第 35 卷 第 4 期 计 算 机 工 程 2009 年 2 月
Vol.35 No.4 Computer Engineering February 2009
·软件技术与数据库· 文章编号：1000—3428(2009)04—0066—03 文献标识码：A 中图分类号：TP311.13

改进的 XML 智能数据清洗策略

翟学敏[1]，刘 渊[1,2]，刘 波[3]，毕蓉蓉[1]
(1. 江南大学信息工程学院数字媒体创意中心，无锡 214122；2. 南京理工大学计算机学院，南京 210094；
3. 中南大学信息学院，长沙 410083)

摘 要：针对 XML 数据的质量问题，以 XML 键为基础，借助多模板隐马尔可夫模型信息抽取策略与粒子群优化算法构建新的 XML 数据清洗方法。为了提高 XML 相似性数据并行检测效率，利用波函数对粒子群优化算法进行优化。仿真实验表明，与其他 XML 数据清洗算法相比，该方法的自适应学习能力强、人工参与程度低、计算量小，时间性能有 94%左右的提升。

关键词：XML 文档集；XML 键；粒子群优化算法；数据清洗；隐马尔可夫模型

Improved XML Intelligence Data Cleaning Strategy

ZHAI Xue-min[1], LIU Yuan[1,2], LIU Bo[3], BI Rong-rong[1]
(1. Digital Media Creative Center, College of Information Engineering, Southern Yangtze University, Wuxi 214122;
2. School of Computer, Nanjing University of Science & Technology, Nanjing 210094;
3. College of Information, Central South University, Changsha 410083)

【Abstract】 Aiming at the quality of XML data, this paper proposes a new XML data cleaning method based on XML key, the information of multiple templates Hidden Markov Model(HMM) draw-out strategy and Particle Swarm Optimization(PSO). For boosting the parallel detection efficiency of the XML similarity records, a wave function is used to give relevant improvements to PSO. Contrasted with other XML data cleaning algorithms, simulation experiments show that the optimized algorithm has powerful adaptive learning capability, lower labor cost, less calculation and better time rate around 94%.

【Key words】 XML document set; XML key; Particle Swarm Optimization(PSO); data cleaning; Hidden Markov Model(HMM)

1 概述

由于 Web 上积累了大量的 XML 数据，形成了许多“脏数据”，阻碍了商业应用，因此需要对它们进行挖掘、清洗，以保证和提高数据质量。国外信息化程度较高，对数据清洗的研究较多，国内的数据清洗研究集中在[1-2]：(1)特殊域清洗，主要解决某类特定应用域的数据清洗，用一个大的预定义规则库处理清洗过程中发现的问题，是目前研究得较多的领域。(2)与特定应用领域无关的数据清洗，主要集中在清洗重复的记录。这些研究或多或少存在不完备的地方，主要表现在：(1)研究主要集中在字符型数据上。(2)检测重复记录遭遇海量数据时，耗时太多，检测效率与精度较差。(3)大多数数据清洗工具都是针对特定领域的，其应用受到一定的限制。(4)研究主要集中在结构化数据上。

针对以上问题，本文以 XML 键为基础讨论构建 XML 数据清洗的过程：首先量化 XML 文档，利用多模板隐马尔可夫算法(Hidden Markov Model, HMM)[3-4]提取 XML 键值信息，利用波函数等方法优化粒子群优化算法(Particle Swarm Optimization, PSO)[5]对相似 XML 数据查找定位，然后完成合并、剔除、转换、补充等操作。重点研究了半结构化 XML 键的选取、知识库的抽取、量化计算方法与 PSO 的融合等问题。

2 基本定义与 XML 键的获取

定义 1 脏数据指不符合数据仓库或上层应用逻辑规格的数据，记为 *DirtyData*。

定义 2 清洗检查指检测出干净数据或脏数据的过程，记为：*cond*: *Data*→*boolean*; *True*: *data*∈*DirtyData*; *False*：*data*∈*CleanData*。

定义 3 数据清洗指从各种原始数据中抽取出干净数据的过程，可形式化为：*DataClean*:*RawData*→*CleanData*。

定义 4 一棵 XML 文档树为 $T=(V,chl,lab,val,Vr)$，其中，V 为 T 中的节点集合；Vr 为根节点；chl 表示子节点的集合；val 表示从集合 V 到 $E\cup S\cup A$ 的映射函数，其中，E 为元素名称集合；S 指代#PCDATA；A 为属性名称集合。

3 多模板隐马尔可夫模型与 XML 文档向量化

一个隐马尔可夫模型[6]是一个五元组：$(\boldsymbol{\Omega}_X, \boldsymbol{\Omega}_O, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$，其中，$\boldsymbol{\Omega}_X=\{q_1,q_2,\cdots,q_N\}$表示状态的有限集合；$\boldsymbol{\Omega}_O=\{v_1,v_2,\cdots,v_M\}$表示观察值的有限集合；$\boldsymbol{A}=\{a_{ij}\}$, $a_{ij}=p(X_{t+1}=q_j|X_t=q_i)$表示转移概率；$\boldsymbol{B}=\{b_{ik}\}$，$b_{ik}=p(O_t=v_k|X_t=q_i)$，表示输出概率；$\boldsymbol{\pi}=\{\pi_i\}$，$\pi_i=p(X_i=q_i)$表示初始状态分布。

$\boldsymbol{\lambda}=\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$是给定的 HMM 的参数，用于解决评估、解码与学习，而 XML 数据抽取需要解决 HMM 中的学习与解码问题，整个操作分为 3 步：(1)以 XML 键组合把数据训练为几个类；(2)训练 HMM 参数；(3)使用学习到的 HMM 参数进行 XML 数据抽取。具体如下：

基金项目：国家部委基础研究基金资助项目；2006 年江苏省教育厅青年骨干教师计划基金资助项目

作者简介：翟学敏(1983－)，女，硕士研究生，主研方向：网络信息管理系统；刘 渊，教授；刘 波，博士研究生；毕蓉蓉，硕士研究生

收稿日期：2008-05-10 **E-mail：**minmin0876@yahoo.com.cn

(1)基于马尔可夫链模型将数据训练为如下 4 个类：

$$\boldsymbol{A}_k=(p_{kij}),\quad p_{kij}=\frac{S_{kij}+\alpha_{kij}}{\sum_{j=1}^{n}(S_{kij}+\alpha_{kij})}\quad,\quad \alpha_{kij}=\frac{\beta}{n\times n} \tag{1}$$

其中，转移概率矩阵 $\boldsymbol{A}_k$ 表示第 k 个已标记的训练 XML 文档序列；S_{kij} 是训练 XML 文档序列中从标记状态 i 转移到状态 j 的次数；通常取 $\beta=n$，n 是模型状态数。

$$\pi_i=\frac{Init(i)}{\sum_{j=1}^{N}Init(j)},1\le i,\ j\le N \tag{2}$$

其中，$Init(i)$是所有训练序列中初始状态为 i 的序列个数。

$$\alpha_{ij}=\frac{C_{ij}}{\sum_{k=1}^{N}C_{ik}},\ 1\le i,\ j\le N \tag{3}$$

其中，C_{ij} 是所有训练序列中从状态 S_i 转换到状态 S_j 的次数。

$$b_j(v_k)=\frac{E_j(V_k)}{\sum_{k=1}^{M}E_j(V_k)},\ 1\le j\le N,\ 1\le k\le M \tag{4}$$

其中，$E_j(V_k)$是所有训练序列中状态 S_j 释放单词 V_k 的次数。

若第 k 个训练序列用矩阵 $\boldsymbol{A}_k$ 表示，第 1 个训练序列用矩阵 $\boldsymbol{A}_1$ 表示，那么这 2 个训练 XML 文档序列之间的距离为

$$D(\boldsymbol{A}_K,\boldsymbol{A}_1)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}p_{kij}\lg\frac{p_{kij}}{p_{lij}}+\sum_{i=1}^{n}\sum_{j=1}^{n}p_{lij}\lg\frac{p_{lij}}{p_{kij}}}{2\times n} \tag{5}$$

对于任意 2 个马尔可夫链，它们之间的动态特征差异越大，距离值就越大，相同时为 0。基于这种距离，通过调节距离的门槛值，可以控制得到的分类个数。

(2)对每一类训练数据，依次利用式(2)~式(4)训练初始概率矩阵 $\boldsymbol{\pi}$、转移概率矩阵 $\boldsymbol{A}$ 以及状态释放概率矩阵 $\boldsymbol{B}$。

(3)使用训练好的模型抽取 XML 数据时，结合每一个初始概率矩阵 $\boldsymbol{\pi}$、转移概率矩阵 $\boldsymbol{A}$ 和统一的释放概率矩阵 $\boldsymbol{B}$，使用马尔可夫模型的韦特比算法找出最优的标记序列，并从中选择一个概率最大的序列作为最终标记序列。即对每种分类模板使用韦特比算法一次，产生一个标记序列，从所有的最优标记序列中选出概率最大的序列作为最终输出结果。

在利用多模板的隐马尔可夫模型构建 XML 数据的清洗知识库后，对海量 XML 数据进行相应的清洗前，先把 XML 文档树向量化。

定义 5 XML 文档向量化：从 XML 数据库中按照 XML 键组合抽取 n 组各包含 k 个 XML 文档的参照集 RS，$RS=\{(r_1,r_2,\cdots,r_k)^1,(r_{k+!},r_{k+2},\cdots,r_{2k})^2,\cdots,(r_{(n-1)k},r_{(n-1)k+1},\cdots,r_{nk})^n\}$，其中，$(r_i,r_{i+1}\cdots,r_t,\cdots,r_{i+k-1})$应尽可能按差异最大化原则选取。通过$(r_i,r_{i+1},\cdots,r_t,\cdots,r_{i+k-1})$将 XML 数据库中的每一个 XML 文档 D 映射为一个 k 维的距离向量 $\boldsymbol{VD}$，$\boldsymbol{VD}$ 中的第 t 维坐标可以通过计算 D 和 r_t 之间的语义距离得到，$\boldsymbol{VD}[t]=ed(D,r_t)$，这样通过参照集 RS，每个 XML 文档 D 可被映射为 n 个 k 维距离向量 $\boldsymbol{VD}$，再将这 n 个 $\boldsymbol{VD}$ 的平均值作为每个 D 在 k 维坐标平面上的投影 $\boldsymbol{VD}[r_1,r_2,\cdots,r_k]$，从而降低 XML 维度。

本文的数据清洗过程如图 1 所示。

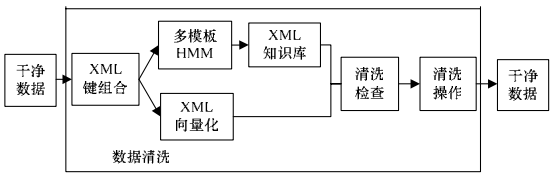

图 1　XML 数据清洗过程

为了更快捷地利用抽取的知识库对量化的 XML 数据进行相似性比较，特引入粒子群优化算法。

4　粒子群优化算法与 XML 数据清洗算法

粒子群优化算法是一种模拟鸟群觅食过程和聚集的全局优化算法，其优化表达式如下：

$$V_i^{t+1}=\omega\times V_i^t+c_1\times r_1\times(p_{bdi}-X_i^t)+c_2\times r_2\times(p_{gd}-X_i^t) \tag{6}$$

$$X_i^{t+1}=X_i^t+\alpha V_i^{t+1} \tag{7}$$

$$\omega=\omega_{\max}-(\omega_{\max}-\omega_{\min})\times\exp(1-sqrt(\frac{run_{\max}}{run})) \tag{8}$$

若抽取的 XML 文档知识库规模为 k，利用 PSO 算法进行相似性测量过程如图 2 所示。

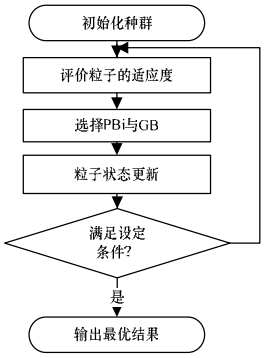

图 2　PSO 操作过程

在此过程中，将向量化的 XML 知识库投影为平面上的样本点，同时将二维平面进行等间距网格划分，并将待检测已向量化的 XML 数据粒子随机分散于一个平面上，即随机赋给每个粒子一对坐标(x,y)，同时以本粒子初始坐标为中心、r 为观察半径，利用式(9)或式(10)计算此模式在观察半径范围内的粒子相似度，同时依次检测该网格周围的邻域网格，若大于最少样本点数，将网格的坐标加入到邻域连接缓冲列表的尾部，否则将网格内的所有样本点丢弃，依次计算已形成的每一个文档类的检测中心。

在应用 PSO 检测相似重复记录之前，需要对数据进行以下操作：

(1)字段分裂。依据 XML 键组合的结构，利用多模板的隐马尔可夫模型抽取各个部分。

(2)验证和改正。根据知识库及相关领域知识验证提取值的正确性，若发现错误，则加以改正。

(3)数据标准化。将同一类型的数据用统一的格式表示，如日期、电话号码、性别。

在计算过程中，可对数字字符与字符串分别套用不同的公式计算相应的距离值。

(1)数字匹配问题

可采用将全部数字字符一一比较的方法得到标识距离，距离计算公式如下：

$$des=1-\frac{\sum_{i=1}^{\min(|a|,|b|)}d(a_i,b_i)+\max(|a|,|b|)-\min(|a|,|b|)}{\min(|a|,|b|)} \tag{9}$$

其中，a_i,b_i 为 2 个标识对应的字符；$d(a_i,b_i)$为对应字符间的距离，若 $a_i=b_i$，则 $d(a_i,b_i)=1$，否则，$d(a_i,b_i)=0$；$\max(|a|,|b|)-\min(|a|,|b|)$为多余的字符相应的距离。

(2)字符串匹配问题

字符串属于长信息字段，再分解后可根据标识的重要性分配权值，利用如下公式计算标识距离：

$$des = 1 - \frac{|a \cap b|}{\min(|a|,|b|)} \tag{10}$$

其中，字段距离为标识距离的加权和；a, b 都是标识。

由于粒子群算法的时间复杂度为 $O(n^3)$，代价较高，因此需根据其量化过程对粒子群操作进行优化，并建立波函数的粒子群(Wave Particle Swarm Optimization, WPSO)并行处理机制。在量子时域空间的框架下，三维空间粒子状态的波函数描述为

$$|\psi|^2 \, \mathrm{d}x\mathrm{d}y\mathrm{d}z = Q\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{11}$$

其中，$|\psi|^2$ 是波函数的概率密度，且满足

$$\int_{-\infty}^{+\infty} |\psi|^2 \, \mathrm{d}x\mathrm{d}y\mathrm{d}z = \int_{-\infty}^{+\infty} Q\mathrm{d}x\mathrm{d}y\mathrm{d}z = 1$$

则粒子群算法可优化成

$$p = (a.p_{id} + b.p_{id})/(a+b) \tag{12}$$

$$L = (1/g).abs(x_{id} - p) \tag{13}$$

$$x_{id} = p \pm L.(\ln(1/u)) \tag{14}$$

其中，p_{id} 是第 i 个粒子在解空间的第 d 维所找到的最优解；x_{id} 是第 i 个粒子在解空间第 d 维的当前值；p_{gd} 是所有粒子在解空间中第 d 维找到的最优解；a,b,u 都是(0,1)之间的随机数；$g>\ln\sqrt{2}$ 就能保证算法收敛。

XML 数据清洗算法如下：

输入 一个 FD_{XML} 集 Σ，路径集 $Paths(T)$

输出 一个干净的 FD_{XML} 集 Σ'

(1)寻找 XML(Σ ,$Paths(T)$)候选键，获取 XML 键组合 Σ_{keyset}。

(2)获取 XML(Σ , Σ_{keyset})知识库，得到动态知识库 Σ_{xmllib}。

(3)在知识库 Σ_{xmllib} 中利用 WPSO 匹配向量化的 XML 数据集：

1)在给定的范围内初始化种群粒子的位置 λ_0 和速度 V_0。

2)利用式(9)或式(10)计算每个粒子的适应度 f_i，并将粒子群的当前位置设为 p_{id}，将适应度最优位置的粒子设为 p_{gd}。

3)根据 2)的结果，利用式(12)~式(14)计算 P, L, x_{id}。

4)用个体粒子的速度产生选择该粒子更新位置方程的数据：

$$rand\_q = 1/(1+|(V_{\max} - V_{id})/(V_{id} - V_{\min})|) \tag{15}$$

5)由 2)产生的数据选择更新粒子位置的方程：如果 $rand\_q>0.5$，则 $x_{id} = p + L.(\ln(1/u))$ ；否则，$x_{id} = p - L.(\ln(1/u))$。

6)根据 2)中 x_{id} 的大小判断是否更新粒子的速度：如果 $x_i^t > x_i^{t-1}$，则 $x_{id} > x_{\max}$ ；如果 $x_{id} > x_{\max}$，则 $x_{id} = x_{\max}$， $v_{id} = -v_{id}$ ；如果 $x_{id} < x_{\min}$， 则 $x_{id} = x_{\min}$， $v_{id} = -v_{id}$。

在更新粒子速度时，即使粒子速度超出了预设范围，也不令 $v_{id} = v_{\max}$ ， $v_{id} = v_{\min}$ ， 避免分母中的 $(v_{\max} - v_{id})$ 和 $(v_{id} - v_{\min})$ 2 项为 0，而式(15)的分母取绝对值，使结果处于[0,1]，保证算法有效地运行。

(4)for each(p_i in Σ){//收集数据成 Σ'}。

在本算法中，WPSO 对 PSO 的改进如下：

(1)适合 XML 多维数据平面处理，而向量投影有利于降低 PSO 的维度灾难并简化计算规模。

(2)利用个体粒子速度产生一个(0,1)间的数代替 PSO 中由计算机随机产生的数。

(3)若个体粒子的位置超出预设范围，则该粒子反方向飞行。规定只要当前代粒子的适应值优于上一代，就不再更新粒子的速度 V_i^{t+1} 与位置 x_i^{t+1} ，减少了计算量。

本算法虽然由 4 个部分组成，但时间性能主要由 WPSO 决定，相对 PSO，循环代码并没有多少变化，但由于采用不同的更新与投影方法，因此迭代次数大大减少了。

5 实验仿真

实验分为 3 个部分：(1)测试多模板的隐马尔可夫模型对 XML 数据的抽取能力；(2)测试一般 PSO 与 WPSO 的性能差异；(3)利用其他数据清洗算法[7-8]对相同的 XML 数据进行对比测试。

5.1 数据集与实验设计

PC：C-M 1.73 GHz CPU，1 GB 内存，80 GB SATA5400 硬盘；OS：Win2000 专业 sp4 版；工具：Mathlab2007a 和 C#。分别选用英文 ACMSIGMOD(http://www.sigmod.org/record/xml/SigmodRecord/SigmodRecord.xml)中 2006, 2005, 2002, 1999 4 个年度的 XML 数据集，每个数据集由数千篇不同 XML 格式的 ACMSIGMOD 论文组成，选用 1 165 个文档，这 4 个数据集都有分类信息，在 ACMSIGMOD 文档中，每个文档通常属于多个分类，同一个分类的 XML 文档具有较强的相似性，因此，随机选取 3 次训练样本，测试结果取平均，并分别采用基于规则、聚类与本算法进行 XML 数据清洗实验评价。

5.2 实验结果分析

(1)多模板的隐马尔可夫模型对 XML 数据的抽取能力

本实验利用 XML 键组合作为模板，利用隐马尔可夫模型提取 XML 数据，根据 XML 数据规模，准确提取的数据及花费时间如表 1 所示。

表 1 提取的不同规模 XML 数据及花费时间

记录大小/个	文档数	有效记录数	花费时间/s
52	2	51	1.195 8
143	7	137	5.795 3
297	13	279	19.551 6
578	22	527	49.059 7
1 573	36	1 409	121.438 4

(2)WPSO 与 PSO 对 XML 数据的检测能力

为了检测各算法对 XML 数据的检测能力，以实验 1 提取的数据为基础投影到二维平面上，针对不同数据规模进行测试，实验结果如表 2 所示。从中可知，由于 WPSO 主要针对 XML 多维数据进行投影操作，相比一般优化的 PSO，在最终距离相近情况下，其迭代次数与时间花费至少降低 1 倍以上，数据量越大，效果越明显。

表 2 XML 数据相似性实验结果

记录大小/个	最终距离/个		迭代次数		花费时间/s	
	PSO	WPSO	PSO	WPSO	PSO	WPSO
52	93.247 9	94.885 1	5 029	2 803	93.247 9	27.274 9
143	362.940 9	363.352 8	3 729	1 356	38.055 6	15.325 4
297	655.425 4	655.896 6	3 564	976	50.950 8	15.327 5
578	1 082.100 0	1 080.000 0	1 447	289	50.713 6	30.011 7
1 573	3 180.900 0	3 180.300 0	2 293	521	152.011 0	52.356 0

(3)不同清洗策略对比实验

为了更好地比较 XML 数据清洗算法，将文献[7-8]提出的规则清洗算法与聚类清洗算法与本文算法进行对比实验。为了达到更好的清洗效果，针对不同数据规模加入不同噪音的测试数据，实验结果如表 3 所示。

(下转第 71 页)