

基于模拟退火算法和隐马尔可夫模型的文本信息抽取

王 宁, 李石君

(武汉大学 计算机学院, 湖北 武汉 430072)

摘 要: 本文提出了一种基于改进的模拟退火算法和二阶隐马尔可夫模型相结合的信息抽取方法. 其中, 改进的模拟退火算法在适当的时机增加了重升温过程, 提高了局部寻优的搜索效率; 二阶隐马尔可夫模型充分考虑了概率与历史状态的关联性, 增加了信息抽取的可靠性. 实验结果表明, 新算法在精确度、召回率和时间性能指标上比基于模拟退火算法和一阶隐马尔可夫模型的信息抽取方法有所提高.

关键词: 文本信息抽取; 改进的模拟退火算法; 二阶隐马尔可夫模型

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2014)12-0052-05

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

WANG Ning, LI Shi-jun

(Computer School, WuHan University, Wuhan 430072, China)

Abstract: An improved combined method embedded with modified simulated annealing algorithm and second-order hidden Markov model was presented. The modified simulated annealing algorithm improves the temperature at the proper time, which increased the efficiency of local search optimization. The second-order hidden Markov model considers the relevance of probability and historical status which improved the reliability of information extraction. Experimental results show that the new method has a better improvement in precision, recall and time performance than the method combined with simulated annealing algorithm and first-order hidden Markov model.

Key words: text information extraction; modified simulated annealing algorithm; second-order hidden markov model

1 引言

信息抽取 (Information Extraction, IE) 是指从文本中自动抽取相关的或指定类型的信息, 并将其以结构化的数据形式存入数据库中以供用户查询和使用的过程. 利用隐马尔可夫模型 (HMM) 进行信息抽取是一种基于统计模型的信息抽取方法. 它因构建简单、适应能力强、抽取精确度高、对词典集和规则集的要求低等优点而逐渐受到研究者的关注. 现在很多基于 HMM 模型的改进算法在文本信息抽取中都取得了不错的效果. 文献[1]在训练语料库不足的情况下, 提出了一种“收缩”的统计技术以改进 HMM 模型中发射概率的参数估计; 文献[2]提

出了一种基于随机优化技术动态搜寻最优 HMM 模型的算法, 并应用于信息抽取; 文献[3]结合二元 HMM 与 SVM 来实现元数据的抽取; 文献[4]提出了一种基于内嵌的遗传算法和二阶 HMM 的信息抽取模型, 在分层的基础上对科研论文头部进行文本信息抽取; 文献[5]提出了一种基于模拟退火算法和一阶 HMM 相结合的信息抽取方法, 并应用于 Web 中论文头部的信息抽取.

针对模拟退火算法 (Simulated Annealing Algorithm, SAA) 在搜寻最优解时耗时较多, 以及一阶 HMM 在计算状态转移概率和释放概率时只考虑模型的当前状态而不考虑状态之间存在的上下文特征信息的不足, 本文提出一种基于改进的 SAA

收稿日期: 2014-04-03; 修回日期: 2014-05-28

基金项目: 国家自然科学基金项目 (61272109)

和二阶 HMM(HMM2)相结合的文本信息抽取方法.实验结果表明,新算法在精确度和召回率方面比传统 SAA 和一阶 HMM(HMM1)相结合的算法具有更好的性能.

2 基于 SAA-HMM 的信息抽取

2.1 模拟退火算法

SAA 是模拟物理热力学中固体退火原理来搜寻最优解的随机搜索算法,现在已经被普遍应用于解决组合优化问题. SAA 在搜索过程中具有概率突跳的能力,能有效地避免在搜索过程中陷入局部最优解,它在退火过程中不但可以接受更优的解,而且还可以一定的概率接受相对较差的解,同时该接受概率还会随着执行退火计划后温度的不断下降而逐渐减小,保证了获得全局最优解的可靠性.

2.2 基于 HMM 的信息抽取

HMM 是一种统计马尔科夫模型,本质上是带有不可观测状态的马尔科夫随机过程,也可以被看作一种简单的动态贝叶斯网络.

在实际应用中 HMM 模型主要用于解决学习问题、解码问题和评估问题.而基于 HMM 的信息抽取方法只需要解决应用中的前两个问题,即学习问题和解码问题.基于 HMM 的信息抽取过程如下:首先是学习问题,即使用 ML(Maximum Likelihood)算法对已标注的训练样本集,或者使用 Baum-Welch 算法对未标注的训练样本集进行模型训练,构建 HMM 模型;然后是解码问题,即利用 Viterbi 算法对测试样本集进行模型求解,获得最大概率的隐藏状态序列.

2.3 基于 SAA-HMM 的信息抽取

由于 SAA 具有较强的局部寻优能力,所以许多文献都将之应用于初始参数的优化.文献[5]首先采用 SAA 对随机设置的 HMM 初始参数进行优化:初始温度依据实验结果进行若干次调整;降温方案采用 $T(k+1) = \lambda \times T(k)$ (λ 为略小于 1 的系数)的方式进行;优化的目标函数采用 HMM 的评估函数 $P(o|\lambda)$;通过内外两层循环来控制算法的终止标准.然后把优化的初始参数和 Baum-Welch 算法相结合进行 HMM 的模型构建并应用于 Web 信息抽取.基于 SAA-HMM 的信息抽取整体框架如图 1 所示.

在基于 SAA-HMM 的信息抽取过程中,首先在 HMM 初始参数的选择中使用 SAA 进行参数优化,以减少 HMM 进入局部最优解的可能;然后把

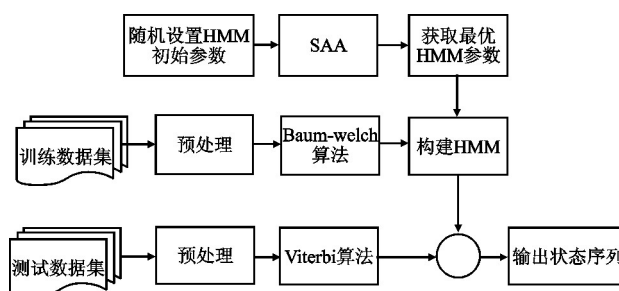


图1 基于 SAA-HMM 模型的信息抽取整体框架

优化的初始参数和 Baum-Welch 算法相结合对未标注的训练样本集进行学习,构建 HMM 模型;最后利用 Viterbi 算法对待抽取的文本序列进行求解,以获取最佳状态输出序列.

3 基于改进 SAA-HMM 的信息抽取

3.1 SAA 改进算法

SAA 虽然在解决组合优化问题中应用广泛,但作为一个启发式的搜索算法,仍然有许多参数需要调整,如退火方案、收敛规则等,以避免陷入局部最优解.文献[6]提出了一种动态改变温度下降比例的快速 SAA;文献[7]通过对 SAA 机理的研究与分析,提出了一种不仅可以保持原算法全局寻优的优点,而且可以提高算法稳定性和搜寻效率的改进算法 MVFSA,并通过实验表明 MVFSA 算法在时间性能上要比 SAA 更加优秀,具有更高的效益和稳健性[8].

针对 SAA 在搜寻最优解时耗时较多的不足,本文提出一种改进的 SAA 算法(Modified SAA, MSAA).改进算法的具体思路是:①由于随机发生器的状态遍历能力要优于其他模型的扰动方式[6],所以高温时应该采用模型的全局扰动方式来产生新解,以提高模型最优解的搜索空间;②定义一个相对低的温度 T_m ,当温度减小到 T_m 以下时,状态被接受的次数 accept 会逐渐减小,此时将温度适当提高,可激活各个状态的接受概率,避免算法陷入局部极小值. MSAA 算法具体描述如下:

- (1) 随机产生一个初始解 X_0 , 令最优解 $X_{best} = X_0$, 并计算优化的目标函数值 $F(X_0)$;
- (2) 初始参数设定: 设定初始温度 T_0 , $i = 0$, $j = 0$, 状态接受次数 $accept = 0$, 内层循环迭代次数 L_{inner} , 温度临界值 T_m ;
- (3) 判断是否满足条件 $i \leq L_{inner}$, 如果满足转向(4), 否则转向(6);
- (4) 使用随机扰动模型产生新解 X_{new} , 计算优

化目标函数值 $F(X_{\text{new}})$, 并计算优化目标函数值的增量值 $\Delta F = F(X_{\text{new}}) - F(X_{\text{best}})$; 如果满足条件 $\Delta F < 0$, 则令 $X_{\text{best}} = X_{\text{new}}$, $\text{accept}++$, $i++$, 转向(3); 否则转向(5);

(5)① 根据下式计算 α : $\alpha = \exp(-\Delta F/T_j)$, ② 产生一个随机数 ϵ , 其中 $\epsilon \in (0, 1)$; ③ 若 $\alpha > \epsilon$, 则令 $X_{\text{best}} = X_{\text{new}}$, $\text{accept}++$, $i++$; 否则, 令 $i++$, 转向(3);

(6) 退温: 如果 $T_j > T_m$, 则 $T_{j+1} = T_j * \exp(-\lambda)$, 其中 λ 为衰减系数; 否则, 按下式计算新的退火方案 ① $\beta = \text{accept}/(\text{accept} + L_{\text{inner}})$; ② 令 $T_{j+1} = T_j * \exp(-\beta)$;

(7) 判断是否满足设定的收敛条件, 若满足, 则转向(8); 否则, 令 $\text{accept} = 0$, 转向(3);

(8) 输出当前最优解 X_{best} , 算法结束.

为了测试改进算法的搜索性能和稳定性, 下面以数学领域著名的 TSP 问题为例来说明. 这里取城市个数 n 为 500, 并把每个城市假定为坐标系中的一个点, 且所有的点都不重合. 根据经验设定初始温度为 2 800, 每个温度下固定迭代次数为 $n * 100$, 降温计划采用算法中的方案, 其中 λ 取 0.92, 临界温度 T_m 是经过多次实验选取的适合温度, 算法结束条件为当连续三次能量函数值相差小于当前结果的 0.000 01 倍.

由表 1 可以看出改进算法 MSAA 的最短路径长度和迭代次数要比原始 SAA 算法更小, 而且在时间性能方面也要比原始算法有所提高.

表 1 2 种方法在抽取时间性能方面的比较

算法	最短路径长度	迭代次数	时间/s
SAA	426.29	33	78
MSAA	417.34	26	63

3.2 二阶 HMM 的引入

HMM1 假设中仅仅考虑概率与当前状态的关系而不考虑与历史状态的关联性, 在一定程度上降低了信息抽取的精度. 然而 HMM2 合理地弥补了这一缺点, 不仅认为概率与当前状态密切相关, 而且充分考虑了概率和模型历史状态的关联性, 有效地利用了状态之间存在的上下文特征信息, 在一定程度上增加了信息抽取的可靠性. HMM2 可以由七元组 $\lambda = (S, O, A_1, A_2, B_1, B_2, PI)$ 进行描述, 其中 S, O, A_1, B_1, PI 与马尔科夫模型中的 S, O, A, B, PI 是完全相同的:

$$A_2 = \{a_{ijk} = p(q_{t+1} = S_k | q_{t-1} = S_i, q_t = S_j),$$

$$1 \leq i, j, k \leq N\},$$

$$B_2 = \{b_{ij}(O_t) = p(O_t \text{ at } t | q_{t-1} = S_i, q_t = S_j),$$

$$1 \leq i, j, k \leq N\},$$

式中, A_2 为状态转移矩阵, a_{ijk} 表示已知 $t-1$ 时刻状态为 S_i 和 t 时刻状态为 S_j , $t+1$ 时刻状态为 S_k 的概率; B_2 为释放概率矩阵, $b_{ij}(O_t)$ 表示已知 $t-1$ 时刻状态为 S_i 和 t 时刻状态为 S_j , t 时刻输出观测值为 O_t 的概率.

文献[9]将 HMM2 应用于论文头部的信息抽取, 其中 ML 算法用于对模型进行训练, Viterbi 算法用于获取最优状态序列, 并通过实验表明 HMM2 比 HMM1 的抽取结果更加符合客观事实, 正确率更高. 对于已标注的训练样本集, 采用 ML 算法对模型进行构建显然是一种较优的选择. 然而目前用于研究的训练语料还是以未经标注的语料居多, 所以此时需采用 Baum-Welch 算法对未标注的训练样本集进行模型训练.

3.3 基于 SAA-HMM2 的信息抽取

针对以上两点, 本文将 MSAA 和 HMM2 相结合, 提出一种新的基于 MSAA-HMM2 的文本信息抽取方法. 基于 MSAA-HMM2 模型的信息抽取整体框架如图 2 所示.

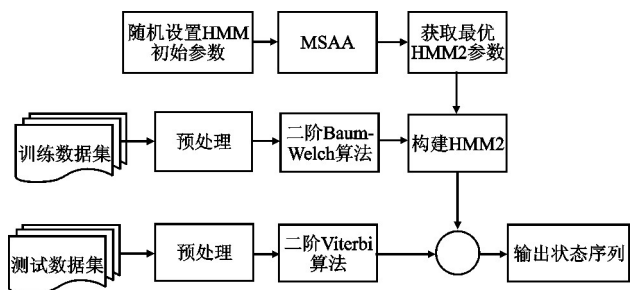


图 2 基于 MSAA-HMM2 模型的信息抽取整体框架

本文以科研论文头部信息作为实验数据进行文本信息抽取, 抽取出用户所关心的某类信息(比如论文题目, 作者等). 基于 MSAA-HMM2 模型的信息抽取分为以下四个步骤:

(1) 文本分块

目前基于 HMM 的信息抽取方法大多以单词或文本分块作为信息抽取的基本单元. 实际应用中的文本信息大多是由许多的文本分块组成的, 而文本分块显然要比单个单词所包含的上下文特征信息更加丰富, 所以以分块为信息抽取的基本单元能最大限度地挖掘和整合上下文特征信息, 提高抽取性能. 本文采用分块的基本思想^[8], 并将之应用于信息抽取. 下面以一篇科研论文的头部信息为例来说明文

本分块的思想:

Web Information Extraction Technology Research Based on Ajax

Zhang Hengru, Cui Chun

School of Computer Science

Southwest Petroleum University

Chengdu, 610500, China

zhanghrswpu@foxmail.com, cuichun006@sina.com

考虑到排版格式、换行、分隔符等信息,这里可以以换行符和逗号为依据把上面的论文头部信息划分为 11 个分组,可以看出有 6 个隐藏的状态,分别为题目、作者、附属机构、地址、邮政编码以及电子邮件地址。其中第 10 和第 11 个分组虽然不属于同一个分组但却都是电子邮件地址,即属于同一个状态。

(2) 参数优化

根据经验对 HMM2 的初始参数值 PI 、 A 、 B 、 A_2 、 B_2 进行初始化。然后利用 MSAA 算法对初始参数进行优化,其中优化目标函数采用评估函数 $p(o|\lambda)$;初始温度 T_0 和临界温度 T_m 根据实验结果进行若干次调整;退火方案采取算法中的规则;选取合适的内层循环迭代次数和收敛规则。经过若干次迭代最终输出最优解,即可得到最优初始参数。

(3) 模型训练

对训练数据进行预处理以生成文本序列;然后将生成的文本序列作为输入,调用 HMM2 下的 Baum-Welch 算法并结合第 2 步得到的最优初始参数对 HMM2 进行模型训练;若训练次数达到事先设定的最大迭代次数,则模型训练终止,否则继续进行迭代直至满足算法终止条件,并输出最优 HMM2 模型。

(4) 信息抽取

对测试数据进行预处理以生成观察序列;之后利用第 3 步得到的 HMM2 模型和 Viterbi 算法求解最优的隐藏状态序列;同时判断该隐藏状态序列是否为正确的状态序列,并根据判断结果对性能参数进行计算。

4 算法实现与结果分析

为确定 MSAA 中的临界温度 T_m ,首先根据经验设置初始温度 T_0 为 2 800;目标函数为评估函数 $P(o|\lambda)$;内层循环迭代次数为 $100 \times n$,其中 n 为可能的观察值的个数;外层循环迭代次数取 200;扰动模型采用全局随机扰动;退火方案采用如 MSAA 算法中所描述的方法。然后设定 T_m 所属区间为 $[0,$

1500],并将该区间分割为若干个区间段,分别计算各个区间段中使目标函数最大的 T_m ,最后取这些 T_m 中最大值作为最终的临界温度值,实验结果如图 3 所示。

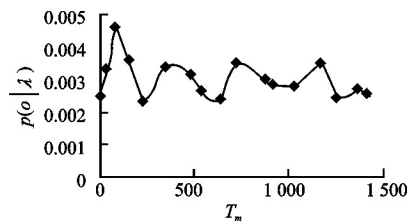


图 3 目标函数 $p(o|\lambda)$ 与 T_m 的关系图

这里设置 T_m 的取值范围为 $[0, 1\ 500]$,而不是区间 $[0, 2\ 800]$,是因为高温时模型扰动空间足够大,不需要进行重升温的过程,只有在低温的情况下模型的扰动空间才逐渐缩小,此时增加重升温过程可以避免模型陷入局部最优。从图 3 可以看出,当临界温度 T_m 达到 90 时,目标函数 $P(o|\lambda)$ 取得最大值,所以本实验取 T_m 为 90。

为了评价新模型的抽取性能,实验以中国期刊全文数据库中随机抽取并经过整理的 2 000 篇未经标注的论文头部信息作为实验语料,以其中的 1 800 篇作为训练样本集,200 篇作为测试样本集。同时将精确度(PRE)、召回率(REC)以及抽取一篇论文头部信息所花费的平均时间,即时间性能,这三个因子作为模型评价的指标。实验分别对基于 HMM 的信息抽取方法,基于 SAA 和一阶 HMM 相结合的 SAA-HMM1 抽取方法和基于 MSAA-HMM2 的信息抽取方法在精确度、召回率和时间性能上进行比较,比较结果如表 2 和表 3 所示。

表 2 3 种方法在精确度和召回率指标上的比较

状态	HMM		SAA-HMM1		MSAA-HMM2	
	PRE	REC	PRE	REC	PRE	REC
题目	0.55	0.61	0.59	0.66	0.67	0.74
作者	0.51	0.54	0.55	0.59	0.59	0.63
机构	0.63	0.70	0.67	0.74	0.68	0.76
地址	0.50	0.53	0.54	0.58	0.58	0.62

表 3 3 种方法在时间性能上的比较

模型	HMM	SAA-HMM1	MSAA-HMM2
平均抽取时间/s	0.76	1.15	0.96

从表 2 可看出基于 HMM 的信息抽取方法在精确度和召回率指标上的抽取效果不理想,这因为基于 HMM 的信息抽取模型的初始参数是随机产生的,而对未标注的文本集进行训练的 Baum-

Welch 算法对初始参数值比较敏感,易于陷入局部最优解;而基于 SAA-HMM1 抽取方法的实验结果比基于 HMM 的实验结果有较好的改善,各个状态的精确度和召回率都有所提高,体现了 SAA 较强的局部寻优能力. 基于 MSAA-HMM2 的信息抽取方法实验效果是最好的,在精确度和召回率方面都要比基于 SAA-HMM1 的信息抽取效果有所提高,体现了 HMM2 在上下文特征信息利用上的充分性.

从表 3 看出在时间性能上,本文基于 MSAA-HMM2 的信息抽取方法比基于 HMM 的信息抽取要差一些,主要是因为新模型采用 MSAA 进行初始参数的优化以及 HMM2 的引入;但与基于 SAA-HMM1 的信息抽取相比,新模型的时间性能有了较好的改善.

5 结束语

本文提出了一种基于改进 SAA 和 HMM2 相结合的信息抽取模型 MSAA-HMM2,实验结果表明,基于新模型的信息抽取在精确度和召回率方面都要比传统 SAA-HMM1 有较好的改善. 为了提高信息抽取的性能,在后续工作中,将对 SAA 进行更好的改进,如在 SAA 中引入自适应、遗传算法等,以提高 SAA 的搜索性能.

参考文献:

- [1] Freitag D, Mc Callum A. Information extraction with HMMs and shrinkage[C]// Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction, Orlando, 1999, 31-36.

- [2] Freitag D, McCallum A. Information extraction structures learned by stochastic optimization[C]// Proceedings of the Eighteenth Conference on Artificial Intelligence, Berlin, 2000:584-589.
- [3] 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): 584-589.
- [4] 李荣, 胡志军, 郑家恒. 基于遗传算法和隐马尔可夫模型的 Web 信息抽取的改进[J]. 计算机科学, 2012, 39(3): 196-199.
- [5] 邹腊梅, 龚向坚, 肖芳, 等. 基于模拟退火算法与隐马尔可夫模型的 Web 信息抽取[J]. 南华大学学报: 自然科学版, 2011, 25(1): 70-74.
- [6] 罗佐, 杜世民, 戚利侠, 等. 基于快速模拟退火算法的可切割布图规划方法[J]. 微电子学与计算机, 2012, 29(9): 109-112.
- [7] 陈华根, 李丽华, 徐惠平, 等. 改进的非常快模拟退火算法[J]. 同济大学学报: 自然科学版, 2006, 34(8): 1121-1125.
- [8] 胡宇舟, 王雷, 顾学道. 基于多模板隐马尔可夫模型的文本信息抽取算法[J]. 计算机应用, 2008, 28(3): 699-702.
- [9] 周顺先, 林亚平, 王耀南, 等. 基于二阶隐马尔可夫模型的文本信息抽取[J]. 电子学报, 2007, 35(11): 2226-2231.

作者简介:

王 宁 男, (1989-), 硕士. 研究方向为信息抽取、数据挖掘. E-mail: ningccuwhu@163.com.

李石君 男, (1964-), 博士. 教授. 研究方向为 Web 搜索与挖掘、数据挖掘、大数据.

(上接第 51 页)

- [6] 许丽, 王晓明. 基于 UC3842 的开关电源短路或过载保护的研究[J]. 兰州交通大学学报: 自然科学版, 2005, 24(4): 114-116.

作者简介:

熊国兴 男, (1988-), 硕士. 研究方向为 DC/DC 功率变换器的设计. E-mail: 2802901413@qq.com.

王俊峰 男, (1964-), 研究员, 硕士生导师. 研究方向为功率变换技术.