文章编号:1674－2869(2015)11－0047－05

基于条件随机场的中文地址行政区划提取方法

段艳会[1,2],李晓林[1,2*],黄　爽[1,2]
1.智能机器人湖北省重点实验室(武汉工程大学),湖北 武汉 430205;
2.武汉工程大学计算机科学与工程学院,湖北 武汉 430205

摘　要:为了在非规范中文地址中有效的提取行政区划信息,提出了一种基于条件随机场的方法. 该方法根据中文地址中行政区划的表达特点和特征,采用判别式概率模型,在观测序列已知的基础上对目标序列建模,通过构建语料训练集和建立相应的特征模板,得到行政区划的表达模型,然后使用该模型对测试集进行测试,并与标注好的测试数据进行比对,验证模型的性能. 实验表明,与最大熵模型相比,条件随机场模型总的性能指标在其之上,地址信息解析的准确率能达到 89.93%.

关键词:位置信息解析;条件随机场;训练语料
中图分类号:TP391.41　　**文献标识码**:A　　**doi**:10. 3969/j. issn. 1674-2869. 2015. 11. 010

0　引　言

一直以来,语言是人们交流与传播知识最重要的工具,而理解语言中包含的信息对人类的交流也起到了重要作用. 随着信息技术的发展,能使人与计算机更好的实现人机交互,是人们长期以来追求的目标. 本文研究对地址信息的解析,在过去很多学者也提出过许多方法去提高地址信息解析[1]的正确度. 于 20 世纪 70 年代,Lawrence 提出的作为一种统计分析的隐马尔科夫模型[2],存在着一种假设,即假设每个元素之间都是彼此独立的,不管何时观察结果仅仅依赖于此时此刻所处的状态. 而这个模型假设的前提只适合较小规模的数据集[3],在现实生活中,很多真实语料的观察序列大部分的表现形式是以多重交互特征表现,观察元素间通常有较长的相关性,即不止与前一个状态相关,可能与多个状态以前的状态有关. 在位置信息解析任务中,因为地址信息自身结构具有一定的复杂性,单一的特征函数难以涵盖它的所有特征,这种情况下,隐马尔科夫模型所提出的假设就使得它使用复杂特征时无能为力,此时它在信息解析中的弊端就显现出来了. 而后由 Jaynes 提出的最大熵模型[4-5],它记录特征是否出现,并不能了解所需特征的强度,因此在信息的分类中并不是最优的. 其次,采用最大似然方法训练出的最大熵模型,其算法存在较慢的收敛速度,所以就致使最大熵模型的计算量大,而且,有很严重的数据稀疏问题,再次由于最大熵对每个词单独进行分类,难以很充足的利用标记之间的关系. 由 Pearl 提出的贝叶斯模型[6]可以任意使用复杂的相关特征,并且能灵便地对约束条件进行设置,模型在未知参数的适应度、已知数据的拟合度方面就是通过约束条件进行调节的,并且能自然地解决模型中参数平滑问题,但其在决策时很容易延误决策的最佳时间导致影响效果. 本文提出的条件随机场模型[7],克服了其它几种模型在信息解析上的缺陷,采用链结式结构,结合了最大熵模型和马尔科夫模型的特点,使用概率图模型,在长距离依赖和交叠性特征方面能力较强,能很好的处理最大熵马尔科夫模型[8]具有的标注偏置等问题,而且与最大熵不同的是它能对所有的特征进行全局归一化[9-11],从而求得全局最优解.

1　模型描述

1.1　位置信息解析实现步骤

位置信息解析具体步骤如图 1 所示.

(1)收集数据:利用网络爬虫等数据挖掘方式,从互联网上提取大量地理位置信息,例如:…|省…|市…|(区)县…|镇…|路…类的地址,之后进行整理,保留至少有一个行政区划名的地址作为研究所需要使用的数据.

收稿日期:2015-10-13
基金项目:国家 863 项目(2013AA12A202);武汉工程大学研究生教育创新基金项目(CX2014090)
作者简介:段艳会(1993-),女,湖北公安人,硕士研究生.研究方向:数据挖掘、机器学习.* 通信联系人

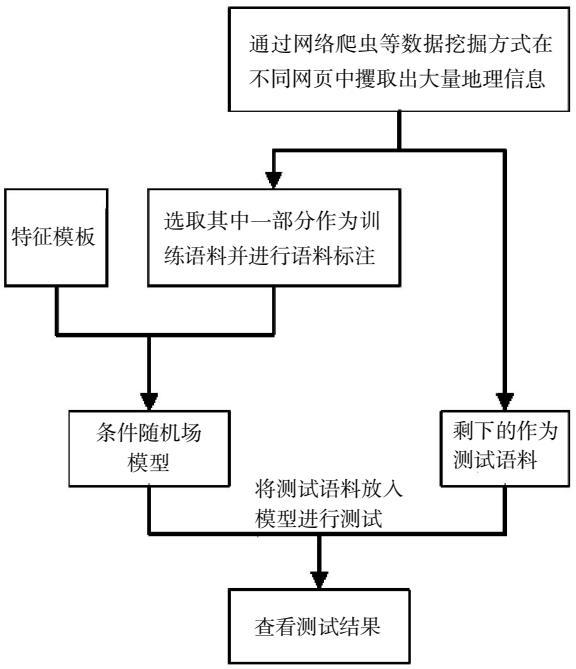

图 1 信息解析流程图

Fig.1 Flowchart of information analysis

(2)分析数据:将数据集里的一部分数据分离出作为训练语料,参考条件随机场特殊文本格式进行人工标注以便生成模型,剩下的一部分作测试语料,并人工对其进行标注,以便测试模型性能.

(3)构建模型:通过对地址信息的特殊结构进行研究制定特征模板,并与测试语料一起生成条件随机场模型.

(4)输出结果:用测试语料检测条件随机场模型,输出 output 文件,并将文本格式规范为 crf++ 特殊文本定义的格式,从而为数据比对提供方便.

(5)测试数据:查看测试结果,对模型性能进行测评.

1.2 条件随机场描述

由于最大熵模型会产生标记偏置问题,无法好好的对信息进行序列标注,针对这个问题,后期提出最大熵马尔科夫模型,是在相邻的变量值之间使用最大熵模型,但这样做会使得得到的结果仅仅是局部的最优解,对整体来说未必是最优的,针对这个问题,提出的条件随机场方法,引出归一化因子,对全局进行归一化,解决了最大熵隐马尔科夫模型存在的不足之处.

条件随机场也是一种判别式模型,所谓判别式模型,学习的是一种条件概率 $p(y|x)$,利用正负例和分类标签,关注在判别模型的边缘分布.

条件随机场是一种无向图性模型,变量 X,Y 分别代表的是观察序列、对应联合分布的随机变量,则以 X 为条件的无向图模型 $G(V,E)$就是条件随机场. 其中 V 表示无向图 G 中节点的集合,E 代表节点 V 之间的无向边集合.

设 $G=(V,E)$是一个无向图,$Y=\{Y_v|v\in V\}$是以无向图中每个随机变量 Y_v 为节点构成的集合,在给定观察序列的条件下,若是随机变量 Y_v 都遵循马尔科夫属性,即如公式(1)所示[6].

$$P(Y_v|X,Y_u,u\neq v)=P(Y_v|X,Y_u,u\sim v) \tag{1}$$

那么(X,Y)就构成一个条件随机场,其中 $u\sim v$ 表示 u 和 v 是相邻的边.

根据 Hammersley-Clifford 定理得到条件随机场分布用公式(2)表示如下[7].

$$p(y|x)=\frac{1}{Z}\prod_{c\in C}\Psi V_c(v_c) \tag{2}$$

这里的 c 为 G 中所有的最大势团的集合,$\Psi V_c(v_c)$为势函数(下一节介绍),通常为了方便计算,$\Psi V_c(v_c)$的形式表示在 Z 的归一化因子中,用式(3)表示.

$$Z=\sum_{v_1\cdots v_n}\prod_{c\in C}\Psi V_c(v_c) \tag{3}$$

条件随机场最常用的模型为线性链结构,如图 2 所示.

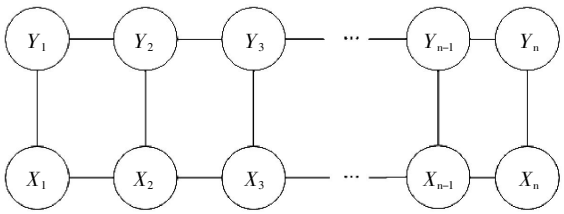

图 2 线性结构条件随机场

Fig.2 Condition random field of linear structural condition

另外观察序列之间并不一定存在某种联系,它只是一个条件,不存在其它任何假设,所以还可以用图 3 表示该模型.

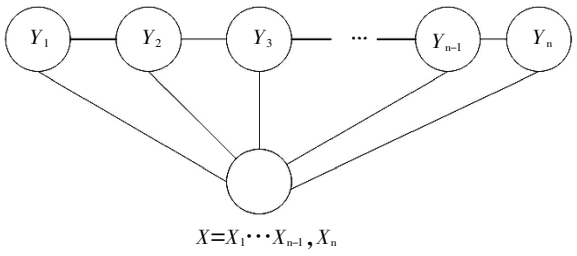

图 3 模型的另一种表示方法

Fig.3 Another expression of the model

该表示方法与线性条件随机场模型的表示方法无异.

1.3　势函数

势函数在条件随机场中是针对每个最大势团做出的定义,如图 4 所示,每个势函数定义如公式(4)[6].

$$\Psi_c(y_c)=\exp\left(\sum_k \lambda_k f_k(y_c,x_c)\right) \tag{4}$$

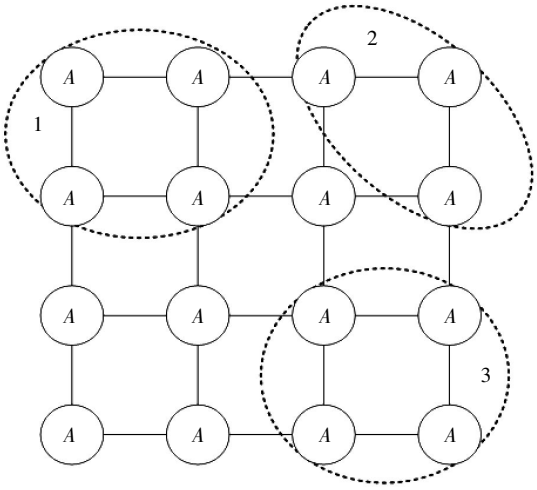

图 4　势函数团示意图

Fig.4　Schematic diagram of potential function

注:1、2、3 表示最大势团;A 表示节点变量

在图 4 中,G 中的所有最大势团已标出,且 $\Psi_c(y_c)$是一个在图 G 中最大势团上严格的势函数,因此,在观察序列 X 的条件下,标记序列 Y 的形式如公式(5)所示[8].

$$p(y|x)=\frac{1}{Z(x)}\exp\left(\sum_k \sum_{c\in C}\lambda_k f_k(y_c,x_c)\right) \tag{5}$$

在线性条件随机场中,G 中最大势团即是相邻的两个随机变量组成的,则公式(4)可以扩写成如下.

$$\Psi_c(y_c)=\exp\left(\sum_k \lambda_k f_k(y_{i=1},y_i,x)+\sum_k u_k g_k(y_i,x)\right) \tag{6}$$

其中 $f_k(y_{i=1},y_i,x)$是表示相邻的观察序列的标记位置之间的特征转移函数,而 $g_k(y_i,x)$表示当前观察位置的状态特征函数.

1.4　特征模板

如果仅仅依靠地理信息本身结构和对字符串的分析,很难取得好的效果. 因此,地理信息中,上下文环境与特征词对提高地理信息的识别效果有比较明显的作用. 而条件随机场有个非常明显的优点,即它可以很轻松的将观察序列里面的特征加入到模型中,表达长距离上下文的依赖关系.

上下文环境,是指包括当前词v_0 在内,以及其前后几个词组成的一个“观察窗口”(v_{-n},$v_{-(n-1)}$,$\cdots v_0,\cdots,v_{n-1},v_n$). 从理论上看,窗口越大,前后词关联越多,可依赖的信息就越多,但窗口过大又会导致效率降低,窗口过小又会导致不能充分利用特征,从而丢失掉一些重要信息. 所以,根据大量数据分析,一般选取的窗口大小为 2,即(v_{-2},v_{-1},v_0,v_1,v_2).

在选取使用地址位置的特征时,考虑到地址都有很明显的行政区划特征词例如省、市、区、县等,所以可以将该特征应用到特征模板中,如表 1 所示.

表 1　特征模板的选取

Table 1　Selection of feature template

序号	意义
L1	v_0 后第一个词是否为地名特征词
L2	v_0 后第二个词是否为地名特征词
L3	v_0 前第一个词是否为地名特征词
L4	v_0 前第一个词是否为地名特征词

以上的各个特征模板要有四个位置偏移,即-2、-1、1、2. 而当特征函数取得特定值,特征模板就会变成一个实例,便会有具体特征出现. 比如当前词后面第二个词 v_2 出现在地址特征词表,可表示为:

$$f_i(x,y)=\begin{cases}0 & \text{else}\\ 1 & \text{if feature}(v_2)=1\ \&\ y=\text{location}\end{cases} \tag{7}$$

2　性能评估

对于一般的分词算法研究,大多数都是以 B、I、E、O 做标注,即以一个词的开始、中间、结束来标记,没有很明显的在标注中加入词性特征,这使在建立模型时,不能很好的利用所研究领域的一些特殊结构与特征来进行更加精准的信息区分,利用条件随机场进行信息的解析,在标注语料时,同时结合地理领域的词结构进行标注,如表 2 所示.

表 2　训练语料标注示例

Table 2　Training corpus tagging examples

地名词	按字符标记
上海市	B_city、M_city、E_city
奉贤县	B_county、M_county、E_county
嘉定区	B_aera、M_area、E_area

选取省、市、县、区等行政区划关键词作为特征词,给每个特征词一个特殊的标记,并结合 B、M、E 作为标注语料的标记,实验表明,加入特征词标记语料之后,利用语义解析去标注,比纯粹的开始、结束标记的正确率增加 5%左右. 条件随机场对信息结果好坏的评估,将与最大熵模型和隐马尔科夫模型作比较,比较性能的指标有如下几个方面:

$$\text{正确率}=\frac{\text{正确解析出的信息的个数}}{\text{解析的信息的总个数}}\times 100\%$$

$$\text{召回率}=\frac{\text{正确解析出的信息的个数}}{\text{标准结果中的信息总数}}\times 100\%$$

$$F=\frac{2\times\text{正确率}\times\text{召回率}}{\text{正确率}+\text{召回率}}$$

3 实验结果及分析

3.1 定性分析

在标注语料时,有一些特殊的语句存在歧义,导致标注时模糊不清从而使得正确率下降,影响地址解析的效果. 在输出 output 文件中,有如表 3 所示例子.

表 3 输出示例

Table 3 Sample output

测试地址	嘉	定	区
输出结果	B_city	M_area	E_area
正确结果	B_area	M_area	E_area

从表 3 可以看到,在测试时,由于训练语料中有些市区的名字带有“嘉”,再通过与模板结合,所以将嘉定区中的“嘉”作为了市级的地名,从而使地址识别出现一些错误. 此时,通过最大移动窗口匹配算法,对于一些不确定的地名进行消歧处理.

移动窗口匹配算法,即首先准备行政区划表,将地址与行政区划表对应进行匹配,得到准确的行政区划信息,实验表明,加入了该算法后,对歧义的消除起到了一定的作用,比对的正确率提升了 4%左右. 当然对于最大移动窗口匹配算法只能进行部分消歧,并不能消除所有歧义,对于未消除歧义里的词,将在后续进行研究.

3.2 定量分析

通过分析地址解析的各种算法,将最大熵分词算法与本文算法相比较,由于最大熵算法知识表达能力较强,具有易于理解、简单、可重用性等优点,被广泛用于分词研究中,但其也具有耗资源、速度训练慢等缺点. 本文算法虽然全局归一化代价高,训练工作量大,但其能将各个新特征进行融合,能兼顾长距离表达,灵活性好,且分词精度高. 选取 2 千条地址进行解析,两者对比结果如表 4 所示:

表 4 结果对比

Table 4 Comparison of results (%)

算法	正确率	召回率	*F*
最大熵	84.43	83.21	83.82
条件随机场	89.93	80.78	85.1

4 结 语

信息解析一直为各个领域研究的热点,本文研究地理领域的信息解析,采用条件随机场的方法,在已知观测序列的条件下学习,从而对目标数据建模,通过学习训练数据并与特征模板结合,得到能解析地址数据的模型,解决了最大熵模型遗留的标注偏置问题,以及最大熵马尔科夫模型的局部优化问题. 它能表达长距离依赖性和交叠性,以序列化形式对全局的参数进行优化. 人工标注训练语料,并利用条件随机场工具 crf++进行语料训练,该模型的信息解析总的性能指标比其它模型优异,取得较满意的结果.

致 谢

感谢武汉工程大学研究生处的资助!

参考文献:

[1] 朱俊.中文标准地址库构建关键技术研究[D].南京:南京师范大学,2013.
ZHU Jun. Reasearch on Key Techniques of constructing Chinese standard address database [D]. Nanjing:Nanjing Normal University, 2013.(in Chinese)

[2] LAWRENCE R, RABOMER. A tutorial on hidden markov models and selected applications in speech recognition[J]. Proceedings of the IEEE,1989,77 (2):257-286.

[3] 申彦.大规模数据集高效数据挖掘算法研究[D].镇江:江苏大学,2013.
SHENG Yan. Research on efficient data mining algorithm for large scale data sets [D]. Zhengjiang:Jiangsu University,2013.(in Chinese)

[4] 周鑫.半监督算法在自然语言处理中应用的研究[D].哈尔滨:哈尔滨工业大学,2014.
ZHOU Xin. Research on Application of semi supervised algorithm in natural language processing [D]. Harbin:Harbin Institute of Technology,2014(in Chinese)

[5] MCCALLUM A, FREITAG D, PEREIRA F. Maximum

Entropy Markov Models for Information Extraction and Segmentation[C]//Proc JcML,2000:591-598.

[6] PEARL J. Probabilistic reasoning in intelligent systems: networks of plausible inference [C]//1th ed, San Mateo, CA: Morgan Kaufmann,1988:117-133.

[7] LAFFERTY J, MCCAI LUMA, PEREIRA F. Conditional Random Fields:Probabilistic Models for Segmenting and Labeling Sequence Data [C]//Proc ICML, 2001.

[8] THOMPSON JD,HIGGINS DG,GIBSON TJ,et al. Improving the sensitivity of progre- ssive multiple sequence alignment through sequence weighting,position specific gap penalties and weight matrix choice [J]. Nucleic Acids Research,1994,22(22):4673-4680.

[9] JIAYI Zhao,XIPENG Qiu,SHU Zhang. Part-of-Speech Tagging for Chinese-English Mixed Texts with Dynamic Features[J]. Journal of Computational Information Systems(JCIS),2012:1379-1388.

[10] 田昕辉,李成基.带有短语切分的中文文本分类方法[J].计算机技术与发展,2010,20 (1):9-13.
TIAN Xin-hui,LI Cheng-ji. Chinese text classification method with phrase segmentation[J]. Computer Technology and Development,2010,20(1):9-13.(in Chinese)

[11] SUN X L, JIA L M, DONG H H, et al. Urban expressway traffic state forecasting based on multimode maximum entropy model [J]. Science China Technological Sciences, 2010, 53(10): 2808-2816.

Extraction of administrative division of Chinese address based on conditional random fields

***DUAN Yan-hui*[1,2],*LI Xiao-lin*[1,2],*HUANG Shuang*[1,2]**

1.Hubei Key Laboratory of Intelligent Robot(Wuhan Institute of Technology), Wuhan 430205, China;
2.School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China

Abstract:To extract the information of administrative division effectively from the non-standard Chinese address, a method based on conditional random fields was proposed. According to the characteristics of administrative division, the model of the target sequence was constructed on the basis of the observation sequence by using the discriminative probability model. Then, the expression model of the administrative division was obtained by constructing the corpus training set and the corresponding feature template. Finally, the performance of the model was verified by testing the test set and comparing its results with the marked test data. Experimental results show that the performance of the model is better than that of the maximum entropy model, and the accuracy rate of analysis of address information reaches 89.93%.

Keywords: location information parsing, condition random fields, training corpus

本文编辑:陈小平