

HMM 在自然语言处理领域中的应用研究

韩 普, 姜 杰

(南京师范大学 教育科学学院, 江苏 南京 210097)

摘 要: 隐马尔可夫模型(HMM)是一种强大的统计机器学习技术,该模型已经成功地应用于连续语音识别、在线手写识别,在生物学信息中也得到了广泛的应用。由于该模型的强大的学习能力,在自然语言处理领域逐渐得到了应用。对隐马尔可夫模型在词性标注、命名实体识别、信息抽取应用中的关键问题进行了分析,着重分析了在信息抽取时使用隐马尔可夫模型的重点和难点问题,期望让更多的研究人员进一步认识和了解 HMM。最后分析了隐马尔可夫模型在应用中的不足之处和改进研究。

关键词: 隐马尔可夫模型;信息抽取;词性标注;命名实体

中图分类号: TP391

文献标识码: A

文章编号: 1673-629X(2010)02-0245-04

Application and Research of Hidden Markov Model in Natural Language Processing Domain

HAN Pu, JIANG Jie

(College of Education Science, Nanjing Normal University, Nanjing 210097, China)

Abstract: Hidden Markov model is a kind of powerful statistical machine learning technology, which has been successfully applied in continuous speech recognition and online character recognition. It has also been widely used in biology information. Because of this model's powerful learning capacity, it is increasingly applied in natural language processing. Analyze the application of hidden Markov model in part of speech tagging, named entity recognition and information extraction, among which the application of hidden Markov model in information extraction is emphatically analyzed, hoping more researchers have a better understanding about HMM. At the end of the paper, make an analysis about the inadequacies and improvement research of HMM in application.

Key words: hidden Markov model; information extraction; part-of-speech tagging; named entity

0 引 言

隐马尔可夫模型(HMM)是一种强有力的概率机器学习过程,已被成功应用于语音识别^[1]、手写体识别、生物信息学等领域。

HMM处理新的数据具有很好的鲁棒性,并且有一套成熟的算法。

隐马尔可夫模型的优点是它有强壮的概率统计作为基础,而这个特点也很适合处理自然语言领域的任务,在自然语言处理中^[2,3],HMM已被应用于词性标注^[4,5]、命名实体识别^[6]、信息抽取^[7-10]等任务。

HMM也有个明显的缺点就是模型的建立比较困难。而模型的构建恰是使用 HMM 的关键步骤。

1 隐马尔可夫模型的概述

1.1 概 述

隐马尔可夫模型(HMM)是一个二重马尔可夫随机过程,包括具有状态转移概率的马尔可夫链和输出观测值的随机过程,其状态只有通过观测序列的随机过程才能表现出来。一个 HMM 包含两层:一个可观察层和一个隐藏层。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,即一个有限状态机,其中每个状态转移都带有转移概率。一阶隐马尔可夫模型做了如下两个重要假设:其前提对于一个随机事件,有一个观察值序列 $O = \{v_1, v_2, \dots, v_M\}$,该事件隐含着—个状态序列 $S = \{s_1, s_2, \dots, s_N\}$ 。

假设 1: t 时刻的状态 q_t ,向 $t+1$ 时刻的状态 q_{t+1} 转移的概率仅仅与 q_t 有关,而与以往任何时刻的状态无关,即隐藏的状态序列构成—阶马尔可夫链,数学表示为: $P(q_i | q_1 \dots q_{i-1}) = P(q_i | q_{i-1})$ 。

假设 2: 在 t 时刻输出观测值 o_t 的概率,只取决于

收稿日期:2009-06-05;修回日期:2009-09-03

基金项目:国家自然科学基金(60873175)

作者简介:韩 普(1983-),男,山东人,硕士,研究方向为信息抽取、自然语言处理。

当前时刻 t 所处的状态 q_t , 而与其他的状态无关。数学表示为: $P(o_i | q_1 \cdots q_n, o_1, \cdots, o_i, \cdots, o_n) = P(o_i | q_i)$ 。

1.2 隐马尔可夫模型(HMM)的组成

HMM 是一个五元组 $\lambda = (S, O, A, B, \Pi)$, 简记为: $\lambda = (A, B, \pi)$, 其中:

$$S = \{s_1, s_2, \cdots, s_N\}$$

$$O = \{v_1, v_2, \cdots, v_M\}$$

$$A = \{a_{ij} = P(q_{t+1} = s_j | q_t = s_i), 1 \leq i, j \leq N\}$$

$$a_{ij} \geq 0, \sum_{j=1}^N a_{ij} = 1$$

$$B = \{b_j(k) = P(o_t = v_k | q_t = s_j), 1 \leq j \leq N,$$

$$1 \leq k \leq M\} b_j(k) \geq 0, \sum_{k=1}^M b_j(k) = 1$$

$$\Pi = \{\pi_i = P(q_1 = s_i), 1 \leq i \leq N\} \pi_i \geq 0, \sum_{i=1}^N \pi_i = 1$$

1.3 隐马尔可夫模型三个问题介绍

HMM 模型主要解决下面三个方面的问题: 评估问题: 给定观察值序列和模型 λ , 即给定模型和观察值序列, 求从模型生成观察值序列的概率 P , 通常采用 Forward 或 backward 算法。学习问题: 对于给定的观察值序列, 如何调整模型参数 λ , 使得观察值出现的概率 P 最大, 通常采用 Maximum Likelihood 算法。解码问题: 给定观察值序列和模型参数 λ , 求最可能的状态序列, 通常采用 Viterbi 算法。

1.4 隐马尔可夫模型解决问题的基本步骤

利用 HMM 解决问题通常有两个步骤, 步骤一: 通过训练样本生成 HMM 模型 λ , 模型的生成一般采用最大似然估计算法, 该算法常用以统计的方法得出 HMM 的模型各参数。计算模型的初始状态概率 π_i 、状态转移概率 a_{ij} 和状态释放概率 $b_j(k)$ 可以通过下面的三个公式进行计算:

$$\pi_i = \frac{\text{Init}(i)}{\sum_{j=1}^N \text{Init}(j)} \quad 1 \leq i \leq N \quad (1)$$

公式(1)中, $\text{Init}(i)$ 表示在所有训练样本中, 初始状态为 i 的序列个数。 $\sum_{j=1}^N \text{Init}(j)$ 表示在所有的训练样本中所有的可能初始状态的序列个数之和。

$$a_{ij} = \frac{c_{ij}}{\sum_{k=1}^N c_{i,k}} \quad 1 \leq i, j \leq N \quad (2)$$

公式(2)中, c_{ij} 表示在所有训练样本中, 从状态 s_i 到状态 s_j 的次数之和。 $\sum_{k=1}^N c_{i,k}$ 表示在所有的训练样本中, 从 s_i 转移到其他所有可能状态的总次数。

$$b_j(v_k) = \frac{E_j(v_k)}{\sum_{i=1}^M E_j(v_i)} \quad 1 \leq j \leq N, 1 \leq k \leq N \quad (3)$$

公式(3)中, $E_j(v_k)$ 表示所有训练样本中, 状态 s_j 发射出观察值 v_k 的次数。 $\sum_{i=1}^M E_j(v_i)$ 表示在所有训练样本中, 状态 s_j 发射出所有观察值的次数。

通过步骤一, HMM 模型已经生成, 步骤二利用已建立好的 HMM 模型对具体任务进行处理。通常以观察值序列 $O = \{v_1, v_2, \cdots, v_M\}$ 作为模型输入, 采用 Viterbi 算法, 找出最大概率的状态序列, 被标记为目标状态的观察文本就是要获取的信息。简单描述为: 在给定的模型下, 从一定观察值序列的所有可能的状态中, 选取概率最大的作为最终的状态序列。

2 隐马尔可夫模型(HMM)在自然语言处理领域中的研究

2.1 隐马尔可夫模型(HMM)在词性标注中的应用

词性标注是根据句子上下文中的信息给句子中的每个词一个正确的词性标记。基于 HMM 的词性标注的突出优点是: 针对一词多义的现象, HMM 可以根据上下文做出正确选择。基于 HMM 词性分标注需满足下面两个假设:

(1) 在统计意义上每个词性的概率分布只与上一个词的词性有关(即词性的二元语法);

(2) 每个单词的概率分布只与其词性相关。

对 HMM 词性标注的各参数分析如下:

① 状态值和观察值: 模型中状态(词性)的数目为词性符号的个数 N , 所有的词性集合构成词性的状态集合; 从每个状态可能输出的不同符号(单词)的数目为词汇的个数 M , 所有的单词集合构成观察值状态集合。目前各研究机构都有自己的词性标注集。所以观察值和状态值也略有差异。

② 状态转移矩阵和释放概率矩阵: 状态转移矩阵, 词性到词性的转移概率矩阵。释放概率矩阵: 从状态(词性)观察到输出符号(单词)的概率分布矩阵。

③ 初始状态概率 $p(i)$ 表示经语料库统计计算得到该词性 s_i 出现的频率。

明确了各参数的在词性标注中所代表的意义, (A, B, π) 可以通过对已经词性标注过的训练样本进行学习, 最后生成统计模型。处理任务时采用 Viterbi 算法进行词性标注, 因同一个词可能会有不同的词性, 需要根据上下文, 就要用到动态规划思想。简单的一阶 HMM 的词性状态仅与该词的上一状态即词性有关。已有系统使用的 HMM 分词系统多采用了改进的

HMM模型,其正确率得到明显提高。

2.2 在命名实体识别中的应用

命名实体识别(NE)任务是指识别文本中具有特定意义的实体,在MUC中提到的命名实体包括人名(Person)、地名(Location)、机构名(Organization)、日期(Data)、时间(Time)、分数(Percentage)、货币(Monetary value)这七类命名实体^[6]。在信息抽取研究中,命名实体识别是目前最有实用价值的一项技术。对HMM命名实体识别的两个假设:

(1)假设下一个NE类别只与前一个NE类别有关;

(2)假设词性的值出现观察值的概率只与当前NE类别有关。

① 状态值和观察值:在建立模型的时候,首先要确定状态的集合 S 以及观察值的集合 O 。然后根据 π , A , B 的定义训练出参数。各研究机构的相关标注也略有差异。这里以哈工大的命名实体为例^[11],将词性作为观察值,词性标注集采用国家863标准,共包含28种词性。NE有四类,包括人名(Nh)、地名(Ns)、机构名(Ni)和专有名词(Nz)。每一种类别根据它的组成部分在NE中出现位置的不同又可以分为NE开头($B-NE$)、NE内部($I-NE$)、NE结尾($E-NE$)以及独立NE($S-NE$),再加上不属于任何NE类别的“其它”类型(O),共17种。这17种组成了HMM中的17种状态集。

② 状态转移矩阵和释放概率矩阵:状态转移概率 $P(t_i | t_{i-1})$:表示从状态 t_{i-1} 到 t_i 的转移概率,在命名实体识别中指的是NE类别的转移概率;释放概率矩阵:从状态(NE类别)观察到输出符号(词性)的概率分布矩阵。

③ 初始状态概率矩阵:初始状态分布,是指一个句子第一个词NE类别的概率分布。

通过以上分析,简单的一阶HMM模型通过训练便可以构建。具体任务处理需要使用Viterbi算法对NE类别进行标注。命名实体识别是一项非常复杂的工作,实际应用中,一般将多种方法结合起来使用,最常见的是将基于统计和基于规则^[12]、基于词典的方法结合起来,以提高准确率。

2.3 在信息抽取中的应用

基于统计的信息抽取是目前信息抽取的一种主流方式,相比基于规则的信息抽取,前者最大的优点在于对领域知识的要求不高,且具备良好的领域移植能力,基于HMM的统计方法是信息抽取研究的重点。相比于HMM在词性标注、命名实体识别中参数的确定,基于HMM的信息抽取最重要的问题是HMM参数的选

择,比如某些状态的选择可以通过观察值更好的被捕获,抽取效果就比较好。目前,信息抽取研究的抽取对象都是格式化程度较高的文本,对于格式比较自由的文本的全篇抽取的研究比较少。使用HMM信息抽取时,状态和观察值的选择是一项很重要的任务,对隐马尔可夫模型的训练(参数估计)是一个非常重要的问题,训练方法的优劣将对整个应用效果产生重要的影响。

这里以高校的学校简介为例,前提是已经将网页中的学校简介文本信息进行了正文抽取,并且进行了词性标注。例如有下面的已经词性标注过的一段学校介绍文字:

学校 /n 占地 /v 面积 /n2009906/m 平方米 /q , /wd 设有 /v 二级 /b 学院 /n23/m 个 /q , /wd 共有 /v 在职 /vn 教职工 /n 3274/m 人 /n , /wd 共有 /v 在校 /b 普通 /a 本科生 /n 15277/m 人 /n , /wd 博士 /n 研究生 /n 684/m 人 /n , /wd 硕士 /n 研究生 /n 5924/m 人 /n , /wd

任务假设要抽取学校简介信息中的学校面积、所有学生、教职工、博士研究生、硕士研究生、二级学院的相关信息,对模型的参数进行分析如下。

2.3.1 观察值和状态值

为给信息抽取模块建立隐马尔可夫模型,首先确定模型应该包含多少状态,共有哪些状态以及状态之间允许如何转移。对于信息源是高校简介的网页来说,需要抽取关于学校的占地面积、博士生人数、硕士生人数、教职工人数、全体学生人数等信息,这些信息又可称之为“抽取域”。这些抽取域对应的状态集合组成的有限状态机便构成了隐马尔可夫模型的隐藏层,待抽取节点中的标识信息则构成了隐马尔可夫模型的可观察层。针对高校简介信息抽取的HMM拓扑图如图1所示,这里的观察值序列要对训练模型要做简单处理,在“占地面积”抽取域中出现的观测值包括数字“2009906”,状态释放的数据一定是待抽取文本中的数据,而文本中是“2009906”还是其他数字对抽取过程本身并没有影响,可以统一作为“数字”来进行处理,这里,“数字”是该抽取域中的重点观测值,可将其附带后面的量词“平方米/q”等抽取出来成为有效的观察值数据。在本例中,可以统一将数字用***取代。通过这样处理,观察值的总个数便有限了,该例中观察值就是:***平方米,***公顷,***人,***个,***名等。

注意,这里加上了两个开始和结束的状态,这恰恰就是HMM模型被称为有限状态自动机的原因,作为有限,要有开始和结束状态,这两个状态只出现在状态

转移矩阵,不存在观察值,也就不存在释放矩阵中。

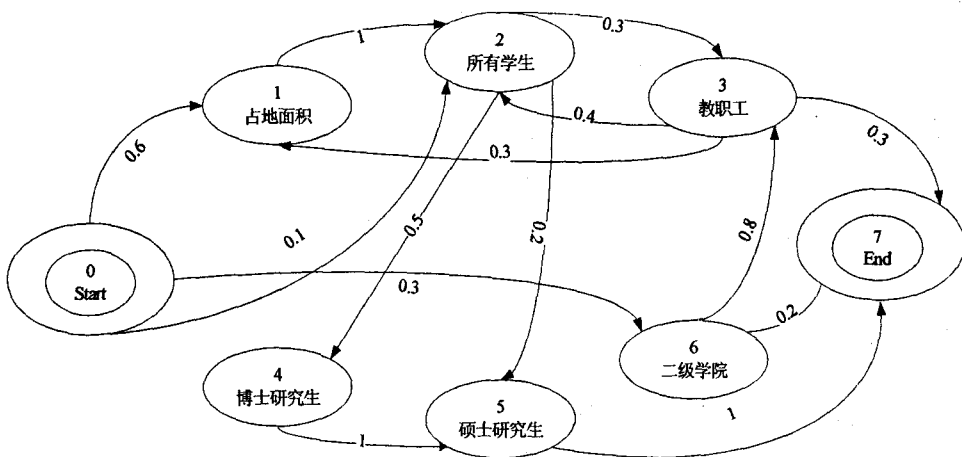


图 1 简单 HMM 示例

2.3.2 状态转移和释放概率矩阵

状态转移矩阵:这里的状态转移就是 6 个抽取域再加上开始和结束状态,共 8 个状态。这样初始状态概率矩阵实际上就包含在这个加上了开始和结束状态的状态转移概率矩阵之中了。释放概率矩阵:从状态观察到输出观察值的概率分布矩阵。初始状态概率矩阵:某一开始状态在所有的训练文档中作为开始的概率,这个矩阵已经包含在总的状态转移概率矩阵之中,不需要单独计算。

这样便建立了一个示范性的隐马尔可夫模型,实际模型要比这复杂得多,包括很多标签和标注信息本身也可能成为状态,为了提高抽取准确度,甚至还需要对每一个状态句子都建立一个 HMM 模型,文中不进行深层次讨论,因为过多的状态和观察值会使模型结构非常庞大,这也是目前尚不容易将隐马尔可夫模型用于非常宽泛的信息抽取的原因。对于具有较高专业性的信息抽取而言,状态的数目可以控制在较好的水平,信息抽取的效率也会比较高。所以建立合理的 HMM 模型是最为重要和关键的一步。

3 隐马尔可夫模型应用的局限和不足

3.1 模型构建需要大量的训练数据

当训练数据的数量不是足够大时,总有可能会出现 0 观测值的现象,对于这种情况的出现,现有研究多是根据不同的情况采用了改进的平滑算法。其次,训练文档跟目标文档的结构越一致,抽取的准确度和精度也越高。

3.2 状态和观察值的确定

实验证明,隐马尔可夫模型的有效性比其它方法具有优越性。但是,隐马尔可夫模型也存在一些问题,如隐马尔可夫模型的结构确定困难等问题。在词性标

注和命名实体识别方面,状态值和观察值还比较容易

确定,但是对自由文本的 HMM 状态值和观察值的确定就比较困难了。状态值的选择决定着 HMM 模型的复杂程度和信息抽取的精度,如何在这两个对立矛盾中建立一种平衡,需要在模型的建立过程中认真把握。

3.3 HMM 模型两个假设的不合理性

近年来隐马尔可夫模型(HMM)在自然语言

处理领域得到很大发展,然而 HMM 模型有着一定的局限性。隐马尔可夫模型默认情况下仅指一阶隐马尔可夫模型,而事实上任一时刻出现的观测值概率不仅依赖于系统当前所处的状态,也可能依赖于系统之前时刻所处的状态,所以上假设并不十分合理,没有考虑到状态之间存在的上下文特征信息,这些信息对于实现更准确的信息抽取是非常有用的。如何克服 HMM 的一阶假设和独立性假设带来的问题一直是研究的热点。

4 隐马尔可夫模型(HMM)改进研究

4.1 与基于规则的相结合

基于规则和 HMM 模型的方面目前使用的最为普遍^[11,12],其抽取正确率和精度有明显提高,基于统计和规则的抽取可以分为两个步骤,首先使用 HMM 进行处理,然后利用具有优先级别的匹配规则对第一步的结果进行修正和转换。在具体的应用中,规则可以根据处理的具体任务适时添加。

4.2 二阶及多阶 HMM 的研究

针对一阶隐马尔可夫模型假设的不合理性,有不少研究对 HMM 进行了改进研究,如将二阶^[10]、三阶 HMM 模型的应用于命名实体识别、文本抽取的研究,多阶 HMM 可以更好地捕捉上下文信息。但其复杂度也越来越大。这里笔者还是建议,根据任务的复杂程度和要求,选用合适的 HMM 模型。

5 结束语

HMM 有着强大的理论基础和成熟的算法支持,目前已经在自然语言处理领域得到了广泛的应用。相比于基于规则的方法,HMM 有较强的系统移植能力。

(下转第 252 页)

速度,同时有利于系统的集成和扩展。该系统能够在诸多复杂因素的影响下,保证监管的正常运行。在跨平台运行和大负荷环境下,具有良好的移植性和运行效率。

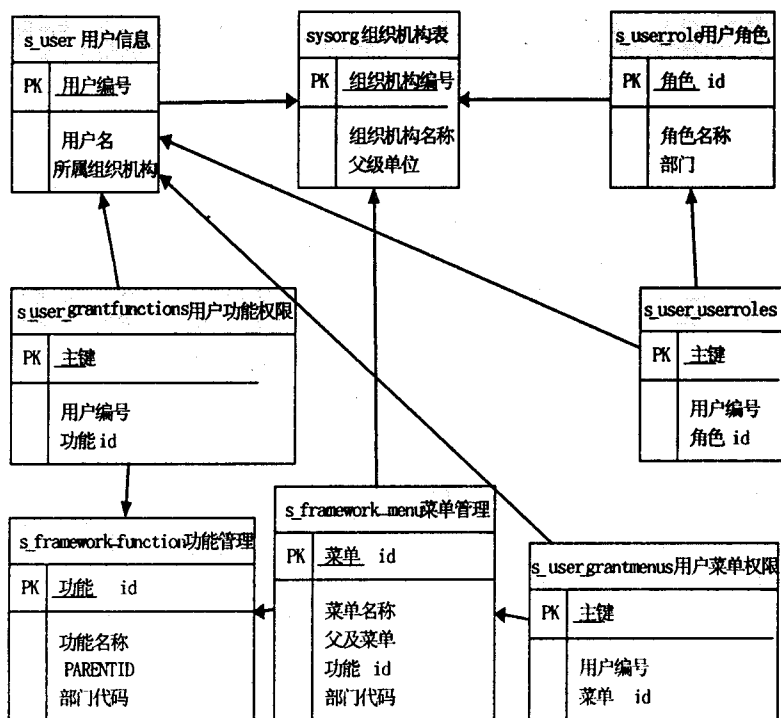


图 4 系统框架的用户关系与实体关系图

参考文献:

- [1] 席晓峰,吕良双,逯 鹏.使用 J2EE 框架技术构建可重用的 Web 应用[J].计算机工程与应用,2005(29):21-22.
- [2] 彭 彬,甘早斌,李志欣.基于 J2EE 的 Web 应用系统的优化设计[J].计算机工程与科学,2005(10):32-34.
- [3] Johnson. J2EE Development Frameworks [J]. Computer,2005,38(1):107-110.
- [4] Yam S. J2EE 编程指南[M].北京:电子工业出版社,2004:98-102.
- [5] 程 洪,钱乐秋,马舜雄.基于 J2EE 体系的 Web 应用框架整合[J].计算机工程,2005(20):105-107.
- [6] 李 敏,黄 强,李 昊,等.基于 J2EE 的客运信息管理系统数据持久层的 Hibernate 解决方案[J].计算机应用,2005(10):32-33.
- [7] 杨兴春,谯 石,董 文,等.基于轻量级 J2EE 构架的高校教务管理系统的设计与实现[J].计算机系统应用,2007(3):42-43.
- [8] 谢运佳,王会进,钟瑞琼,等.一种轻量级的 J2EE 解决方案及其应用[J].微计算机信息,2006(9):36-37.

(上接第 248 页)

文中对 HMM 在自然语言处理领域中的几个方面进行了分析,着重分析了这些方面在使用 HMM 时应该注意的重要问题。HMM 也不是一个非常完美的模型,由于 HMM 是在假设的前提下成立的,这与实际情况并不相符,研究者提出了不少改进的研究方案,研究者和对 HMM 进行了改进,并提出了多阶的 HMM 方法,取得了较好的成果。具体的应用中,很少有使用单一的 HMM 的方法,目前基于 HMM 和规则相结合的方法在自然语言处理领域中处于主流地位。HMM 在自然语言处理领域的应用才刚刚起步,相信在不久的将来,HMM 可以在自然语言处理领域的作用可以得到更充分的挖掘。

参考文献:

- [1] Rabiner L E. A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition[J]. Proceedings of The IEEE,1989,77(2):257-286.
- [2] Freitag D, McCallum A. Information extraction with HMM structures learned by stochastic optimization[C]//Proceedings of the Eighteenth Conference on Artificial Intelligence. [s.l.]: [s.n.],2000:584-589.
- [3] Freitag D, McCallum A. Information Extraction with HMMs and shrinkage[C]//In: Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction. [s.l.]: [s.n.],1999:31-36.
- [4] 王 敏,郑家恒.基于改进的隐马尔科夫模型的汉语词性标注[J].计算机应用,2006,26(12):197-198.
- [5] 胡春静,韩兆强.基于隐马尔科夫模型(HMM)的词性标注的应用研究[J].计算机工程与应用,2002(6):62-64.
- [6] 赵琳瑛.基于隐马尔科夫的中文命名实体识别研究[D].西安:西安电子科技大学,2008.
- [7] 王 雷,顾学道.基于多模板隐马尔科夫模型的文本信息抽取算法[J].计算机应用,2008,28(3):699-702.
- [8] 林亚平,刘云中,陈治平.基于最大熵的隐马尔科夫模型文本信息抽取[J].电子学报,2005,33(2):236-241.
- [9] 刘云中,林亚平,陈治平.基于隐马尔科夫模型的文本信息抽取[J].系统仿真学报,2004,16(3):507-510.
- [10] 周顺先,林亚平,王耀南,等.基于二阶隐马尔科夫模型的文本信息抽取[J].电子学报,2007,35(11):2226-2232.
- [11] 廖先桃,于海滨,秦 兵,等.HMM 与自动规则提取相结合的中文命名实体识别[C]//第二届全国学生计算语言学研讨会.北京:[出版者不详],2004:232-237.
- [12] 向晓雯,史晓东,曾华琳.一个统计与规则相结合的中文命名实体识别系统[J].计算机应用,2005,25(10):2404-2406.