文章编号:1673 - 0062(2011)01 - 0070 - 05

基于模拟退火算法与隐马尔可夫模型的 Web 信息抽取

邹腊梅[1],龚向坚[1],肖 芳[2],马淑萍[1]

(1. 南华大学 计算机科学与技术学院,湖南 衡阳 421001; 2. 衡阳技师学院 信息技术系,湖南 衡阳 421007)

摘 要:典型隐马尔可夫模型对初始参数非常敏感,采用随机参数训练隐马尔可夫模型时常陷入局部最优,应用于 Web 信息抽取时效果不佳. 文中提出基于模拟退火算法与隐马尔可夫模型的 Web 信息抽取算法. 通过实验比较选择最佳的模拟退火算法参数,结合 Baum-Welch 算法优化隐马尔可夫模型并应用于 Web 信息抽取. 实验结果表明新算法在信息抽取的精确率和召回率都有明显的提高.

关键词:模拟退火算法;隐马尔可夫模型;Web 信息抽取

中图分类号:TP391.1 **文献标识码**:B

Web Information Extraction Based on Simulated Annealing Algorithm and Hidden Markov Model

ZOU La-mei[1], GONG Xiang-jian[1], XIAO Fang[2], MA Shu-ping[1]

(1. School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China;
2. Department of Information Technology, Hengyang Technician College, Hengyang, Hunan 421007, China)

Abstract: Typical HMM is sensitive to the initial model parameters and often leads to sub-optimal when training it with random parameters. It is ineffective when extracting Web information with typical HMM. The artical proposes web information extraction algorithm based on SA and HMM. The algorithm chooses the best SA parameters by experiment and optimizes HMM combining Baum-Welch during the course of extracting Web information. Experimental results show that the new algorithm significantly improves the performance in precision and recall.

key words: simulated annealing algorithm; hidden Markov model; Web information extraction

随着计算机技术的发展,不同领域均产生和 存储了大量的文本数据,因特网的发展也导致全

收稿日期:2010 - 12 - 20

基金项目:湖南省教育厅基金资助项目(07C637)

作者简介:邹腊梅(1977 -),女,湖南衡阳人,南华大学计算机科学与技术学院讲师,硕士. 主要研究方向:计算机网络、数据挖掘、信息检索.

球网页数惊人的增长,而网页上 80% 的内容都是文本信息. 怎样快速、准确从这些网络文本中找出自己需要的信息成为人们急需解决的难题,目前隐马尔可夫模型(Hidden Markov Model,HMM)在 Web 信息抽取方法中占重要地位[1-3]. 利用隐马尔可夫模型进行 Web 信息抽取是一种基于统计学习理论的信息抽取方法. 由于 Web 信息是未标记的数据,在训练构建隐马尔可夫模型时常使用 Baum-Welch 算法,采用 Viterbi 算法将待抽取 Web 文本中人们需要的特定的信息标注出来. 在利用未标记数据集进行 HMM 训练时,通常采用随机设定初始参数的方法. 但 Baum-Welch 算法本身是一种局部搜索算法,对参数的初值十分敏感,因此 HMM 的训练容易陷入局部极值而得不到最优模型. 模拟退火算法是模拟热力学中物理淬火过程的一种学习规则,该算法既能向目标函数优化的方向迭代,又能以一定的概率接受目标函数劣化的情况,从而避免了陷入局部最优点,保证获得全局最优解的可靠性,而且模拟退火算法在手写体数字识别、非数值并行计算、环境质量评估等方面已经取得成功[4-7],因此本文将模拟退火算法引入 HMM 的训练中,提出基于模拟退火算法与隐马尔可夫模型的 Web 信息抽取算法,通过实验选择最佳的模拟退火算法参数,结合 Baum-Welch 算法优化隐马尔可夫模型并应用于 Web 信息抽取,从而提高 Web 信息抽取的准确率和召回率.

1　模拟退火算法

模拟退火算法(Simulated Annealing Algorithm,简称为 SA 算法)[8] 是一个全局最优化算法. 算法的思想最先由 Metropolis 在 1953 年提出. SA 是模拟热力学中物理淬火过程一种学习规则,在某一初始温度下,伴随温度参数的不断下降,结合概率突跳特性在解空间中随机寻找目标函数的全局优解,该算法既能向目标函数优化的方向迭代,又能以一定的概率接受目标函数劣化的情况,即在局部最优解中能概率性地跳出并最终趋于全局最优,从而避免了陷入局部最优点[9].

SA 具体步骤描述如下:

(1)随机产生一个初始最优点,以它作为当前最优点,并计算目标函数值;

(2)设置初始温度:$\theta \leftarrow T_0$;

(3)设置循环计数器初值:$k \leftarrow 1$;

(4)对当前最优点作随机扰动,产生一个新的最优点,计算新的目标函数值,并计算目标函数值的增量 Δ;

(5)如果 $\Delta < 0$,则接受该新产生的最优点作为当前最优点;如果 $\Delta \geqslant 0$,则以概率 $p = \exp(-\Delta/\theta)$ 接受该新产生的最优点为当前最优点;

(6)如果 $k <$ 终止步数,则 $k \leftarrow k+1$,转向第(4)步;

(7)按照 $T(k+1) = \lambda \times T(k)$(λ 为正、略小于 1.00 的常数,k 为降温的次数)降低控制温度 $T(k+1)$,如果未到达冷却状态,则 $\theta \leftarrow T(k+1)$,转第(4)步;否则转(8);

(8)当前解作为最优解输出.

2　基于隐马尔可夫模型(HMM)的 Web 信息抽取

HMM 为一个五元组[10]:$\lambda = (S, O, A, B, \pi)$:$S$ 表示模型中状态集合,共 N 个状态. O 表示模型中输出观察值集,每个状态上对应的可能的观察值的数目为 M; $A = \{a_{ij}\}$ 为状态转移概率矩阵,表示从状态 i 转移到状态 j 的概率;$B = \{b_j(k)\}$ 为输出观察值概率分布矩阵,表示在 S_j 状态下,t 时刻出现 w_k 的概率;$\pi = \{\pi_i\}$ 为初始状态分布向量,表示 $t = 1$ 时处于状态 s_i 的概率. 建立隐马尔可夫模型需要解决的三个问题及相应算法为[11-12]:

1)评估问题:对于给定模型 λ 和某个观察值序列 $O = (O_1, O_2, \cdots, O_T))$,如何求概率 $P(O|\lambda)$. 常用算法有前向算法和后向算法.

2)学习问题(训练问题):对于给定的观察值序列 $O = (O_1, O_2, \cdots, O_T)$,调整参数 λ,使得观察值出现的概率 $P(O|\lambda)$ 最大. 常用算法有最大似然(Maximum Likelihood ML)算法和 Baum-Welch 算法

3)解码问题:对于给定模型 λ 和观察值序列 $O = (O_1, O_2, \cdots, O_T)$,求可能性最大的状态序列. 常用算法有 Viterbi 算法.

HMM 在自然语言理解中主要应用在词性标注、词语切分方面. 本文将 HMM 应用于 Web 信息抽取的目的是从众多的 Web 信息中抽取人们关注的特定的信息,如从网页中的论文列表中提取论文的作者姓名、出版社、发表时间、作者所在工作单位等信息. 在论文列表中存在大量不同的词、短语、句子,字典中所有可能出现的单词都有可能出现,而要找出某个单词潜在的意义(如属于作者、出版社、时间、工作单位等),这是一个费时费

力的工作,而隐马尔可夫模型提供了描述复杂现象的一种可能机制. 本文将论文列表看成是由多文本序列连在一起构成“词串”(或者是短语串,句子串),也就是可观察的符号序列,将作者、出版社、发表时间、作者所在工作单位等信息看成隐藏在文本背后的状态,“词串”序列所对应的隐藏的状态序列在标记前是隐藏的,是需要求解的目标序列. 因此基于 HMM 的 Web 信息抽取过程是可以简要定义为:利用 Web 信息集建立 HMM(利用 Baum-Welch 算法)和符号序列,利用 Viterbi 算法寻找能使产生该符号序列概率最大的状态序列,即完成 Web 信息特定信息的标注. 基于初始参数为随机数的 HMM 的 Web 信息抽取框架如图 1 所示.

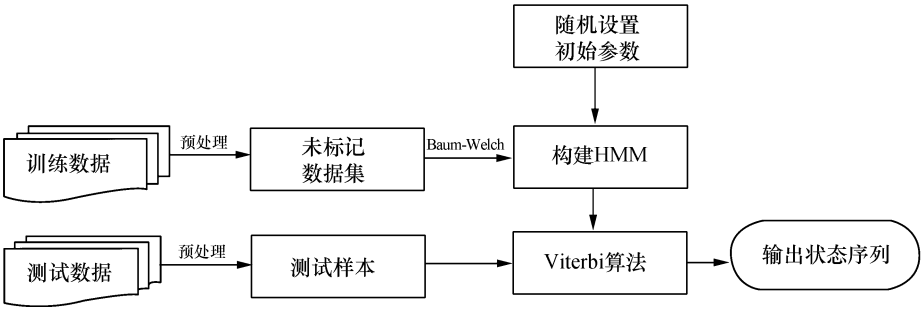

图 1 基于 HMM 的 Web 信息抽取框架

Fig. 1 Framework of Web information extraction based on HMM

3 基于 SA 与 HMM 的 Web 信息抽取

训练 HMM 常用算法有最大似然(Maximum Likelihood ML,针对已标注数据)算法和 Baum-Welch 算法(针对未标注数据),Web 页面中的数据为未标记数据,进行 HMM 训练时,只能随机设定初始参数,而 Baum-Welch 算法本身是一种局部搜索算法,对初值十分敏感,所以 HMM 的训练常陷入局部极值而得不到最优模型. SA 是一个全局最优化算法,它可以很好地解决 HMM 训练时对初值敏感的问题. 基于 SA-HMM 的 Web 信息抽取整体框架如图 2 所示.

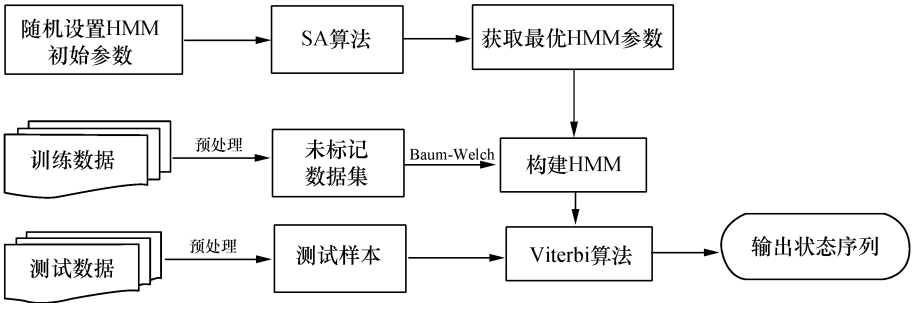

图 2 基于 SA_HMM 的 Web 信息抽取框架

Fig. 2 Framework of Web information extraction based on SA_HMM

抽取过程中 SA 参数设置:

1)初始温度 T_0 的设置. 温度 T 的初始值设置是影响模拟退火算法全局搜索性能的重要因素之一,初始温度太高会花费高昂的计算时间,太低会拒绝劣解的接受,会丢失 SA 全局优化的优点,初始温度一般需要依据实验结果进行若干次调整.

2)温度降低策略. 温度降低越快,陷入局部的概率就越大. 然而,温度降低太慢会导致算法速度慢得不能接受. 假定时刻 t 的温度 $T(k)$ 来表示,则经典的模拟退火算法的降温方式为:$T(k)=\frac{T_0}{\lg(1+k)}$;快速模拟退火算法的降温方式为:$T(k)=\frac{T_0}{1+k}$;考虑到计算复杂度,本文采用 $T(k+1)=\lambda\times T(k)$($\lambda$ 为略小于 1 的系数)方式进行降温管理.

3)能量函数. 能量函数也就是需要进行优化计算的目标函数,其最小点为所求的最优解,本文以 HMM 的评估函数 $P(O|\lambda)$ 作为优化计算的目标函数.

4)终止标准. SA 终止通常通过内外 2 层循环来控制,内循环是 Marcob 链的长度,外循环取某个温度 t 作为算法终止标准.

4　实验结果分析

本文以 Web 中的论文头部作为实验数据进行不同域的提取,首先需要收集若干包含论文头部的 Web,Web 中的论文头部常用格式如下:

J. Eckmann and C. Pillet and L. Rey-Bellet, "Entropy production in nonlinear thermally driven hamiltonian systems", Proceedings of the 3rd International Workshop on Inductive Logic Programming, CRC Press, May 1994.

为了提取单条论文头部,需要对收集的网页进行结构分析,建立相应的 HTML 结构树. 然后通过估计结构树中每一个内部节点的 Shannon 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. 本文使用从 www. jaist. ac. jp/ ~ hieuxuan/softwares/peweb 网站下载的 peweb 工具来完成 Web 中记录的提取. 通过对该论文记录的分析,确定该记录共包含了 5 个域,进一步分解为:

author = " J. Eckmann and C. Pillet and L. Rey-Bellet",

title = " Entropy production in nonlinear thermally driven hamiltonian systems",

booktitle = " Proceedings of the 3rd International Workshop on Inductive Logic Programming",

publisher = " CRC Press",

year = " May 1994",

实验采用 2000 篇未标注的论文头部作为训练集,共包含单词 49 651 个. 100 篇作为测试样本,共包含单词 2 437 个. 为了确定 SA 的初始参数,对初始温度、温度衰减、Mapkob 链长度取不同的值做实验,实验结果如表 1 所示.

表 1　训练 HMM 的 SA 参数比较

Table 1　Comparing of SA parameters training HMM

初始温度	2 000	2 000	2 000	1 500	1 500	1 500	1 000	1 000	1 000
温度衰减	0.9	0.6	0.4	0.9	0.6	0.4	0.9	0.6	0.4
Mapkob 链长度	500	1 000	1 500	500	1 000	1 500	500	1 000	1 500
$P(O\|\lambda)$	7.552 2	7.234 5	7.646 8	7.726 2	8.544 3	8.015 1	7.862 6	7.023 8	7.103 4

实验中设置的最高温度为 2000,在实验中,初始温度越高,衰减的足够慢,Mapkob 链长度足够长,越接近最优解. 在实验过程中,当衰减到一定程度后,Markov 链中解已无任何改变时可终止算法,已经产生了最优解. 实验证明当初始温度为 1500,衰减温度为 0.6,Mapkob 链长度为 1000 时,$P(O|\lambda)$ 达到了 8.5443,达到最优.

表 2 是基于随机 HMM 和基于 SA-HMM 的 Web 信息抽取实验结果比较.

表 2　基于 HMM 和基于 SA_HMM 的实验结果比较

Table 2　Comparing of experiment results based on HMM and based on SA_HMM

REC/PRE	基于 HMM		基于 SA-HMM	
域　名	REC	PRE	REC	PRE
author	0.842 65	0.864 53	0.893 21	0.949 3
title	0.738 63	0.461 31	0.817 43	0.792 18
booktitle	0.564 37	0.702 16	0.745 23	0.853 18
publisher	0.732 45	0.563 78	0.823 82	0.783 43
year	0.823 41	0.824 26	0.894 36	0.943 16
平均	0.740 302	0.683 208	0.834 81	0.864 25

从表 2 可以看出,对于未标记的训练数据集,Baum-Welch 本身是梯度下降的算法,对初始参数非常敏感,因此容易产生局部极小,信息抽取的精确率和召回率都不太高,特别在 title 域中,精确率仅为 0.461 31,booktitle 域中召回率也只有 0.564 37,实验结果不是很理想,平均精确率为 0.740 302,平均召回率为 0.683 208. 引入 SA 算法后,实验效果有了明显的改善,平均精确率为 0.834 81,平均召回率为 0.864 25在 author 和 year 的召回率接近 95%,实验结果体现了 SA 全局寻优的特点,同时也证明基于 SA-HMM 的 Web 信息抽取是有效的.

参考文献:

[1] Fabien Salzenstein, Wojciech Pieczynski. Parameter estimation in hidden fuzzy markov random fields and image segmentation[J]. Graphical Models and Image Processing, 1997, 59(4): 205-220.

[2] Xuan-Hieu Phan, Susumu Horiguchi, Tu-Bao Ho. Automated data extraction from the web with conditional models[J]. Int. J. Business Intelligence and Data Mining, 2005, 1(2): 210-228.

[3] Freitag D, McCallum A, Pereira F. Maximum entropy markov models for information extraction and segmentation[C]//processing of ICML, 2000, 1(1): 591-598.

[4] 贾德香,唐国庆,韩净. 基于改进模拟退火算法的电网无功优化[J]. 继电器,2004,32(4):32-36.

[5] 洪沛霖,张佑生,邢燕. 基于改进模拟退火算法的手写体数字识别[J]. 计算机技术与发展,2007,17(9):15-17.

[6] 吴月,刘忠明,刘永祺. 模拟退火算法在大气环境质量综合评价中的应用[J]. 四川环境,2008,27(3):53-55.

[7] 林慧君,彭宏. 模拟退火算法在全局查询优化中的应用[J]. 计算机技术与发展,2006,16(4):155-157.

[8] Metropolis N, Rosenbluth A. Rosenbluth metal, equation of state calculations by fast computing machines[J]. Journal of Chemical Physics, 1953, 56(21): 1087-1092.

[9] Kirkpatrick S, Jr Gelatt C D, Vecchi M P. Optimization by simulated annealing[J]. Science, 1983, 220(11): 650-671.

[10] Tobias Scheffer, Christian Decomain, Stefan Wrobel. Mining the web with active hidden markov models[C]//San Jose. Proceedings of the IEEE Internationl Conference on Data Mining. California: IEEE Computer Society, 2001: 309-318.

[11] Liu Lifang, Huo Hongwei, Wang Baoshu. A novel optimization of profile HMM by a hybrid genetic algorithm[C]//Vilanova Barcelona. IWANN 2005. Spain: Springer, 2005: 734-741.

[12] Liu Jianghua, Cheng Junshi, Chen Jiapin. Optimization of HMM Parameters based on chaos and genetic algorithm for hand gesture recognition[J]. Journal of Systems Engineering and Electronics, 2002, 3(4): 79-84.