引用格式:刘晨玥,李兵,吴卫星. 基于罪名相关成分标注的刑事裁判文书概要信息提取[J]. 山东科技大学学报(自然科学版), 2018, 37(4):92-101.

LIU Chenyue, LI Bing, WU Weixing. Information extraction of judical documents based on crime-related tags[J]. Journal of Shandong University of Science and Technology (Natural Science), 2018, 37(4):92-101.

基于罪名相关成分标注的刑事裁判文书概要信息提取

刘晨玥[1],李 兵[2],吴卫星[1]

(1. 对外经济贸易大学 金融学院,北京 100029;2. 对外经济贸易大学 信息学院,北京 100029)

摘 要:近年来,互联网信贷市场迅猛发展,多角度多信息源充分了解贷款申请人的信用情况显得愈发重要。法院的案件裁判文书的权威性、规范性以及其官方可得性,使其成为贷款申请人信用评估的重要数据源。命名实体识别技术在司法领域的应用亟待探索。针对网上公开的刑事裁判文书进行概要信息提取,构建基于罪名相关成分标注语料库的隐马尔科夫模型和最大熵马尔科夫模型,并利用其识别提取裁判文书中的被告人及其罪名等关键司法信息,可以为互联网信贷平台的信用风险管理工作提供更充分的信息资源。开放性测试结果显示基于罪名相关成分标注的 HMM 和 MEMM 的平均 F 值分别达到了 87.79%、90.25%,说明提出的方法克服了裁判文书格式的差异和罪名实体识别的困难,具有较好的刑事裁判文书概要信息提取效果。

关键词:罪名实体识别;刑事裁判文书;隐马尔科夫模型;最大熵马尔科夫模型;罪名相关成分标注

中图分类号:TP391 **文献标志码**:A **文章编号**:1672-3767(2018)04-0092-10

DOI:10.16452/j.cnki.sdkjzk.2018.04.012

Information Extraction of Judical Documents Based on Crime-related Tags

LIU Chenyue[1], LI Bing[2], WU Weixing[1]

(1. School of Banking and Finance, University of International Business and Economics, Beijing 100029, China;
2. School of Information, University of International Business and Economics, Beijing 100029, China)

Abstract: The past few years have seen the rapid development of Internet finance market, which attaches a growing importance to the comprehensive understanding of the credit of money borrowers. At the same time, the authority, standardization and availability of judical documents have become an essential data source for the credit rating of money borrowers. In this paper, the application of named entity recognition (NER) technology on judical documents was explored. Information extraction was first carried out from judicial documents on the internet. A hidden Markov model (HMM) and max entropy Markov model (MEMM) based on crime-related tags were then established. With the help of the HMM and MEMM, key judicial information such as defendants and charges was finally extracted from these judicial documents to provide sufficient information source for the credit risk management of internet finance platforms. Experiment results show that the average F values of HMM and MEMM reach 87.79% and

收稿日期: 2017-12-14

基金项目:北京市自然科学基金项目(9142014);国家社会科学基金项目(16BTQ065);对外经济贸易大学国内外联合培养研究生项目

作者简介:刘晨玥(1994—),女,山东新泰人,硕士研究生,主要从事机器学习与金融工程研究.
E-mail:chenyue.liu@columbia.edu
李 兵(1970—),男,辽宁沈阳人,教授,主要从事社会网络分析与数据挖掘研究. E-mail:lb0501@126.com
吴卫星(1974—),男,湖北荆门人,教授,博士生导师,主要从事金融工程与金融风险管理研究.

90.28% respectively. The proposed method can overcome the difficulty of recognizing crime entity and the difference of judicial document format and has a better performance in extracting judicial information from documents of criminal judgment.

Key words: crime entity recognition; judical documents; HMM; MEMM; crime-related tags

2014 年以来,我国的互联网信贷市场发展迅猛,互联网金融迎来了黄金发展期。但由于 P2P 网贷等互联网金融平台低门槛、高收益的特点,在繁荣发展的同时坏账现象频现,亟需建立严谨有效的信用风险控制模式。

管理信用风险需要充分了解贷款申请人的信用情况,以避免为了取得借款而虚构个人信息、隐瞒违法犯罪。然而,相比于财务报表等结构化数据,贷款申请人犯罪记录等司法领域的非结构化数据,由于非直观性、非精简性的特点,并没有在信用风险管理中得到足够的重视。

司法公开化的深入推行提供了权威的犯罪记录信息,使基于这一非结构化数据的有效信用风险管理成为可能:自 2014 年 1 月 1 日起,《最高人民法院关于人民法院在互联网公布判决文书的规定》正式实施,明确各级人民法院的生效判决文书在中国判决文书网上统一公布。如果能够提取这些由法院等权威机构公开的刑事裁判文书中的关键信息,则可以为 P2P 网贷平台的贷款申请人审核提供更加全面充分的参考数据,有助于互联网金融企业更好地管理其信用风险。

法院公开的刑事案件裁判文书的关键信息主要包括被告人姓名、案件发生地点、时间以及被告人所获罪名。对裁判文书中被告人及其所获罪名实体的识别与提取,体现了司法领域文本挖掘的特殊性,是本文的重点工作。对命名实体的识别方法主要分为两类:

一是人工总结待提取字段的识别规则和专用词典:王宁等(2002)[1]基于语料库对金融领域公司实体名称的识别工作、梅奥诊所(2011)[2]基于 UMLS 和 SNOMED CT 两大受控术语词典的 cTAKES 系统对病历文本的挖掘、Zhang 等(2013)[3]对建筑类文本的信息提取、Derczynski(2016)[4]基于大型消歧语料库对推特文本的命名实体识别,均表现出在开放性测试中准确率较低的缺陷。虽然灵活性欠佳,但针对格式固定的文本时不失高效简便。因此,对于裁判文书中的格式较为固定的案件发生时间与地点,本研究将直接使用基于规则的方法来提取。

二是基于机器学习,尤其是监督学习的方法,由于其灵活性与可塑性逐渐成为近年来的命名实体识别主流算法。机器学习的方法主要分为监督学习、半监督学习和无监督学习。半监督与无监督学习的主要手段为自展和聚类。Elsner 等[5]使用基于聚类的无监督学习来识别命名实体;Munro 等[6]针对非对齐双语语料,通过计算两门语言之间的局部编辑距离偏差实现自展法,实现了无监督学习下的命名实体识别。文献[7]提出了基于自展法的半监督学习,得到平均 75%的 F 值。虽然对训练预料的数量要求较少,但半监督学习的效果对训练语料的代表性有一定要求;而无监督学习参数的优化,仍需要监督学习的辅助;虽然不需要大规模人工标记的训练语料,无监督学习和半监督学习均有耗时较长的缺陷。本研究借助最高人民法院发布的权威刑法罪名,提出罪名相关成分标注算法,提供训练语料,使速度较快、准确率较高的监督学习成为可能。

常用的监督学习模型包括 Bikel 等使用的隐马尔可夫模型(HMM)[8]、Sekine 等[9]使用的决策树、Borthwick 等[10]使用的最大熵模型、Asahara 等[11]使用的支持向量机模型、Lafferty 等[12]提出的 CRF 模型等。Ekbal 等[13]、Ritter 等[14]基于支持向量机模型从推特中提取特定事件的相关信息,然而支持向量机模型局限于二元判断,不适用更为复杂的命名实体识别工作。张华平等[15]基于隐马尔科夫模型在中国人名识别开放测试中达到了 91%的召回率;Ye 等[16-18]建立条件随机场模型从非结构化的道桥报告文本中提取道桥相关信息,其准确率达 97%。在多数研究中 CRF 模型的准确率要高于 HMM 和 MEMM,但其解码算法时间复杂度与特征空间的规模呈正比、收敛速度缓慢;相比之下,HMM 和 MEMM 在保证一定的准确率的同时可操作性较强,因此,将采取 HMM 和 MEMM 来学习罪名实体的特征,对裁判文书中被告人及所获罪名实体进行识别与提取。

本研究在对分析格式较为固定的案件发生时间与地点基于规则的提取进行分析的基础上,针对被告人与所获罪名难以关联、罪名难以作为命名实体被完整识别和公文书写存在差异等困难,将基于罪名相关成分标注语料库建立隐马尔科夫模型和最大熵马尔科夫模型,使用该模型解码并识别被告人及所获罪名实体,最终进行提取实验并分析结果。

1 刑事裁判文书案件发生时间与地点的识别提取

刑事案件裁判文书中的案件发生时间与地点具有较为固定的格式,故人民法院的名称或者人民检察院的名称中包含案件发生地点的信息,而刑事案件裁判文书的编号,如"(2015)南刑初字第 23 号"和如下裁判文书样例所示的"(2011)嘉平刑初字第 11 号"等,则包含了案件发生年份这一时间信息。已隐去被告人姓名的裁判文书样例如下:

《陆 XX 容留他人吸毒罪一审刑事判决书》

(2011)嘉平刑初字第 11 号

公诉机关平湖市人民检察院。

被告人陆 XX,绰号:全糖小蛮子,农民。因吸食毒品,于 2003 年 5 月 7 日被本市公安局罚款 2 000 元,并被强制戒毒四个月。因本案,于 2010 年 10 月 16 日被嘉兴市公安局港区分局刑事拘留,同年 11 月 19 日被依法逮捕。现羁押于本市看守所。

……

被告人陆 XX 犯容留他人吸毒罪,判处有期徒刑七个月,并处罚金 3 000 元(刑期从判决执行之日计算,判决执行以前先行羁押的,羁押一日折刑期一日。即自 2010 年 10 月 16 日起至 2011 年 5 月 15 日止;罚金款限本判决生效后十日内缴纳)。

本文选择用正则表达式从公诉机关或者法院的所在市中提取案件发生地点信息,从刑事案件裁判文书的编号中提取案件时间信息。正则表达式用事先定义好的一些特定字符及其组合,从字符串中提取特定部分。在本文工作中,用到了如"(.*)((.*))(.*)公诉机关(.*)"以对文书编号中的年份进行提取,再如"(.*)公诉机关(.*?)人民检察院(.*)",以对公诉机关所在地进行提取。

正则表达式等基于规则的方法,应用于文本结构相对固定的案件发生时间与地点时,能够实现高效准确的识别提取。但对于文本结构较为复杂的被告人及其所获罪名,该方法不再适用,需要采用基于机器学习的方法来识别提取被告人及其所获罪名信息。

2 刑事裁判文书被告人与被告人所获罪名的识别提取

裁判文书中被告人及其所获罪名实体的识别与提取工作,主要面临如下难点:①虽然现有的分词系统如中国科学院中文词法分析系统 ICTCLAS 能识别出裁判文书中的被告人姓名,但是仅识别出人名并不能满足司法文本挖掘的需求:一份裁判文书中出现的人名可能是被告人姓名,也可能是辩护人姓名;一份判决文书中亦可同时为多个被告人定罪。因此,被告人姓名的识别提取,需要在识别过程中实现被告人姓名与其所获罪名的关联。②对于被告人三字以上的罪名,现有的分词软件并不能将其作为一个整体识别出来。根据中华人民共和国最高人民法院《关于执行〈中华人民共和国刑法〉确定罪名的规定》及其补充规定,可以得到最高人民法院发布的最新刑法罪名,但基于词典的方法并不可行:如当裁判文书内容为"非法提供答案罪"时,对应在《刑法修正案(九)》第二十五条第三款中的刑法罪名全名则为"非法出售、提供试题、答案罪",遍历所有刑法罪名并对罪名内部各成分排列组合后进行对比,用时过长。③直接使用基于规则和词典来进行提取的方法有很大缺陷:尽管公文写作遵循一定规范,但历年来各地各级法院的裁判文书书写仍在格式上存在诸多差异,一一总结工作量巨大。

针对以上难点,选择借助最高人民法院发布的权威刑法罪名,对罪名相关成分进行标注,通过基于此训练得到的 HMM 和 MEMM 实现对被告人及罪名实体的识别提取,工作流程如图 1 所示。

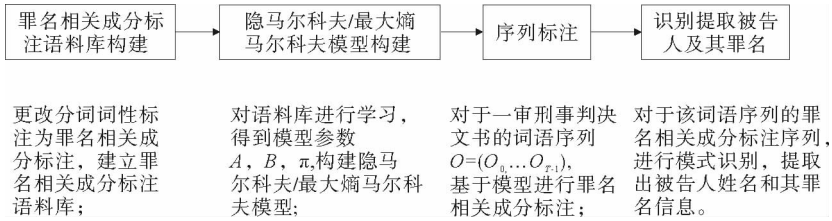

图 1　使用基于罪名相关成分标注语料库的 HMM 和 MEMM 进行被告人及其所获罪名识别提取流程

Fig. 1　The recognition and extraction process of names of the accused and the offense he/she commited using HMM and MEMM trained on crime-related component tags

2.1　罪名相关成分标注语料库构建

2.1.1　罪名相关成分标注集

对裁判文书中被告人及其所获罪名的识别与提取是司法文本挖掘工作的难点,而建立罪名相关成分标注集、根据该标注集建立训练语料库,则是克服该难点的关键所在。

通过观察传统的分词软件如 ICTCLAS 对刑事裁判文书的词性标注结果可以发现,除了“盗窃罪”等三字短罪名能够被识别出来,其余较长的罪名常常被划分为两个或两个以上的词语,例如“交通肇事罪”的分词结果为“交通”和“肇事罪”。故需要根据罪名实体的分词情况,对罪名实体内部的构成词语建立标注。同时,为了实现被告人姓名与其所获罪名的关联,也需要对罪名实体的上下文词语建立标注。因此,本文将一个词语序列中的所有词划分为罪名的内部构成词语和罪名上下文词语,统称为罪名相关成分。总结罪名相关成分标注集如表 1 所示。

表 1　罪名相关成分标注集(注:文中隐去被告人姓名)

Tab. 1　The crime-related component tag set(Note: names of the accused have been hidden)

标注	成分含义	例子
A	罪名前第三个词语	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
B	罪名前第二个词语	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
C	罪名前第一个词语	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
D	罪名后第一个词语	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
E	罪名后第二个词语	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
F	罪名前部	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
G	罪名内部	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
H	罪名尾部	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
I	罪名后缀	被告人/陆 XX/犯/容留/他人/吸毒/罪/,/判处/……
J	罪名尾部和后缀成词	被告人/陆 XX/犯/交通/肇事罪/,/判处/……
K	可直接识别的罪名	被告人/蔡 XX/犯/盗窃罪/,/判处/……
Z	其他与罪名无关词语	

2.1.2　罪名相关成分标注语料库构建

借助最高人民法院发布的权威罪名列表,本文重新标注语料,将 ICTCLAS 的词性标注修改为如表 1 所示的罪名相关成分标注,构建面向刑事裁判文书的语料库,具体算法如表 2 所示。

采用上述算法,对多条刑事裁判文书进行罪名相关成分标注,得到多条标注后的裁判文书,如“被告人/A 陆 XX/B 犯/C 容留/F 他人/G 吸毒/H 罪/I,/D 判处/E 有期徒刑/Z 七/Z 个/Z 月/Z”,从而建立罪名相关成分标注语料库。

表 2 罪名相关成分标注算法

Tab. 2 The crime-related component tagging algorithm

Step 1	对最高人民法院发布的权威罪名列表中的分词后罪名 splitedCrime,判断分词后的裁判文书词语序列 splitedObservation 中是否含有该分词后罪名 splitedCrime
Step 2	如果含有罪名 splitedCrime,则记录 splitedCrime 在 splitedObservation 中的每次出现时的前一个词语的位置到 crimePosList,并且根据分词后的词语数量、各词语在 splitedCrime 中的位置以及词语自身的长度特性,来对 splitedCrime 内部构成词语更改标注:当数组大小为 1 时,可知该罪名可由分词软件直接识别,故将词性标注更改为 K;否则,将位置为 0 的词语的词性标注改为 F。如果数组中最后一个词语长度为 5 且包含"罪"这个字,可知这属于罪名尾部与罪名后缀被划分成一个词的情况,故将其词性标注改为 J;对最后一个位置以前到第一个位置之后的词语的词性标注改为 G。否则,可知最后一个词即为罪名后缀"罪"字,故将其词性标注改为 I;如果罪名的长度大于 2,则继续将倒数第二个词的词性标注改为 H;对倒数第二个位置以前到第一个位置之后的词语词性标注改为 G。通过该步骤,得到更改标注后的罪名 relabeledCrime
Step 3	根据 crimePosList,可以对 splitedObservation 中的罪名上下文词语进行标注。根据该步骤,得到更改了罪名上下文词语标注但尚未更改罪名内部构成词语的裁判文书词语序列 contextRelabeledObservation
Step 4	根据 crimePosList,将 contextRelabeledObservation 中的 splitedCrime 替换为 relabeledCrime。通过该步骤,得到对这一罪名的相关成分更改过标注的词语序列 relabeledObservation
Step 5	将 relabeledObservation 送入下一循环,遍历最高人民法院发布的权威罪名列表,对所有罪名重复 Step 1-4

2.2 基于罪名相关成分标注语料库的隐马尔科夫模型

隐马尔科夫模型在自然语言处理领域的应用,通常是将观察值序列 $O=\{o_0,o_1\cdots\cdots,o_{T-1}\}$ 相对应的状态序列 $S=\{s_0,s_1\cdots\cdots,s_{T-1}\}$ 抽象成马尔科夫链,针对相邻状态序列的转移情况进行建模。隐马尔科夫模型的 5 个模型参数 $Q,V,\boldsymbol{A},\boldsymbol{B},\boldsymbol{\pi}$ 说明如下:$Q=\{q_0,q_1,\cdots\cdots,q_{N-1},\}$,为所有状态的集合,即如表 1 所示的罪名相关成分标注;$V=\{v_0,v_1,\cdots\cdots,v_{M-1},\}$,为所有可能观察值的集合,即所有词语的集合;$\boldsymbol{A}=\{a_{ij}\}_{N\times N}$,其中 $a_{ij}=P(s_{t+1}=q_j\mid s_t=q_i)$,$0\leqslant t\leqslant T-1$,为状态转移概率矩阵;$\boldsymbol{B}=\{b_j(k)\}_{N\times M}$,其中 $b_j(k)=P(o_t=v_k\mid s_t=q_j)$,$0\leqslant t\leqslant T-1$,为状态——观察值发射概率矩阵;$\boldsymbol{\pi}=\{\pi_i\}_{N\times 1}$,其中 $\pi_i=P(s_1=q_i)$,为状态初始分布概率向量。

隐马尔科夫的模型参数 $\boldsymbol{A},\boldsymbol{B},\boldsymbol{\pi}$ 通过学习罪名相关成分标注语料库最大似然估计得到。转移概率矩阵 $\boldsymbol{A}$ 中由罪名相关成分q_i 转移到罪名相关成分 q_j 的转移概率 a_{ij} 的计算公式为

$$a_{ij}\approx C(q_i,q_j)/C(q_i)。\tag{1}$$

其中:$C(q_i,q_j)$ 为罪名相关成分 q_i 且下一个罪名相关成分为 q_j 的次数,$C(q_i)$ 为罪名相关成分 q_i 在语料库中出现的次数。

发射概率矩阵 $\boldsymbol{B}$ 中罪名相关成分 q_j 下词语 v_k 的发射概率 $b_j(k)$ 的计算公式为

$$b_j(k)\approx C(v_k,q_j)/C(q_j)。\tag{2}$$

其中:$C(v_k,q_j)$ 为语料库中罪名相关成分为 q_j 的词语 v_k 的出现次数,$C(q_j)$ 为罪名相关成分 q_j 的词语在语料中出现的总次数。

在实际计算过程中,发射概率矩阵 $\boldsymbol{B}$ 并不预先在参数训练步骤算出,因为在实验所用词语序列中出现的词语未必已经登录在语料库中,故仅仅统计频数 $C(v_k,q_j)$ 和 $C(q_j)$ 即可。解决该问题的办法如下:在使用 Viterbi 算法来标注序列时,首先需要判断语料库中是否已经存在该词语。如果存在,则根据统计的频数计算其发射概率,如果不存在,则默认返回 1,以避免出现概率为 0 的情况。

2.3 基于罪名相关成分标注语料库的最大熵马尔科夫模型

最大熵马尔科夫模型由 Mccallum 等(2000)[19]在隐马尔科夫模型的基础上改造而得。最大熵马尔科夫模型的两个参数 $Q,\boldsymbol{\pi}$ 均与隐马尔科夫模型的相同。与隐马尔科夫模型相比,其不同之处在于:①最大熵马尔科夫模型对特征的使用不再局限于前一个标注状态,而是从观察值中抽取若干非独立特征,更加充分灵

活地学习训练集。故最大熵马尔科夫模型的参数 $V=\{v_0,v_1,\cdots\cdots,v_{M-1},\}$ 不再是所有可能观察值的集合,而是从所有观察值中抽取出来的特征构成的集合。②隐马尔科夫模型中,当前状态只与前一状态有关;而最大熵马尔科夫模型则通过对转换函数 $P(s_t=q_j\mid s_{t-1}=q_i,o_t=v_k)$ 建模,用该状态转换函数代替了隐马尔科夫模型中的状态转移概率和状态——观察值发射概率,使得当前状态不仅依赖于前一状态,还依赖于当前观察值的特征,从而克服了隐马尔科夫模型观察值之间相互独立的假设,将上下文信息引入到模型中来。故最大熵马尔科夫模型参数中不再有转移概率矩阵 $\boldsymbol{A}$ 和发射概率矩阵 $\boldsymbol{B}$,取而代之的是由 s_{t-1} 的函数 $P(s_t=q_j\mid s_{t-1}=q_i,o_t=v_k)$ 构成的集合。

最大熵马尔科夫模型依据最大熵原理对 s_{t-1} 的函数 $P(s_t=q_j\mid s_{t-1}=q_i,o_t=v_k)$ 进行建模,认为在满足训练数据约束的同时拥有最大信息熵的概率分布是最佳模型。为了表示训练数据所面临的约束,最大熵马尔科夫模型定义了特征函数 $f_{v_k,q_i}(o_t,s_t)$,o_t 为当前观察值的特征,s_t 为当前观察值可能对应的标注状态。

$$f_{v_k,q_i}(o_t,s_t)=\begin{cases}1, & \text{若当前观察值的特征}o_t\text{为}v_k\text{,且当前标注状态}s_t=q_i;\\ 0, & \text{否则。}\end{cases}\tag{3}$$

基于特征函数,可以写出建模时应满足的约束条件:特征函数 $f_{v_k,q_i}(o_t,s_t)$ 关于经验分布 $\widetilde{P}(o_t,s_t)$ 的期望 $E_{\tilde{p}}(f)$ 应当与特征函数关于模型 $P(s_t=q_j\mid s_{t-1}=q_i,o_t=v_k)$ 和经验分布 $\widetilde{P}(o_t)$ 的期望 $E_p(f)$ 相等,即

$$\frac{1}{m_{s_t}}\sum_{n=1}^{m_{s_t}}f_{v_k,q_i}(o_{t_n},s_{t_n})=\frac{1}{m_{s_t}}\sum_{n=1}^{m_{s_t}}\sum_{s\in S}P(s_t=q_j\mid s_{t-1}=q_i,o_{t_n}=v_k)f_{v_k,q_i}(o_{t_n},s_t)。\tag{4}$$

其中,由于是对 s_{t-1} 的函数 $P(s_t=q_j\mid s_{t-1}=q_i,o_t=v_k)$ 建模,故 t_1,t_2,……,t_n,…… $t_{m_{s_t}}$ 表示的是序列中 $s_{t_n}=s_{t-1}$ 时对应的状态转移步次。

在该约束下最大化信息熵,即可得到。

$$P(s_t=q_j\mid s_{t-1}=q_i,o_t=v_k)=\frac{1}{Z(o_t,s_{t-1})}\exp\left(\sum_{v_k,q_i}\lambda_{v_k,q_i}f_{v_k,q_i}(o_t,s_t)\right)。\tag{5}$$

其中,λ_{v_k,q_i} 是可以通过 IIS 算法训练得到的参数,$Z(o_t,s_{t-1})$ 是正则化因子。

2.4　基于 Viterbi 算法的罪名相关成分标注

在建立隐马尔科夫模型和最大熵马尔科夫模型后,基于其序列标注均可以通过 Viterbi 算法来求解。考虑到计算过程中有可能出现概率值过小而面临的数据下溢问题,对概率值取负对数,将最大化问题转换为最小化问题来进行求解。

对于一条刑事裁判文书的词语序列 $O=\{o_0,o_1\cdots\cdots,o_{T-1}\}$,以隐马尔科夫模型为例,应用 Viterbi 算法解码其罪名相关成分标注序列 $S=\{s_0,s_1\cdots\cdots,s_{T-1}\}$ 的具体算法如表 3 所示。

最大熵马尔模型下应用 Viterbi 算法对裁判文书词语序列进行罪名相关成分标注的步骤与隐马尔可夫模型下的类似,只需将上述算法中的转移概率和发射概率替代为状态转换函数即可。

表 3　隐马尔科夫模型下应用 Viterbi 算法对裁判文书词语序列进行罪名相关成分标注

Tab. 3　Labeling word sequences in judicial documents with crime-related component tags using Viterbi Algorithm under HMM

Step 1	令 $\delta_0(i)=-\ln(\pi_i)-\ln(b_i(o_0))$,$\psi_0(i)=0$,对于 $i=0,1,\cdots\cdots,N-1$
Step 2	对于 $t=1,2,\cdots\cdots,T-1$ 和 $i=0,1,\cdots\cdots,N-1$,计算 $\delta_t(i)=\min_{j\in\{0,1,\cdots\cdots,N-1\}}[-\ln(\delta_{t-1}(j))-\ln(a_{ji})-\ln(b_i(o_t))]$,记录使 $\delta_t(i)$ 取到最小值的标注状态 $\psi_t(i)=arg\min_{j\in\{0,1,\cdots\cdots,N-1\}}[-\ln(\delta_{t-1}(j))-\ln(a_{ji})-\ln(b_i(o_t))]$
Step 3	当 $t=T-1$ 时终止上一步,总体最优路径的概率为 $P^*=\min_{j\in\{0,1,\cdots,N-1\}}[\delta_{T-1}(j)]$,$s_{T-1}=arg\min_{j\in\{0,1,\cdots,N-1\}}[\delta_{T-1}(j)]$
Step 4	回溯之前记录的每一步计算过程中选择的最优标注状态,对于 $t=T-2,T-3,\cdots\cdots0$,有 $s_t=\psi_{t+1}(s_{t+1})$,最终可以从终点回溯得到整条最优标注序列 $S=\{s_0,s_1\cdots\cdots,s_{T-1}\}$

2.5 识别并提取被告人及其罪名

对已经进行过罪名相关成分标注的刑事裁判文书词语序列进行简单的模式识别,从而提取出该条判决文书对应的一条或多条被告人及其罪名信息。据表 1 罪名相关成分标注集可知,罪名自身结构主要有表 4 中的三种形式,找出标注序列为"$FG*H?\ I$"、"FJ"、"K"的词语序列,即可识别并提取出罪名实体。

表 4 罪名内部结构

Tab. 4 The inner structure of crime name

罪名内部结构	对应模式	例子
罪名前部+罪名内部(0 次或多次)+罪名尾部+罪名后缀	$FG*H?\ I$	/容留/他人/吸毒/罪/
罪名前部+罪名尾部和后缀成词	FJ	/交通/肇事罪/
可直接识别的罪名	K	盗窃罪/

罪名自身和其上下文主要为:"ABC+罪名+DE"。通过观察和统计各个标注的词语,发现罪名前第 2 个词语(即标注为 B 的词语)为人名的概率为 76.16%,最高,而将近 20%的误差主要来源于"因"、"涉嫌"两词。因此选用罪名前第二个词语作为该罪名对应的被告人,并设置规则去掉"因"和"涉嫌"以修正提取结果。通过对罪名及其上下文进行简单模式识别,实现了被告人姓名及被告人获罪罪名的关联,最终可以提取得到一条裁判文书中所含有的一条或多条被告人及其获罪罪名信息。

3 实验与分析

3.1 实验语料

将从中国裁判文书网(http://wenshu.court.gov.cn/)抓取的 5 000 篇的刑事裁判文书,经过 ICTCLAS 词性标注后分成两个部分,前 4 000 篇含有的 6 378 条概要信息用作训练数据,随机抽取并建立大小分别为 1 000、2 000、3 000、4 000、5 000 和 6 000 条概要信息的罪名相关成分标注训练集,用于第 3 节中隐马尔科夫模型和最大熵马尔科夫模型的参数学习;剩余的 1 000 篇作为测试集,测试集的刑事裁判文书中约有概要信息 1 632 条,存放在测试文书数据库中;用于本节中的裁判文书概要信息提取实验。具体实验流程如图 2 所示。

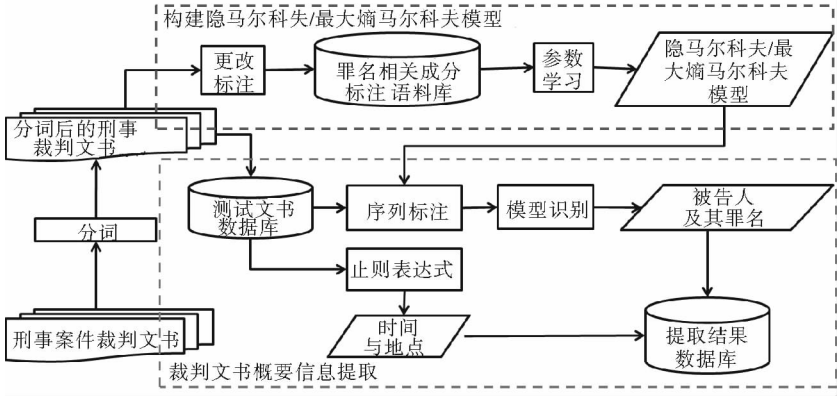

图 2 裁判文书概要信息提取实验流程

Fig. 2 The process of experiments on information extraction of judical documents

3.2 评价指标

实验把基于最高人民法院发布的权威刑法罪名和相关规则直接识别提取罪名实体的方法作为 Baseline 方法,与本文提出的基于罪名相关成分标注语料库的隐马尔科夫、最大熵马尔科夫模型来进行罪名实体识别的算法进行比较。使用准确率 P、召回率 R 和 F 值来评价刑事裁判文书概要信息的提取情况,具体参数为:

$$P = \text{本文算法正确提取出的概要信息条数} / \text{本文算法提取出的概要信息条数}, \tag{6}$$

$$R = \text{本文算法正确提取出的概要信息条数} / \text{刑事裁判文书内实际概要信息条数}, \tag{7}$$

$$F = R \times P \times (1+\beta^2)/R + P \times \beta^2。 \tag{8}$$

其中,β 是准确率与召回率之间的权衡因子。本研究认为准确率和召回率同样重要,因此取 β 为 1。

由于在较充分总结规则的前提下,使用基于规则的方法直接提取案件发生时间与地点的准确率极高,几乎没有任何错误,因此对概要信息的正确提取等同于对被告人及被告人所获罪名的正确识别。

除了准确率 P、召回率 R 和 F 值,由于提出的算法主要应用于司法信息挖掘与互联网信贷平台建设,需要对海量的裁判文书进行操作,因此,算法的运行时间也是本研究关注的一个重要评价指标。由于算法的运行速度取决于硬件条件等多种因素,其绝对运行时间并不具有可比性。故使用对其运行时间进行相对衡量。

$$\text{相对运行时间} = \text{实际运行时间} / t_0。 \tag{9}$$

其中,t_0 为学习最小训练集得到的模型运行时间。相对运行时间代表了模型在不同大小训练集下的时间复杂度。

3.3　实验及结果分析

利用基于罪名相关成分标注语料库的 HMM 和 MEMM 模型,来对刑事裁判文书中的被告人及被告人所获罪名进行识别和提取。设计实验一以观察提出的算法基于不同大小的罪名相关成分标注语料库的 HMM 和 MEMM 的性能。实验二是提出的算法和直接基于规则提取被告人及其罪名的 Baseline 方法的比较。

3.3.1　实验一

在实验一中,使用不同大小的罪名相关成分标注语料库训练出的隐马尔科夫模型、最大熵马尔科夫模型,对刑事裁判文书进行概要信息提取,并通过准确率、召回率、F 值以及相对运行时间比较不同语料库大小下该算法的性能。

如图 3～5 所示,随着罪名相关成分训练集的增大,准确率、召回率、F 值逐步提高,但是提高幅度趋于平缓,最终维持在一个相对稳定的水平。再对比最大熵马尔科夫与隐马尔科夫模型的表现,可以发现,总体而

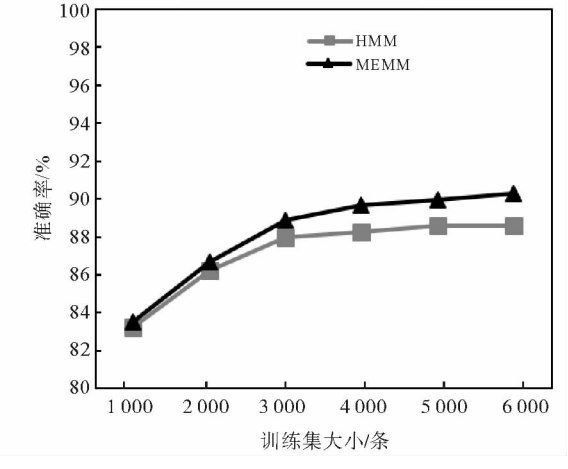

图 3　基于不同大小的罪名相关成分标注语料库的 HMM 和 MEMM 进行概要信息提取的准确率变化曲线

Fig. 3　The curve of the change of precisionof information extraction based on HMM and MEMM with respect to the size of training set labelled by crime-related tags

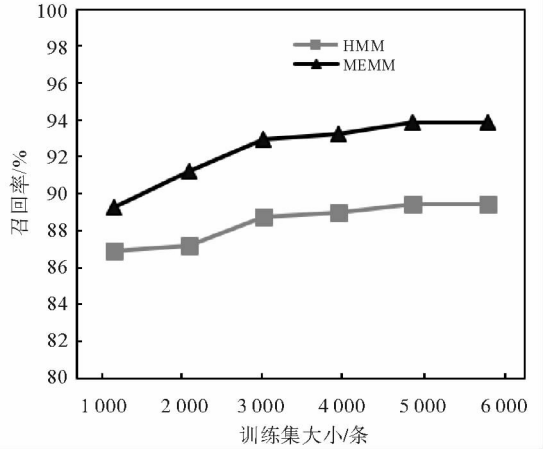

图 4　基于不同大小的罪名相关成分标注语料库的 HMM 和 MEMM 进行概要信息提取的召回率变化曲线

Fig. 4　The curve of the change of recallof information extraction based on HMM and MEMM with respect to the size of training set labelled by crime-related tags

言,最大熵马尔科夫模型的准确率、召回率和 F 值要高于隐马尔科夫模型,随着训练集的增大,二者差距亦逐渐增大。这是由于隐马尔科夫模型假设词语序列的各个词语相互独立,每个时刻的词语只依赖于此时刻的标注状态。该假设在较小的训练集上较为合适,但随着训练集的增大,隐马尔科夫模型无法再覆盖训练数据的更多特征。因此隐马尔科夫模型相较于最大熵马尔科夫模型,在准确率、召回率方面的劣势随着用于训练的罪名相关成分语料库的增大而凸显。

如图 6 所示,虽然随着训练集的增大,最大熵马尔科夫在准确率等方面的表现明显优于隐马尔科夫模型,但是由于其覆盖的特征增多,模型参数的训练收敛速度明显减慢。因此,考虑到刑事裁判文书概要信息提取主要用于 P2P 网贷公司的贷款申请人的审核和信用风险管理,要对海量的裁判文书进行操作,在选择使用本文提出的两种机器学习模型时需在准确率与速度之间做一定的权衡。

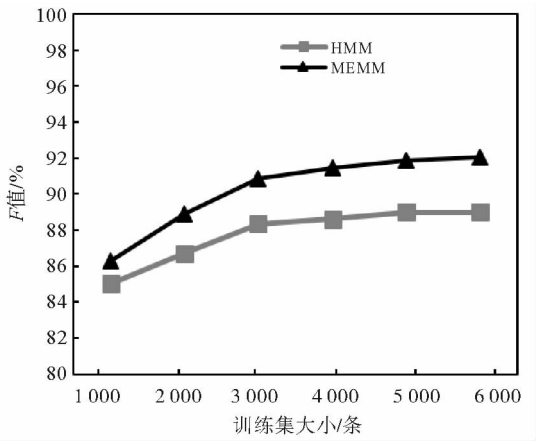

图 5 基于不同大小的罪名相关成分标注语料库的 HMM 和 MEMM 进行概要信息提取的 F 值变化曲线

Fig. 5 The curve of the change of F values of information extraction based on HMM and MEMM with respect to the size of training set labelled by crime-related tags

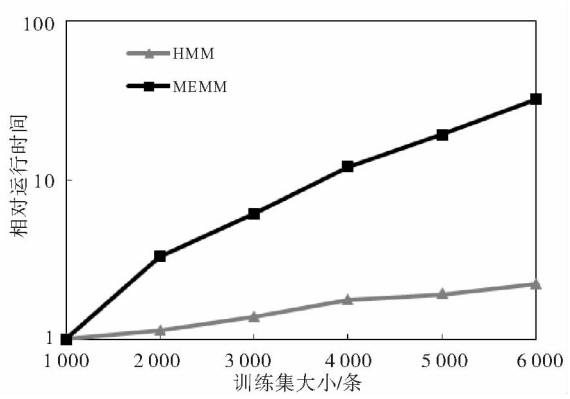

图 6 基于不同大小的罪名相关成分标注语料库的 HMM 和 MEMM 进行概要信息提取的相对运行时间变化曲线(注:对数刻度)

Fig. 6 The curve of the change of running time of information extraction based on HMM and MEMM with respect to the size of training set labelled by crime-related tags(Note: logarithmic scale)

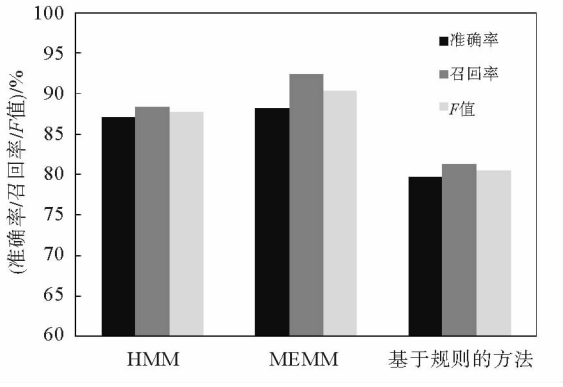

图 7 提出算法和基于规则的 Baseline 方法的准确率、召回率、F 值对比

Fig. 7 Comparison between the precision, recall and F values of method proposed by us and Baseline method based on rules

3.3.2 实验二

实验二比较了提出的算法和直接基于规则的 Baseline 方法,使用了最大熵马尔科夫模型、隐马尔科夫模型在不同大小训练集下的准确率、召回率、F 值的平均值,来与 Baseline 方法进行比较。

如图 7 所示,对于刑事裁判文书的概要信息提取,基于罪名相关成分标注的 HMM 进行的被告人及其所获罪名识别准确率达到 87.14%、召回率为 88.45%和 F 值 87.79%,基于 MEMM 进行的识别达到了 88.17%的准确率、92.41%的召回率和 90.25%的 F 值,相对于直接使用相关规则和最高人民法院发布的权

威罪名的 Baseline 方法,克服了裁判文书格式的差异和罪名实体识别的困难,具有更好的效果。

4　结论及下一步研究工作

针对司法文本挖掘的难点,探索了基于机器学习的命名实体识别技术在该领域的一种应用:针对刑事案件裁判文书,借助最高人民法院发布的权威罪名列表,先建立罪名相关成分标注语料库;通过对该语料库的学习得到 HMM 和 MEMM 的模型参数,从而实现了对司法文本的概要信息,尤其是对被告人及其所获罪名实体的识别与提取。实验证明,基于罪名相关成分标注语料库训练得到的隐马尔科夫模型和最大熵马尔科夫模型对刑事裁判文书概要信息的提取效果较好,其 F 值分别达到 87.79%与 90.25%,为进一步的司法领域文本挖掘奠定基础,亦可用于互联网信贷平台的信用风险管理。

由于 MEMM 能够更为灵活地学习训练数据集,故 MEMM 特征函数的选取是重要的,在训练最大熵马尔科夫模型(MEMM)时仅使用了较为简单的低维特征。在以后的研究中,探索可为 MEMM 所利用的刑事案件裁判文书文本特征,以进一步提高刑事案件裁判文书概要信息提取的精度。

参考文献:

[1]王宁,葛瑞芳,苑春法,等. 中文金融新闻中公司名的识别[J]. 中文信息学报,2002,16(2):1-6.
WANG Ning, GE Ruifang, YUAN Chunfa, et al. Companyname identification in chinese financial domain[J]. Journal of Chinese Information Processing, 2002, 16(2): 1-6.

[2]SAVOVA G K, MASANZ J J, OGREN P V, et al. Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications[J]. Journal of the American Medical Informatics Association Jamia, 2010, 17(5): 507 - 513.

[3]ZHANG J, EL-GOHARY N M. Semantic nlp-based information extraction from construction regulatory documents for automated compliance checking[J]. Journal of Computing in Civil Engineering, 2013, 30(2): 710-711.

[4]DERCZYNSKI L, MAYNARD D, RIZZO G, et al. Analysis of named entity recognition and linking for tweets[J]. Information Processing & Management, 2015, 51(2): 32-49.

[5]ELSNER M, CHARNIAK E, JOHNSON M. Structured generative models for unsupervised named-entity clustering[C]// Human Language Technologies Proceedings: Conference of the North American Chapter of the Association of Computational Linguistics. Boulder, Colorado, May 31-June 5, 2009: 164-172.

[6]MUNRO R, MANNING C D. Accurate unsupervised joint named-entity extraction from unaligned parallel text[C]// Named Entity Workshop. Association for Computational Linguistics, Jeju, July 10-12, 2012: 21-29.

[7]THENMALAR S, BALAJI J, GEETHA T V. Semi-supervised bootstrapping approach for named entity recognition[J]. International Journal on Natural Language Computing, 2015, 10(4): 1-14.

[8]BIKEL D M, MILLER S, SCHWARTZ R, et al. Nymble: Ahigh-performance learning name-finder[C]// Proceedings of the Fifth Conference on Applied Natural Language Processing. Washington D C, March 31-April 3, 1997: 194-201.

[9]SEKINE S. NYU: Description of the japanese ne system used for MET-2[C]// Proceedings of the 7th Message Understanding Conference. Fairfax, Virginia, April 29-May 1, 1998: 28-32.

[10]BORTHWICK A, STERLING J, AGICHTEIN E, et al. NYU: Description of the MENE named entity system as used in MUC-7[C]// Proceedings of the 7th Message Understanding Conference. Fairfax, Virginia, April 29-May 1, 1998: 56-77.

[11]ASAHARA M, MATSUMOTO Y. Japanese named entity extraction with redundant morphological analysis[C]// Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edonton, May 27-June 1, 2003: 8-15.

[12]LAFFERTY J, MCCALLUM A, PEREIRA F, et al. Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the 18th International Conference. on Machine Learning. San Francisco, California, June 28-July 1, 2001: 282-289.

[13]EKBAL A, BANDYOPADHYAY S. Named entity recognition using support vector machine: A language independent approach[J]. International Journal of Computer Systems Science & Engineering, 2010(2): 155.

(下转第 124 页)

ZHAO Zhigao, YANG Jianmin, WANG Lei, et al. The development and research of dynamic positioning system[J]. Oceanographic Engineering, 2002, 20(1): 91-97.

[4]谢彬. 深水半潜式钻井平台设计与建造技术[M]. 北京: 石油工业出版社, 2013: 196-197.

[5]TANNURI E A, AGOSTINHOA C, MORISHITA H M, et al. Dynamic positioning systems: An experimental analysis of sliding mode control[J]. Control Engineering Practice, 2010, 18(10): 1121-1132.

[6]BESSA W M, DUTRA M S, KREUZER E. Dynamic positioning of underwater robotic vehicles with thruster dynamics compensation[J]. International Journal of Advanced Robotic Systems, 2013, 10: 1-8.

[7]李娇娇, 施小成, 孟羽泽, 等. 基于非线性滤波器的船舶动力定位系统[J]. 应用科技, 2017, 44(2): 23-28.
LI Jiaojiao, SHI Xiaocheng, MENG Yuze, et al. The dynamic positioning system of ships based on nonlinear filter[J]. Applied Technology, 2017, 44(2): 23-28.

[8]BALCHEN J G, JENSSEN N A, MATHISEN E, et al. Dynamic positioning of floating vessles based on Kalman Filtering and optimal control [C]// IEEE Conference on Decision and Control Including the Symposium on adaptive Processes. 1980, 107(3): 852-864.

[9]边信黔, 付明玉, 王元慧. 船舶动力定位[M]. 北京: 科学出版社, 2011: 51-53.

[10]贾欣乐, 杨盐生. 船舶运动数学模型-机理建模与辨识建模[M]. 大连: 大连海事大学出版社, 1999: 106-108.

[11] 郑大钟. 线性系统理论[M]. 2 版. 北京: 清华大学出版社, 2002: 218-230.

[12]WEI L, PANG Y J. Research on dynamic simulation of DP for a deep water semi-submersible platform[C]// Proceedings of 2012 International Conference on Mechanical Engineering and Material Science. 2012: 642-645.

[13]杨杨. 动力定位船舶非线性自适应控制研究[D]. 大连: 大连海事大学, 2013: 26-29.

[14]刘金琨. 滑模变结构控制 MATLAB 仿真[M]. 北京: 清华大学出版社, 2015: 217-222.

(责任编辑: 李 磊)

(上接第 101 页)

[14]RITTER A, WRIGHT E, CASEY W, et al. Weakly supervised extraction of computer security events from twitter[C]// Proceedings of the 24th International Conference on World Wide Web. Florence, May 18-22, 2015: 896-905.

[15]张华平, 刘群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报, 2004, 27(1): 85-91.
ZHANG Huaping, LIU Qun. Automatic recognition of Chinese personal name based on role tagging[J]. Chinese Journal of Computers, 2004, 27(1): 85-91.

[16]YE F, CHEN Y, ZHOU G, et al. Intelligent recognition of named entity in electronic medical records[J]. Chinese Journal of Biomedical Engineering, 2011, 30(2): 256-262.

[17]PENG N, DREDZE M. Named entity recognition for chinese social media with jointly trained embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Sep. 17-21, 2015: 548-554.

[18]LIU K, EL-GOHARY N. Ontology-based sequence labelling for automated information extraction for supporting bridge data analytics[J]. Procedia Engineering, 2016, 145(1): 504-510.

[19]MCCALLUM A, FREITAG D, PEREIRA F C N. Maximum entropy Markov models for information extraction and segmentation[C]// Proceedings of the 17th International Conference on Machine Learning. San Francisco, June 29-July 2, 2000: 591-598.

(责任编辑: 傅 游)