(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
2019 年 4 月 25 日 (25.04.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/075969 A1

(51) 国际专利分类号:
G06F 17/24 (2006.01)

(21) 国际申请号: PCT/CN2018/076167

(22) 国际申请日: 2018 年 2 月 10 日 (10.02.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710962303.5 2017年10月16日 (16.10.2017) 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八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。

(72) 发明人: 苏晓明 (SU, Xiaoming); 中国广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。 汪伟 (WANG, Wei); 中国广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。 肖京 (XIAO, Jing); 中国广东省深圳市福田区八卦岭工业区平安大厦六楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所（普通合伙） (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路10号中浩大厦1528-1530室于志光, 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, 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,

(54) **Title:** METHOD FOR EXTRACTING FORM INFORMATION IN A STRUCTURED MANNER, ELECTRONIC DEVICE, AND COMPUTER-READABLE STORAGE MEDIUM

(54) **发明名称**: 表格信息结构化提取方法、电子设备及计算机可读存储介质

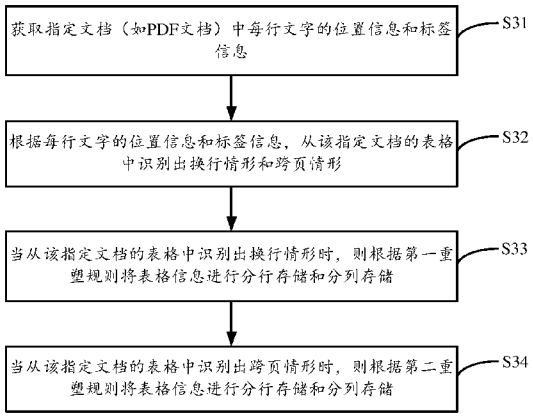


图 3

S31 ACQUIRE POSITION INFORMATION AND LABEL INFORMATION ABOUT EACH ROW OF CHARACTERS IN A SPECIFIED DOCUMENT (SUCH AS A PDF DOCUMENT)
S32 ACCORDING TO THE POSITION INFORMATION AND LABEL INFORMATION ABOUT EACH ROW OF CHARACTERS, RECOGNIZE A LINE WRAP SITUATION AND A PAGE-CROSSING SITUATION FROM A FORM OF THE SPECIFIED DOCUMENT
S33 WHEN A LINE WRAP SITUATION IS RECOGNIZED FROM THE FORM OF THE SPECIFIED DOCUMENT, STORE INFORMATION IN THE FORM IN ROWS AND IN COLUMNS ACCORDING TO A FIRST RECONSTRUCTION RULE
S34 WHEN A PAGE-CROSSING SITUATION IS RECOGNIZED FROM THE FORM OF THE SPECIFIED DOCUMENT, THEN STORE INFORMATION IN THE FORM IN ROWS AND IN COLUMNS ACCORDING TO A SECOND RECONSTRUCTION RULE

(57) **Abstract:** A method for extracting form information in a structured manner. The method comprises the following steps: acquiring position information and label information about each row of characters in a specified document (such as a PDF document) (S31); according to the position information and label information about each row of characters, recognizing a line wrap situation and a page-crossing situation from a form of the specified document (S32); when a line wrap situation is recognized from the form of the specified document, storing information in the form in rows and in columns according to a first reconstruction rule (S33); and when a page-crossing situation is recognized from the form of the specified document, then storing information in the form in rows and in columns according to a second reconstruction rule (S34). By means of the method, data can be extracted and stored in a structured manner.

(57) **摘要**: 一种表格信息结构化提取方法，该方法包括步骤：获取指定文档（如PDF文档）中每行文字的位置信息和标签信息（S31）；根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形（S32）；当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储和分列存储（S33）；当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储（S34）。通过该方法可以进行结构化数据提取和存储。

WO 2019/075969 A1

[见续页]

PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(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))。

表格信息结构化提取方法、电子设备及计算机可读存储介质

本申请要求于2017年10月16日提交中国专利局、申请号为201710962303.5、发明名称为“表格信息结构化提取方法、电子设备及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在申请中。

技术领域

本申请涉及计算机信息技术领域，尤其涉及一种表格信息结构化提取方法、电子设备及计算机可读存储介质。

背景技术

现有针对PDF年报中的表格信息抽取，一般是基于OCR技术。但是，在出现换行、换页、特殊字符干扰情况下，OCR技术无法还原和重塑原始表格信息，及进一步的结构化整合，并对使用者造成理解上的困难，也不利于后续的信息比对。故，现有技术中的表格信息提取方法设计不够合理，亟需改进。

发明内容

有鉴于此，本申请提出一种表格信息结构化提取方法、电子设备及计算机可读存储介质，通过分析指定文档（如PDF文档）中表格文字内容的位置信息和标签信息，可以识别出表格（如PDF年报表格）中的换行情形和跨页情形，并对出现换行情形和跨页情形的表格进行结构化数据提取并存储。

首先，为实现上述目的，本申请提出一种电子设备，所述电子设备包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的表格信息结构化提取系统，所述表格信息结构化提取系统被所述处理器执行时实现如下步骤：

获取指定文档中每行文字的位置信息和标签信息；

根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形；

当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储和分列存储；及

当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储。

优选地，所述第一重塑规则包括：将上边沿坐标相同的文字存储为同一行，将左边沿坐标相同的文字存储为同一列；

所述第二重塑规则包括：

删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉；

将删除页脚后的前一表格文字内容与删除页眉后的下一表格文字内容进行拼接形成拼接表格；及

将所述拼接表格中上边沿坐标相同的文字存储为同一行，并将所述拼接表格中左边沿坐标相同的文字存储为同一列。

优选地，所述删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉包括：

根据前一页面和下一页面的标签信息、及该指定文档的特定规则，对该指定文档中前一页面的页脚范围和下一页面的页眉范围进行定位，并根据定位确定的页脚范围和页眉范围删除前一页面的页脚和下一页面的页眉；

其中，前一页面的页脚范围的确定规则为：依据从下至上的顺序，选取所述前一页面的页面长度的第一比例内容，将选取的第一比例内容作为所述前一页面的页脚范围；及

下一页面的页眉范围的确定规则为：依据从上至下的顺序，选取所述下一页面的页面长度的第二比例内容，将选取的第二比例内容作为所述下一页面的页眉范围。

优选地，所述换行情形包括行内换行和行尾换行；

其中，所述行内换行的识别包括：

获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的文字内容位置信息包括每个单元格的文字内容的上边沿坐标；及

从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置和最后一次出现的单元格位置，将第一次出现的单元格位置到最后一次出现的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

优选地，所述行尾换行的识别包括：

若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息，其中，所述剩余单元格的文字内容位置信息包括所述剩余单元格的文字内容的上边沿坐标；

计算所述剩余单元格的文字内容上边沿坐标与当前行及下一行所有单元格的文字内容上边沿坐标的距离，或计算所述剩余单元格的文字内容上边沿坐标与当前行及上一行所有单元格的文字内容上边沿坐标的距离；及

若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

此外，为实现上述目的，本申请还提供一种表格信息结构化提取方法，该方法应用于电子设备，所述方法包括：

获取指定文档中每行文字的位置信息和标签信息；

根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形；

当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储和分列存储；及

当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储。

优选地，所述第一重塑规则包括：将上边沿坐标相同的文字存储为同一行，将左边沿坐标相同的文字存储为同一列；

所述第二重塑规则包括：

删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉；

将删除页脚后的前一表格文字内容与删除页眉后的下一表格文字内容进行拼接形成拼接表格；及

将所述拼接表格中上边沿坐标相同的文字存储为同一行，并将所述拼接表格中左边沿坐标相同的文字存储为同一列。

优选地，所述删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉包括：

根据前一页面和下一页面的标签信息、及该指定文档的特定规则，对该指定文档中前一页面的页脚范围和下一页面的页眉范围进行定位，并根据定位确定的页脚范围和页眉范围删除前一页面的页脚和下一页面的页眉；

其中，前一页面的页脚范围的确定规则为：依据从下至上的顺序，选取所述前一页面的页面长度的第一比例内容，将选取的第一比例内容作为所述前一页面的页脚范围；及

下一页面的页眉范围的确定规则为：依据从上至下的顺序，选取所述下一页面的页面长度的第二比例内容，将选取的第二比例内容作为所述下一页面的页眉范围。

优选地，所述跨页情形的识别包括：

针对该指定文档中相邻的前一表格和下一表格，获取前一表格文字内容的位置信息、标签信息和下一表格文字内容的位置信息、标签信息；

比对下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标；

当下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标都相同时，比对下一表格每行文字的页码与前一表格每行文字的页码；及

若下一表格每行文字的页码与前一表格每行文字的页码存在不同，则判定下一表格与前一表格为存在跨页情形的同一表格。

进一步地，为实现上述目的，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存储有表格信息结构化提取系统，所述表格信息结构化提取系统可被至少一个处理器执行，以使所述至少一个处理器执行如上述的表格信息结构化提取方法的步骤。

相较于现有技术，本申请所提出的电子设备、表格信息结构化提取方法及计算机可读存储介质，通过分析指定文档（如 PDF 文档）中表格文字内容的位置信息和标签信息，可以识别出表格（如 PDF 年报表格）中的换行情形

和跨页情形，并对出现换行情形和跨页情形的表格进行结构化数据提取并存储。该方法无需将 PDF 文件转化为 word、excel 等结构化文档，数据提取效率高，大规模数据集下召回率和准确率较高，且利于后续横向比对分析、纵向比对分析及数据建模。

附图说明

图1是本申请电子设备一可选的硬件架构的示意图；

图2是本申请电子设备中表格信息结构化提取系统一实施例的程序模块示意图；

图3为本申请表格信息结构化提取方法一实施例的实施流程示意图。

附图标记：

电子设备	2
存储器	21
处理器	22
网络接口	23
表格信息结构化提取系统	20
获取模块	201
识别模块	202
存储模块	203
流程步骤	S31-S34

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。

进一步需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者装置不仅包括那些要素，而且还包括没有明确列出的其他要素，或

者是还包括为这种过程、方法、物品或者装置所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

首先，本申请提出一种电子设备 2。

参阅图 1 所示，是本申请电子设备 2 一可选的硬件架构的示意图。本实施例中，所述电子设备 2 可包括，但不限于，可通过系统总线相互通信连接存储器 21、处理器 22、网络接口 23。需要指出的是，图 1 仅示出了具有组件 21-23 的电子设备 2，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

其中，所述电子设备 2 可以是机架式服务器、刀片式服务器、塔式服务器或机柜式服务器等计算设备，该电子设备 2 可以是独立的服务器，也可以是多个服务器所组成的服务器集群。

所述存储器 21 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、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。在一些实施例中，所述存储器 21 可以是所述电子设备 2 的内部存储单元，例如该电子设备 2 的硬盘或内存。在另一些实施例中，所述存储器 21 也可以是所述电子设备 2 的外部存储设备，例如该电子设备 2 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。当然，所述存储器 21 还可以既包括所述电子设备 2 的内部存储单元也包括其外部存储设备。本实施例中，所述存储器 21 通常用于存储安装于所述电子设备 2 的操作系统和各类应用软件，例如所述表格信息结构化提取系统 20 的程序代码等。此外，所述存储器 21 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。

所述处理器 22 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 22 通常用于控制所述电子设备 2 的总体操作，例如执行与所述电子设备 2 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 22 用于运行所述存储器 21 中存储的程序代码或者处理数据，例如运行所述的表格信息结构化提取系统 20 等。

所述网络接口 23 可包括无线网络接口或有线网络接口，该网络接口 23 通常用于在所述电子设备 2 与其他电子设备之间建立通信连接。例如，所述网络接口 23 用于通过网络将所述电子设备 2 与外部数据平台相连，在所述电子设备 2 与外部数据平台之间的建立数据传输通道和通信连接。所述网络可以是企业内部网（Intranet）、互联网（Internet）、全球移动通讯系统（Global System of Mobile communication，GSM）、宽带码分多址（Wideband Code Division Multiple Access，WCDMA）、4G 网络、5G 网络、蓝牙（Bluetooth）、

Wi-Fi 等无线或有线网络。

至此，已经详细介绍了本申请各个实施例的应用环境和相关设备的硬件结构和功能。下面，将基于上述应用环境和相关设备，提出本申请的各个实施例。

参阅图 2 所示，是本申请电子设备 2 中表格信息结构化提取系统 20 一实施例的程序模块图。本实施例中，所述的表格信息结构化提取系统 20 可以被分割成一个或多个程序模块，所述一个或者多个程序模块被存储于所述存储器 21 中，并由一个或多个处理器（本实施例中为所述处理器 22）所执行，以完成本申请。例如，在图 2 中，所述的表格信息结构化提取系统 20 可以被分割成获取模块 201、识别模块 202、以及存储模块 203。本申请所称的程序模块是指能够完成特定功能的一系列计算机程序指令段，比程序更适合于描述所述表格信息结构化提取系统 20 在所述电子设备 2 中的执行过程。以下将就各程序模块 201-203 的功能进行详细描述。

所述获取模块 201，用于获取指定文档（如 PDF 文档）中每行文字的位置信息和标签信息。在本实施例中，可以使用特定的文字识别工具（如 pdf2html 工具）获取该指定文档中每行文字的位置信息和标签信息。该特定的文字识别工具可以将 PDF 文档解析为文本文件（如 XML 文件），同时解析出每行文本在 PDF 文档中每行文字的位置信息和标签信息。

优选地，在本实施例中，每行文字的位置信息包括，但不限于，每行文字的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。其中，该指定文档中表格的每一行存储在相邻位置，即依据每行文字的位置信息（如左边沿坐标）依次存储。进一步地，每行文字的标签信息包括，但不限于，每行文字在所述指定文档（如 PDF 文档）的页码（即每行文字所在页面的序号）、页面长度、页面宽度等。

所述识别模块 202，用于根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形。

具体而言，所述从该指定文档的表格中识别出换行情形包括如下步骤 A1-A2。

（A1）定位该指定文档中的特定表格，获取该特定表格的位置信息，如该特定表格的左边沿坐标、表格宽度（表格高度）、和表格长度等。在本实施例中，可以通过该指定文档的特定规则，对该指定文档中的表格进行定位。例如，若该指定文档为 PDF 年报，则年报发布有明确的格式要求，可以根据类似下述的年报规则对特定表格进行判断：

如介绍主要客户和供应商时，表格标题会设为“主要销售客户和主要供应商情况”，因此这一标题后面就是客户供应商的特定表格。根据特定表格的标题关键词，则可以对介绍特定内容的表格进行定位，方便后续的解析。同理，PDF 年报中的其它表格都有类似的格式。

（A2）根据该特定表格的位置信息依次读取该特定表格中的一行文字，并根据该行文字的位置信息从该行文字中识别出行内换行的单元格。在本实施例中，可以从该特定表格的左边沿坐标开始，根据该特定表格的表格长度读取第一行，并根据该特定表格的表格宽度直至读取到该特定表格的最后一行。

优选地，在本实施例中，该特定表格的每行文字包括多个单元格，如第 1 单元格、第 2 单元格、第 3 单元格、第 4 单元格。更具体而言，所述换行情形包括行内换行和行尾换行。所述行内换行是指：在该特定表格的一行文字的内部单元格中存在换行情形。所述行尾换行是指：在该特定表格的一行文字的尾部单元格中存在换行情形。

优选地，在本实施例中，所述根据该行文字的位置信息从该行文字中识别出行内换行的单元格包括如下步骤 A21-A22。

（A21）获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的文字内容位置信息包括，但不限于，每个单元格的文字内容的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。

（A22）从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置和最后一次出现的单元格位置（即找到相同上边沿坐标第一次出现的单元格位置和最后一次出现的单元格位置），将第一次出现的单元格位置到最后一次出现的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

优选地，在其它实施例中，所述表格信息换行识别还包括步骤：（A3）根据当前行文字的位置信息从当前行文字中识别出行尾换行的单元格。

具体而言，所述根据当前行文字的位置信息从当前行文字中识别出行尾换行的单元格包括如下步骤 A31-A33。

（A31）若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息。其中，所述剩余单元格的文字内容位置信息包括，但不限于，所述剩余单元格的文字内容的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。

（A32）计算所述剩余单元格的文字内容上边沿坐标与当前行（如第一行）及下一行（如第二行）所有单元格的文字内容上边沿坐标的距离。

（A33）若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

进一步地，若距离最小值出现在下一行，则将所述剩余单元格的文字内容并入下一行，并判定所述剩余单元格为下一行的行尾换行单元格。

需要说明的是，在其它实施例中，所述根据当前行文字的位置信息从当前行文字中识别出行尾换行的单元格还可以包括如下步骤 A34-A36。

（A34）若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩

余单元格的文字内容位置信息。其中，所述剩余单元格的文字内容位置信息包括，但不限于，所述剩余单元格的文字内容的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。

（A35）计算所述剩余单元格的文字内容上边沿坐标与当前行（如第二行）及上一行（如第一行）所有单元格的文字内容上边沿坐标的距离。

（A36）若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

进一步地，若距离最小值出现在上一行，则将所述剩余单元格的文字内容并入上一行，并判定所述剩余单元格为上一行的行尾换行单元格。

需要说明的是，上述步骤 A1-A2、A21-A22、A31-A36 是以在 PDF 文件的特定表格（如客户供应商表格）中识别出表格信息换行情形为例进行说明，本领域技术人员可以理解，在其它实施例中，上述表格信息换行识别方法也可以针对 PDF 文件的所有表格进行换行情形识别，在此不再赘述。

进一步地，所述从该指定文档的表格中识别出跨页情形包括如下步骤 B1-B3（方法一）。

（B1）针对该指定文档中相邻的前一表格和下一表格，获取前一表格文字内容的位置信息、标签信息和下一表格文字内容的位置信息、标签信息。

优选地，在本实施例中，所述前一表格文字内容的位置信息包括，但不限于，前一表格每行文字的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息，及前一表格每列文字的左边沿坐标等。所述前一表格文字内容的标签信息包括，但不限于，前一表格每行文字在所述指定文档（如 PDF 文档）的页码（即每行文字所在页面的序号）、页面长度、页面宽度等。

进一步地，所述下一表格文字内容的位置信息包括，但不限于，下一表格每行文字的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息，及下一表格每列文字的左边沿坐标等。所述下一表格文字内容的标签信息包括，但不限于，下一表格每行文字在所述指定文档（如 PDF 文档）的页码（即每行文字所在页面的序号）、页面长度、页面宽度等。

（B2）比对下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标。举例而言，比对下一表格第 1 列文字的左边沿坐标与前一表格第 1 列文字的左边沿坐标，比对下一表格第 2 列文字的左边沿坐标与前一表格第 2 列文字的左边沿坐标，依此类推。

（B3）当下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标都相同时（代表下一表格与前一表格为同一表格），则比对下一表格每行文字的页码与前一表格每行文字的页码。举例而言，第一页的页尾包括前一表格，第二页的页首包括下一表格，其中，下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标都相同，则判定下一表格与前一表格为同一表格。

若下一表格每行文字的页码与前一表格每行文字的页码存在不同，则判定下一表格与前一表格为存在跨页情形的同一表格。若下一表格每行文字的页码与前一表格每行文字的页码都相同，则判定下一表格与前一表格为不存在跨页情形的同一表格，即下一表格与前一表格为位于同一页面的同一表格。

优选地，在本实施例中，若下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标之间的差值都小于预设阈值(如2个像素单位值)，则判定下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标都相同。

需要说明的是，上述步骤B1-B3（方法一）是以在PDF文件的相邻两个表格（前一表格和下一表格）中识别出表格信息跨页情形为例进行说明，本领域技术人员可以理解，在其它实施例中，上述表格信息跨页识别也可以针对PDF文件的特定表格（如财务表格）进行跨页情形识别（方法二），方法二包括如下步骤B4-B5。

（B4）定位该指定文档中的特定表格，获取该特定表格文字内容的位置信息和标签信息。其中，该特定表格文字内容的位置信息包括，但不限于，该特定表格每行文字的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。该特定表格文字内容的标签信息包括，但不限于，该特定表格每行文字在所述指定文档（如PDF文档）的页码（即每行文字所在页面的序号）、页面长度、页面宽度等。

具体而言，可以通过该指定文档的特定规则，对该指定文档中的特定表格进行定位。例如，若该指定文档为PDF年报，则年报发布有明确的格式要求，可以根据类似下述的年报规则对特定表格进行判断。

如介绍主要客户和供应商时，表格标题会设为“主要销售客户和主要供应商情况”，因此这一标题后面就是客户供应商的特定表格。根据特定表格的标题关键词，则可以对介绍特定内容的表格进行定位，方便后续的解析。同理，PDF年报中的其它特定表格都有类似的格式。

（B5）根据该特定表格文字内容的位置信息（如上边沿坐标）依次读取该特定表格的每行文字（如上边沿坐标相同的文字为同一行），并根据该特定表格文字内容的标签信息获取每行文字的页码。

若该特定表格每行文字的页码存在不同，则判定该特定表格存在跨页情形（即识别出该特定表格位于不同页面的前一表格和下一表格）。若该特定表格每行文字的页码都相同，则判定该特定表格不存在跨页情形。

所述存储模块203，用于当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储（分行提取表格数据并存储）和分列存储（分列提取表格数据并存储），形成结构化的表格数据。

优选地，在本实施例中，所述第一重塑规则包括：将上边沿坐标相同的文字存储为同一行（分行存储），将左边沿坐标相同的文字存储为同一列（分

列存储）。

所述存储模块 203，还用于当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储，形成结构化的表格数据。

优选地，在本实施例中，所述第二重塑规则包括：

删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉；

将删除页脚后的前一表格文字内容与删除页眉后的下一表格文字内容进行拼接形成拼接表格（即形成同一页的表格）；及

将所述拼接表格中上边沿坐标相同的文字存储为同一行（分行存储），并将所述拼接表格中左边沿坐标相同的文字存储为同一列（分列存储）。

具体而言，所述删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉包括：根据前一页面和下一页面的标签信息、及该指定文档的特定规则，对该指定文档中前一页面的页脚范围和下一页面的页眉范围进行定位，并根据定位确定的页脚范围和页眉范围删除前一页面的页脚和下一页面的页眉。

其中，所述前一页面的标签信息包括，但不限于，前一页面的页码、页面长度、页面宽度等；所述下一页面的标签信息包括，但不限于，下一页面的页码、页面长度、页面宽度等。该指定文档的特定规则包括，但不限于，前一页面的页脚所占页面长度的第一比例（如 8%）、下一页面的页眉所占页面长度的第二比例（如 9%）。可以理解，所述第一比例和第二比例也可以相同。

进一步地，所述前一页面的页脚范围的确定规则为：依据从下至上的顺序，选取所述前一页面的页面长度的第一比例内容，将选取的第一比例内容作为所述前一页面的页脚范围。所述下一页面的页眉范围的确定规则为：依据从上至下的顺序，选取所述下一页面的页面长度的第二比例内容，将选取的第二比例内容作为所述下一页面的页眉范围。

优选地，在其它实施例中，所述表格信息结构化提取系统 20 还用于：针对存储的结构化表格数据进行横向比对分析和纵向比对分析。

其中，所述横向比对分析包括：比对同一时间范围内（如同一年）同一行业不同公司的结构化表格数据（如应收账款数据），以便分析不同公司的债务情况、资金状况等经营信息。所述纵向比对分析包括：比对同一公司在不同时间范围内（如近三年）的结构化表格数据（如应收账款数据），以便分析该公司的债务情况、资金状况等经营信息（如应收账款变化）。

通过上述程序模块 201-203，本申请所提出的表格信息结构化提取系统 20，通过分析指定文档（如 PDF 文档）中表格文字内容的位置信息和标签信息，可以识别出表格（如 PDF 年报表格）中的换行情形和跨页情形，并对出

现换行情形和跨页情形的表格进行结构化数据提取并存储。该方法无需将PDF文件转化为word、excel等结构化文档，数据提取效率高，大规模数据集下召回率和准确率较高，且利于后续横向比对分析、纵向比对分析及数据建模。

此外，本申请还提出一种表格信息结构化提取方法。

参阅图3所示，是本申请表格信息结构化提取方法一实施例的实施流程示意图。在本实施例中，根据不同的需求，图3所示的流程图中的步骤的执行顺序可以改变，某些步骤可以省略。

步骤S31，获取指定文档（如PDF文档）中每行文字的位置信息和标签信息。在本实施例中，可以使用特定的文字识别工具（如pdf2html工具）获取该指定文档中每行文字的位置信息和标签信息。该特定的文字识别工具可以将PDF文档解析为文本文件（如XML文件），同时解析出每行文本在PDF文档中每行文字的位置信息和标签信息。

优选地，在本实施例中，每行文字的位置信息包括，但不限于，每行文字的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。其中，该指定文档中表格的每一行存储在相邻位置，即依据每行文字的位置信息（如左边沿坐标）依次存储。进一步地，每行文字的标签信息包括，但不限于，每行文字在所述指定文档(如PDF文档)的页码(即每行文字所在页面的序号)、页面长度、页面宽度等。

步骤S32，根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形。

具体而言，所述从该指定文档的表格中识别出换行情形包括如下步骤A1-A2。

（A1）定位该指定文档中的特定表格，获取该特定表格的位置信息，如该特定表格的左边沿坐标、表格宽度（表格高度）、和表格长度等。在本实施例中，可以通过该指定文档的特定规则，对该指定文档中的表格进行定位。例如，若该指定文档为PDF年报，则年报发布有明确的格式要求，可以根据类似下述的年报规则对特定表格进行判断：

如介绍主要客户和供应商时，表格标题会设为“主要销售客户和主要供应商情况”，因此这一标题后面就是客户供应商的特定表格。根据特定表格的标题关键词，则可以对介绍特定内容的表格进行定位，方便后续的解析。同理，PDF年报中的其它表格都有类似的格式。

（A2）根据该特定表格的位置信息依次读取该特定表格中的一行文字，并根据该行文字的位置信息从该行文字中识别出行内换行的单元格。在本实施例中，可以从该特定表格的左边沿坐标开始，根据该特定表格的表格长度读取第一行，并根据该特定表格的表格宽度直至读取到该特定表格的最后一行。

优选地，在本实施例中，该特定表格的每行文字包括多个单元格，如第1单元格、第2单元格、第3单元格、第4单元格。更具体而言，所述换行情形包括行内换行和行尾换行。所述行内换行是指：在该特定表格的一行文字的内部单元格中存在换行情形。所述行尾换行是指：在该特定表格的一行文字的尾部单元格中存在换行情形。

优选地，在本实施例中，所述根据该行文字的位置信息从该行文字中识别出行内换行的单元格包括如下步骤A21-A22。

（A21）获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的文字内容位置信息包括，但不限于，每个单元格的文字内容的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。

（A22）从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置和最后一次出现的单元格位置（即找到相同上边沿坐标第一次出现的单元格位置和最后一次出现的单元格位置），将第一次出现的单元格位置到最后一次出现的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

优选地，在其它实施例中，所述表格信息换行识别方法还包括步骤：（A3）根据当前行文字的位置信息从当前行文字中识别出行尾换行的单元格。

具体而言，所述根据当前行文字的位置信息从当前行文字中识别出行尾换行的单元格包括如下步骤A31-A33。

（A31）若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息。其中，所述剩余单元格的文字内容位置信息包括，但不限于，所述剩余单元格的文字内容的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。

（A32）计算所述剩余单元格的文字内容上边沿坐标与当前行（如第一行）及下一行（如第二行）所有单元格的文字内容上边沿坐标的距离。

（A33）若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

进一步地，若距离最小值出现在下一行，则将所述剩余单元格的文字内容并入下一行，并判定所述剩余单元格为下一行的行尾换行单元格。

需要说明的是，在其它实施例中，所述根据当前行文字的位置信息从当前行文字中识别出行尾换行的单元格还可以包括如下步骤A34-A36。

（A34）若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息。其中，所述剩余单元格的文字内容位置信息包括，但不限于，所述剩余单元格的文字内容的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。

（A35）计算所述剩余单元格的文字内容上边沿坐标与当前行（如第二行）及上一行（如第一行）所有单元格的文字内容上边沿坐标的距离。

（A36）若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

进一步地，若距离最小值出现在上一行，则将所述剩余单元格的文字内容并入上一行，并判定所述剩余单元格为上一行的行尾换行单元格。

需要说明的是，上述步骤A1-A2、A21-A22、A31-A36是以在PDF文件的特定表格（如客户供应商表格）中识别出表格信息换行情形为例进行说明，本领域技术人员可以理解，在其它实施例中，上述表格信息换行识别方法也可以针对PDF文件的所有表格进行换行情形识别，在此不再赘述。

进一步地，所述从该指定文档的表格中识别出跨页情形包括如下步骤B1-B3（方法一）。

（B1）针对该指定文档中相邻的前一表格和下一表格，获取前一表格文字内容的位置信息、标签信息和下一表格文字内容的位置信息、标签信息。

优选地，在本实施例中，所述前一表格文字内容的位置信息包括，但不限于，前一表格每行文字的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息，及前一表格每列文字的左边沿坐标等。所述前一表格文字内容的标签信息包括，但不限于，前一表格每行文字在所述指定文档（如PDF文档）的页码（即每行文字所在页面的序号）、页面长度、页面宽度等。

进一步地，所述下一表格文字内容的位置信息包括，但不限于，下一表格每行文字的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息，及下一表格每列文字的左边沿坐标等。所述下一表格文字内容的标签信息包括，但不限于，下一表格每行文字在所述指定文档（如PDF文档）的页码（即每行文字所在页面的序号）、页面长度、页面宽度等。

（B2）比对下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标。举例而言，比对下一表格第1列文字的左边沿坐标与前一表格第1列文字的左边沿坐标，比对下一表格第2列文字的左边沿坐标与前一表格第2列文字的左边沿坐标，依此类推。

（B3）当下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标都相同时（代表下一表格与前一表格为同一表格），则比对下一表格每行文字的页码与前一表格每行文字的页码。举例而言，第一页的页尾包括前一表格，第二页的页首包括下一表格，其中，下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标都相同，则判定下一表格与前一表格为同一表格。

若下一表格每行文字的页码与前一表格每行文字的页码存在不同，则判定下一表格与前一表格为存在跨页情形的同一表格。若下一表格每行文字的页码与前一表格每行文字的页码都相同，则判定下一表格与前一表格为不存在跨页情形的同一表格，即下一表格与前一表格为位于同一页面的同一表格。

优选地，在本实施例中，若下一表格每列文字的左边沿坐标与前一表格

对应每列文字的左边沿坐标之间的差值都小于预设阈值（如2个像素单位值），则判定下一表格每列文字的左边沿坐标与前一表格对应每列文字的左边沿坐标都相同。

需要说明的是，上述步骤B1-B3（方法一）是以在PDF文件的相邻两个表格（前一表格和下一表格）中识别出表格信息跨页情形为例进行说明，本领域技术人员可以理解，在其它实施例中，上述表格信息跨页识别方法也可以针对PDF文件的特定表格（如财务表格）进行跨页情形识别（方法二），方法二包括如下步骤B4-B5。

（B4）定位该指定文档中的特定表格，获取该特定表格文字内容的位置信息和标签信息。其中，该特定表格文字内容的位置信息包括，但不限于，该特定表格每行文字的左边沿坐标、上边沿坐标、文本宽度、文本长度等坐标信息。该特定表格文字内容的标签信息包括，但不限于，该特定表格每行文字在所述指定文档（如PDF文档）的页码（即每行文字所在页面的序号）、页面长度、页面宽度等。

具体而言，可以通过该指定文档的特定规则，对该指定文档中的特定表格进行定位。例如，若该指定文档为PDF年报，则年报发布有明确的格式要求，可以根据类似下述的年报规则对特定表格进行判断。

如介绍主要客户和供应商时，表格标题会设为“主要销售客户和主要供应商情况”，因此这一标题后面就是客户供应商的特定表格。根据特定表格的标题关键词，则可以对介绍特定内容的表格进行定位，方便后续的解析。同理，PDF年报中的其它特定表格都有类似的格式。

（B5）根据该特定表格文字内容的位置信息（如上边沿坐标）依次读取该特定表格的每行文字（如上边沿坐标相同的文字为同一行），并根据该特定表格文字内容的标签信息获取每行文字的页码。

若该特定表格每行文字的页码存在不同，则判定该特定表格存在跨页情形（即识别出该特定表格位于不同页面的前一表格和下一表格）。若该特定表格每行文字的页码都相同，则判定该特定表格不存在跨页情形。

步骤S33，当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储（分行提取表格数据并存储）和分列存储（分列提取表格数据并存储），形成结构化的表格数据。

优选地，在本实施例中，所述第一重塑规则包括：将上边沿坐标相同的文字存储为同一行（分行存储），将左边沿坐标相同的文字存储为同一列（分列存储）。

步骤S34，当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储，形成结构化的表格数据。

优选地，在本实施例中，所述第二重塑规则包括：

删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉；

将删除页脚后的前一表格文字内容与删除页眉后的下一表格文字内容进行拼接形成拼接表格（即形成同一页的表格）；及

将所述拼接表格中上边沿坐标相同的文字存储为同一行（分行存储），并将所述拼接表格中左边沿坐标相同的文字存储为同一列（分列存储）。

具体而言，所述删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉包括：根据前一页面和下一页面的标签信息、及该指定文档的特定规则，对该指定文档中前一页面的页脚范围和下一页面的页眉范围进行定位，并根据定位确定的页脚范围和页眉范围删除前一页面的页脚和下一页面的页眉。

其中，所述前一页面的标签信息包括，但不限于，前一页面的页码、页面长度、页面宽度等；所述下一页面的标签信息包括，但不限于，下一页面的页码、页面长度、页面宽度等。该指定文档的特定规则包括，但不限于，前一页面的页脚所占页面长度的第一比例（如8%）、下一页面的页眉所占页面长度的第二比例（如9%）。可以理解，所述第一比例和第二比例也可以相同。

进一步地，所述前一页面的页脚范围的确定规则为：依据从下至上的顺序，选取所述前一页面的页面长度的第一比例内容，将选取的第一比例内容作为所述前一页面的页脚范围。所述下一页面的页眉范围的确定规则为：依据从上至下的顺序，选取所述下一页面的页面长度的第二比例内容，将选取的第二比例内容作为所述下一页面的页眉范围。

优选地，在其它实施例中，所述表格信息结构化提取方法还包括步骤：针对存储的结构化表格数据进行横向比对分析和纵向比对分析。

其中，所述横向比对分析包括：比对同一时间范围内（如同一年）同一行业不同公司的结构化表格数据（如应收账款数据），以便分析不同公司的债务情况、资金状况等经营信息。所述纵向比对分析包括：比对同一公司在不同时间范围内（如近三年）的结构化表格数据（如应收账款数据），以便分析该公司的债务情况、资金状况等经营信息（如应收账款变化）。

通过上述步骤 S31-S34 及其它相关步骤，本申请所提出的表格信息结构化提取方法，通过分析指定文档（如 PDF 文档）中表格文字内容的位置信息和标签信息，可以识别出表格（如 PDF 年报表格）中的换行情形和跨页情形，并对出现换行情形和跨页情形的表格进行结构化数据提取并存储。该方法无需将 PDF 文件转化为 word、excel 等结构化文档，数据提取效率高，大规模数据集下召回率和准确率较高，且利于后续横向比对分析、纵向比对分析及数据建模。

进一步地，为实现上述目的，本申请还提供一种计算机可读存储介质（如 ROM/RAM、磁碟、光盘），所述计算机可读存储介质存储有表格信息结构化

提取系统20，所述表格信息结构化提取系统20可被至少一个处理器22执行，以使所述至少一个处理器22执行如上所述的表格信息结构化提取方法的步骤。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件来实现，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，空调器，或者网络设备等）执行本申请各个实施例所述的方法。

以上参照附图说明了本申请的优选实施例，并非因此局限本申请的权利范围。上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。另外，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

本领域技术人员不脱离本申请的范围和实质，可以有多种变型方案实现本申请，比如作为一个实施例的特征可用于另一实施例而得到又一实施例。凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1. 一种电子设备，其特征在于，所述电子设备包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的表格信息结构化提取系统，所述表格信息结构化提取系统被所述处理器执行时实现如下步骤：

获取指定文档中每行文字的位置信息和标签信息；

根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形；

当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储和分列存储；及

当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储。

2. 如权利要求1所述的电子设备，其特征在于，所述第一重塑规则包括：将上边沿坐标相同的文字存储为同一行，将左边沿坐标相同的文字存储为同一列；

所述第二重塑规则包括：

删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉；

将删除页脚后的前一表格文字内容与删除页眉后的下一表格文字内容进行拼接形成拼接表格；及

将所述拼接表格中上边沿坐标相同的文字存储为同一行，并将所述拼接表格中左边沿坐标相同的文字存储为同一列。

3. 如权利要求 2 所述的电子设备，其特征在于，所述删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉包括：

根据前一页面和下一页面的标签信息、及该指定文档的特定规则，对该指定文档中前一页面的页脚范围和下一页面的页眉范围进行定位，并根据定位确定的页脚范围和页眉范围删除前一页面的页脚和下一页面的页眉；

其中，前一页面的页脚范围的确定规则为：依据从下至上的顺序，选取所述前一页面的页面长度的第一比例内容，将选取的第一比例内容作为所述前一页面的页脚范围；及

下一页面的页眉范围的确定规则为：依据从上至下的顺序，选取所述下一页面的页面长度的第二比例内容，将选取的第二比例内容作为所述下一页面的页眉范围。

4. 如权利要求 2 所述的电子设备，其特征在于，所述换行情形包括行内换行和行尾换行；

其中，所述行内换行的识别包括：

获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的

文字内容位置信息包括每个单元格的文字内容的上边沿坐标；及

从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置和最后一次出现的单元格位置，将第一次出现的单元格位置到最后一次出现的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

5. 如权利要求 3 所述的电子设备，其特征在于，所述换行情形包括行内换行和行尾换行；

其中，所述行内换行的识别包括：

获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的文字内容位置信息包括每个单元格的文字内容的上边沿坐标；及

从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置和最后一次出现的单元格位置，将第一次出现的单元格位置到最后一次出现的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

6. 如权利要求 4 所述的电子设备，其特征在于，所述行尾换行的识别包括：

若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息，其中，所述剩余单元格的文字内容位置信息包括所述剩余单元格的文字内容的上边沿坐标；

计算所述剩余单元格的文字内容上边沿坐标与当前行及下一行所有单元格的文字内容上边沿坐标的距离，或计算所述剩余单元格的文字内容上边沿坐标与当前行及上一行所有单元格的文字内容上边沿坐标的距离；及

若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

7. 如权利要求 5 所述的电子设备，其特征在于，所述行尾换行的识别包括：

若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息，其中，所述剩余单元格的文字内容位置信息包括所述剩余单元格的文字内容的上边沿坐标；

计算所述剩余单元格的文字内容上边沿坐标与当前行及下一行所有单元格的文字内容上边沿坐标的距离，或计算所述剩余单元格的文字内容上边沿坐标与当前行及上一行所有单元格的文字内容上边沿坐标的距离；及

若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

8. 一种表格信息结构化提取方法，应用于电子设备，其特征在于，所述方法包括：

获取指定文档中每行文字的位置信息和标签信息；

根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形；

当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储和分列存储；及

当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储。

9. 如权利要求 8 所述的表格信息结构化提取方法，其特征在于，所述第一重塑规则包括：将上边沿坐标相同的文字存储为同一行，将左边沿坐标相同的文字存储为同一列；

所述第二重塑规则包括：

删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉；

将删除页脚后的前一表格文字内容与删除页眉后的下一表格文字内容进行拼接形成拼接表格；及

将所述拼接表格中上边沿坐标相同的文字存储为同一行，并将所述拼接表格中左边沿坐标相同的文字存储为同一列。

10. 如权利要求 9 所述的表格信息结构化提取方法，其特征在于，所述删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉包括：

根据前一页面和下一页面的标签信息、及该指定文档的特定规则，对该指定文档中前一页面的页脚范围和下一页面的页眉范围进行定位，并根据定位确定的页脚范围和页眉范围删除前一页面的页脚和下一页面的页眉；

其中，前一页面的页脚范围的确定规则为：依据从下至上的顺序，选取所述前一页面的页面长度的第一比例内容，将选取的第一比例内容作为所述前一页面的页脚范围；及

下一页面的页眉范围的确定规则为：依据从上至下的顺序，选取所述下一页面的页面长度的第二比例内容，将选取的第二比例内容作为所述下一页面的页眉范围。

11. 如权利要求 9 所述的表格信息结构化提取方法，其特征在于，所述换行情形包括行内换行和行尾换行；

其中，所述行内换行的识别包括：

获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的文字内容位置信息包括每个单元格的文字内容的上边沿坐标；及

从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置

和最后一次出现的单元格位置，将第一次出现的单元格位置到最后一次出现的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

12. 如权利要求 10 所述的表格信息结构化提取方法，其特征在于，所述换行情形包括行内换行和行尾换行；

其中，所述行内换行的识别包括：

获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的文字内容位置信息包括每个单元格的文字内容的上边沿坐标；及

从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置和最后一次出现的单元格位置，将第一次出现的单元格位置到最后一次出现的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

13. 如权利要求 11 所述的表格信息结构化提取方法，其特征在于，所述行尾换行的识别包括：

若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息，其中，所述剩余单元格的文字内容位置信息包括所述剩余单元格的文字内容的上边沿坐标；

计算所述剩余单元格的文字内容上边沿坐标与当前行及下一行所有单元格的文字内容上边沿坐标的距离，或计算所述剩余单元格的文字内容上边沿坐标与当前行及上一行所有单元格的文字内容上边沿坐标的距离；及

若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

14. 如权利要求 12 所述的表格信息结构化提取方法，其特征在于，所述行尾换行的识别包括：

若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息，其中，所述剩余单元格的文字内容位置信息包括所述剩余单元格的文字内容的上边沿坐标；

计算所述剩余单元格的文字内容上边沿坐标与当前行及下一行所有单元格的文字内容上边沿坐标的距离，或计算所述剩余单元格的文字内容上边沿坐标与当前行及上一行所有单元格的文字内容上边沿坐标的距离；及

若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

15. 一种计算机可读存储介质，所述计算机可读存储介质存储有表格信息

结构化提取系统，所述表格信息结构化提取系统可被至少一个处理器执行，所述表格信息结构化提取系统被所述处理器执行时实现如下步骤：

获取指定文档中每行文字的位置信息和标签信息；

根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形；

当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储和分列存储；及

当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储。

16. 如权利要求15所述的计算机可读存储介质，其特征在于，所述第一重塑规则包括：将上边沿坐标相同的文字存储为同一行，将左边沿坐标相同的文字存储为同一列；

所述第二重塑规则包括：

删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉；

将删除页脚后的前一表格文字内容与删除页眉后的下一表格文字内容进行拼接形成拼接表格；及

将所述拼接表格中上边沿坐标相同的文字存储为同一行，并将所述拼接表格中左边沿坐标相同的文字存储为同一列。

17. 如权利要求 16 所述的计算机可读存储介质，其特征在于，所述删除前一表格所在前一页面的页脚和下一表格所在下一页面的页眉包括：

根据前一页面和下一页面的标签信息、及该指定文档的特定规则，对该指定文档中前一页面的页脚范围和下一页面的页眉范围进行定位，并根据定位确定的页脚范围和页眉范围删除前一页面的页脚和下一页面的页眉；

其中，前一页面的页脚范围的确定规则为：依据从下至上的顺序，选取所述前一页面的页面长度的第一比例内容，将选取的第一比例内容作为所述前一页面的页脚范围；及

下一页面的页眉范围的确定规则为：依据从上至下的顺序，选取所述下一页面的页面长度的第二比例内容，将选取的第二比例内容作为所述下一页面的页眉范围。

18. 如权利要求 16 所述的计算机可读存储介质，其特征在于，所述换行情形包括行内换行和行尾换行；

其中，所述行内换行的识别包括：

获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的文字内容位置信息包括每个单元格的文字内容的上边沿坐标；及

从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置和最后一次出现的单元格位置，将第一次出现的单元格位置到最后一次出现

的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

19. 如权利要求 17 所述的计算机可读存储介质，其特征在于，所述换行情形包括行内换行和行尾换行；

其中，所述行内换行的识别包括：

获取该行文字中每个单元格的文字内容位置信息，其中，每个单元格的文字内容位置信息包括每个单元格的文字内容的上边沿坐标；及

从文字内容上边沿坐标相同的单元格中，获取第一次出现的单元格位置和最后一次出现的单元格位置，将第一次出现的单元格位置到最后一次出现的单元格位置的所有单元格确定为同一行，并判定第一次出现的单元格位置与最后一次出现的单元格位置之间文字内容上边沿坐标不同的单元格为行内换行的单元格。

20. 如权利要求 18 或 19 所述的计算机可读存储介质，其特征在于，所述行尾换行的识别包括：

若行内换行识别后当前行文字中存在剩余单元格，则获取所述剩余单元格的文字内容位置信息，其中，所述剩余单元格的文字内容位置信息包括所述剩余单元格的文字内容的上边沿坐标；

计算所述剩余单元格的文字内容上边沿坐标与当前行及下一行所有单元格的文字内容上边沿坐标的距离，或计算所述剩余单元格的文字内容上边沿坐标与当前行及上一行所有单元格的文字内容上边沿坐标的距离；及

若距离最小值出现在当前行，则将所述剩余单元格的文字内容并入当前行，并判定所述剩余单元格为当前行的行尾换行单元格。

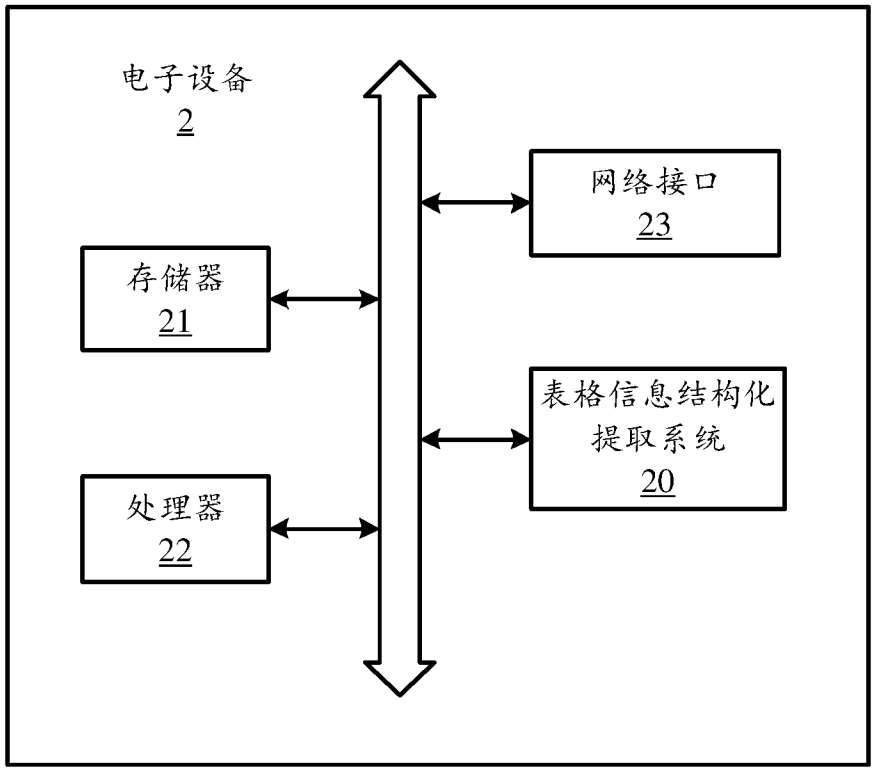

电子设备
2
网络接口
23
存储器
21
表格信息结构化
提取系统
20
处理器
22

图 1

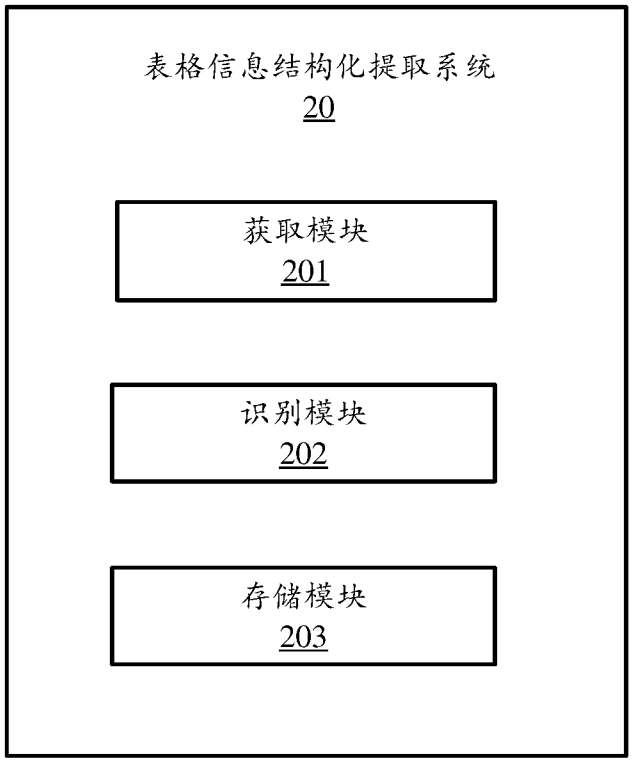

表格信息结构化提取系统
20
获取模块
201
识别模块
202
存储模块
203

图 2

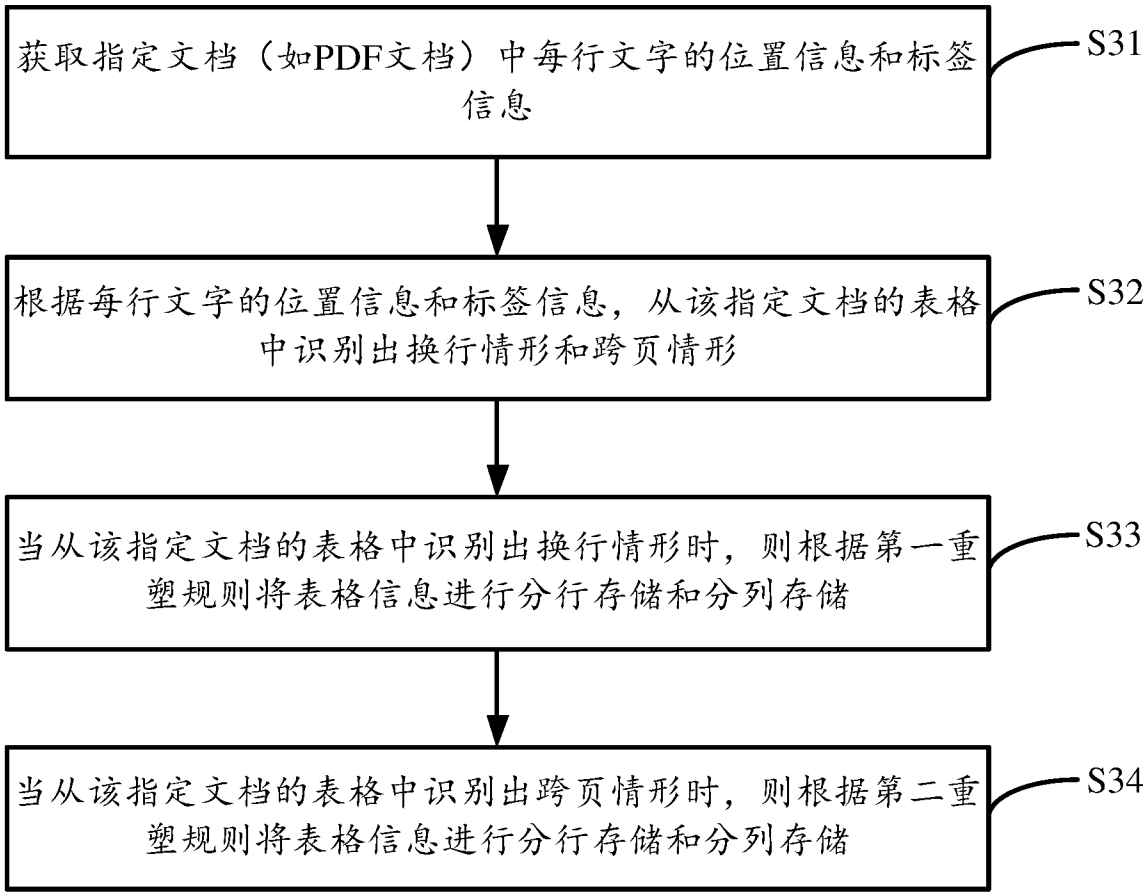

获取指定文档（如PDF文档）中每行文字的位置信息和标签信息
S31
根据每行文字的位置信息和标签信息，从该指定文档的表格中识别出换行情形和跨页情形
S32
当从该指定文档的表格中识别出换行情形时，则根据第一重塑规则将表格信息进行分行存储和分列存储
S33
当从该指定文档的表格中识别出跨页情形时，则根据第二重塑规则将表格信息进行分行存储和分列存储
S34

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/076167

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/24 (2006.01) 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)

CNPAT, WPI, EPODOC, CNKI, IEEE, GOOGLE: 表格, 文档, 识别, 位置, 行, 页, 页眉, 页脚, pdf, excel, table, form, recognize, row, column, header, footer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102722475 A (SHENZHEN WONDERSHARE SOFTWARE CO., LTD.) 10 October 2012 (10.10.2012), description, paragraphs [0044], [0055] and [0061]	1-20
Y	CN 102855232 A (TONGFANG DISC INC.) 02 January 2013 (02.01.2013), description, paragraphs [0006]-[0011] and [0018]-[0029]	1-20
A	CN 104268127 A (CNKI (BEIJING) TECHNOLOGY CO., LTD.) 07 January 2015 (07.01.2015), entire document	1-20
A	US 2009282009 A1 (TAGS LTD.) 12 November 2009 (12.11.2009), entire document	1-20
A	US 2012265759 A1 (XEROX CORPORATION) 18 October 2012 (18.10.2012), 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	Date of mailing of the international search report
08 May 2018	05 July 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer KONG, Xin Telephone No. (86-10) 53961371

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/076167

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102722475 A	10 October 2012	None	
CN 102855232 A	02 January 2013	None	
CN 104268127 A	07 January 2015	None	
US 2009282009 A1	12 November 2009	None	
US 2012265759 A1	18 October 2012	None	

国际检索报告

国际申请号 PCT/CN2018/076167

A. 主题的分类

G06F 17/24(2006.01)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）)

CNPAT,WPI,EPODOC,CNKI,IEEE,GOOGLE：表格，文档，识别，位置，行，页，页眉，页脚，pdf，excel，table，form，recognize，row，column，header，footer

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	CN 102722475 A (深圳市万兴软件有限公司) 2012年 10月 10日 (2012-10-10) 说明书第[0044]，[0055]，[0061]段	1-20
Y	CN 102855232 A (同方光盘股份有限公司) 2013年 1月 2日 (2013-01-02) 说明书第[0006]-[0011]，[0018]-[0029]段	1-20
A	CN 104268127 A (同方知网北京技术有限公司) 2015年 1月 7日 (2015-01-07) 全文	1-20
A	US 2009282009 A1 (TAGS LTD.) 2009年 11月 12日 (2009-11-12) 全文	1-20
A	US 2012265759 A1 (XEROX CORPORATION) 2012年 10月 18日 (2012-10-18) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 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
2018年 5月 8日	2018年 7月 5日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	孔昕
传真号 (86-10)62019451	电话号码 86-(10)-53961371

表 PCT/ISA/210 (第2页) (2015年1月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/076167

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	102722475	A	2012年 10月 10日	无	
CN	102855232	A	2013年 1月 2日	无	
CN	104268127	A	2015年 1月 7日	无	
US	2009282009	A1	2009年 11月 12日	无	
US	2012265759	A1	2012年 10月 18日	无	