

类案检索技术研究综述

谢永峰¹, 尹华^{1,2}, 乔丹¹

(1. 广东财经大学信息学院; 2. 广东省智能商务工程技术研究中心, 广东 广州 510320)

摘要: 类案检索是法学和计算机相关学科的交叉应用, 类案检索技术旨在通过类案判定获得候选检索结果。当前研究分别从法律和计算机两个视角展开, 通过技术与领域的深度融合实现类案检索以精准服务法官判案需要, 尽管类案检索技术研究逐渐体系化, 但尚未存在相关技术发展的系统性综述。为此, 首先分析在法学上类案界定标准的变化, 以浅层和深层文本表示方法为划分标准, 论述相似案例匹配技术研究进展; 其次比较中国裁判文书网、“法信”“北大法宝”3类案检索平台的特点; 最后提出构建大型标注数据集、设计可解释性算法和提升检索效率是未来类案检索技术的主要发展方向。实验表明, 现阶段类案检索平台数量多, 但大都处于起步阶段, 缺乏具备可解释的稳定算法是影响其发展的重要因素。

关键词: 类案检索; 相似案例匹配; 文本表示; 类案; 类案检索平台

DOI: 10.11907/tj.231102

开放科学(资源服务)标识码(OSID):

中图分类号: TP391.3

文献标识码: A

文章编号: 1672-7800(2024)006-0198-07



A Survey on Law Case Retrieval Technology

XIE Yongfeng¹, YIN Hua^{1,2}, QIAO Dan¹

(1. School of Information, Guangdong University of Finance and Economics; 2. Guangdong Intelligent Business Engineering Technology Research Center, Guangzhou 510320, China)

Abstract: Case retrieval is an interdisciplinary application of law and computer science. Case retrieval technology aims to obtain candidate search results through case determination. The current research is conducted from two perspectives: legal and computer science. Through the deep integration of technology and fields, case retrieval is achieved to accurately serve the needs of judges in making judgments. Although the research on case retrieval technology is gradually becoming systematic, there is no systematic review of the development of related technologies. To this end, first analyze the changes in the standards for defining similar cases in law, using shallow and deep text representation methods as classification criteria, and discuss the research progress of similar case matching technology; Then compare the characteristics of three major case retrieval platforms: China Judgment Document Network, "Faxin", and "Peking University Magic Treasure"; Finally, it is proposed that building large-scale annotated datasets, designing interpretable algorithms, and improving retrieval efficiency are the main development directions of future case retrieval technologies. Experiments have shown that there are many similar case retrieval platforms at present, but most of them are in the early stages, and the lack of stable algorithms with interpretability is an important factor affecting their development.

Key Words: law case retrieval; similar case matching; text representation; similar case; case retrieval platform

0 引言

类案检索是智慧司法建设的一项重要应用。最高人民法院明确法官办理案件时, 应当依托中国裁判文书网、审判案例数据库等进行类案检索, 并对检索的真实性、

准确性负责。类案是指与待决案件在基本事实、争议焦点、法律适用问题等方面具有相似性, 且已被人民法院裁判生效的案件。目前, 类案检索方法包含关键词检索、法条关联检索及案例关联检索^[1]。其中, 类案检索技术是实施检索方法, 判定类案并提供检索结果的技术支持; 关键词检索和法条关联检索是目前类案检索平台所采用的主

收稿日期: 2023-02-13

基金项目: 教育部人文社会科学研究青年基金项目(21YJCZH202)

作者简介: 谢永峰(1997-), 男, 广东财经大学信息学院硕士研究生, 研究方向为自然语言处理; 尹华(1981-), 女, 博士, 广东财经大学信息学院副教授、硕士生导师, 研究方向为自然语言处理、机器学习、知识图谱; 乔丹(1997-), 男, 广东财经大学信息学院硕士研究生, 研究方向为自然语言处理。本文通讯作者: 尹华。

流方法,但受限于检索者的关键词抽象能力差异、检索平台的标签自动识别技术水平及人工标注标签规范性等因素,检索效果不佳,亟需新的智能检索技术支撑。案例关联检索根据案情描述在裁判文书库中寻找相似案例,是一种更为直观的检索模式,检索精度依赖于相似案例匹配算法及检索算法的有效性,也是目前研究者们重点关注的类案检索技术。

作为一项交叉学科应用,有关类案检索的研究成为当前法学和计算机相关学科的热点。研究者们分别从法学^[2]、法律实务^[3]、技术创新^[4]等角度展开研究。从法律视角而言,主要聚焦类案判定标准和类案检索应用瓶颈^[5,6]。从技术视角而言,研究者重点突破支持精准类案检索的关键技术,其中以相似案例匹配技术为核心重点^[7,8]。由于跨学科特性,类案检索存在“领域—技术鸿沟”问题。法律专家侧重法律需求,但并不涉及实现技术,某些需求可能在当前技术下无法解决,而计算机领域专家重点关注某一项具体技术的实现,往往会将领域需求简化,更关注问题背后的算法创新。实际上,只有技术与领域的深度融合,才能使得实现类案检索真正服务法官办案。

通过文献梳理发现,在类案检索技术领域的相关文献正逐年增加,但尚未系统地综述该领域的研究进展。目前,类案检索技术是法律应用背景下的信息技术创新,涉及自然语言理解、信息检索等前沿技术,其中类案的技术判定是难点。因此,本文采用跨学科交叉视角,从类案界定标准出发,分析类案检索法律功能需求,进而以文本表示方法为划分标准,重点论述相似案例匹配技术研究进展;然后分析现有类案检索平台的应用情况;最后提出类案检索技术发展的方向。

1 类案界定

类案检索的前提是界定何为类案。在《关于统一法律适用加强类案检索的指导意见(试行)》(简称《指导意见》)提出之前,已经有研究者对类案界定做过研究。李红海^[5]通过三段论思想总结任何司法过程都需要弄清楚事实真相和规范依据,而类案判定就是判断事实真相和规范依据的相似程度。其中,规范依据的选择取决于案件关键事实,也就是对于案件有重要影响的描述,因此类案判定即为关键事实相似判定。张骥^[6]认为争议焦点,即诉讼过程中当事人之间争论的问题是界定类案的关键,关键事实和争议焦点之间不可分割的,因此需要综合考虑关键事实和争议焦点作为类案界定的标准。《指导意见》则增加了法律适用作为类案界定标准之一,孙海波^[9]将《指导意见》提出的标准进一步细化为关键事实、争议焦点和规范目的的相似。由于规范目过于抽象,雷槟硕^[10]中将规范目的改为法律要件,即以一般方式描述的案件事实,例如案件事实通常由主体、客体等组成。

综上可知,在法学研究中类案界定逐渐清晰,类案检索技术的实施具有可操作性。目前,类案检索技术研究一般基于裁判文书,采用自然语言处理技术进行类案判定后提供检索。裁判文书是记录人民法院审理过程和结果的文档,在书写上具有一定的格式和内容要求。根据裁判文书结构可抽取关键事实和争议焦点。其中,关键事实为“本院审理查明”之后,“上述事实”之前的文本段;争议焦点为“本院认为”之后,“如不服本判决”之前的文本段。如图1、图2所示。

人陈兵红到庭参加诉讼。本案现已审理终结。

公诉机关指控并经本院审理查明:

1、2020年10月28日,被告人陈兵红在苏州工业园区某广场,将其私自通过广告店复制的5200张某某冰场门票谎称真票卖给被害人王某,骗得人民币129480元。

2、2021年1月18日,被告人陈兵红在苏州市吴中区郭巷,将其私自通过广告店复制的10800张某某冰场门票谎称真票卖给被害人王某,骗得人民币216000元。案发前,被告人回收了部分假票并退款给王某人民币16万元。被告人实际得款人民币185480元。被告人陈兵红被抓后如实供述了自己罪行,且自愿认罪认罚,其另行退赔了被害人4620元。

关键事实

上述事实,被告人陈兵红在庭审中不持异议,并有被害人王某的陈述笔录,证人毛某等人的证言笔录,聊天记录、转账记录,抓获经过、发破案经过,收条,认罪认罚具结书,户籍资料等证据经庭审质证予以证实,足以认定。

Fig. 1 Key fact distribution

图1 关键事实分布

破案经过,收条,认罪认罚具结书,户籍资料等证据经庭审质证予以证实,足以认定。

本院认为,被告人陈兵红以非法占有为目的,骗取他人财物,数额巨大,其行为已构成诈骗罪。被告人陈兵红归案后如实供述自己罪行,且自愿认罪认罚,依法从轻处罚。公诉机关指控被告人陈兵红犯诈骗罪的事实清楚,证据确实、充分,罪名成立,应予支持。据此,本院认为,被告人陈兵红退赔被害人王某人民币十八万零八百六十元。

依照《中华人民共和国刑法》第二百六十六条、第六十七条第三款、第六十四条及《中华人民共和国刑事诉讼法》第十五条之规定,判决如下:

一、被告人陈兵红犯诈骗罪,判处有期徒刑三年十一个月,并处罚金人民币一万元。(刑期自判决执行之日起计算;判决执行以前先行羁押的,羁押一日折抵刑期一日,即自2021年3月4日至2025年2月3日止;罚金自判决生效后第二日起一个月内缴纳,并上缴国库)。

争议焦点

如不服本判决,可在接到判决书的第二日起十日内,通过本院或者直接向江苏省苏州市中级人民法院提出上诉。书面上诉的,应当提交上诉状正本一份,副本两份。

Fig. 2 Distribution of controversy focus

图2 争议焦点分布

类案检索技术研究通常结合法律研究提出的类案判定方法,基于裁判文书进行类案匹配以供检索,其中类案匹配或相似案件匹配是核心。由于现有信息检索技术相对成熟,所以技术研究文献通常侧重于裁判文书理解。

2 相似案件匹配技术

由于法律案件以文档形式呈现,因此相似案件匹配本

质上是研究文档相似匹配的技术。文档通常以自然语言形式描述,不具有可计算性,通过文本表示方法将文本表示为理论意义上的等价形式,然后用于后续的相似计算。然而,采用的文本表示方法不同,相似案件匹配技术方法也会存在区别,因此本文按照文本表示发展,以提取文本信息的粒度为分类依据,将相似案件匹配技术划分为基于浅层文本表示的方法和基于深层文本表示的方法。

2.1 基于浅层文本表示的方法

浅层文本表示方法以提取字、词、短语等小粒度文本信息为特征,主要考虑字词本身所包含的信息。现有基于浅层文本表示的相似案例匹配技术也遵循这一过程,主要分为基于向量空间模型的方法和基于主题模型的方法。

2.1.1 基于向量空间模型的方法

基于向量空间的表示方法是一种词表示方法,根据文本中的词,将文本表示为一个向量,向量中各分量通常为词频大小等,向量维度则是词表大小,再通过向量来计算两个文本之间的相似度,常用的分量计算方法包括 TF-IDF。Wong^[11]使用 TF-IDF 反映每个词对整个系统文档的重要性,由于 TF-IDF 存在很多无意义的维度,通过降维 TF-IDF 后比较矩阵中每个文档和搜索关键字的相似性来计算搜索结果。王君泽等^[12]根据裁判文书的特点,在分词时根据词性的不同对词的重要程度进行分析,考虑了专有名词和未登录词对分词的影响。在计算相似度时,考虑到基础的向量空间模型会忽略词序的影响,将文档级相似度计算拆分为句子级相似度计算,即两个文档之间相似的句子占比越大,文档之间的相似度越高。张云婷等^[13]基于法理论中通过构成要件确定类案的思想,根据公共法律服务案例建立要素库,使用正则表达和人工提取相结合的方式提取裁判文书的案例要素,最后基于向量空间模型计算裁判文书之间的相似度。

向量空间模型的优点是简单直接,每个分量的维度都意义明确,但存在稀疏问题、未登录词问题和检索范围问题。例如,裁判文书中存在大量的低频词,导致词表过大从而向量变得高度稀疏;在检索过程中添加文档可能会出现未录用词,对词表进行修改时需要修改所有向量维度;仅考虑在裁判文书中出现过的词而未考虑到近义词和同义词,因此会缩小检索宽度。

2.1.2 基于主题模型的方法

基于向量空间模型的方法虽然能将文档映射到一个高维空间中,但维度较多时容易造成数据稀疏。主题模型尝试从概率生成模型的角度实现文本表示,每一个维度是一个“主题(topic)”,在降维的同时让每个维度都存在自身含义。Blei 等^[14]提出的 LDA 模型不仅能有效降低数据维度,还可识别文本隐藏的主题信息。Lyu 等^[15]认为基于 LDA 的模型在法律领域时应该给予不同“主题”不同的权重,于是在点互信息中调节每个特征词的权重,将其扩展到 LDA 模型中得到隐含权重信息的特征词,并将其表示为

特征向量。同时,为了弥补 LDA 模型过于泛化,在文本细节上表示能力不足的缺点,将 TF-IDF 值在阈值以上的特征词的 TF-IDF 值也拼接到表示文本的特征向量上,相较于传统 LDA 而言更适合运用在法律领域。张扬武等^[16]认为在法律领域使用 LDA 模型时,不同特征词需要不同的权重,但与文献[15]的不同之处在于基于 LDA 模型训练得到主题,将各个词的主题权重输入到 SVM 分类器中,以文本中带有法律意义的关键词作为正样本,其他词作为负样本进行训练,得到裁判文书在加权后的各个主题上的分布概率,最后通过加权的主题模型过滤裁判文书中的无用词语向量,以提升相似度计算的精确度和效率。然而,主题模型的主题数一般人工选定,主题数的差异也会对结果产生较大影响,进而影响模型泛化能力。

2.2 基于深层文本表示方法

基于浅层文本表示的方法仅考虑字词本身的分布信息,未充分考虑上下文语境和外部知识,无法考虑深层的语义信息。基于深层文本表示方法通常使用神经网络训练合适的模型进行文本表示,相较于浅层文本表示方法而言没有大量的人工干预,泛化能力相对较好。目前,根据法律知识的融合程度可将基于深层文本表示方法划分为基于神经网络的方法、融合法律知识的方法及基于知识图谱的方法。

2.2.1 基于神经网络的方法

严悦等^[17]提出使用神经网络来解决相似案例匹配问题,探索神经网络和相似案例匹配之间结合的可能性,但当时计算机性能较差,无法训练大规模的神经网络模型。李兰君等^[7]设计了第一个公开的中文法律案例文档相似度标注数据集,提出一个基于层级注意力机制的孪生网络模型用于计算文档相似度。具体为,通过词嵌入得到单词级的向量,利用双向 GRU 和注意力机制得到关注上下文信息的句子级的向量,然后通过相同方式获得文档级的向量,最后基于孪生网络计算相似度。此外,为了防止关键信息的遗漏,还使用 TextRank 对文档内容进行压缩,实验表明该模型能解决几百字以内的文本数据稀疏问题,但并未考虑局部语义信息。

由于在相似案例匹配中局部语义信息十分重要,许梓涛等^[18]设计了一个复杂的混合模型,由 BERT、LSTM 及 CNN 拼接而成,使用 BERT 进行文本表示,再利用双向 LSTM 和 CNN 提升本文表述的上下文信息和局部语义特征,使裁判文书转化为矩阵后包含的信息更全面^[19]。同时,使用三元组损失函数提升模型的鲁棒性和泛化性,具体模型如图 3 所示。虽然,上述两种方法都在类案检索中取得了不错的效果,文献[7]更是创建了中文法律案例文档相似度标注数据集,但并未考虑到现有裁判文书字数已达到数千字的规模,而长文本带来的问题是如果在长文本匹配中套用短文本匹配的模型将导致细粒度的匹配十分稀疏、噪声较大,难以捕捉有用的信息。

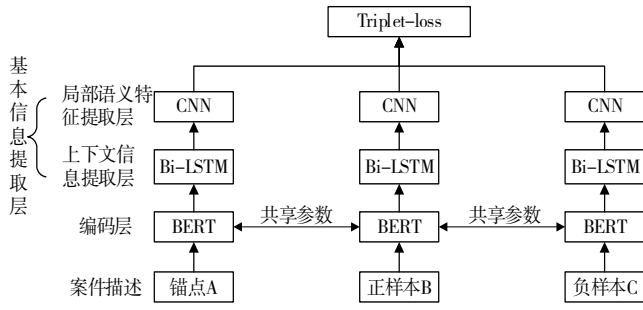


Fig. 3 Context aware case matching and recommendation model architecture

图3 上下文感知类案匹配与推荐模型架构

针对长文本问题,赖华等^[20]采用分段编码思想,使用 Word2Vec 将文本以词为单位进行向量表示后,将文本矩阵分为 5 段;然后利用不同卷积核的 CNN 模型进行编码,使用最大池化得到文本向量包含的关键信息;最后设计了一种文本向量之间相互交互的全连接层,让裁判文书之间产生更多交互。该方法虽然成功解决了文本过长带来的噪声过大问题,但与之前的模型存在一个共同问题,并未考虑到裁判文书中包含的法律信息。

(1)融合法律知识的方法。基于神经网络的方法最初是将神经网络模型应用于文本相似匹配中,并不关注文本中蕴含的法律信息,仅利用神经网络的学习能力。深入研究发现,仅采用神经网络模型的方法不具有可解释性,难以验证模型的可靠性,需要在现有方法中更多地融入法律知识。刘权等^[4]提出要从案例文档中抽取涵盖关键信息的案件要素,使用全连接层将案件要素和 BERT 作为编码器的文本表示进行信息融合,以加重关键语义信息在相似度计算中的权重,提升关键信息的对相似度计算的影响。为了解决案例文档过长无法使用 BERT 的问题,根据案例文书的特性从后往前截取文档,但文献^[4]提出的关于民间借贷的案件要素只有 6 个,文献^[21]在此基础上进行了完善,将民间借贷的案件要素提升至 10 个,并修改了模型

的损失函数,将其改为三元组损失函数和对比损失函数相结合的方式,进一步增强了模型区分细节特征的能力与鲁棒性。其中,三元组损失函数由式(1)所示, h^a, h^b, h^c 为编码器提取的 3 个特征表示;对比损失函数由式(2)所示,由 KL 散度对数据进行扰动,以达到数据增强的目的;最终损失函数由式(3)所示,文献中将 α, β 均设置为 1。实验表明,在 CAIL2019-SCM^[22]的相似案例匹配数据集中,该方法相较于文献^[4]的模型效果更好。

$$L_{tri} = \frac{1}{N} \sum_i^n \left[|h_i^a - h_i^c|^2 - |h_i^a - h_i^b|^2 + \gamma \right] \quad (1)$$

$$L_{con} = \frac{1}{3} \left(D_{KL}(h^a \| h_{adv}^a) + D_{KL}(h^b \| h_{adv}^b) + D_{KL}(h^c \| h_{adv}^c) \right) \quad (2)$$

$$L = \alpha L_{tri} + \beta L_{con} \quad (3)$$

由于上述模型的案例要素均为人工选定和提取,消耗了大量的人力资源,限制了模型泛化性。为此,不少学者将法条预测或要素预测作为模型的子任务,以相似案例匹配作为主任务的方式,让模型同时包含语义信息和法律要素信息。

刘博阳等^[23]以法律要素为标签训练神经网络,将网络结构输出层中学习到的法律要素信息的前一层的输出向量作为案例文本的向量表示,并通过该向量计算案例文本之间的相似度。

Peng 等^[24]采取多任务的方式实现显性知识的融入,将模型分为知识蒸馏层、信息融合层和文本匹配层 3 个部分。在知识蒸馏层通过 BERT 与全连接层的模型进行多标签预测,使 BERT 得到的文本表示包含更多的显性知识;在信息融合层将 BERT 的输出作为 CNN 的输入得到语义知识的文本表示;全连接层的输出作为 Bi-LSTM 的输入得到法律知识的文本表示,然后将二者拼接起来得到融合语义知识和法律知识的文本表示,最后在文本匹配层计算文本间相似度,具体模型如图 4 所示。然而,模型的多标签预测仅考虑到民间借贷这一个背景,在多背景任务上还有待提升。

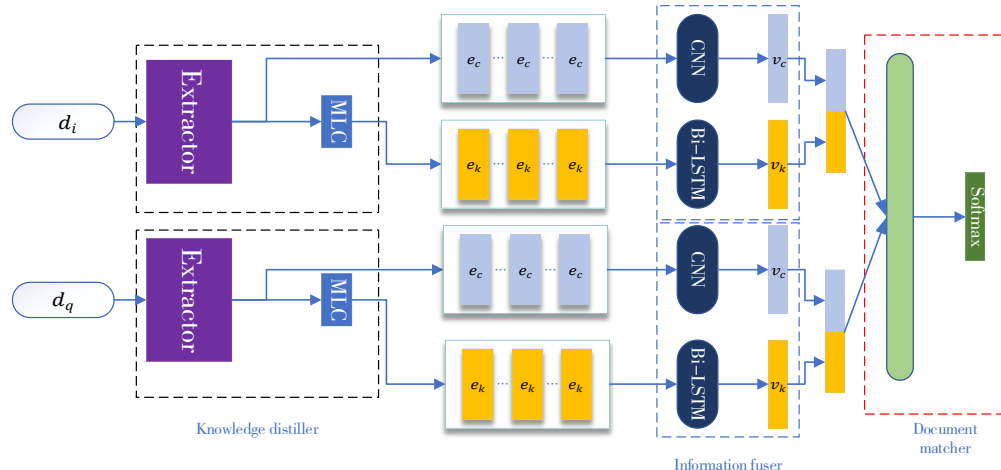


Fig. 4 Structure of explicit knowledge-enhanced model

图4 显式知识增强模型结构

由于大部分模型都使用了BERT预训练模型,现有裁判文书字数已达到数千字的规模,远超BERT能够处理的长度。为了在不丢失关键信息的情况下缩短文本,Li等^[25]利用裁判文书的结构特征截取文本中关键的4个部分,使用BERT得到文本表示后进行拼接,然后通过训练哈希函数得到各个特征的权重,最后使用三元损失函数对模型进行训练。Hu等^[26]简单将文本进行分段,在Bert编码的基础上通过LDA模型将每段的主题进行抽取,还提取出其中的构成要件,使向量表示更适合法律场景。

(2)基于知识图谱表示学习的方法。知识图谱表示学习是在不改变知识图谱固有结构的条件下将知识图谱中的实体和关系嵌入向量空间,得到知识图谱的低维向量表示,并对语义信息进行保留^[27]。

此外,利用知识图谱表示学习来实现相似案例匹配,也是现在相似案例匹配的一个研究方向,通常需要构筑一个精准的知识图谱表示学习来匹配相似案例。乔钢柱等^[28]以《刑事法律文书制作指南与范例》《量刑指导建议书》等主要知识源构筑了盗窃案例法律文书本体,包含300多个实体,提出本体和知识图谱间的映射关系是树与树、树与图的映射,通过这种映射规则将本体映射为知识图谱。洪文兴等^[8]提出以自然语言处理为核心技术的司法案件案情知识图谱自动构建技术,针对实体识别任务比较研究了两种基于预训练的实体识别模型,并针对关系抽取任务提出融合平移嵌入的多任务联合的语义关系抽取模型,获得了结合上下文的案情知识表示学习。然后,以此为基础实现了对数十万判决书案情知识图谱的自动构建,为类案检索提供了语义支撑,随后通过向量表示进行类案匹配。

黄治纲等^[29]针对现有知识图谱向量表示精度不高的问题,提出一种基于图聚类向量优化的案件知识图谱表征学习方法(FU-TransH)。通过图聚类算法针对TransH在负采样上存在的不足进行改进,以提升实体向量化的准确度,进而提升类案检索的准确度。表1为在CAIL2019-SCM数据集中的基准模型及上文提到的部分效果较好的模型准确率作了汇总。

Table1 Accuracy comparison of different models on the CAIL2019-SCM

表1 不同模型在CAIL2019-SCM数据集上的准确率比较

模型	准确率/%
CNN	69.53
LSTM	68.00
BERT	67.32
文献[4]	72.45
CASCMR ^[18]	73.74
文献[20]	76.04
文献[21]	74.39

3 类案检索应用平台

类案检索技术应用的载体是检索平台,而检索平台能实现的技术一般是经过实验验证较为成熟的算法和技术。目前,类案检索平台大致可分为最高人民法院平台、地方人民法院平台和商业平台。其中,最高人民法院平台主要有中国裁判文书网和“法信”等,地方人民法院平台包括上海刑事案件智能辅助办案系统(206系统)等,商业检索平台包含“无讼”“北大法宝”等。

截至2023年2月10日13时,中国裁判文书网总文书量已经高达约1.39亿篇,访问次数达992亿次,主要检索模式包含关键词检索和目录树检索。关键词检索主要包含案由、法院、当事人、律师或文书中出现过的原词等要素进行检索;目录树检索则根据裁判文书的案由、类别和罪名进行检索,例如选择“刑事案件”,然后根据法律要素、法院层级、法院地域、文书类型等进行检索,但中文裁判文书网尚未部署基于案例关联检索的类案检索,即中文裁判文书网目前无法实现以文搜文。

“法信”是2016年中国推出的首个法律知识和案例大数据融合平台,相较于中国裁判文书网的检索机制更为细致,实现了全文检索,但实现的逻辑还是依托原文关键词进行检索,虽然能从争议焦点等法律要素进行检索,但专业性强,非法律行业人员可能无法适用。“北大法宝”检索平台依赖“北大法宝”数据库而建立,数据库雏形的诞生可追溯到1985年,是北大法律系统创建的全国第一个法学信息数据库^[30]。1998年,北京大学法制信息中心依托“北大法宝”数据库开发了《中国法律检索系统》浏览器版,正式取名为“北大法宝”,检索方式相较中国裁判文书网而言更全面,既包含与中文裁判文书网类似的关键词检索,也有在近几年专门部署的类案检索窗口,通过输入基本事实、争议焦点或法律适用的事件对类案进行检索,相较于中文裁判文书网的检索方式准确率更高、适用范围更广,更能满足法律专业人士的需求;相较于“法信”“北大法宝”类案检索窗口而言,可以输入文本段,降低了用户的使用门槛。

4 类案检索技术发展

类案检索本质上由法律需求推动,最终目的是服务于司法体系。最高院希望通过智慧法院实现司法公平,要求法官进行类案检索,但法官的类案检索思路和现有类案检索技术实现思路并不一致,存在技术一领域鸿沟,尽管类案检索技术已经发展到一定程度,但仍然存在一些问题:

(1)缺乏用于技术研究的大型公共数据集。目前,已知的带标注的公共数据集只有CAIL2019-SCM中关于民间借贷的相似案例匹配数据集和CAIL2021中关于刑事案

件的类案检索数据集,数据规模不大且并非真实裁判文书。数据集规模和质量能直接影响模型训练的效果,但法律数据集人工标注需要标注者拥有法律领域相关知识,导致标注人力成本较高,且类案判定存在个人主观判断,将影响数据集质量。因此,研究无监督学习技术、自动标注技术及数据增强技术是未来的研究方向。

(2)算法黑箱问题。基于深层文本表示的相似案例匹配算法通常以神经网络为基础,其构成及参数优化是法官的知识盲点,因此研究可解释性的算法也是一个研究方向。

(3)检索效率问题。目前,有关类案检索技术的研究重点关注相似案例匹配技术,对于检索技术的创新研究还不够充分。以裁判文书网中文书为例,截至2023年2月10日,刑事案件文书已有约1亿篇,如何设计高效的检索技术并应用于类案检索平台是提升用户体验的关键之一。

5 结语

类案检索技术在《指导意见》颁布后,开始快速发展。法律领域首先从类案界定出发研究类案判定的标准;技术领域则重点研究相似案例匹配技术。基于向量空间模型的方法虽然操作简单、实现快速,但存在向量维度过高、未录用词和多义词影响较大的问题。

基于主题模型的方法对向量进行了降维,但“主题”个数只作为超参数,在一定程度上影响了模型的泛化性。LDA是在类案检索领域主要使用的基于主题模型的方法,但上述两种方法无法获得深层次的语义信息,而复杂的神经网络模型却可弥补此缺陷。通过BERT预训练模型缩减训练时间、CNN对关键语义进行提取、LSTM对上下文信息进行提取可极大提升匹配精度。同时,将法律知识融入模型中也能提升神经网络模型在类案检索中的精度。

现阶段,类案检索平台数量多,但大都处于起步阶段,缺乏具备可解释的稳定算法是影响发展的重要因素,解决数据集问题、可解释性问题和检索速度问题将对类案检索技术和平台发展产生巨大的积极意义。

参考文献:

- [1] Supreme People's Court. Guiding opinions on unifying the application of laws and strengthening the search of similar cases (trial implementation) [EB/OL]. <https://www.court.gov.cn/zixun-xiangqing-243981.html>. 最高人民法院. 关于统一法律适用加强类案检索的指导意见(试行) [EB/OL]. <https://www.court.gov.cn/zixun-xiangqing-243981.html>.
- [2] CAO L, LIU X Y. The difficulties and solutions of similar case retrieval applications—from the perspective of helping judges to adjudicate and write documents[J]. China Review of Administration of Justice, 2021(5): 162–172. 曹磊,刘晓燕. 类案检索应用的困境与破解——以助力法官裁决及文书撰写为视角[J]. 中国应用法学, 2021(5): 162–172.
- [3] TAN X T, PANG L. Application status and optimizing approach of similar case retrieval system [J]. Shandong Judges Training Institute Journal, 2022, 38(1): 93–104. 谭小彤, 庞丽. 类案检索系统的应用现状与优化进路[J]. 山东法官培训学院学报, 2022, 38(1): 93–104.
- [4] LIU Q, YU Z T, GAO S X, et al. Incorporating case elements for case matching[J]. Journal of Chinese Information Processing, 2022, 36(11): 140–147. 刘权, 余正涛, 高盛祥, 等. 融合案件要素的相似案例匹配[J]. 中文信息学报, 2022, 36(11): 140–147.
- [5] LI H H. The future of case guidance system and judicial governance capability[J]. Peking University Law Journal, 2018, 30(2): 493–511. 李红海. 案例指导制度的未来与司法治理能力[J]. 中外法学, 2018, 30(2): 493–511.
- [6] ZHANG Q. The expression and application of case judgment rules [J]. Modern Law Science, 2020, 42(5): 35–49. 张骥. 论案例裁判规则的表达与运用[J]. 现代法学, 2020, 42(5): 35–49.
- [7] LI L J, ZHOU J S, GU Y H, et al. Similar legal case retrieval based on improved siamese network[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2019, 55(1): 84–90. 李兰君, 周俊生, 顾颜慧, 等. 基于改进孪生网络结构的相似法律案例检索研究[J]. 北京大学学报(自然科学版), 2019, 55(1): 84–90.
- [8] HONG W X, HU Z Q, WENG Y, et al. Automated knowledge graph construction for judicial case facts[J]. Journal of Chinese Information Processing, 2020, 34(1): 34–44. 洪文兴, 胡志强, 翁洋, 等. 面向司法案件的案情知识图谱自动构建[J]. 中文信息学报, 2020, 34(1): 34–44.
- [9] SUN H B. Rediscovering the "same case": constructing the criteria for judging the similarity of cases [J]. Peking University Law Journal, 2020 (6): 262–281. 孙海波. 重新发现“同案”: 构建案件相似性的判断标准[J]. 中国法学, 2020(6): 262–281.
- [10] LEI B S. On the methodology of the similar cases judgement [J]. Modern Law Science, 2022, 44(6): 24–41. 雷棕硕. 类案判断的方法论[J]. 现代法学, 2022, 44(6): 24–41.
- [11] WONG T F. Documents retrieval system [D]. Hong Kong: City University of Hong Kong, 2021.
- [12] WANG J Z, MA H J, ZHANG Y, et al. A case similarity calculation modeling case pushing of judicial documents [J]. Computer Engineering & Science, 2019, 41(12): 2193–2201. 王君泽, 马洪晶, 张毅, 等. 裁判文书类案推送中的案情相似度计算模型研究[J]. 计算机工程与科学, 2019, 41(12): 2193–2201.
- [13] ZHANG Y T, YE L, FANG B X, et al. A similar case retrieval algorithm based on TF-IDF and law ontology [J]. Intelligent Computer and Applications, 2021, 11(5): 229–235. 张云婷, 叶麟, 方滨兴, 等. 基于词频-逆文档频率和法律本体的相似案例检索算法[J]. 智能计算机与应用, 2021, 11(5): 229–235.
- [14] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993–1022.
- [15] LYU B, HOU W L. Typical case recommendation of court texts based on

- topic model[J]. *Microelectronics & Computer*, 2018, 35(2): 128-132.
- 吕宾, 侯伟亮. 基于主题模型的法院文本典型案例推荐[J]. *微电子学与计算机*, 2018, 35(2): 128-132.
- [16] ZHANG Y W, LI G H, WANG L M. Research on topic model of legal texts based on keyword weighting[J]. *Computer & Digital Engineering*, 2019, 47(5): 1170-1174, 1228.
- 张扬武, 李国和, 王立梅. 基于关键词加权的法律文本主题模型研究[J]. *计算机与数字工程*, 2019, 47(5): 1170-1174, 1228.
- [17] YAN R, HA J B. Optimizing similar case matching method using ART neural network[J]. *China CIO News*, 2012(3): 70-74.
- 严悦, 哈进兵. 利用 ART 神经网络优化相似案例匹配方法[J]. *信息系统工程*, 2012(3): 70-74.
- [18] XU Z T, HUANG B S, PAN W K, et al. A novel context-aware similar case matching and recommendation method[J]. *Journal of Taiyuan University of Technology*, 2022, 53(1): 80-88.
- 许梓涛, 黄炳森, 潘微科, 等. 一个新的上下文感知类案匹配与推荐方法[J]. *太原理工大学学报*, 2022, 53(1): 80-88.
- [19] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bi-directional transformers for language understanding[C]// *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019: 4171-4186.
- [20] LAI H, ZHANG H T, XIAN Y T, et al. A similarity case matching method combining segment encoding and affine-mechanism[J]. *Journal of Shandong University(Natural Science)*, 2023, 58(1): 40-47.
- 赖华, 张恒滔, 线岩团, 等. 融合分段编码与仿射机制的相似案例匹配方法[J]. *山东大学学报(理学版)*, 2023, 58(1): 40-47.
- [21] ZHENG J, HUANG H, QIN Y B. Matching similar cases with legal knowledge fusion[J]. *Data Analysis and Knowledge Discovery*, 2022, 6(7): 99-106.
- 郑洁, 黄辉, 秦永彬. 一种融合法律知识的相似案例匹配模型[J]. *数据分析与知识发现*, 2022, 6(7): 99-106.
- [22] XIAO C J, ZHONG H X, GUO Z P, et al. CAIL2019-SCM: a dataset of similar case matching in legal domain[DB/OL]. <https://arxiv.org/abs/1911.08962>.
- [23] LIU B Y, LI S, YE L, et al. Similar case recommendation algorithm based on legal elements[J]. *Intelligent Computer and Applications*, 2021, 11(6): 1-4, 13.
- 刘博阳, 李尚, 叶麟, 等. 基于法律要素引导的相似案例推荐算法[J]. *智能计算机与应用*, 2021, 11(6): 1-4, 13.
- [24] PENG D, YANG J, LU J. Similar case matching with explicit knowledge-enhanced text representation[J]. *Applied Soft Computing*, 2020, 95: 106514.
- [25] LI J, LIU X, NIE X, et al. Weighted-attribute triplet hashing for large-scale similar judicial case matching[J]. *Computational Intelligence and Neuroscience*, 2021: 6650962.
- [26] HU W, ZHAO S, ZHAO Q, et al. BERT_LF: a similar case retrieval method based on legal facts[J]. *Wireless Communications and Mobile Computing*, 2022: 2511147.
- [27] YIN H, XIAO S R, CHEN Z Q, et al. Knowledge graph completion method based on multi-semantic relation embedding[J]. *Journal of Frontiers of Computer Science and Technology*, 2023, 17(2): 134-477.
- 尹华, 肖石冉, 陈智全, 等. 多语义关系嵌入的知识图谱补全方法[J]. *计算机科学与探索*, 2023, 17(2): 134-477.
- [28] QIAO G Z, FENG T T, ZHANG G C. Research on intelligent reasoning of legal documents in theft case based on knowledge graph[J]. *Computer Systems & Applications*, 2019, 28(7): 206-213.
- 乔钢柱, 冯婷婷, 张国晨. 基于知识图谱的盗窃案件法律文书智能推理研究[J]. *计算机系统应用*, 2019, 28(7): 206-213.
- [29] HUANG Z G, XIE X Q, XING T J, et al. Case recommendation based on knowledge graph of judicial cases[J]. *Journal of Nanjing University(Natural Science)*, 2021, 57(6): 1053-1063.
- 黄治纲, 谢新强, 邢铁军, 等. 基于司法案例知识图谱的类案推荐[J]. *南京大学学报(自然科学)*, 2021, 57(6): 1053-1063.
- [30] LI Y S. Laws & regulations database-chinalawinfo: exploring the "smart rule of law"[J]. *Faren Magazine*, 2021(11): 83-85.
- 李韵石. 北大法宝: 探寻“智慧法治”[J]. *法人*, 2021(11): 83-85.

(责任编辑: 刘嘉文)