[HMM中文分词](http://www.cnblogs.com/xiongmao-cpp/p/4910368.html)

**一、研究背景**

因为中文句子的特点——没有分隔符来分离句子中的词，所以在进行中文处理的时候，首先要做的就是如何对中文语句进行分词。系统的主要的内容就是建立隐马尔科夫模型，用《人民日报语料库》进行训练得到模型参数，然后再用维特比算法求出最可能的隐含序列，最后将输入的句子分成一个个词的形式。

**二、**[模型方法](http://wenku.baidu.com/link?url=zgMeHvVxzxNyVe4pGtJc4FQhZzFvluitlzjr8QKKIzkGqwkie8GWjJH8DadUX3_E3QT7SWXf5DhjCwYZmZ0hPqxUMmrA29_pEx_nqIpTMJW)

目标：给每个词打上最有可能的词状态标注，根据词的状态标注切分字串。（单字成词切开，遇到词尾切开），于是问题就变成了对最大概率词状态序列的求解，这正是隐马尔可夫模型要解决的三个问题之一。本工程主要使用的是隐马尔科夫模型和维特比算法。隐马尔科夫模型是一个统计模型，它可以用一个5元组来表示：｛S，O，π，A，B｝。下面对隐马尔科夫模型的五元组的学术含义和工程含义进行说明，通过对比直观的了解五元组在实际工程中的含义：

|  |  |  |
| --- | --- | --- |
| HMM五元素 | 学术含义 | 工程含义 |
| S | 隐含转态 | 词中4种状态：词头、词中、词尾、单字成词 |
| O | 观察状态 | 语料库中的全部汉字 |
| π | 初始状态概率矩阵 | 各种隐含状态的初始概率 |
| A | 隐含状态转移概率矩阵 | 4种隐含状态的转移概率 |
| B | 观察状态转移概率矩阵 | 每一个汉字到四种状态的概率 |

O：一个可观察序列,对于一个分词实例就是一个汉字字串，其中𝑂1，𝑂2，…，𝑂𝑘代表字串的每个字。（之后的k就代表单字总数）

S：S=｛/B、/M、/E、/S｝

{词头（[/B]Begin）、词中（[/M]Middle）、词尾（[/E]End）和单字成词（[/S]Single）}

π=｛P(我|B)、P(我|M)、P(我|E)、P(我|S)｝。假设输入的语句为：我是中国人

A矩阵就是对语料库求一个状态转移概率。将语料库中所有的表示状态的英文字母按顺序提出，存放到一个字符串中，然后进行逐字扫描，观察每个英文字母后面跟着的字母是什么，并分别记录出现个数，作为状态转移的概率。

A=

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | /B | /M | /E | /S |
| /B |  |  |  |  |
| /M |  |  |  |  |
| /E |  |  |  |  |
| /S |  |  |  |  |

**B矩阵的求解：B也是分析语料库求解，对特定分词实例，逐字读入，并分别对其扫描语料库，求解对于每个该字，后面跟的词状态的类别，同时统计个数作为B阵的值。**

B=

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 欢 | 迎 | 光 | 临 |
| /B |  |  |  |  |
| /M |  |  |  |  |
| /E |  |  |  |  |
| /S |  |  |  |  |

上述涉及到的概率均可从语料库中根据统计得到。

**三、系统设计**

1、语料库处理

(1)去掉原语料库中的词性(2)统计每个状态中出现的字及其个数

(3)统计状态间的转换，求得状态转移矩阵

A、统计出各个状态间转换在语料库中出现的次数及状态转换的总次数，计算出对应的概率

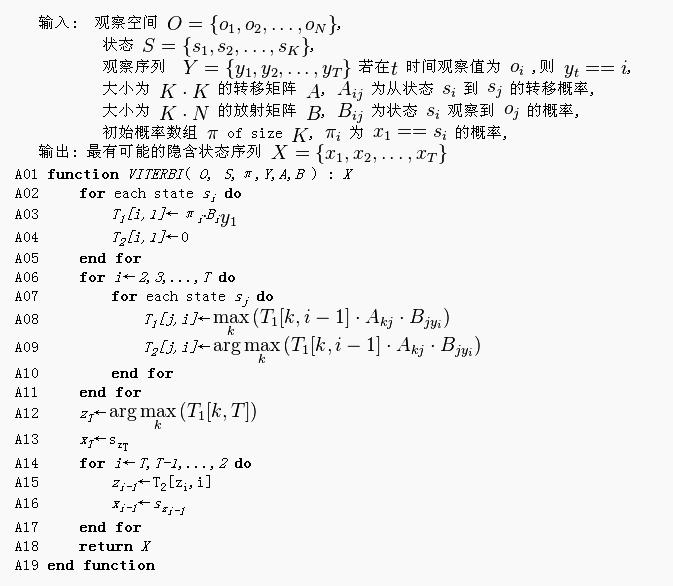
B、该过程结束后可以得到4\*4的状态转移矩阵

2、viterbi算法解码，求最佳隐含序列

（1）维特比算法是一种动态规划算法。在本工程中，通过当前状态的前一个状态，计算出在前面状态出现的条件下出现当前状态的概率，并取最大值作为当前状态出现的概率。通过迭代可以计算出到最后一个字时，哪个状态出现的概率最大。最后通过回溯得到最佳的隐含状态序列。

（2）Viterbi算法伪代码如下：

定义两个n×k维矩阵𝑇1，𝑇2。𝑇1(𝑖,j)保存前一词的所有可能状态到当前词的状态j的最大概率，𝑇2保存了这个最大概率时对应的前一个词的状态。



四，系统演示与分析

五，参考资料

1、http://www.tuicool.com/articles/FRZ77b 利用统计进行中文分词与词性分析

2、基于N最短路径和隐马尔科夫模型的中文POI分词系统的研究 唐霄

3、基于逆向隐马尔可夫模型的中文分词方法研究

4、http://blog.csdn.net/sight\_/article/details/43307581  隐马尔科夫模型详解