基于 Maximum Likelihood 与 HMM 的文本挖掘

邹腊梅,肖基毅,龚向坚
(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:随着信息技术、数据库技术、网络技术的发展,各行各业均存储了大量的文本数据,怎样从这些文本数据中发掘有价值的信息和知识成为人们急需解决的问题。提出基于 Maximum Likelihood 与 HMM 的文本挖掘方法,利用 Maximum Likelihood 构建隐马尔可夫模型,对论文条目进行特定信息的发掘,并克服了实验过程中"零概率"的缺陷。实验结果表明准确率平均达到0.9,召回率平均达到0.85,从理论和实践上证明该方法是有效的。

关键词:隐马尔可夫模型;最大似然;文本挖掘;信息抽取
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2007)12-0110-03

Text Information Mining Based on Maximum Likelihood and Hidden Markov Model

ZOU La-mei, XIAO Ji-yi, GONG Xiang-jian
(Department of Computer Science and Technology, Nanhua University, Hengyang 421001, China)

Abstract: With the development of information technology, database technology and network technology, a large number of texts are produced in all kinds of fields, the question should be solved quickly that how to mine useful information and knowledge from texts. Introduces how to mine information using maximum likelihood and hidden Markov model. It constructs HMM with maximum likelihood and mines customizing messages from thesis entries with HMM. During the process of extracting, it deals with the questing of "zero probability". The experiment results indicate that the average precise rate arrives to 0.9 and the average recall rate arrives to 0.85. Both in theory and in practice the method are effective.

Key words: hidden Markov model; maximum likelihood; text mining; information extraction

0 引 言

随着信息技术、数据库技术、网络技术的快速发展,不同领域均产生和存储了大量的文本数据,怎样从这些文本数据中挖掘出有用的信息已成为一个人们急切需要解决的问题,也是促使文本挖掘技术不断发展的动力。

文本挖掘(Text Mining)是指从文本数据中抽取有价值的信息和知识的计算机处理技术。文本数据挖掘主要包括文本信息抽取、文本分类、文本聚类、文本数据压缩、文本数据处理五大技术。文本信息抽取就是从大量的文本数据中抽取人们关注的特定的信息。文本信息抽取的方法有:最大熵方法(Maximum Entropy)、隐马尔可夫模型方法(Hidden Markov Model, HMM)[1,2]、条件随机场方法(CRFs)、基于核(kernel)的机器学习方法。文中将应用隐马尔可夫模型方法对文本信息进行抽取,实验结果表明该方法是有效的。

收稿日期:2007-03-03
基金项目:湖南省自然科学基金资助项目(04JJ40051);湖南省教育厅资助项目(06C724)
作者简介:邹腊梅(1977-),女,湖南衡阳人,硕士,讲师,研究方向为数据挖掘。

1 隐马尔可夫模型

20世纪60年代末,Baum 等人提出了隐马尔可夫模型的基本理论。20世纪70年代到80年代,Baker 等人将隐马尔可夫模型应用到语音信号处理领域,并逐渐成为语音识别领域居主导地位的方法。从90年代初起,隐马尔可夫模型开始被用于图像信号处理以及视频信号处理等领域。到目前为止,隐马尔可夫模型及其各种推广形式在语音识别、计算机语言学、基因识别、命名实体识别等领域取得了一定的成果[3~7]。

定义 一个 HMM 为一个五元组:$\lambda=(S,O,\boldsymbol{A},\boldsymbol{B},\pi)$

(1)S 表示模型中状态集合,共 N 个状态。虽然状态是隐藏的,但这些状态之间相互联系,而且可以从一种状态转移到其它状态。所有独立的状态定义为 $S=$

$[S_1, S_2, \cdots, S_N]$,且用 q_t 来表示 t 时刻的状态。

(2)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。观察值对应于模型系统的实际输出,记为:$W = [w_1, w_2, \cdots, w_N]$。

(3)$\boldsymbol{A} = \{a_{ij}\}$ 为状态转移概率矩阵,表示从状态 i 转移到状态 j 的概率。

$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i), 1 \leqslant i, j \leqslant N$,$a_{ij}$ 满足约束条件:$\forall\ i, j, a_{ij} \geqslant 0$,且 $\forall\ i, \sum_j a_{ij} = 1$

(4)$\boldsymbol{B} = \{b_j(k)\}$ 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。$b_j(k) = P(O_t = W_k \mid q_t = S_j), 1 \leqslant j \leqslant N, 1 \leqslant k \leqslant M$,$b_j(k)$ 满足约束条件:$\forall\ j, k, b_j(k) \geqslant 0$,且 $\forall\ j, \sum_k b_j(k) = 1$。

(5)$\boldsymbol{\pi} = \{\pi_i\}$ 为初始状态分布向量,表示 $t = 1$ 时处于状态 S_i 的概率。

$\pi_i = P(q_l = S_i), 1 \leqslant i \leqslant N$,$\pi_i$ 满足约束条件 $\sum_i \pi_i = 1$。

为简便起见,一般使用形式 $\boldsymbol{\lambda} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$ 来描述 HMM 模型。HMM 模型包含双重随机过程,一个是描述状态转移的基本随机过程,另一个描述状态和观察值之间的统计对应关系。由于 HMM 的状态是隐藏在内部的,所以观察者只能看到观察值,不能直接看到状态。观察者只能通过对外部观察序列的研究,去感知内部状态的存在及其特性。

在隐马尔可夫模型方法应用于特定信息发掘过程中,把待提取的每个语义项称之为域,每个域对应一个或多个状态,原始文本中的符号作为状态的输出符号。隐马尔可夫模型方法是一种有指导的机器学习过程,整个过程分为模型训练(学习问题)和信息抽取两个过程。模型训练过程为:收集若干标记的数据样本(训练数据集),每一个样本由文字单元序列及相应的标记序列(状态序列)组成。机器学习系统基于已有的标记数据构建一个模型,由一个条件概率分布来表示。信息提取过程为:信息抽取系统参照学习得到的条件概率分布模型,搜索最可能创建符号序列的状态序列,也就是说对一些新的文字单元序列(测试数据集)进行域标记的过程。为了完成模型的构建和信息抽取的过程,需要解决两个问题:

① 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 $\boldsymbol{\lambda}$,使得观察值出现的概率 $P(O \mid \boldsymbol{\lambda})$ 最大。在此将采用 Maximum Likelihood(最大似然)算法。

② 解码问题:对于给定模型 $\boldsymbol{\lambda} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}\}$ 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。文中将采用 Viterbi 算法[8]。

2　实验数据

文中将隐马尔可夫模型应用于对论文头部不同域的信息提取,一条论文头部描述如下:

Title = "Implementing Distributed Server Groups for the World Wide Web".

Author = "Michael Garland. Robert Monroe. Siddhartha Puri".

Date = "25 January 1995".

Affiliation = "School of Computer Science Carnegie Mellon University".

进一步将该论文头部标记为:

<title> Implementing </title><title> Distributed </title> <title> Server </title><title> Groups </title><title> for </title><title> the </title><title> World </title> <title> Wide </title><title> Web </title> <author> Michael </author> <author> Garland </author><author> Robert </author><author> Monroe </author><author> Siddhartha </author> <author> Puri </author> <date> 25 </date> <date> January </date> <date> 1995 </date> <affiliation> School </affiliation><affiliation> of </affiliation><affiliation> Computer </affiliation> <affiliation> Science </affiliation> <affiliation> </affiliation> Carnegie <affiliation> Mellon <affiliation> University </affiliation>

一条论文头部包含 4 个域,域名及含义如表 1 如示。

表 1　数据域名及含义

域名	含义
title	论文的标题
author	论文的作者
date	论文发表的年月
affiliation	作者所在单位

3　隐马尔可夫模型参数设定

确定隐马尔可夫模型的结构时,定义 $N = 4$,M 为训练数据集中包含的不同词汇数。由于训练数据集已经逐一标记了,在确定模型的 $\pi_i, a_{ij}, b_j(k)$ 三个参数时,使用 Maximum Likelihood 方法来估计。

(1)$\pi_i = P(q_1 = S_i) =$

$$\frac{\text{文本序列中首个单词属于状态 } S_i \text{ 出现的次数}}{\sum_{i=1}^{4} \text{文本序列中首个单词属于状态 } S_i \text{ 出现的次数}}$$

(2)$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i) =$

$$\frac{\text{状态 } S_i \text{ 转移到状态 } S_j \text{ 的次数}}{\sum_{j=1}^{M} \text{状态 } S_i \text{ 转移到状态 } S_j \text{ 的次数}}$$

(3)$b_j(k) = P(O_t = W_k \mid q_t = S_j) =$

$$\frac{\text{状态 } S_j \text{ 释放单词 } W_k \text{ 的次数}}{\text{文本序列中状态 } S_j \text{ 出现的总的次数}}$$

当训练数据集中没有包含测试数据集中出现的单词的时候,就会出现“零概率”的问题(即在训练数据集中未出现的单词的概率为零),这时需要进行概率的平滑。常用的平滑方法有加一法、绝对减值法、线性减值法、绝对折扣法。考虑到在文本序列中,前后单词属于同一状态的可能性很大,同时当单词量较大时,出现“零概率”的单词数不多。所以在实验过程中,笔者采用当出现某个单词概率为零时,就使用后面单词的概率来代替该单词概率的方法。经过实验表明,发现该方法更简单、效果也很好。

4 实验评价标准

在 1991 年第三届 MUC(Message Understanding Conference)会议中制定了恒定信息抽取性能的评价指标:召回率(REC)和准确率(PRE)。文中采用的评价指标定义如下:

$$\text{REC}=\frac{\text{Correct}-\text{Marked Words}}{\text{Words in Data Set}}$$

$$\text{PRE}=\frac{\text{Correct}-\text{Marked Words}}{\text{Words in Extracted Result Set}}$$

Correct - Marked Words 为提取结果中本域标记正确的单词个数,Words in Data Set 为数据集中本域总的单词个数,Words in Extracted Result Set 为提取结果中本域总的单词个数。

5 实验结果及结论

文中所使用的实验数据集共包括 944 篇论文头部,每次选用其中一部分作为训练数据集,一部分作为测试集。得到两组实验结果如下:

第一组实验数据使用 500 条论文头部作为训练集,共包含单词 11472 个。100 条作为测试集,共包含单词 2203 个。实验结果如表 2 所示。

表 2 第一组数据实验结果

域名 / 单词个数	title	author	date	affiliation
Correct - Marked Words	749	557	85	621
Words in Data Set	806	621	93	683
Words in Extracted Result Set	835	575	101	691
REC	0.929	0.897	0.914	0.909
PRE	0.897	0.969	0.842	0.899

第二组实验数据使用 600 条论文头部作为训练集,共包含单词 13909 个。200 条作为测试集,共包含单词 4934 个。实验结果如表 3 所示。

实验结果看出,四个域的准确率、召回率都较高,表明 Maximum Likelihood 与 HMM 应用于已标记数据集的信息抽取是有效的。其中以 title 的准确率最高,平均准确率达到 0.934,author 的平均召回率最大,达到 0.961。随着训练集和测试集的语料量变大,date 的召回率有所提升,原因在于 date 域本身包含的数据量不大,随着测试集数据的递增,date 的召回率相应增加。

表 3 第二组数据实验结果

域名 / 单词个数	title	author	date	affiliation
Correct - Marked Words	1705	1182	164	1356
Words in Data Set	1816	1425	175	1518
Words in Extracted Result Set	1907	1242	232	1609
REC	0.939	0.829	0.937	0.893
PRE	0.894	0.952	0.707	0.843

基于最大似然和隐马尔可夫模型用于文本的信息抽取的实验获得不错的实验效果。但实验所需要的标记语料库需要大量的人工标记时间,怎样从大量的文本数据中选择一定具有代表性的文本作为训练样本,从而减少实验所需的时间和人的干预,更好地提高实验的效率是以后研究的重点。

参考文献:

[1] Frietag D, McCallum A. Information Extraction with HMMs and shrinkage[C]//In Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction. Orlando, US: AAAI Press/The MIT Press, 1999:31 - 36.

[2] Scheffer T. Decomain C, Wrobel, S. Mining the Web With Active Hidden Markov Models Data Mining[C]//Proceedings of the 2001 IEEE International Conference on Data Mining. Washington, DC, USA: IEEE Computer Society, 2001:645 - 646.

[3] Rabiner L. A tutorial on hidden markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2):257 - 285.

[4] Li J, Najmi A, Gray R M. Image classification by a two - dimensional hidden Markov model[J]. IEEE Transactions on Signal Processing, 2000, 48(2):517 - 523.

[5] 李 珩,杨 峰. 基于增益的隐马尔可夫模型的文本组块分析[J]. 计算机科学,2004,31(2):152 - 154.

[6] 曹胜玉,刘来福. 隐马模型及其在基因识别中的应用[J]. 数学的实践与认识,2006,36(9):212 - 218.

[7] 杜世平,李 海. 二阶隐马尔可夫模型及其在计算语言学中的应用[J]. 四川大学学报:自然科学版,2004,41(2):284 - 289.

[8] 刘河生,高小榕,杨福生. 隐马尔可夫模型的原理与实现[J]. 国外医学生物医学工程分册,2002,25(6):253 - 259.