

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 3 月 10 日 (10.03.2022)



(10) 国际公布号
WO 2022/048211 A1

(51) 国际专利分类号:
G06F 40/151 (2020.01)

(21) 国际申请号: PCT/CN2021/097070

(22) 国际申请日: 2021 年 5 月 30 日 (30.05.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010919654.X 2020 年 9 月 3 日 (03.09.2020) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 刘玉(LIU, Yu); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。徐国强(XU, Guoqiang); 中国广东省深圳市福田区

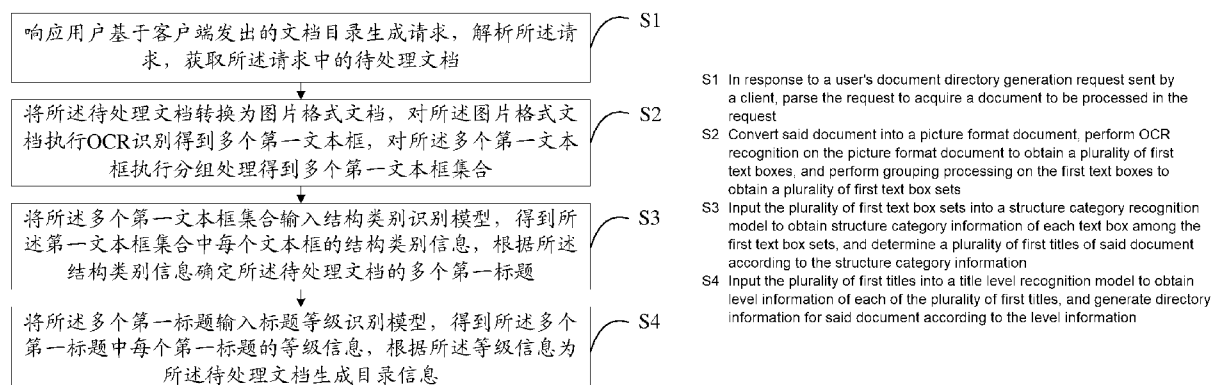
福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理有限公司(普通合伙)(SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区梅林街道中康路 128 号卓越城 1 期 2 栋 301B 于志光, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(54) **Title:** DOCUMENT DIRECTORY GENERATION METHOD AND APPARATUS, ELECTRONIC DEVICE AND READABLE STORAGE MEDIUM

(54) 发明名称: 文档目录生成方法、装置、电子设备及可读存储介质



(57) **Abstract:** A document directory generation method, relating to data processing. Said method comprises: converting a document to be processed into a picture format document, performing OCR on the picture format document to obtain a plurality of first text boxes, and performing grouping processing on the first text boxes to obtain a plurality of first text box sets; inputting the plurality of first text box sets into a structure category recognition model to obtain structure category information of each text box among the first text box sets, and determining a plurality of first titles of said document according to the structure category information; and inputting the plurality of first titles into a title level recognition model to obtain level information of each first title, and generating directory information for said document according to the level information. Further provided are a document directory generation apparatus, an electronic device and a readable storage medium. The document directory generation efficiency can be improved, and the accuracy and integrity of directory information extraction are ensured.

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 一种文档目录生成方法, 涉及数据处理, 该方法包括: 将待处理文档转换为图片格式文档, 对图片格式文档执行OCR识别得到多个第一文本框, 对第一文本框执行分组处理得到多个第一文本框集合; 将多个第一文本框集合输入结构类别识别模型, 得到第一文本框集合中每个文本框的结构类别信息, 根据结构类别信息确定待处理文档的多个第一标题; 将多个第一标题输入标题等级识别模型, 得到每个第一标题的等级信息, 根据等级信息为待处理文档生成目录信息。还提供一种文档目录生成装置、电子设备及可读存储介质。可提高文档目录生成效率, 保证目录信息提取的准确性、完整性。

文档目录生成方法、装置、电子设备及可读存储介质

本申请要求于 2020 年 9 月 3 日提交中国专利局、申请号为 CN202010919654.X、名称为“文档目录生成方法、装置、电子设备及可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及数据处理领域，尤其涉及一种文档目录生成方法、装置、电子设备及可读存储介质。

背景技术

随着信息技术的发展，各个领域每天产生数以万计的文件，且产生出的文件格式众多，例如：PDF 格式文件、doc 格式文件、纸质文件、图片格式文件，对于数据量较大的文件，我们希望能通过目录信息迅速定位到自己感兴趣的内容，然而，如何生成目录信息呢？

当前，通常分别读取段落信息以确定文件结构，再根据文件结构确定目录信息，发明人意识到这种方式目录信息生成效率低下，且当文件中有多级标题的情况下，现有方法无法准确、完整的将目录信息提取出来。因此，亟需一种文档目录生成方法，以提高文档目录生成效率，保证目录信息提取的准确性、完整性。

发明内容

本申请提供的文档目录生成方法，包括：

响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待处理文档；

将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合；

将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题；

将所述多个第一标题输入标题等级识别模型，得到所述多个第一标题中每个第一标题的等级信息，根据所述等级信息为所述待处理文档生成目录信息。

本申请还提供一种文档目录生成装置，所述装置包括：

获取模块，用于响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待处理文档；

分组模块，用于将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合；

确定模块，用于将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题；

生成模块，用于将所述多个第一标题输入标题等级识别模型，得到所述多个第一标题中每个第一标题的等级信息，根据所述等级信息为所述待处理文档生成目录信息。

本申请还提供一种电子设备，所述电子设备包括：

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文档目录生成程序，所述文档目录生成程序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如下步骤：

响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待

处理文档；

将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合；

将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题；

将所述多个第一标题输入标题等级识别模型，得到所述多个第一标题中每个第一标题的等级信息，根据所述等级信息为所述待处理文档生成目录信息。

本申请还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有文档目录生成程序，所述文档目录生成程序可被一个或者多个处理器执行，以实现如下步骤：

响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待处理文档；

将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合；

将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题；

将所述多个第一标题输入标题等级识别模型，得到所述多个第一标题中每个第一标题的等级信息，根据所述等级信息为所述待处理文档生成目录信息。

附图说明

图 1 为本申请一实施例提供的文档目录生成方法的流程示意图；

图 2 为本申请一实施例提供的文档目录生成装置的模块示意图；

图 3 为本申请一实施例提供的实现文档目录生成方法的电子设备的结构示意图；

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本申请，并不用于限定本申请。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

需要说明的是，在本申请中涉及“第一”、“第二”等的描述仅用于描述目的，而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施例之间的技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在，也不在本申请要求的保护范围之内。

本申请提供一种文档目录生成方法。参照图 1 所示，为本申请一实施例提供的文档目录生成方法的流程示意图。该方法可以由一个电子设备执行，该电子设备可以由软件和/或硬件实现。

本实施例中，文档目录生成方法包括：

S1、响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待处理文档；

S2、将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合。

本实施例中，所述待处理文档的格式包括 PDF 格式、doc 格式、纸质格式，所述将所述待处理文档转换为图片格式文档包括：

A1、将 PDF 格式待处理文档中的每一页文档转换为一张图片，得到以页码命名的多

张图片，汇总所述多张图片得到图片格式文档；

A2、将 doc 格式（或 docx 格式）待处理文档转换为 PDF 格式文档，将 PDF 格式文档转换为以页码命名的多张图片，汇总所述多张图片得到图片格式文档；

A3、将纸质格式待处理文档按文档顺序拍摄为以页码命名的多张照片，汇总所述多张照片得到图片格式文档。

本实施例中，采用 OCR 识别技术对图片格式的待处理文档进行文字识别，可识别得到每张图片中包含的多个文本框、每个文本框中的文字信息及每个文本框在对应图片中的坐标。

所述对所述多个第一文本框执行分组处理得到多个第一文本框集合包括：

B1、为所述多个第一文本框中的每个第一文本框建立一个三元数组，所述三元数组包括第一文本框所属图片的页码、第一文本框的左上角和右下角坐标、第一文本框中的文字信息；

B2、基于所述三元数组将位于同一页同一行的第一文本框合并，得到多个行文本框；

B3、将所述多个行文本框划分为包含预设数量文本框的多个第一文本框集合。

将 OCR 识别得到的文本框中位于同一页同一行的文本框合并后，每页图片至少对应从上往下排列的 15 个文本框，当待处理文档页数较多时，得到的文本框数量较多，现有的序列模型无法一次性建模这么多数据（大部分序列模型的输入长度为 512，当输入文字长度大于 512 时，需要对输入文字进行拆分），将文本框划分为包含预设数量（例如，10 个）文本框的一个或多个文本框集合，可避免信息在输入模型时溢出，保证了后续文本结构类别及标题等级识别的准确性。

S3、将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题。

本实施例中，所述结构类别识别模型为训练好的 LayoutLM 模型，用于识别第一文本框集合中每个文本框的结构类别（结构类别包括：居中标题、分级标题、正文），根据结构类别可确定所述待处理文档的多个标题（包括居中标题、分级标题），但是不能区分分级标题为几级分级标题。

LayoutLM 模型是一种通用文档预训练模型，对文档结构信息和视觉信息进行建模，文档中文字的位置关系蕴含着丰富的语义信息，例如，居中标题通常位于文档首行正中位置；文字格式所呈现的视觉信息也体现了语义信息，例如，文章的分级标题通常会放大加粗呈现，特殊概念名称会以斜体呈现。

本实施例中，所述结构类别识别模型的确定过程包括：

C1、对携带结构类别标注信息的第一图片样本执行文字识别得到多个第二文本框，对所述多个第二文本框执行分组处理得到多个第二文本框集合；

C2、将所述第二文本框集合输入第一 LayoutLM 模型，得到所述第二文本框集合中每个文本框的预测结构类别；

C3、基于所述结构类别标注信息确定所述第二文本框集合中每个文本框的真实结构类别，通过最小化预测结构类别与真实结构类别之间的第一损失值确定所述第一 LayoutLM 模型的结构参数，得到结构类别识别模型。

所述将所述第二文本框集合输入第一 LayoutLM 模型包括：

D1、基于所述三元数组将所述第二文本框集合中的文本框按照位置顺序通过连接符进行连接，得到第一文本序列；

D2、在所述第一文本序列的开头增加句首符，在所述第一文本序列的结尾增加连接符，得到第二文本序列，将所述第二文本序列输入第一 LayoutLM 模型。

假设第二文本框集合中包括 10 个文本框，本实施例中句首符用 CLS 表示，连接符用 SEP 表示，则第二文本序列为

CLS, W11, W12, ..., W1n₁, SEP, W21, W22, ..., W2n₂, SEP, ..., SEP, W101, W102, ..., W10n₁₀, SEP。其中 W_{ij} 表示第二文本框集合中第 i 个文本框的第 j 个文字, n_i 表示第二文本框集合中第 i 个文本框中文字的总数量。

当输入经过 LayoutLM 模型后, 会得到相同长度的输出, 每个文本框后面的 SEP 所输出的结构类别即为该文本框的结构类别。

所述第一损失值的计算公式为:

$$\text{loss}(q_{mn}, p_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c p_{mn} (\log q_{mn})$$

其中, q_{mn} 为第 m 个第二文本框集合中第 n 个文本框的预测结构类别, p_{mn} 为第 m 个第二文本框集合中第 n 个文本框的真实结构类别, $\text{loss}(q_{mn}, p_{mn})$ 为第 m 个第二文本框集合中第 n 个文本框的预测结构类别和真实结构类别之间的第一损失值, c 为第 m 个第二文本框集合中文本框的总数量, t 为结构类别的总数量(本实施例中, 结构类别的总数量为 3, 包括: 居中标题、分级标题、正文)。

S4、将所述多个第一标题输入标题等级识别模型, 得到所述多个第一标题中每个第一标题的等级信息, 根据所述等级信息为所述待处理文档生成目录信息。

本实施例中, 所述标题等级识别模型也为训练好的 LayoutLM 模型, 用于识别分级标题为几级分级标题(例如, 一级标题、二级标题、三级标题...)。本申请中标题等级识别模型与文本结构类别识别模型输入的样本不同, 训练目标不同, 训练得到的模型参数也不同。

本实施例在识别得到居中标题和各级分级标题后即可创建出较为详细的目录信息, 用户根据目录信息可确定文档的整体布局, 快速定位自己感兴趣的部分。

本实施例中, 不将文本结构类别识别模型和标题等级识别模型结合在一起作为一个模型进行文本结构类别识别和标题等级识别的目的在于: 对于大数据量的文档, OCR 识别后得到的文本框数量较多, LayoutLM 模型无法一次性将所有文本框输入, 所以将文本框进行分组, 得到多个文本框集合, 然而因居中标题和分级标题在一个文档中的数量比较稀少, 一个文本框集合中可能只有一个文本框对应分级标题, 也可能没有一个文本框对应分级标题, 而分级标题的等级的识别需要关注上下文信息, 故而无法根据一个文本框集合来确定分级标题具体为几级分级标题, 采用一个模型来识别结构类别和标题等级会使标题等级识别准确度不够高。本实施例通过结构类别识别模型得到多个第一标题, 因第一标题的数量不会太多, 可一次性输入标题等级识别模型, 完整的保留了上下文信息, 保证了标题等级识别模型的高准确性。

所述标题等级识别模型的确定过程包括:

E1、对携带标题等级标注信息的第二图片样本执行文字识别得到多个第三文本框, 对所述多个第三文本框执行分组处理得到多个第三文本框集合;

E2、将所述第三文本框集合输入文本结构类别识别模型, 得到所述第三文本框集合中每个文本框的预测结构类别, 基于所述预测结构类别得到所述第二图片样本对应的多个第二标题;

E3、将所述多个第二标题输入第二 LayoutLM 模型, 得到每个第二标题对应的预测标题等级, 基于所述标题等级标注信息确定每个第二标题的真实标题等级, 通过最小化预测标题等级与真实标题等级之间的第二损失值确定所述第二 LayoutLM 模型的结构参数, 得到标题等级识别模型。

所述第二损失值的计算公式为:

$$\text{loss}(q_{uv}, p_{uv}) = -\frac{1}{h} \sum_v^g \sum_u^h p_{uv} (\log q_{uv})$$

其中, q_{uv} 为第 u 个第二图片样本的第 v 个第二标题的预测标题等级, p_{uv} 为第 u 个第二图片样本的第 v 个第二标题的真实标题等级, $\text{loss}(q_{uv}, p_{uv})$ 为第 u 个第二图片样本的第

v 个第二标题的的预测标题等级和真实标题等级之间的第二损失值, h 为第 u 个第二图片样本的第二标题的总数量, g 为标题等级的总数量(本实施例中, 标题等级的总数量为 8, 包括: 一级标题、二级标题、...、八级标题)。

由上述实施例可知, 本申请提出的文档目录生成方法, 首先, 将待处理文档转换为图片格式文档, 对图片格式文档执行文字识别得到多个第一文本框, 为每个第一文本框建立一个三元数组, 基于三元数组对第一文本框进行分组得到多个第一文本框集合, 本步骤实现了将不同格式的文档统一转换为第一文本框, 并分组得到多个第一文本框集合, 使得每个第一文本框集合的数据量适中, 可避免后续在将数据输入模型时溢出, 保证了信息完整性; 接着, 将第一文本框集合输入结构类别识别模型得到每个文本框的结构类别, 基于结构类别确定待处理文档的多个第一标题, 本步骤可完整获取待处理文档的每个第一标题; 最后, 将所述多个第一标题一次性输入标题等级识别模型, 保留了上下文信息, 提高了标题等级的识别准确率, 再根据等级信息即可生成目录信息, 使得目录信息生成效率较高。故而, 本申请提高了文档目录生成效率, 保证了目录信息提取的准确性、完整性。

如图 2 所示, 为本申请一实施例提供的文档目录生成装置的模块示意图。

本申请所述文档目录生成装置 100 可以安装于电子设备中。根据实现的功能, 所述文档目录生成装置 100 可以包括获取模块 110、分组模块 120、确定模块 130 及生成模块 140。本申请所述模块也可以称之为单元, 是指一种能够被电子设备处理器所执行, 并且能够完成固定功能的一系列计算机程序段, 其存储在电子设备的存储器中。

在本实施例中, 关于各模块/单元的功能如下:

获取模块 110, 用于响应用户基于客户端发出的文档目录生成请求, 解析所述请求, 获取所述请求中的待处理文档;

分组模块 120, 用于将所述待处理文档转换为图片格式文档, 对所述图片格式文档执行 OCR 识别得到多个第一文本框, 对所述多个第一文本框执行分组处理得到多个第一文本框集合。

本实施例中, 所述待处理文档的格式包括 PDF 格式、doc 格式、纸质格式, 所述将所述待处理文档转换为图片格式文档包括:

A1、将 PDF 格式待处理文档中的每一页文档转换为一张图片, 得到以页码命名的多张图片, 汇总所述多张图片得到图片格式文档;

A2、将 doc 格式(或 docx 格式)待处理文档转换为 PDF 格式文档, 将 PDF 格式文档转换为以页码命名的多张图片, 汇总所述多张图片得到图片格式文档;

A3、将纸质格式待处理文档按文档顺序拍摄为以页码命名的多张照片, 汇总所述多张照片得到图片格式文档。

本实施例中, 采用 OCR 识别技术对图片格式的待处理文档进行文字识别, 可识别得到每张图片中包含的多个文本框、每个文本框中的文字信息及每个文本框在对应图片中的坐标。

所述对所述多个第一文本框执行分组处理得到多个第一文本框集合包括:

B1、为所述多个第一文本框中的每个第一文本框建立一个三元数组, 所述三元数组包括第一文本框所属图片的页码、第一文本框的左上角和右下角坐标、第一文本框中的文字信息;

B2、基于所述三元数组将位于同一页同一行的第一文本框合并, 得到多个行文本框;

B3、将所述多个行文本框划分为包含预设数量文本框的多个第一文本框集合。

将 OCR 识别得到的文本框中位于同一页同一行的文本框合并后, 每页图片至少对应从上往下排列的 15 个文本框, 当待处理文档页数较多时, 得到的文本框数量较多, 现有的序列模型无法一次性建模这么多数据(大部分序列模型的输入长度为 512, 当输入文字长度大于 512 时, 需要对输入文字进行拆分), 将文本框划分为包含预设数量(例如, 10

个)文本框的一个或多个文本框集合,可避免信息在输入模型时溢出,保证了后续文本结构类别及标题等级识别的准确性。

确定模块 130,用于将所述多个第一文本框集合输入结构类别识别模型,得到所述第一文本框集合中每个文本框的结构类别信息,根据所述结构类别信息确定所述待处理文档的多个第一标题。

本实施例中,所述结构类别识别模型为训练好的 LayoutLM 模型,用于识别第一文本框集合中每个文本框的结构类别(结构类别包括:居中标题、分级标题、正文),根据结构类别可确定所述待处理文档的多个标题(包括居中标题、分级标题),但是不能区分分级标题为几级分级标题。

LayoutLM 模型是一种通用文档预训练模型,对文档结构信息和视觉信息进行建模,文档中文字的位置关系蕴含着丰富的语义信息,例如,居中标题通常位于文档首行正中位置;文字格式所呈现的视觉信息也体现了语义信息,例如,文章的分级标题通常会放大加粗呈现,特殊概念名称会以斜体呈现。

本实施例中,所述结构类别识别模型的确定过程包括:

C1、对携带结构类别标注信息的第一图片样本执行文字识别得到多个第二文本框,对所述多个第二文本框执行分组处理得到多个第二文本框集合;

C2、将所述第二文本框集合输入第一 LayoutLM 模型,得到所述第二文本框集合中每个文本框的预测结构类别;

C3、基于所述结构类别标注信息确定所述第二文本框集合中每个文本框的真实结构类别,通过最小化预测结构类别与真实结构类别之间的第一损失值确定所述第一 LayoutLM 模型的结构参数,得到结构类别识别模型。

所述将所述第二文本框集合输入第一 LayoutLM 模型包括:

D1、基于所述三元数组将所述第二文本框集合中的文本框按照位置顺序通过连接符进行连接,得到第一文本序列;

D2、在所述第一文本序列的开头增加句首符,在所述第一文本序列的结尾增加连接符,得到第二文本序列,将所述第二文本序列输入第一 LayoutLM 模型。

假设第二文本框集合中包括 10 个文本框,本实施例中句首符用 CLS 表示,连接符用 SEP 表示,则第二文本序列为 CLS,W11,W12,...,W1n₁,SEP,W21,W22,...,W2n₂,SEP,...,SEP,W101,W102,...,W10n₁₀,SEP。其中 W_{ij} 表示第二文本框集合中第 i 个文本框的第 j 个文字,n_i 表示第二文本框集合中第 i 个文本框中文字的总数量。

当输入经过 LayoutLM 模型后,会得到相同长度的输出,每个文本框后面的 SEP 所输出的结构类别即为该文本框的结构类别。

所述第一损失值的计算公式为:

$$\text{loss}(q_{mn}, p_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c p_{mn} (\log q_{mn})$$

其中,q_{mn}为第 m 个第二文本框集合中第 n 个文本框的预测结构类别,p_{mn}为第 m 个第二文本框集合中第 n 个文本框的真实结构类别,loss(q_{mn},p_{mn})为第 m 个第二文本框集合中第 n 个文本框的预测结构类别和真实结构类别之间的第一损失值,c 为第 m 个第二文本框集合中文本框的总数量,t 为结构类别的总数量(本实施例中,结构类别的总数量为 3,包括:居中标题、分级标题、正文)。

生成模块 140,用于将所述多个第一标题输入标题等级识别模型,得到所述多个第一标题中每个第一标题的等级信息,根据所述等级信息为所述待处理文档生成目录信息。

本实施例中,所述标题等级识别模型也为训练好的 LayoutLM 模型,用于识别分级标题为几级分级标题(例如,一级标题、二级标题、三级标题...)。本申请中标题等级识别

模型与文本结构类别识别模型输入的样本不同，训练目标不同，训练得到的模型参数也不同。

本实施例在识别得到居中标题和各级分级标题后即可创建出较为详细的目录信息，用户根据目录信息可确定文档的整体布局，快速定位自己感兴趣的部分。

本实施例中，不将文本结构类别识别模型和标题等级识别模型结合在一起作为一个模型进行文本结构类别识别和标题等级识别的目的在于：对于大数据量的文档，OCR 识别后得到的文本框数量较多，LayoutLM 模型无法一次性将所有文本框输入，所以将文本框进行分组，得到多个文本框集合，然而因居中标题和分级标题在一个文档中的数量比较稀少，一个文本框集合中可能只有一个文本框对应分级标题，也可能没有一个文本框对应分级标题，而分级标题的等级的识别需要关注上下文信息，故而无法根据一个文本框集合来确定分级标题具体为几级分级标题，采用一个模型来识别结构类别和标题等级会使标题等级识别准确度不够高。本实施例通过结构类别识别模型得到多个第一标题，因第一标题的数量不会太多，可一次性输入标题等级识别模型，完整的保留了上下文信息，保证了标题等级识别模型的高准确性。

所述标题等级识别模型的确定过程包括：

E1、对携带标题等级标注信息的第二图片样本执行文字识别得到多个第三文本框，对所述多个第三文本框执行分组处理得到多个第三文本框集合；

E2、将所述第三文本框集合输入文本结构类别识别模型，得到所述第三文本框集合中每个文本框的预测结构类别，基于所述预测结构类别得到所述第二图片样本对应的多个第二标题；

E3、将所述多个第二标题输入第二 LayoutLM 模型，得到每个第二标题对应的预测标题等级，基于所述标题等级标注信息确定每个第二标题的真实标题等级，通过最小化预测标题等级与真实标题等级之间的第二损失值确定所述第二 LayoutLM 模型的结构参数，得到标题等级识别模型。

所述第二损失值的计算公式为：

$$\text{loss}(q_{uv}, p_{uv}) = -\frac{1}{h} \sum_v^g \sum_u^h p_{uv} (\log q_{uv})$$

其中， q_{uv} 为第 u 个第二图片样本的第 v 个第二标题的预测标题等级， p_{uv} 为第 u 个第二图片样本的第 v 个第二标题的真实标题等级， $\text{loss}(q_{uv}, p_{uv})$ 为第 u 个第二图片样本的第 v 个第二标题的预测标题等级和真实标题等级之间的第二损失值， h 为第 u 个第二图片样本的第二标题的总数量， g 为标题等级的总数量（本实施例中，标题等级的总数量为 8，包括：一级标题、二级标题、...、八级标题）。

如图 3 所示，为本申请一实施例提供的实现文档目录生成方法的电子设备的结构示意图。

所述电子设备 1 是一种能够按照事先设定或者存储的指令，自动进行数值计算和/或信息处理的设备。所述电子设备 1 可以是计算机、也可以是单个网络服务器、多个网络服务器组成的服务器组或者基于云计算的由大量主机或者网络服务器构成的云，其中云计算是分布式计算的一种，由一群松散耦合的计算机集组成的一个超级虚拟计算机。

在本实施例中，电子设备 1 包括，但不限于，可通过系统总线相互通信连接的存储器 11、处理器 12、网络接口 13，该存储器 11 中存储有文档目录生成程序 10，所述文档目录生成程序 10 可被所述处理器 12 执行。图 3 仅示出了具有组件 11-13 以及文档目录生成程序 10 的电子设备的结构，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。

其中，存储器 11 包括内存及至少一种类型的可读存储介质。内存为电子设备 1 的运

行提供缓存；可读存储介质可为如闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、随机访问存储器（RAM）、静态随机访问存储器（SRAM）、只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、可编程只读存储器（PROM）、磁性存储器、磁盘、光盘等的非易失性存储介质。在一些实施例中，可读存储介质可以是电子设备 1 的内部存储单元，例如该电子设备 1 的硬盘；在另一些实施例中，该非易失性存储介质也可以是电子设备 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。本实施例中，存储器 11 的可读存储介质通常用于存储安装于电子设备 1 的操作系统和各类应用软件，例如存储本申请一实施例中的文档目录生成程序 10 的代码等。此外，存储器 11 还可以用于暂时地存储已经输出或者将要输出的各类数据。

处理器 12 在一些实施例中可以是中央处理器（Central Processing Unit, CPU）、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器 12 通常用于控制所述电子设备 1 的总体操作，例如执行与其他设备进行数据交互或者通信相关的控制和处理等。本实施例中，所述处理器 12 用于运行所述存储器 11 中存储的程序代码或者处理数据，例如运行文档目录生成程序 10 等。

网络接口 13 可包括无线网络接口或有线网络接口，该网络接口 13 用于在所述电子设备 1 与客户端（图中未画出）之间建立通信连接。

可选的，所述电子设备 1 还可以包括用户接口，用户接口可以包括显示器（Display）、输入单元比如键盘（Keyboard），可选的用户接口还可以包括标准的有线接口、无线接口。可选的，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。

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。

所述电子设备 1 中的所述存储器 11 存储的文档目录生成程序 10 是多个指令的组合，在所述处理器 12 中运行时，可以实现：

响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待处理文档；

将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合；

将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题；

将所述多个第一标题输入标题等级识别模型，得到所述多个第一标题中每个第一标题的等级信息，根据所述等级信息为所述待处理文档生成目录信息。

具体地，所述处理器 12 对上述文档目录生成程序 10 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。需要强调的是，为进一步保证上述待处理文档的私密和安全性，上述待处理文档还可以存储于一区块链的节点中。

进一步地，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，可以存储在一个计算机可读存储介质中。所述计算机可读存储介质可以是非易失性的，也可以是非易失性的。所述计算机可读存储介质可以包括：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）。

所述计算机可读存储介质上存储有文档目录生成程序 10，所述文档目录生成程序 10 可被一个或者多个处理器执行以实现如下步骤：

响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待

处理文档；

将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合；

将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题；

将所述多个第一标题输入标题等级识别模型，得到所述多个第一标题中每个第一标题的等级信息，根据所述等级信息为所述待处理文档生成目录信息。

在本申请所提供的几个实施例中，应该理解到，所揭露的设备，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。

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。

另外，在本申请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。

对于本领域技术人员而言，显然本申请不限于上述示范性实施例的细节，而且在不背离本申请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申请。

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，本申请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申请内。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。

本申请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。

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。

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

权利要求书

1、一种文档目录生成方法，其中，所述方法包括：

响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待处理文档；

将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合；

将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题；

将所述多个第一标题输入标题等级识别模型，得到所述多个第一标题中每个第一标题的等级信息，根据所述等级信息为所述待处理文档生成目录信息。

2、如权利要求 1 所述的文档目录生成方法，其中，所述对所述多个第一文本框执行分组处理得到多个第一文本框集合包括：

为所述多个第一文本框中的每个第一文本框建立一个三元数组，所述三元数组包括第一文本框所属图片的页码、第一文本框的左上角和右下角坐标、第一文本框中的文字信息；

基于所述三元数组将位于同一页同一行的第一文本框合并，得到多个行文本框；

将所述多个行文本框划分为包含预设数量文本框的多个第一文本框集合。

3、如权利要求 2 所述的文档目录生成方法，其中，所述结构类别识别模型的确定过程包括：

对携带结构类别标注信息的第一图片样本执行文字识别得到多个第二文本框，对所述多个第二文本框执行分组处理得到多个第二文本框集合；

将所述第二文本框集合输入第一 LayoutLM 模型，得到所述第二文本框集合中每个文本框的预测结构类别；

基于所述结构类别标注信息确定所述第二文本框集合中每个文本框的真实结构类别，通过最小化预测结构类别与真实结构类别之间的第一损失值确定所述第一 LayoutLM 模型的结构参数，得到结构类别识别模型。

4、如权利要求 3 所述的文档目录生成方法，其中，所述将所述第二文本框集合输入第一 LayoutLM 模型包括：

基于所述三元数组将所述第二文本框集合中的文本框按照位置顺序通过连接符进行连接，得到第一文本序列；

在所述第一文本序列的开头增加句首符，在所述第一文本序列的结尾增加连接符，得到第二文本序列，将所述第二文本序列输入第一 LayoutLM 模型。

5、如权利要求 4 所述的文档目录生成方法，其中，所述标题等级识别模型的确定过程包括：

对携带标题等级标注信息的第二图片样本执行文字识别得到多个第三文本框，对所述多个第三文本框执行分组处理得到多个第三文本框集合；

将所述第三文本框集合输入文本结构类别识别模型，得到所述第三文本框集合中每个文本框的预测结构类别，基于所述预测结构类别得到所述第二图片样本对应的多个第二标题；

将所述多个第二标题输入第二 LayoutLM 模型，得到每个第二标题对应的预测标题等级，基于所述标题等级标注信息确定每个第二标题的真实标题等级，通过最小化预测标题等级与真实标题等级之间的第二损失值确定所述第二 LayoutLM 模型的结构参数，得到标题等级识别模型。

6、如权利要求 5 所述的文档目录生成方法，其中，所述第一损失值的计算公式为：

$$\text{loss}(q_{mn}, p_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c p_{mn} (\log q_{mn})$$

10

其中, q_{mn} 为第 m 个第二文本框集合中第 n 个文本框的预测结构类别, p_{mn} 为第 m 个第二文本框集合中第 n 个文本框的真实结构类别, $\text{loss}(q_{mn}, p_{mn})$ 为第 m 个第二文本框集合中第 n 个文本框的预测结构类别和真实结构类别之间的第一损失值, c 为第 m 个第二文本框集合中文本框的总数量, t 为结构类别的总数量;

所述第二损失值的计算公式为:

$$\text{loss}(q_{uv}, p_{uv}) = -\frac{1}{h} \sum_v^g \sum_u^h p_{uv} (\log q_{uv})$$

其中, q_{uv} 为第 u 个第二图片样本的第 v 个第二标题的预测标题等级, p_{uv} 为第 u 个第二图片样本的第 v 个第二标题的真实标题等级, $\text{loss}(q_{uv}, p_{uv})$ 为第 u 个第二图片样本的第 v 个第二标题的预测标题等级和真实标题等级之间的第二损失值, h 为第 u 个第二图片样本的第二标题的总数量, g 为标题等级的总数量。

7、一种文档目录生成装置, 其中, 所述装置包括:

获取模块, 用于响应用户基于客户端发出的文档目录生成请求, 解析所述请求, 获取所述请求中的待处理文档;

分组模块, 用于将所述待处理文档转换为图片格式文档, 对所述图片格式文档执行 OCR 识别得到多个第一文本框, 对所述多个第一文本框执行分组处理得到多个第一文本框集合;

确定模块, 用于将所述多个第一文本框集合输入结构类别识别模型, 得到所述第一文本框集合中每个文本框的结构类别信息, 根据所述结构类别信息确定所述待处理文档的多个第一标题;

生成模块, 用于将所述多个第一标题输入标题等级识别模型, 得到所述多个第一标题中每个第一标题的等级信息, 根据所述等级信息为所述待处理文档生成目录信息。

8、如权利要求 7 所述的文档目录生成装置, 其中, 所述对所述多个第一文本框执行分组处理得到多个第一文本框集合包括:

为所述多个第一文本框中的每个第一文本框建立一个三元数组, 所述三元数组包括第一文本框所属图片的页码、第一文本框的左上角和右下角坐标、第一文本框中的文字信息;

基于所述三元数组将位于同一页同一行的第一文本框合并, 得到多个行文本框;

将所述多个行文本框划分为包含预设数量文本框的多个第一文本框集合。

9、一种电子设备, 其中, 所述电子设备包括:

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文档目录生成程序, 所述文档目录生成程序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如下步骤:

响应用户基于客户端发出的文档目录生成请求, 解析所述请求, 获取所述请求中的待处理文档;

将所述待处理文档转换为图片格式文档, 对所述图片格式文档执行 OCR 识别得到多个第一文本框, 对所述多个第一文本框执行分组处理得到多个第一文本框集合;

将所述多个第一文本框集合输入结构类别识别模型, 得到所述第一文本框集合中每个文本框的结构类别信息, 根据所述结构类别信息确定所述待处理文档的多个第一标题;

将所述多个第一标题输入标题等级识别模型, 得到所述多个第一标题中每个第一标题的等级信息, 根据所述等级信息为所述待处理文档生成目录信息。

10、如权利要求 9 所述的电子设备, 其中, 所述对所述多个第一文本框执行分组处理得到多个第一文本框集合包括:

为所述多个第一文本框中的每个第一文本框建立一个三元数组, 所述三元数组包括第一文本框所属图片的页码、第一文本框的左上角和右下角坐标、第一文本框中的文字信息;

基于所述三元数组将位于同一页同一行的第一文本框合并，得到多个行文本框；
将所述多个行文本框划分为包含预设数量文本框的多个第一文本框集合。

11、如权利要求 10 所述的电子设备，其中，所述结构类别识别模型的确定过程包括：

对携带结构类别标注信息的第一图片样本执行文字识别得到多个第二文本框，对所述多个第二文本框执行分组处理得到多个第二文本框集合；

将所述第二文本框集合输入第一 LayoutLM 模型，得到所述第二文本框集合中每个文本框的预测结构类别；

基于所述结构类别标注信息确定所述第二文本框集合中每个文本框的真实结构类别，通过最小化预测结构类别与真实结构类别之间的第一损失值确定所述第一 LayoutLM 模型的结构参数，得到结构类别识别模型。

12、如权利要求 11 所述的电子设备，其中，所述将所述第二文本框集合输入第一 LayoutLM 模型包括：

基于所述三元数组将所述第二文本框集合中的文本框按照位置顺序通过连接符进行连接，得到第一文本序列；

在所述第一文本序列的开头增加句首符，在所述第一文本序列的结尾增加连接符，得到第二文本序列，将所述第二文本序列输入第一 LayoutLM 模型。

13、如权利要求 12 所述的电子设备，其中，所述标题等级识别模型的确定过程包括：

对携带标题等级标注信息的第二图片样本执行文字识别得到多个第三文本框，对所述多个第三文本框执行分组处理得到多个第三文本框集合；

将所述第三文本框集合输入文本结构类别识别模型，得到所述第三文本框集合中每个文本框的预测结构类别，基于所述预测结构类别得到所述第二图片样本对应的多个第二标题；

将所述多个第二标题输入第二 LayoutLM 模型，得到每个第二标题对应的预测标题等级，基于所述标题等级标注信息确定每个第二标题的真实标题等级，通过最小化预测标题等级与真实标题等级之间的第二损失值确定所述第二 LayoutLM 模型的结构参数，得到标题等级识别模型。

14、如权利要求 13 所述的电子设备，其中，所述第一损失值的计算公式为：

$$\text{loss}(q_{mn}, p_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c p_{mn} (\log q_{mn})$$

其中， q_{mn} 为第 m 个第二文本框集合中第 n 个文本框的预测结构类别， p_{mn} 为第 m 个第二文本框集合中第 n 个文本框的真实结构类别， $\text{loss}(q_{mn}, p_{mn})$ 为第 m 个第二文本框集合中第 n 个文本框的预测结构类别和真实结构类别之间的第一损失值， c 为第 m 个第二文本框集合中文本框的总数量， t 为结构类别的总数量；

所述第二损失值的计算公式为：

$$\text{loss}(q_{uv}, p_{uv}) = -\frac{1}{h} \sum_v^g \sum_u^h p_{uv} (\log q_{uv})$$

其中， q_{uv} 为第 u 个第二图片样本的第 v 个第二标题的预测标题等级， p_{uv} 为第 u 个第二图片样本的第 v 个第二标题的真实标题等级， $\text{loss}(q_{uv}, p_{uv})$ 为第 u 个第二图片样本的第 v 个第二标题的预测标题等级和真实标题等级之间的第二损失值， h 为第 u 个第二图片样本的第二标题的总数量， g 为标题等级的总数量。

15、一种计算机可读存储介质，其中，所述计算机可读存储介质上存储有文档目录生成程序，所述文档目录生成程序可被一个或者多个处理器执行，以实现如下步骤：

响应用户基于客户端发出的文档目录生成请求，解析所述请求，获取所述请求中的待处理文档；

将所述待处理文档转换为图片格式文档，对所述图片格式文档执行 OCR 识别得到多个第一文本框，对所述多个第一文本框执行分组处理得到多个第一文本框集合；

将所述多个第一文本框集合输入结构类别识别模型，得到所述第一文本框集合中每个文本框的结构类别信息，根据所述结构类别信息确定所述待处理文档的多个第一标题；

将所述多个第一标题输入标题等级识别模型，得到所述多个第一标题中每个第一标题的等级信息，根据所述等级信息为所述待处理文档生成目录信息。

16、如权利要求 15 所述的计算机可读存储介质，其中，所述对所述多个第一文本框执行分组处理得到多个第一文本框集合包括：

为所述多个第一文本框中的每个第一文本框建立一个三元数组，所述三元数组包括第一文本框所属图片的页码、第一文本框的左上角和右下角坐标、第一文本框中的文字信息；

基于所述三元数组将位于同一页同一行的第一文本框合并，得到多个行文本框；

将所述多个行文本框划分为包含预设数量文本框的多个第一文本框集合。

17、如权利要求 16 所述的计算机可读存储介质，其中，所述结构类别识别模型的确定过程包括：

对携带结构类别标注信息的第一图片样本执行文字识别得到多个第二文本框，对所述多个第二文本框执行分组处理得到多个第二文本框集合；

将所述第二文本框集合输入第一 LayoutLM 模型，得到所述第二文本框集合中每个文本框的预测结构类别；

基于所述结构类别标注信息确定所述第二文本框集合中每个文本框的真实结构类别，通过最小化预测结构类别与真实结构类别之间的第一损失值确定所述第一 LayoutLM 模型的结构参数，得到结构类别识别模型。

18、如权利要求 17 所述的计算机可读存储介质，其中，所述将所述第二文本框集合输入第一 LayoutLM 模型包括：

基于所述三元数组将所述第二文本框集合中的文本框按照位置顺序通过连接符进行连接，得到第一文本序列；

在所述第一文本序列的开头增加句首符，在所述第一文本序列的结尾增加连接符，得到第二文本序列，将所述第二文本序列输入第一 LayoutLM 模型。

19、如权利要求 18 所述的计算机可读存储介质，其中，所述标题等级识别模型的确定过程包括：

对携带标题等级标注信息的第二图片样本执行文字识别得到多个第三文本框，对所述多个第三文本框执行分组处理得到多个第三文本框集合；

将所述第三文本框集合输入文本结构类别识别模型，得到所述第三文本框集合中每个文本框的预测结构类别，基于所述预测结构类别得到所述第二图片样本对应的多个第二标题；

将所述多个第二标题输入第二 LayoutLM 模型，得到每个第二标题对应的预测标题等级，基于所述标题等级标注信息确定每个第二标题的真实标题等级，通过最小化预测标题等级与真实标题等级之间的第二损失值确定所述第二 LayoutLM 模型的结构参数，得到标题等级识别模型。

20、如权利要求 19 所述的计算机可读存储介质，其中，所述第一损失值的计算公式为：

$$\text{loss}(q_{mn}, p_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c p_{mn} (\log q_{mn})$$

其中， q_{mn} 为第 m 个第二文本框集合中第 n 个文本框的预测结构类别， p_{mn} 为第 m 个第二文本框集合中第 n 个文本框的真实结构类别， $\text{loss}(q_{mn}, p_{mn})$ 为第 m 个第二文本框集合中第 n 个文本框的预测结构类别和真实结构类别之间的第一损失值， c 为第 m 个第二文本框集合中文本框的总数量， t 为结构类别的总数量；

所述第二损失值的计算公式为：

$$\text{loss}(q_{uv}, p_{uv}) = -\frac{1}{h} \sum_v^g \sum_u^h p_{uv} (\log q_{uv})$$

其中， q_{uv} 为第 u 个第二图片样本的第 v 个第二标题的预测标题等级， p_{uv} 为第 u 个第二图片样本的第 v 个第二标题的真实标题等级， $\text{loss}(q_{uv}, p_{uv})$ 为第 u 个第二图片样本的第 v 个第二标题的的预测标题等级和真实标题等级之间的第二损失值， h 为第 u 个第二图片样本的第二标题的总数量， g 为标题等级的总数量。

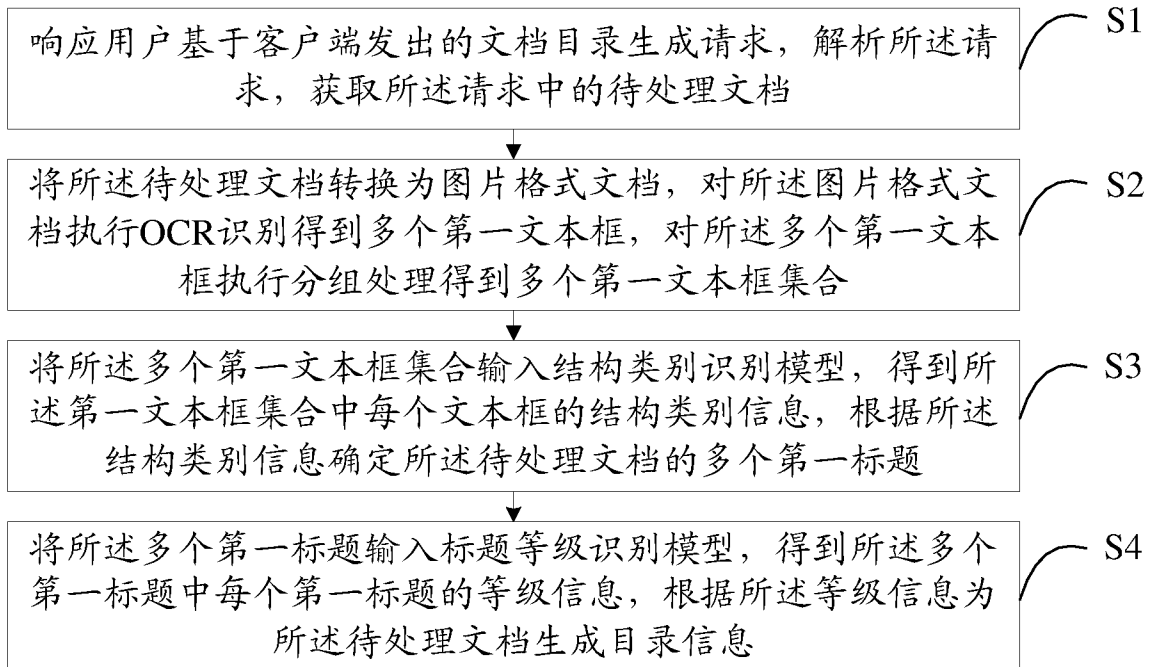


图 1

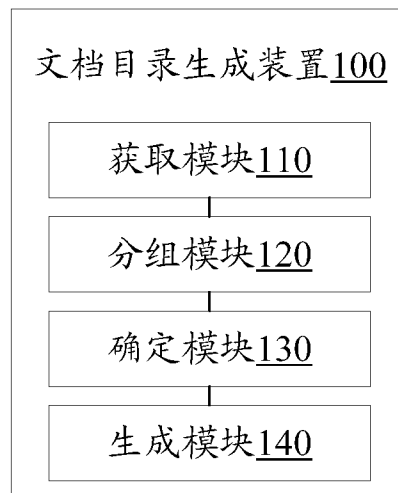


图 2

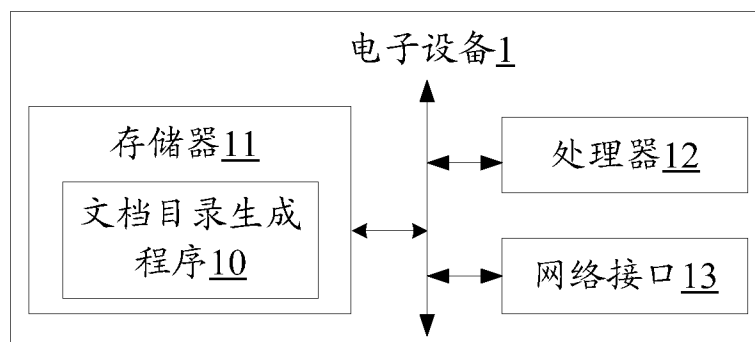


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/097070

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/151(2020.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI: 文档, 目录, 结构, 标题, 模型, 等级, 级别, 文本框, 坐标, 类别, document, title, directory, model, textbox, ocr, structure, category

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112016273 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 01 December 2020 (2020-12-01) claims 1-10, description paragraphs [0093]-[0149]	1-20
Y	CN 110852079 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 28 February 2020 (2020-02-28) description, paragraphs [0044]-[0087]	1, 7, 9, 15
Y	CN 110427884 A (DAERGUAN INFORMATION TECHNOLOGY (SHANGHAI) CO., LTD.) 08 November 2019 (2019-11-08) description, paragraphs [0028]-[0087]	1, 7, 9, 15
A	US 2019171704 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 06 June 2019 (2019-06-06) entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

10 August 2021

Date of mailing of the international search report

26 August 2021

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/097070

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	112016273	A	01 December 2020	None	
CN	110852079	A	28 February 2020	None	
CN	110427884	A	08 November 2019	None	
US	2019171704	A1	06 June 2019	None	

国际检索报告

国际申请号

PCT/CN2021/097070

A. 主题的分类 G06F 40/151 (2020.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																	
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) WPI, EPDOC, CNPAT, CNKI: 文档, 目录, 结构, 标题, 模型, 等级, 级别, 文本框, 坐标, 类别, document, title, directory, model, textbox, ocr, structure, category																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 112016273 A (平安科技深圳有限公司) 2020年 12月 1日 (2020 - 12 - 01) 权利要求1-10, 说明书第[0093]-[0149]段</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 110852079 A (平安科技深圳有限公司) 2020年 2月 28日 (2020 - 02 - 28) 说明书第[0044]-[0087]段</td> <td>1, 7, 9, 15</td> </tr> <tr> <td>Y</td> <td>CN 110427884 A (达观信息科技上海有限公司) 2019年 11月 8日 (2019 - 11 - 08) 说明书第[0028]-[0087]段</td> <td>1, 7, 9, 15</td> </tr> <tr> <td>A</td> <td>US 2019171704 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2019年 6月 6日 (2019 - 06 - 06) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 112016273 A (平安科技深圳有限公司) 2020年 12月 1日 (2020 - 12 - 01) 权利要求1-10, 说明书第[0093]-[0149]段	1-20	Y	CN 110852079 A (平安科技深圳有限公司) 2020年 2月 28日 (2020 - 02 - 28) 说明书第[0044]-[0087]段	1, 7, 9, 15	Y	CN 110427884 A (达观信息科技上海有限公司) 2019年 11月 8日 (2019 - 11 - 08) 说明书第[0028]-[0087]段	1, 7, 9, 15	A	US 2019171704 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2019年 6月 6日 (2019 - 06 - 06) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
PX	CN 112016273 A (平安科技深圳有限公司) 2020年 12月 1日 (2020 - 12 - 01) 权利要求1-10, 说明书第[0093]-[0149]段	1-20															
Y	CN 110852079 A (平安科技深圳有限公司) 2020年 2月 28日 (2020 - 02 - 28) 说明书第[0044]-[0087]段	1, 7, 9, 15															
Y	CN 110427884 A (达观信息科技上海有限公司) 2019年 11月 8日 (2019 - 11 - 08) 说明书第[0028]-[0087]段	1, 7, 9, 15															
A	US 2019171704 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2019年 6月 6日 (2019 - 06 - 06) 全文	1-20															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
<table border="0"> <tr> <td> * 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 </td> <td> “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件 </td> </tr> </table>			* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件													
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件	“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																
国际检索实际完成的日期		国际检索报告邮寄日期															
2021年 8月 10日		2021年 8月 26日															
ISA/CN的名称和邮寄地址		受权官员															
中国国家知识产权局 (ISA/CN) 中国 北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		田晶 电话号码 86-(10)-53961334															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/097070

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	112016273	A	2020年 12月 1日	无	
CN	110852079	A	2020年 2月 28日	无	
CN	110427884	A	2019年 11月 8日	无	
US	2019171704	A1	2019年 6月 6日	无	