基于遗传算法和隐马尔可夫模型的Web信息抽取的改进

李 荣[1] 胡志军[1] 郑家恒[2]
(忻州师范学院计算机系 忻州 034000)[1] (山西大学计算机与信息技术学院 太原 030006)[2]

摘 要 为了进一步提高Web信息抽取的准确性和效率,针对Web信息抽取的遗传算法和一阶隐马尔可夫模型混合方法在初值选取和参数寻优上的不足,提出了一种遗传算法和二阶隐马尔可夫模型内嵌结合的改进方法。在分层预处理阶段,利用格式信息和文本特征将文本切分成文本行、块或单个的词等恰当的层次;然后采用内嵌的遗传算法和二阶隐马尔可夫混合模型训练参数,保留最优和次优染色体,修正Baum-Welch算法的初始参数,多次使用遗传算法微调二阶隐马尔可夫模型;最后用改进的Viterbi算法实现Web信息抽取。实验结果表明,改进方法在精确度、召回率指标和时间性能上均比遗传算法和一阶隐马尔可夫模型的混合方法具有更好的性能。

关键词 Web信息抽取,遗传算法,二阶隐马尔可夫模型,分层
中图法分类号 TP391 **文献标识码** A

Improvement of Web Information Extraction Based on Genetic Algorithm and Hidden Markov Model

LI Rong[1] HU Zhi-jun[1] ZHENG Jia-heng[2]
(Computer Department, Xinzhou Teachers' University, Xinzhou 034000, China)[1]
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)[2]

Abstract In order to further enhance the accuracy and efficiency of Web information extraction, for the shortcomings of hybrid method of genetic algorithm and first-order hidden Markov model in the initial value selection and parameter optimization, an improved combined method embedded with genetic algorithm and second-order hidden Markov model was presented. In the hierarchical preprocessing phase, text was segmented hierarchically into proper lines, blocks and words by using the format information and text features. And then the embedded genetic algorithm and second-order hidden Markov hybrid model were adopted to train parameters, and the optimal and sub-optimal chromosomes were all retained to modify initial parameters of Baum-Welch algorithm and genetic algorithm was used repeatedly to fine-tune the second-order hidden Markov model. Finally the improved Viterbi algorithm was used to extract Web information. Experimental results show that the new method improves the performance in precision, recall and time.

Keywords Information extraction, Genetic algorithm, Second-order hidden markov model, Hierarchy

1 引言

随着互联网技术的迅速发展,网络信息呈几何级数增长,如何快速而准确地从海量非结构Web文本中自动抽取出潜在、有用的结构化信息成为一个重要的研究课题[1]。Web信息抽取是加快查找速度,提高查找准确率的重要手段之一,在信息查询、文本深层挖掘、问题自动回答等方面都起着非常重要的作用。近年来,基于统计模型的机器学习方法成为Web信息抽取新的研究热点。这类统计方法主要有隐马尔可夫模型(Hidden Markov Model, HMM)[2,3]、最大熵模型[4]、最大熵马尔可夫模型[5]、条件随机场[6]、支持向量机[7]和HMM混合方法[8-11]等。

隐马尔可夫模型是Web信息抽取的一种重要统计方法。HMM易于建立,不需大规模的样本集,适应性好,抽取精度较高。但是HMM对初值十分敏感,在训练中极易得到局部最优模型参数;而且HMM没有考虑文本特征信息,尤其一阶HMM(HMM1)未考虑状态转移概率和观察值输出概率与模型历史状态的关联性,因而HMM混合方法得到了广泛关注和研究。文献[8]提出一种基于最大熵的HMM信息抽取方法,将每个观察文本单元所有特征加权之和用来调整HMM中的转移概率参数;文献[9]结合二元HMM与SVM来实现元数据的抽取;文献[10]提出了基于改进的粒子群优化算法的HMM参数优化模型,用于Web信息抽取;文献[11]用遗传算法(Genetic Algorithm, GA)优化一阶HMM参数实现Web信息抽取,得到了优于传统HMM的抽取效果,但是其GA运算只在HMM训练的前端作初始化调整,而且采用的是一阶HMM,也没有考虑HMM的特征利用问题,所以在抽取性能上仍有很大改进空间。

本文把遗传算法与二阶HMM(HMM2)相结合,提出了一种改进的内嵌式GA-HMM2模型。将GA的最优和次优

到稿日期:2011-04-30 返修日期:2011-08-17 本文受国家自然科学基金(60775041),山西省高校科技开发项目(20101120)资助。
李 荣(1974—),女,硕士,副教授,主要研究方向为人工智能和信息处理等,E-mail:lirong_1217@126.com;胡志军(1974—),男,副教授,主要研究方向为人工智能;郑家恒(1948—),女,教授,主要研究方向为人工智能和中文信息处理。

染色体均保留下来，以调整 Baum-Welch 算法的初始参数，训练中多次使用 GA 微调二阶 HMM 参数，实现全局最优，减少了迭代次数，解决了前面所述 HMM 的第一个问题；同时结合 Web 信息的特点，利用分隔符、换行符、行首字符等格式信息和文本特征，将文本切分成文本行、块或单个的词等恰当的层次，这种基于文本行的分层抽取与 HMM2 相结合，解决了上述 HMM 的第二个问题；最后用改进的 Viterbi 算法[12]来实现 Web 信息抽取。实验结果表明，新方法在精确度、召回率指标和效率上均比 GA-HMM 模型[11]具有更好的性能。

2 基于 GA-HMM 的 Web 信息抽取

2.1 HMM 模型

一个 HMM 包含两层：可观察层和隐藏层，可观察层是待识别的观察序列，隐藏层是一个马尔可夫过程，其中每个状态转移都带有转移概率。一个 HMM 可记为一个五元组(S, O, A, B, Π)，其中：①S 是状态集，模型共有 N 个状态，记为 $S=\{S_1, S_2, \cdots, S_N\}$；②$O$ 是符号集，共有 M 个可能输出的符号，记为 $O=\{O_1, O_2, \cdots O_M\}$；③$A$ 为状态转移概率矩阵，记为 $A=\{a_{ij}=p(q_{t+1}=S_j \mid q_t=S_i), 1\leqslant i, j\leqslant N\}$；④$B$ 为观察值输出概率分布矩阵，记为 $B=\{b_j(O_k)=P(O_k \text{ at } t \mid q_t=S_j), 1\leqslant i, j\leqslant N, 1\leqslant k\leqslant M\}$；⑤$\Pi$ 为初始状态分布向量，记为 $\Pi=\{\pi_i=P(q_1=S_i), 1\leqslant i\leqslant N\}$，$\pi_i=P(q_1=S_i)$。应用 HMM 模型，主要解决 3 个方面的问题：评估问题、学习问题和解码问题。

2.2 基于 HMM 的 Web 信息抽取

为讨论 HMM 在信息抽取中的应用，本文以科研论文头部作为对象予以研究，将一篇论文头部中的标题、作者、地址等文本信息抽取出来。应用 HMM 模型进行科研论文头部信息的抽取时，每个状态和要抽取的一个域相关联，这些域可以是标题、作者、单位、摘要等十多项，则模型隐藏层是由标题、作者等状态组成的有限状态机模型，可观察层是待抽取的论文头部文本序列。需要从可观察到的科研论文头部，即那些词序列或语义块序列来确定状态序列，也就是为这些头部语义块标记相关联的域。这里需要解决 HMM 模型中的学习问题和解码问题。进行 Web 信息抽取时，一般采用最大似然算法对已标记训练样本集或 Baum-Welch 算法对未标记训练样本集解决 HMM 模型中的学习问题，得出 HMM 模型参数。然后采用 Viterbi 算法给待抽取的输入 Web 信息找出最大概率的状态标签序列，被标记为目标状态标签的观察 Web 信息即为信息抽取的内容。

2.3 基于 GA-HMM 的 Web 信息抽取

GA 在搜索全局最优上具有突出的优势。文献[11]采用实数矩阵编码的方法，将 A、B、Π 分别用实数矩阵表示，将编码后的参数序列连成基因串，构成一个个体(即染色体)；采用 $P(O|\lambda)$作为适应度函数；选择算子采用比例选择法；交叉算子采用算术交叉；变异算子采用单点随机变异。

基于 GA-HMM 的信息抽取过程为：首先在 HMM 初始参数的选择中采用 GA，以减小 HMM 进入局部极小的概率；然后利用修正的 Baum-Welch 算法进行 HMM 的训练，构建 HMM；最后利用 Viterbi 算法进行测试样本最佳状态序列的求解。

3 GA-HMM 模型在 Web 信息抽取中的改进

针对 GA-HMM 模型在 Web 信息抽取中的实际应用，对其做了如下改进。

3.1 信息的分层抽取

目前文献中的信息抽取模型，多以单词或块为基本抽取单位。基于块的方法在针对一个抽取域对应一个文本块的情况下，上下文特征信息的利用率明显高于基于词的方法。但是，抽取效果仍不是很好。若采用基于文本行来分层抽取，能增加抽取单位在粒度上的灵活性，最大限度地挖掘和整合上下文特征信息，提高抽取性能。

基于文本行的分层信息抽取是以文本行为默认抽取单位，但以下两种情况需灵活处理。

(1)针对多文本块组成一个待抽取域的情况，需要根据版面格式信息，把开头不为空格的行与前一个文本行合并。如单词数较多的抽取域〈address〉、〈title〉、〈abstract〉、〈intro〉、〈note〉等，需要把同属一个抽取域的多个文本块合并为一个整体，即横向合并，使得大量属于同一个域的文本块和单词只需要进行一次标注，这样不但提高了效率，而且降低了错误抽取率。

(2)针对有的文本块中包含复合抽取域的情况，还需用空格符分割，按词抽取；结合文本和格式信息，利用部分确定性特征，用分隔符把文本行切分成文本块，甚至还需用空格符分割，切分成单词。确定性特征描述了文本自身特征和上下文信息特征，如是否以冠词打头、是否全为数字、是否含有“@”字符〈email〉、是否人名〈author〉、是否国名地名〈address〉、是否月份或其缩写〈date〉、是否以空格打头、首字符是否大写、是否包含“college”、“institute”、“university”等单词〈affiliation〉。

下面以一篇 Web 论文头部为例来说明分层信息抽取的思想：

Learning Hidden Markov Model Structure for Information Extraction

Kristie Seymore kseymore@ri. cmu. edu

Aindrew McCallum mccallum@justresearch. com

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213

可以看到，该论文头部可以划分为 6 个文本行、7 个文本块，一共包含了 5 个抽取域。其中，第 4、5 和 6、7 块是不同的文本块，却分别属于同一个抽取域〈affiliation〉和〈address〉；而第 2、3 个文本块，都包含了〈author〉和〈email〉两个抽取域。所以，单纯以块为抽取单位，仍然不够准确。若采用基于文本行的分层抽取，可以发现第 2、6 个文本行不是以空格开头，均与上一文本行同属一个抽取域，需要合并形成大的文本行。则新形成的论文头部包含 4 个文本行，新头部的第 2、3 个文本行中均含有〈author〉和〈email〉两个域的特征，应该细化到单词一级上对其进行分割和抽取，从而得到包含在同一文本块中的两个不同的域；再分析新头部的第 4 个文本行，利用前面所说的部分确定性特征，可以发现同时含有〈affiliation〉和〈address〉两个域的特征，因此，可以将第 4 行利用分隔符分

割为新头部的第4、5、6、7文本块，根据特征第4、5文本块和第6、7文本块需要再进行合并，从而分层抽取到5个域。

3.2 二阶 HMM(HMM2)的确定

HMM1假设状态转移概率和观察值输出概率仅依赖于模型当前的状态，这种HMM1假设的不合理性一定程度上降低了信息抽取的精确度，而HMM2合理地考虑了概率和模型历史状态的关联性，可以很好地捕捉状态之间存在的上下文特征信息，对错误有更强的识别能力。HMM2与HMM1的区别在于两个重要假设：

假设1　在HMM2中，隐藏的状态序列是一个二阶Markov链，即在$t+1$时刻的状态的转移概率不仅依赖t时刻的状态q_t，同时依赖于$t-1$时刻的状态q_{t-1}。

假设2　在t时刻释放观察值O_k的输出概率，不仅依赖于系统当前所处的状态S_j，同时依赖于系统前一时刻所处的状态S_i。

HMM2可以看成一个七元组$(S,O,A_1,A_2,B_1,B_2,\Pi)$；其中，$S$、$O$、$A_1$、$B_1$、$\Pi$与HMM1相同，$A_2=(a_{ijl})_{N\times N\times N}$，$B_2=(b_{ij}(O_l))_{N\times N\times M}$，其中，$a_{ijl}=P(q_{t+1}=S_l\mid q_t=S_j,q_{t-1}=S_i)$，$1\leqslant i,j,k\leqslant N$，$b_{ij}(O_l)=P(O_t=O_l\mid q_t=S_j,q_{t-1}=S_i)$，$1\leqslant i,j\leqslant N$，$1\leqslant l\leqslant M$。

上面例中Web论文头部信息包含〈title〉、〈author〉、〈email〉、〈affiliation〉、〈address〉共5个状态，经过分层预处理后，得到9个观察值，其对应HMM的状态序列、观察值序列在时间序列上的关系如表1所列。由此，构建的HMM2如图1所示。与HMM2相比，HMM1示意图只是缺少形如图1中S_1向S_3的相隔状态的转移曲线。

表1　论文头部信息的HMM状态转移及观察值输出关系表

时刻	1	2	3	4	5	6	7	8	9
状态	S_1	S_2	S_3	S_2	S_3	S_4	S_4	S_5	S_5
观察	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9

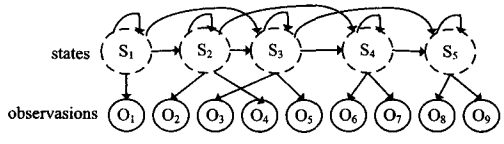

图1　HMM2示意图

现以计算机科研论文头部信息抽取为背景，以观察值O_6为例，分析比较HMM1和HMM2抽取信息的正确性。假设O_6可能存在以下两种单词组合：O_6^1="School of Computer Science"；O_6^2="School of Foreign Language"，对于计算机科研论文，显然，O_6^2不能作为〈affiliation〉(状态S_4)的正确输出，下面分析比较HMM1和HMM2对这种可能存在的错误信息的识别能力。

应用HMM1时，O_6^1和O_6^2的输出概率分别为：$b_4(O_6^1)=p(O_t=O_6^1\mid q_t=S_4)$，$b_4(O_6^2)=p(O_t=O_6^2\mid q_t=S_4)$；此时，$O_6^1$和$O_6^2$作为$S_4$的输出概率是均等的，所以"School of Foreign Language"将作为正确的输出信息被抽取出来，显然是错误的。应用HMM2时，$b_{34}(O_6^1)=p(O_t=O_6^1\mid q_t=S_4,q_{t-1}=S_3)$，$b_{34}(O_6^2)=p(O_t=O_6^2\mid q_t=S_4,q_{t-1}=S_3)$；显然，由于$O_6^2$作为$S_3$的输出可能性很小，$b_{34}(O_6^1)>b_{34}(O_6^2)$，因此，"School of Foreign Language"将作为正确的输出信息被抽取出来的可能性很小，使得抽取结果中错误信息更少。可见，HMM2在Web信息抽取中的有效性。

3.3 基于内嵌式GA-HMM2模型的参数估计

一般的GA-HMM混合模型均采用前端GA-HMM，即GA运算只在HMM的训练前端作初始化调整。本文采用内嵌GA-HMM2模型，多次使用GA微调HMM2，以保证Baum-Welch(B-W)训练算法跳出局部最优，使训练参数以较大概率、较快速度收敛于全局最优点。改进模型仍采用实数矩阵编码的方式，将编码后的HMM2参数序列构成一个染色体g_i，以$f(g_i)=\frac{1}{n}(\sum_{k=1}^{n}\log p(o_i\mid\lambda))$作为适应度函数，其中，$n$是观察序列$\{O^{(1)},O^{(2)},\cdots,O^{(n)}\}$的长度。然后，对每一代种群个体按其适应度排序，前10%的个体按"适者生存"被保留到下一代，其余90%则用轮盘赌选择父体进行交叉和变异运算，交叉运算使用单点交叉算子，进行变异运算时，先随机确定变异个体，然后将小片段上的基因作循环左移一位，以保证个体的合法性。

对于每个观察序列$O_k(1\leqslant k\leqslant M)$，传统方法只选取使得$f(g_k^*)$最大的最优染色体$g_k^*$进行训练。文中认为对于次优染色体$g_k^{**}$，若$(f(g_k^*)-f(g_k^{**}))/f(g_k^*)$的值大于某一门限值$\omega$，表示在这一代中$g_k^*$和$g_k^{**}$都是优良染色体，对于$O_k$都较为适应，简单地去除就是抛弃了部分有用的信息，这对于参数训练是有影响的。虽然次优染色体可能在若干代后重现，成为优良染色体，但是这将降低算法的收敛速度。文中将次优染色体也保留，进行参数训练。令q_k^t表示g_k^*的第t位对应的状态值，保留的g_k^{**}共有N个，q_k^{t*}表示g_k^{**}的第t位对应的状态值，定义算子$R(a)=\begin{cases}1, & a=\text{true}\\ 0, & a=\text{false}\end{cases}$，则参数$\lambda=(A_1,A_2,B_1,B_2,\Pi)$的迭代公式修正如下：

$$\hat{\pi}=\frac{\sum_{k=1}^{M}R(q_k^1=s_i)+\sum_{N}R(q_k^{1*}=s_i)}{M+N}\tag{1}$$

$$\hat{a}_{ij}=\frac{\sum_{k=1}^{M}R(q_k^t=s_i,q_k^{t+1}=s_j)+\sum_{N}R(q_k^{t*}=s_i,q_k^{(t+1)*}=s_j)}{\sum_{k=1}^{M}R(q_k^t=s_i)+\sum_{N}R(q_k^{t*}=s_i)}\tag{2}$$

$$\hat{b}_{ij}=\frac{\sum_{k=1}^{M}R(q_k^t=s_i,o_k^t=o_j)+\sum_{N}R(q_k^{t*}=s_i,o_k^{t*}=o_j)}{\sum_{k=1}^{M}R(q_k^t=s_i)+\sum_{N}R(q_k^{t*}=s_i)}\tag{3}$$

$$\hat{a}_{ijl}=\frac{\sum_{k=1}^{M}R(q_k^{t-1}=s_i,q_k^t=s_j,q_k^{t+1}=s_l)+\sum_{N}R(q_k^{(t-1)*}=s_i,q_k^{t*}=s_j,q_k^{(t+1)*}=s_l)}{\sum_{k=1}^{M}R(q_k^{t-1}=s_i,q_k^t=s_j)+\sum_{N}R(q_k^{(t-1)*}=s_i,q_k^{t*}=s_j)}\tag{4}$$

$$\hat{b}_{ijl}=\frac{\sum_{k=1}^{M}R(q_k^{t-1}=s_i,q_k^t=s_j,o_k^t=o_l)+\sum_{N}R(q_k^{(t-1)*}=s_i,q_k^{t*}=s_j,o_k^{t*}=o_l)}{\sum_{k=1}^{M}R(q_k^{t-1}=s_i,q_k^t=s_j)+\sum_{N}R(q_k^{(t-1)*}=s_i,q_k^{t*}=s_j)}\tag{5}$$

内嵌GA-HMM2模型的训练算法如下：

Step 1　随机初始化HMM2的模型参数λ_0，将其编码为染色体，作为GA的初始人口；

Step 2　对个体按适应度函数排序，选择排名在后90%的个体，用"车轮"选择法选取父辈到培育池；

Step 3　作父辈的交叉和变异运算；

Step 4　对于每一个观察序列 O_k，选取该代中的最优染色体 g_k^* 和次优染色体 g_k^{**}，若$(f(g_k^*)-f(g_k^{**}))/f(g_k^*)>\omega$，则将 g_k^{**} 保留。将训练集带入式(1)－式(5)，进行参数训练，得到子代 λ_0'；

Step 5　将子代 λ_0'作为 Baum-Welch 训练初始参数，作迭代次数为 3 次的 B-W 训练；

Step 6　根据适应度的大小来更新人口；

Step 7　遗传 20 代，全体人口作迭代次数为 20 次的 B-W 训练；

Step 8　重复 3 次 Step 2－Step 7，得到最终 HMM2 的模型参数。

4　基于 GA-HMM2 的信息抽取过程及实验结果

本文采用改进的内嵌式 GA-HMM2 混合模型，以网上收集的论文头部作为实验数据，进行不同域的抽取。混合模型的信息抽取过程如下：

(1)网页预处理：对收集的 Web 页进行结构分析，建立相应的 HTML 结构树。然后估计其中每一个内部节点的熵，定位包含数据记录的数据域，将找到的记录转换成数据段序列，以提取单条论文头部。

(2)信息分层处理：采用基于文本行的分层抽取方法，根据排版信息和特征信息将头部文本切分成文本行、块或单个词等恰当的层次。

(3)模型训练：采用内嵌式 GA-HMM2 重估 Baum-Welch 参数，用 GA 多次微调 HMM2 参数，并保留次优染色体，修正 B-W 迭代公式，从而减少迭代次数，使 B-W 算法跳出局部最优解。

(4)抽取信息：将网页预处理和分层后的实验文本，采用 HMM2 中改进的 Viterbi 算法输出最佳状态序列。

表 2　3 种方法的抽取域精确度和召回率比较

State	HMM		GA-HMM		本文方法	
	Precision	Recall	Precision	Recall	Precision	Recall
Title	0.465851	0.753084	0.706492	0.876554	0.856101	0.891679
Author	0.693954	0.696533	0.742991	0.892484	0.898551	0.906533
Affiliation	0.733176	0.813294	0.757374	0.915298	0.948264	0.917692
Address	0.603797	0.800010	0.723905	0.854124	0.906123	0.855108
Email	0.860298	0.620807	0.880087	1.000000	0.959547	1.000000
Note	0.774795	0.666338	0.782381	0.716479	0.901789	0.696424
Web	0.902169	0.261619	0.924587	0.547015	0.978457	0.547619
Phone	0.941667	0.406889	0.946309	0.913801	0.906667	0.913580
Date	0.636676	0.800945	0.868138	0.909109	0.908718	0.930909
Abstract	0.814545	0.989916	0.880089	0.999895	0.934545	0.998856
Intro	0.833418	0.982469	0.873418	0.9794329	0.887543	0.979956
Keyword	0.823279	0.987395	0.823279	0.937395	0.867279	0.937395
Degree	0.128821	0.838235	0.517651	0.813047	0.723492	0.813795
Pubnum	0.847641	0.483939	0.847177	0.703596	0.906091	0.783427
Page	0.119617	0.731600	0.801474	0.853192	0.820217	0.891374

表 3　3 种方法的平均 $F_{\beta=1}$值比较

	HMM	GA-HMM	本文算法
$F_{\beta=1}$值	0.639311142	0.8239557	0.8763079

表 4　3 种方法的时间性能比较

	HMM	GA-HMM	本文算法
t/s	13.46	17.25	13.57

实验采用 2480 篇未标注的论文头部作为训练集，共包含单词 54912 个；其中 200 篇作为测试样本，包含单词 4078 个。在训练中，GA 主要参数为：群体规模 $M=200$，迭代次数 $T=60$，交叉率 $P_c=0.2$，变异率 $P_m=0.01$。保留次优染色体能减少迭代次数，当 $\omega=0.1$ 时，收敛所用迭代次数最少。实验中，状态数 $N=15$。

实验分别对基于 HMM、基于 GA-HMM 和本文的混合方法进行了信息抽取比较，比较结果如表 2－表 4 所列。

实验中采用的训练数据是未标记的，HMM 初始参数只能随机产生，而训练算法 Baum-Welch 对迭代的初值较为敏感，易陷入局部最小。从表 2 可以看出，HMM 方法实验结果不太理想。基于 GA-HMM 的实验效果较 HMM 有所改改善，各个域的精确率和召回率都有所提高，体现了 GA 全局寻优的特性。而采用本文方法的抽取精确度和召回率总体上又比传统 GA-HMM 方法的精确度和召回率高。尤其是各状态的精确率平均提高约 9%，是因为改进算法使用了二阶 HMM 和文本特征信息，它们能提高对错误信息的识别能力。由于采用信息分层抽取的思想，〈title〉的精确率和召回率分别提高了约 15%和 2%，而〈Keyword〉、〈degree〉状态的召回率没有提高，主要是因为这两个状态内单词的内聚性不是很强。从表 3 可见，在综合性指标 $F_{\beta=1}$值的比较上，本文算法的平均 $F_{\beta=1}$值比前两种方法分别提高了 24%和 5%。从表 4 可见，在抽取时间性能上，本文算法比 HMM 所用时间稍长，主要是因为本文算法采用的是 HMM2，而与 GA-HMM 的时间相比，具有非常明显的优势，这体现了改进算法的有效性和参数寻优过程的全局性和快速性。

结束语　提出了一种内嵌式 GA-HMM2 混合模型的 Web 信息抽取方法。信息的分层预处理，提高了抽取单位在粒度上的灵活性；考虑到概率和模型历史状态的关联，构建了二阶 HMM，并结合文本特征，提高了抽取准确率；然后使用 GA 微调二阶 HMM，优化 B-W 参数，并改进了优良染色体的选择策略，这样不但克服了 B-W 算法过早陷入局部最优和 GA 算法收敛速度缓慢的缺点，而且减少了迭代次数，提高了效率。实验结果表明，该方法能迅速在全局范围内找到一个最优解，证明了改进方法的有效性。后续研究可以着眼于：(1)结合模拟退火算法，解决遗传算法的早熟问题；(2)让系统自动从候选集中筛选特征，制定更加准确的特征函数。

参 考 文 献

[1]　冀高峰，汤庸，道炜，等. 基于 XML 的自动学习 Web 信息抽取[J]. 计算机科学，2008，35(3)：87-90

[2]　Skounakis M, Craven M, Ray S. Hierarchical hidden markov models for information extraction[C]//Proceedings of the 18th International Joint Conference on Artificial Intelligence. Acapulco, Mexico: Morgan Kaufmann, 2003: 427-433

[3]　祝伟华，卢熠，刘斌斌，等. 基于 HMM 的 Web 信息抽取算法的研究与应用[J]. 计算机科学，2010，37(2)：203-205

[4]　韦小丽，孙涌，张书奎，等. 基于最大熵模型的本体概念获取方法[J]. 计算机工程，2009，35(24)：114-116

(下转第 215 页)

$$(x)_1^F=\{x_3,x_5,x_6,x_{10}\} \tag{37}$$

利用第2节中的式(1)与式(3)得到属性集合 α_2^F 和具有属性集合 α_2^F 的信息 $(x)_2^F$，它们分别是

$$\alpha_2^F=\alpha_1^F\cup\{\beta_2\}=(\alpha\cup\{\beta_1\})\cup\{\beta_2\}=\{\alpha_1,\alpha_2,\alpha_3,\beta_1,\beta_2\} \tag{38}$$

$$(x)_2^F=\{x_6\} \tag{39}$$

显然，$(x)_1^F$ 是来自上海市的学生 x_3,x_5,x_6,x_{10} 构成的信息；$(x)_2^F$ 是来自上海市虹桥区的学生 x_6 构成的信息。由式(21)得到：$(x)_1^F,(x)_2^F$ 是 (x) 的内收敛信息；式(37)，式(39)给出的信息 $(x)_1^F,(x)_2^F$ 满足第2节中的内P-推理式(15)；或者，

if $\alpha\Rightarrow\alpha_1^F$, then $(x)_1^F\Rightarrow(x)$

if $\alpha_1^F\Rightarrow\alpha_2^F$, then $(x)_2^F\Rightarrow(x)_1^F$

事实上，利用式(34)，式(35)分别得到式(37)，式(39)；获取式(37)，式(39)的过程中包含着一个推理过程。这个推理过程很早就被人们使用，却没有人对这个推理给出理论认识与理论研究(在文献[3]发表之前)。式(37)与式(34)，式(39)与式(34)分别满足式(31)与内收敛信息的存在原理；或者，$(x)_1^F,(x)_2^F$ 在 (x) 内被发现，在 (x) 内被辨识是由内P-推理得到的。$(x)_1^F,(x)_2^F$ 的内收敛系数 ρ_1^F,ρ_2^F 满足定理3与推论4。具有属性集合 α_2^F 的信息 $(x)_2^F$ 是 (x) 的内收敛信息核(式(22))，$(x)_2^F=(x)^{F,*}$。信息 $(x)_1^F,(x)_2^F$ 的获取-辨识满足第3节中的内收敛信息的存在原理。

在未对属性集合 α 内给予补充属性 β_1,β_2 之前，内收敛信息 $(x)_1^F,(x)_2^F$ 是未知的，它们不被人们事先知道。

作者选择了42名大学二年级的学生，利用这个例子给出的方法，得到了很好的结果，得到的结果与实际相符合。

结束语 P-集合是文献[1,2]中提出的一个新的数学模型，P-推理是利用P-集合提出的一个新的推理结构。P-集合、P-推理的一个重要特性与优点是：P-集合、P-推理具有动态特征，它们的动态特征与动态信息系统的动态特征相同。因此，P-集合、P-推理能够被应用到动态信息研究中[4-20]。本文给出的讨论与结果还能被应用到动态信息系统中的其它应用领域中，或许人们从本文给出的研究中获得这样的结论：P-集合与P-推理是研究动态信息系统的一个新模型、新方法。

参 考 文 献

[1] 史开泉. P-集合[J]. 山东大学学报:理学版,2008,43(11):77-84

[2] Shi Kai-quan. P-sets and its applications[J]. An International Journal Advances in Systems Science and Applications,2009,9(2):209-219

[3] 史开泉. P-推理与信息的P-推理发现-辨识[J]. 计算机科学,2011,38(7):1-9

[4] 史开泉. P-集合与它的应用特征[J]. 计算机科学,2010,37(8):1-10

[5] Shi Kai-quan, Li Xiu-hong. Camouflage information identification and its application[J]. An International Journal Advances in Systems Science and Applications,2010,10(2):157-167

[6] Zhang Li,Cui Yu-quan,Shi Kai-quan. Outer P-sets and data internal recovery[J]. An International Journal Advances in Systems Science and Applications,2010,10(2):189-199

[7] Zhang Li,Xiu Ming,Shi Kai-quan. P-sets and applications of internal-outer data circle[J]. Quantitative Logic and Soft Computing,2010,1(2):581-592

[8] Li Yu-ying,Zhang Li,Shi Kai-quan. Generation and recovery of compressed data and redundant data[J]. Quantitative Logic and Soft Computing,2010,1(2):661-672

[9] Xiu Ming,Shi Kai-quan,Zhang Li. P-sets and $\overline{F}$-data selection-discovery[J]. Quantitative Logic and Soft Computing,2010,1(2):791-800

[10] 张丽,崔玉泉,史开泉. 外P-集合与数据内-发现[J]. 系统工程与电子技术,2010,32(6):1233-1238

[11] 张飞,陈萍,张丽. P-集合的P-分离与应用[J]. 山东大学学报:理学版,2010,45(3):18-22

[12] 史开泉,张丽. 内P-集合与数据外恢复[J]. 山东大学学报:理学版,2009,44(4):8-14

[13] 周玉华,张冠宇,张丽. 内外数据圆与动态数据发现[J]. 山东大学学报:理学版,2010,45(8):21-26

[14] Li Hong-kang,Li Yu-ying. P-sets and its P-separation theorems[J]. An International Journal Advances in Systems Science and Applications,2010,10(2):209-215

[15] 周玉华,张冠宇,史开泉. P-集合与双信息规律生成[J]. 数学的实践与认识,2010,40(13):71-80

[16] 于秀清. P-集合的识别与筛选[J]. 山东大学学报:理学版,2010,45(1):94-98

[17] 汪洋,张冠宇,史开泉. P-集合的F-记忆信息特征-应用[J]. 计算机科学,2011,38(2):246-249,266

[18] 耿红芹,张冠宇,史开泉. F-信息伪装与伪装还原-辨识[J]. 计算机科学,2011,38(2):241-245,256

[19] Liu Ji-qin. P-probabilities and its applications[J]. An International Journal of Advances in systems Science and Applications,2010,10(2):200-208

[20] 汤积华,陈保会,史开泉. P-集合与$(\overline{F},F)$-数据生成-辨识[J]. 山东大学学报:理学版,2009,44(11):83-92

(上接第199页)

[5] Freitag D, McCallum A, Pereira F. Maximum Entropy Markov models for information extraction and segmentation[C]// Proceedings of the Seven teenth International Conference on Machine Learning. San Francisco:Morgan Kaufmann,2000:591-598

[6] Bundschus M, Dejoril M, Stetter M, et al. Extraction of semantic biomedical relations from text using conditional random fields [J]. BioMed Central(BMC)Bioinformatics,2008,9:207-220

[7] Martens D, Baesens B, et al. Decompositional Rule Extraction from Support Vector Machines by Active Learning[J]. Knowledge and Data Engineering,2008,21(2):178-191

[8] 林亚平,刘云中,周顺先,等. 基于最大熵的隐马尔可夫模型文本信息抽取[J]. 电子学报,2005,33(2):236-240

[9] 张铭,银平,邓志鸿,等. SVM+BiHMM:基于统计方法的元数据抽取混合模型[J]. 软件学报,2008,19(2):358-367

[10] 王川,段德全,王晓东. 基于改进的PSO和HMM的Web信息抽取算法[J]. 河南师范大学学报:自然科学版,2010,38(5):65-68

[11] 肖基毅,邹腊梅,李伟琦. 混合遗传算法和隐马尔可夫模型的Web信息抽取[J]. 计算机工程与应用,2008,44(18):132-135

[12] 刘洁彬,宋茂强,赵方,等. 基于上下文的二阶隐马尔可夫模型[J]. 计算机工程,2010,36(10):231-232