的新颖信任模型,合格声誉系统允许 agents 去积极地提供第三方关于过去行为的参考作为建立信任关系的平均值,这样只需一点花费就可以迅速建立信任关系。Teacy[5] 提出了 TRAVOS-C 系统,试图解决陌生智能体间交互的可信性问题。Shi[6] 利用非参数贝叶斯模型,提出了情形敏感的方法来模型化和学习多维可信参数,并应用该方法在没有先前交互关系的情况下建立针对陌生交互主体的可信。Haller 和 Teacy 的方法不具备自动和直观地学习可信参数的能力;Shi 的方法可以简单预测未来的交互事件,但没有考虑到对交互事件序列组成的交互行为的预测问题。

上述的信任模型主要从分析信任特点和交互的结构出发,对陌生交互实体的行为可信性研究较少。

根据普适计算环境的特点和陌生实体的交互行为,笔者认为,采用先进的统计的机器学习方法对运行时实体的行为状态和行为踪迹的规律性进行分析,建立普适计算软件的行为预测模型,对交互实体元素的意图进行推断,对将来行为态势进行主动预测,可以为普适计算软件系统的全方位管理和灵活的动态调控提供坚实基础。

2　行为分析与态势预测模型

2.1　软件行为和行为可信的定义

在普适计算环境下,实体的交互行为是通过软件的交互行为来表现,因此关注软件的行为就是实体的行为。文献[7,8]指出了软件行为可信——运行期实体元素间的交互行为及其结果总是可预期与可控制的,即能够做到行为状态可监测、行为结果可评估、异常行为可控制。运行期交互实体可信性包括身份和行为可信。用户身份可信是交互实体的身份可以被准确鉴定,不被冒充,即实体的身份真实有效。实体的行为可信是指实体的行为是否可预测、评估和管理;是否产生了欺骗、不可信等恶意行为;是否对数据会造成破坏或毁坏。

笔者认为普适环境下实体的行为可信定义为:系统运行期实体 α 对实体 β 的行为可信是指:α 的行为序列应在实体 β 的各个事件的输出结果范围内;然后根据 α 的行为在结果范围中的分布情况 D_β;再用统计的机器学习技术分析事件序列或行为意图以确定交互行为的可信性,并对后续可能的行为态势进行主动预测,这可有效预防交互过程中欺骗、欺诈等反常行为。

2.2　行为分析与态势预测模型

行为分析与态势预测模型如图 1 所示。针对普适计算环境下运行期的陌生交互实体(1),用监测器(2)监测实体交互过程中的接口数据;通过数据预处理(3)对监测获得的原始数据进行过滤、降维和归一化,然后得到能够表示实体交互行为的可信参数;用 HDP 对交互过程中的可信参数进行聚类(4),以确定交互事件的类型;根据聚类结果确定每个时刻实体的行为状态,再利用无限隐马尔可夫模型对一段时间内的状态序列进行识别(6),如果该序列与原知识库(5)中的某条序列匹配,则可以根据事先定义的实体行为语义对交互过程中实体行为分析与预测。如果在知识库中找不到与观察序列匹配的状态序列,则采用无限隐马尔可夫模型学习(7),将聚类得到的新状态序列加入知识库中,再人为地给此序列定义其行为语义。最后根据事先定义好的实体行为语义进行风险分析和意图分析(8),然后再根据意图和风险分析的结果指导实体的运行。

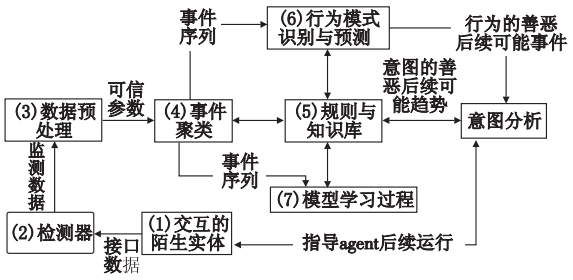

图1　行为分析与态势预测模型

2.3　实体交互过程中的监测和数据预测处理

为了后期的分析、预测等环节的进行,必须监测实体交互过程中的参数。软件的监测技术已经发展了三十多年,目前已较为成熟,主要有 AOP 方法、封装方法、反射方法、插装方式、截获器方式和监控 API 方式等方式。本文将充分利用这些成熟技术。

在进行 Dirichlet 过程对参数聚类之前,需要对监测获取的原始信息进行预处理。过滤、降维和归一化等过程是目前参数预处理的主要方法。对于监测内容过多的交互过程中,本文采取过滤方法,去除诊断和预测不需要的信息,以提高聚类参数的效率;针对信息参数的多维化,根据应用目的的不同需求,降低要处理信息的维数;针对格式的多样化,对监测信息进行归一化处理,将规格不同的数据用规范的形式进行描述。经过预处理的监测数据内容精炼、格式统一、形式规范,便于实体行为的分析与态势预测。

2.4　Dirichlet process 和 hierarchical Dirichlet process 对参数聚类[9]

假定,可能输出结果 X 的空间是离散和有限的,当进行 n 个观察之后,第 i 个观察是 X_i。本文认为被观察变量 X 是离散的,有 r ($r\geq 2$)个可能状态 $x_1,\cdots,x_r$,用多项式抽样所得的简单共轭先验 $p(\theta|\xi)$ 就是 Dirichlet 分布。

$$p(\theta|\xi)=\mathrm{dir}(\theta|\alpha_1,\cdots,\alpha_r)\equiv\frac{\Gamma(\alpha)}{\prod_{k=1}^{r}\Gamma(\alpha_k)}\prod_{k=1}^{r}\theta_k^{\alpha_k-1}$$

其中:$\alpha=\sum\alpha_k$ 且 $\alpha_k>0$, $k=1,\cdots r$;数量 $\alpha_1,\cdots,\alpha_r$ 指区别于 θ 的超参数;$\theta=\{\theta_1,\theta_2,\cdots,\theta_r\}$ 与物理分布 $x^1,\cdots,x^r$ 相对应。

Dirichlet 过程(Dirichlet process, DP)混合模型无须事先设定混合分量个数,DP 为混合分量个数提供一个非参数先验,而其聚类过程则被形象地描述为中餐馆过程(Chinese restaurant process, CRP),其优点是能将选择模型分量个数和确定模型参数联合进行。

多层 Dirichlet 过程(hierarchical dirichlet process, HDP)着重考虑多个相关数据组的学习问题,每个组用一个无限混合模型描述,混合成分(component)可以跨组被共享。单个组的观察 $\{x_1,\cdots,x_N\}$,每个 x_N 由分布 $x_i=F(\theta_i)$ 产生。参数 θ_i 从未知混合分布 G 提取,G 符合 Dirichlet 过程 $G\sim\mathrm{DP}(\gamma,H)$。其中:γ 是一个正实数;$H$ 是 G 的基本分布。该模型被称做 Dirichlet 过程混合模型,该模型具有混合成分数量无上界且能够从数据中自动推理出来。

2.5　iHMM

本文的行为分析预测模型中采取的是无限隐马尔可夫模型(iHMM)作为行为识别和分析的模型,由于隐马尔可夫模型结构必须事先指定,即状态数量固定。在一些应用程序中,要

获得模型复杂性需要花很大代价进行模型选择,有的应用程序可能没有这样固定的“正确模型”。然而 iHMM 更适合于行为复杂的普适计算环境,状态数的随时增加,可以随时计算模型的初始条件——状态转移概率矩阵 A。

Beal 等人首先提出了 iHMM,并提供了一个用于推理的合适的抽样模式。Teh 则认为 HDP 能被用于重构 iHMM,并提供了一个合适的抽样模式。

3　智能商场描述和仿真实验

3.1　情景描述

假设商场是一个普适场景,实体与实体之间的交互是通过无线网络交互,商场的监测主机是这个环境的拥有者(O),每个携带智能手机进入商场的顾客都可以看成一个实体(A),因此可以将顾客看成该系统环境下的陌生实体(相对于环境拥有者来说)。A 到商场的任何一个地方,O 都可以识别他的状态(位置)。商场的每一件商品都有可以识别的惟一标签 ID(S)。

A 进入商场后,可能有一系列位置状态序列。假设 A 进入商场,带走了某件商品(通过 ID 识别),但是主机没有发现 A 有去收银台的状态,而是直接走出了商场。此时,可以判定 A 有偷东西行为。

通过检测智能商场环境下每天交互的数据,并对获取的数据进行预处理;再利用聚类方法得到的位置状态序列作为训练样本,对 iHMM 进行训练;最后再将训练后得到的样本——位置状态序列储存在知识库中,并人为地给出序列的语义,如非常可信、可信、一般可信、不可信、非常不可信等级别。训练后的模型就可以投入试运行。通过监测、预处理和聚类得到顾客的位置状态序列,然后再利用模型分析顾客的状态序列,得出最佳匹配的解析;根据分析的结果再指导主机实体的运行。比如得出的分析结果为顾客有不可信的行为,主机可以提前发放广播或者发送短信给顾客随身所带的智能手机,可以提前阻止顾客的不良行为,即可以维护商场的利益。如果在知识库中找不出与聚类得到的观察序列匹配的状态序列,则可以重新对模型学习,再将此条序列加入知识库中。

3.2　Matlab 仿真过程

3.2.1　无限隐马尔可夫模型训练(学习过程)

无限隐马尔可夫模型训练过程首先是利用 Dirichlet 过程对在智能商场环境下实体过程中监测得到的参数进行聚类;然后再将聚类后的状态序列对无限隐马尔可夫模型进行训练;最后调整模型参数,使得 $P(O/\lambda)$ 最大。图 2 是行为分析与态势预测流程。

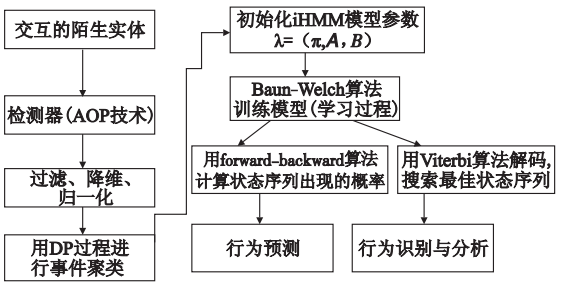

图2　行为分析与态势预测流程图

其中:$\lambda=(\pi,A,B)$,A 为状态转移概率矩阵,B 为给定状态下的概率分布,π 为初始状态空间的概率分布。由于普适环境的随时随地性,在实体与实体交互过程中随时都会出现新的状态,此时又要用 HDP 过程来确定新的状态转移概率矩阵 A'。训练过程可以通过 Baum-Welch 算法估计。如图 3 所示,当训练样本抽样 200 次时,状态数 K、状态之间的相似度 joint log likelihood、Dirichlet 分布超参数 alpha0 和 gamma 慢慢地趋向稳定,可以证明无限隐马尔可夫模型训练成功。

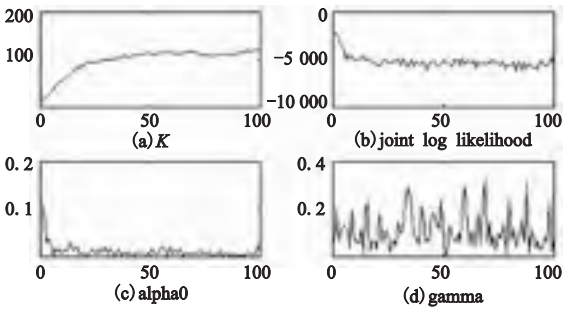

图3　iHMM_HDP训练结果

3.2.2　行为识别与分析(解码过程)

行为识别与分析过程反映的是实体行为识别过程。根据 DP 或者 HDP 过程聚类所获得的实体观察序列 $O=O_1,O_2\cdots,O_T$ 和模型 $\lambda=(\pi,A,B)$,在知识库中找出与观察序列最匹配的实体实际行为状态序列 $Q=q_1,q_2,\cdots,q_T$。该过程可以理解为对聚类所得到的观察序列的最佳解释即找出环境中交互的陌生实体的行为。然而可以根据识别出来的实际实体行为来进行意图分析与风险估计。图 4 是识别与分析过程的仿真实验。其中,(a)是观察序列,(b)是隐含状态序列,(c)是解码得到的序列。

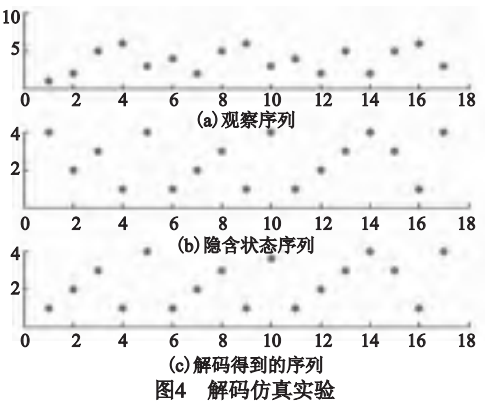

图4　解码仿真实验

3.2.3　行为预测(评估问题)

行为预测过程实际上是在交互过程中根据聚类所获得的实体行为观察序列和模型,计算 $P(O/\lambda)$。即给定模型和聚类观察序列,计算交互过程中下一步出现观察序列 $O'=o_1',o_2',\cdots,o_T'$ 的概率,然后根据其概率值对实体后续可能产生的行为进行主动预测,对不可信的行为提前干预,避免产生严重的后果。

4　结束语

本文着重对陌生实体交互过程中实体行为的可信问题进行研究。结合普适环境的特点,提出了一个基于嵌套的狄利克雷过程和无限隐马尔可夫模型的行为分析与态势预测模型,提供了一条解决普适环境中陌生交互过程中,判定实体的行为是否可信的有效途径。后续的研究思路主要侧重于在商场中添设传感器和通信设备收集实验数据,通过大量的样本训练行为分析与态势预测模型,改进目前在仿真分析中的偏差和效率低等问题。另外,构造适合智能商场的普适计算系统的原型,将实体的行为分析与态势预测模型添加在原有 (下转第 4571 页)

(f)当 $r \in R_v \wedge d_i r d_j \wedge d_i, d_j \in C_v$ 时

ABFS(c_i, c_j, P_{ci}, P_{cj}, TL), lookfor(d_i, P_{ci}), lookfor(d_j, P_{dj})

$\rho(c_i, d_i) = 1 - D(c_i, d_i)/\max(c_i)$, $\rho(c_j, d_j) = 1 - D(c_j, d_j)/\max(c_j)$

令 $\rho'_{ci} = \rho'_{ci} \cup \rho(c_i, d_i)$, $\rho'_{cj} = \rho'_{cj} \cup \rho(c_j, d_j)$

$R_w = R_w \cup \{r\}$, $C_w = C_w \cup \{c_i, c_j\}$

(g)转(c)

(h) $(\forall c) c \in C_w$, 令 $c.\rho_c = \max \rho'_c$

(i)结束

PMSD 算法的输入为 u、v 和概念格 TL,输出 w。从第(c)~(g)步对 R_u 循环,循环次数为 $|R_u|$。每次循环中,需在 R_v 中查找对应关系 r,最坏情况下循环 $|R_v|$ 次。因此总的循环次数为 $|R_u||R_v|$。第(f)步中的 ABFS( )是一个改进的宽度优先搜索算法,在 TL 中搜索 c_i 和 c_j 并返回所在路径 P_{ci} 和 P_{cj},再通过 lookfor( )分别在 P_{ci} 和 P_{cj} 中查找 d_i 和 d_j。设 TL 深度为 n,最大宽度为 m,则 ABFS( )最多查找 nm 次,lookfor( )最多查找 n 次,总查找次数最大上界为 $|R_u||R_v|(nm+n)$。第(h)步是对 w 中概念求最大匹配值,w 中概念匹配的最多形式为 $|C_u||C_v|$,所以这一步时间代价为 $|C_w||C_u||C_v|$。因此,最坏情况下算法 PMSD 总时间代价为 $|R_u||R_v|(nm+n) + |C_w||C_u||C_v|$。

一般情况下,概念图中的关系为二元关系,每个关系同两个概念关联。因此在一个概念图中,概念节点数近似为关系节点数的两倍。另外,根据投影匹配的同构性,有 $|C_w| = |C_u|$。因此 PMSD 最坏情况下总时间代价为 $|R_u||R_v|(nm+4|R_u|)$,时间复杂度为 $O(nm)$。

PMSD 算法空间复杂度为概念图 u、v 和 w 中所有节点所需要的存储空间以及概念类型格的存储空间,即 $|R_w| + |R_u| + |R_v| + |C_w| + |C_u| + |C_v| + nm$。对于二元关系的概念图,则存储空间近似为 $6|R_u| + 3|R_v| + nm$。一般情况下,关系式 $nm >> |R_u|$ 和 $nm >> |R_v|$ 成立,所以空间复杂度为 $O(nm)$。

4　实验分析

本文在研究过程中,以《计算机文化基础》课程为考察对象,建立了课程的概念图、概念关系和概念分类结构。在课堂上通过 4 个简答题对 96 名同学当堂测试,试卷总题目数为 384,平均每道试题 15 个概念。在测试中,有 86 道试题完全正确,正确率 100%,占总试卷的 22.4%;有 145 份试卷中出错的概念个数在两个以内,正确率为 86.7%,占总试卷的 37.76%;有 96 份试卷中出错的概念个数在 4 个以内,正确率为 73.33%,占总试卷的 25%;有 36 份试卷中出错的概念个数在 8 个以内,正确率为 46.67%,占总试卷的 9.38%;其余 21 分试卷因同学对问题的理解有一定偏差、题目的文字叙述不规范、对概念图不太熟练、没有及时掌握课程相关知识等原因出错较多,占总试卷的 5.47%。

从初步的实验分析可以看出,约 14.8% 的试题出错较多,有 85.2% 的考试结果能满足实际考试要求。只要在题目的叙述等方面进一步规范化,题目批阅的正确率可达到 90%,可用于实际的考试中,另外约 10% 的题目可通过人工方式解决。

5　结束语

高等院校的《计算机文化基础》课程是高等院校中非计算机专业学生的一门必修课程。本文从提高教学考核的基本方式出发,以自然语言理解的观点对主观题中简答题自动阅卷的关键问题进行了研究,问题包括:模糊含权概念图的定义、概念图的模糊不确定匹配、基于语义距离的投影匹配算法 PMSD 设计、概念类型格的定义以及《计算机文化基础》课程中概念类型格的设计。在文中,对 PMSD 算法的时间复杂度和空间复杂度进行了定量计算。该课题的研究,解决了《计算机文化基础》课程中简答题自动阅卷问题,对该课程中其他主观题和其他课程主观题的自动阅卷也具有一定参考价值。

参考文献:

[1] 高思丹,袁春风. 主观试题的计算机自动批改技术研究[J]. 计算机应用研究,2004,21(2):181-184.

[2] 刘培奇,李增智. 基于模糊含权概念图的主观题自动阅卷方法研究[J]. 计算机应用研究,2009,26(12):4565-4567,4584.

[3] SOWA J F. Conceptual structure[M]. New Jersey: Addison Welslely, 1984.

[4] WUWONGSE V, MANZANO M. Fuzzy conceptual graph[C]//Proc of Conceptual Graphs for Knowledge Representation. London: Springer-Verlag, 1993:430-449.

[5] 黄康,袁春风. 基于领域概念网络的自动批改技术[J]. 计算机应用研究,2004,21(11):260-262.

[6] 张蕾. 概念结构及其应用[D]. 西安:西北工业大学, 2001.5.

[7] 刘培奇. 新一代专家系统关键问题的研究[D]. 西安:西安交通大学,2005.

(上接第 4564 页)普适计算系统中,作为一个功能模块分析实体的行为可信性。最后研究了普适系统在实际智能商场中的应用。

参考文献:

[1] 徐光祐,史元春,谢伟凯. 普适计算[J]. 计算机学报,2003,26(9):1042-1051.

[2] WEISER M. The computer for twenty-first century [J]. Scientific American,1991,4(3-4):8-19.

[3] WHITBY A, JØSANG A, INDULSKA J. Filtering out unfair ratings in Bayesian reputation systems [C]//Proc of the 7th International Workshop on Trust in Agent Societies. Roma: Autonomous Agents & Multi Agent Systems, 2004:106-117.

[4] HUYNH T D, JENNINGS N R, SHADBOLT N R. Certified reputation: how an agent can trust a stranger[C]//Proc of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems. 2006:1217-1224.

[5] TEACY W T L. Agent-based trust and reputation in the context of inaccurate information sources [D]. Southampton: Southampton Electronics and Computer Science, University of Southampton, 2006.

[6] SHI Jian-qiang. A trust model with statistical foundation[D]. Ottawa: University of Ottawa, 2005.

[7] 林闯, 田立勤, 王元卓. 可信网络中用户行为可信的研究[J]. 计算机研究与发展, 2008, 45(12): 2033-2043.

[8] 田立勤, 林闯. 可信网络中一种基于行为信任预测的博弈控制机制[J]. 计算机学报, 2007, 30(11):1930-1938.

[9] NI Kai, QI Yu-ting, CARIN L. Muti-aspect target detection via the infinite hidden Markov model [J]. Journal of the Acoustical Society of America, 2007,121:2731-2742.