“智慧法院”数据融合分析与集成应用

秦永彬[1,2]，冯丽[1]，陈艳平[1,2]，黄瑞章[1,2]，刘于雷[3]，丁红发[1]
1. 贵州大学计算机科学与技术学院，贵州 贵阳 550025；
2. 贵州大学贵州省公共大数据重点实验室，贵州 贵阳 550025；3. 多彩贵州网有限责任公司，贵州 贵阳 550000

摘要
针对“智慧法院”建设中存在的共性问题和实际需求，介绍了“智慧法院”数据融合分析及集成应用示范平台的架构。从司法大数据深度语义学习、基于知识图谱的司法数据融合、司法数据安全防护与隐私保护以及司法数据融合分析的可视化4个方面，探讨了“智慧法院”建设中共性关键技术的研究思路和实现路径。最后，以证据抽取、犯罪行为链构建和法律条文推荐为例，展现了数据融合分析及集成应用示范平台的应用效果。研究成果对实现以法院司法数据为核心的新一代“智慧法院”建设目标具有一定的参考价值。

关键词
智慧法院；知识图谱；数据融合；融合分析应用

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2019022

“Intelligent Court” data fusion analysis and integrated application

QIN Yongbin[1,2], FENG Li[1], CHEN Yanping[1,2], HUANG Ruizhang[1,2], LIU Yulei[2], DING Hongfa[1]
1. College of Computer Science & Technology, Guizhou University, Guiyang 550025, China
2. Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China
3. Colorful Guizhou Net Company with Limited Liability, Guiyang 550000, China

Abstract
In view of the common problems and practical needs in the construction of “Intelligent Court”, the framework of “Intelligent Court” data fusion analysis and integration application demonstration platform was introduced. The research ideas and implementation paths of common key technologies in the construction of “Intelligent Court” were put forward from four aspects: deep semantic learning of judicial big data, judicial data fusion based on knowledge map, judicial data security protection and privacy protection, and visualization of judicial data fusion analysis. Finally, taking evidence extraction, criminal chain construction and legal provisions recommendation as examples, the application effect of data fusion analysis and integrated application demonstration platform was analyzed. The research results have certain reference value for realizing the goal of building a new generation of “Intelligent Court” with the judicial data of courts as the core.

Key words
Intelligent Court, knowledge map, data fusion, application of fusion analysis

1 引言

2016年,《国家信息化发展战略纲要》将建设“智慧法院”列入国家信息化发展的战略中。同年,《“十三五”国家信息化规划》中明确指出:支持“智慧法院”建设,推行电子诉讼,建设完善公正司法信息化工程;提高案件受理、审判、执行、监督等各环节的信息化水平;推动执法司法信息公开,促进司法公平正义。目前,各级人民法院以“智慧法院”建设为依托,加快推进人民法院信息化建设,取得了重大进展:各级法院都在积极推动互联网、物联网、大数据、人工智能与法院工作的深度融合,围绕“智慧法院”的业务办理、信息检索、文书审阅、案件预测、智能咨询等信息化应用,不断地进行探索和实践;各级法院的全业务网上办理的网络化格局基本形成;利用互联网推动全流程依法公开的阳光化目标基本实现。这些进展为人民法院的智能化建设打下了坚实基础。

在此过程中,以大数据和人工智能技术为支撑的“智慧法院”系统不断发展,有效推动了司法领域的智能化实践。例如,以人工智能为基础的庭审语音转文字、裁判文书或起诉书的自动生成、法律文件辅助审阅、裁判文书证据材料自动抽取等,为“智慧法院”的审判质量评估、精准分案、自动量刑和辅助审判等应用需求提供了核心技术支撑。基于大数据和人工智能的辅助办案可以显著提高法院办案的工作效率,一方面可以帮助当事人形成最佳的诉讼策略,节约诉讼成本;另一方面,可以帮助法官实现同案同判,确保判决一致性,增强司法公信力,确保司法公平正义。此外,法院通过人工智能客服机器人,可以为公众提供在线的实时法律咨询服务。

然而,“智慧法院”建设涉及的数据来源广泛、结构复杂、动态实时,法院数据在对数据进行有效利用的过程中急需解决以下问题:

- 数据来源和存储结构的多样性造成了部门内部或体系内部的“数据孤岛”;
- 数据具有鲜明的领域性和专业性,数据特征隐藏较深,导致数据挖掘分析性能较差;
- 由于数据的多源异构性,司法知识体系难以构建;
- 数据涉及面广、敏感度高,导致数据安全保障和隐私保护问题迫切。

笔者团队针对这些问题,基于贵州省高级人民法院的司法大数据,开展了理论研究和技术攻关,探索了司法领域数据的融合分析和集成应用方法,目标是通过整合法院现有的审判数据、业务数据以及人员数据等,提升法院审判过程的智能化水平,实现法院的自动分案、人案关联分析、审判态势数据智能统计与分析等业务需求。本文分析了“智慧法院”在司法数据融合、数据安全与隐私保护、数据可视化方面的研究现状,针对“智慧法院”建设中存在的实际问题,提出了司法大数据深度语义学习方法、基于知识图谱的司法数据融合方法、司法数据安全防护与隐私保护以及数据融合分析的可视化应用示范的研究思路和技术路线;并以证据抽取、犯罪行为链构建和法律条文推荐为例,阐述了本文所做的研究在实际应用中的有效性,为审判工作的智能化、精细化提供了有效的支撑。

2 国内外研究进展

“智慧法院”是2016年提出的司法领域的信息化发展战略,到2018年,智慧法院3.0版形成,相关研究与应用取得了重大进展。

近年来，以大数据和人工智能技术为基础的应用已经在政府治理[1-3]、生态建设[4]、创新平台[5]、智慧城市[6]等领域取得了广泛应用，也为法院业务办理、信息检索、文书审阅、案件预测、智能咨询等信息化应用带来了巨大便利[7]。早在1986年，美国的法律体系中就出现了人工智能[8]一词，现如今，美国的机器人律师已经可以提供简单的业务办理服务，这标志着人工智能在法律实践中迈出了重要一步。Remus D等人[9]认为机器人正在慢慢取代律师，Payne S等人[10]认为计算机技术在法律行业的应用对法学教育产生了重大影响。国内学者对“智慧法院”的研究早在其概念提出之前就开始了。2001年，张保生[11]从法律推理与人工智能的关系的角度探讨了人工智能法律系统的历史及发展动力。2018年，季卫东[12]指出科技手段只是司法实践中的辅助手段。郝铁川[13]指出人们应理性对待人工智能在司法应用中的作用。

在司法数据融合方面，目前采用的较多的是传统信息检索技术，通过向量空间模型提供面向文档的分析。该技术的缺点是不能提供面向文档内容的信息融合与分析。传统的信息提取、知识融合方法性能相对偏低。为有效地支撑法官画像、自动量刑、法律条文推荐等面向文档内容的分析与应用，需要充分挖掘句子的结构信息和语义特征，提升知识融合的性能。在已有研究中，获取句子结构特征的方法主要有4类：N元语法[14]、基于树核的方法[15-16]、序列模型（如隐马尔可夫模型、条件随机场、循环神经网络等）[17]、复合特征[18-19]。另外，Jacobs D W等人[20]提出了一种广义多视角的特征提取方法。Zhou J T等人[21]提出一种域自适应学习方法。Zhu Y等人[22]通过文本语义提升图像特征抽取的质量。Liu K等人[23]提出一种基于重构错误累计矢量的跨媒体融合方法。Zhang X M等人[24]提出一种基于跨媒体数据协作的社会图像标注方法。在特征选择方面，Liu M等人[25]将链接限制条件作为区分能力的正则项，提出一种基于成对限制稀疏学习的特征方法。相对于其他领域的数据融合，司法数据融合主要聚焦案件分析。国内针对司法数据融合的研究相对来说还比较少[26]。

在数据融合与应用的过程中，数据的安全与隐私保护也是必须解决的问题[27-28]。匿名和泛化是用于隐私保护的经典方法，能够在保证对外发布数据真实性的同时保护隐私信息。Amro B等人[29]提出了一种基于用户偏好的假名机制，该机制通过用户自发地更换假名的方式，达到隐私保护的目的。Ghinita G等人[30]利用k-匿名的思想，用空间匿名区域的位置代替用户的精确位置，从而达到隐私保护的目的。Wu S等人[31]将k-匿名应用于众包数据库中的数据隐私保护。目前，k-匿名采用泛化技术对原始数据进行匿名处理，当属性的维数较高时，为满足泛化要求，必将导致更多的信息丢失，使得发布的匿名数据可用性变差。

数据可视化是数据应用示范的重要途径，可视化技术将数据挖掘过程中抽取出的潜在的、有价值的知识展示给用户，从而满足用户的不同需求。可视化数据挖掘技术已经成为数据挖掘领域的研究热点[32]。汪加才等人[33]设计了一个交互式可视化数据挖掘系统，允许用户以交互的方式从标记图或距离图中选定感兴趣的区域，加以深入分析。陈涛等人[34]研究了关联数据的可视化技术。袁海等人[35]研究了中文文本的可视化技术，文本可视化提高了数据的可视化效率和转换成利益的效率。

在司法领域的数据融合分析与集成应用过程中，还没有形成相对系统化的研究体系或应用体系，许多针对司法文本的分析大多使用开放域或传统的文本分析

的方法。目前，也没有针对司法数据的专门的数据安全和隐私保护的体系或模式。

3 “智慧法院”数据融合分析及集成应用示范平台架构

“智慧法院”数据融合分析及集成应用示范平台架构如**图1**所示。以贵州省高级人民法院为例，通过分析其在“智慧法院”建设过程中存在的实际问题和对智慧办案、智慧办公、智慧运维、智慧监督的需求，本文重点研究“司法大数据深度语义学习”“基于知识图谱的司法数据融合”“司法数据安全防护与隐私保护”和“司法数据融合分析的可视化”4个内容。为贵州省高级人民法院面向“数字法官-金字团队-智慧法院”的司法大数据应用提供理论依据和技术支撑。

3.1 司法大数据深度语义学习

司法数据包含各种结构化和半结构化数据。司法数据中的各类文档因撰写格式、使用措辞、时间、法院甚至法官和团队的不同而存在很大差异。这种差异会导致相似案件的裁判文书在表达方式上存在明显的区别。大量的案件信息隐藏在非结构化的办案文件中，例如，裁判文书中的案情特征对法条推荐结果、案件审判结果有显著影响；涉及具体的案件审判时，裁判文书中的案情特征的顺序可能直接影响案件的审判结果；相同特征的案件，由于案情特征序列的不同也可能导致审判结果的不同。司法数据融合分析的目标是有效地挖掘法院多源异构数据中的隐含知识，以支撑法院的数据应用。为了支撑“智慧法院”司法数据的融合应用，应重点研究针对司法数据的深度语义学习方法。通过深度学习模型，挖掘文本中的语义信息，识别其中的案件知识要素，从而有效地解决数据融合过程中的语义理解问题。

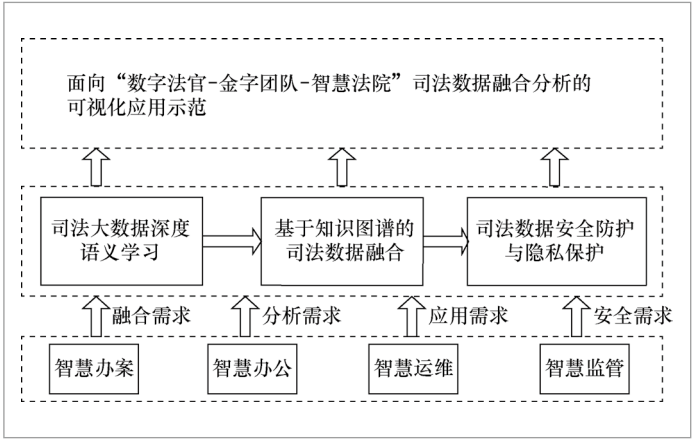

图1 “智慧法院”数据融合分析及集成应用示范平台架构

3.2 基于知识图谱的司法数据融合

在司法数据深度语义分析的基础上，构建“智慧法院”数据知识图谱，融合司法数据中的案件要素，并研究案情的演化分析方法。司法数据多元化、深度化、层次化的特点导致案件的演化分析和案件知识的转化非常困难。针对这些特点，笔者采用了基于知识图谱的司法数据融合方法。在知识图谱构建的过程中，需要研究不同粒度实体的提取方法；然后，识别实体之间的关联关系。在实体识别和关系识别的基础上，构建“智慧法院”应用中的司法大数据知识图谱。在融合与分析的过程中，各类算法需要具备良好的可扩展性和实时性，满足系统平台对知识图谱的实时检索、快速更新处理的需求。在应用过程中，应通过深入分析多层次知识图谱的演化性质，支撑法院审判工作中的案情演化分析。

3.3 司法数据安全防护与隐私保护

“智慧法院”信息化系统中积累了大量的司法敏感数据和个人隐私数据。数据安全防护和个人隐私保护是“智慧法院”建设与应用的关键基础，也是数据融合分析的核心需求。在数据融合过程中，需要精确定位案件数据中敏感数据的位置和安全需求等级，设计针对司法敏感数据的访问控制、数据传输安全、数据访问接入安全认证、个人隐私数据自动化识别、隐私度量等防护方案；通过数字签名、访问控制、对称加密、Hash算法等数据安全防护中常用的方案，结合差分隐私、隐私量化、匿名技术、泛化技术等隐私保护技术，构建司法领域的数据安全和隐私保护体系，实现多源数据融合过程中的数据安全防护与隐私保护。

3.4 司法数据融合分析的可视化

针对贵州省高级人民法院建设“数字法官-金子团队-智慧法院”示范应用的需求，借助深度语义分析、特征画像、法院知识图谱构建和可视化等关键技术，实现从“法官”到“团队”再到“法院”的可视化展示及全方位评价，从而支撑“精准分案”和“智能化推荐”等应用，提高司法审判的效率和质量，促进专业化审判团队的发展。

4 共性关键技术

为了支撑“智慧法院”建设中的数据融合分析与集成应用研究，笔者提出了基于深度神经网络边界组合实体识别方法、多通道实体关系识别方法、证据识别方法、犯罪行为识别方法和句法要素识别方法；构建了以犯罪行为为中心的知识图谱，有效支撑了审判质量评估、精准分案和自动量刑等具体应用需求。相关研究内容为“智慧法院”数据融合分析及集成应用示范提供了理论依据和技术支持。其中涉及的共性技术和研究思路，具体如下。

4.1 司法数据的特征表示技术

传统司法数据的特征表示主要采用向量空间模型，该模型把文档空间映射到一个测度空间，文档的相似度对应测度空间中文档向量的距离。法院各类文书的异质性（如产生的时间不同、法院不同、法官不同等）使得传统的向量空间模型容易产生高维的稀疏特征，不利于针对司法文档的语义分析。为此，本文研究了一种能够处理异质数据的特征抽取与自适应匹配的方法，即基于深度语义特征提取技术。该方法利用深度学习方法挖掘司法大数据中的深度语义特征，可以有效支撑司法数据知识图谱中的案件要素抽取。其技术路线如**图2**所示[36]。

深度语义特征提取技术的研究思路如下。

- 语义结构空间生成：利用神经网络把浅层特征映射到一个深度语义空间，利用特征组合产生潜在的语义结构空间，提高数据可分性。
- 语义结构排序：通过计算候选语义结构和目标语义结构的距离，进行排序。
- 特征选择：根据句子的结构信息和语法功能，利用先验知识操作划分后的特征集合。
- 特征画像：针对法院的各知识要素，建立实体的特征体系，建设特征实体算法库，实现准确、高效的法院大数据的实体画像拓扑集。

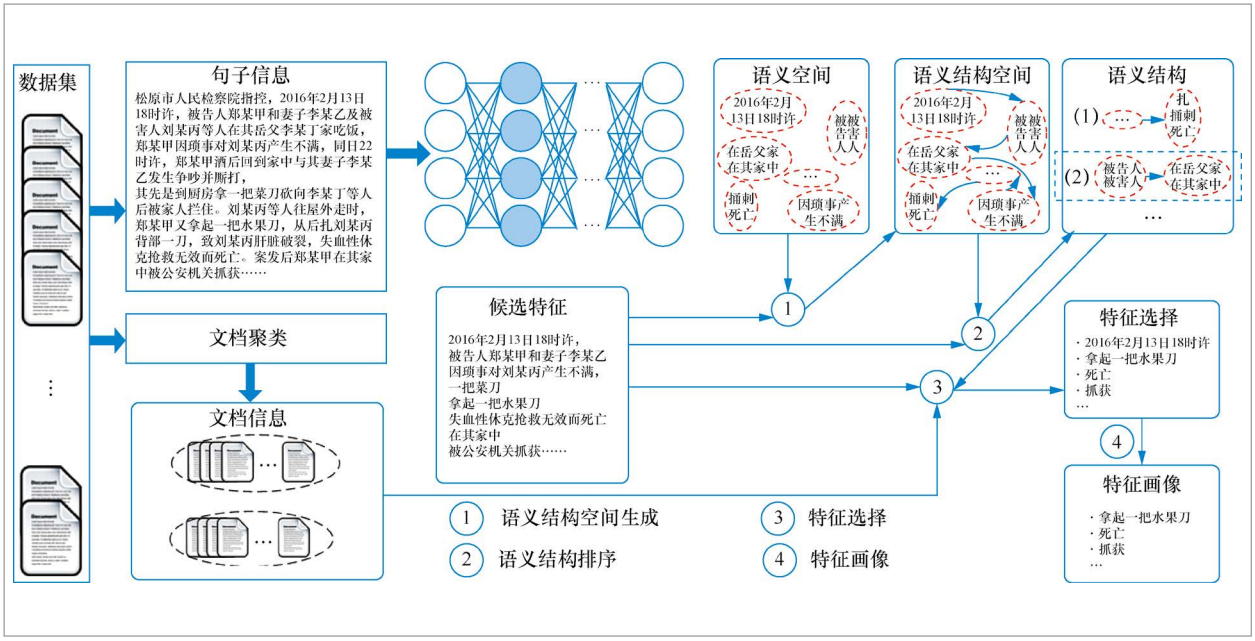

图 2 深度语义特征提取技术路线

4.2 基于知识图谱的司法数据融合技术

基于知识图谱的司法数据融合技术实现过程共分6个步骤，如**图3**所示。第一步，在实体特征画像的基础上，对相似实体进行合并和消歧。第二步，识别实体的关联关系，建立表述实体间关系的知识图谱。第三步，基于同类实体之间的强关联关系，合并同类实体，构建多粒度实体。第四步，挖掘多粒度实体之间的关联关系，构建多层次知识图谱。第五步，利用时间特征，针对实体进行特征的演变识别，利用动态数据的进化算法进行知识进化学习。第六步，利用知识图谱的链接预测方法，衡量实体间的全局和局部相似度，推断实体与实体间的间接关系。

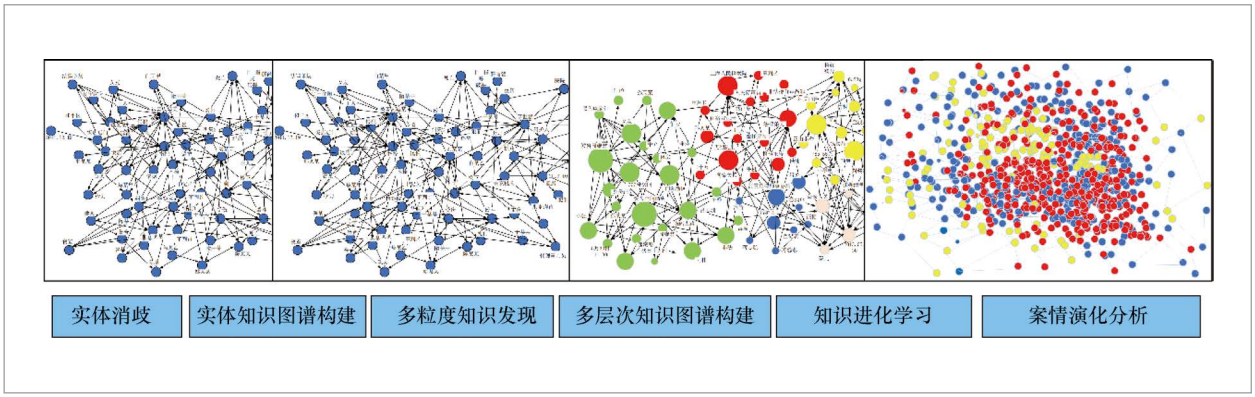

图 3 基于知识图谱的司法数据融合技术路线

4.3 融合过程中的数据安全与隐私保护技术

数据安全与隐私保护技术的研究思路如**图4**所示。第一步，制定对法院大数据进行数据安全和隐私保护数据的分级、分类标示方法；第二步，在数据安全防护中采用数字签名、对称加密、Hash算法等技术，在隐私保护中采用隐私风险控制、隐私量化、匿名技术、泛化技术等；第三步，根据数据安全标示确定数据的安全访问权限，确定数据脱敏方法，并进行数据的并行脱敏处理；第四步，根据多源异构数据的索引结构和过滤算法，建设脱敏后数据的索引方法。

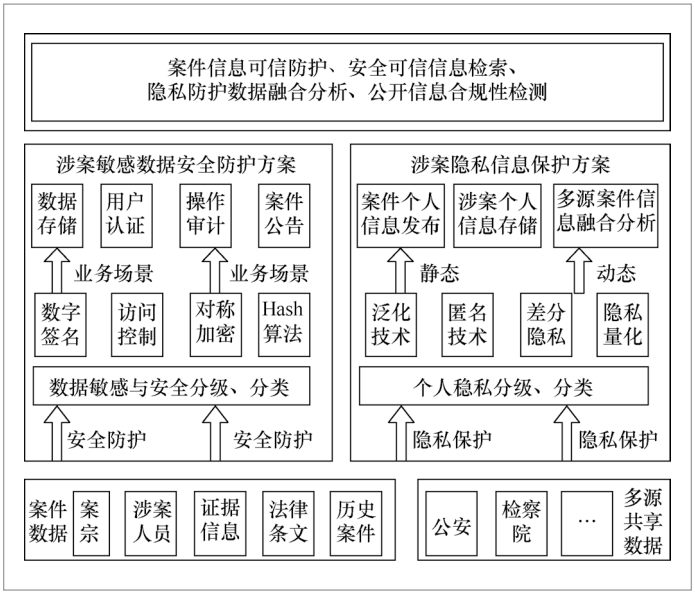

图 4 数据安全与隐私保护技术路线

4.4 “智慧法院”的可视化应用示范

“智慧法院”的可视化应用示范平台工作流程如**图5**所示。第一，通过法院数据融合与分析平台进行集成数据标准管理，通过特征管理系统建立实体特征体系；第二，通过数据抽取、转换、加载（ETL）和应用程序编程接口（API），实现法院内部数据与外部数据的实时与批量导入，并进行初步的数据整合；第三，利用基于知识图谱的分析技术、实体特征体系、集成数据标准进行数据融合与分析；第四，将分析后的数据导入基

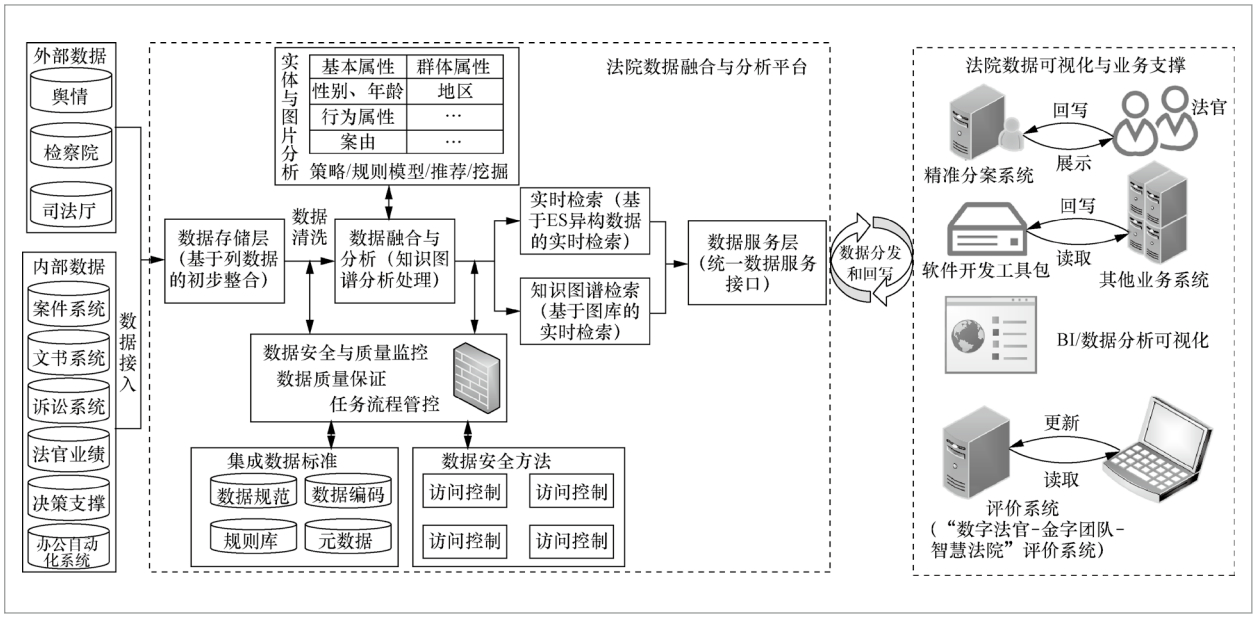

图 5 “智慧法院”的可视化应用示范平台工作流程

于搜索引擎 (Elasticsearch, ES)、图库的检索系统, 通过统一的数据服务接口对外提供数据服务; 第五, 法院数据可视化与服务支撑平台、法院数据融合与分析平台进行数据的分发与回写。

5 应用案例

以证据抽取、犯罪行为链构建和法律条文推荐为例, 本节简单介绍笔者提出的“智慧法院”的数据融合分析与集成应用的研究进展和应用效果。

5.1 证据抽取

根据法官办案的要求, 裁判文书中的证据条目过少或过多会导致量刑中的轻判或重判。对裁判文书中的证据进行自动抽取, 并与案件卷宗中提供的证据列表进行一一对比, 可以对案件的审判质量进行评估。在实现中, 首先选择650篇裁判文书, 标注其中的证据部分①。然后, 采用基于深度神经网络的命名实体识别技术进行证据的自动抽取。研究发现, 裁判文书中的证据在表达上与传统命名实体存在一定的差异, 其结构性比较强、字数多、中心词明显 (如××复印件、××结婚证等), 而且存在互相嵌套的问题。传统的序列标注模型 (如隐马尔可夫模型、条件随机场、循环神经网络等) 主要依赖局部特征, 不能有效识别嵌套证据, 在证据抽取中的效果比较差。根据裁判文书中证据条目的文本特点, 笔者提出了基于神经网络的边界组合证据抽取模型。该方法首先利用循环神经网络模型 (Bi-LSTM-CRF) 识别证据的边界 (如边界的开始、边界的结束), 然后通过组合证据的边界产生证据候选词, 再根据证据候选词的上下文特征, 采用卷积神经网络 (convolutional neural network, CNN) 识别出其中的正确证据。与直接使用Bi-LSTM-CRF模型的方法相比, 该方法能够在性能上提升10%以上的F值。

5.2 犯罪行为链构建

在法院具体的案件审判工作中, 需要对案件的案情进行分析。裁判文书中的案情描述和预测案件涉及的相关法条、罪名等信息对辅助法官办案有重要的作用。在传统的审判辅助工作中, 司法人员主要采用案件的文本特征进行法条、量刑、案由识别。在这种情况下, 通常不考虑案情要素之间的顺序关系。在实际应用中, 案件要素之间的时间序列、行为序列会直接影响最终的审判结果。对于相同案件要素的案件, 案情要素序列的不同会导致审判结果的不同。比如, “强奸杀人”和“杀人辱尸”, 犯罪行为的顺序不同会导致判决结果的差异。针对这一问题, 在知识图谱构建中, 笔者通过对案情行为序列关键词进行提取, 建立与之对应的犯罪“行为链”, 展现案情的主要情况、时序关系和发展趋势, “行为链”以“行为词”为中心, 围绕“行为词”提取关键案情要素特征, 通过行为序列构建行为词与其他以“行为”为中心的特征词之间的关联关系, 进而构建以“犯罪行为”为中心的司法数据间的内在关联和知识体系。如此, 笔者可以借助犯罪“行为链”实现对案情语义的深层分析。

5.3 法律条文推荐

传统的法律条文推荐主要依靠法律文本中的案情特征进行预测。这种情况无法考虑裁判文书中案情要素的时间特征。在裁判文书分析中, 案件要素的时间特征对审判结果会有较大的影响。为了有效地利用

① 案件裁判文书为半结构化数据, 其中包括首部、事实 (案情描述)、理由 (证据部分)、判决结果等部分。根据裁判文书的特点, 笔者只标注其中的案件描述部分

案件要素的时间信息提升法律条文预测的性能，辅助法官办案，笔者利用案情的事实描述和犯罪行为序列预测案件涉及的相关法律条文，从而验证基于犯罪行为序列的法律条文预测的有效性。在实现上，笔者利用TextCNN模型对裁判文书中案情描述部分的信息进行处理，获得文本中的语义信息。由于卷积神经网络模型不能有效地获取文本中案情要素之间的语义依赖关系，笔者采用Bi-LSTM模型对文本中抽取的犯罪行为过程进行建模，获取裁判文书中案情要素之间的时间特征，然后与TextCNN模型的输出进行拼接，最后通过全连接层和softmax函数得到法律条文的预测结果。与直接采用TextCNN的方法相比，该模型在法律条文的推荐上能提升6%的F值。

6 结束语

“智慧法院”的建设是司法改革的重要抓手。本文基于贵州省高级人民法院的案件数据，探索了司法大数据深度语义学习方法、基于知识图谱的数据融合与分析、司法数据安全防护与隐私保护等围绕智慧法院应用的相关研究，并提出了研究思路和技术实现路径，以期帮助法院实现从信息化到数据化、智能化的提升，实现以法院司法数据为核心的新一代的“智慧法院”建设。此外，笔者重点阐述了以数据为核心的应用理念，从司法领域的层面为政务领域大数据的应用提供了多源异构政务数据的融合分析与应用案例。

参考文献：

[1] 安小米，郭明军，洪学海，等. 政府大数据治理体系的框架及其实现的有效路径[J]. 大数据，2019，5(3)：3-12.
AN X M，GUO M J，HONG X H，et al. Framework of government big data governance system and effective way of implementation[J]. Big Data Research，2019，5(3)：3-12.

[2] 陈刚. 运用大数据思维和手段提升政府治理能力[J]. 大数据时代，2017(3)：6-13.
CHEN G. Using big data thinking and means to improve governance ability[J]. Big Data Time，2017(3)：6-13.

[3] 连玉明. “人在干、云在算”——“数据铁笼”助力政府治理[J]. 大数据时代，2017(3)：50-54.
LIAN Y M. “Man is doing things，cloud is calculating”-data cage helps government governance[J]. Big Data Time，2017(3)：50-54.

[4] 鲍旭华，曲晓东，郑新华. 大数据驱动的安全协同生态建设[J]. 大数据，2018，4(3)：93-100.
BAO X H，QU X D，ZHENG X H. Big data driven security collaborative ecological construction[J]. Big Data Research，2018，4(3)：93-100.

[5] 张平文，鄂维南，袁晓如，等. 大数据分析与应用技术创新平台[J]. 大数据，2018，4(4)：86-93.
ZHANG P W，E W N，YUAN X R，et al. Big data analysis and application technology innovation platform[J]. Big Data Research，2018，4(4)：86-93.

[6] 卢英华. 大数据在智慧城市规划中的应用[J]. 智能建筑与智慧城市，2019(1)：75-76.
LU Y H. Application of big data in intelligent urban planning[J]. Intelligent Building and Smart City，2019(1)：75-76.

[7] 和荒. 人工智能进法院：对科技应用于司法的思考[J]. 科技与法律，2018(6)：77-88.
HE Y. Artificial intelligence entering the court：thoughts on the application of science and technology to justice[J]. Science Technology and Law，2018(6)：77-88.

[8] WATERMAN D A，PAUL J，PETERSON M A，et al. Expert systems for legal decision making[J]. Expert Systems，1986，3(4)：212-226.

[9] REMUS D，LEVY F S. Can robots be lawyers? Computers，lawyers，and the practice of law[J]. SSRN Electronic Journal，2015.

[10] PAYNE S，KOOPS K. Opening remarks：a potpourri of recent developments affecting the teaching of transactional law and skills[J].

Transactions: The Tennessee Journal of Business Law, 2016, 18(2): 333-342.

[11] 张保生. 人工智能法律系统的法理学思考[J]. 法学评论, 2001(5): 11-21.
ZHANG B S. Jurisprudential thoughts on artificial intelligence legal system[J]. Law Review, 2001(5): 11-21.

[12] 季卫东. 人工智能时代的司法权之变[J]. 东方法学, 2018(1): 125-133.
JI W D. Change of judicial power in the age of artificial intelligence[J]. Oriental Law, 2018(1): 125-133.

[13] 郝铁川. 不可幻想和高估人工智能对法治的影响[N]. 法制日报, 2018-01-03.
HAO T C. Can't overestimate and fantasy the artificial intelligence influences on the rule of law[N]. Legal Daily, 2018-01-03.

[14] DAMASHEK M. Gauging similarity with n-grams: language-independent categorization of text[J]. Science, 1995, 267(5199): 843-848.

[15] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 3(3): 1083-1106.

[16] SAHAMI M, HEILMAN T D. A Web-based kernel function for measuring the similarity of short text snippets[C]//The 15th International Conference on World Wide Web, May 23-26, 2006, Edinburgh, Scotland. New York: ACM Press, 2006: 377-386.

[17] 吴帅, 潘海珍. 基于隐马尔可夫模型的中文分词[J]. 现代计算机, 2018(33): 27-30.
WU S, PAN H Z. Chinese word segmentation based on hidden Markov model[J]. Modern Computer, 2018(33): 27-30.

[18] Kambhatla, Nanda. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//The 42nd Annual Meeting of the Association for Computational Linguistics, July 21-26, 2004, Barcelona, Spain. [S.l.:s.n.], 2004.

[19] CHEN Y, ZHENG Q, ZHANG W. Omni-word feature and soft constraint for Chinese relation extraction[C]// The 52nd Annual Meeting of the Association for Computational Linguistics, June 22-27, 2014, Baltimore, USA. [S.l.:s.n.], 2014: 572-581.

[20] JACOBS D W, DAUME H, KUMAR A, et al. Generalized multiview analysis: a discriminative latent space[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, USA. Piscataway: IEEE Press, 2012: 2160-2167.

[21] ZHOU J T, TSANG I W, PAN S J, et al. Heterogeneous domain adaptation for multiple classes[C]//The 17th International Conference on Artificial Intelligence and Statistics, April 22-25, 2014, Reykjavik, Iceland. [S.l.:s.n.], 2014: 1095-1103.

[22] ZHU Y, CHEN Y, LU Z, et al. Heterogeneous transfer learning for image classification[C]// The 25th AAAI Conference on Artificial Intelligence, August 7-11, 2011, San Francisco, California. Palo Alto: AAAI Press, 2011: 1304-1309.

[23] LIU K, WEI S, ZHAO Y, et al. Accumulated reconstruction error vector (AREV): a semantic representation for cross-media retrieval[J]. Multimedia Tools and Applications, 2015, 74(2): 561-576.

[24] ZHANG X M, LI Z J, CHAO W H. Improving image tags by exploiting web search results[J]. Multimedia Tools and Applications, 2013, 62(3): 601-631.

[25] LIU M, ZHANG D. Pairwise constraint-guided sparse learning for feature selection[J]. IEEE Transactions on Cybernetics, 2017, 46(1): 298-310.

[26] 郭增茂. 知识管理中RS-CBR案例检索研究[D]. 郑州: 郑州大学, 2014.
GUO Z M. Research on RS-CBR case retrieval in knowledge management[D]. Zhengzhou: Zhengzhou University, 2014.

[27] 铁共. 大数据应用安全挑战与实践[J]. 大数据时代, 2018(4): 43-49, 42.
TIE G. Security challenges and practice of big data applications[J]. Big Data Time, 2018(4): 43-49, 42.

[28] 方贤进, 肖亚飞, 杨高明. 大数据及其隐私保护[J]. 大数据, 2017, 3(5): 45-56.
FANG X J, XIAO Y F, YANG G M. Privacy preserving in the age of big data[J]. Big Data Research, 2017, 3(5): 45-56.

[29] AMRO B, SAYGIN Y, LEVI A, et al. PA-CTM: privacy aware collaborative traffic monitoring system using autonomous location update mechanism[C]//The 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, November 1, 2011, Chicago, USA. New York: ACM Press, 2011: 1-8.

[30] GHINITA G, ZHAO K, PAPADIAS D, et al. A reciprocal framework for spatial K-anonymity[J]. Information Systems, 2010, 35(3): 299-314.

[31] WU S, WANG X, WANG S, et al. K-anonymity for crowdsourcing database[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2207-2221.

[32] 施惠娟. 可视化数据挖掘技术的研究与实现[D]. 上海: 华东师范大学, 2010.
SHI H J. Research and implementation of visual data mining technology[D]. Shanghai: East China Normal University, 2010.

[33] 汪加才, 陈奇, 赵杰煜, 等. VISMiner: 一个交互式可视化数据挖掘原型系统[J]. 计算机工程, 2003(1): 17-19.
WANG J C, CHEN Q, ZHAO J Y, et al. VISMiner: an interactive visual data mining prototyped system[J]. Computer Engineer, 2003(1): 17-19.

[34] 陈涛, 夏翠娟, 刘炜, 等. 关联数据的可视化技术研究与实现[J]. 图书情报工作, 2015, 59(17): 113-119.
CHEN T, XIA C J, LIU W, et al. Research and implementation of visualization technology for linked data[J]. Library and Information Service, 2015, 59(17): 113-119.

[35] 袁海, 陈康, 陶彩霞, 等. 基于中文文本的可视化技术研究[J]. 电信科学, 2014, 30(4): 114-122.
YUAN H, CHEN K, TAO C X, et al. Research on visualization techniques based on chinese texts[J]. Telecommunications Science, 2014, 30(4): 114-122.

[36] CHEN Y, LIU S, ZHENG Q, et al. Discovery of rare key phrases[C]//The 15th International Conference on e-Business Engineering(ICEBE), October 12-14, Xi'an, China. Piscataway: IEEE Press, 2018: 127-132.

作者简介

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治理与应用、多源数据融合与应用、企业信息化与电子政务。

作者简介

冯丽 (1993-), 女, 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、数据融合分析。

陈艳平 (1980-), 男, 博士, 贵州大学计算机科学与技术学院副教授, 主要研究方向为数据融合分析、自然语言处理、知识发现。

黄瑞章 (1979-), 女, 博士, 贵州大学计算机科学与技术学院副教授, 主要研究方向为数据融合分析、文本挖掘、网络挖掘、知识发现。

刘于雷 (1978-), 男, 多彩贵州网有限责任公司互联网大数据研发部负责人, 贵州省互联网大数据与传播应用研发中心副主任, 主要研究方向为互联网大数据应用、数据融合分析。

丁红发 (1988-), 男, 贵州大学贵州省公共大数据重点实验室博士生, 主要研究方向为隐私保护与数据安全、密码算法及应用。

收稿日期: 2019-01-29

基金项目: 国家自然科学基金资助项目 (No.91746116, No.U1836205) ; 贵州省重大应用基础研究基金资助项目 (No.JZ20142001) ; 贵州省科技重大专项计划基金资助项目 (No.[2017]3002) ; 贵州省自然科学基金资助项目 (No.[2018]1035)

Foundation Items: The National Natural Science Foundation of China(No.91746116, No.U1836205), The Major Applied Basic Research Program of Guizhou Province(No.JZ20142001), Special Science and Technology Projects of Guizhou Province (No.[2017]3002), The Natural Science Foundation of Guizhou Province(No.[2018]1035)