(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 112579727 B
(45)授权公告日 2022.03.22

(21)申请号 202011487916.6
(22)申请日 2020.12.16
(65)同一申请的已公布的文献号
申请公布号 CN 112579727 A
(43)申请公布日 2021.03.30
(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层
(72)发明人 曾凯 路华
(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 韩海花
(51)Int.Cl.
G06F 16/31(2019.01)
G06F 16/33(2019.01)
G06F 40/205(2020.01)
G06N 20/00(2019.01)

(56)对比文件
CN 110516048 A,2019.11.29
US 2019087414 A1,2019.03.21
CN 101877004 A,2010.11.03
CN 104111913 A,2014.10.22
CN 111930895 A,2020.11.13
CN 110659346 A,2020.01.07
CN 102959538 A,2013.03.06
CN 110334346 A,2019.10.15
CN 110888965 A,2020.03.17
CN 111832396 A,2020.10.27

审查员 何蒙蒙

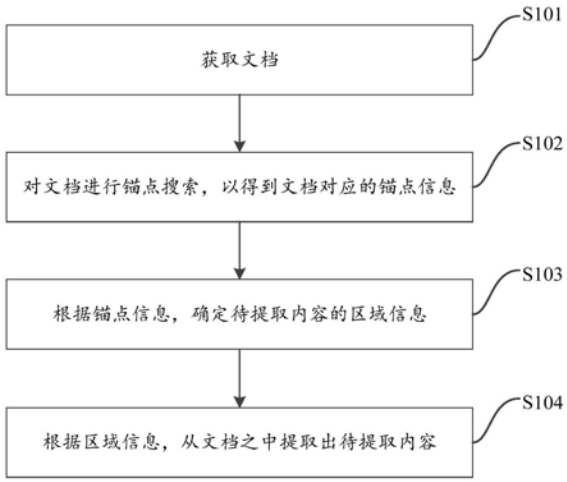
权利要求书2页 说明书10页 附图4页

(54)发明名称

文档内容的提取方法、装置、电子设备及存储介质

(57)摘要

本申请公开了文档内容的提取方法、装置、电子设备及存储介质，涉及自然语言处理、深度学习、知识图谱等人工智能技术领域。具体实现方案为：获取文档；对文档进行锚点搜索，以得到文档对应的锚点信息；根据锚点信息，确定待提取内容的区域信息；以及根据区域信息，从文档之中提取出待提取内容，能够有效避免受到文档内容布局的限制，有效地提升文档内容提取的准确性和提取效率，提升文档内容的提取效果。



CN 112579727 B

1.一种文档内容的提取方法,包括:

获取文档;

对所述文档进行锚点搜索,以得到所述文档对应的锚点信息;

根据所述锚点信息,确定待提取内容的区域信息;以及

根据所述区域信息,从所述文档之中提取出所述待提取内容;

其中,所述对所述文档进行锚点搜索,以得到所述文档对应的锚点信息,包括:

采用预先生成的空间索引搜索树对所述文档进行锚点搜索,以得到所述文档对应的锚点信息;

所述空间索引搜索树包括多个结点,以及多条边,所述结点表示参考锚点中的字符,所述边表示其所连接结点对应的字符之间的相关性向量;

所述参考锚点为参考键,

其中,所述采用预先生成的空间索引搜索树对所述文档进行锚点搜索,以得到所述文档对应的锚点信息,包括:

采用所述空间索引搜索树搜索所述文档中的每个字符,以从所述文档之中搜索得到与所述参考键相匹配的目标键;

确定所述参考键和与其对应的参考值在样本文档之中的相对布局信息;

将所述目标键作为搜索得到的与所述文档对应的锚点,并将所述相对布局信息作为与所述锚点对应的锚点信息。

2.根据权利要求1所述的方法,所述参考锚点的数量为多个,其中,所述从所述文档之中搜索得到与所述参考键相匹配的目标键,包括:

根据所述相关性向量确定匹配路径,所述匹配路径包括至少两个所述参考锚点;

根据所述相关性向量遍历所述匹配路径上的各个所述参考锚点;以及

从所述文档之中搜索得到与各个所述参考键相匹配的目标键。

3.根据权利要求1所述的方法,其中,所述根据所述锚点信息,确定待提取内容的区域信息,包括:

确定候选提取模板,所述候选提取模板具有对应的候选锚点信息;

确定与所述锚点信息匹配的候选锚点信息所属的候选提取模板,并将所述所属的候选提取模板作为目标提取模板;

根据所述目标提取模板,确定所述待提取内容的区域信息。

4.根据权利要求3所述的方法,其中,所述根据所述目标提取模板,确定所述待提取内容的区域信息,包括:

确定所述目标键对应于所述目标提取模板中的基准布局信息;

根据所述基准布局信息结合所述相对布局信息确定所述区域信息。

5.根据权利要求3所述的方法,其中,所述确定与所述锚点信息匹配的候选锚点信息所属的候选提取模板,包括:

将所述锚点信息和所述候选锚点信息输入至预训练的图模型之中,以得到所述图模型输出的所述所属的候选提取模板。

6.一种文档内容的提取装置,包括:

获取模块,用于获取文档;

搜索模块，用于对所述文档进行锚点搜索，以得到所述文档对应的锚点信息；

确定模块，用于根据所述锚点信息，确定待提取内容的区域信息；以及

提取模块，用于根据所述区域信息，从所述文档之中提取出所述待提取内容；

所述搜索模块，具体用于：

采用预先生成的空间索引搜索树对所述文档进行锚点搜索，以得到所述文档对应的锚点信息；

所述空间索引搜索树包括多个结点，以及多条边，所述结点表示参考锚点中的字符，所述边表示其所连接结点对应的字符之间的相关性向量；

所述参考锚点为参考键，

其中，所述搜索模块，具体用于：

采用所述空间索引搜索树搜索所述文档中的每个字符，以从所述文档之中搜索得到与所述参考键相匹配的目标键；

确定所述参考键和与其对应的参考值在样本文档之中的相对布局信息；

将所述目标键作为搜索得到的与所述文档对应的锚点，并将所述相对布局信息作为与所述锚点对应的锚点信息。

7.根据权利要求6所述的装置，所述参考锚点的数量为多个，其中，所述搜索模块，还用于：

根据所述相关性向量确定匹配路径，所述匹配路径包括至少两个所述参考锚点；

根据所述相关性向量遍历所述匹配路径上的各个所述参考锚点；以及

从所述文档之中搜索得到与各个所述参考键相匹配的目标键。

8.根据权利要求6所述的装置，其中，所述确定模块，包括：

第一确定子模块，用于确定候选提取模板，所述候选提取模板具有对应的候选锚点信息；

第二确定子模块，用于确定与所述锚点信息匹配的候选锚点信息所属的候选提取模板，并将所述所属的候选提取模板作为目标提取模板；

第三确定子模块，用于根据所述目标提取模板，确定所述待提取内容的区域信息。

9.根据权利要求8所述的装置，其中，所述第三确定子模块，具体用于：

确定所述目标键对应于所述目标提取模板中的基准布局信息；

根据所述基准布局信息结合所述相对布局信息确定所述区域信息。

10.根据权利要求8所述的装置，其中，所述第二确定子模块，具体用于：

将所述锚点信息和所述候选锚点信息输入至预训练的图模型之中，以得到所述图模型输出的所述所属的候选提取模板。

11.一种电子设备，其特征在于，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-5中任一项所述的方法。

12.一种存储有计算机指令的非瞬时计算机可读存储介质，其特征在于，所述计算机指令用于使所述计算机执行权利要求1-5中任一项所述的方法。

文档内容的提取方法、装置、电子设备及存储介质

技术领域

[0001] 本申请涉及计算机技术领域，具体涉及自然语言处理、深度学习、知识图谱等人工智能技术领域，尤其涉及文档内容的提取方法、装置、电子设备及存储介质。

背景技术

[0002] 人工智能是研究使计算机来模拟人的某些思维过程和智能行为(如学习、推理、思考、规划等)的学科，既有硬件层面的技术也有软件层面的技术。人工智能硬件技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理等技术；人工智能软件技术主要包括计算机视觉技术、语音识别技术、自然语言处理技术以及机器学习/深度学习、大数据处理技术、知识图谱技术等几大方向。

[0003] 文档中通常包含键值对和表格等，文档提取，即对文档进行内容识别，得到需求的键值对和表格等对应的实际内容。

发明内容

[0004] 提供了一种文档内容的提取方法、装置、电子设备、存储介质及计算机程序产品。

[0005] 根据第一方面，提供了一种文档内容的提取方法，包括：获取文档；对所述文档进行锚点搜索，以得到所述文档对应的锚点信息；根据所述锚点信息，确定待提取内容的区域信息；以及根据所述区域信息，从所述文档之中提取出所述待提取内容。

[0006] 根据第二方面，提供了一种文档内容的提取装置，包括：获取模块，用于获取文档；搜索模块，用于对所述文档进行锚点搜索，以得到所述文档对应的锚点信息；确定模块，用于根据所述锚点信息，确定待提取内容的区域信息；以及提取模块，用于根据所述区域信息，从所述文档之中提取出所述待提取内容。

[0007] 根据第三方面，提供了一种电子设备，包括：至少一个处理器；以及与所述至少一个处理器通信连接的存储器；其中，所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行本申请实施例的文档内容的提取方法。

[0008] 根据第四方面，提出了一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令用于使所述计算机执行本申请实施例公开的文档内容的提取方法。

[0009] 根据第五方面，提出了一种计算机程序产品，包括计算机程序，当所述计算机程序由处理器执行时实现本申请实施例公开的文档内容的提取方法。

[0010] 应当理解，本部分所描述的内容并非旨在标识本申请的实施例的关键或重要特征，也不用于限制本申请的范围。本申请的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0011] 附图用于更好地理解本方案，不构成对本申请的限定。其中：

[0012] 图1是根据本申请第一实施例的示意图；

[0013] 图2是本申请实施例中空间索引搜索树的结构示意图；

[0014] 图3是根据本申请第二实施例的示意图；

[0015] 图4是根据本申请第三实施例的示意图；

[0016] 图5是根据本申请第四实施例的示意图；

[0017] 图6是用来实现本申请实施例的文档内容的提取方法的电子设备的框图。

具体实施方式

[0018] 以下结合附图对本申请的示范性实施例做出说明，其中包括本申请实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本申请的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0019] 图1是根据本申请第一实施例的示意图。

[0020] 其中，需要说明的是，本实施例的文档内容的提取方法的执行主体为文档内容的提取装置，该装置可以由软件和/或硬件的方式实现，该装置可以配置在电子设备中，电子设备可以包括但不限于终端、服务器端等。

[0021] 本申请实施例涉及自然语言处理、深度学习、知识图谱等人工智能技术领域。

[0022] 其中，人工智能(Artificial Intelligence)，英文缩写为AI。它是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。

[0023] 深度学习是学习样本数据的内在规律和表示层次，这些学习过程中获得的信息对诸如文字，图像和声音等数据的解释有很大的帮助。深度学习的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据。

[0024] 自然语言处理，能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。

[0025] 而知识图谱，是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合目的的现代理论。

[0026] 如图1所示，该文档内容的提取方法包括：

[0027] S101：获取文档。

[0028] 其中，该文档为任意一张待提取其内容的文档，该文档中可以是包含键值对、表格、图片、文字等内容，对此不做限制。

[0029] 本申请实施例中，可以经由电子设备提供文本输入界面，接收用户输入的一段文本，并根据该段文本形成标准化的文档，或者，也可以解析用户语音录入的一段语音，将该段语音转换为相应的标准化的文档，对此不做限制。

[0030] S102：对文档进行锚点搜索，以得到文档对应的锚点信息。

[0031] 上述在获取文档之后，可以对文档进行锚点搜索，以得到文档对应的锚点信息。

[0032] 其中，锚点比如可以是文档内键值对中的键，键值对例如：银行名称-工商银行，则键是“银行名称”，值是“工商银行”，键值对还可以例如，表头和与表头对应的表格内容，则键可以是表头，值可以是相应的表格内容，对此不做限制。

[0033] 本申请实施例中的锚点，可以是上述两种示例中的键，键“银行名称”可以被称为

字符键，而表头形式的键，可以被称为表头键，字符键和表头键可以标识本申请实施例中描述的键的概念，对此不做限制。

[0034] 由此，对文档进行锚点搜索，可以具体是搜索文档中的字符键和表头键，也即是说，本申请在提取文档内容时，首先可搜索文档中的字符键和表头键，而后，根据搜索到的字符键和表头键辅助进行内容提取，而不是搜索整个文档包含的全部的实际内容，由此，能够有效地提升提取效率。

[0035] 一些实施例中，对文档进行锚点搜索，以得到文档对应的锚点信息，可以是采用预先生成的空间索引搜索树对文档进行锚点搜索，以得到文档对应的锚点信息，从而可以有效地提升搜索效率，并且保障了搜索的准确性。

[0036] 其中，空间索引搜索树可以是预先生成的，比如，可以获取海量的样本文档(也可以成为模板文档)，对各个样本文档进行内容识别，框选出需要提取的内容，并确定需要提取的内容对应的参考键(从样本文档中预先标注出的键，可以被称为参考键)，和与参考键对应的参考值(样本文档中，与预先标注的参考键对应的值，可以被称为参考值，而参考键和参考值的举例说明可以具体参见上述，在此不再赘述)，上述在提取了各个样本文档对应的参考键和参考值之后，可以将参考键作为参考锚点，从而将各个参考锚点中的字符作为结点，并且，在相互之间具有搜索相关性的字符之间构建边，根据各个参考锚点中的字符以及相应的边，形成空间索引搜索树。

[0037] 上述构建空间索引搜索树的过程可以被称为人工标注的过程，举例而言，人工标注的过程，指通过标注工具在各个样本文档上标注期望提取的结构化内容，比如可以通过画矩形框+输入标签实现：针对字符键值对(字符键-对应的值)：可以是用方框框选字符键部分的全部内容，并输入k1的标签；用方框框选对应的值部分的全部内容，并输入v1的标签；对于第二个字符键值对，重复以上步骤，差异为输入标签变为k2和v2，相同数字表示了字符键和对应的值的一一匹配关系。

[0038] 又比如，针对表头形式的键(表头键-对应的值)：可以用方框框选一个表头键对应的表头单元格的全部内容，并输入h1的标签；用方框框选出该表头键对应行和/或列的剩余单元格全部内容，并输入v1的标签；对于表格第二个表头单元格的标注，重复以上步骤，差一位输入标签变为h2和v2，相同数字表示了表头和行和/或列的一一匹配关系。

[0039] 上述在针对样本文档标注出字符键和表头键之后，可以相应地将字符键和表头键中的字符作为结点构建空间索引搜索树，举例而言：

[0040] 对于同一类文档，人工标注的字符键和表头键可以被视为是固定不变的，变化的是对应的内容，由此，可以将字符键和表头键作为参考锚点，根据字符键和表头键中的字符构建一棵空间索引搜索树，从而使得后续可根据该空间索引搜索树在实际的文档中用于锚点搜索，以搜索得到文档中的字符键和表头键。

[0041] 可选地，一些实施例中，空间索引搜索树包括多个结点，以及多条边，结点表示参考锚点中的字符，边表示其所连接结点对应的字符之间的相关性向量。

[0042] 例如，空间索引搜索树可以定义为一棵前缀树，树上的结点表示参考锚点中的字符，而树中从根节点到叶节点的一条路径表示一个参考锚点，相同前缀的参考键可以共享空间索引搜索树上从根结点开始的部分路径，树上结点之间的边表示从前一个字符向后一个字符的向量(该向量可以描述字符之间的相关性，由此，该向量可以被称为相关性向量)。

[0043] 另外一些实施例中，上述构建空间索引搜索树，使得空间索引搜索树包括多个结点，以及多条边，结点表示参考锚点中的字符，边表示其所连接结点对应的字符之间的相关性向量，还可以根据字符的尺寸，来归一化相关性向量，标注简单，从而能够减少标注数据量，有效降低文档提取所需要的软硬件资源消耗，并且避免文档排版过程中尺寸缩放时对内容提取造成影响，当将空间索引搜索树应用在实际的文档内容提取过程中时，具有较好的通用性，提升文档内容提取的灵活性。

[0044] 参见图2，图2是本申请实施例中空间索引搜索树的结构示意图，图2中模块21中表示从样本文档中标注出的字符，各个字符之间配置了相关性向量，从而将各个字符作为结点，将具有相关性的字符之间的相关性向量作为边构建得到空间索引搜索树(图2中模块22)，而后，在实际应用当中，可以结合图2中的空间索引搜索树，对文档中的内容逐个字符进行匹配，以识别得到文档中的锚点。

[0045] 另外一些实施例中，参考锚点包括：参考键，则采用预先生成的空间索引搜索树对文档进行锚点搜索，以得到文档对应的锚点信息，可以是采用空间索引搜索树搜索文档中的每个字符，以从文档之中搜索得到与参考键相匹配的目标键；确定参考键和与其对应的参考值在样本文档之中的相对布局信息；将目标键作为搜索得到的与文档对应的锚点，并将相对布局信息作为与锚点对应的锚点信息。

[0046] 也即是说，本申请实施例中还可以配置参考键作为参考锚点，并且，由于参考键和参考值是从样本文档中相应的键值对匹配得出的，则相应地，参考键和参考值，映射于样本文档之中会有一个相对布局信息，比如参考键和参考值映射于样本文档之中的相对的布局位置，和尺寸大小信息等，则这些相对的布局位置，和尺寸大小信息等，则可以被称为相对布局信息。

[0047] 可以理解的是，由于参考键和参考值是预先基于海量的样本文档标注得到的，并且，参考键和参考值之间具有相应的映射于样本文档之中的相对布局信息，则本申请实施例中可以采用空间索引搜索树搜索文档中的每个字符，以从文档之中搜索得到与参考键相匹配的目标键(文档之中与参考键相匹配的键，可以被称为目标键)，确定参考键和参考值在样本文档之中的相对布局信息；将目标键作为搜索得到的与文档对应的锚点，并将相对布局信息作为与锚点对应的锚点信息。

[0048] 则可以采用上述相对布局信息和目标键，辅助进行后续文档内容的提取，举例而言，可以使用空间索引搜索树从文档中每一个字开始沿着记录的下一个字符的相关性向量搜索，当沿着该相关性向量能够找到下一个字符时，则再沿着下一字的相关性向量继续搜索，直到根据各个字符之间的相关性向量搜索到一个完整的目标键(字符键或者表头键)，则将目标键作为搜索到的锚点，将相应的参考键和参考值对应的相对布局信息记录为该锚点的锚点信息，用于下一步的抽取。

[0049] 当每一个目标键都作为起始搜索一遍之后，就能够得到锚点序列(锚点序列中可以包括多个锚点)，该锚点序列中的各个锚点的锚点信息可以被用于指导下一步的内容提取过程。

[0050] 由于是采用空间索引搜索树从每个字符开始搜索锚点，因此每个锚点可以认为是相互独立的，从而使得各种因素导致的文档布局改变不影响空间索引搜索树对于锚点的搜索，另外，在搜索时每个锚点还可支持大小写匹配的搜索方法，避免英文字符的大小写对文

档布局产生的影响，使得文档在页面上的绝对位置、缩放大小、旋转角度、英文字符大小写等不影响提取效果，保障识别锚点的灵活性，从而拓展了文档内容的提取方法的应用范围。

[0051] 另外一些实施例中，参考锚点的数量为多个，其中，从文档之中搜索得到与参考键相匹配的目标键，可以是根据相关性向量确定匹配路径，匹配路径包括至少两个参考锚点，并根据相关性向量遍历匹配路径上的各个参考锚点；以及从文档之中搜索得到与各个参考键相匹配的目标键。

[0052] 也即是说，本申请实施例还提供了另外一种从文档中搜索锚点的方法，可以首先基于各个相关性向量确定匹配路径(该匹配路径，可以是由具有相关性向量的各边构成的)，而后，直接基于匹配路径上的各个参考锚点(参考锚点，即为参考键)的字符来搜索确定文档中的目标键并作为搜索到的锚点，能够减少搜索用的已标注的参考锚点的数据量，从而提升搜索效率。

[0053] S103：根据锚点信息，确定待提取内容的区域信息。

[0054] 上述在将目标键作为搜索到的锚点，将相应的参考键和参考值对应的相对布局信息(该相对布局信息也可以是预先标注参考键和参考值时，一并标注的，对此不做限制)记录为该锚点的锚点信息，可以直接根据目标键和相对布局信息来确定待提取内容的区域信息。

[0055] 其中，针对文档希望提取的内容，可以被称为待提取内容。

[0056] 比如，可以将目标键和相对布局信息输入至预训练的模型当中，以根据模型的输出来确定待提取内容的区域信息，或者，也可以采用其他任意可能的方式来根据锚点信息，确定待提取内容的区域信息，比如，工程的方式、数学运算的方式等等，对此不做限制。

[0057] S104：根据区域信息，从文档之中提取出待提取内容。

[0058] 上述在确定出待提取内容的区域信息之后，可以对文档进行内容识别，将识别到的内容当中，映射到区域信息所覆盖区域之内的内容，作为待提取内容，对此不做限制。

[0059] 本实施例中，通过获取文档，对文档进行锚点搜索，以得到文档对应的锚点信息，根据锚点信息，确定待提取内容的区域信息，以及根据区域信息，从文档之中提取出待提取内容，能够有效避免受到文档内容布局的限制，有效地提升文档内容提取的准确性和提取效率，提升文档内容的提取效果。

[0060] 图3是根据本申请第二实施例的示意图。

[0061] 如图3所示，该文档内容的提取方法包括：

[0062] S301：获取文档。

[0063] S302：对文档进行锚点搜索，以得到文档对应的锚点信息。

[0064] S301-S302的说明可以具体参见上述实施例，在此不再赘述。

[0065] S303：确定候选提取模板，候选提取模板具有对应的候选锚点信息。

[0066] 其中，候选提取模板，可以是预先标注的，该候选提取模板中可以包括提取处理逻辑，也即是说，该候选提取模板可以被调用，从而基于其中包含的提取处理逻辑，从文档中提取出待提取内容。

[0067] 与候选提取模块对应的锚点信息，可以被称为候选锚点信息，则候选提取模板，可以被用于提取出与候选锚点信息匹配的锚点信息所属的文档中的内容。

[0068] 候选提取模板的数量可以是多个，则本实施例中，可以支持从多个候选提取模板

之中选取出与搜索到的锚点信息相匹配的目标提取模板。

[0069] S304:确定与锚点信息匹配的候选锚点信息所属的候选提取模板,并将所属的候选提取模板作为目标提取模板。

[0070] 上述确定多个候选提取模板,并确定了每个候选提取模板对应的候选锚点信息之后,可以从多个候选提取模板之中选取出与搜索到的锚点信息相匹配的目标提取模板。

[0071] 其中,与搜索到的锚点信息相匹配的候选锚点信息所属的候选提取模板,可以被称为目标提取模板,且由于目标提取模板的候选锚点信息是与从文档中搜索到的锚点信息相匹配的,从而实现候选提取模板的自动管理,可以实现自动化地选取出提取效果最优的目标提取模板。

[0072] 一些实施例中,确定与锚点信息匹配的候选锚点信息所属的候选提取模板,可以是将锚点信息和候选锚点信息输入至预训练的图模型之中,以得到图模型输出的所属的候选提取模板。

[0073] 其中图模型,可以是深度学习中的图模型,或者也可以是人工智能技术领域中其他任意可能架构形式的图模型,对此不做限制。

[0074] 本申请实施例中采用的图模型为概率分布的图形表示,一个图由结点和它们之间的链接组成,在概率图模型中,每个结点表示一个随机变量(或一组随机变量),链接表示这些变量之间的概率关系。这样,图模型描述了联合概率分布在所有随机变量上能够分解为一组因子乘积的方式,每个因子只依赖于随机变量的一个子集。

[0075] 比如,可以首先将锚点信息和候选锚点信息输入至预训练的图模型之中,基于预训练的图模型建立以锚点信息为结点,两两锚点信息的连线为边的图G(V,E),其中V表示结点,E表示边,按照同样的方法可以将所有的候选提取模板也抽象为图,而后,基于预训练的图模型度量文档$G_i(V,E)$与候选提取模板$G_j(V,E)$的相似度(i表示文档中搜索到的锚点的数量,j表示每个候选提取模板之中候选锚点的数量),而后,确定相似度最大的候选提取模板为目标提取模板。

[0076] 而基于预训练的图模型度量文档$G_i(V,E)$与候选提取模板$G_j(V,E)$的相似度的公式可以是相关技术中任意可能的相似度计算公式,对此不做限制。

[0077] 而另外一些实施例中,由于采用了图相似匹配算法,不仅可以度量文档与候选提取模板的相似度,对于文本内容相同的锚点,还可以根据锚点在文档中布局上的差异,通过构建以冲突锚点为中心的子图,并根据图相似度算法区分各个冲突的锚点,从而允许存在多个相同的键,实现对冲突锚点的区分检测。

[0078] 上述在确定候选提取模板,并确定与锚点信息匹配的候选锚点信息所属的候选提取模板,并将所属的候选提取模板作为目标提取模板之后,可以直接基于该目标提取模板从文档之中提取出待提取内容,从而实现采用一张目标提取模板抽取文档中的内容,并且该目标提取模板的候选锚点和文档之中的锚点的布局具有较匹配的相似度,从而有效提升提取准确性。

[0079] S305:根据目标提取模板,确定待提取内容的区域信息。

[0080] 其中,区域信息比如该待提取内容在文档之中所占据区域的位置、尺寸等信息,比如该待提取内容所占据的区域A,相对于文档的整个区域的相对位置坐标、长宽比例等。

[0081] 一些实施例中,当根据目标提取模板,确定待提取内容的区域信息,可以是确定目

标键对应于目标提取模板中的基准布局信息;根据基准布局信息结合相对布局信息确定区域信息。

[0082] 由于目标键是从文档中搜索出的锚点,而该搜索出的锚点与目标提取模板的候选锚点具有较高的相似度,由此,本实施例中,为了在提取的过程中,直接快速地基于目标提取模板去提取文档中的内容,可以将文档中搜索出的锚点,与目标提取模板进行匹配,将文档中搜索出的目标键,对应于目标提取模板中的布局位置、尺寸等作为基准布局信息,而后,结合相对布局信息(参考键和参考值映射于样本文档之中的相对的布局位置,和尺寸大小信息等)确定区域信息。

[0083] 比如,可以将基准布局进行与相对布局信息作加和运算,从而运算出待提取内容在文档之中所占据区域的位置、尺寸等信息,对此不做限制。

[0084] S306:根据区域信息,从文档之中提取出待提取内容。

[0085] 举例而言,当确定了目标提取模板之后,由于对每个目标键都有对应的一个匹配的参考键,并且针对该参考键,均预先标注了参考值,以及参考键与对应的参考值之间的相对布局信息,因此,可以根据锚点在目标提取模板中的基准布局,结合参考键与对应的参考值之间的相对布局信息,能够在文档中计算出待抽取内容的区域信息(内容占据区域的大小和位置),而后,从该区域信息描述的区域当中提取出待提取内容(比如该区域信息描述的区域当中的键值对和表格的表头、行或者列的结构中实际的内容)。

[0086] 由于确定目标键对应于目标提取模板中的基准布局信息,并且根据基准布局信息结合相对布局信息确定区域信息,从而辅助后续直接抽取出区域信息所描述区域中的待提取内容,实现简便,具有较好的适用性和实用性,提升提取效率和提升准确性。

[0087] 本申请实施例中,当候选提取模板的数量为多个时,还可以根据实际应用的需求对多个候选提取模板进行组合、拼接,或者对候选提取模板进行拆分,本申请实施例中在匹配提取模板时,还可以支持局部模板匹配,因此,具有较好的提取灵活性。

[0088] 本实施例中,由于目标提取模板的候选锚点信息是与从文档中搜索到的锚点信息相匹配的,从而实现候选提取模板的自动管理,可以实现自动化地选取出提取效果最优的目标提取模板。由于采用了图相似匹配算法,不仅可以度量文档与候选提取模板的相似度,对于文本内容相同的锚点,还可以根据锚点在文档中布局上的差异,通过构建以冲突锚点为中心的子图,并根据图相似度算法区分各个冲突的锚点,从而允许存在多个相同的键,实现对冲突锚点的区分检测。在确定候选提取模板,并确定与锚点信息匹配的候选锚点信息所属的候选提取模板,并将所属的候选提取模板作为目标提取模板之后,可以直接基于该目标提取模板从文档之中提取出待提取内容,从而实现采用一张目标提取模板抽取文档中的内容,并且该目标提取模板的候选锚点和文档之中的锚点的布局具有较匹配的相似度,从而有效提升提取准确性。

[0089] 图4是根据本申请第三实施例的示意图。

[0090] 如图4所示,该文档内容的提取装置40,包括:

[0091] 获取模块401,用于获取文档;

[0092] 搜索模块402,用于对文档进行锚点搜索,以得到文档对应的锚点信息;

[0093] 确定模块403,用于根据锚点信息,确定待提取内容的区域信息;

[0094] 提取模块404,用于根据区域信息,从文档之中提取出待提取内容。

[0095] 在本申请的一些实施例中，其中，搜索模块402，具体用于：

[0096] 采用预先生成的空间索引搜索树对文档进行锚点搜索，以得到文档对应的锚点信息。

[0097] 在本申请的一些实施例中，其中，空间索引搜索树包括多个结点，以及多条边，结点表示参考锚点中的字符，边表示其所连接结点对应的字符之间的相关性向量。

[0098] 在本申请的一些实施例中，其中，参考锚点包括：参考键，

[0099] 其中，搜索模块402，具体用于：

[0100] 采用空间索引搜索树搜索文档中的每个字符，以从文档之中搜索得到与参考键相匹配的目标键；

[0101] 确定参考键和与其对应的参考值在样本文档之中的相对布局信息；

[0102] 将目标键作为搜索得到的与文档对应的锚点，并将相对布局信息作为与锚点对应的锚点信息。

[0103] 在本申请的一些实施例中，参考锚点的数量为多个，其中，搜索模块402，还用于：

[0104] 根据相关性向量确定匹配路径，匹配路径包括至少两个参考锚点；

[0105] 根据相关性向量遍历匹配路径上的各个参考锚点；以及

[0106] 从文档之中搜索得到与各个参考键相匹配的目标键。

[0107] 在本申请的一些实施例中，如图5所示，图5是根据本申请第四实施例的示意图，该文档内容的提取装置50，包括：获取模块501、搜索模块502、确定模块503，以及提取模块504，其中，确定模块503，包括：

[0108] 第一确定子模块5031，用于确定候选提取模板，候选提取模板具有对应的候选锚点信息；

[0109] 第二确定子模块5032，用于确定与锚点信息匹配的候选锚点信息所属的候选提取模板，并将所属的候选提取模板作为目标提取模板；

[0110] 第三确定子模块5033，用于根据目标提取模板，确定待提取内容的区域信息。

[0111] 在本申请的一些实施例中，其中，第三确定子模块5033，具体用于：

[0112] 确定目标键对应于目标提取模板中的基准布局信息；

[0113] 根据基准布局信息结合相对布局信息确定区域信息。

[0114] 在本申请的一些实施例中，其中，第二确定子模块5032，具体用于：

[0115] 将锚点信息和候选锚点信息输入至预训练的图模型之中，以得到图模型输出的所属的候选提取模板。

[0116] 可以理解的是，本实施例附图5中的文档内容的提取装置50与上述实施例中的文档内容的提取装置40，获取模块501与上述实施例中的获取模块401，搜索模块502与上述实施例中的搜索模块402，确定模块503与上述实施例中的确定模块403，提取模块504与上述实施例中的提取模块404，可以具有相同的功能和结构。

[0117] 需要说明的是，前述对文档内容的提取方法的解释说明也适用于本实施例的文档内容的提取装置，此处不再赘述。

[0118] 本实施例中，通过获取文档，对文档进行锚点搜索，以得到文档对应的锚点信息，根据锚点信息，确定待提取内容的区域信息，以及根据区域信息，从文档之中提取出待提取内容，能够有效避免受到文档内容布局的限制，有效地提升文档内容提取的准确性和提取

效率，提升文档内容的提取效果。

[0119] 根据本申请的实施例，本申请还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0120] 图6是用来实现本申请实施例的文档内容的提取方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0121] 如图6所示，设备600包括计算单元601，其可以根据存储在只读存储器(ROM)602中的计算机程序或者从存储单元608加载到随机访问存储器(RAM)603中的计算机程序，来执行各种适当的动作和处理。在RAM 603中，还可存储设备600操作所需的各种程序和数据。计算单元601、ROM 602以及RAM 603通过总线604彼此相连。输入/输出(I/O)接口605也连接至总线604。

[0122] 设备600中的多个部件连接至I/O接口605，包括：输入单元606，例如键盘、鼠标等；输出单元607，例如各种类型的显示器、扬声器等；存储单元608，例如磁盘、光盘等；以及通信单元609，例如网卡、调制解调器、无线通信收发机等。通信单元609允许设备600通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0123] 计算单元601可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元601的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元601执行上文所描述的各个方法和处理，例如，文档内容的提取方法。

[0124] 例如，在一些实施例中，文档内容的提取方法可被实现为计算机软件程序，其被有形地包含于机器可读介质，例如存储单元608。在一些实施例中，计算机程序的部分或者全部可以经由ROM 602和/或通信单元609而被载入和/或安装到设备600上。当计算机程序加载到RAM 603并由计算单元601执行时，可以执行上文描述的文档内容的提取方法的一个或多个步骤。备选地，在其他实施例中，计算单元601可以通过其他任何适当的方式(例如，借助于固件)而被配置为执行文档内容的提取方法。

[0125] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0126] 用于实施本申请的文档内容的提取方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数

据处理装置的处理器或控制器，使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0127] 在本申请的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0128] 为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置(例如，CRT(阴极射线管)或者LCD(液晶显示器)监视器)；以及键盘和指向装置(例如，鼠标或者轨迹球)，用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈(例如，视觉反馈、听觉反馈、或者触觉反馈)；并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0129] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如，作为数据服务器)、或者包括中间件部件的计算系统(例如，应用服务器)、或者包括前端部件的计算系统(例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如，通信网络)来将系统的部件相互连接。通信网络的示例包括：局域网(LAN)、广域网(WAN)、互联网及区块链网络。

[0130] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器，又称为云计算服务器或云主机，是云计算服务体系中的一项主机产品，以解决了传统物理主机与VPS服务("Virtual Private Server"，或简称"VPS")中，存在的管理难度大，业务扩展性弱的缺陷。服务器也可以为分布式系统的服务器，或者是结合了区块链的服务器。

[0131] 应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本发公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本申请公开的技术方案所期望的结果，本文在此不进行限制。

[0132] 上述具体实施方式，并不构成对本申请保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等，均应包含在本申请保护范围之内。

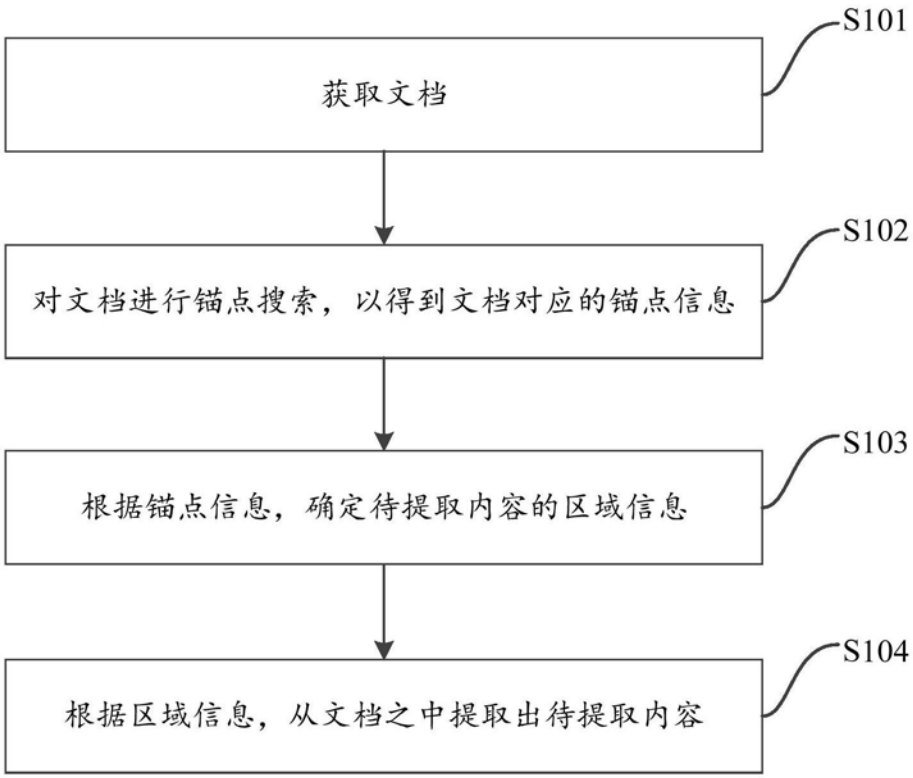

S101
获取文档
S102
对文档进行锚点搜索，以得到文档对应的锚点信息
S103
根据锚点信息，确定待提取内容的区域信息
S104
根据区域信息，从文档之中提取出待提取内容

图1

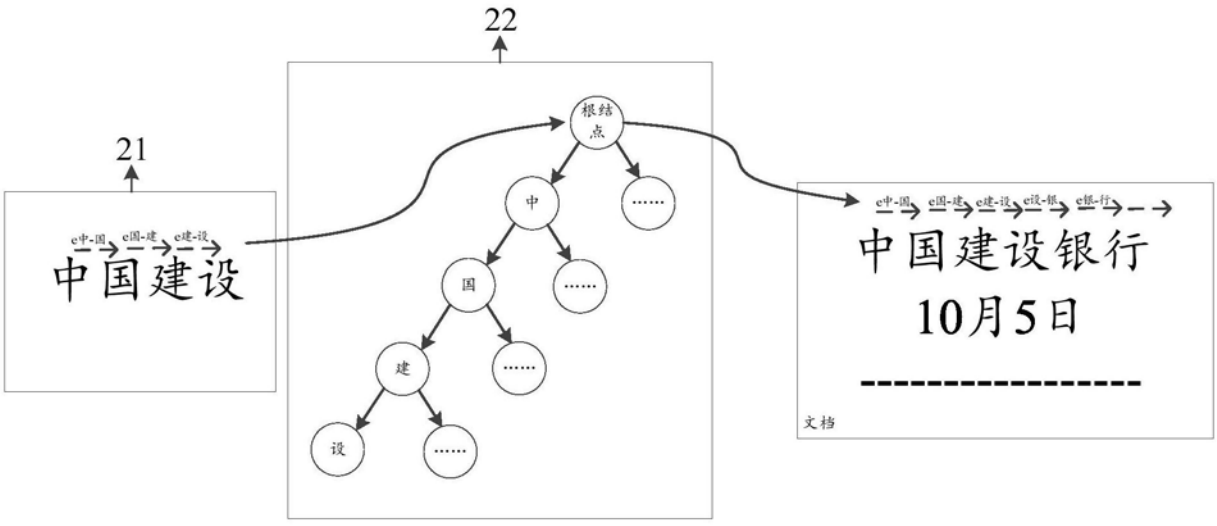

22
21
e中-国 e国-建 e建-设
中国建设
根结点
中
……
国
……
建
……
设
……
e中-国 e国-建 e建-设 e设-银 e银-行
中国建设银行
10月5日
文档

图2

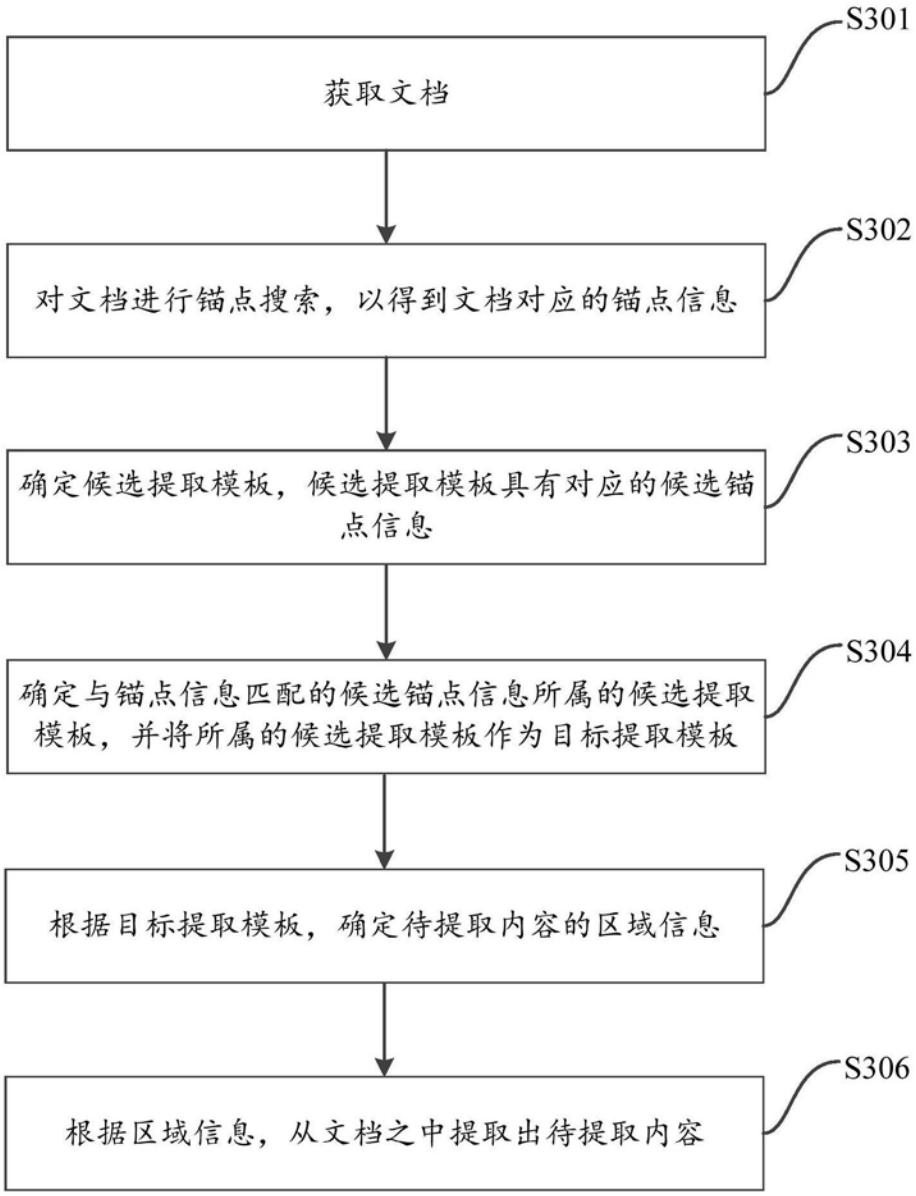

S301
获取文档
S302
对文档进行锚点搜索，以得到文档对应的锚点信息
S303
确定候选提取模板，候选提取模板具有对应的候选锚点信息
S304
确定与锚点信息匹配的候选锚点信息所属的候选提取模板，并将所属的候选提取模板作为目标提取模板
S305
根据目标提取模板，确定待提取内容的区域信息
S306
根据区域信息，从文档之中提取出待提取内容

图3

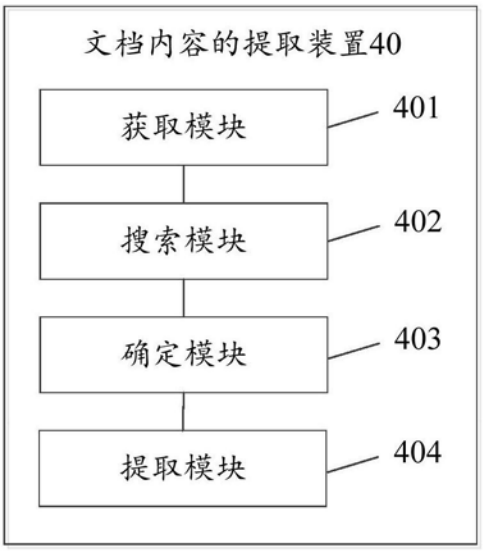

文档内容的提取装置40
获取模块
401
搜索模块
402
确定模块
403
提取模块
404

图4

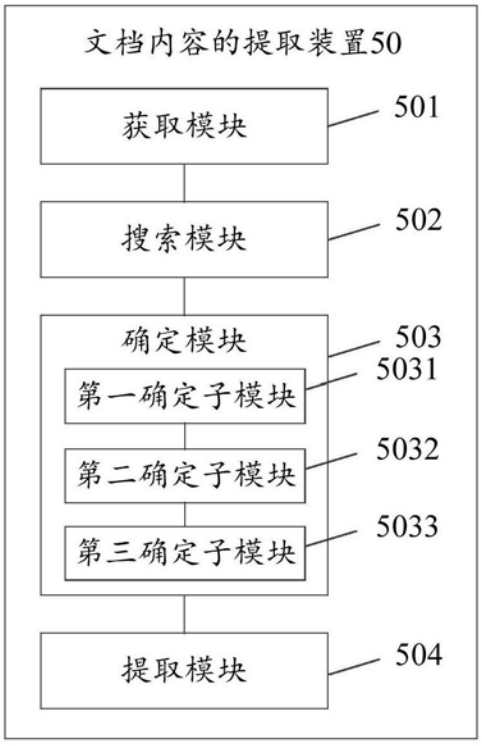

文档内容的提取装置50
获取模块
501
搜索模块
502
确定模块
503
第一确定子模块
5031
第二确定子模块
5032
第三确定子模块
5033
提取模块
504

图5

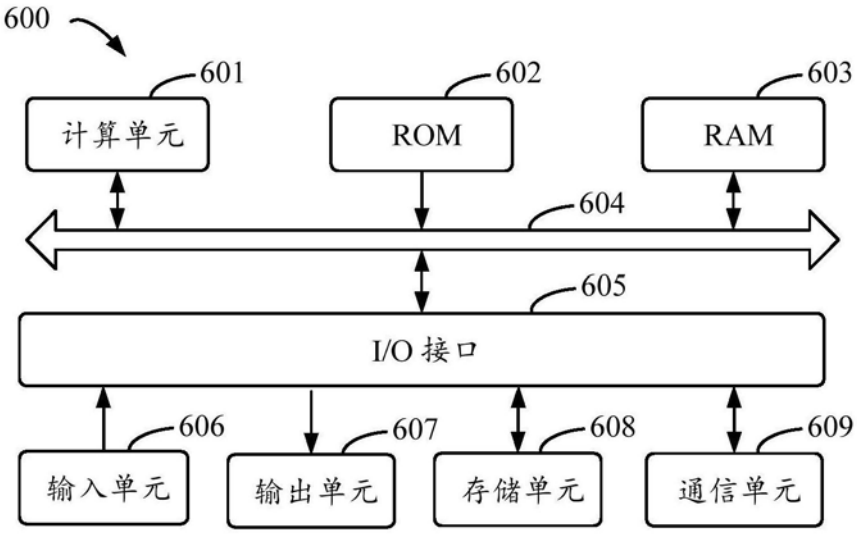

600
601
计算单元
602
ROM
603
RAM
604
605
I/O 接口
606
输入单元
607
输出单元
608
存储单元
609
通信单元

图6