基于卷积神经网络的辅助分案方法研究

敖绍林[1]，秦永彬[1,2]，黄瑞章[1,2]，陈艳平[1,2]，刘丽娟[3]，郑庆华[4]，陈昌恒[5]，程少芬[5]
1. 贵州大学计算机科学与技术学院，贵州 贵阳 550025；
2. 公共大数据国家重点实验室，贵州 贵阳 550025；
3. 贵州师范学院，贵州 贵阳 550025；
4. 西安交通大学计算机科学与技术学院，陕西 西安 710049；
5. 贵州省高级人民法院，贵州 贵阳 550081

摘要

法院系统中主要有人工指定分案和简单随机分案两种模式。这两种模式无法做到人案的自动匹配，存在金钱案、关系案等弊端。目前分案方法的相关研究主要存在法官表示和案件匹配两个难点。结合法官历史审判数据，在法官表示中融合法官擅长的审判领域，提出一种融合审判质量的法官表示方法。然后，通过卷积神经网络学习案件表示和法官表示中不同粒度的抽象语义特征向量，计算案件和多个法官的特征向量间的余弦相似度，用向量相似度表示案件与法官的匹配度，输出匹配值较高的前*N*个法官作为案件的推荐法官。在贵州省某法院真实数据下进行实验，结果表明该方法推荐法官的正确率比传统方法高80%。

关键词

文本表示；卷积神经网络；智能分案；智慧法院

中图分类号：TP183 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022020

Research on auxiliary division method based on convolutional neural network

AO Shaolin[1], QIN Yongbin[1,2], HUANG Ruizhang[1,2], CHEN Yanping[1,2], LIU Lijuan[3], ZHENG Qinghua[4], CHEN Changheng[5], CHENG Shaofen[5]
1. School of Computer Science and Technology, Guizhou University, Guiyang 550025, China
2. State Key Laboratory of Public Big Data, Guiyang 550025, China
3. Guizhou Education University, Guiyang 550025, China
4. School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
5. Guizhou Higher People's Court, Guiyang 550081, China

Abstract

The court system mainly has two modes: manual designated division and simple random division. The above method

cannot achieve automatic matching of persons and cases, and there are drawbacks such as money cases and relationship cases. At present, the research on division method mainly has two difficulties: judge's representation and case matching. Combining the judge's historical trial data, the judge's expertise in the judge's representation was integrated, and a judge representation method that integrates the quality of the trial was proposed. Then, the abstract semantic feature vectors of different granularities in the case representation and the judge representation were learned through the convolutional neural network, the cosine similarity between the case and the feature vectors of multiple judges was calculated, and vector similarity was used to indicate the matching degree between the case and the judge, the top N judges with high matching value were output as recommended judges for the case. Experiments with real data from a court in Guizhou Province, and the results show that the accuracy of the method for recommending judges is 80% higher than the traditional method.

Key words
text representation, convolutional neural network, smart division, smart court

0 引言

案件分配是诉讼程序的重要环节，也是审判管理的重要内容，对合理调配法院审判资源、激发法官办案积极性具有关键作用。我国法院分案制度改革的历史脉络大致是从人工指定分案发展为计算机随机分案。目前我国各级法院的分案方法还是简单随机分案，具体表现是以法官分配到的案件数量相等为目标的均衡分案和不考虑法官专业能力、案件性质的完全随机分案（比如摇号分案），存在人案不适问题。

随着国家提出建设智慧法院[1]和实行员额制改革，现有分案方法已无法适应新型办案机制。在员额制改革的背景下，法官团队进一步实现专业化、精英化、职业化。笔者认为应将案件分配给擅长审判这类案件的法官。针对上述问题，本文的研究目标是将案件自动分配给擅长审判此类案件的法官，形成专业化的办案模式，避免司法腐败，提高办案质效。然而，实现自动分案目前还存在以下两个研究难点。

- 表示困难。法院系统存储了法官的基本信息和历史审判数据，其中多为文本信息和元数据。如何在法官表示中融合法官抽象语义特征并体现法官擅长的审判领域是实现自动分案的一个难点。
- 匹配困难。如何将案件表示和法官表示自动映射到一个高阶语义空间，自动获取案件表示和法官表示中的关联语义信息，计算案件和法官匹配度是实现自动分案的另一个难点。

针对以上难点，本文提出融合审判质量的法官表示方法，以突出法官擅长的审判领域。利用案情事实描述表示案件，然后利用三元组损失（triplet loss）技术调节卷积神经网络（convolutional neural network，CNN），使其更好地学习法官表示和案件表示的语义特征向量。本文主要贡献如下。

- 提出一种融合案件审判质量的法官表示方法。通过审判质量评价指标，得出法官在各类案件下的审判质量权重。利用法官审判质量高的案件语义特征表示法官擅长的审判领域，从而在法官表示中融合法官擅长领域的抽象语义信息。
- 提出利用CNN学习案件和法官的语义特征向量，通过相似性函数自动计算案件和法官的匹配度。该方法构造了一个三元组，该三元组由案件表示、擅长审判此案件的法官表示和不擅长审判此案件的法

官表示组成。在高阶语义空间中，利用三元组损失技术调节CNN，使其更好地学习案件表示和法官表示的语义特征向量，然后在非线性空间中计算案件和法官特征向量间的余弦相似度，用向量相似度表示案件和法官的匹配度。

本文的分案方法不同于均衡分案，更不同于庭长指定分案。该方法通过同时考虑法官擅长审判领域和案件信息实现自动分案，减少了案件分配过程中的人为干扰因素，以人案匹配为目标，实现公正、合理的分案。

1 相关工作

本文利用融合案件审判质量表示法官，利用案情事实描述表示案件。基于卷积神经网络和余弦相似度方法实现案件和法官的自动匹配。其主要工作涉及文本表示、裁判文书的分析与应用两个方面。

早期的文本表示主要基于向量空间模型。代表方法是词频-逆文档频率（term frequency-inverse document frequency，TF-IDF）[2]。TF-IDF根据词在文档集中的重要度来表示文档，忽略了词的上下文关系，无法表示语义。这一时期的文本表示方法无法表示文本间词的位置信息及上下文语义信息。文本分布式表示的提出旨在解决上述缺陷，早期主要是基于主题模型的方法，这类方法从文本库中发现文本的代表性主题，由此计算每篇文档的主题分布，代表方法有概率潜在语义分析（probabilistic latent semantic analysis，PLSA）模型[3]和隐含狄利克雷分布（latent Dirichlet allocation，LDA）模型[4]。裁判文书数据具有逻辑关系严谨、时序关系与因果关系明显等典型特征，利用基于向量空间模型和主题模型的文本表示方法无法较好地体现其语义关系。近年来，深度学习技术的发展较好地提升了文本表征能力，这一类方法可被统称为基于神经网络的文本表示方法。Bengio Y等人[5]在2003年提出神经网络语言模型（neural network language model，NNLM），用神经网络建模n-gram，进而得到表征单词语义的词向量。2013年，Mikolov T等人[6-7]提出著名的word2vec模型来训练词向量，语义上相似或相关的词得到的表示向量相近。在word2vec之后词的分布式表示技术得到了长足的发展。2014年Pennington J等人[8]提出Glove模型，对词向量进行全局意义上的学习。2017年Bojanowski P等人[9]提出FastText模型，以学习词的形态学信息。直到ELMo[10]、BERT（bidirectional encoder representation from transformer）[11]等模型被提出，文本语义表示才开始在考虑词的形态学信息的同时兼顾上下文语义信息。另外，由于TF-IDF在特征提取方面存在缺点，2014年Kim Y[12]提出了Text-CNN模型，利用CNN捕捉局部特征能力强的特点，将句子经过卷积层、池化层得到句子的表示，在文本分类任务上取得了不错的效果。2019年冯兴杰等人[13]提出基于多注意力的CNN问题相似度计算模型，与基于循环神经网络的模型相比，该模型对问句的识别能力更强。Chiu J P C等人[14]设计了一种双向长短期记忆（bi-directional long short-term memory，Bi-LSTM）和CNN结合的神经网络模型，实验证明了该模型能较好地获得句子的结构化表示。

本文研究如何表示法官以突出法官擅长的审判领域，以及如何实现法官和案件的自动匹配。这涉及表示向量在非线性空间的高阶语义匹配问题，本文研究利用深度学习方法获取句子的抽象语义表示。

随着国家司法信息化建设的推进，提高案件受理、审判、执行、监督等各环节的信息化水平，促进司法公平正义成为必然的趋势，分案过程自动化、智能化对于促进国家司法信息化体系建设具有重要的推动作用。关于分案制度，最高人民法院在相关文件中多次对建立“随机分案为主，指定分案为辅”的分案方式提出指导意见[15]。目前世界各国都在积极探索随机分案制度。

在美国，州法院使用计算机进行随机分案，而联邦法院通过人为考虑案件的争议点和复杂性实现分案[16-17]。在德国[18]，先对案件的分配工作做出安排，在之后的一个审判年度内，新受理案件都必须按照预先安排进行分配。在中国，北京国双科技有限公司[19]通过建立案件、法官实体数据以及两者之间的关系，利用分案因素，从法官实体数据中匹配法官列表，将待分配案件随机分配给法官列表中的法官。广州大学[20]基于机器学习方法和人工决策相结合的模式实现自动分案。陈芳序[21]以法官工作量为导向，打破以往以案件为导向的分案思维，利用统计学相关知识进行Pearson相关性分析，实现对案件工作量的评估，从而确认法官工作量，在工作量较小的$N-1$个法官中随机选择一个法官分配案件。该方法为了平衡法官工作量，只考虑案件因素实现分案，往往会造成人案不适，降低公众对司法的信任。王小新[22]以江苏省法院试行的刑事案件难易程度权重为基础，通过将案件审理难度看作二分类问题，利用Logistic回归方法构建个案审理难度评估模型，判断不同法官审理同一案件的难度系数，以全院审理案件难度系数最小为约束实现案件的最优分配。

在这些研究的基础上，本文提出基于卷积神经网络模型的辅助分案方法。针对现有方法存在的人案不适、人情案、关系案等弊端，提出一种融合案件审判质量的法官表示方法，利用法官审判质量高的案件语义特征表示法官擅长的审判领域，从而在法官表示中融合法官擅长领域的抽象语义信息。最后，利用卷积神经网络实现案件与法官的匹配，从而实现高效率分案。

2 分案模型的实现

本文基于卷积神经网络获取法官和案件的语义特征表示，利用相似性函数计算案件和任何一个法官的匹配度，再通过分案模块得出推荐法官。本文方法的分案流程如**图1**所示，主要包括案件审判质量评价模块、表示模块、匹配度估算模块、分案模块。

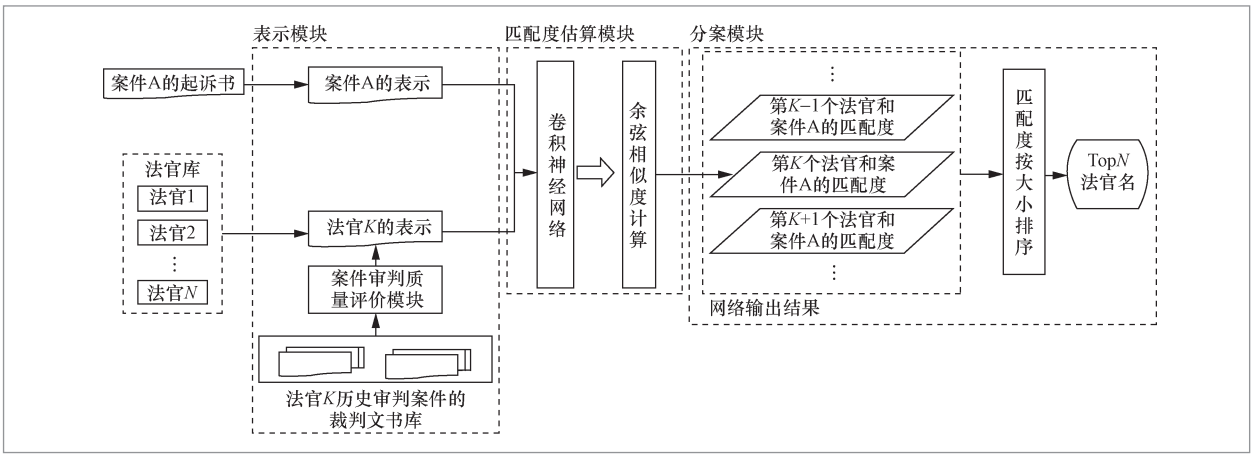

图1 本文方法的分案流程

本文首先对数据进行预处理，通过表示模块得到案件和法官的表示。通过匹配度估算模块自动计算案件和法官的匹配度，得出案件和任何一个法官的匹配度。最后基于分案模块将案件和所有法官的匹配度按大小排序，输出匹配度较大的前N个法官，即案件推荐的Top$N$个承办法官。

2.1 融合案件审判质量的法官表示方法

传统案件分配仅将案由作为唯一分案标准，忽略了法官擅长的审判领域，无法保证为案件分配的法官擅长审判此类案件，常常造成人案不适。为了解决这一问题，本文提出融合案件审判质量的法官表示方法，以突出法官专长。法官的历史审判案件众多，不同案件的审判质量高低不同。本文认为法官审判质量较高的案件就是法官擅长审判的案件，利用这类案件语义信息能反映法官的专长、判案思维、审判习惯。

2011年最高人民法院在《关于开展案件质量评估工作的指导意见》[23]中公布了31项用于评估法院整体案件审判质量的指标。本文从中选取一审改判发回重审率、案均审理时间、法定正常审限内结案率3个指标评估法官个人对案件的审判质量。法官对任何一类案件的审判质量权重计算如下：

$$w_j = \frac{\vartheta \cdot \theta_j + 1}{\gamma \cdot \alpha_j + \mu_j \cdot \beta_j + 1} \qquad (1)$$

其中，w_j表示法官对任何一类案件的审判质量权重；$j=1,\cdots,m$表示案由数，则每一个法官在m类案由下有m个权重值；ϑ、γ μ_j是调节因子；θ表示法定正常审限内结案率；α表示一审改判发回重审率；β表示案均审理时间。分子分母加1的目的是对式子进行平滑处理。

在m类案由下，比较同一法官不同类别下的w_j值，权重值越高，法官对该类案件的审判质量越高。本文认为审判质量最高的这类案件是法官擅长审判的案件。由此可以得到任何一个法官擅长审判的案件类型和不擅长审判的案件类型，保证后续实验合理构建三元组数据。通过对裁判文书的分析，笔者发现案情事实描述是案件判决的主要依据。由此，本文抽取案情事实描述构成案件的表示。法官的表示则由多个案件的案情特征构成。

实验时考虑到法官表示文本的长度，规定构成法官表示的案件个数为5。本文通过设定参数ε改变构成法官表示的案件语义特征。ε表示构成法官表示的案件中有多少案件属于其审判质量较高的案件类别。当ε的值大于0.5时，构成法官表示的案件中超过50%的案件属于其审判质量较高的案件。笔者认为，ε值越接近1，法官表示的抽象语义越能体现法官擅长的审判领域。

2.2 基于CNN的案件与法官自动匹配方法

CNN是一类包含卷积计算且具有深度结构的前馈神经网络，其特有的卷积和池化结构能以较小的计算量提取有价值的特征。句子中具有丰富的语义信息，CNN可以利用多个不同尺寸的卷积核从不同角度获取句子丰富的语义特征。本文采用CNN处理案件和法官表示文本，获取案件和法官的抽象语义表示，用相似度函数计算向量相似度，自动评估法官和案件的匹配度。基于CNN的匹配模型结构如**图2**所示。

本文利用CNN获取案件和法官表示，借鉴人脸识别的思想，采用三元组损失来调节CNN的网络结构，使其更好地学习案件和法官的语义特征表示。人脸识别任务指输入一张人脸图像，在数据集中寻找同一个人的图像。通常的做法是构建三元组

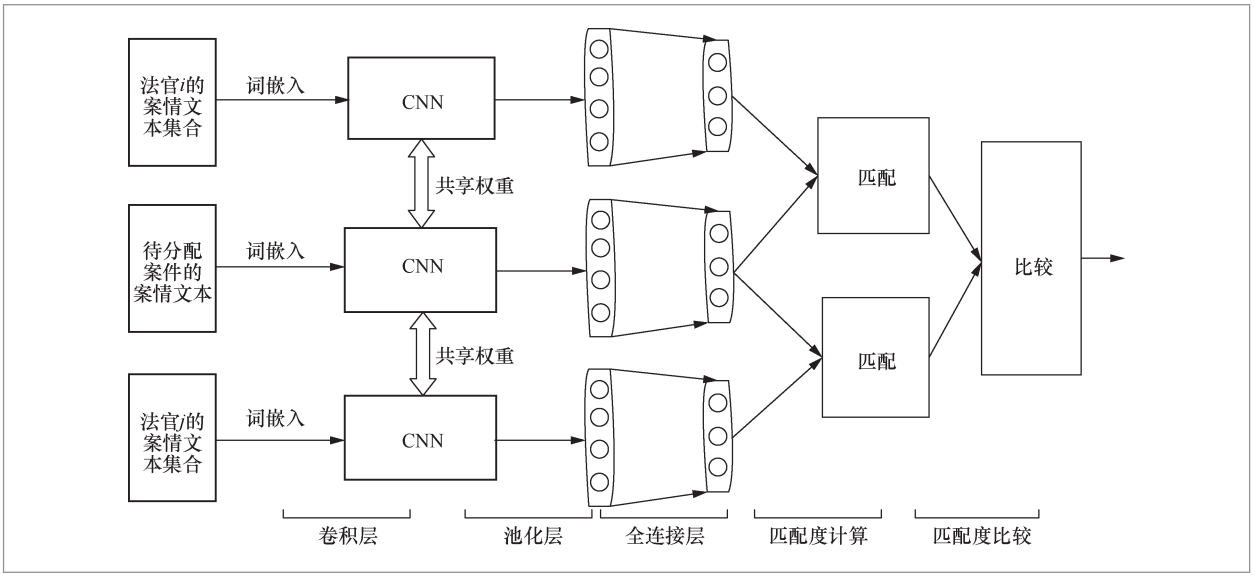

图 2 基于 CNN 的匹配模型结构

（a,p,n）。其中，a是基准正例，表示输入图像；p是正例，表示与输入图像中的人物是同一个人的图像；n是负例，表示与输入图像中的人物是不同人的图像。此类检索任务常用三元组损失调整网络参数，目标是使同类照片在编码空间中的距离尽量小，使不同类照片在编码空间中的距离尽量大。上述三元组损失的目标函数是：

$$L=\max(d(a,p)-d(a,n)+\text{margin},0) \quad (2)$$

其中，$d(a,p)$、$d(a,n)$分别表示a和p、a和n的向量的距离，margin是阈值参数。通过最小化L，$d(a,p)$趋于0，$d(a,n)$远大于$d(a,p)$与margin的和。

基于此思想，类比案件$\boldsymbol{x}_i^a$为基准正例，擅长审判此类案件的法官$\boldsymbol{x}_i^p$为正例，不擅长审判此类案件的法官$\boldsymbol{x}_i^n$为负例，通过CNN学习到案件表示和法官表示，构建三元组$(\boldsymbol{f}(\boldsymbol{x}_i^a),\boldsymbol{f}(\boldsymbol{x}_i^p),\boldsymbol{f}(\boldsymbol{x}_i^n))$，使$\boldsymbol{f}(\boldsymbol{x}_i^a)$与$\boldsymbol{f}(\boldsymbol{x}_i^p)$的匹配度大于$\boldsymbol{f}(\boldsymbol{x}_i^a)$与$\boldsymbol{f}(\boldsymbol{x}_i^n)$的匹配度，目标函数如下：

$$Y=\sum_{i=1}^{N}\left[\cos(\boldsymbol{f}(\boldsymbol{x}_i^a),\boldsymbol{f}(\boldsymbol{x}_i^n))-\cos(\boldsymbol{f}(\boldsymbol{x}_i^a),\boldsymbol{f}(\boldsymbol{x}_i^p))+\text{margin}\right] \quad (3)$$

其中，$\cos(\boldsymbol{x},\boldsymbol{y})$表示向量$\boldsymbol{x}$和$\boldsymbol{y}$的余弦相似度，用相似度表示$\boldsymbol{x}$和$\boldsymbol{y}$的匹配度；margin是阈值参数；$N$表示案件个数与法官个数的乘积。最小化损失$Y$，使$\cos(\boldsymbol{f}(\boldsymbol{x}_i^a),\boldsymbol{f}(\boldsymbol{x}_i^p))$的值远大于$\cos(\boldsymbol{f}(\boldsymbol{x}_i^a),\boldsymbol{f}(\boldsymbol{x}_i^n))$与margin的和。

将上述$\boldsymbol{x}$到$\boldsymbol{f}(\boldsymbol{x})$的映射变换过程进行形式化表示。令$L=l_1,l_2,l_3,\cdots$，表示案情事实描述，其中$l_i$表示文本中的第$i$个字。在嵌入层，基于预训练的中文维基百科字向量表$\boldsymbol{W}$，每一个$l_i$都被映射成一个向量。其中，$\boldsymbol{W}\in\boldsymbol{R}^{S\times K}$，$S$表示字典大小，$K$表示向量维度。假设输入模型的文本序列长度为$s$，经嵌入表示得到文本的向量序列为$\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_s]$，其中$\boldsymbol{x}_i\in\boldsymbol{R}^K$。该过程可表示为：

$$\boldsymbol{x}_i=\text{Embedding}(l_i) \quad (4)$$

然后将$\boldsymbol{x}$输入卷积层提取局部特征，卷积操作由滤波器完成，令滤波器尺寸为$\boldsymbol{W}_c\in\boldsymbol{R}^{h\times K}$，其中，$h$表示滤波器移动的窗口大小。该过程可表示为：

$$\boldsymbol{c}=f(\boldsymbol{W}_c*\boldsymbol{x}+\boldsymbol{b}) \quad (5)$$

其中，$\boldsymbol{b}\in\boldsymbol{R}$表示偏置量；$f$表示非线性函数，将卷积输出结果做一次非线性映射，

本文使用ReLU激活函数；*表示卷积。在文本操作中，常设置大小不同的多个窗口以获取不同粒度信息的特征向量，例如$\boldsymbol{h}$=(3,5,7)时，获得特征向量表示为$[\boldsymbol{c}_1,\boldsymbol{c}_2,\boldsymbol{c}_3]$。对每组特征向量进行池化操作以获取文本中更有价值的特征，本文采用最大池化操作。该过程形式化表示为：

$$f(\boldsymbol{x})=[\max(\boldsymbol{c}_1),\max(\boldsymbol{c}_2),\cdots,\max(\boldsymbol{c}_t)] \quad (6)$$

这里$f(\boldsymbol{x})$就是文本被CNN学习到的语义特征表示，其中，t表示卷积窗口的数量。

对于任何一个待分配案件$\boldsymbol{x}$，基于法官库，构建得到三元组$(\boldsymbol{x}^a,\boldsymbol{x}_i^p,\boldsymbol{x}_i^n)$数据，经过卷积匹配模型，得到$(\boldsymbol{f}(\boldsymbol{x}^a),\boldsymbol{f}(\boldsymbol{x}_i^p),\boldsymbol{f}(\boldsymbol{x}_i^n))$，计算$\cos(\boldsymbol{f}(\boldsymbol{x}^a),\boldsymbol{f}(\boldsymbol{x}_i^p))$与$\cos(\boldsymbol{f}(\boldsymbol{x}^a),\boldsymbol{f}(\boldsymbol{x}_i^n))$。若有$M$个法官就有$M$个相似度值，用相似度值表示匹配值，通过比较$M$个值的大小，输出前$N(N<M)$个匹配值较高的法官作为案件的推荐法官。

3 实验与结果

为了验证本文方法的高效性和有效性，在相同数据集下将本文分案方法和传统分案方法、传统机器学习算法和深度学习方法进行了实验对比分析。通过改变参数ε的值来改变法官表示的组成特征，分析ε对结果的影响。

3.1 实验数据

实验数据来源于贵州省某法院的买卖合同纠纷和民间借贷纠纷两类案件。本文抽取2016—2019年间买卖合同纠纷案件（共2 096个）和民间借贷纠纷案件（共3 110个）作为实验原始数据。数据质量在很大程度上会影响模型的训练效果，本文首先对5 206个案件卷宗数据进行预处理。首先，删除数据源中的传票、通知书等图片数据以及非判决书文本数据；其次，根据关键字正则匹配提取案件案情描述，删除无法有效提取案情的案件；最后，利用哈尔滨工业大学的语言技术平台（language technology platform，LTP）对案情要素进行分析处理，将“公诉机关指控”等词语以及人名、车牌号、电话号码、地名等归一化。经过清洗，数据中涉及法官18个，共有案件1 546个，其中民间借贷纠纷案件1 114个，买卖合同纠纷案件432个。每个案件只有一个审判法官。

本文的研究目标是将案件分配给更擅长审判此类案件的法官。抽取案情事实描述表示案件。为了保证分案结果具有实际意义以及防止实验出现过拟合问题。生成数据集时，首先将案件按8∶1∶1切分成训练集、验证集以及测试集。接着，在切分得到的训练集案件下，按照第2.1节的审判质量权重计算方法计算审判质量权重，通过比较两类案件的审判质量权重，得到法官擅长审判和不擅长审判的案件类型。通过设置参数的值，得到法官表示。基于此，利用不同部分的案件对应构建三元组数据数据集。在数据集中，保证每一个案件的审判法官都是擅长审判此类案件的法官。数据集情况见**表1**。

3.2 评价指标

本文的实验目标是给案件推荐N个法官。采用正确率ACC作为实验结果的评价指标。正确率的大小取决于推荐法官的个数，推荐法官个数越多，正确率越高。正确

表1 数据集情况

数据集类别	案件数/个	三元组数/个
训练集	1 238	21 046
验证集	152	2 584
测试集	156	2 652

率的计算方法如下：

$$ACC=\frac{count}{Z}\times 100\% \quad (7)$$

其中，Z表示测试集案件的个数，count表示在测试集中为每一个案件推荐的N个法官中包含该案件的原审法官的案件个数，这里Z恒等于156。在生成数据集时保证了原审法官一定擅长审判此案件。count的值由N的大小决定。N越大，推荐法官个数越多，那么推荐法官中包含原审法官的可能性就越大。

3.3 本文分案方法的实验结果

本文基于Triplet CNN在N为1、3、5、7时进行实验，实验结果见**表2**。

从**表2**可以看出，只推荐1个法官的正确率只有86.54%。随着推荐法官个数的增多，正确率逐渐增大，当推荐法官个数为7时，正确率已经高达98.72%。这证明了本文的法官推荐方法是高效并且高精准度的，融合审判质量的法官表示方法确实能很好地体现法官擅长领域信息。根据第2.1节和第2.2节可知，这样的结果是合理的。因为任何一个案件只有一个法官擅长，针对每个案件，只有一个法官的匹配度最高。当N逐渐增大时，推荐法官中包含原审法官的概率随之增大，正确率自然越高。

3.4 本文方法与传统分案方法的实验对比分析

目前我国法院系统中的分案方法主要为简单随机分案。在实践中，简单随机分案方法主要为摇号分案和均衡分案两种分案方法。摇号分案指将法院所有法官编号，每个法官的编号都是唯一的。法院接收到新的案件后，利用计算机程序随机产生一个号码，号码对应的法官就是程序为案件分配的法官。均衡分案是在摇号分案的基础上，增加保证每个法官在一段时间内的承办案件数量基本相等这一约束条件，即每次分案优先将案件分给现有案件承办数较少的法官。本文在同一数据集上对摇号分案和均衡分案方法进行了实验。本文方法与传统分案方法的实验结果对比见**表3**。

表 2 本文方法实验结果

N	正确率
1	86.54%
3	96.15%
5	98.07%
7	98.72%

实验时，在任何一个N值下，摇号分案和均衡分案都做100组实验，然后计算平均值得到该N值下的正确率。从**表3**可以看出，摇号分案在推荐Top1法官时，正确率只有5.40%。这是合理的。因为摇号分案是不考虑法官擅长领域和案件信息的完全随机分案，本文实验数据中有18个法官，每次随机分配正确的概率只有十八分之一，即正确率只有5.56%左右。由此可知，本文的实验结果是拟合于实际结果的。同样，均衡分案本质上也是随机分案，它的实验结果与摇号分案相差不大。但均衡分案增加了在一段时间内保证每个法官承办的案件数量基本相等这一约束条件，因此其正确率稍高一点。

从**表3**可以看出，本文方法的实验结果明显优于传统分案方法。在Top1时，本文方法推荐法官的精准度比摇号分案和均衡分案高80%以上。这一实验结果证明使用本文分案方法不仅可以实现案件的自动分配，还能显著提高案件分配的精准度，实现人案相适。

3.5 本文方法与机器学习算法的实验对比分析

为了进一步验证本文方法的有效性，

本文另选取TF-IDF以及BM25算法获取案件和法官特征向量，结合余弦相似度计算案件和法官的匹配度，并对匹配度进行排序，从而实现分案。本文在同一数据集上进行对应的实验，分案结果见**表4**。通过实验发现，本文方法的性能明显优于传统机器学习算法。BM25算法是改进的TF-IDF算法，TF-IDF是根据词在文本中的重要度来获取文本特征表示的，TF值在理论上可以无限大，但BM25算法在TF计算方法中增加了常量以限制TF值的增长极限，并且考虑了文档长度，因此BM25算法对文档的表征能力要优于TF-IDF算法。本文使用卷积神经网络从多角度捕捉文本特征，卷积神经网络考虑了词的上下文信息，对文本的表征能力明显优于传统机器学习算法。

3.6 本文方法与深度学习方法的实验对比分析

本组实验在同一数据集情况下，将本文方法与现有常用的深度学习方法进行比较，实验结果见**表5**。基准模型如下。

- Triplet Bi-LSTM（Tri-BiLSTM）：利用Bi-LSTM获取案件和法官的特征表示，设置最大序列长度为512。
- Triplet BERT（Tri-BERT）：利用BERT获取案件和法官的特征表示，设置最大序列长度为512。
- Triplet AlBERT（Tri-ALBERT）[24]：该模型是谷歌提出的基于BERT的改进模型，本文利用该模型获取案件和法官的特征表示，设置最大序列长度为512。

由**表5**可知，在Top1下本文方法的分案效果优于其他方法。BERT及ALBERT都是大规模语料的预训练模型，预想对文本的表征能力应该优于CNN。但通过对裁判文书数据的分析，案件的案情文本长度大多在1 500以上，个别案情文本长度甚至超过3 000，而BERT能接收的最大序列长度为512，因此用BERT模型获取案情文本特征表示时会丢失较多语义信息。并且根据裁判文书的书写规范，案情事实描述通常以“公诉机关指控”“某某地某某区检察院指控”“经审理查明”等固定短语开头，以“上述事实，有公诉机关当庭出示，并经庭审质证的被告人MM在公安机关的供述及户籍证明，xxx等证据证实，足以认定”等固定句式结尾。由此可以看出，案情的关键信息应集中在案情文本的中间部分，而不是案情描述的开头和结尾，而BERT的序列长度约定使其提取的关键信息受到限制，从而在分案效果上不如基于CNN的分案方法。另外，预想LSTM的表

表3 本文方法与传统分案方法的实验结果对比

分案方法	正确率			
	Top1	Top3	Top5	Top7
均衡分案	5.85%	18.72%	30.77%	43.07%
摇号分案	5.40%	18.02%	27.27%	35.85%
本文方法	86.54%	96.15%	98.07%	98.72%

表4 本文方法与机器学习算法的实验结果对比

分案方法	正确率			
	Top1	Top3	Top5	Top7
TF-IDF+cosine	1.28%	24.36%	57.05%	66.03%
BM25+cosine	16.67%	43.59%	66.67%	90.38%
本文方法	86.54%	96.15%	98.07%	98.72%

表5 本文方法与深度学习方法的实验结果对比

分案方法	正确率			
	Top1	Top3	Top5	Top7
Tri-BiLSTM	49.36%	75.64%	88.46%	96.15%
Tri-BERT	78.21%	91.67%	96.15%	97.44%
Tri-ALBERT	76.92%	94.23%	97.44%	98.72%
本文方法	86.54%	96.15%	98.07%	98.72%

现能力应该优于CNN。根据上述分析，虽然LSTM能捕捉序列的长距离依赖关系，但由于案情文本长度过长，LSTM的循环机制决定其对较长文本的特征提取更关注序列的末尾，而案情描述结尾的内容不能较好地体现案情信息，故本文方法是优于基于LSTM的分案方法的。

3.7 参数ε对实验结果的影响

从第2.1节可知，构成法官表示的案情特征是由参数ε决定的。随着ε的变化，构成法官表示的案情事实描述可能有部分或全部是法官审判质量较低的案件类型。基于本文模型，保证模型参数设置不变，本文在参数ε=1.0、0.5、0.1下分别进行了实验。当ε=1.0时，构成法官表示的案件都属于审判质量较高的案件类别。当ε=0.5时，构成法官表示的案件个数中有50%从其审判质量较高的案件类别中随机选择，另外50%从其案件审判质量较低的案件类别中随机选择。同样，当ε=0.1时，构成法官表示的案件个数中有10%从其审判质量较高的案件类别中随机选择，另外90%从其案件审判质量较低的案件类别中随机选择。不同参数值下本文方法的实验结果见**表6**。

从**表6**可以看出，在ε=1、0.5、0.1情况下，ε=1时实验效果最好。这一实验结果表明，用法官审判质量较高的案件来表示法官能更好地体现法官擅长的领域信息，用CNN提取法官表示特征，能实现更加精准的分案。随着ε的变化，构成法官表示的特征也发生变化。若法官表示中包含其审判质量较低的案件，会导致CNN提取的抽象语义特征向量不能很好地突出法官擅长的领域，无法更精准地匹配案件，最终导致正确率降低。

表 6 不同参数值下本文方法的实验结果

参数ε	正确率			
	Top1	Top3	Top5	Top7
0.1	79.49%	94.23%	97.44%	98.08%
0.5	82.69%	92.95%	97.44%	99.36%
1.0	86.54%	96.15%	98.07%	98.72%

4 结束语

笔者希望打破法院系统中传统人定分案的局面，解决人为干扰案件分配、人案不适、人情案等问题，探索一种以人案相适为目标的辅助分案方法。本文提出了一种融合案件审判质量的法官表示方法，利用法官审判质量较高的案件语义特征，综合反映法官擅长的审判领域，从而在法官表示中融合法官擅长领域的抽象语义信息。用案情事实描述表示案件。采用卷积神经网络学习案件表示和法官表示中不同粒度的抽象语义特征表征向量，计算案件和多个法官的表征向量间的余弦相似度，用向量相似度表示案件与法官的匹配度，输出前N个匹配值较高的法官作为案件的推荐法官。该方法可为案件推荐擅长审判此类型案件的法官，实现专案专办，形成专业化办案模式，避免关系案、金钱案等弊端，提高办案质效。本文分案方法避免了在分案过程中的人为因素干扰，保证了分案过程留痕可查，促进司法公开、公正。未来笔者将结合繁简分流思想进行分案，并拟融合推荐系统方法，以取得更好的分案效果。

参考文献:

[1] 秦永彬, 冯丽, 陈艳平, 等. “智慧法院”数据融合分析与集成应用[J]. 大数据, 2019, 5(3): 35-46. QIN Y B, FENG L, CHEN Y P, et al.

"Intelligent Court" data fusion analysis and integrated application[J]. Big Data Research, 2019, 5(3): 35–46.

[2] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513–523.

[3] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391–407.

[4] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2001, 3: 601–608.

[5] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003, 3: 1137–1155.

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, 2013, arXiv:1301.3781.

[7] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference in Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2013: 3111–3119.

[8] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1532–1543.

[9] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135–146.

[10] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv preprint, 2018, arXiv:1802.05365.

[11] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv:1810.04805.

[12] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv preprint, 2014, arXiv:1408.5882.

[13] 冯兴杰, 张乐, 曾云泽. 基于多注意力CNN的问题相似度计算模型[J]. 计算机工程, 2019, 45(9): 284–290.
FENG X J, ZHANG L, ZENG Y Z. Question similarity calculation model based on multi-attention CNN[J]. Computer Engineering, 2019, 45(9): 284–290.

[14] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357–370.

[15] 孙国荣. 智能分案: 兼顾公平与效率[J]. 人民法治, 2018(2): 49–51.
SUN G R. Smart division: taking into account fairness and efficiency[J]. People·Rule of Law, 2018(2): 49–51.

[16] 王崇璋. 法院案件分配标准体系实证研究: 以HB市中基层法院为例[D]. 淮北: 淮北师范大学, 2018.
WANG C Z. An empirical study on the standard system of court case allocation-take the example of the HB municipal intermediate and basic courts[D]. Huaibei: Huaibei Normal University, 2018.

[17] 王建华, 张宁. 美国联邦法院民事案件的立案规则[N]. 人民法院报, 2015-10-23(8).
WANG J H, ZHANG N. Rules for filing civil cases in the U.S. Federal Court[N]. People's Court Daily, 2015-10-23(8).

[18] 兰世民, 兰馨, 缪新森. 法院分案若干问题研究[J]. 法律适用, 2012(6): 95–100.
LAN S M, LAN X, MIAO X S. Research on several Issues of court division[J]. Journal of Law Application, 2012(6): 95–100.

[19] 贾炜, 常鹏, 石鹏. 一种实现法院分案的方法

及装置: CN107665212A[P]. 2018-02-06.
JIA W, CHANG P, SHI P. Method and device for case allocation of court: CN107665212A[P]. 2018-02-06.
[20] 顾钊铨, 方滨兴, 韩伟红, 等. 一种法院案件智能化分案辅助方法及系统: CN109872052A[P]. 2019-06-11.
GU Z Q, FANG B X, HAN W H, et al. Court case intelligent case division auxiliary method and system: CN109872052A[P]. 2019-06-11.
[21] 陈芳序. 法官工作量均等视角下法院随机分案系统的检视与重构: 以A法院刑事分案为切入点[J]. 海峡法学, 2019, 21(2): 87-95.
CHEN F X. Inspection and reconstruction of the court random division system from perspective of equal workload for judges-based on criminal division of court A[J]. Cross-Strait Legal Science, 2019, 21(2): 87-95.
[22] 王小新. 法院分案系统的检视与重构: 以X法院刑事案件分配为例[J]. 法律适用, 2016(4): 112-116.
WANG X X. Review and reconstruction of the court division system: taking X court criminal case distribution as an example[J]. Journal of Law Application, 2016(4): 112-116.
[23] 孙晓东. 中国司法评估制度完善研究[J]. 广东社会科学, 2018(6): 231-242.
SUN X D. An analysis on improvement of judicial performance evaluation system of China[J]. Social Sciences in Guangdong, 2018(6): 231-242.
[24] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[J]. arXiv preprint, 2019, arXiv:1909.11942.

作者简介

敖绍林(1995-), 男, 贵州大学计算机科学与技术学院硕士生, 主要研究方向为自然语言处理、文本分析。

秦永彬(1980-), 男, 博士, 贵州大学计算机科学与技术学院教授, 主要研究方向为大数据处理、云计算、文本挖掘。

黄瑞章(1979-), 女, 博士, 贵州大学计算机科学与技术学院副教授, 主要研究方向为信息检索、文本挖掘。

陈艳平（1980- ），男，博士，贵州大学计算机科学与技术学院副教授，主要研究方向为人工智能、自然语言处理。

刘丽娟（1980- ），女，贵州师范学院讲师，主要研究方向为法学与思想政治教育。

郑庆华（1969- ），男，博士，西安交通大学计算机科学与技术学院教授，主要研究方向为多媒体远程教育、计算机网络安全。

陈昌恒（1978- ），男，贵州省高级人民法院信息技术处处长、三级调研员，主要研究方向为司法审判应用。

程少芬（1982- ），女，贵州省高级人民法院信息技术处应用推广科科长、四级调研员，主要研究方向为司法审判应用。

收稿日期： 2021-01-18

通信作者： 秦永彬，ybqin@foxmail.com

基金项目： 国家自然科学基金资助项目（No.U1836205，No.91746116，No.62066007，No.62066008，No. 62166007）；贵州省科技重大专项计划（No.[2017]3002）；贵州省科学技术基金重点项目（No.[2020]1Z055）；贵州省研究生科研基金立项课题（No.YJSCXJH[2019]102）

Foundation Items: The National Natural Science Foundation of China (No.U1836205, No.91746116, No.62066007, No.62066008, No. 62166007), The Major Special Science and Technology Projects of Guizhou Province (No.[2017]3002), The Key Projects of Science and Technology of Guizhou Province (No.[2020]1Z055), Project of Guizhou Province Graduate Research Fund (No.YJSCXJH[2019]102)