计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(2):63−69 [doi: 10.15888/j.cnki.csa.007639]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的“一人多案”风险预警系统①

马志柔[1], 马新宇[1,2], 刘　杰[1], 叶　丹[1]

[1](中国科学院 软件研究所 软件工程技术研究开发中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
通讯作者: 马志柔, E-mail: mazhirou@otcaix.iscas.ac.cn

摘　要: 针对在法院立案-审判-执行全流程阶段, 多起案件中存在的当事人或者案件事实相同的情况, 即“一人多案”, 造成了司法资源浪费与不合理使用, 设计实现了基于深度学习的“一人多案”风险预警系统. 该系统基于深度学习技术和海量裁判文书数据, 通过对案件文本的向量表示建模, 提出了面向法律文书的案由识别和相似度量方法, 结合法律业务规则进行“一人多案”关联识别, 并给出风险预警报告. 该系统能够为司法资源统筹提供技术支持, 为法院公正、高效地审理案件提供保障.

关键词: 深度学习; 案由识别; 相似度量; 一人多案; 风险预警

引用格式: 马志柔,马新宇,刘杰,叶丹.基于深度学习的“一人多案”风险预警系统.计算机系统应用,2021,30(2):63–69. http://www.c-s-a.org.cn/1003-3254/7639.html

Risk Pre-Warning System of “One Person with Multiple Cases” Based on Deep Learning

MA Zhi-Rou[1], MA Xin-Yu[1,2], LIU Jie[1], YE Dan[1]

[1](Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: During the court’s case-trial-enforcement process, the parties involved in multiple cases or the facts of the case are sometimes the same, namely “one person with multiple cases”, resulting in the waste and unreasonable allocation of judicial resources. Then a risk pre-warning system of one person with multiple cases based on deep learning is designed and implemented. This system is based on deep learning technology and massive judgment documents. Case identification and similarity measure for legal documents are proposed by modeling the vector representation of the case text, and one person with multiple cases is associated with the legal business rules, providing the risk pre-warning report. This system can offer technical support for judicial resource coordination, supporting courts to try cases fairly and efficiently.

Key words: deep learning; case identification; similarity measure; one person with multiple cases; risk pre-warning

随着知识经济的迅猛发展和民主法制建设的不断完善, 人民法院维护社会稳定的职能和任务不断增多和加重, 各级法院在完成繁重审判任务的同时, 还需更好地解决审判质量问题. 在法院立案-审判-执行全流程阶段, 多起案件中存在当事人或者案件事实相同的情况, 即“一人多案”的处理情况, 比如一人起诉多案或被诉多案、相同当事人之间的多起类似案件等[1]. 对此类案件合并处理有利于此类案件的快速解决, 不仅能够有效提升司法机关的工作效率、减少司法资源的无端消耗、优化社会资源的合理使用, 而且在面临问题时可以从相同的案件类型中获取一个可靠的参照. 同样也有利于当事人相关问题的解决, 从社会与公民个人

① 基金项目: 国家重点研发计划 (2018YFC0831302)
Foundation item: National Key Research and Development Program of China (2018YFC0831302)
收稿时间: 2020-03-10; 修改时间: 2020-04-10; 采用时间: 2020-04-21; csa 在线出版时间: 2021-01-27

两个层面实现经济利益的最大化.

为了提高司法人员在案件处理环节的效率, 本文提出了一种基于深度学习的“一人多案”智能风险预警系统, 利用深度学习与自然语言处理技术对法律文书进行案由识别预测和相似度量匹配, 并与法律业务规则相结合, 解决当前信息化系统无法有效分辨“一人多案”的技术难题, 实现法院立案-审判-执行全流程阶段的“一人多案”的关联识别, 为跨区域跨层级的司法资源统筹提供技术支持, 为法院公正、高效地审理和执行案件提供保障.

1 相关工作

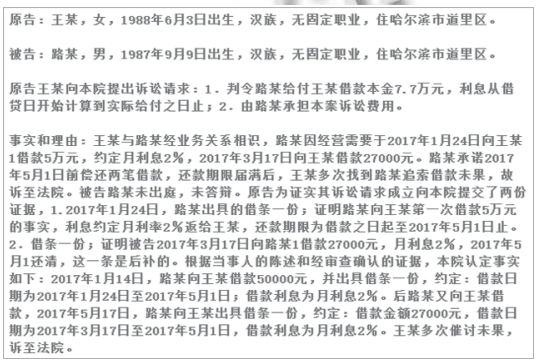
我国法院信息化水平在世界处于领先水平, 近年来我国法院信息化领域建成了全面覆盖各级人民法院和法庭的网络设施、业务应用、数据管理和安全保障体系, 极大地提升了审判执行、司法为民和司法管理质效. 但在法院内部协同智能水平相对薄弱, 尤其诉讼服务中“一人多案”方面的信息融合共享和服务高效协同需要亟待提升. 从法学的角度, “一人多案”的理论日趋完备, 依据诉讼法对重复起诉识别的不同学说进行分析评述, 给出了民事重复起诉的识别要素、判别规则及处置方法的法律释明[1,2]. 重复起诉的判断要素包括当事人、案由和诉讼请求, 关键要判断是否是同一当事人基于同一法律关系、同一法律事实提出的同一诉讼请求[3].

当前深度学习技术日趋成熟, 其端到端的学习避免了繁重的特征工程和自然语言处理工具带来的错误传播问题, 在文本处理任务中取得了显著的成功, 达到了远超传统方法的性能[4,5]. 在文本特征表示方面, Mikolov 等提出了通过神经网络训练词向量的方法 Word2Vec[6]; 之后 Joulin 等基于词向量提出了一种高效的文本分类和表征学习的方法 fastText[7], 使用 *n*-gram 模型可以更有效的表达词前后的之间的关系; 而 BERT[8] 预训练模型的提出将文本特征表示推向顶峰. 在文本分类匹配方面, Kim 提出了 TextCNN 方法[9] 将卷积神经网络应用于文本分类任务, 该网络通过一维卷积核捕获句子中类似 *n*-gram 的关键信息; Liu 等的工作提出了将 RNN 用于分类问题的网络设计[10], 考虑文本的时序特征; 之后涌现一些网络变体 LSTM、RCNN, 以及引入 attention[11] 机制的网络模型.

近些年来, 随着以裁判文书为代表的司法大数据不断公开, 以及自然语言处理技术的不断突破, 如何将人工智能技术应用在司法领域来提高司法人员在案件处理环节的效率逐渐成为法律科技研究的热点, 一些学者已经在研究与深度学习相关的法律文本处理技术. Luo 等[12] 提出了一种基于注意力机制的神经网络方法, 在罪名预测任务中融入法条信息, 使罪名预测更具有合理性, 有助于提高法律助理系统效率. Hu 等[13] 提出一个属性-注意力罪名预测模型, 根据法律属性把罪名分类, 通过人工将相关罪名属性信息进行标记, 显著提升低频罪名与易混淆罪名的预测精度. Zhong 等[14] 利用有向无环拓扑图来建模多任务之间的逻辑依赖关系, 将法条、罪名与刑期之间的依赖关系融合到统一的司法判决框架中, 所有任务的效果在多个数据集上取得了一致和显著的提升.

2 关键技术研究

在“一人多案”的关联识别时, 其关键技术是如何判别两个案件是否同一个案由和两个案件的诉讼请求是否相似, 以起诉状文本语义理解为核心, 利用自然语言处理技术与机器学习方法实现对起诉书中的当事人、案由、诉讼请求等关键信息进行智能识别与语义理解. 通常起诉状文本的内容格式如图 1 所示, 包含原告、被告、诉讼请求、事实与理由 4 部分信息.


原告：王某，女，1988年6月3日出生，汉族，无固定职业，住哈尔滨市道里区。

被告：路某，男，1987年9月9日出生，汉族，无固定职业，住哈尔滨市道里区。

原告王某向本院提出诉讼请求：1．判令路某给付王某借款本金7.7万元，利息从借贷日开始计算到实际给付之日止；2．由路某承担本案诉讼费用。

事实和理由：王某与路某经业务关系相识，路某因经营需要于2017年1月24日向王某1借款5万元，约定月利息2%，2017年3月17日向王某借款27000元。路某承诺2017年5月1日前偿还两笔借款，还款期限届满后，王某多次找到路某追索借款未果，故诉至法院。被告路某未出庭，未答辩。原告为证实其诉讼请求成立向本院提交了两份证据，1. 2017年1月24日，路某出具的借条一份；证明路某向王某第一次借款5万元的事实，利息约定月利率2%返给王某，还款期限为借款之日起至2017年5月1日止。2．借条一份；证明被告2017年3月17日向路某1借款27000元，月利息2%，2017年5月1还清，这一条是后补的。根据当事人的陈述和经审查确认的证据，本院认定事实如下：2017年1月14日，路某向王某借款50000元，并出具借条一份，约定：借款日期为2017年1月24日至2017年5月1日；借款利息为月利息2%。后路某又向王某借款，2017年5月17日，路某向王某出具借条一份，约定：借款金额27000元，借款日期为2017年3月17日至2017年5月1日，借款利息为月利息2%。王某多次催讨未果，诉至法院。

图 1 起诉状文本内容

2.1 文本案由识别算法

案由识别是指通过诉讼文书中的事实与理由描述文本来认定民事起诉状的法律纠纷, 比如民间借贷纠纷、房屋买卖合同纠纷、物业服务合同纠纷等. 将此类问题看作是分类问题, 深度学习方法自动抽取文本特征, 可端到端地解决文本分类问题. 但在案由识别任务中, 诉讼事实和理由的文本长度在 400~600 字左右,

面临文本过长难以分析的问题. 单一的网络结构造成了语义匹配的不完善, CNN 网络无法很好获取问题的全局信息, 而 RNN 网络存在无法并行和梯度消失的问题, 训练速度不佳. 本系统提出一种多粒度融合模型, 能够结合 CNN 和 RNN 各自的优势, 利用 CNN 处理语法层面的局部匹配信息, 抽取不同位置上的特征, 处理与空间相关的数据; 利用 RNN 对句子整体进行编码, 提取到语义层面的匹配信息, 处理语句中的前后序列信息. 该模型结构如图 2 所示, 模型中 CNN 采用 TextCNN 来实现, RNN 采用 Bi-LSTM 来实现.

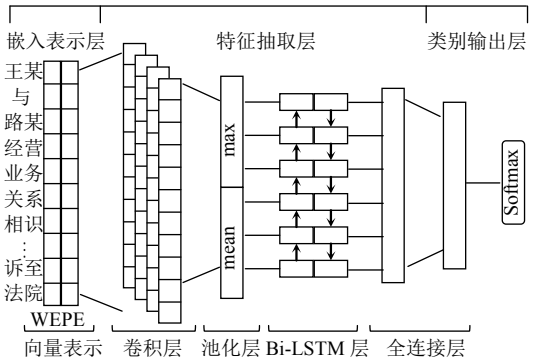


图 2 案由识别算法模型结构图

(1) 嵌入表示层, 对法律文本进行词级别的特征表示. 将法律文本中的每一个词表示为一个固定维度的向量, 其由词向量 (Word Embedding, WE) 和位置向量 (Position Embedding, PE) 的拼接而成. 其中, 词向量表示词汇的语义特征, 位置向量表示词在文本中的位置信息与相对距离特征. 所用的词向量由提前使用 Word2Vec 在中文维基百科语料预训练得到, 训练过程中词向量并不是固定的, 会随着模型的训练而更新. 所用的位置向量由不同维度上使用不同的模型函数学习得到, 这样高维的表示空间才有意义. 在偶数位置使用式 (1) 计算位置编码, 在奇数位置使用式 (2) 计算位置编码. 位置向量为随机初始化, 并通过模型训练得到最终的参数值.

$$PE_{2i}(pos)=\sin\left(pos/10\,000^{2i/dim}\right) \tag{1}$$

$$PE_{2i+1}(pos)=\cos\left(pos/10\,000^{2i/dim}\right) \tag{2}$$

(2) 特征抽取层, 对法律文本进行文档级别的特征表示. 通过特征抽取器提取文档的特征, 包括句子结构、句子语义、上下句子关系等. 该部分由两个子层组成, 第一子层是 TextCNN 网络层, 第二子层是 Bi-LSTM 网络层. TextCNN 网络层为句子特征抽取器, 卷积核大小决定了网络能够提取的局部特征范围, 大尺寸的卷积核可以提取较长距离的特征, 小尺寸的卷积核可以提取细粒度的特征, 使用多个大小不同的卷积核, 可抽取更长更复杂的句子特征. 同样, 卷积核数量的提升可以使网络从多角度进行特征提取, 但是计算量会随之上升, 网络复杂度增加, 容易导致过拟合. 为了能够从多角度、多范围内提取文本中包含的特征又不过分增加计算量, 保证泛化能力, 本文选用了大小分别为{2, 3, 5, 7}的卷积核各 128 个. 同时在每层的卷积后面加入了批量标准化层 (batch normalization) 和线性整流激活函数 (ReLU), 避免了梯度消失问题, 加速模型训练的收敛速度与稳定性. 为了从文本序列中得到的句子表示, 对每个卷积核的输出使用了 max-mean 池化, 即将最大池化 (max pooling) 与平均池化 (mean pooling) 的结果拼接. 其中, 最大池化得到的句子表示包含了当前文本序列的最大贡献, 平均池化得到的句子表示包含了整个文本中每个词的贡献. 将所有卷积核的结果拼接得到法律文本中句子的表示向量. Bi-LSTM 网络层为文档特征抽取器, 将得到的各个句子向量作为输入. 该层由两个 LSTM 网络组合而成, 一个向前传播、一个向后传播, 可以有效地利用文本上下文语义信息, 学习到句子之间的时序特征, 从而得到法律文本的文档级别特征表示向量.

(3) 类别输出层, 通过法律文本的表示向量分类得到每个法律纠纷的概率大小. 模型的输出层由双层的全连接网络、Dropout 和 Softmax 组成. 其中双层的网络结构, 提高了本层网络的非线性表达能力, 并将结果映射到每个相应的类别. Dropout 有效缓解了网络的过拟合问题, Softmax 归一化得到每个法律纠纷的概率.

2.2 文本相似度量算法

案件的特征量很多, 很难通过具体的规则来判断两个案件是否相同或相似, 需要研究案件相似度度量, 由于近几年深度学习在众多领域获得了突破性的成果, 已经有许多将深度学习和度量学习算法结合的尝试, 并且在许多数据集上得到了先进的结果. 本系统中提出了一种基于深度度量学习的案件相似匹配算法, 将有监督的距离度量学习的优化目标与深度学习强大的特征表示能力结合, 从而更加准确且符合法律语义地刻画案件之间的相似性.

为了方便后续的描述, 这里首先给出案件类型相关的定义:

定义 1. 同构案件: 记为D_i^+, 是与案件 D_i 具有完全

相同语义表述的案件样本.

定义 2. 异构案件: 记为D_i^-, 是与案件 D_i 具有完全不同语义表述的案件样本.

本系统提出的算法相比于传统度量学习更好地进行案件特征与算法模型的结合, 更加适用于法律文本语义度量匹配的场景使用, 如图 3 所示其算法框架由 3 部分组成.

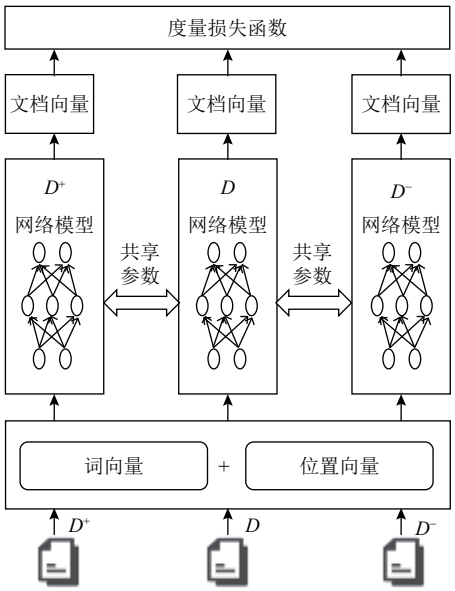


图 3 案件相似度算法框架图

(1) 输入层, 首先将案件文本分词, 然后计算每个词的词向量与位置向量, 最后将词向量和位置向量拼接得到案件文本的分布式向量化表示. 这里的案件文本的嵌入式表示方法可参见 2.1 小节文本案由识别算法中的法律文本特征表示描述. 每次输入一个案件文本三元组$\{D_i、D_i^+、D_i^-\}$, 案件文本之间的相似性满足以下公式:

$$sim\left(D_i,D_i^+\right)>sim\left(D_i,D_i^-\right) \tag{3}$$

(2) 表示层, 对应案件文本三元组输入设置 3 个网络, 即图 3 中的 D 网络、D^+网络、D^-网络, 网络之间共享参数, 该层的网络作为非线性变换表示函数将案件的原始特征转换为分布式表示 (embedding) 特征, 即案件文本的特征抽取器. 在案件相似匹配任务中, 诉讼请求文本长度在 50~100 字之间, 上下文之间逻辑性较强. 为了高效学习文本表示, 需要对特征抽取器进行仔细选择, 常用的特征抽取器为 LSTM 或 CNN. LSTM 虽然擅长对于序列建模, 然而由于其本身的序列依赖结构导致很难进行并行计算, 运算效率低; CNN 的卷积核滑动窗口位置之间则没有依赖关系, 可以并行计算, 故其运算效率高, 但其缺点在于难以捕获长距离特征, 大小为 k 的卷积核只能覆盖 k-gram 片段的特征. Transformer 的核心思想是基于自注意力机制, 不存在序列依赖问题, 能够通过并行计算提高运算效率. 在注意力计算中, 通过每个词与其他词的交互解决了长距离依赖特征获取问题; 根据不同词之间的相似度计算得到权重的方法, 使得模型能够捕获特征内部的相关性, 结合多头机制, 从不同角度捕获特征, 增强了语义特征提取能力. 故该层子网络使用 Transformer 网络作为特征提取器代替 LSTM 和 CNN 的编码方式, 既能对句子整体进行编码, 提取到语义层面的匹配信息; 又能提取语法层面的局部匹配信息. 针对法律诉讼请求文本设计一种基于多注意机制的网络结构, 该网络由两个子层组成, 第一子层是 multi-head 的自注意力结构, 第二子层是 position-wise 的全连接前馈网络. Multi-head 的自注意力结构从不同视角匹配计算案件序列各位置的权重, 其中加法方法 (additive attention) 考虑了位置的匹配程度, 乘法方法 (multiplicative attention) 能够捕捉文本摘要信息, 序列注意力方法 (sequential attention) 考虑了位置上下文的信息. 对文本序列进行多个不同的线性变换, 然后通过自注意力机制学习不同子空间下文本的表示, 最后将多个文本表示向量拼接起来作为输出. Position-wise 的全连接前馈网络由两个线性变换组成, 并且线性变换在不同位置上参数相同, 类似于卷积核为 1 的两层 CNN 网络.

(3) 度量层, 在 embedding 空间上对案件向量计算距离来刻画相似度, 使用 triplet loss[15] 作为整体框架的优化目标, 最终通过该层得到案件特征在 embedding 空间上的表示, 从而在诉讼请求的度量中得到应用.

采用曼哈顿距离度量两个案件之间的距离, 即在欧几里德空间的固定直角坐标系上两点所形成的线段对轴产生的投影的距离总和, 其计算公式如下:

$$d(x,y)=\sqrt{\sum_{k=1}^{n}|x_k-y_k|} \tag{4}$$

其中, x, y 表示两个不同案件的文档向量, n 表示文档向量的维度, x_k, y_k 表示文档向量的第 k 个元素.

采用式 (5) 作为 triplet loss 损失函数, 训练的目标是让相似案件在新的编码空间里的距离尽可能小, 让不相似案件在新的编码空间里的距离尽可能大, 即$d(D_i,D_i^-)$大于$d(D_i,D_i^+)+margin$, 其中$d(x,y)$表示两个案件之间的距离, $margin$ 为阈值. 在训练过程中对于某

一个案件, 将同构邻居拉近, 将异构邻居推远, 从而学习出一个间隔.

$$Loss = \max\left(d(D_i, D_i^+) - d\left(D_i, D_i^-\right) + margin, 0\right) \quad (5)$$

3 系统设计与实现

为了解决“一人多案”的处理问题, 本文设计一套基于深度学习的“一人多案”风险预警系统. 本系统实现采用的程序开发语言为 Python、深度学习框架为 Tensorflow、Web 服务框架为 Flask. 整个系统的组织架构如图 4 所示, 分为线下训练模块和线上预警模块两部分.

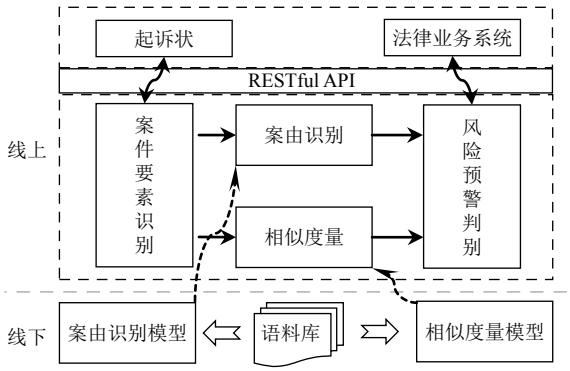


图 4 “一人多案”风险预警系统架构图

为了不破坏原有业务系统的结构, 系统通过 RESTful API 的方式提供用户与现有业务系统的数据交互与分析功能, 包括立案阶段风险判别、审理阶段风险判别、执行阶段风险判别、按当事人查询案件、按律师查询案件、按财产查询案件等. 也提供若干 API 供系统二次开发使用, 包括文本案由识别、文本相似度量等.

3.1 线下训练模块

该模块通过构造法律文本标注语料, 利用 2.1 节文本案由识别算法和 2.2 节文本相似度量算法, 训练得到案由识别模型和相似度量模型, 为线上预警模块提供算法模型支持.

3.1.1 案由识别模型

数据语料库构建: 从中国裁判文书网抓取民事案件文书 10 万篇, 涉及判决书、裁定书、决定书、调解书 4 类文书, 其中裁定书、决定书、调解书这 3 类文书中没有案情描述, 无法预测文书案由, 不能作为训练模型的数据集. 经筛选后, 取 83 979 篇民事判决书作为案由识别模型的数据语料库, 按照 18: 1: 1 的比例划分训练集、验证集和测试集, 总共涉及 78 类案由, 涵盖了常用的民事纠纷.

模型验证与分析: 采用准确率来评价, 取 top1、top3、top5 三种情况计算, 分别代表分类概率最高的前 1 名、前 3 名、前 5 名中类别包含正确的类别, 其在测试集上的准确率分别为 97.18%、99.45%、99.59%. 经验证分析, 民事案由识别效果满足需求. 而预测错误的纠纷类型有两种, 一种是样本数量太少 (没超过 5 篇), 一种是语义混淆 (比如侵害发明专利权纠纷与侵害外观设计专利权纠纷), 可以通过增加语料改进.

3.1.2 相似度量模型

数据语料库构建: 从上文中的民事判决书数据集中提取消费者权益保护纠纷类文书 2000 篇, 两两对比文书诉讼请求描述的语义, 构造三元组对数据集 8000 个, 按照 8: 1: 1 的比例划分训练集、验证集和测试集. 其中每份数据由 3 篇法律文书组成, 以三元组集合形式存储, 对于每份数据用 (d_0, d_1, d_2) 来代表该组数据, 约定文书 d_0 和文书 d_1 的相似度比文书 d_0 和文书 d_2 的相似度高, 即 $sim(d_0, d_1) > sim(d_0, d_2)$, 不符合的需要调整 d_1 和 d_2 的顺序.

模型训练及验证: 采用准确率来衡量模型的好坏. 对于测试数据集, 打乱 (d_0, d_1, d_2) 的顺序, 不再保证 $sim(d_0, d_1) > sim(d_0, d_2)$. 模型需要预测最终的结果是 $sim(d_0, d_1) > sim(d_0, d_2)$ 还是 $sim(d_0, d_1) < sim(d_0, d_2)$. 如果预测正确, 那么该测试点可以得到 1 分, 否则是 0 分. 实验对比了传统度量学习方法与深度度量学习方法, 在传统度量学习方法中用 TF-IDF 算法的准确率为 53.76%; 而在深度度量学习方法中的准确率为 70.76%. 经对比分析, 该模型方法比传统方法提高了将近 17 个百分点, 可以对法律文本进行细粒度的相似度量, 满足诉讼请求相似判断的需求.

3.2 线上预警模块

该模块实现了“一人多案”的关联识别和风险预警, 输入一个起诉状文本, 首先通过案件要素识别模型得到案件要素信息, 然后利用案由识别模型和相似度量模型对案件要素信息进行相似预测, 最后到风险预警判别模型中判断该案件是否属于“一人多案”, 并给出风险预警报告和协同处置方案. 该系统主要包括案件要素识别模块、案由识别模块、相似度量模块和风险预警模块, 其中案由识别模块和相似度量模块见上文, 这里不再赘述.

3.2.1 案件要素识别模块

案件要素识别模块是该系统的基础模块, 主要是对起诉状文本进行分析, 利用机器学习和自然语言处

理技术得到案件要素信息, 包括当事人信息、诉讼请求、事实与理由. 其过程分两步:

⑴ 基本信息识别. 起诉状文本的内容格式如图 1 所示, 其文字描述带有一定的格式. 通过“原告”、“被告”、“诉讼请求”、“起诉请求”、“事实”、“理由”等关键词, 将文本拆分为当事人文本信息、诉讼请求文本信息、事实与理由文本信息 3 部分.

⑵ 当事人信息识别. 当事人信息包括原告和被告, 其有可能是自然人、也可能是企业. 文本格式如下“某某, 男, xxxx 年 xx 月 xx 日出生, 某族, 住 xxx 省 xxx 市”或“被告: 某某有限公司, 住 xxx 省 xxx 市”. 利用正则表达式建立模式匹配, 从中提取当事人的人名、地名、机构名; 以及自然人的性别、民族、出生日期.

3.2.2 风险预警判别模块

风险预警判别模块是该系统的核心模块, 主要是将当事人信息构造成查询语句, 从法律业务系统中得到候选案件集合进行判别, 给出具有“一人多案”风险的处置建议. 其过程分两步:

(1) 获取候选案件集合. 调用待关联的法律业务系统 API 查询构建候选案件集合, 比如法院立案系统、执行办案系统等. 查询语句由当事人信息 (自然人、法人、其他组织) 的姓名、性别、住址以及企业名称等构成, 执行查询语句从系统中检索出原告和被告符合当事人信息条件的案件, 案件文本包含案号、当事人、代理律师、审理法院、案由、诉讼请求等文本信息, 形成候选案件集合以待进一步分析.

(2) “一人多案”判别分析. 调用案由识别模型和相似度量模型进行“一人多案”判别, 并给出风险预警报告.

为了方便后续的描述, 这里给出“一人多案”相关概念及判定规则:

定义 3. “一人多案”: “一人多案”情况主要是指重复立案, 重复立案与重复起诉有关, 重复起诉是指当事人就已经提起诉讼的事项在诉讼过程中或者裁判生效后再次起诉. 特别针对相同当事人、同一诉案由、同一法律关系以及主要诉讼请求相同. 按照当事人之间的纠纷类型不同, “一人多案”的判定规则可分为 3 种情况, 见表 1.

表 1 “一人多案”判定规则

规则名称		判定要素	处置建议
规则1	重复起诉	当事人相同、案由相同、诉讼请求相同, 即同一纠纷判断.	不予受理
规则2	一人起诉多案或被诉多案	案由相同、原告相同或被告相同, 即一方当事人相同纠纷.	预警、建议合并审理
规则3	串案	原告相同、被告相同、案由相同或相关, 即相同当事人之间的多起纠纷.	预警、建议合并审理

根据“一人多案”的判定规则, 将新起诉状和候选案件集进行各个案件要素的相似判定, 返回该案件是否存在“一人多案”, 具体流程如图 5 所示.

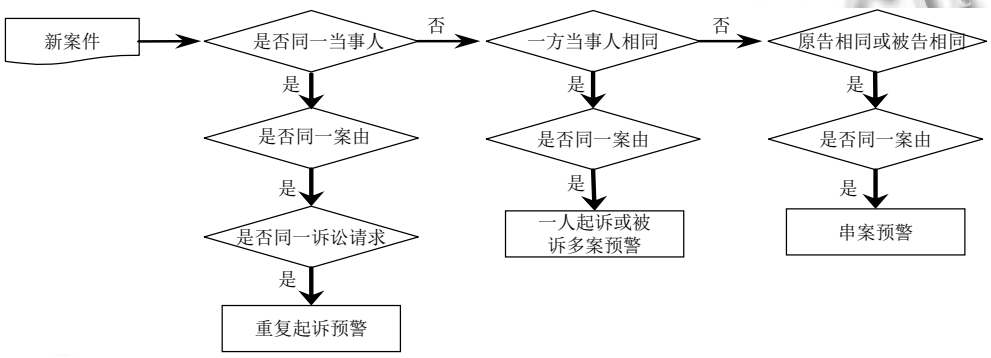


图 5 “一人多案”判定流程图

(1) 判断当事人是否相同: 如果存在身份证号或统一社会信用代码, 则可以直接判断是否同一人. 如果没有身份证号或统一社会信用代码, 则先根据姓名、性别、出生日期、企业名称等结构化信息进行判断是否是同一个人; 然后将地址信息拆分成省、市、区县、乡镇、村 5 级行政区划, 作为辅助信息进一步排查同名同姓的人, 如果两个地址在同一区县或经纬度距离小于 25 公里, 则可以认为是同一个人.

(2) 判断案由是否相同: 将获取到的新起诉状中的事实和理由文本输入到文本案由识别算法模型中, 得到一个案由, 并和候选案件集合中的案件案由对比, 判断是否有相同案由的案件. 根据案由的级别, 案由相同又可分为案由强相同和案由弱相同. 比如“人格权纠纷”的子案由包含“姓名权纠纷”、“肖像权纠纷”、“名

誉权纠纷”等. 如果两个案由同为“姓名权纠纷”, 则属于案由强相同; 如果一个案由为“肖像权纠纷”, 一个案由为“名誉权纠纷”, 两个案由同属“人格权纠纷”的子案由, 则属于案由弱相同.

(3) 判断诉求请求是否相同: 调用案件相似度量模块得到新起诉状与候选案件的诉讼请求的特征向量表示, 通过计算曼哈顿距离来判定两者之间是否相似. 这里不仅要判断一个案件与其他案件是否相似, 还要计算一个案件与其他案件的相似度值是多少, 能够按照相似度值大小排序, 并设定阈值筛选出相似案件.

(4) 建立“一人多案”关联预警: 通过对新起诉状和候选案件集之间的案件要素进行相似认定, 判定是否同一当事人认定、是否同一案由认定、是否同一诉讼请求认定, 利用“一人多案”判定规则建立以当事人为中心的案件之间的关联, 根据要素相似性的高低, 设立高、中、低不同级别的风险等级, 针对不同情况给出不同的风险提示和处置建议.

4 结束语

本文介绍了基于深度学习的“一人多案”风险预警系统的设计与实现. 该系统充分利用现有的司法大数据资源, 采用深度学习和自然语言处理技术对裁判文书文本挖掘分析, 设计了文本案由识别和文本相似度量算法, 解决了法律长文本的分类和细粒度度量问题, 实现了“一人多案”的关联识别和风险预警, 帮助法院合理分配案件审理、法官识别立案风险, 具有重要的应用价值. 在接下来的工作中, 将研究如何运用法律知识, 设计深度学习与知识图谱相结合的方法, 对法律文本进行深入挖掘分析. 此外, 文本预训练模型的司法应用也是一个有价值的研究方向.

参考文献

1 夏璇. 论民事重复起诉的识别及规制——对《关于适用〈中华人民共和国民事诉讼法〉的解释》第 247 条的解析. 法律科学 (西北政法大学学报), 2016, 34(2): 167–174. [doi: 10.16290/j.cnki.1674-5205.2016.02.017]

2 张卫平. 重复诉讼规制研究: 兼论“一事不再理”. 中国法学, 2015, (2): 43–65. [doi: 10.14111/j.cnki.zgfx.2015.02.004]

3 李静. 不当得利纠纷与合同纠纷的重复起诉规制. 中国社会科学院研究生院学报, 2019, (5): 77–88.

4 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]

5 Young T, Hazarika D, Poria S, *et al*. Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine, 2018, 13(3): 55–75. [doi: 10.1109/MCI.2018.2840738]

6 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 3111–3119.

7 Joulin A, Grave E, Bojanowski P, *et al*. Bag of tricks for efficient text classification. Proceedings of the 15th Conference of the European Chapter of The Association for Computational Linguistics. Valencia, Spain. 2017. 427–431. [doi: 10.18653/v1/e17-2068]

8 Devlin J, Chang M W, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, MN, USA. 2019. 4171–4186.

9 Kim Y. Convolutional neural networks for sentence classification. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1746–1751. [doi: 10.3115/v1/D14-1181]

10 Liu PF, Qiu XP, Huang XJ, *et al*. Recurrent neural network for text classification with multi-task learning. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, NY, USA. 2016. 2873–2879.

11 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA. 2017. 5998–6008.

12 Luo BF, Feng YS, Xu JB, *et al*. Learning to predict charges for criminal cases with legal basis. Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. 2017. 2727–2736. [doi: 10.18653/v1/d17-1289]

13 Hu ZK, Li X, Tu CC, *et al*. Few-shot charge prediction with discriminative legal attributes. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, NM, USA. 2018. 487–498.

14 Zhong HX, Guo ZP, Tu CC, *et al*. Legal judgment prediction via topological learning. Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. 2018. 3540–3549. [doi: 10.18653/v1/d18-1390]

15 Hoffer E, Ailon N. Deep metric learning using triplet network. Proceedings of the 3rd International Workshop on Similarity-Based Pattern Recognition. Copenhagen, Denmark. 2015. 84–92. [doi: 10.1007/978-3-319-24261-3_7]