

文章编号:1001-9081(2010)09-2348-03

## 基于隐马尔可夫模型的文本分类算法

杨健<sup>1,2</sup>, 汪海航<sup>1</sup>

(1. 同济大学 电子与信息工程学院, 上海 201804; 2. 大理学院 数学与计算机学院, 云南 大理 671003)

(sbjc1215@126.com)

**摘要:**自动文本分类领域近年来已经产生了若干成熟的分类算法,但这些算法主要基于概率统计模型,没有与文本自身的语法和语义建立起联系。提出了将隐马尔可夫序列分析模型(HMM)用于自动文本分类的算法,首先构造表示文档类别的特征词集合,并以文档类别的特征词序列作为不同HMM分类器的观察序列,而HMM的状态转换序列则隐含地表示了不同类别文档内容的形成演化过程。分类时,具有最大生成概率的HMM分类器类标即为测试文档的分类结果。该算法构造的分类器模型一定程度上体现了不同类别文档的语法和语义特征,并可以实现多类别的自动文本分类,分类效率较高。

**关键词:**文本分类;隐马尔可夫模型;信息增益; $\chi^2$ 检验;词频—反文档频率

**中图分类号:** TP182 **文献标志码:** A

### Text classification algorithm based on hidden Markov model

YANG Jian<sup>1,2</sup>, WANG Hai-hang<sup>1</sup>

(1. School of Electronics and Information, Tongji University, Shanghai 201804, China;

2. School of Mathematics and Computer Science, Dali University, Dali Yunnan 671003, China)

**Abstract:** A number of sophisticated automatic text classification algorithms have been proposed in recent years, but those algorithms are mainly based on the probability and statistical models and have not established a relationship with the syntax and semantic of text. In this paper, a new automatic text classification algorithm using Hidden Markov Model (HMM) was proposed. At first, a feature set was built to distinguish the document types. Then the different sequences of feature words were regarded as the different observations generated by HMM classifiers. The state transition sequence of a specific HMM classifier implied the process of document's formation and evolution in a specific document type. When a document was classified, the result was created by the HMM classifier which could get the greatest generation probability according to the document. To some extent, some syntactic and semantic features of different document were represented by the classification model. The model can be applied to automatic multi-category text classification, and it has high classification efficiency.

**Key words:** text classification; Hidden Markov Model (HMM); information gain;  $\chi^2$  test; Term Frequency-Inverse Document Frequency (TF-IDF)

## 0 引言

文本分类是现代信息处理中一个重要任务,是文本挖掘中的一个重要环节,在建立面向特定领域的主题搜索引擎构建中也非常重要。自动文本分类经过多年的研究已经产生了许多优秀的模型和算法。研究中比较多的有支持向量机(Support Vector Machine, SVM)、K近邻(K-Nearest Neighbor, KNN)、朴素贝叶斯(Naive Bayes, NB)等算法,并且这些算法的一些改进模型和分类过程中涉及的文本表示、分词、特征选择等领域也得到了进一步的探索。

可以看到,统计学理论在文本分类算法中有非常重要的地位。然而传统统计学模型将特征词的出现频率作为分类算法的直接依据,没有充分考虑特征词在不同类型文本中或上下文中的本质作用,即包含的语法和语义信息。并且,在语义网及文本表示本体技术仍需进一步发展的条件下,需要将新型的概率统计模型与文本分类等信息处理研究结合起来。本文建立基于隐马尔可夫模型(Hidden Markov Model, HMM)的文本分类模型,提出了模型结构,并详细介绍了HMM参数学

习方法和分类步骤。模型隐含了特征词的语义关联,能够进行多类别文本自动分类,分类效率较高。

## 1 HMM模型及相关的研究工作

### 1.1 隐马尔可夫模型

HMM是一种基于统计的序列分析和学习模型,是由Baum等人在20世纪60年代末至70年代初建立起来的,近年来在信息技术领域颇受重视,在语音识别、自然语言处理和文本挖掘上已经有了相应的应用。基于HMM信息处理模型易于建立,不需大规模的词典集和规则集,并且将词汇的概率分布特征用于模型建立,比人工神经网络(Artificial Neural Networks, ANN)等模型易于理解,能部分反映出文本的语义性质。

一个HMM模型是不确定的、随机的有限状态自动机,由不可观测的状态转移过程(一个Markov链)和可观测的观察生成过程组成。按观察值是离散还是连续的,HMM可分为离散型HMM和连续型HMM。在文本分类中,计算的是离散的词的分布特征,因此下面只介绍离散HMM表示。

离散HMM是一个五元组: $\lambda = \{N, M, \pi, A, B\}$ ,可以简

收稿日期:2010-03-08;修回日期:2010-04-27。 基金项目:上海市科委科技支撑计划项目(072712036)。

作者简介:杨健(1976-),男,浙江上虞人,讲师,博士研究生,CCF会员,主要研究方向:智能信息系统;汪海航(1965-),男,浙江奉化人,教授,博士生导师,主要研究方向:信息安全、智能信息系统、电子商务。

写为  $\lambda = \{A, B, \pi\}$ , 其中  $N$  是 Markov 链状态数;  $M$  是状态可能生成的观察值数;  $\pi = (\pi_1, \dots, \pi_N)$  表示初始状态概率向量, 其中  $\pi_i = P(q_1 = S_i)$ ,  $1 \leq i \leq N$ 。

$A$  为状态转换概率分布, 表示  $i$  状态向  $j$  状态转换的概率:  $A = (a_{ij})_{N \times N}$ , 其中:  $a_{ij} = P(q_{t+1} = s_j | q_t = s_i)$ ;  $1 \leq i, j \leq N$ 。

$B$  为观察值概率分布, 表示在  $j$  状态输出  $k$  观察值的概率:  $B = (b_j(k))_{N \times M}$ , 其中:  $b_j(k) = P(o_t = v_k | q_t = s_j)$ ;  $1 \leq j \leq N, 1 \leq k \leq M$ 。

按照隐状态转换特征进行分类, HMM 有几种典型的结构, 图 1 所示的即为典型的两种左-右型 Markov 链。注意到图中并没有表示观察值的输出。

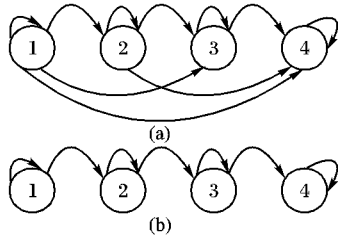


图1 左-右型 Markov 链

HMM 有 3 个基本问题(评估问题、解码问题、学习问题)需要解决, 它们是 HMM 理论的主要内容<sup>[1]</sup>。其中评估问题是给出观察值序列和模型, 评估观察序列是由模型给出的概率; 解码问题是给出观察序列和模型, 如何选择对应的状态序列, 以较好地解释观察值; 学习问题是如何调整模型参数以使得观察序列是由模型给出的概率最大化。在建立好 HMM 模型后, 可以使用评估问题算法进行待分类文档的分类, 常用的有前向算法和后向算法。

## 1.2 与本文相关研究工作

在信息处理领域, HMM 用于文本信息抽取的研究比较深入: 文献[2-4]利用文本排版格式信息对文本进行分块, 并结合隐马尔可夫模型进行文本信息抽取。为了获得较优的 HMM 模型的初始参数, 文献[5]利用遗传算法以减少 HMM 进入局部极小的概率; 文献[6]采用主动学习来选择最有价值的训练文本, 以减少标注工作量; 文献[7]提出一种将命名实体识别方法 NER 集成到文本分类特征选择中的方法, 除了统计特征, 还保留了单词作为命名实体的分类特征。在信息抽取时, 文献[8]描述了二阶 HMM 在论文头部信息抽取中的基本步骤。文献[9]则将反向动态规划和正向 A\* 算法结合, 以确定部分文本块的状态。然而, 上述研究主要集中于某些特定领域具有特定结构的文本, 例如计算机科研论文头部的信息抽取, 对于非结构化文本涉及较少。

在非结构化文本分类研究中, 与本文的研究工作更为近似的有: 面向多页面文档(例如书或杂志)研究页面序列提高分类精度的问题<sup>[10]</sup>; HMM 中的每个状态与一个唯一的页面类别(如封面、序、目录等)关联, 而观察输出概率通过页面中单词的多项式分布来建模。对于每个新的待分类文档, 算法输出有最大后验概率的页面类别的序列。但该研究要解决的是页面在整个文档中所属结构类别的问题, 并不能实现整个文档从语义角度上的类别的自动识别。文献[11]提出了一种基于 HMM 的文本分类方法。但该方法有如下缺陷: 1) 每个训练文本对应一个 HMM 模型, 训练时间难以保证; 2) HMM 分类器的概率分布学习算法没有明确说明, 而这是一个分类器模型的重要组成部分; 3) HMM 分类器参数确定与文本类的语法、语义和结构没有明显关系。

本文提出一种基于 HMM 的文本分类方法。在将 HMM 用于文本分类时, 对不同类别分别建立 HMM 分类器。另外, 鉴别文本是否属于同一类别, 可以通过观察它们的特征词组成及频率。因此, HMM 中的观察输出就是特征词的组成。一个 HMM 分类器中的状态转换, 可以看做是从与该类别不是很相关的词组成的文档输出分布, 向与该类别非常相关的词组成的文档输出分布转化的一种过程。因此, 状态从起始点向终结点转化应对应着类别相关词汇的强化。基于以上的考虑, 提出了基于 HMM 的文本分类模型及分类算法实现。模型以有类标的文本集为训练集, 训练得到不同类别的 HMM 分类器, 然后检测待分类文本对于各个类别 HMM 的生成概率, 生成概率最大的类别的类标号作为最后的分类结果。

## 2 基于隐马尔可夫模型的文本分类

### 2.1 模型及分类器训练

在基于 HMM 的文本分类模型中, 假设训练文档集的类标集合为  $C = \{c_1, c_2, \dots, c_k\}$ , 在经过分词、去除停用词等文本预处理过程后, 文档被转换为词集。基于 HMM 的文本分类模型分类器训练要经过如下步骤。

1) 特征选择及降维。特征选择有多种方法, 如词频-反文档频率(Term Frequency-Inverse Document Frequency, TF-IDF), 信息增益(Information Gain, IG), 互信息(Mutual Information, MI), 期望交叉熵(Expected Cross Entropy, CE), 开方检验( $\chi^2$  检验), 粗糙集(Rough Set, RS)等。本文采用信息增益(IG)值作为特征选择标准:

$$IG(t) = - \sum_{i=1}^{|C|} P(c_i) \lg P(c_i) + P(t) \sum_{i=1}^{|C|} P(c_i | t) \lg P(c_i | t) + P(\bar{t}) \sum_{i=1}^{|C|} P(c_i | \bar{t}) \lg P(c_i | \bar{t})$$

计算各个词  $t$  在分类中的 IG 值后, 设定阈值, 选择信息增益大的词作为特征词, 构成特征集  $W$ 。

2) 类别特征集抽取。不同类别判断需要的特征词集不同, 因此, 需要对每个类别抽取最有效的类别判断特征词。利用  $\chi^2$  检验, 找到不同类别  $\{c_1, c_2, \dots, c_k\}$  对应的类别特征集  $\{W_1, W_2, \dots, W_k\}$ , 即类别  $c$  有  $W$  的特征子集  $W_c$ 。具体方法如下所示。

假设包含词  $t$  且属于  $c$  的文档个数、包含词  $t$  且不属于  $c$  的文档个数、不包含词  $t$  且属于  $c$  的文档个数、不包含词  $t$  且不属于  $c$  的文档个数分别为  $A, B, C, D$ , 则有:

$$\chi^2(t, c) = \frac{N(AD - BC)^2}{(A + C)(A + B)(B + D)(C + D)} \quad (1)$$

其中  $N = A + B + C + D$ 。如果给定训练集和类别, 则  $N, M, N - M$  对于同一类别文档中所有词来说都是一样的。这里  $M = A + C, N - M = B + D$ 。所以, 公式可简化为:

$$\chi^2(t, c) = \frac{(AD - BC)^2}{(A + B)(C + D)} \quad (2)$$

其值越大, 表明  $t$  与  $c$  类越相关。通过设定阈值, 可以找到  $c$  类别最相关的特征词的集合  $W_c = \{w_{c1}, w_{c2}, \dots, w_{cd}\}$ 。

3) 分类器模型及参数学习。设类  $c$  的分类器为  $HMM_c$ , 如图 2 所示。

图 2 中,  $S_0, S_t$  分别表示开始和结束状态。在分类器工作时, 状态的转换表示词集合的统计特征向该类别的逼近, 因此, 状态转换概率有如下性质:  $P\{S_{t+1} | S_t\} = 1$ , 由此可以得到第  $c$  个类别分类器的状态转移概率分布  $A_c$ 。图中的各个特

征值  $w_{c1}, w_{c2}, \dots, w_{1cl}$  不是表示具体的输出,而是表示在特征值  $w_{c1}, w_{c2}, \dots, w_{1cl}$  的基础上,各个特征词的输出概率。并且  $w_{c1}, w_{c2}, \dots, w_{1cl}$  是通过式(2)计算值的一个升序排列,表明该类文档是经过特征词由不相关到逐渐相关的一个渐变过程而形成的,也即文档特征向类别  $c$  的逼近。

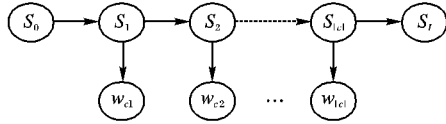


图2 类  $c$  的 HMM 分类器模型

在中间状态  $S_i$  (除  $S_0, S_t$  以外的状态)下,  $c$  类的观察值分布由式(3)计算而得,即:

$$b_i^{(c)}(k) = P_c[w_k \text{ att} | q_t = S_i] = P_c(w_k | w_i) = X_c(w_k, w_i) \cdot TFIDF_c(i) \quad (3)$$

$w_k$  表示特征词集  $W$  中第  $k$  个特征词,而  $w_i$  则是  $c$  类特征词集  $W_c$  按相关度(式(2)计算)升序排列的第  $i$  个特征词。 $X_c(w_k, w_i)$  可以用开方检验  $\chi_c^2(w_k, w_i)$  计算而得:

$$X(w_k, w_i) = \frac{\chi_c^2(w_k, w_i)}{\sum_{1 \leq i \leq 1cl} \chi_c^2(w_k, w_i)} \quad (4)$$

其中  $\chi_c^2(w_k, w_i)$  衡量了  $w_k$  与类别  $c$  中  $w_i$  的相关程度,也即  $w_k$  与隐状态  $S_i$  的相关性。 $\sum_{1 \leq i \leq 1cl} \chi_c^2(w_k, w_i)$  为归一化分母,  $\chi_c^2(w_k, w_i)$  的计算方法如下。

设类别  $c$  中同时包含  $w_k$  和  $w_i$  的文档个数、包含  $w_k$  但不包含  $w_i$  的文档个数、不包含  $w_k$  且包含  $w_i$  的文档个数、不包含  $w_k$  且不包含  $w_i$  的文档个数分别为  $A, B, C, D$ , 类似于式(2)有:

$$\chi_c^2(w_k, w_i) = \frac{(AD - BC)^2}{(A + B)(C + D)} \quad (5)$$

$TFIDF_c(i)$  函数则是特征词  $w_i$  的 TF-IDF,不过这里与传统的 TF-IDF 略有不同:

$$TFIDF_c(i) = \frac{n_{ic}}{\sum_c n_{ic}} \cdot \log \frac{|D|}{|D_c, D_c \ni w_i|} \quad (6)$$

其中:  $D_c$  表示类别  $c$  的文档集合,  $n_{ic}$  表示特征词  $w_i$  在类别  $c$  中的文档中出现的次数。即此时的  $TFIDF_c(i)$  用以评估  $w_i$  对于训练集中的类别  $c$  的重要程度。可以看到,这里在某个类别下的状态与观察输出的概率关系通过特征词之间的关联度量来表示,也即隐含了特定类别下特征词之间的语义关系。

由式(3)得到了  $HMM_c$  的观察值输出概率分布  $B_c$ 。这样,类  $c$  的 HMM 分类器可由  $\{A_c, B_c, S_0\}$  表示。

4) 用不同类标号训练文档训练相应的 HMM 分类器,对应于  $k$  个类别得到分类器集合  $\{HMM_1, HMM_2, \dots, HMM_k\}$ 。

## 2.2 使用 HMM 分类器分类步骤

经过上述的 HMM 分类器训练过程后,训练集中每一个类标号  $c$  对应有一个分类器  $HMM_c$ 。待分类文档  $d$  的分类则按如下步骤进行。

1) 首先经过分词和去除停用词等文本预处理过程,并按照上节所述的分类特征选择时计算的信息增益大小将出现在分类特征集合中的  $d$  中的特征词表示为升序集合  $W_d = \{w_1, w_2, \dots, w_d\}$ 。

2) 对于某个类别  $c$ , 设  $HMM_c$  的状态相关特征词有  $k$  个,则在  $W_d$  集合中选取信息增益最大的  $k$  个作为该文档的特征集合,也即在  $HMM_c$  中的输出序列  $Q$ 。

3) 利用前向或后向算法,计算类别  $c$  的  $P_c(Q | \lambda)$ ,  $\lambda$  表示

$c$  的分类器  $HMM_c$ 。

4) 依次计算每个类别的  $P(Q | \lambda)$ , 最后得到待分类文档的类标号  $c = \arg \max_{1 \leq c \leq |C|} [P_c(Q | \lambda)]$ , 注意到  $|C|$  是类别总数。

## 3 算法分析

文本由一系列特征词集合而成,特征词权重及词频等统计特征在不同类中表现不同,且不同类别需要观察的特征词的类别子集不同,所以算法对不同类别分别建立 HMM 模型。另外,鉴别文本是否属于同一类别,可以通过观察它们的特征词组成及频率。因此,模型中的观察输出就是特征词。而某个 HMM 分类器中的状态转换可以看做是从与该类别不是很相关的词组成的文档输出分布向与该类别更为相关的词组成的文档输出分布转化的一种过程。因此,状态从起始点向终结点转化对应着特定类别相关词汇的强化,这种强化的随机过程用观察输出概率来表示。

在特征选择阶段,所有训练集不是共用一个特征集,而是通过两个阶段来实现特征选择及降维:第一阶段利用信息增益对所有训练文档进行特征选择,生成一个特征集,使得特征词集合对分类信息的贡献较大,这个集合也是 HMM 中观察值的合集;第二阶段利用开方检验找到不同类别对应的类别特征集,类别特征集在表示上以特征表示,而本质上反映了不同类别文档隐含状态的转换,即由一般文档向类别特征突出文档的转换。

此外,在不同类别分类器中,状态对应的观察值输出概率由式(3)表示。该公式不但反映了词频与文档的关系(通过改进的 TF-IDF 值计算得到的  $c$  类文档下的 TF-IDF,由式(6)表示),而且反映了特征词与当前类别的文档状态转换之间的关联(式(4)计算的关联值)。

由上述分析可以看出,本文提出的模型有以下特点:1) 在模型的观察值输出概率分布矩阵的学习过程中,通过将  $\chi^2$  检验和 TF-IDF 结合,使得观察值输出反映了一定的语义关联信息;2) 根据训练集类标号不同,可以实现多类别的自动文本分类;3) 分类器训练好以后,进行分类所需时间主要集中在文档预处理等环节,对于多个待分类文档,其处理时间是线性增长的,因而分类效率较高。

## 4 结语

本文提出了基于隐马尔可夫模型的文本分类方法。在训练阶段,每个类别的文档集合用于训练一个分类器。分类器中的状态转换表示文本由不相关内容向相关内容的逐渐转换,而输出序列则用文本的特征词的有序集合表示。在特征选择阶段和观察输出概率分布学习阶段,采用了多种特征选择和文档相关性评估方法。在分类阶段,首先寻找待分类文档的特征词集,然后代入各个类别的 HMM 分类器,利用前向或后向算法进行评估,最后采用最大概率的类标号作为分类结果。

模型还有很多需要改进的内容,如:1) 特征词选择用人工设定阈值的形式,没有自动学习的机制;2) 待分类文档的表示是使用特征词集的统计特征,没有深入考虑文本本身的语义信息;3) 用不同类别训练集训练不同的分类器,容易造成过学习的现象。下阶段的工作主要包括:将文本的本体表示和本体进化功能与 HMM 分类器的参数学习和分类相结合;实践检验算法模型的效率,与其他经典文本分类算法进行比较;进一步改进模型的状态转移概率和观察输出概率分布的学习算法,解决过学习问题。

(下转第 2361 页)

对于通道对  $\langle r2i, a2i \rangle$ , 有  $0 \leq (s \downarrow r2i - s \downarrow a2i) \leq 1$ , 满足基本 Client-Server 进程的条件 3) 中的 c)、d)。

对任意的  $s \in \text{traces}(\text{GUEST}_i)$ , 有

$$s = \langle R1i, R2i, A2i, A1i, R1i, R2i, A2i, A1i, \dots \rangle$$

对于通道对  $\langle R1i, A1i \rangle$ , 有  $0 \leq (s \downarrow R1i - s \downarrow A1i) \leq 1$ 。

对于通道对  $\langle R2i, A2i \rangle$ , 有  $0 \leq (s \downarrow R2i - s \downarrow A2i) \leq 1$ , 满足基本 Client-Server 进程的条件 3) 中的 c)、d)。

明显地, 对进程 GLUE, 基本 Client-Server 进程的条件 3) 中的 c)、d) 也满足。

总之, 增加了设计准则的连接子 XEN\_BLOCK\_IO 不死锁。

#### 4 实验和结论

以第3章提出的约束 IDD 和 GUEST<sub>i</sub> 交互行为的设计准则为指导, 在 IDD 和 GUEST 的 OS 内核的标准实现中增加了相关程序代码。

设计了两个实验, 用 I/O 和文件系统性能测评工具 Iozone<sup>[8]</sup> 来评估由于优化实现而带来的块设备 I/O 吞吐量变化。Iozone 的输入是文件大小 (KB), 输出则是创建、读、写等文件操作所产生的 I/O 吞吐量 (KBps)。两个实验都是比对 Iozone 的随机读、写文件操作在优化设计 (IO-OPT) 和标准实现 (IO) 上所产生的 I/O 吞吐量。

实验的硬件平台是带一个 P4 1.8 GHz CPU, 256 MB RAM, 80 GB 硬盘, 100 Mbps 以太网卡的 PC 机。为简化实验而又不失一般性, 把控制虚拟机和 IDD 合二为一。

第一个实验配置如下虚拟机。IDD (控制虚拟机) 有 128 MB RAM, 10 GB 虚拟硬盘; 客户虚拟机 GUEST1 有 50 MB RAM, 10 GB 虚拟硬盘。

IDD 运行带 Linux-2.6.9-xen0 (特权) 内核映象和 Redhat 9.0 发布的客户 OS, 客户虚拟机 GUEST1 运行带 Linux-2.6.9-xenU (非特权) 内核映象和相同发布的客户 OS。如表 1 所示。

第二个实验增加一个客户虚拟机 GUEST2, 配置与 GUEST1 相同。GUEST1、GUEST2 同时运行 Iozone, 文件大小为 10 MB, 以模拟多个客户虚拟机并发、频繁访问 IDD 的场景。如表 2 所示。

实验表明, 与标准实现相比, 优化实现会带来块设备 I/O 吞吐量降低, 文件较小时 (小于 256 KB), 降低平均为 7%, 文件较大时 (大于 256 KB), 降低平均为 4%, 而且有文件越大、I/O 吞吐量降低越小的趋势。

优化的目的在于防止并发系统死锁, 提高系统可靠性, 而不是提高 I/O 吞吐量。尽管 I/O 吞吐量略有降低, 但与提高

系统可靠性相比, 还是值得的。

表 1 随机读写操作 I/O 吞吐量对比 KBps

文件 大小/KB	IO (随机读)	IO-OPT (随机读)	IO (随机写)	IO-OPT (随机写)
4	666 673	606 672	571 423	513 709
8	727 274	667 638	666 667	600 000
16	888 887	897 776	761 902	739 045
32	941 175	865 881	888 890	773 334
64	969 698	901 819	914 286	857 600
128	1 007 873	932 283	948 148	891 259
256	977 099	918 473	948 148	891 259
512	881 239	845 989	766 466	743 472
1 024	828 478	811 908	595 348	586 418
10 240	763 267	755 634	322 438	316 312

表 2 并发随机读写 10 MB 文件 I/O 吞吐量对比 KBps

客户机	IO (随机读)	IO-OPT (随机读)	IO (随机写)	IO-OPT (随机写)
GUEST1	755 440	725 222	315 805	306 331
GUEST2	400 046	392 045	215 189	204 430

#### 参考文献:

- [1] GOLDBERG R. Survey of virtual machine research [J]. IEEE Computer, 1974, 7(6): 34-45.
- [2] BARHAM P, DRAGOVIC B, FRASER K, *et al.* Xen and the art of virtualization [C]// Proceedings of the 19th ACM Symposium on Operating Systems Principles. New York: ACM Press, 2003: 164-177.
- [3] HOARE C. Communicating sequential processes [M]. Upper Saddle River, New Jersey: Prentice-Hall, 1991.
- [4] ALLEN R, GARLAN D. A formal basis for architectural connection [J]. ACM Transactions on Software Engineering and Methodology, 1997, 6(3): 213-249.
- [5] MARTIN J. The design and construction of deadlock-free concurrent systems [D]. Buckingham, UK: University of Buckingham, 1996.
- [6] EAST I, MARTIN J, WELCH P, *et al.* Prioritised service architecture [EB/OL]. [2009-12-10]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.8941&rep=rep1&type=pdf>.
- [7] FRASER K, HAND S, NEUGEBAUER R, *et al.* Reconstructing IO, UCAM-CL-TR-596 [R]. Cambridge, UK: University of Cambridge, 2004.
- [8] IOzone filesystem benchmark [EB/OL]. [2010-01-12]. <http://www.iozone.org/>.

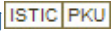
(上接第 2350 页)

#### 参考文献:

- [1] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.
- [2] 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.
- [3] 王雷, 陈治平, 李志成. 基于文本分块的多模板隐马尔可夫模型的文本信息抽取[J]. 山东大学学报: 理学版, 2006, 41(3): 21-24.
- [4] 刘云中, 林亚平, 陈治平. 基于隐马尔可夫模型的文本信息抽取[J]. 系统仿真学报, 2004, 16(3): 507-510.
- [5] 肖基毅, 邹腊梅, 李传琦. 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.

- [6] 周顺先, 林亚平, 王耀南. 基于主动学习隐马尔可夫模型的文本信息抽取[J]. 湖南大学学报: 自然科学版, 2007, 34(6): 74-77.
- [7] 施德明, 林洋港, 陈恩红. 一种集成 NER 的文本分类特征选择方法[J]. 计算机工程与科学, 2007, 29(11): 152-156.
- [8] 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.
- [9] 吴芬芬, 刘磊, 肖宪. 一种启发式的信息抽取算法[J]. 吉林大学学报: 理学版, 2007, 45(1): 73-76.
- [10] FRASCONI P, SODA G, VULLO A. Hidden Markov models for text categorization in multi-page documents [J]. Journal of Intelligent Information Systems, 2002, 18(2): 195-217.
- [11] 罗双虎, 欧阳为民. 基于隐 Markov 模型的文本分类[J]. 计算机工程与应用, 2007, 43(30): 179-181, 227.

# 基于隐马尔可夫模型的文本分类算法

作者: 杨健, 汪海航, YANG Jian, WANG Hai-hang  
作者单位: 杨健, YANG Jian (同济大学电子与信息工程学院, 上海, 201804; 大理学院数学与计算机学院, 云南, 大理, 671003), 汪海航, WANG Hai-hang (同济大学电子与信息工程学院, 上海, 201804)  
刊名: 计算机应用   
英文刊名: JOURNAL OF COMPUTER APPLICATIONS  
年, 卷(期): 2010, 30(9)

## 参考文献(11条)

1. 王雷;陈泊平;李志成 基于文本分块的多模板隐马尔可夫模型的文本信息抽取[期刊论文]-山东大学学报(理学版) 2006(03)
2. 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[期刊论文]-计算机应用 2008(03)
3. RABINER L R A tutorial on hidden Markov models and selected applications in speech recognition[外文期刊] 1989(02)
4. 罗双虎;欧阳为民 基于隐Markov模型的文本分类[期刊论文]-计算机工程与应用 2007(30)
5. FRASCONI P;SODA G;VULLO A Hidden Markov models for text categorization in multi-page documents 2002(02)
6. 吴芬芬;刘磊;肖宪 一种启发式的信息抽取算法[期刊论文]-吉林大学学报(理学版) 2007(01)
7. 周顺先;林亚平;王耀南 基于二阶隐马尔可夫模型的文本信息抽取[期刊论文]-电子学报 2007(11)
8. 施德明;林洋港;陈恩红 一种集成NER的文本分类特征选择方法[期刊论文]-计算机工程与科学 2007(11)
9. 周顺先;林亚平;王耀南 基于主动学习隐马尔可夫模型的文本信息抽取[期刊论文]-湖南大学学报(自然科学版) 2007(06)
10. 肖基毅;邹腊梅;李传琦 混合遗传算法和隐马尔可夫模型的Web信息抽取[期刊论文]-计算机工程与应用 2008(18)
11. 刘云中;林亚平;陈泊平 基于隐马尔可夫模型的文本信息抽取[期刊论文]-系统仿真学报 2004(03)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_jsjyy201009020.aspx](http://d.g.wanfangdata.com.cn/Periodical_jsjyy201009020.aspx)