(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111932413 B
(45)授权公告日 2021.01.12

(21)申请号 202010957947.7

(22)申请日 2020.09.14

(65)同一申请的已公布的文献号
申请公布号 CN 111932413 A

(43)申请公布日 2020.11.13

(73)专利权人 平安国际智慧城市科技股份有限公司
地址 518000 广东省深圳市前海深港合作区妈湾兴海大道3048号前海自贸大厦1-34层

(72)发明人 娟娟 肖畅

(74)专利代理机构 深圳市赛恩倍吉知识产权代理有限公司 44334
代理人 何春兰 迟珊珊

(51)Int.Cl.
G06Q 50/18(2012.01)
G06F 16/27(2019.01)
G06F 16/34(2019.01)
G06F 21/64(2013.01)
G06N 3/04(2006.01)
G06F 40/174(2020.01)
G06F 40/186(2020.01)

(56)对比文件
CN 111475513 A,2020.07.31
CN 109657039 A,2019.04.19
CN 107679234 B,2020.02.11
CN 108268444 A,2018.07.10
US 10169315 B1,2019.01.01
US 2019065467 A1,2019.02.28

审查员 赵静

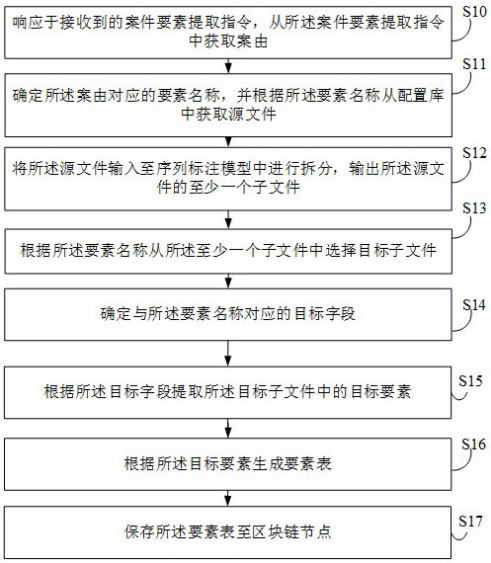
权利要求书2页 说明书12页 附图2页

(54)发明名称

案件要素提取方法、装置、设备及介质

(57)摘要

本申请涉及人工智能及数据处理技术领域，提供一种案件要素提取方法、装置、设备及介质，其方法能够从案件要素提取指令中获取案由，确定案由对应的要素名称以从配置库中获取源文件，以确保源文件的全面性，将源文件输入至序列标注模型中进行拆分，输出子文件，并根据要素名称选择目标子文件，以缩小要素提取的范围，提高要素提取的效率，确定与要素名称对应的目标字段以提取目标要素，并根据目标要素生成要素表，以便基于人工智能手段自动提取要素，并将提取的要素生成要素表，释放人力，同时有效保证了数据处理的高效性及准确性，进一步保存要素表至区块链节点，以基于区块链技术保证数据的安全性及隐私性，避免数据被恶意篡改。



CN 111932413 B

1.一种案件要素提取方法,其特征在于,所述案件要素提取方法包括:

响应于接收到的案件要素提取指令,其中,从所述案件要素提取指令中获取案由;

确定所述案由对应的要素名称,并根据所述要素名称从配置库中获取源文件;

将所述源文件输入至序列标注模型中进行拆分,输出所述源文件的至少一个子文件,其中,在将所述源文件输入至序列标注模型中进行拆分前,获取历史案由数据;根据所述历史案由数据配置拆分模式;将所述拆分模式添加到Bi-LSTM+CRF模型中,得到所述序列标注模型;

根据所述要素名称从所述至少一个子文件中选择目标子文件;

确定与所述要素名称对应的目标字段;

根据所述目标字段提取所述目标子文件中的目标要素;

根据所述目标要素生成要素表;

保存所述要素表至区块链节点。

2.如权利要求1所述的案件要素提取方法,其特征在于,在将所述源文件输入至序列标注模型中进行拆分前,所述案件要素提取方法还包括:

检测所述源文件的文件格式;

当所述源文件的文件格式为PDF格式或者jpg格式时,对所述源文件进行OCR识别。

3.如权利要求1所述的案件要素提取方法,其特征在于,所述将所述源文件输入至序列标注模型中进行拆分,输出所述源文件的至少一个子文件包括:

利用所述序列标注模型抽取所述源文件的子要素;

将所述子要素与所述拆分模式进行匹配,确定目标拆分模式;

根据所述目标拆分模式对所述源文件进行拆分,输出至少一个页码范围的文件及每个页码范围的文件对应的文件名称;

以所述至少一个页码范围的文件及每个页码范围的文件对应的文件名称生成所述至少一个子文件。

4.如权利要求1所述的案件要素提取方法,其特征在于,所述根据所述目标字段提取所述目标子文件中的目标要素包括:

采用正则表达式以所述目标字段在所述目标子文件中进行匹配,以提取所述目标子文件中的目标要素;

当检测到有目标子文件匹配失败时,采用NER模型,并基于所述目标字段对检测到的目标子文件进行识别,以提取所述检测到的目标子文件中的目标要素。

5.如权利要求1所述的案件要素提取方法,其特征在于,所述根据所述目标要素生成要素表包括:

获取预先配置的表单模板;

识别所述表单模板对应的要素匹配规则;

根据所述要素匹配规则将所述目标要素填充至所述表单模板,生成所述要素表。

6.如权利要求1所述的案件要素提取方法,其特征在于,在保存所述要素表至区块链节点前,所述案件要素提取方法还包括:

将所述要素表反馈至校验平台;

接收所述校验平台返回的要素表,并保存返回的要素表至区块链节点。

7.一种案件要素提取装置,其特征在于,所述案件要素提取装置包括:

获取单元,用于响应于接收到的案件要素提取指令,其中,从所述案件要素提取指令中获取案由;

所述获取单元,还用于确定所述案由对应的要素名称,并根据所述要素名称从配置库中获取源文件;

拆分单元,用于将所述源文件输入至序列标注模型中进行拆分,输出所述源文件的至少一个子文件,其中,在将所述源文件输入至序列标注模型中进行拆分前,获取历史案由数据;根据所述历史案由数据配置拆分模式;将所述拆分模式添加到Bi-LSTM+CRF模型中,得到所述序列标注模型;

选择单元,用于根据所述要素名称从所述至少一个子文件中选择目标子文件;

确定单元,用于确定与所述要素名称对应的目标字段;

提取单元,用于根据所述目标字段提取所述目标子文件中的目标要素;

生成单元,用于根据所述目标要素生成要素表;

保存单元,用于保存所述要素表至区块链节点。

8.一种电子设备,其特征在于,所述电子设备包括:

存储器,存储至少一个指令;及

处理器,执行所述存储器中存储的指令以实现如权利要求1至6中任意一项所述的案件要素提取方法。

9.一种计算机可读存储介质,其特征在于:所述计算机可读存储介质中存储有至少一个指令,所述至少一个指令被电子设备中的处理器执行以实现如权利要求1至6中任意一项所述的案件要素提取方法。

案件要素提取方法、装置、设备及介质

技术领域

[0001] 本发明涉及人工智能及智能决策技术领域，尤其涉及一种案件要素提取方法、装置、设备及介质。

背景技术

[0002] 目前，法官在审理案件的时候为了快速、准确地审理案件，通常会对影响案件定性的要素及影响案件审理的程序性事项重点关注。而每个审理的案件均有差异性，意味着法官每审理一个案件，就需要翻阅所有案件材料进行阅读并核对要素，耗费了大量的时间，并且，因法官人数有限，行业内还普遍存在案多人少的情况。

[0003] 另外，现有技术中还无法做到辅助法官对案件要素进行自动提取。

发明内容

[0004] 鉴于以上内容，有必要提供一种案件要素提取方法、装置、设备及介质，能够基于人工智能手段全面地自动提取要素，并生成要素表，释放人力，同时有效保证了数据处理的高效性及准确性。

[0005] 一种案件要素提取方法，所述案件要素提取方法包括：

[0006] 响应于接收到的案件要素提取指令，其中，从所述案件要素提取指令中获取案由；

[0007] 确定所述案由对应的要素名称，并根据所述要素名称从配置库中获取源文件；

[0008] 将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件；

[0009] 根据所述要素名称从所述至少一个子文件中选择目标子文件；

[0010] 确定与所述要素名称对应的目标字段；

[0011] 根据所述目标字段提取所述目标子文件中的目标要素；

[0012] 根据所述目标要素生成要素表；

[0013] 保存所述要素表至区块链节点。

[0014] 根据本发明优选实施例，在将所述源文件输入至序列标注模型中进行拆分前，所述案件要素提取方法还包括：

[0015] 检测所述源文件的文件格式；

[0016] 当所述源文件的文件格式为PDF格式或者jpg格式时，对所述源文件进行OCR识别。

[0017] 根据本发明优选实施例，在将所述源文件输入至序列标注模型中进行拆分前，所述案件要素提取方法还包括：

[0018] 获取历史案由数据；

[0019] 根据所述历史案由数据配置拆分模式；

[0020] 将所述拆分模式添加到Bi-LSTM+CRF模型中，得到所述序列标注模型。

[0021] 根据本发明优选实施例，所述将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件包括：

[0022] 利用所述序列标注模型抽取所述源文件的子要素；

[0023] 将所述子要素与所述拆分模式进行匹配，确定目标拆分模式；

[0024] 根据所述目标拆分模式对所述源文件进行拆分，输出至少一个页码范围的文件及每个页码范围的文件对应的文件名称；

[0025] 以所述至少一个页码范围的文件及每个页码范围的文件对应的文件名称生成所述至少一个子文件。

[0026] 根据本发明优选实施例，所述根据所述目标字段提取所述目标子文件中的目标要素包括：

[0027] 采用正则表达式以所述目标字段在所述目标子文件中进行匹配，以提取所述目标子文件中的目标要素；

[0028] 当检测到有目标子文件匹配失败时，采用NER模型，并基于所述目标字段对检测到的目标子文件进行识别，以提取所述检测到的目标子文件中的目标要素。

[0029] 根据本发明优选实施例，所述根据所述目标要素生成要素表包括：

[0030] 获取预先配置的表单模板；

[0031] 识别所述表单模板对应的要素匹配规则；

[0032] 根据所述要素匹配规则将所述目标要素填充至所述表单模板，生成所述要素表。

[0033] 根据本发明优选实施例，在保存所述要素表至区块链节点前，所述案件要素提取方法还包括：

[0034] 将所述要素表反馈至校验平台；

[0035] 接收所述校验平台返回的要素表，并保存返回的要素表至区块链节点。

[0036] 一种案件要素提取装置，所述案件要素提取装置包括：

[0037] 获取单元，用于响应于接收到的案件要素提取指令，其中，从所述案件要素提取指令中获取案由；

[0038] 所述获取单元，还用于确定所述案由对应的要素名称，并根据所述要素名称从配置库中获取源文件；

[0039] 拆分单元，用于将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件；

[0040] 选择单元，用于根据所述要素名称从所述至少一个子文件中选择目标子文件；

[0041] 确定单元，用于确定与所述要素名称对应的目标字段；

[0042] 提取单元，用于根据所述目标字段提取所述目标子文件中的目标要素；

[0043] 生成单元，用于根据所述目标要素生成要素表；

[0044] 保存单元，用于保存所述要素表至区块链节点。

[0045] 一种电子设备，所述电子设备包括：

[0046] 存储器，存储至少一个指令；及

[0047] 处理器，执行所述存储器中存储的指令以实现所述案件要素提取方法。

[0048] 一种计算机可读存储介质，所述计算机可读存储介质中存储有至少一个指令，所述至少一个指令被电子设备中的处理器执行以实现所述案件要素提取方法。

[0049] 由以上技术方案可以看出，本发明能够响应于接收到的案件要素提取指令，从所述案件要素提取指令中获取案由，确定所述案由对应的要素名称，并根据所述要素名称从

配置库中获取源文件。首先获取到所有可能与所述案由相关的源文件供后续使用，以确保所述源文件的全面性；再将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件，根据所述要素名称从所述至少一个子文件中选择目标子文件，以缩小要素提取的范围，提高要素提取的效率；之后确定与所述要素名称对应的目标字段，根据所述目标字段提取所述目标子文件中的目标要素，并根据所述目标要素生成要素表，以便基于人工智能的手段自动提取要素，并将提取的要素生成要素表，释放人力，同时有效保证了数据处理的高效性及准确性，进一步保存所述要素表至区块链节点，以基于区块链技术保证数据的安全性及隐私性，避免数据被恶意篡改。

附图说明

[0050] 图1是本发明案件要素提取方法的较佳实施例的流程图。

[0051] 图2是本发明案件要素提取装置的较佳实施例的功能模块图。

[0052] 图3是本发明实现案件要素提取方法的较佳实施例的电子设备的结构示意图。

具体实施方式

[0053] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0054] 如图1所示，是本发明案件要素提取方法的较佳实施例的流程图。根据不同的需求，该流程图中步骤的顺序可以改变，某些步骤可以省略。

[0055] 所述案件要素提取方法应用于一个或者多个电子设备中，所述电子设备是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit，ASIC)、可编程门阵列(Field－Programmable Gate Array，FPGA)、数字处理器(Digital Signal Processor，DSP)、嵌入式设备等。

[0056] 所述电子设备可以是任何一种可与用户进行人机交互的电子产品，例如，个人计算机、平板电脑、智能手机、个人数字助理(Personal Digital Assistant，PDA)、游戏机、交互式网络电视(Internet Protocol Television，IPTV)、智能式穿戴式设备等。

[0057] 所述电子设备还可以包括网络设备和/或用户设备。其中，所述网络设备包括，但不限于单个网络服务器、多个网络服务器组成的服务器组或基于云计算(Cloud Computing)的由大量主机或网络服务器构成的云。

[0058] 所述电子设备所处的网络包括但不限于互联网、广域网、城域网、局域网、虚拟专用网络(Virtual Private Network，VPN)等。

[0059] S10，响应于接收到的案件要素提取指令，其中，从所述案件要素提取指令中获取案由。

[0060] 在本实施例中，所述案件要素提取指令可以由相关工作人员触发，以实时满足用户需求。

[0061] 所述案由可以是与法律相关的各种案件，例如：信用卡纠纷、金融借款纠纷、机动车交通事故责任纠纷等。

[0062] S11，确定所述案由对应的要素名称，并根据所述要素名称从配置库中获取源文

件。

[0063] 在本发明的至少一个实施例中,所述要素名称是指所述案由包含的案件要点关键词。

[0064] 例如:所述要素名称可以包括:原告名称、被告名称、诉讼请求、本金金额等。

[0065] 在本实施例中,所述配置库是指存储各种案由的相关资料的数据库,所述配置库可以是一个或者多个。

[0066] 进一步地,所述电子设备从所述配置库中获取包含所述要素名称的所有文件作为所述源文件,以首先获取到所有可能与所述案由相关的源文件供后续使用,以确保所述源文件的全面性。

[0067] S12,将所述源文件输入至序列标注模型中进行拆分,输出所述源文件的至少一个子文件。

[0068] 可以理解的是,在一份源文件中,由于一个案由对应的个案或批量案件从立案到审判到结案中间要经过多个不同的流程,因此,一份源文件通常还可以包括多种类型的小文件。

[0069] 例如:在一份源文件中,可以同时包括判决书、起诉状、答辩状、证据材料等多种不同类型的文件。

[0070] 可以理解的是,所述源文件并不限于文本格式,可以是PDF格式或者jpg格式等,而对于非文本格式的源文件,由于无法直接被所述序列标注模型处理,因此,还需要对所述源文件进行文字识别。

[0071] 具体地,在将所述源文件输入至序列标注模型中进行拆分前,所述案件要素提取方法还包括:

[0072] 检测所述源文件的文件格式;

[0073] 当所述源文件的文件格式为PDF格式或者jpg格式时,对所述源文件进行OCR(Optical Character Recognition,光学字符识别)识别。

[0074] 在本发明的至少一个实施例中,所述电子设备还需要首先训练所述序列标注模型,所述序列标注模型为Bi-LSTM+CRF模型。

[0075] 具体地,在将所述源文件输入至序列标注模型中进行拆分前,所述案件要素提取方法还包括:

[0076] 获取历史案由数据;

[0077] 根据所述历史案由数据配置拆分模式;

[0078] 将所述拆分模式添加到Bi-LSTM+CRF模型中,得到所述序列标注模型。

[0079] 例如:根据获取的历史案由数据,确定处于结构模式A,且带有关键字“起诉书”的为《起诉书》的首页,则对应的拆分模式可以配置为:获取与所述结构模式A相匹配,且带有关键字“起诉书”的页面,并将所述页面确定为《起诉书》的首页,记录该页面的页码为《起诉书》的起始页码。

[0080] 需要说明的是,Bi-LSTM(Bidirectional Long Short Term Memory,双向长短时记忆)层提供了长距离的依赖建模,加强了每个字符与上下文字符间的联系,而CRF(conditional random field,条件随机场)可以容纳任意的上下文信息,特征设计灵活,CRF层可以构建字符间的特征转移及对应关系,并同时考虑输出标签之间的顺序性,进而达

到更准确的识别与拆分效果。

[0081] 在本发明的至少一个实施例中，所述将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件包括：

[0082] 利用所述序列标注模型抽取所述源文件的子要素；

[0083] 将所述子要素与所述拆分模式进行匹配，确定目标拆分模式；

[0084] 根据所述目标拆分模式对所述源文件进行拆分，输出至少一个页码范围的文件及每个页码范围的文件对应的文件名称；

[0085] 以所述至少一个页码范围的文件及每个页码范围的文件对应的文件名称生成所述至少一个子文件。

[0086] 例如：当抽取到子要素“起诉状”时，获取与所述“起诉状”对应的拆分模式，并根据所述拆分模式拆分所述源文件，当确定所述“起诉状”的起始页为15页，结束页为25页时，以所述“起诉状”为文件名称，并以所述源文件的第15-25页作为所述“起诉状”的页码范围，形成《起诉状》作为所述源文件的其中一个子文件。

[0087] 通过上述实施方式，即可利用序列标注模型将所述源文件自动拆分为多个子文件，且得到的子文件拆分的准确度较高。

[0088] S13，根据所述要素名称从所述至少一个子文件中选择目标子文件。

[0089] 可以理解的是，一份源文件的覆盖面很广，并非每个子文件都是需要被使用的，因此，所述电子设备根据所述要素名称从所述至少一个子文件中选择目标子文件，以缩小要素提取的范围，提高要素提取的效率。

[0090] S14，确定与所述要素名称对应的目标字段。

[0091] 例如：所述目标字段可以是本金金额、利息、年利率等。

[0092] 所述目标字段用于定位及获取到所述案由项下案件的要素。

[0093] S15，根据所述目标字段提取所述目标子文件中的目标要素。

[0094] 在本实施例中，以所述目标字段在所述目标子文件中进行定位匹配，即可获取到所述目标子文件中的目标要素。

[0095] 具体地，所述根据所述目标字段提取所述目标子文件中的目标要素包括：

[0096] 采用正则表达式以所述目标字段在所述目标子文件中进行匹配，以提取所述目标子文件中的目标要素；

[0097] 当检测到有目标子文件匹配失败时，采用NER模型(Named Entity Recognition，命名实体识别)，并基于所述目标字段对检测到的目标子文件进行识别，以提取所述检测到的目标子文件中的目标要素。

[0098] 通过上述实施方式，结合了正则表达式与NER模型提取所述目标子文件中的目标要素，首先利用正则表达式快速且准确的特性，采用正则表达式先执行要素提取，并在正则表达式提取失败时，再基于NER模型进行二次提取，以提高要素提取的全面性及召回率，有效避免了要素的遗漏。

[0099] S16，根据所述目标要素生成要素表。

[0100] 在本发明的至少一个实施例中，所述根据所述目标要素生成要素表包括：

[0101] 获取预先配置的表单模板；

[0102] 识别所述表单模板对应的要素匹配规则；

[0103] 根据所述要素匹配规则将所述目标要素填充至所述表单模板，生成所述要素表。

[0104] 例如：上述要素表可以是一份裁判文书，在执行了前述的要素采集及梳理后，辅助法官生成最后的裁判文书。具体地，根据前端针对不同案件梳理的裁判文书模板，将提取的共性信息及当事人信息等前端已有的目标要素对号入座到裁判文书模板中，以便自动形成一份裁判文书供法官参考，省却了大量的重复信息填入工作，提升了法官文书的写作效率。

[0105] S17，保存所述要素表至区块链节点。

[0106] 通过上述实施方式，能够保证数据的安全性及隐私性，避免数据被恶意篡改。

[0107] 在本发明的至少一个实施例中，在保存所述要素表至区块链节点前，所述案件要素提取方法还包括：

[0108] 将所述要素表反馈至校验平台；

[0109] 接收所述校验平台返回的要素表，并保存返回的要素表至区块链节点。

[0110] 通过上述实施方式，能够在保存前首先对所述要素表进行校验，避免所述要素表生成错误。

[0111] 需要说明的是，本发明对所述校验平台的校验方式不做限制，所述校验平台可以采用人工校验，也可以采用机器自动校验。

[0112] 由以上技术方案可以看出，本发明能够响应于接收到的案件要素提取指令，从所述案件要素提取指令中获取案由，确定所述案由对应的要素名称，并根据所述要素名称从配置库中获取源文件。首先获取到所有可能与所述案由相关的源文件供后续使用，以确保所述源文件的全面性；再将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件，根据所述要素名称从所述至少一个子文件中选择目标子文件，以缩小要素提取的范围，提高要素提取的效率；之后确定与所述要素名称对应的目标字段，根据所述目标字段提取所述目标子文件中的目标要素，并根据所述目标要素生成要素表，以便基于人工智能的手段自动提取要素，并将提取的要素生成要素表，释放人力，同时有效保证了数据处理的高效性及准确性，进一步保存所述要素表至区块链节点，以基于区块链技术保证数据的安全性及隐私性，避免数据被恶意篡改。

[0113] 如图2所示，是本发明案件要素提取装置的较佳实施例的功能模块图。所述案件要素提取装置11包括获取单元110、拆分单元111、选择单元112、确定单元113、提取单元114、生成单元115、保存单元116、检测单元117、识别单元118、配置单元119、添加单元120、反馈单元121。本发明所称的模块/单元是指一种能够被处理器13所执行，并且能够完成固定功能的一系列计算机程序段，其存储在存储器12中。在本实施例中，关于各模块/单元的功能将在后续的实施例中详述。

[0114] 响应于接收到的案件要素提取指令，其中，获取单元110从所述案件要素提取指令中获取案由。

[0115] 在本实施例中，所述案件要素提取指令可以由相关工作人员触发，以实时满足用户需求。

[0116] 所述案由可以是与法律相关的各种案件，例如：信用卡纠纷、金融借款纠纷、机动车交通事故责任纠纷等。

[0117] 所述获取单元110确定所述案由对应的要素名称，并根据所述要素名称从配置库中获取源文件。

[0118] 在本发明的至少一个实施例中，所述要素名称是指所述案由包含的案件要点关键词。

[0119] 例如：所述要素名称可以包括：原告名称、被告名称、诉讼请求、本金金额等。

[0120] 在本实施例中，所述配置库是指存储各种案由的相关资料的数据库，所述配置库可以是一个或者多个。

[0121] 进一步地，所述获取单元110从所述配置库中获取包含所述要素名称的所有文件作为所述源文件，以首先获取到所有可能与所述案由相关的源文件供后续使用，以确保所述源文件的全面性。

[0122] 拆分单元111将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件。

[0123] 可以理解的是，在一份源文件中，由于一个案由对应的个案或批量案件从立案到审判到结案中间要经过多个不同的流程，因此，一份源文件通常还可以包括多种类型的小文件。

[0124] 例如：在一份源文件中，可以同时包括判决书、起诉状、答辩状、证据材料等多种不同类型的文件。

[0125] 可以理解的是，所述源文件并不限于文本格式，可以是PDF格式或者jpg格式等，而对于非文本格式的源文件，由于无法直接被所述序列标注模型处理，因此，还需要对所述源文件进行文字识别。

[0126] 具体地，在将所述源文件输入至序列标注模型中进行拆分前，检测单元117检测所述源文件的文件格式；

[0127] 当所述源文件的文件格式为PDF格式或者jpg格式时，识别单元118对所述源文件进行OCR(Optical Character Recognition，光学字符识别)识别。

[0128] 在本发明的至少一个实施例中，还需要首先训练所述序列标注模型，所述序列标注模型为Bi-LSTM+CRF模型。

[0129] 具体地，在将所述源文件输入至序列标注模型中进行拆分前，所述获取单元110获取历史案由数据；

[0130] 配置单元119根据所述历史案由数据配置拆分模式；

[0131] 添加单元120将所述拆分模式添加到Bi-LSTM+CRF模型中，得到所述序列标注模型。

[0132] 例如：根据获取的历史案由数据，确定处于结构模式A，且带有关键字“起诉书”的为《起诉书》的首页，则对应的拆分模式可以配置为：获取与所述结构模式A相匹配，且带有关键字“起诉书”的页面，并将所述页面确定为《起诉书》的首页，记录该页面的页码为《起诉书》的起始页码。

[0133] 需要说明的是，Bi-LSTM(Bidirectional Long Short Term Memory，双向长短时记忆)层提供了长距离的依赖建模，加强了每个字符与上下文字符间的联系，而CRF(conditional random field，条件随机场)可以容纳任意的上下文信息，特征设计灵活，CRF层可以构建字符间的特征转移及对应关系，并同时考虑输出标签之间的顺序性，进而达到更准确的识别与拆分效果。

[0134] 在本发明的至少一个实施例中，所述拆分单元111将所述源文件输入至序列标注

模型中进行拆分，输出所述源文件的至少一个子文件包括：

[0135] 利用所述序列标注模型抽取所述源文件的子要素；

[0136] 将所述子要素与所述拆分模式进行匹配，确定目标拆分模式；

[0137] 根据所述目标拆分模式对所述源文件进行拆分，输出至少一个页码范围的文件及每个页码范围的文件对应的文件名称；

[0138] 以所述至少一个页码范围的文件及每个页码范围的文件对应的文件名称生成所述至少一个子文件。

[0139] 例如：当抽取到子要素“起诉状”时，获取与所述“起诉状”对应的拆分模式，并根据所述拆分模式拆分所述源文件，当确定所述“起诉状”的起始页为15页，结束页为25页时，以所述“起诉状”为文件名称，并以所述源文件的第15-25页作为所述“起诉状”的页码范围，形成《起诉状》作为所述源文件的其中一个子文件。

[0140] 通过上述实施方式，即可利用序列标注模型将所述源文件自动拆分为多个子文件，且得到的子文件拆分的准确度较高。

[0141] 选择单元112根据所述要素名称从所述至少一个子文件中选择目标子文件。

[0142] 可以理解的是，一份源文件的覆盖面很广，并非每个子文件都是需要被使用的，因此，所述电子设备根据所述要素名称从所述至少一个子文件中选择目标子文件，以缩小要素提取的范围，提高要素提取的效率。

[0143] 确定单元113确定与所述要素名称对应的目标字段。

[0144] 例如：所述目标字段可以是本金金额、利息、年利率等。

[0145] 所述目标字段用于定位及获取到所述案由项下案件的要素。

[0146] 提取单元114根据所述目标字段提取所述目标子文件中的目标要素。

[0147] 在本实施例中，以所述目标字段在所述目标子文件中进行定位匹配，即可获取到所述目标子文件中的目标要素。

[0148] 具体地，所述提取单元114根据所述目标字段提取所述目标子文件中的目标要素包括：

[0149] 采用正则表达式以所述目标字段在所述目标子文件中进行匹配，以提取所述目标子文件中的目标要素；

[0150] 当检测到有目标子文件匹配失败时，采用NER模型(Named Entity Recognition，命名实体识别)，并基于所述目标字段对检测到的目标子文件进行识别，以提取所述检测到的目标子文件中的目标要素。

[0151] 通过上述实施方式，结合了正则表达式与NER模型提取所述目标子文件中的目标要素，首先利用正则表达式快速且准确的特性，采用正则表达式先执行要素提取，并在正则表达式提取失败时，再基于NER模型进行二次提取，以提高要素提取的全面性及召回率，有效避免了要素的遗漏。

[0152] 生成单元115根据所述目标要素生成要素表。

[0153] 在本发明的至少一个实施例中，所述生成单元115根据所述目标要素生成要素表包括：

[0154] 获取预先配置的表单模板；

[0155] 识别所述表单模板对应的要素匹配规则；

[0156] 根据所述要素匹配规则将所述目标要素填充至所述表单模板，生成所述要素表。

[0157] 例如：上述要素表可以是一份裁判文书，在执行了前述的要素采集及梳理后，辅助法官生成最后的裁判文书。具体地，根据前端针对不同案件梳理的裁判文书模板，将提取的共性信息及当事人信息等前端已有的目标要素对号入座到裁判文书模板中，以便自动形成一份裁判文书供法官参考，省却了大量的重复信息填入工作，提升了法官文书的写作效率。

[0158] 保存单元116保存所述要素表至区块链节点。

[0159] 通过上述实施方式，能够保证数据的安全性及隐私性，避免数据被恶意篡改。

[0160] 在本发明的至少一个实施例中，在保存所述要素表至区块链节点前，反馈单元121将所述要素表反馈至校验平台；

[0161] 保存单元116接收所述校验平台返回的要素表，并保存返回的要素表至区块链节点。

[0162] 通过上述实施方式，能够在保存前首先对所述要素表进行校验，避免所述要素表生成错误。

[0163] 需要说明的是，本发明对所述校验平台的校验方式不做限制，所述校验平台可以采用人工校验，也可以采用机器自动校验。

[0164] 由以上技术方案可以看出，本发明能够响应于接收到的案件要素提取指令，从所述案件要素提取指令中获取案由，确定所述案由对应的要素名称，并根据所述要素名称从配置库中获取源文件。首先获取到所有可能与所述案由相关的源文件供后续使用，以确保所述源文件的全面性；再将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件，根据所述要素名称从所述至少一个子文件中选择目标子文件，以缩小要素提取的范围，提高要素提取的效率；之后确定与所述要素名称对应的目标字段，根据所述目标字段提取所述目标子文件中的目标要素，并根据所述目标要素生成要素表，以便基于人工智能的手段自动提取要素，并将提取的要素生成要素表，释放人力，同时有效保证了数据处理的高效性及准确性，进一步保存所述要素表至区块链节点，以基于区块链技术保证数据的安全性及隐私性，避免数据被恶意篡改。

[0165] 如图3所示，是本发明实现案件要素提取方法的较佳实施例的电子设备的结构示意图。

[0166] 所述电子设备1可以包括存储器12、处理器13和总线，还可以包括存储在所述存储器12中并可在所述处理器13上运行的计算机程序，例如案件要素提取程序。

[0167] 本领域技术人员可以理解，所述示意图仅仅是电子设备1的示例，并不构成对电子设备1的限定，所述电子设备1既可以是总线型结构，也可以是星形结构，所述电子设备1还可以包括比图示更多或更少的其他硬件或者软件，或者不同的部件布置，例如所述电子设备1还可以包括输入输出设备、网络接入设备等。

[0168] 需要说明的是，所述电子设备1仅为举例，其他现有的或今后可能出现的电子产品如可适应于本发明，也应包含在本发明的保护范围以内，并以引用方式包含于此。

[0169] 其中，存储器12至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、移动硬盘、多媒体卡、卡型存储器(例如：SD或DX存储器等)、磁性存储器、磁盘、光盘等。存储器12在一些实施例中可以是电子设备1的内部存储单元，例如该电子设备1的移动硬盘。存储器12在另一些实施例中也可以是电子设备1的外部存储设备，例如电子设备1上配备的插

接式移动硬盘、智能存储卡(Smart Media Card, SMC)、安全数字(Secure Digital, SD)卡、闪存卡(Flash Card)等。进一步地,存储器12还可以既包括电子设备1的内部存储单元也包括外部存储设备。存储器12不仅可以用于存储安装于电子设备1的应用软件及各类数据,例如案件要素提取程序的代码等,还可以用于暂时地存储已经输出或者将要输出的数据。

[0170] 处理器13在一些实施例中可以由集成电路组成,例如可以由单个封装的集成电路所组成,也可以是由多个相同功能或不同功能封装的集成电路所组成,包括一个或者多个中央处理器(Central Processing unit,CPU)、微处理器、数字处理芯片、图形处理器及各种控制芯片的组合等。处理器13是所述电子设备1的控制核心(Control Unit),利用各种接口和线路连接整个电子设备1的各个部件,通过运行或执行存储在所述存储器12内的程序或者模块(例如执行案件要素提取程序等),以及调用存储在所述存储器12内的数据,以执行电子设备1的各种功能和处理数据。

[0171] 所述处理器13执行所述电子设备1的操作系统以及安装的各类应用程序。所述处理器13执行所述应用程序以实现上述各个案件要素提取方法实施例中的步骤,例如图1所示的步骤。

[0172] 示例性的,所述计算机程序可以被分割成一个或多个模块/单元,所述一个或者多个模块/单元被存储在所述存储器12中,并由所述处理器13执行,以完成本发明。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机程序指令段,该指令段用于描述所述计算机程序在所述电子设备1中的执行过程。例如,所述计算机程序可以被分割成获取单元110、拆分单元111、选择单元112、确定单元113、提取单元114、生成单元115、保存单元116、检测单元117、识别单元118、配置单元119、添加单元120、反馈单元121。

[0173] 或者,所述处理器13执行所述计算机程序时实现上述各装置实施例中各模块/单元的功能,例如:

[0174] 响应于接收到的案件要素提取指令,其中,从所述案件要素提取指令中获取案由;

[0175] 确定所述案由对应的要素名称,并根据所述要素名称从配置库中获取源文件;

[0176] 将所述源文件输入至序列标注模型中进行拆分,输出所述源文件的至少一个子文件;

[0177] 根据所述要素名称从所述至少一个子文件中选择目标子文件;

[0178] 确定与所述要素名称对应的目标字段;

[0179] 根据所述目标字段提取所述目标子文件中的目标要素;

[0180] 根据所述目标要素生成要素表;

[0181] 保存所述要素表至区块链节点。

[0182] 上述以软件功能模块的形式实现的集成的单元,可以存储在一个计算机可读取存储介质中。上述软件功能模块存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机、计算机设备,或者网络设备等)或处理器(processor)执行本发明各个实施例所述案件要素提取方法的部分。

[0183] 所述电子设备1集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明实现上述实施例方法中的全部或部分流程,也可以通过计算机程序来指示相关的硬件设备来

完成,所述的计算机程序可存储于一计算机可读存储介质中,该计算机程序在被处理器执行时,可实现上述各个方法实施例的步骤。

[0184] 其中,所述计算机程序包括计算机程序代码,所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM,Read-Only Memory)。

[0185] 进一步地,计算机可用存储介质可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序等;存储数据区可存储根据区块链节点的使用所创建的数据等。

[0186] 本发明所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain),本质上是一个去中心化的数据库,是一串使用密码学方法相关联产生的数据块,每一个数据块中包含了一批次网络交易的信息,用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0187] 总线可以是外设部件互连标准(peripheral component interconnect,简称PCI)总线或扩展工业标准结构(extended industry standard architecture,简称EISA)总线等。该总线可以分为地址总线、数据总线、控制总线等。为便于表示,在图3中仅用一根箭头表示,但并不表示仅有一根总线或一种类型的总线。所述总线被设置为实现所述存储器12以及至少一个处理器13等之间的连接通信。

[0188] 尽管未示出,所述电子设备1还可以包括给各个部件供电的电源(比如电池),优选地,电源可以通过电源管理装置与所述至少一个处理器13逻辑相连,从而通过电源管理装置实现充电管理、放电管理、以及功耗管理等功能。电源还可以包括一个或一个以上的直流或交流电源、再充电装置、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。所述电子设备1还可以包括多种传感器、蓝牙模块、Wi-Fi模块等,在此不再赘述。

[0189] 进一步地,所述电子设备1还可以包括网络接口,可选地,所述网络接口可以包括有线接口和/或无线接口(如WI-FI接口、蓝牙接口等),通常用于在该电子设备1与其他电子设备之间建立通信连接。

[0190] 可选地,该电子设备1还可以包括用户接口,用户接口可以是显示器(Display)、输入单元(比如键盘(Keyboard)),可选地,用户接口还可以是标准的有线接口、无线接口。可选地,在一些实施例中,显示器可以是LED显示器、液晶显示器、触控式液晶显示器以及OLED(Organic Light-Emitting Diode,有机发光二极管)触摸器等。其中,显示器也可以适当的称为显示屏或显示单元,用于显示在电子设备1中处理的信息以及用于显示可视化的用户界面。

[0191] 应该了解,所述实施例仅为说明之用,在专利申请范围上并不受此结构的限制。

[0192] 图3仅示出了具有组件12-13的电子设备1,本领域技术人员可以理解的是,图3示出的结构并不构成对所述电子设备1的限定,可以包括比图示更少或者更多的部件,或者组合某些部件,或者不同的部件布置。

[0193] 结合图1,所述电子设备1中的所述存储器12存储多个指令以实现一种案件要素提取方法,所述处理器13可执行所述多个指令从而实现:

[0194] 响应于接收到的案件要素提取指令，其中，从所述案件要素提取指令中获取案由；

[0195] 确定所述案由对应的要素名称，并根据所述要素名称从配置库中获取源文件；

[0196] 将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件；

[0197] 根据所述要素名称从所述至少一个子文件中选择目标子文件；

[0198] 确定与所述要素名称对应的目标字段；

[0199] 根据所述目标字段提取所述目标子文件中的目标要素；

[0200] 根据所述目标要素生成要素表；

[0201] 保存所述要素表至区块链节点。

[0202] 具体地，所述处理器13对上述指令的具体实现方法可参考图1对应实施例中相关步骤的描述，在此不赘述。

[0203] 在本发明所提供的几个实施例中，应该理解到，所揭露的系统，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

[0204] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0205] 另外，在本发明各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

[0206] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。

[0207] 因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

[0208] 此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第二等词语用来表示名称，而并不表示任何特定的顺序。

[0209] 最后应说明的是，以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或等同替换，而不脱离本发明技术方案的精神和范围。


响应于接收到的案件要素提取指令，从所述案件要素提取指令中获取案由
S10
确定所述案由对应的要素名称，并根据所述要素名称从配置库中获取源文件
S11
将所述源文件输入至序列标注模型中进行拆分，输出所述源文件的至少一个子文件
S12
根据所述要素名称从所述至少一个子文件中选择目标子文件
S13
确定与所述要素名称对应的目标字段
S14
根据所述目标字段提取所述目标子文件中的目标要素
S15
根据所述目标要素生成要素表
S16
保存所述要素表至区块链节点
S17

图1

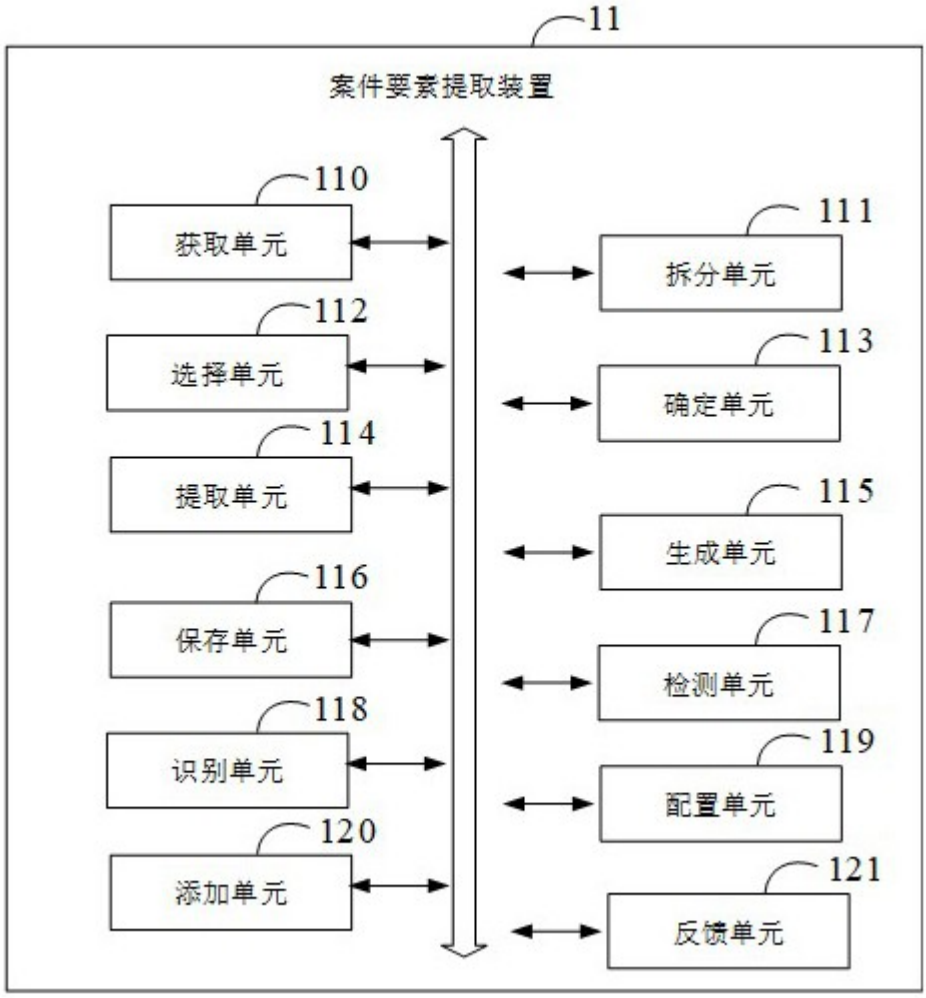

11
案件要素提取装置
110
获取单元
111
拆分单元
112
选择单元
113
确定单元
114
提取单元
115
生成单元
116
保存单元
117
检测单元
118
识别单元
119
配置单元
120
添加单元
121
反馈单元

图2

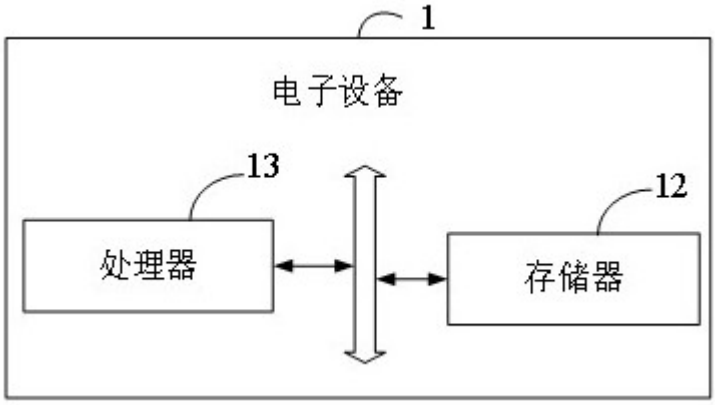

1
电子设备
13
处理器
12
存储器

图3