ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.2, February 2008, pp.358−368
DOI: 10.3724/SP.J.1001.2008.00358

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

SVM+BiHMM:基于统计方法的元数据抽取混合模型*

张 铭+, 银 平, 邓志鸿, 杨冬青

(北京大学 信息科学技术学院,北京 100871)

SVM+BiHMM: A Hybrid Statistic Model for Metadata Extraction

ZHANG Ming+, YIN Ping, DENG Zhi-Hong, YANG Dong-Qing

(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

+ Corresponding author: Phn: +86-10-62765825, Fax: +86-10-62765822, E-mail: mzhang@db.pku.edu.cn, http://db.pku.edu.cn/mzhang

Zhang M, Yin P, Deng ZH, Yang DQ. SVM+BiHMM: A hybrid statistic model for metadata extraction. *Journal of Software*, 2008,19(2):358−368. http://www.jos.org.cn/1000-9825/19/358.htm

Abstract: This paper proposes SVM+BiHMM, a hybrid statistic model of metadata extraction based on SVM (support vector machine) and BiHMM (bigram HMM (hidden Markov model)). The BiHMM model modifies the HMM model with both Bigram sequential relation and position information of words, by means of distinguishing the beginning emitting probability from the inner emitting probability. First, the rule based extractor segments documents into line-blocks. Second, the SVM classifier tags the blocks into metadata elements. Finally, the SVM+BiHMM model is built based on the BiHMM model, with the emitting probability adjusted by the Sigmoid function of SVM score, and the transition probability trained by Bigram HMM. The SVM classifier benefits from the structure patterns of document line data while the Bigram HMM considers both words' Bigram sequential relation and position information, so the complementary SVM+BiHMM outperforms HMM, BiHMM, and SVM methods in the experiments on the same task.

Key words: metadata extraction; rule based information extraction; SVM (support vector machine); HMM (hidden Markov model); BiHMM (bigram hidden Markov model)

摘 要: 提出了一种 SVM+BiHMM 的混合元数据自动抽取方法.该方法基于 SVM(support vector machine)和二元 HMM(bigram HMM(hidden Markov model),简称 BiHMM)理论.二元 HMM 模型 BiHMM 在保持模型结构不变的前提下,通过区分首发概率和状态内部发射概率,修改了 HMM 发射概率计算模型.在 SVM+BiHMM 复合模型中,首先根据规则把论文粗分为论文头、正文以及引文部分,然后建立 SVM 模型把文本块划分为元数据子类,接着采用 Sigmoid 双弯曲函数把 SVM 分类结果用于拟合调整 BiHMM 模型的单词发射概率,最后用复合模型进行元数据抽取.SVM 方法有效考虑了块间联系,BiHMM 模型充分考虑了单词在状态内部的位置信息,二者的元数据抽取结果得到了很好的互补和修正,实验评测结果表明,SVM+BiHMM 算法的抽取效果优于其他方法.

关键词: 元数据抽取;基于规则的信息抽取;支持向量机;隐马尔科夫模型;二元 HMM 模型

* Supported by the National Natural Science Foundation of China under Grant Nos.90412010, 60573166 (国家自然科学基金); the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No.2007108 (高等学校博士学科点专项科研基金); the HP University Collaborative Foundation of China under Grant No.HLCFY08-001 (惠普大学合作基金)

Received 2006-03-28; Accepted 2007-06-07

中图法分类号: TP311 文献标识码: A

元数据是“关于数据的结构化数据”,为数字图书馆和语义网提供了一种精确描述数据内容、语义和服务的机制.元数据可以帮助公司更有效地查找、验证、组织资源,有效地组织元数据每年每个员工大约可以节省 8 200 美元的相应开销.但是,若给 100 万个文档标注元数据,则需要 60 人年的工作,标注元数据代价昂贵[1].元数据的自动抽取,意味着在无人工干预的情况下,依据文档元数据规范提取用户感兴趣的元数据信息,并将提取结果进行合理组织与存储的全过程.目前在语义网和数字图书馆的相关研究中,元数据标注仍然是一个公认的瓶颈和难题,尤其对于元数据自动抽取的技术研究非常必要,因为采用手工标注的方法工作量大,而且容易出错.

1 元数据自动抽取策略

论文是自由格式的元数据,元数据的抽取有一定的难度.但论文的信息也有一定的规律.一篇数字化科技文档通常由标题(含副标题)、作者信息、文档摘要、关键词、文档主体、引文、附录等部分组成,但也未尽然.一些数字化文档还包括目录、写作背景、致谢等其他内容,可以在提取过程中根据具体处理的文档集合扩充定义.有些元数据内容是可选出现的,同一内容也可能在文档中出现多次.元数据抽取是信息抽取在论文元数据自动获取中的一种应用,在信息抽取的研究领域,有两条主要的技术路线:基于规则的路线与基于统计模型的路线.

1.1 基于规则的信息抽取

规则的获取一般通过两种途径:手工制定和自动生成.手工制定通常需要专业的人员,通常称他们为知识工程师(knowledge engineer).由于信息源可能会有比较迅速的变化(如网页上的信息)以及信息源格式的多样性(如来自多个网站),因此,知识工程师的工作是很繁琐的,也可以说是一项很繁重的体力劳动.为了把工程师从这种困境中解放出来,近几年,人们逐渐发展出一些可以通过某些方法自动生成抽取规则的系统,这方面的研究通常称为 Wrapper Generation[2].Cornell 大学数字图书馆项目组实现的 OpCit 引文元数据提取模块“至少 80%的准确率”[3].Klink 等人对 979 篇期刊的测试结果是:page-number: 90% recall, 98% precision;abstract: 35% recall, 90% precision;biography: 80% recall, 35% precision[4].Kim 等人对 76 篇生物医学文档进行测试,准确率为:title 100%,author 95.64%,abstract 95.85%,affiliation 63.13%[5].我们在 PKUSpace 中应用规则来进行抽取,准确率超过 83.3%[6,7].当前,上述基于规则的技术还未发展到十分成熟的地步,wrapper classes 也是由人工建立,信息抽取的能力与 wrapper classes 的完整性有很大的关系.

1.2 基于统计的信息抽取

另一条近几年新发展起来的技术路线是基于统计模型的方向.这方面的研究普遍应用了概率、统计、模糊推理、应用随机过程等领域的模型和技巧,其基本思想是,寻找一个合适的模型,通过改变模型的参数和训练样本集合来适应不同的应用领域.训练后的模型即可用于新的信息源(文本)的信息抽取.这个思想与人工智能中的机器学习、模式识别等领域的思想是一脉相承的.与基于规则的路线相比,基于统计的方法最突出的优点是具有学习能力,适应变化的能力强,从而有可能真正实现信息的自动化抽取.HMM(hidden Markov model)和 SVM(support vector machine)是典型的基于统计模型的方法.

基于 HMM 模型的元数据抽取把文档看作由一些隐藏状态产生的词组序列(例如 title,author 等),从中找到最可能的状态序列[8−11],模型的参数从样例中学习.Seymore 等人用 HMM 实现了对论文头(paper header)的元数据抽取,包括标题、作者、摘要、关键词等,总的准确率达到 90.1%[9].

在 HMM 中,观察值只与当前状态相关,未能有效利用文本上下文信息.而 MEMM(maximum entropy Markov model)通过最大熵框架得到一系列势函数的对应值,从而计算出给定观察序列和前一状态下,得到当前状态的概率.这样,文本中的上下文信息得到了很好的体现[12].文献[12]网络下载的 FAQ 信息进行分段,取得了很好的效果.

Lafferty 和 McCallum 提出的 CRFs(conditional random fields)是一种用于在给定输入结点值时计算指定输出结点值的条件概率的无向图模型,它具有表达元素长距离依赖性和交叠性特征的能力,通常用于处理全局性关联较强的信息抽取工作[13].CRF 模型展现了强于 HMM 很多的提取效果,避免了 HMM 模型中的强相关性假设,而且避免了像 MEMM 等基于有向图的模型中会出现的偏移(元数据标注偏置)问题.Peng 和 McCallum 将 CRF 应用于论文元数据抽取,取得了 90%以上准确度这样的良好结果[14].

Han 等人应用 SVM 来抽取元数据,每种元数据被看作一个类,元数据抽取就是对每个文档块进行分类的工作,总的准确率达到 92.9%[15].

1.3 本文的元数据抽取工作

虽然基于启发式规则与正则匹配算法的信息抽取技术抽取结果比较精确、高效,但是还有很多不足之处.毕竟没有任何规则可以涵盖现实世界中的所有情况,总会有规则之外的元数据格式出现,使得模块的元数据抽取精度降低.规则库也不可能根据元数据抽取的动态结果实时更新,这就使得新出现的元数据抽取规则不能马上利用到后继的元数据抽取过程中去,因此也缺乏一定的灵活性.

HMM 方法的精度已经比较高了,而且比基于规则的方法更灵活.但是,HMM 方法中的单词泛化做得不够好,因为 HMM 方法存在以下一些缺点:(1) 对分类起关键作用的只是有少数一些关键词,需要尽量把其他不起作用的词泛化;(2) HMM 可能把某个短语分割到两个不同的类中,因为 HMM 每次只能发射 1 个单词,但又不能把整个短语作为一个特征整体发射(很多短语并不是固定搭配).最大熵 MEMM 还是存在全局信息不够丰富的缺点,而 CRF 是一种更为复杂的全局 HMM 模型,其精度与其他模型相比较高,但训练时间也较多.

单纯采用 SVM 的效果也不是很好,因为 SVM 分类的方法只能根据文本本身的特征,而孤立了各文本块之间的联系.对元数据抽取来说,各文本块之间的联系(比如各文本块出现的顺序的模式、文本块之间起分隔作用的词或字符)是非常重要的,其重要程度有时甚至超过了文本块本身的内容.Han 等人[15]将上下行的分类信息加入本行的特征向量中,这正是加入块之间联系信息的一种尝试.

本文提出的 SVM+BiHMM 模型把规则、SVM、HMM 方法结合起来,研究论文元数据抽取.其中,HMM 采用的是我们改进的 BiHMM(二元 HMM(bigram HMM)),在保持模型结构不变的前提下,通过区分首发概率和状态内部发射概率,修改了 HMM 发射概率计算模型,有效地克服了传统 HMM 忽略了单词位置信息的缺点,从而提高了抽取精度.在 SVM+BiHMM 模型中,首先根据训练集分别建立独立的 SVM 模型和 BiHMM 模型,采用 Sigmoid 双弯曲函数把 SVM 分类结果拟合为 BiHMM 模型的单词发射概率,再采用 SVM+BiHMM 复合模型进行元数据抽取.该混合模型结合了 SVM 的全局信息优势和 BiHMM 的上下文和单词位置信息的优势.

本文采用 Seymore 定义的 15 个论文头元数据标签(title,author,pubnum,date,abstract,affiliation,address,email,degree,note,phone,intro,keyword,web,page)[9].

本文第 2 节和第 3 节介绍特征的泛化、单独 SVM 模型和 BiHMM 模型的训练.第 4 节利用抽取规则和 SVM+BiHMM 模型进行混合元数据抽取.第 5 节给出实验评测.第 6 节是总结和展望.

2 SVM方法的元数据自动抽取

SVM 是近年来机器学习研究中的一项重大成果.它主要用于解决二值分类的模式识别问题.支持向量机是在统计学习理论(statistical learning theory,简称 SLT)的基础上发展出来的一种新的通用学习方法,其核心内容是 Vapnik 等人在 1992 年~1995 年间提出的[16].支持向量机在众多领域的成功应用表现了它很多优于现有各种方法的性能.

对于线性可分问题,支持向量机的主要思想是,在向量空间中找到一个决策平面(decision surface) $\vec{w}\cdot\vec{x}+b=0$,这个平面能够“最好”地分割两个类别中的数据点.其中,$\vec{x}$ 是待分类的数据点,向量 $\vec{w}$ 和常数 b 从线性可分的训练集中学习得到.假设 $T=\{(\vec{x_i}, y_i)\}$ 为训练集,其中,$y_i\in\{\pm1\}$ 是向量 $\vec{x}$ 的类别(+1 为正样本,–1 为负样本),SVM 就是要找到满足以下限制的 $\vec{w}$ 和 b,使得向量 $\vec{w}$ 的欧式模||w||最小:

$$\vec{w}\cdot\overrightarrow{x_i}+b\geq+1,\ y_i=+1 \quad (1)$$

$$\vec{w}\cdot\overrightarrow{x_i}+b\leq-1,\ y_i=-1 \quad (2)$$

应用 SVM 来抽取元数据,每种元数据被看作一个类,元数据抽取就是对每个文档块进行分类的工作.首先需要确定元数据的类别以及各类别的特征向量.本文抽取 15 个论文主体元数据标签(tag),以向量形式表示为 $\vec{v}$ =(title,author,pubnum,date,abstract,affiliation,address,email,degree,note,phone,intro,keyword,web,page). 设向量的维度为 k,则本文定义的论文主体向量的维度为 15.

本文参考 Han 的方法进行特征的泛化[15].

首先建立 country,city,state,location,person name,week,month,number 特征数据库.通过正则表达式,识别 email、url、数字串等.通过统计各个类别中高词频的词建立领域词典.

然后进行特征的概化.对训练集特征进行概化的过程如下:

(1) 首先用 rule 对数据进行处理,比如,将识别出来的代表 email 的字符串替换为单词 email;

(2) 然后使用上面提到的各种数据库对数据里的字符串进行替换,如将字符串 Ming Zhang 替换为单词 name;将 China 替换为 country;

(3) 最后统计新数据中属于各个类别的高词频的词语,建立领域词典;

(4) 利用领域词典对数据的特征进一步概化,方法同步骤 2.

对测试集只需进行上面的(1),(2),(4)步.另外,在构建特征向量时,忽略了所有文档频率(DF)小于 3 的词.

根据 Han[15]的方法,首先把一篇待处理文本划分为块序列 $f_1,f_2,\ldots,f_n$,再对每一块进行特征选取,建立每行的特征向量,然后对测试集进行分类.根据分类结果,根据该块的前后 d 块来建立该块的新特征(例如 d=3),对原特征向量进行修改,然后用迭代的 SVM 分类器对修改后的特征向量重新进行分类.

设 d 为 SVM 算法迭代时考察的前后特征块个数,k 为向量维度,特征数组 A 是二维数组,根据分类结果修改特征向量的过程如下:对块 L,在原来特征向量的基础上加入新的特征,共$(2d+1)\cdot k$ 个特征(其实是二维数组的线性化),迭代修改特征数组的元素 $A[i][j]$(其中,i 取值区间为$[-d,d]$,表示本块的前或后第$|i|$块;j 取值区间为$[0,k-1]$,表示类别号);若块 $L+i$ 属于类别 j,则 $A[i][j]$=0.5;否则,$A[i][j]$=0.如此进行多次,直到分类结果收敛为止.

经过特征选取和分类器训练,对块序列 $f_1,f_2,\ldots,f_n$ 应用 one-vs-rest 的 SVM 分类,将得到一个得分序列向量 $v_1,v_2,\ldots,v_n$.可以根据文档块分类后所得到的结果,把文档块划归为相应的元数据.

SVM 的结果可以进一步融入到 HMM 模型中.本文第 4 节将讨论结合的 SVM+BiHMM 方法,第 5 节中报告了实验评测结果.

3 基于BiHMM的元数据抽取

隐马尔可夫模型是一阶马尔可夫链的扩展,与一阶马尔可夫链不同,它的观察信号不是模型的状态本身,而是状态的概率函数.真正的模型状态变迁是看不到的,只能看到每个时刻状态所发射的观察信号,通过这些观察信号去推断内在的状态变迁.这也正是模型名称中“隐”字的含义.

HMM 应用于信息抽取实际上是将信息抽取问题转换为文本分类问题(单标注分类).基于 HMM 模型的元数据抽取,把文档看作由一些隐藏状态产生的词组序列(例如 title,author 等),从中找到最可能的状态序列.HMM 模型的参数从样例中学习.

3.1 确定BiHMM模型结构

传统 HMM 不能充分利用某些信息,比如孤立了单词间的直接联系,导致其识别词组的能力很弱,比如状态产生字符串“a b”和“b a”的概率是一样的,这不符合实际情况.HMM 也忽略了单词在状态中的位置信息,没有考虑到位于状态之首的单词和状态内部的单词的重要程度应该是不同的.

其实,影响抽取效果的主要是一些关键单词,特别是状态之首的单词或标点以及一些有特殊意义的单词.比如,在标题部分出现的很多单词是无关紧要的,只需要知道标题在哪里开始和结束,而不需要知道标题的内容;

而像另外一些单词却是很重要的,而且这些单词出现的位置信息也是不可忽略的,比如,“pp.”通常出现在引文页码部分,而且标记着页码的开始.因此在实际处理中,“pp.”出现在页码状态之首的概率应该很高,而位于页码状态内部的概率应该很低.但传统的 HMM 却忽略了单词的位置信息,影响了抽取精度.

为了弥补这些不足,一类方法是修改模型拓扑结构而实现层次模型.Nymble 采用二层 HMM 模型抽取“named entity”(名词实体),待抽取的域作为“name-class”(名词类属)状态,其内层为全连接的模型,不过,其模型由人工设定[8].DATAMOLD 采用二层 HMM 模型把文本分为结构化的记录,其外部模型和内部模型都从训练数据中学习得到,内部 HMM 可以获得更为细致的结构[10].

本文介绍了二元 HMM 模型 BiHMM.在保持模型结构不变的前提下,通过改变发射概率计算模型,有效地克服了传统 HMM 的缺点.

首先作如下约定,$q\uparrow\sigma$表示状态 q 发射单词σ, $q\nwarrow\sigma$ 表示σ为状态 q 的首单词, $q\nearrow\sigma$ 表示状态 q 发射单词σ且σ非 q 的首单词.P 表示发射概率,显然有 $P(q\uparrow\sigma)=P(q\nwarrow\sigma)+P(q\nearrow\sigma)$.

HMM 的参数主要包括转移概率和发射概率,单词间的顺序关系信息以及单词在状态中的位置信息只能通过参数体现出来,由于与单词有关,因此只能通过单词发射概率体现出来.为此,需要改变单词发射概率的计算模型,使发射概率能够反映这两方面的信息.在传统的 HMM 中,发射概率仅仅是单词频率的函数,如果将发射概率变为单词频率、单词顺序关系和单词位置信息的函数,就能满足要求.

单词间的顺序关系可以通过 N 元文法(N-gram)来捕捉,为了便于表示与计算,本文采用二元文法(bigram)模型,即单词的发射概率不仅与该单词的频率有关,还与上一个单词有关,这也是本文称改进后的模型为 BiHMM(二元 HMM)的原因.另外,为了使发射概率够反映单词的位置信息,本文根据单词所处位置的不同(位于状态之首还是状态内部),采用不同的发射概率计算方法.具体说来,就是根据单词在状态中的位置的不同以及前一个单词的不同,将发射概率 $P(q\uparrow\sigma)$分成两部分,即状态之首发射概率 $P(q\nwarrow\sigma)$ 和状态内部概率 $P(q\nearrow\sigma)$. $P(q\uparrow\sigma)$可以表示为 $P(\sigma|q)$,σ_{-1} 表示σ的前一个单词.状态内部发射概率的 BiHMM 模型为

$$P(\sigma\mid q)=\begin{cases}P(q\nwarrow\sigma), & \sigma\text{出现在状态}q\text{的开头}\\ P(q\nearrow\sigma)=P(\sigma\mid\sigma_{-1},q), & \sigma\text{出现在状态}q\text{的内部},\sigma_{-1}\text{是前一个单词}\end{cases} \tag{3}$$

可以利用训练集来计算 $P(q\nwarrow\sigma)$ 和 $P(\sigma|\sigma_{-1},q)$.设 $c(x)$表示事件 x 在训练集中出现的次数,为了计算 $c(x)$,需要在模型合并过程中保存状态的一些必要信息,包括状态中各个单词出现的频率以及顺序关系、状态到其他状态的转换关系.

$$P(q\nwarrow\sigma)=\frac{c(q\nwarrow\sigma)}{\sum_{\rho\in\Sigma}c(q\nwarrow\rho)};\ P(\sigma\mid\sigma_{-1},q)=\frac{c(q\uparrow\sigma_{-1}\sigma)}{c(q\uparrow\sigma_{-1})} \tag{4}$$

本文从训练集自动学习模型结构.根据标注训练集构建一个初始模型,这个模型能够产生训练集的所有观察符号序列,但不能产生任何不属于训练集的观察符号序列,可称为训练集的最确定模型(most specific model).又由于这个模型能够以最大概率产生整个训练集,因此也称为训练集的最大似然模型(maximum likelihood model).

构造方法如下:(1) 首先建立状态,状态的元数据标签与单词的标签相同.假设训练集大小为 n,则可以得到 n 个状态序列,对于论文主体,建立 15 种状态(title,author,pubnum,date,abstract,affiliation,address,email,degree,note,phone,intro,keyword,web,page).再设定一个起始状态和一个结束状态,起始状态以概率 $1/n$ 转换到每个状态序列的第 1 个状态,每个状态序列的最后一个状态以概率 1 转换到结束状态;(2) 将标注训练集的每个单词序列翻译为一个唯一与之对应的状态序列,每个状态以概率 1 发射其对应的单词,并以概率 1 转换到与其相邻的下一个状态.

为了让模型有实际的意义,需要对模型进行状态合并,使其能够产生更多的单词序列甚至无穷多的序列.这个合并过程称为模型的泛化.一方面,随着状态的不断合并,模型产生单词序列的能力越来越强,但是,它产生训练集的概率在逐渐降低,也就是说,模型与训练集的拟合程度在不断降低;另一方面,随着泛化的进行,模型 M 越来越有实际意义,因此,它出现的概率会不断增加,但也不能无限制地泛化:一方面，过度泛化会导致模型与训练

集的拟合程度太低,失去了训练的意义;另一方面,模型的出现概率并不是一直增加的,考虑一个极端情况,当模型被泛化到只剩下 1 个状态时,模型也就失去了实际的意义,其出现的概率也会很低.因此,需要找到泛化的临界点,使模型与训练集的拟合程度和模型出现的概率达到一个平衡点.

能够使用的合并技术很多,主要有 3 种:贝叶斯合并(Bayesian merging)[17]、N 合并(neighbor-merging)和 V 合并(vertical-merging)[18].

贝叶斯合并通过下面的公式:

$$P(M|D)\sim P(D|M) \tag{5}$$

迭代合并状态,直到在数据和模型规模之间达到一种优化的平衡.在公式(5)中,M=Model,D=Data.

贝叶斯合并实现比较麻烦,在信息抽取领域用得较多的是 N 合并和 V 合并,特别是它们二者结合的应用.一方面,它们的实现都相当简单;另一方面,它们所采用的方法比较符合人的直觉,容易让人接受;再一方面,由于信息抽取领域的特殊性,采用这两种方法也能取得较好的效果.

N 合并将相邻且标签相同的两个状态进行合并.在 N 合并中,多个相邻且拥有同样标签的状态被合并成了一个状态,需要引入状态的自我转换.

V 合并是指合并来自同一个状态或去向同一个状态的两个状态.V 合并能够极大地减少模型中的分支,比如,它能够将来自 start 状态的所有标签为 author 的状态合为 1 个状态,从而减少了很多分支.至于合并的终点,一般是直到不能再合并为止.

通过标注集{“⟨title⟩ a ⟨/title⟩⟨author⟩ b ⟨/author⟩”,“⟨title⟩ c d ⟨/title⟩⟨email⟩ e ⟨/email⟩”}自动学习 BiHMM 模型如图 1 所示.

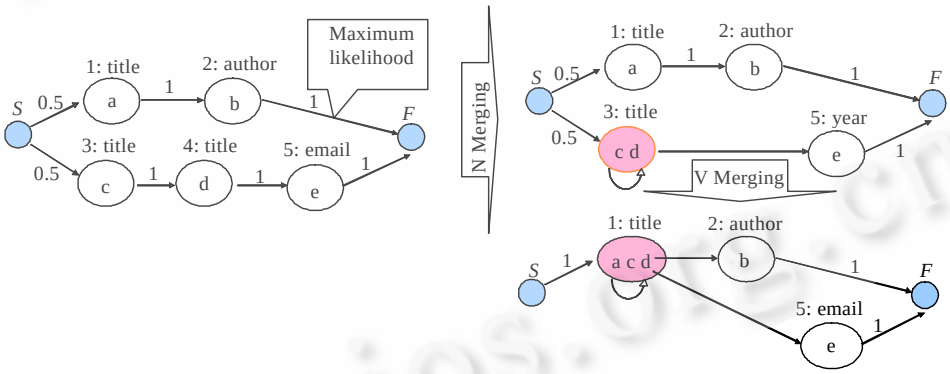


Fig.1 Learning structure and merging of HMM models

图 1 HMM 模型学习结构及合并

3.2 参数估计和寻找最优状态序列

在模型的建立过程中,最有价值的训练集是标注训练集(labeled training data),即为训练集中所有单词序列的每个单词注上标签.标注训练集能够很好地确定模型的结构,并估计模型的参数.通过统计标注训练集中状态 q 发射单词 σ 的出现频率 $c(q\uparrow\sigma)$ 和状态间的转换频率 $c(q\uparrow q')$,得到模型参数的最大似然估计(maximum likelihood estimation).

$$P(q\to q')=\frac{c(q\to q')}{\sum_{s\in Q}c(q\to s)} \tag{6}$$

$$P(q\uparrow\sigma)=\frac{c(q\uparrow\sigma)}{\sum_{\rho\in\Sigma}c(q\uparrow\rho)} \tag{7}$$

由于单词发射概率的计算采用了二元模型,因此,平滑方法也作了相应的改变.本文参考 Nymble 系统的实现[8],借用 N-gram 模型常用的一种模型退化思想来进行参数平滑.它其实也是某种形式的 shrinkage 平滑,称为

back off-shrinkage.

关于找出给定观察序列 $O=O_1O_2\ldots O_n$ 的最优状态序列 $Q=q_1q_2\ldots q_n$ 的问题,本文考虑最大概率产生观察序列意义上的最优.有关最优问题的详细讨论参见文献[19].

单纯采用 BiHMM 方法也可以得到比较好的抽取结果[11].而本文采用 SVM+BiHMM 动态修正的方法,得到了更好的结果,详见第 5 节的评测.

4 SVM+BiHMM自动抽取的流程

本文对论文主体信息进行抽取.如图 2 所示,首先根据第 2 节和第 3 节介绍的算法,训练出独立的 SVM 和 BiHMM 模型.利用这两个模型,结合了以下 3 个主要步骤对每一篇文档进行元数据自动抽取:

1. 根据规则把论文粗分为论文头、正文以及引文部分,接下来对论文头部分进行细分.将论文头的行记为 $f_1, f_2,\ldots, f_n$.
2. 采用 SVM 方法对论文头的行进行分类[15].
   (1) 采用大量人名、地名词典,将单词泛化;
   (2) 从训练集中根据词频,为每个类别建立相关的 word-list;
   (3) 建立行特征向量,采用 SVM 分类器给每一行分类;
   (4) 采用迭代的方法,用上一次每一行的分类结果来修改每一行的特征向量.根据该行前后 d 行(本实验采用 5 行)的类别,同时考虑前后几个类别作为特征(因为相邻的几行可能属于同一个类).该迭代过程不断进行,直到分类结果收敛为止.收敛的判断参见文献[15].
   (5) 对属于多个类的行,再进行细分.
3. 最后采用 BiHMM 对 SVM 分类结果进行修正.

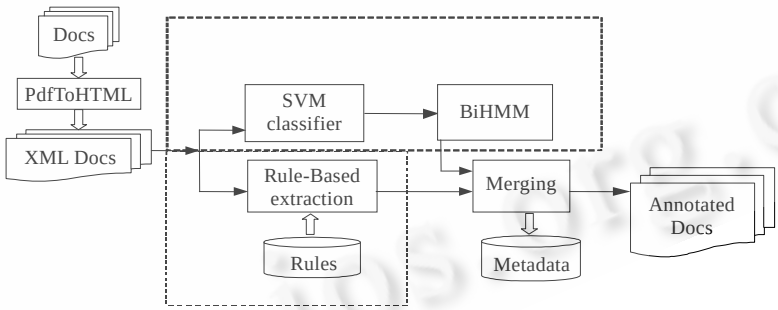


Fig.2 The model of SVM+BiHMM metadata extraction

图 2 SVM+BiHMM 元数据抽取模型

One-vs-Rest 的 SVM 分类器对于行序列 $f_1,f_2,\ldots,f_n$ 得到一个得分序列向量 $v_1,v_2,\ldots,v_n$.然后利用公式(8)所示的 Zadrozny 和 Gannapathiraju 等人在语音识别领域采用的 Sigmoid 双弯曲线函数(S 形曲线函数),将 SVM 分类的距离值转换为对应的概率[20].Sigmoid 双弯曲线函数如公式(8)所示:

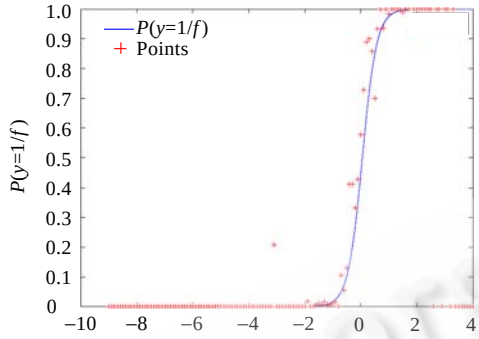
$$\hat{P}(q\mid\sigma)=\frac{1}{1+e^{A\cdot v(\sigma)+B}} \tag{8}$$

其中,$v(\sigma)$为得分值, $\hat{p}(q\mid\sigma)$ 为估计概率值.参数 A 和 B 需要根据训练集进行动态调整,具体调整方法见文献[20].

本文采用公式(8)计算调整公式(7)中的发射概率 $P(q\uparrow\sigma)$,得到调整后的概率,如图 3 所示.计算方法如公式(9)所示,其中,c,d 是调整参数,根据训练集进行动态调整.具体方法为,对训练集采用 4-fold 交叉验证(cross validation),将其分为 4 份,依次将其中的 1 份作为验证集,其余 3 份作为训练集.通过对 4 次评测结果的平均,得到最终结果.不断调整参数 c 和 d,直到该最终结果最优为止.

$$P'(q\uparrow\sigma)=c\cdot\hat{P}(q\mid\sigma)+d\cdot P(q\uparrow\sigma) \tag{9}$$

保持 BiHMM 的状态转移概率不变.应用调整后的 SVM+BiHMM 模型,再进行一次调整.



Parameters for sigmoid function to map SVM score (*Tag*=2): *A*=−3.88120729675265, *B*=0.215057297010456

Fig.3 Mapping SVM scores into the emitting probability of BiHMM model using a sigmoid function

图 3 用双曲线函数把 SVM 得分映射为 BiHMM 模型的发射概率

与语音识别领域结合 SVM 和 HMM 的工作不同[20−22],本文的元数据抽取工作有如下特色:

(1) 处理的数据不同,因此,本文根据规则进行文本块的划分以及 SVM 特征提取工作不能借鉴语言领域的工作.Han 等人应用 SVM 来抽取元数据[15],而本文是对规则、SVM、BiHMM 三者的结合.

(2) HMM 建模方面,本文采用了 BiHMM 方法,考虑了单词之间的顺序.

(3) 在 SVM 的得分拟合为后验概率方面,本文的工作具有一定的创新.语音识别工作首先生成一个完全图的 HMM 模型,其转换概率都设置为 1,其发射概率根据 SVM 分类的距离值直接得到.而本文则首先根据训练集得到基本的 BiHMM 模型,这个模型的转换概率和发射概率都是通过训练集得到的. BiHMM 模型的发射概率根据 SVM 的得分(而不是距离值)进行修正.

5 元数据自动抽取实验评测

由前面的介绍可知,SVM,HMM 应用于信息抽取实际上是将信息抽取问题转换为文本分类问题(单标注分类),因此,实验的性能评估也采用文本分类评估的标准,包括查全率(recall)、查准率(precision)、*F*1 值、准确率(accuracy)等作为评估标准[11].

本文将每个单词看成一篇文档,将单词的标签看作单词所属的类别,将 SVM 和 HMM 看成一个分类器,于是,便将抽取性能评估问题转变成文本分类的性能评估问题.

形象地说,查准率是指正确分到某类别集下的文档数占实际分到该类别集下的文档数的比率;查全率是指正确分到某类别集下的文档数占应该分到该类别集下的文档数的比率.所谓“实际分到”是指采用分类器进行分类的结果,“应该分到”是指标准的分类结果.准确率是指被正确地判断为“正”(“正”分到正)以及被正确地判断为“负”(“负”分到负)的文档数占所有文档数的比率.

表 1 所示符号的含义如下:TP_c 表示 *M* 正确分出属于标签类别 *c* 的文档数目;FP_c 表示本不属于标签 *c*,但 *M* 却分出属于 *c* 的文档数目;FN_c 表示本应该属于标签 *c*,但 *M* 却分出不属于 *c* 的文档数目;TN_c 表示本来不属于标签 *c*,而 *M* 也正确地分为不属于 *c* 的文档数目.

Table 1 The contingent judgment of tag *c*

表 1 对标签类别 *c* 的判别划分

Tag *c*		Expert judgments	
		YES	NO
Classifier judgments	YES	TP_c	FP_c
	NO	FN_c	TN_c

给定测试文档集 *D*、标签 *c* 和分类器 *M*,在 *c* 下,*M* 的分类查准(P_c)和分类查全(R_c)、*F*1 测度、分类准确率

(A_c)的计算公式如下:

$$P_c=\frac{TP_c}{TP_c+FP_c},\ R_c=\frac{TP_c}{TP_c+FN_c},\ F1_c=\frac{2\cdot P_c\cdot R_c}{P_c+R_c},\ A_c=\frac{TP_c+TN_c}{TP_c+FN_c+FP_c+TN_c} \tag{10}$$

在各个标签类别的测度基础上,可以计算宏观平均(macro averaging):先按照每个标签来计算分类效果指标(查全、查准等),然后用所有标签的指标平均值来代表整体丰的分类效果.

本文采用 Seymore 的数据集[9]并采用与 Seymore 和 Han 相同的训练集划分方法[9,15],将 935 篇论文的 header 部分的前 500 篇作为训练集,后 435 篇作为测试集.提取 15 个标签,包括 title,author,pubnum,date,abstract,affiliation,address,email,degree,note,phone,intro,keyword,web 和 page.抽取结果见表 2.

Table 2 The metadata extraction result of SVM+BiHMM

表 2 SVM+BiHMM 的元数据抽取结果

Class	Precision	Recall	*F*1 measure	Accuracy
title	0.930	0.970	0.949	0.991
author	0.887	0.930	0.908	0.986
affiliation	0.937	0.955	0.946	0.989
address	0.947	0.955	0.951	0.993
note	0.949	0.814	0.876	0.987
email	0.955	0.994	0.974	0.997
date	0.823	0.990	0.899	0.997
abstract	0.982	0.997	0.989	0.989
intro	0.996	0.954	0.974	0.998
phone	0.913	0.976	0.943	0.999
keyword	0.909	0.810	0.857	0.995
web	0.925	1.000	0.961	0.999
degree	0.979	0.753	0.852	0.996
pubnum	0.800	0.962	0.873	0.998
page	0.991	1.000	0.995	0.999
macro	0.928	0.937	0.930	0.994
micro	0.957	0.966	0.961	0.994

图 4 是各种抽取算法的效果对比,其中,MySVM 是本文重复 Han 的独立行分类方法所得到的结果,尽管这个结果并不理想,但是经过 BiHMM 修正后,SVM+BiHMM 的结果达到了最佳性能.图 5 是与 Han 的结果相比的各项具体 tag 的结果.从这些数据可以看出,本文提出的 SVM+BiHMM 方法的抽取效果是最好的.

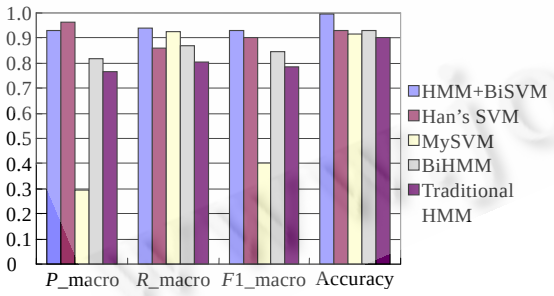


Fig.4 Comparison of different extraction models

图 4 各种抽取模型的比较

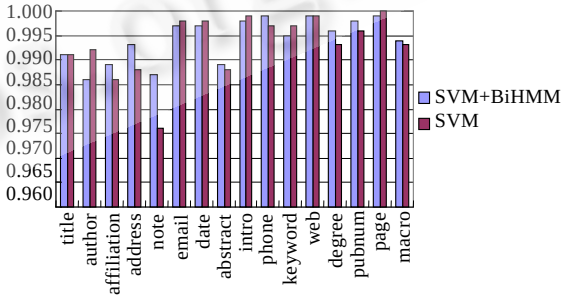


Fig.5 *F*1 measure comparison of SVM+BiHMM and SVM

图 5 SVM+BiHMM 与 SVM 的 *F*1 测度比较

6 结束语

本文提出了结合 SVM 和 BiHMM 自动抽取科技文献元数据的方法.BiHMM 模型充分考虑了单词的首发信息,利用了单词间的联系.BiHMM 模型的转换概率和发射概率都是通过训练集得到的.在 SVM+BiHMM 复合模型中,首先根据规则把论文粗分为论文头、正文以及引文部分;然后采用大量的行排版特征属性,根据人名、地名词典将单词泛化,形成特征向量,在此基础上建立 SVM 模型,把文本块划分为元数据子类;然后采用 Sigmoid

双弯曲函数,把 SVM 分类结果用于拟合调整 BiHMM 模型的单词发射概率;最后用复合模型进行元数据抽取.该模型弥补了元数据抽取中 SVM 不能很好地利用上下文的联系以及传统 HMM 只能利用词频计算概率的不足.实验表明,SVM+BiHMM 的元数据抽取精确度优于单纯的 HMM 和 SVM 方法.

我们将把本文的研究扩展到各种信息抽取模型的结合上,并应用于引文详细元数据和网络资源元数据的抽取.

References:

[1] Morville P, Rosenfeld L. Information Architecture for the World Wide Web: Designing Large-Scale Web Site. 3rd ed., Sebastopol: 0'Reilly&Associates, 2006.

[2] Chidlovskii B Wrapping web information providers by transducer induction. In: Raedt L, Flach P, eds. Proc of the 12th Int'l of European Conf. on Machine Learning (ECML 2001). LNCS 2167, Heidelberg: Springer-Verlag, 2001. 61−72.

[3] Hitchcock S, Carr L, Jiao Z, Bergmark D, Hall W, Lagoze C, Harnad S. Developing services for open eprint archives: Globalisation, integration and the impact of links. In: Proc. of the 5th ACM Conf. on Digital Libraries (ACMDL 2000). New York: ACM Press, 2000. 143−151.

[4] Klink S, Dengel A, Kieninger T. Rule-Based document structure understanding with a fuzzy combination of layout and textual features. Int'l Journal on Document Analysis and Recognition, 2001,4(1):18−26.

[5] Kim J, Le DX, Thoma GR. Automated labeling algorithms for biomedical document images. In: Proc. of the 7th World Multiconference on Systemics, Cybernetics and Informatics. Orlando: IIIS, 2003. 352−357.

[6] Zhang M, Yang DQ, Deng ZH, Feng Y, Wang WQ, Zhao PX, Wu S, Wang SA, Tang SW. PKUSpace: A collaborative platform for scientific researching. In: Liu WY, Shi YC, Li Q, eds. Proc of the Int'l Conf. of Web-based Learning (ICWL 2004). LNCS 3143, Heidelberg: Springer-Verlag, 2004. 120−127.

[7] Zhao PX, Zhang M, Yang DQ, Tang SW. Automatic extraction of metadata from digital documents. Computer Science, 2003, 30(10):217−204 (in Chinese with English abstract).

[8] Bikel DM, Miller S, Schwartz R, Weischedel R. Nymble: A high performance learning name finder. In: Proc. of the 5th Conf. on Applied Natural Language Processing (ANLC'97). San Francisco: Morgan Kaufmann Publishers, 1997. 194−201.

[9] Seymore K, McCallum A, Rosenreid R. Learning hidden Markov model structure for information extraction. In: Califf ME, Freitag D, Kushmerick N, Muslea I, eds. Proc. of the AAAI'99 Workshop on Machine Learning for Information Extraction. Cambridge: MIT Press, 1999. 37−42.

[10] Borkar VR, Deshmukh K, Sarawagi S. Automatic segmentation of text into structured records. In: Aref WG, ed. Proc. of the ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD 2001). New York: ACM Press, 2001. 175−186.

[11] Yin P, Zhang M, Deng ZH, Yang DQ. Metadata extraction from bibliographies Using bigram HMM. In: Chen Z, Chen H, Miao Q, Fu Y, Fox E, Lim E, eds. Proc. of the Int'l Conf. of Asian Digital Libraries (ICADL 2004). LNCS 3334, Heidelberg: Springer-Verlag, 2004. 310−319.

[12] McCallum A, Freitag D, Pereira F. Maximum entropy Markov models for information extraction and segmentation. In: Langley P, ed. Proc. of the Int'l Conf. on Machine Learning (ICML 2000). San Francisco: Morgan Kaufmann Publishers, 2000. 591−598.

[13] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Brodley C, Danyluk A, eds. Proc. of the Int'l Conf. on Machine Learning (ICML 2001). San Francisco: Morgan Kaufmann Publishers, 2001. 282−289.

[14] Peng F, McCallum A. Accurate information extraction from research papers using conditional random fields. In: Dumais S, Marcu D, Roukos S, eds. Proc. of the Human Language Technology Conf. and North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004). New York: ACM Press, 2004. 329−336.

[15] Han H, Giles CL, Mnavoglu E, Zha HY, Zhang ZY, Fox EA. Automatic document metadata extraction using support vector machine. In: Proc. of the ACM/IEEE Joint Conf. on Digital Libraries (JCDL 2003). New York: ACM Press, 2003. 37−48.

[16] Stitson MO, Weston JAE, Gammerman A, Vovk V, Vapnik V. Theory of support vector machines. Technical Report, CSD-TR-96-17, London: University of London, 1996.

[17] Stolcke A, Omohundro SM. Best-First model merging for hidden Markov model induction. Technical Report, TR-94-003, Computer Science Division, University of California at Berkeley, Int'l Computer Science Institute, 1994.

[18] McCallum AK, Nigam K, Rennie J, Seymore K. Automating the construction of internet portals with machine learning. Information Retrieval Journal, 2000,3(2):127−163.

[19] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 1989,77(2): 257−285.

[20] Ganapathiraju A, Hamaker JE, Picone J. Applications of support vector machines to speech recognition. IEEE Trans. on Signal Processing, 2004,52(8):2348−2355.

[21] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Hand D, Keim D, Ng R, eds. Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2002). New York: ACM Press, 2002. 694−699.

[22] Venkataramani V, Byrne V. Lattice segmentation and support vector machines for large vocabulary continuous speech recognition. In: Petropulu AP, Xia XG, eds. Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005). Washington: IEEE Computer Society, 2005. 817−820.

附中文参考文献:

[7] 赵培翔,张铭,杨冬青,唐世渭.数字化文档元数据的自动提取.计算机科学,2003,30(10):217−204.

张铭(1966－),女,山西交城人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,数据挖掘,数字图书馆,语义网,网格计算.

银平(1982－),男,硕士,主要研究领域为数据库,信息系统数据挖掘.

邓志鸿(1973－),男,博士,副教授,主要研究领域为数据库,数据挖掘.

杨冬青(1945－),女,教授,博士生导师,CCF 高级会员,主要研究领域为数据库,信息系统.