一种完备数据流的不确定数据择优算法

徐雪松[1] 徐 佳[2] 郭立玮[3] 张 宏[4] 周金海[1]

(南京中医药大学信息技术学院 南京 210046)[1] (南京邮电大学计算机学院 南京 210003)[2]
(南京中医药大学中医药研究院 南京 210046)[3] (南京理工大学计算机科学与技术学院 南京 210094)[4]

摘 要 针对射频识别(RFID)数据与上层应用需求之间存在的信息鸿沟及其需要实时处理的特征,提出了一种完备数据流的不确定数据择优算法。分析了常规粒子滤波方法存在的不足之处,采用基于熵的方法推导属性最优权重,并利用可能度矩阵选择最佳粒子,从不确定RFID数据流上有效捕获对象的当前状态。算法的优化结果使得采样集向后验概率密度分布取值较大的区域运动,从而提高了算法计算效率并且显著地减少了精确定位所需的粒子数。最后,通过实例表明了该方法能够有效度量RFID数据中蕴含的不确定性。

关键词 物联网,射频识别数据流,优化估计,粒子滤波

中图法分类号 TP311 **文献标识码** A

Algorithm for Choosing Optimal Uncertain Data of Complete Data Streams

XU Xue-song[1] XU Jia[2] GUO Li-wei[3] ZHANG Hong[4] ZHOU Jin-hai[1]

(College of Information and Technology, Nanjing University of Chinese Medicine, Nanjing 210046, China)[1]
(College of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)[2]
(Research Institute of Chinese Medicine, Nanjing University of Chinese Medicine, Nanjing 210046, China)[3]
(Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China)[4]

Abstract To address the information gap between RFID data and the requirements of upstream applications, the character of real time of sensor data, an algorithm for choosing the optimal uncertain data of complete data streams was proposed. The drawbacks of generic particle filter were analyzed. Then an entropy-based method was adopted to estimate the most likely attribute weight for each object, by using possibility degree matrix to select optimal particles, to efficiently capture the possible locations and containment for tagged objects. The performance of the generic particle filter is improved. In this method, though particle optimization, particles are moved to the regions where they have larger values of posterior density function. The experimental results show the accuracy and efficiency and the number of particles needed for accurate location are reduced dramatically. Finally, a numerical example was given to show the feasibility and effectiveness in terms of measurement of underlying uncertainties over RFID data.

Keywords Internet of things, Radio frequency identification data streams, Optimal estimation, Particle filter

目前,射频识别(Radio Frequency dentification,RFID)技术在很多领域获得了越来越广泛的应用,但仍然有许多问题有待解决,其中一个重要问题是RFID传感器数据与上层应用需求之间存在信息鸿沟[1],原因是RFID应用场合中数据的准确度受到传感器自身及周围环境中信号散射、相互干扰的影响。因此有效地处理RFID数据的不确定性对于提高数据精度以及解决信息鸿沟至关重要。

在现代物流商品服务行业应用中,由于RFID识读本身以及识读中存在的关联关系,查询结果存在许多不确定性,如第X趟车,第Y号物品,经过哪些中转,延迟交付的可能性多大等等。对于这些概率数据库,由于可能实例的存在,相同查询计划返回查询结果的概率值可能不同,其根本原因是设计查询计划的过程中未考虑到数据的概率之间的相关性,导致重复计算[2]。要解决此问题,必须在传感数据首次流入处理系统时为它们附加上可能的概率。

目前,对RFID数据的不确定性度量研究比较少,现有的概率估计方法不能直接应用于RFID数据管理。文献[3]首次讨论了利用血统模拟数据的不确定性与查询结果;文献[4]采用元组上的全局布尔公式计算元组的概率,研究怎样提高不确定数据上的处理效率。然而已有的工作都是假设原始数

到稿日期:2012-09-06 返修日期:2013-01-01 本文受国家自然科学基金资助项目(30873449),江苏省科技支撑计划项目(BE2011012,BE2012184),江苏省中医药科技项目(LZ11203)资助。

徐雪松(1975-),男,博士,讲师,主要研究方向为传感数据处理、智能空间交互技术、模式识别等,E-mail:xxs105@yahoo.cn;徐 佳(1980-),男,博士,副教授,主要研究方向为物联网技术及其应用、无线网络路由技术等;郭立玮(1948-),男,研究员,博士生导师,主要研究方向为中药制剂、生物药剂学、制药工程技术;张 宏(1956-),男,教授,博士生导师,主要研究方向为信息安全、模式识别,E-mail:zhhong@mail.njust.edu.cn(通信作者);周金海(1958-),男,教授,硕士生导师,主要研究方向为人工智能、医药信息工程。

据概率已知(即预设好了不确定性),未提及原始数据概率的度量,尤其是多目标传感器数据概率的度量问题。文献[5-7]提出各种优化重采样粒子质量的方法,其根据 RFID 数据进化情况自适应调整粒子数目以改善粒子退化现象,同时采用改进的粒子群算法 PSO(Particle Swarm Optimization)优化重采样性能,改善粒子贫化问题,从不确定性本源上提供原始数据中蕴含的不确定性度量。但粒子滤波算法的核心是利用一系列随机样本的加权和来表示所需的后验概率密度,具有最大权重值的粒子表示系统最可能处于的状态。这需要额外增加粒子数目,而粒子数目增加,经过几次迭代后很难收敛到真实状态,同时算法的计算效率极大降低,不能满足 RFID 数据流的实时性要求。为此,本文提出一种完备数据流的不确定数据择优算法,该算法利用元组对属性的依赖性进行加权集成,以具有最优权重值的粒子从不确定 RFID 数据流上有效捕获对象的当前状态,同时减少所需粒子数目,适应完备数据流一次扫描算法,满足实时处理的要求。

1 完备数据流模型与定义

定义 1 设 $S=(X,A)$是一个数据流,X 表示对象的非空有限集合,A 表示属性的非空有限集合,任意属性 $a\in A$,a:$X\to V_a$ 是一个映射,其中 V_a 称为 a 的值集。对于给定数据流 $S=(X,A)$,若任意属性 $a\in A$,且 V_a 不含空值,则称 S 是一个完备数据流。

定义 2 RFID 数据流集 $\bar{S}$ 由多条完备数据流 S 组成,滑动窗口中的 S 部分相当于关系数据模型中的表,元组 $x\in S$,由多个属性值组成。

由于数据流潜在的无限性,不可能将所有的数据全部存储后才计算,因此必须确定合适的数据流窗口(滑动窗口)。大小适中的滑动窗口既保留了窗口中的采样样本统计完备性,又降低了数据流的存储代价。令 p_i^{avg} 表示在一个间隙中标签 i 的观测概率,若在平滑窗口中间隙的个数 n_i 满足不等式 $n_i\geqslant\left|\frac{\ln(1/\rho)}{p_i^{avg}}\right|$,则可以保证在窗口 n_i 中以大于 $1-\rho$ 的概率读取标签 i[8]。

定义 3 基于测量值和控制值的粒子滤波,用一个含权的点集$\{(x^i,\hat{w}^i),i=1,2,\cdots,N\}$近似后验概率分布,根据 RFID 数据流的应用背景,采用如下带有倍增噪声的非线性模式[9]:

$$x_t=f(x_{t-1})(1+v_{t-1}) \tag{1}$$

$$y_t=h(x_t)(1+u_t) \tag{2}$$

式中,y_t 为测量值,$f(x_t)=0.5x_t+\frac{25x_t}{1+x_t^2}+8\cos(1.2t)$,$h(x_t)=x_t^2/20$,$v_t$ 和 u_t 是均值为零,方差为 Q 与 R 的白噪声。

2 属性权重优化模型与定义

RFID 完备数据流的数据不确定性可细分为元组级不确定性和属性级不确定性。元组级不确定性描述元组的存在与否,而属性级不确定性并不涉及整个元组的不确定性。以数据流形式返回查询结果,会产生数据相同但概率值不同的查询结果,导致重复计算,因此,必须考虑不确定元组对属性的依赖程度,给属性增加相应的权重,然后得到每个元组的集值优化矩阵,进而根据集值优化矩阵对元组进行择优选取。

对于一个完备数据流 $S=(X,A)$,流值 $S(t)$开始于 t 时刻,在滑动窗口中每隔 $\triangle t$ 秒采样概率值不同但数据相同的不确定元组$\{x_1,x_2,\cdots,x_m\}$,其属性集 $A=\{a_1,a_2,\cdots,a_n\}$,则元组对属性的依赖度可用不确定变量 B 表示,$B=\{b_i,i=-\lambda,\cdots,\lambda\}$,且满足条件(1)若 $i>j$,则 $b_i>b_j$;(2)若 $b_i\geqslant b_j$,则 $\max(b_i,b_j)=b_i$;(3)若 $b_i\leqslant b_j$,则 $\min(b_i,b_j)=b_i$;即 B 可定义为$\{b_{-4}=$极低,$b_{-3}=$很低,$b_{-2}=$低,$b_3=$很高,$b_4=$极高$\}$。

定义 4[10] 存在 $b_1=|b_{\alpha 1},b_{\beta 1}|$和 $b_2=|b_{\alpha 2},b_{\beta 2}|$两个不确定变量,其中$(b_\alpha,b_\beta)$分别表示下限和上限,若 $len(b_1)=\beta_1-\alpha_1$,$len(b_2)=\beta_2-\alpha_2$ 表示两个不确定变量的长度,则称:

$$P(b_1\geqslant b_2)=\frac{\max(0,len(b_1)+len(b_2)-\max(\beta_2-\alpha_1,0))}{len(b_1)+len(b_2)} \tag{3}$$

为 $b_1>b_2$ 的可能度,且具有如下性质:

(1)$0\leqslant P(b_1\geqslant b_2)\leqslant 1,0\leqslant P(b_2\geqslant b_1)\leqslant 1$;

(2)$P(b_1\geqslant b_2)+P(b_2\geqslant b_1)=1$,特例,$P(b_1\geqslant b_1)=P(b_2\geqslant b_2)=0.5$。

定义 5 b_1 与 b_2 之间的距离为:

$$d(b_1,b_2)=1/2(\beta_1-\beta_2+\alpha_1-\alpha_2) \tag{4}$$

定义 6 在滑动窗口中每隔 Δt 秒采样概率值不同但数据相同的不确定元组$\{x_1,x_2,\cdots,x_m\}$,其属性集 $A=\{a_1,a_2,\cdots,a_n\}$,对于元组 $x_i\in X$,按第 j 个属性 a_j 进行测度,得到不确定数据流权重优化矩阵 $R=(r_{ij})_{m\times n}$,其中 $w=(w_1,w_2,w_3,\cdots,w_n)^{\mathrm{T}}$ 为属性权重向量,$w_j\in(0,1)$,$j=1,2,\cdots,n$,$\sum_1^n w_j=1$。

由于完备数据流的数据不确定性,查询计划往往难以给出明确的属性权重,甚至属性权重完全未知,因此需要事先确定属性的权重。

对于概率值不同但数据相同的采样元组,各元组在属性 a_i 下的属性值差异越小,说明该属性对各元组优劣比较的作用越小;反之,如果属性值能使所有元组的属性值有较大的偏差,则说明该属性对元组优劣选择将起重要作用。因此,从正确捕获当前元组的角度看,偏差越大应赋予越大的权重,偏差越小就应该赋予越小的权重。特别地,若所有元组在属性 a_i 下的属性值无差异,则属性 a_i 对各元组的比较不起作用,可令其权重为零。

对于属性 a_j,若元组 x_i 与其他所有元组的偏差定义为:

$$D_{ij}(w)=\sum_1^m d(r_{ij},r_{kj})w_j,i\in M,J\in N \tag{5}$$

那么,对属性 a_j 而言,各元组的属性值的偏差为:

$$D_j(w)=\sum_1^m D_{ij}(w)=\sum_{i=1}^m\sum_{k=1}^m d(r_{ij},r_{kj})w_j,j\in N \tag{6}$$

而对属性 a_j,各元组属性值的标准差和平均差分别为:

$$S_i(w)=\overline{\frac{1}{m}\sum_{i=1}^m d^2\left[r_{ij}w_j,\frac{1}{m}\sum_{k=1}^m r_{kj}w_j\right]}=w_j\overline{\frac{1}{m}\sum_{i=1}^m d^2(r_{ij},r_j)},j\in N \tag{7}$$

$$V_j(w)=\frac{1}{m}\sum_{i=1}^m d\left[r_{ij}w_j,\frac{1}{m}\sum_{k=1}^m r_{kj}w_j\right]=\frac{w_j}{m}\sum_{i=1}^m d(r_{ij},r_j),j\in N \tag{8}$$

式中,$r_j=\frac{1}{m}\sum_{k=1}^m r_{kj}$ 表示属性 a_j 下各元组的平均属性值;$d(r_{ij},$

r_j)表示元组 x_i 的属性值与属性 a_j 的平均值之间的距离。

同时，对于权重优化矩阵，采用熵作为不确性的一种度量，小的熵意味着变量更为确定，即熵很小时，就可认为该变量是确定的，并可将返回概率最大的对应值作为该不确定变量的值。具有 n 个可能取值的 w_i 不确定变量修正值可用熵表示为：

$$H(w)=-\sum_{i=1}^{n}w_iT(w_i)\log_cT(w_i) \tag{9}$$

式中，$T(b_i)$为 B 取值 b_i 的概率质量函数。基于熵的定义，熵对不确定变量 B 可能取值的个数是敏感的。例如具有 10 个相同概率取值的不确定变量的熵会远大于具有 3 个相同概率取值的不确定变量的熵。为了降低这种影响，利用具有 n 个可能取值的最大熵对拥有 n 个可能取值的不确定变量的熵归一化处理得下式：

$$H(w)=(-\sum_{i=1}^{n}w_iT(w_i)\log_cT(w_i))/(-\frac{1}{n}\log_c(\frac{1}{n})) \tag{10}$$

引入了权重优化问题，其解是建立在权重目标函数值的基础上，并通过权重向量进行某种上、下关系的比较来定义。考虑属性权重向量 w 的选择应使在各属性下所有元组的总的组合偏差之和最大，可构造目标函数：

$$\begin{aligned}\max F(w)&=\sum_{j=1}^{n}(\hat{a}D_j(w)+\hat{b}S_j(w)+\hat{c}V_i(w)+\hat{d}H_j(w))\\&=\sum_{j=1}^{n}w_j[\hat{a}\sum_{i=1}^{m}\sum_{k=1}^{m}d(r_{ij},r_{kj})+\hat{b}\overline{\frac{1}{m}\sum_{i=1}^{m}d^2(r_{ij},r_j)}\\&\quad+\frac{\hat{c}}{m}\sum_{i=1}^{m}d(r_{ij},r_j)+\hat{d}H(w)]\\&\hat{a}+\hat{b}+\hat{c}+\hat{d}=1,\hat{a}\geqslant0,\hat{b}\geqslant0,\hat{c}\geqslant0,\hat{d}\geqslant0\end{aligned} \tag{11}$$

$\hat{a}=0$ 表示依赖度只考虑平均差而不考虑标准差，$\hat{b}=0$ 表示依赖度只考虑标准差而不考虑平均差，若两者都为零，则表示标准差和平均差两者兼而考虑。记

$$\chi_j=\sum_{i=1}^{m}\sum_{k=1}^{m}d(r_{ij},r_{kj})$$

$$\gamma_j=\frac{1}{m}\sum_{i=1}^{m}d^2(r_{ij},r_j),\eta_j=\frac{1}{m}\sum_{i=1}^{m}d(r_{ij},r_j)$$

$$\mu_j=(-\sum_{i=1}^{n}T(w_i)\log_cT(w_i))/(-\frac{1}{n}\log_c(\frac{1}{n}))$$

由此，求解 w 等价于求解如下单目标最优化问题：

$$\max F(w)=\sum_{j=1}^{n}w_j(\hat{a}\chi_j+\hat{b}\gamma_j+\hat{c}\eta_j+\hat{d}\mu_j)$$

$$\text{s.t.}\sum_{j=1}^{n}w_j^2=1,w_j\geqslant0,j\in N \tag{12}$$

解该函数得：

$$w_j=\frac{\hat{a}\chi_j+\hat{b}\gamma_j+\hat{c}\eta_j+\hat{d}\mu_j}{\sum_{j=1}^{n}[\hat{a}\chi_j+\hat{b}\gamma_j+\hat{c}\eta_j+\hat{d}\mu_j]^2},j\in N \tag{13}$$

对上述权重向量作归一化处理，得：

$$w_j=\frac{\hat{a}\chi_j+\hat{b}\gamma_j+\hat{c}\eta_j+\hat{d}\mu_j}{\sum_{j=1}^{n}\hat{a}\chi_j+\hat{b}\gamma_j+\hat{c}\eta_j+\hat{d}\mu_j},j\in N \tag{14}$$

在求出属性的最优权重向量 w_j 后，可得到各元组的综合属性值 $z_i=\sum_{j=1}^{n}r_{ij}w_j,i\in M$。可利用式(3)计算 z_i 之间的可能度，并建立可能度互补矩阵 $P=(p_j)_{m\times m}$，其中 $p_i=p((z_i(w)\geqslant z_j(w)),p_{ij}\geqslant0,p_{ij}+p_{ji}=1,p_{ii}=0.5,i,j\in M$，该互补判断矩阵最优元组即为所求解。

定理 互补判断矩阵 P 必有最优元组。

证明：通过调整偏差变量的大小，一定可以得到满足互补判断矩阵 P 相关约束式的权重，即互补判断矩阵 P 的可行域必为非空，所以互补判断矩阵 P 必有最优元组。

证毕。

3 算法描述

为满足 RFID 数据不确定性的在线度量，本文对粒子滤波算法进行了改进，提出了一种完备数据流的不确定数据择优算法。算法描述如下：

步骤 1 取得量测值，采用文献[5]定义的适应度函数初始化粒子，设计合适的滑动窗口大小，其输入为标签平均概率与置信度概率，输出为合适的窗口尺寸 F。

步骤 2 在 Δt 时刻，从重要性密度函数采样 L 个粒子，用$\{x_k^i,\hat{w}_k^i\}_{i=1}^{L}$表示，令每个样本的初始权值为$\hat{w}_k^i=1/L,i=1,2,\cdots,L$，重要性密度函数取转移先验概率 $x_k^i\sim q(x_k^i|x_{k-1}^i,y_k)=p(x_k^i|x_{k-1}^i)$。

步骤 3 粒子权值更新：根据最新量测值更新当前粒子权值：

$$\begin{aligned}\hat{w}&=\hat{w}_{k-1}^i p(y_k|x_{k-1}^i)\\&=\hat{w}_{k-1}^i\frac{p(y_k|x_k^i)p(x_k^i|x_{k-1}^i)}{q(x_k^i|x_{k-1}^i,y_k)}\end{aligned}$$

步骤 4 利用文献[7]的 PSO 算法更新每个粒子的速度与位置，使粒子不断地向真实状态靠近。设定阈值 ε，判断方差是否小于阈值 ε，若方差不小于阈值，则转至步骤 6，否则向下进行。

步骤 5 通过比较求出粒子的个体极值，再利用 PSO 算法更新粒子速度与位置，摆脱次优位置，跳出局部最优，驱动粒子向全局最优位置靠近。

步骤 6 粒子权值归一化：$\hat{w}_k^i=\hat{w}_k^i/\sum_{i=1}^{L}\hat{w}_k^i$。

步骤 7 重采样：当 $L_{eff}=1/\sum_{i=1}^{L}(\hat{w}_k^i)^2<L_{thrshold}$ 时，对$\{x_k^i,\hat{w}_k^i\}_{i=1}^{L}$原来带权粒子进行重采样，得到等权粒子$\{x_k^i,L^{-1}\}_{i=1}^{L}$。

步骤 8 将滑动窗口中概率值不同、数据相同的 m 个粒子，按照其 n 个属性，得到不确定权重优化矩阵 R。

步骤 9 利用式(10)熵的归一化处理和式(14)权重向量归一化处理，求出属性的最优权重向量 w_j，并求得各粒子的综合属性值 $z_i=\sum_{j=1}^{n}r_{ij}w_j$。

步骤 10 利用不确定变量之间的可能度式(4)进行两两比较，并构造互补判断矩阵 P。

步骤 11 可根据互补判断矩阵 P 的性质，构造一个简洁公式：

$$w=\frac{1}{m(m-1)}[\sum_{j=1}^{m}p_{ij}+\frac{m}{2}-1],i\in M \tag{15}$$

求得矩阵 P 的排序向量 $w=\{w_1,w_2,\cdots,w_n\}^{\mathrm{T}}$，按其分量大小对粒子进行排序并择优。

步骤 12 粒子状态估计：$\hat{x}=\sum_{i=1}^{L}\hat{w}_k^i x_k^i$；粒子方差估计：

$p_k=\sum_{i=1}^{L}\hat{w}_k^i(x_k^i-\hat{x}_i)(x_k^i-\hat{x})^T$。

步骤 13　判断 t 时刻是否为目标最后时刻，若是，则算法结束，否则，令 $t=t+1$，返回步骤 1，递推估计下一 Δt 时刻目标状态的后验概率。

其中，粒子即是完备数据流中的元组。引用 PSO 算法的实质是利用本身信息、个体极值信息和全局极值信息这 3 种信息指导粒子的下一步迭代位置。而通过改进优化后，可使得粒子集在权重值更新前更加趋向高似然区域，利于解决粒子贫化问题，同时，自适应调整粒子数目，改善粒子退化现象。步骤 8 至步骤 11 在滑动窗口中采样，在不增加粒子数目的同时，利用互补判断矩阵择优选择最佳元组，进一步降低粒子数目，从而提高算法效率，满足 RFID 数据流的实时性要求。

4　实验验证

设 20 个 RFID 传感器随机分布于 8m×8m 的正方形实验室内，以反映位置信息的 RFID 数据样本为粒子，让携带标签的学生在 20 个阅读器识别范围内做随机匀速运动，针对 3 种不同环境进行实验。采样间隔为 0.4s，阅读器的识读率为 0.5～1，上位机配置为 CPU：Intel CoreTM 2 Duo(2.9GHz)/主存 4GB。每个 RFID 样本为粒子，其属性包括温度、湿度、高度、速度、位置信息等。采用式(1)和式(2)定义的模型测试本文算法，滑动窗口采用 60 个时间步骤(1，2，…，60)。为了验证本文提出算法的有效性，采用文献[7]提出的 PSOPF 算法作为对比。

实验 1　不同识读率的性能测试。

本文提出的算法可充分利用最近某个滑动窗口时间周期的观测值，即使识读率较低，也可获得相当准确的粒子信息。由于低识读率会影响元组对属性关系的有效确认，同时还会造成邻接元组之间的同位置信息丢失，因此元组与属性间的依赖关系推导的准确性会随着识读率的降低而较为明显地下降，如图 1 所示。

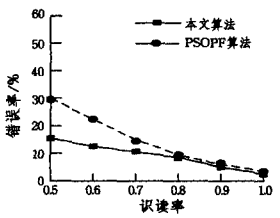

图 1　变化识读率测试

两种算法在识读率不小于 0.8 时错误率均低于 10%，相比于 PSOPF 算法，改进后的方法在识读率为 50%时确认粒子准确性的提高可达 9%。该方法采用概率熵的修正权重确认方法，在一定程度上增强了可能度矩阵对于粒子信息确认的促进作用。

实验 2　在线处理性能测试。

不确定数据流处理是 RFID 数据不确定性进化的主要特征，由于数据流到达的速度极快，算法需要满足一次性在线处理要求。两种算法的测试结果如图 2 所示，可以看出样本数量选择代价与在线处理约束之间的平衡，样本选择不足会导致漏读率提高，进而无法确认粒子的真实信息；另一方面，如果选择样本过多，算法的每次更新需持续数秒，从而也会提高漏读率。相比较而言，本文算法在两种极端情况下，仍然能获得较好的粒子真实信息。

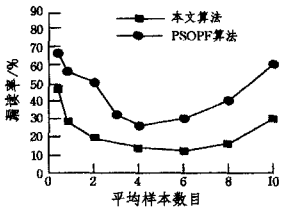

图 2　不同平均样本数目时的漏读率变化

对于式(1)和式(2)定义中给出的方差为 Q、R 的白噪声，分别比较两种算法的滤波性能。由表 1 的统计数据可知，在相同噪声环境下，相比于 PSOPF 算法，本文提出的算法的有效样本数最多，在抑制粒子退化和增加粒子多样性方面效率较高。将粒子数增加到 800 的 PSOPF 算法也不及本文算法的估计精度，且 PSOPF 算法估计所用时间长，说明在同等精度要求下，本文算法提高了算法的效率。同时，在噪声增加的情况下，本文算法的抗噪性能最好，在噪声增加的情况下仍能很好地抑制粒子退化，增加粒子多样性，保持算法估计的高精度。

表 1　两种算法滤波性能比较

条件	算法	粒子数	平均有效样本数	均方误差	运行时间 s
R=10 Q=1	PSOPF	400	48	2.63	1.05
	PSOPF	800	86	2.46	3.22
	本文算法	400	98	1.69	2.36
	本文算法	800	182	1.21	3.32
条件	算法	粒子数	平均有效样本数	均方误差	运行时间 s
R=20 Q=1	PSOPF	400	52	4.56	1.34
	PSOPF	800	94	4.15	3.56
	本文算法	400	106	2.86	2.62
	本文算法	800	194	1.33	3.55

结束语　粒子滤波作为近年来新兴的滤波算法受到很大的关注，但需要修正具有最大权重值的粒子来表示系统最可能处于的状态，这会额外增加粒子数目，同时由于 RFID 数据流的特点，需要在实际应用中对缺陷进行改进。本文通过基于熵的方法推导属性最优权重，并利用可能度矩阵选择最佳粒子，与现有算法进行比较，可获得更好的粒子信息确认，同时，排除综合权重值低的粒子，提高了算法计算效率。实验结果表明，本文所提出的方法可获得较准确的数据，同时具有良好的执行效率。

参 考 文 献

[1]　聂艳明，李战怀，陈群. 针对不确定射频识别数据流的改进概率推导方法[J]. 西安交通大学学报，2011，45(12)：45-52

[2]　Sarma A D，Theobald M，Widom J. Exploiting lineage for confidence computation in uncertain and probabilistic databases[A]// Proceedings of the 24th IEEE International Conference on Data Engineering[C]. Washington，DC：IEEE Computer Society Cancun，2008：1023-1032

[3]　Benjelloun O，Sarma A，Halevy A，et al. Uldbs：Databases with uncertainty and lineage[A]// Proceeding of the 32th International Conferance on Very Large Data Base (VLDB06) [C]. Seoul：VLDB Endowment，2006：953-964

[4] Sarma A D, Theobald M, Widom J. Exploiting lineage for confidence computation in uncertain and probabilistic databases[A]// Proceedings of the 24th IEEE International Conference on Data Engineering[C]. Washington, DC: IEEE Compu ter Society Cancun, 2008: 1023-1032

[5] 王永利,钱江波,等. 一种 REID 数据不确定性的自适应度量算法[J]. 电子学报,2011,39(3):579-584

[6] Christopher Re, Letchner J, Balazinksa M, et al. Event queries on correlated probabilistic streams[A]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data[C]. New York, NY: ACM, 2008: 715-728

[7] Fang Zheng, Tong Guo-feng, Xu Xin-he. Particle swarm optimized particle filter[J]. Control and Decision, 2007, 22(3): 273-277

[8] Shawn R J, Minos G, Michael J F. Adaptive cleaning for RFID data streams[A]// Proceeding s of the 32nd International Conference on Very Large Data Bases (VLD B06) [C]. Seoul: VLDB Endowment, 2006: 167-174

[9] Gordon N J, Salmond D J, Smith A F M. Novel approach to nonlinear/ non gaussian bayesian state estimation[J]. IEE Proceedings F In Radar and Signal Processing, 2002, 140(2): 107- 113

[10] Wu Z B, Chen Y H. The maximizing deviation method for group multiple attribute decision making under linguistic environment [J]. Fuzzy Sets and Systems, 2007, 158(14): 1608-1617

(上接第 146 页)

异算子产生的测试用例的总数。图 6 给出各变异算子的有效率,可以看出 FVS 和 IIV 变异算子比其它的变异算子发现错误的数量多,说明它发现错误的效率比较高,对大多数的 Web 服务而言,这两种变异算子是通用的;其次是 SVB 变异算子,它也是针对大部分的数据来产生测试用例,这种变异算子在现实中也是非常实用的;有效性最不理想的是 SSI 算子,因为它的适用范围比较窄,只适用于某个特定范围。系统使用的变异算子大部分可以较好地发现错误,只是某种变异算子针对某种参数类型有更好的效用。从实验结果可以看出,变异算子总体的有效率约为 24%,验证了测试用例生成算法的可行性和有效性。

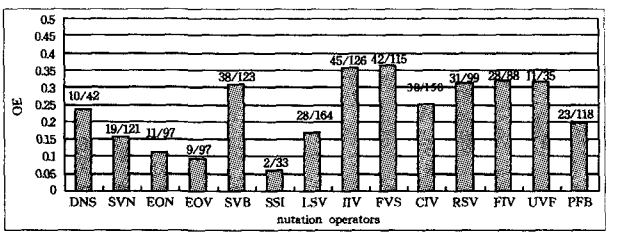

图 6 各变异算子的有效率

结束语 在基于 Web Service 的软件的开发和维护活动中,测试 Web Service 的脆弱性是至关重要的一环。研制针对 WS 脆弱性的自动化、半自动化的测试工具是 Web Service 亟待解决的课题。目前针对脆弱性的测试工具不多,主要有 SOAPUI 和 SMAT-WS,与这两类工具相比,WSVTS 测试工具原型具有如下优点:

(1)不同于 SOAPUI 的数据手工输入,WSVTS 的测试数据通过算法自动生成,WSVTS 基于 SOAP 消息参数的个数和类型实现了两种测试用例生成算法 TCFN 和 FDMA,所生成的测试用例具有很高的故障检测能力。

(2)实现了 15 种针对 SOAP 消息的变异算子,并通过实验验证了变异算子的有效率。

(3)工具的自动化程度较高,只需要少量的人工参与。

(4)具备较全面的功能模块,包括 SOAP 消息生成器模块、测试用例生成器模块、SOAP 消息变异模块、测试流管理模块、测试执行模块及安全分析模块等,几个模块协同工作完成测试过程。

(5)测试工程以项目形式管理,有利于测试的组织和管理,可以重现测试活动,得到的测试报告可以进行对比和分析。

参 考 文 献

[1] 陈锦富,卢炎生,谢晓东,等. 一个组件安全自动化测试平台的设计与实现[J]. 计算机科学,2008,35(12):229-233

[2] Lourival F, de Almeida J, Vergilio S R. Exploring Perturbation Based Testing for Web Services[C]// ICWS 2006. IEEE Computer Society, Los Alamitos, 2006: 717-726

[3] The Eviware SOAPUI 官方网站[EB/OL]. http://www.SOAPUI.org/2007

[4] Sourceforge Org[EB/OL]. http://sourceforge.net/forum/

[5] 罗作民,朱燕,程明. Web 服务测试工具 SOAPUI 及其分析[J]. 计算机应用和软件,2010,27(5):155-157

[6] Chen T Y, Eddy G, et al. Adaptive Random Testing Through Dynamic Partitioning[C]// Proceedings of the Fourth International Conference on Quality Software. 2004: 79-86

[7] Chen T Y, Leung H, Mak I K. Adaptive Random Testing[J]. LNCS, 2004, 3321: 320-329

[8] 李博涵,郝忠孝. 反向最远邻的有效过滤和查询算法[J]. 小型微型计算机系统,2009,30(10):1948-1951

[9] Kim H C, Choi Y H, Lee D H. Efficient File Fuzz Testing Using Automated Analysis of Binary File Format[J]. Journal of Systems Architecture, 2011, 57(3): 259-268

[10] Chan K P, Chen T Y, Towey D. Normalized Restricted Random Testing [C]// Springer-Verlag 2003, 2655: 368-381

[11] 陈锦富,卢炎生,谢晓东. 一种采用接口错误注入的构件安全性测试方法[J]. 小型微型计算机系统,2010,31(6):1090-1096