(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111046096 B
(45)授权公告日 2023.11.24

(21)申请号 201911296540.8

(22)申请日 2019.12.16

(65)同一申请的已公布的文献号
申请公布号 CN 111046096 A

(43)申请公布日 2020.04.21

(73)专利权人 北京信息科技大学
地址 100085 北京市海淀区清河小营东路12号
专利权人 王长胜

(72)发明人 田英爱 王长胜 李宁 施运梅 李海波 陈亚军

(74)专利代理机构 北京唯智勤实知识产权代理事务所(普通合伙) 11557
专利代理师 陈佳

(51)Int.Cl.
G06F 16/25(2019.01)
G06F 40/189(2020.01)

(56)对比文件
CN 102262618 A,2011.11.30
CN 109657221 A,2019.04.19
CN 101308488 A,2008.11.19
CN 104111922 A,2014.10.22
US 2016247020 A1,2016.08.25

审查员 王琦瑶

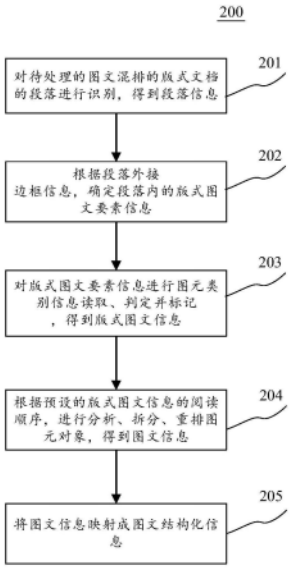
权利要求书2页 说明书8页 附图5页

(54)发明名称

用于生成图文结构化信息的方法和装置

(57)摘要

本公开的实施例公开了用于生成图文结构化信息的方法和装置。该方法的一具体实施方式包括:对待处理的图文混排的版式文档的段落进行识别,得到段落信息,其中,上述段落信息包括用于表征上述段落所在范围的段落外接边框信息;根据上述段落外接边框信息,确定上述段落内的版式图文要素信息;对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息;根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息;将上述图文信息映射成图文结构化信息。该实施方式实现了图文结构化信息准确生成,进而增强了文档在不同设备上显示的自适应性。



CN 111046096 B

1.一种用于生成图文结构化信息的方法,包括:

对待处理的图文混排的版式文档的段落进行识别,得到段落信息,其中,所述段落信息包括用于表征所述段落所在范围的段落外接边框信息,对每一段落施加一外接边框;

根据所述段落外接边框信息,确定所述段落内的版式图文要素信息;

对所述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息;

根据预设的所述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息;

将所述图文信息映射成图文结构化信息;

其中,所述根据所述段落外接边框信息,确定所述段落内的版式图文要素信息,包括:

根据所述段落外接边框信息,识别段落外接边框范围内的图元信息和图元所在版式页面的版式页面块信息,形成图元信息列表和版式页面块信息列表,其中,所述图元信息至少包括图元类别信息、图元标识信息以及是否跨页信息,所述版式页面块信息至少包括版式页面排版边框的标识信息,并与所述图元信息列表相关联;

其中,所述根据预设的所述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息,包括:

基于预设阅读方向,对所述段落内的行进行行高分析;

确定所述行中的文本图元对象对应的纵坐标值偏差是否大于预设偏差值;

响应于确定大于所述预设偏差值,依据行高将所述段落拆分为多个独立文本图元数据;

依据各个文本图元数据的行内横坐标确定各个文本图元的阅读顺序;

若所述各个文本图元数据内行的横坐标范围内图元对象与所述文本图元对象不同,则进一步拆分所述文本图元数据的文本图元,以得到图元信息顺序列表;

重新调整所述图元信息顺序列表中图元所在的版式页面块的阅读顺序。

2.根据权利要求1所述的方法,其中,所述对所述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息,包括:

分析所述段落内的版式页面块中的图元信息列表中对应的图元类别;

响应于所述版式页面块中同时存在文本图元和其它类型图元,则标记分析结果为0;

响应于所述版式页面块中只存在某一种相同类型图元,则标记分析结果为1;

响应于所述版式页面块中存在其它情形,则标记分析结果为2。

3.根据权利要求2所述的方法,其中,所述将所述图文信息映射成图文结构化信息,包括:

初始化结构化版式文档的根节点,以及生成对应的结构化图文段落节点;

在所述根节点下增加所述结构化图文段落节点,以及在所述结构化图文段落节点下增加段落片段节点,以及将待处理的开始页码映射到所述段落片段节点;

依据所排序完成的图元信息顺序列表,依次将所述图元信息顺序列表中的图元映射到对应的段落片段节点,至此图文结构化信息映射完毕。

4.根据权利要求1-3之一所述的方法,其中,所述依据所排序完成的图元信息顺序列表,依次将所述图元信息顺序列表中的图元映射到对应的段落片段节点,包括:

响应于第一个图元是跨页的,则新建段落片段节点,以及将所跨页的跨页码映射到所

述新建段落片段节点；

依次映射所述图元信息顺序列表中的图元到对应的结构化段落片段节点下的块节点，以及增加对应图元类别；

若图元类别为非文本类别的，则增加所述图元类别所对应的图元的文字绕排属性，关联所述图元类别所对应的图元对象。

5.一种用于生成图文结构化信息的装置，包括：

识别单元，被配置成对待处理的图文混排的版式文档的段落进行识别，得到段落信息，其中，所述段落信息包括用于表征所述段落所在范围的段落外接边框信息，对每一段落施加一外接边框；

确定单元，被配置成根据所述段落外接边框信息，确定所述段落内的版式图文要素信息；

读取判定单元，被配置成对所述版式图文要素信息进行图元类别信息读取、判定并标记，得到版式图文信息；

分析拆分重排单元，被配置成根据预设的所述版式图文信息的阅读顺序，进行分析、拆分、重排图元对象，得到图文信息；

映射单元，被配置成将所述图文信息映射成图文结构化信息；

其中，所述确定单元，进一步被配置成：根据所述段落外接边框信息，识别段落外接边框范围内的图元信息和图元所在版式页面的版式页面块信息，形成图元信息列表和版式页面块信息列表，其中，所述图元信息至少包括图元类别信息、图元标识信息以及是否跨页信息，所述版式页面块信息至少包括版式页面排版边框的标识信息，并与所述图元信息列表相关联；

其中，所述分析拆分重排单元，进一步被配置成：基于预设阅读方向，对所述段落内的行进行行高分析；确定所述行中的文本图元对象对应的纵坐标值偏差是否大于预设偏差值；响应于确定大于所述预设偏差值，依据行高将所述段落拆分为多个独立文本图元数据；依据各个文本图元数据的行内横坐标确定各个文本图元的阅读顺序；若所述各个文本图元数据内行的横坐标范围内图元对象与所述文本图元对象不同，则进一步拆分所述文本图元数据的文本图元，以得到图元信息顺序列表；重新调整所述图元信息顺序列表中图元所在的版式页面块的阅读顺序。

6.根据权利要求5所述的装置，其中，所述确定单元，包括：

识别子单元，被配置成根据所述段落外接边框信息，识别段落外接边框范围内的图元信息和图元所在版式页面的版式页面块信息，形成图元信息列表和版式页面块信息列表，其中，所述图元信息至少包括图元类别信息、图元标识信息以及是否跨页信息，所述版式页面块信息至少包括版式页面排版边框的标识信息，并与所述图元信息列表相关联。

用于生成图文结构化信息的方法和装置

技术领域

[0001] 本公开的实施例涉及计算机技术领域,具体涉及用于生成图文结构化信息的方法和装置。

背景技术

[0002] 数字出版物,常见为电子书,可以在电脑、手机、大型号立柜式的触摸屏、电纸书等数字阅读设备上呈现。当前数字出版资源加工的输入多是专业排版软件的中间产物,如PDF等纯版式文档。

[0003] 但是由于不包含流式的图文结构化信息,或者经过自动化的智能版面识别后得到的流式的图文结构化信息质量较差,无法准确的生成图文结构化信息。因而在不同尺寸屏幕的设备之间的自适应性阅读效果差,无法完全满足“一次出版,多平台应用,多途径传播”的目标。

[0004] 而从版式文档提取正确的流式信息则不尽人意,原因很多,主要体现在版式文档版面的复杂性。图4-14给出了各种类型的图文混排版式文档段落以及对应的图文结构化信息所呈现出来的效果,版面的不规整将影响阅读体验。

发明内容

[0005] 本公开的内容部分用于以简要的形式介绍技术方案,这些技术方案将在后面的具体实施方式部分被详细描述。本公开的内容部分并不旨在表示要求保护的技术方案的关键特征或必要特征,也不旨在用于限制所要求的保护的技术方案的范围。

[0006] 本公开的一些实施例提出了用于生成图文结构化信息的方法和装置,来解决以上背景技术部分提到的技术问题。

[0007] 第一方面,本公开的一些实施例提供了一种用于生成图文结构化信息的方法,该方法包括:对待处理的图文混排的版式文档的段落进行识别,得到段落信息,其中,上述段落信息包括用于表征上述段落所在范围的段落外接边框信息;根据上述段落外接边框信息,确定上述段落内的版式图文要素信息;对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息;根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息;将上述图文信息映射成图文结构化信息。

[0008] 在一些实施例中,上述根据上述段落外接边框信息,确定上述段落内的版式图文要素信息,包括:根据上述段落外接边框信息,识别段落外接边框范围内的图元信息和图元所在版式页面的版式页面块信息,形成图元信息列表和版式页面块信息列表,其中,上述图元信息至少包括图元类别信息、图元标识信息以及是否跨页信息,上述版式页面块信息至少包括版式页面排版边框的标识信息,并与上述图元信息列表相关联。

[0009] 在一些实施例中,上述对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息,包括:分析上述段落内的版式页面块中的图元信息列表中对应的图元类别;响应于上述版式页面块中同时存在文本图元和其它类型图元,则标记分析结果

为0;响应于上述版式页面块中只存在某一种相同类型图元,则标记分析结果为1;响应于上述版式页面块中存在其它情形,则标记分析结果为2。

[0010] 在一些实施例中,上述根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息,包括:基于预设阅读方向,对上述段落内的行进行行高分析;确定上述行中的文本图元对象对应的纵坐标值偏差是否大于预设偏差值;响应于确定大于上述预设偏差值,依据行高将上述段落拆分为多个独立文本图元数据,依据各个文本图元数据的行内横坐标确定各个文本图元的阅读顺序,以及若上述各个文本图元数据内行的横坐标范围内图元对象与上述文本图元对象不同,则进一步拆分上述文本图元数据的文本图元,以得到图元信息顺序列表;重新调整上述图元信息顺序列表中图元所在的版式页面块的阅读顺序。

[0011] 在一些实施例中,上述将上述图文信息映射成图文结构化信息,包括:初始化结构化版式文档的根节点,以及生成对应的结构化图文段落节点;在上述根节点下增加上述结构化图文段落节点,以及在上述结构化图文段落节点下增加段落片段节点,以及将待处理的开始页码映射到上述段落片段节点;依据所排序完成的图元信息顺序列表,依次将上述图元信息顺序列表中的图元映射到对应的段落片段节点,至此图文结构化信息映射完毕。

[0012] 在一些实施例中,上述依据所排序完成的图元信息顺序列表,依次将上述图元信息顺序列表中的图元映射到对应的段落片段节点,包括:响应于第一个图元是跨页的,则新建段落片段节点,以及将所跨页的跨页码映射到上述新建段落片段节点;依次映射上述图元信息顺序列表中的图元到对应的结构化段落片段节点下的块节点,以及增加对应图元类别;若图元类别为非文本类别的,则增加上述图元类别所对应的图元的文字绕排属性,关联上述图元类别所对应的图元对象。

[0013] 第二方面,本公开的一些实施例提供了一种用于生成图文结构化信息的装置,装置包括:识别单元,被配置成对待处理的图文混排的版式文档的段落进行识别,得到段落信息,其中,上述段落信息包括用于表征上述段落所在范围的段落外接边框信息;确定单元,被配置成根据上述段落外接边框信息,确定上述段落内的版式图文要素信息;读取判定单元,被配置成对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息;分析拆分重排单元,被配置成根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息;映射单元,被配置成将上述图文信息映射成图文结构化信息。

[0014] 在一些实施例中,上述确定单元,包括:识别子单元,被配置成根据上述段落外接边框信息,识别段落外接边框范围内的图元信息和图元所在版式页面的版式页面块信息,形成图元信息列表和版式页面块信息列表,其中,上述图元信息至少包括图元类别信息、图元标识信息以及是否跨页信息,上述版式页面块信息至少包括版式页面排版边框的标识信息,并与上述图元信息列表相关联。

[0015] 本公开的上述各个实施例中的一个实施例具有如下有益效果:通过对待处理的图文混排的版式文档的段落进行识别,可以得到段落信息,其中,上述段落信息包括用于表征上述段落所在范围的段落外接边框信息。之后,根据上述段落外接边框信息,可以确定上述段落内的版式图文要素信息。然后,对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息。随后,根据预设的上述版式图文信息的阅读顺序,进行分析、拆

分、重排图元对象,得到图文信息。最后,将上述图文信息映射成图文结构化信息。由于结构化图文信息描述了图文信息中的结构层次与阅读的顺序,进而,通过生成图文结构化信息,可以使图文内容进行重排。从而,可以提高文档在不同设备上显示的自适应性。通过对版式图文要素信息的阅读顺序进行分析,可以提高图文信息的准确率。进而,可以增强图文结构化信息所呈现出来的显示效果,提高用户阅读体验。

附图说明

[0016] 结合附图并参考以下具体实施方式,本公开各实施例所做的详细描述,本公开的其他特征、优点及目的将变得更加明显。贯穿附图中,相同或相似的附图标记表示相同或相似的元素。应当理解附图是示意性的,原件和元素不一定按照比例绘制。

[0017] 图1是本公开的一些实施例可以应用于其中的示例性系统的架构图;

[0018] 图2是根据本公开的用于生成图文结构化信息的方法的一些实施例的流程图;

[0019] 图3是根据本公开的用于生成图文结构化信息的装置的一些实施例的结构示意图;

[0020] 图4-14是示例性的图文混排版式文档的排版方式以及对应的段落结构化信息所呈现的效果。

具体实施方式

[0021] 下面将参照附图更详细地描述本公开的实施例。虽然附图中显示了本公开的某些实施例,然而应当理解的是,本公开可以通过多种形式来实现,而且不应该被解释为限于这里阐述的实施例。相反,提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是,本公开的附图及实施例仅用于示例性作用,并非用于限制本公开的保护范围。

[0022] 另外还需要说明的是,为了便于描述,附图中仅示出了与有关发明相关的部分。在不冲突的情况下,本公开中的实施例及实施例中的特征可以相互组合。

[0023] 需要注意,本公开中提及的“第一”、“第二”等概念仅用于对不同的装置、模块或单元进行区分,并非用于限定这些装置、模块或单元所执行的功能的顺序或者相互依存关系。

[0024] 需要注意,本公开中提及的“一个”、“多个”的修饰是示意性而非限制性的,本领域技术人员应当理解,除非在上下文另有明确指出,否则应该理解为“一个或多个”。

[0025] 本公开实施方式中的多个装置之间所交互的消息或者信息的名称仅用于说明性的目的,而并不是用于对这些消息或信息的范围进行限制。

[0026] 下面将参考附图并结合实施例来详细说明本公开。

[0027] 图1示出了可以应用本公开的一些实施例的用于生成图文结构化信息的方法或用于生成图文结构化信息的装置的示例性系统架构100。

[0028] 如图1所示,系统架构100可以包括终端设备101、102、103,网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线、无线通信链路或者光纤电缆等等。

[0029] 用户可以使用终端设备101、102、103通过网络104与服务器105交互,以接收或发送消息等。终端设备101、102、103上可以安装有各种通讯客户端应用,例如文档类应用。

[0030] 需要说明的是,本公开的实施例所提供的用于生成图文结构化信息的方法可以由

终端设备101、102、103执行,也可以由服务器105执行。相应地,用于生成图文结构化信息的装置可以设置于终端设备101、102、103中,也可以设置于服务器105中。在此不做具体限定。

[0031] 继续参考图2,示出了根据本公开的用于生成图文结构化信息的方法的一些实施例的流程200。该用于生成图文结构化信息的方法,包括以下步骤:

[0032] 步骤201,对待处理的图文混排的版式文档的段落进行识别,得到段落信息。

[0033] 在一些实施例中,用于生成图文结构化信息的方法的执行主体可以通过版面分析算法对待处理的图文混排的版式文档的段落进行识别,得到段落信息。其中,上述待处理的图文混排的版式文档可以是存储在本地的文档,上述待处理的图文混排的版式文档可以由技术人员指定,也可以根据一定的条件筛选。上述待处理的图文混排的版式文档可以包括段落。实践中,版式文档可以是一种独立于软件、硬件、操作系统等显示设备或打印设备的文档。作为示例,可以是PDF、CEBX、OFD等格式的文档。上述版面分析算法常常是指对版面进行分析的算法。上述版面分析算法可以包括但不限于:版面分割与区域识别算法。上述版面分割与区域识别算法常常是指识别得到版式文档的段落以及段落内版式图文要素的算法。

[0034] 其中,上述版式文档可以包括但不限于:全文的书写/阅读顺序,文档度量单位,文档总页数,当前待处理页码(通常第一页开始循环处理),页面大小等等。以文档的页面为单位,以上述版面分析算法得到的段落为待处理图文混排版式文档的段落,并对每一段落施加一外接边框。其中,给定全文书写/阅读顺序readDirection,如l2r—表示从左到右,至上而下书写/阅读顺序;文档度量单位docUnit,如mm毫米;文档总页数pageCount;当前待处理页码pageNumber;页面大小pageSize;段落外接边框paraBox。即<readDirection,docUnit,pageCount,pageNumber,pageSize,paraBox>。

[0035] 步骤202,根据上述段落外接边框信息,确定上述段落内的版式图文要素信息。

[0036] 在一些实施例中,基于步骤201中得到的段落外接边框信息,上述执行主体可以通过上述版面分析算法识别确定上述段落内的版式图文要素信息。其中,上述版式图文要素信息可以包括但不限于版式文本图元信息,版式图像图元信息,版式图形图元信息。

[0037] 在一些实施例的一些可选的实现方式中,根据上述段落外接边框信息,上述执行主体通过上述版面分析算法可以识别段落外接边框范围内的图元信息和图元所在版式页面的版式页面块信息,形成图元信息列表和版式页面块信息列表,其中,上述图元信息至少包括图元类别信息、图元标识信息以及是否跨页信息,上述版式页面块信息至少包括版式页面排版边框的标识信息,并与上述图元信息列表相关联。

[0038] 其中,上述图元信息可以包括但不限于图元类别信息type,图元标识信息id,外接矩形边框box,是否跨页bCrossPage以及跨页时页码pageNumber,即图元pageObject<type,id,box,bCrossPage,pageNumber,fontSize,charSpace,wordSpace,x,y,strText>。若图元类别为文本图元时,上述图元信息还可以包括记录字体大小fontSize和字符/文本间距值charSpace/wordSpace以及其文本内容信息,如起始绘制点x,y,文本字符串strText。上述版式页面块信息可以包括但不限于版式页面排版边框的标识信息pageBlockId,图元信息列表pageObjList(此图元信息列表pageObjectList由若干图元信息pageObject构成),图元区域信息pageObjectRefId(此图元区域信息pageObjectRefId是上述图元id的引用),CTM转换矩阵,裁剪区ClipArea。即版式页面块pageBlock<pageBlockId,pageObjList,pageObjectRefId,CTM,ClipArea>。

[0039] 步骤203,对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息。

[0040] 在一些实施例中,上述执行主体首先可以对上述版式图文要素信息进行图元类别信息的读取,得到图元类别信息。之后,可以对上述图元类别信息进行判定以及进行标记,得到版式图文信息。

[0041] 作为示例,上述执行主体可以执行以下步骤得到版式图文信息:首先,可以通过分析上述段落内的版式页面块pageBlock中的pageObjList图元信息列表中对应的图元类别type,若上述版式页面块pageBlock中同时存在文本图元和其它类型图元,则标记分析结果nFlag为0;若上述版式页面块pageBlock中全部为某一种相同类型图元(例如文本、图像、图形),则标记分析结果nFlag为1;若上述版式页面块pageBlock中存在其它情形,则标记分析结果nFlag为2。其中,上述其他情形可以是指除去图元信息列表pageObjList中对应的图元类别全部为某一种类别和同时存在文本图元和其它类型图元的两种情形之外的情形。

[0042] 其中,上述nFlag为0情形时,若pageBlock版式页面块数量为1,其对应的pageObject图元数量也为1并且为复合对象,则拆分该复合对象为单一类型的图元对象。上述nFlag为0情形时,分析计算pageObject图元中非文本图元对象的文字绕排类型:首先依据非文本图元对象的外接矩形边框box以及上述初始化上下文中的段落外接边框paraBox,计算其四周文本布局情况,若该box高度范围内存在大于1行的文本图元(至少两行图元的Y坐标不同,而且差值不小于上一行文本图元高度),则标记文字绕排类型wrap为四周绕排布局around;其它情形标记文字绕排类型wrap为随文布局follow。

[0043] 步骤204,根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息。

[0044] 在一些实施例中,上述执行主体可以根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息。其中,上述图文信息也可以包括文本图元信息、图像图元信息和图形图元信息。

[0045] 作为示例,上述执行主体可以执行以下步骤得到图文信息:第一,基于预设阅读方向,可以对上述段落内的行进行识别,进而可以进行行高分析;第二,可以确定上述行中的文本图元对象对应的纵坐标值偏差是否大于预设偏差值;第三,响应于确定大于预设偏差值,依据行高将上述段落拆分为多个独立文本图元数据;第四,可以依据各个文本图元数据的行内横坐标确定各个文本图元的阅读顺序;第五,若上述各个文本图元数据内行的横坐标范围内图元对象与上述文本图元对象不同,则进一步拆分上述文本图元数据的文本图元,以得到图元信息顺序列表;第六,重新调整上述图元信息顺序列表中图元所在的版式页面块的阅读顺序,得到图文信息。通过重新调整上述图元信息顺序列表中图元所在的版式页面块的阅读顺序可以确保版面数据正确的呈现。

[0046] 步骤205,将上述图文信息映射成图文结构化信息。

[0047] 在一些实施例中,上述执行主体可以将上述图文信息映射成图文结构化信息。其中,作为示例,上述执行主体可以利用一些现有的图文结构化信息生成工具将得到的图文信息映射为图文结构化信息。结构化信息可用于实现版面内容的重排(Reflow),以适应不同屏幕尺寸的设备特别是移动设备的需求。

[0048] 在一些实施例的一些可选的实现方式中,上述执行主体可以执行以下步骤得到图

文结构化信息:首先,初始化结构化版式文档的根节点,以及生成对应的结构化图文段落节点;其次,在上述根节点下增加上述结构化图文段落节点,以及在上述结构化图文段落节点下增加段落片段节点,以及将待处理的开始页码映射到上述段落片段节点;最后,依据所排序完成的图元信息顺序列表,依次将上述图元信息顺序列表中的图元映射到对应的段落片段节点,至此图文结构化信息映射完毕。

[0049] 可选地,上述执行主体还可以执行以下步骤依次将上述图元信息顺序列表中的图元映射到对应的段落片段节点:第一,响应于第一个图元是跨页的,则新建段落片段节点,以及将所跨页的跨页码映射到上述新建段落片段节点;第二,依次映射上述图元信息顺序列表中的图元到对应的结构化段落片段节点下的块节点,以及增加对应图元类别;第三,若图元类别为非文本类别的,则增加上述图元类别所对应的图元的文字绕排属性,关联上述图元类别所对应的图元对象。

[0050] 本公开的一些实施例提供的方法通过对待处理的图文混排的版式文档的段落进行识别,可以得到段落信息,其中,上述段落信息包括用于表征上述段落所在范围的段落外接边框信息。之后,根据上述段落外接边框信息,可以确定上述段落内的版式图文要素信息。然后,对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息。随后,根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息。最后,将上述图文信息映射成图文结构化信息。由于结构化图文信息描述了图文信息中的结构层次与阅读的顺序,进而,通过生成图文结构化信息,可以使图文内容进行重排。从而,可以提高文档在不同设备上显示的自适应性。通过对版式图文要素信息的阅读顺序进行分析,可以提高图文信息的准确率。进而,可以增强图文结构化信息所呈现出来的显示效果,提高用户阅读体验。

[0051] 进一步参考图3,作为对上述各图所示方法的实现,本公开提供了一种用于生成图文结构化信息的装置的一些实施例,这些装置实施例与图2所示的那些方法实施例相对应,该装置具体可以应用于各种电子设备中。

[0052] 如图3所示,一些实施例的用于生成图文结构化信息的装置300包括:识别单元301、确定单元302、读取判断单元303、分析拆分重排单元304和映射单元305。其中,识别单元301被配置成对待处理的图文混排的版式文档的段落进行识别,得到段落信息,其中,上述段落信息包括用于表征上述段落所在范围的段落外接边框信息;确定单元302被配置成根据上述段落外接边框信息,确定上述段落内的版式图文要素信息;读取判定单元303被配置成对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息;分析拆分重排单元304被配置成根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息;而映射单元305被配置成将上述图文信息映射成图文结构化信息。

[0053] 在一些实施例的可选实现方式中,用于生成图文结构化信息的装置300的确定单元302包括:识别子单元,被配置成根据上述段落外接边框信息,识别段落外接边框范围内的图元信息和图元所在版式页面的版式页面块信息,形成图元信息列表和版式页面块信息列表,其中,上述图元信息至少包括图元类别信息、图元标识信息以及是否跨页信息,上述版式页面块信息至少包括版式页面排版边框的标识信息,并与上述图元信息列表相关联。

[0054] 在一些实施例的可选实现方式中,用于生成图文结构化信息的装置300的读取判

定单元303被进一步配置成分析上述段落内的版式页面块中的图元信息列表中对应的图元类别;响应于上述版式页面块中同时存在文本图元和其它类型图元,则标记分析结果为0;响应于上述版式页面块中只存在某一种相同类型图元,则标记分析结果为1;响应于上述版式页面块中存在其它情形,则标记分析结果为2。

[0055] 在一些实施例的可选实现方式中,用于生成图文结构化信息的装置300的分析拆分重排单元304被进一步配置成基于预设阅读方向,对上述段落内的行进行行高分析;确定上述行中的文本图元对象对应的纵坐标值偏差是否大于预设偏差值;响应于确定大于上述预设偏差值,依据行高将上述段落拆分为多个独立文本图元数据;依据各个文本图元数据的行内横坐标确定各个文本图元的阅读顺序;若上述各个文本图元数据行的横坐标范围内图元对象与上述文本图元对象不同,则进一步拆分上述文本图元数据的文本图元,以得到图元信息顺序列表;重新调整上述图元信息顺序列表中图元所在的版式页面块的阅读顺序,得到图文信息。

[0056] 在一些实施例的可选实现方式中,用于生成图文结构化信息的装置300的映射单元305包括:生成子单元,增加子单元和映射子单元。其中,生成子单元被配置成初始化结构化版式文档的根节点,以及生成对应的结构化图文段落节点;增加子单元被配置成在上述根节点下增加上述结构化图文段落节点,以及在上述结构化图文段落节点下增加段落片段节点,以及将待处理的开始页码映射到上述段落片段节点;映射子单元被配置成依据所排序完成的图元信息顺序列表,依次将上述图元信息顺序列表中的图元映射到对应的段落片段节点,至此图文结构化信息映射完毕。

[0057] 在一些实施例的可选实现方式中,用于生成图文结构化信息的装置300的映射子单元被进一步配置成响应于第一个图元是跨页的,则新建段落片段节点,以及将所跨页的跨页码映射到上述新建段落片段节点;依次映射上述图元信息顺序列表中的图元到对应的结构化段落片段节点下的块节点,以及增加对应图元类别;若图元类别为非文本类别的,则增加上述图元类别所对应的图元的文字绕排属性,关联上述图元类别所对应的图元对象。

[0058] 本公开的一些实施例提供的装置,通过对待处理的图文混排的版式文档的段落进行识别,可以得到段落信息,其中,上述段落信息包括用于表征上述段落所在范围的段落外接边框信息。之后,根据上述段落外接边框信息,可以确定上述段落内的版式图文要素信息。然后,对上述版式图文要素信息进行图元类别信息读取、判定并标记,得到版式图文信息。随后,根据预设的上述版式图文信息的阅读顺序,进行分析、拆分、重排图元对象,得到图文信息。最后,将上述图文信息映射成图文结构化信息。由于结构化图文信息描述了图文信息中的结构层次与阅读的顺序,进而,通过生成图文结构化信息,可以使图文内容进行重排。从而,可以提高文档在不同设备上显示的自适应性。通过对版式图文要素信息的阅读顺序进行分析,可以提高图文信息的准确率。进而,可以增强图文结构化信息所呈现出来的显示效果,提高用户阅读体验。

[0059] 特别地,根据本公开的实施例,上文参考流程图描述的过程可以被实现为计算机软件程序。

[0060] 需要说明的是,本公开的实施例上述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。

[0061] 以上描述仅为本公开的一些较佳实施例以及对所运用技术原理的说明。本领域技

术人员应当理解,本公开的实施例中所涉及的发明范围,并不限于上述技术特征的特定组合而成的技术方案,同时也应涵盖在不脱离上述发明方法的情况下,由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开的实施例中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

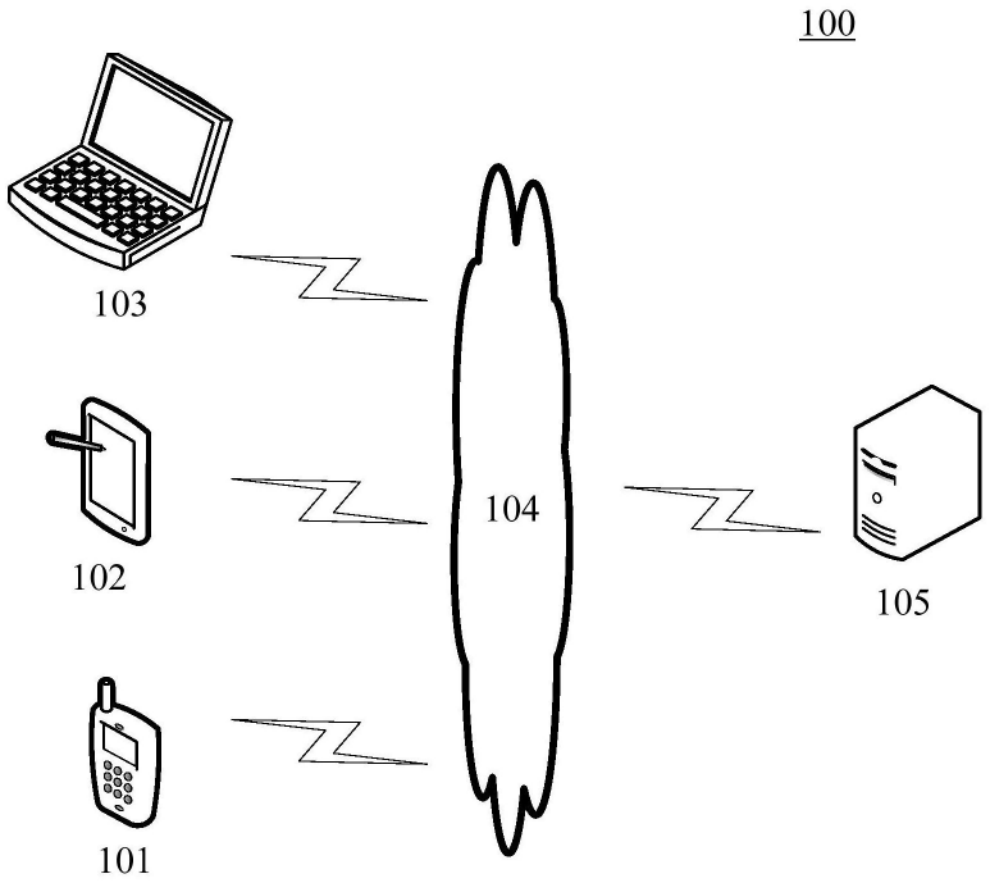


图1

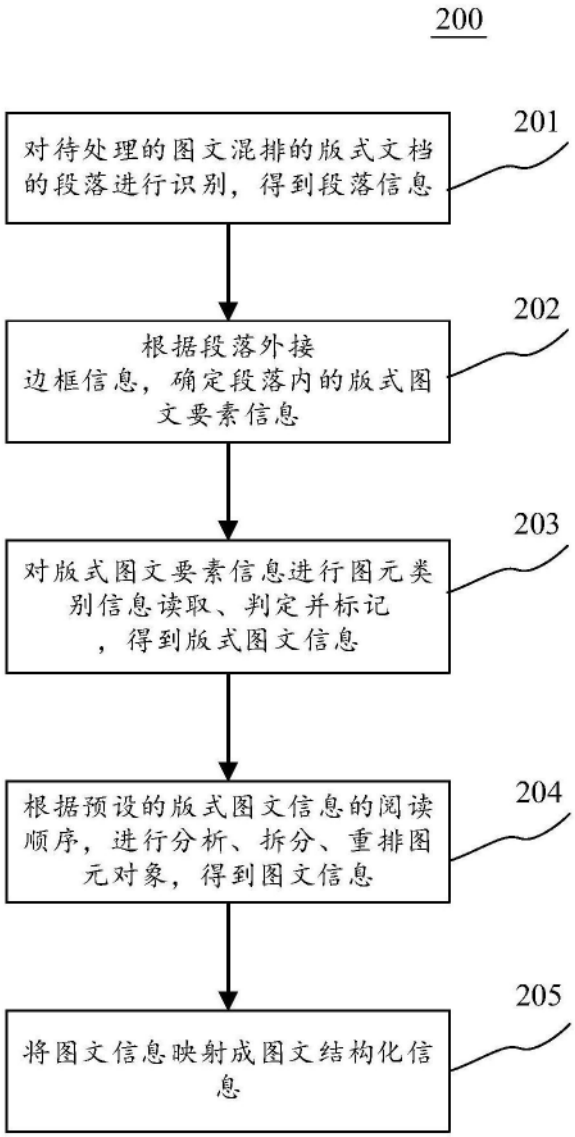

200
对待处理的图文混排的版式文档的段落进行识别，得到段落信息
201
根据段落外接边框信息，确定段落内的版式图文要素信息
202
对版式图文要素信息进行图元类别信息读取、判定并标记，得到版式图文信息
203
根据预设的版式图文信息的阅读顺序，进行分析、拆分、重排图元对象，得到图文信息
204
将图文信息映射成图文结构化信息
205

图2

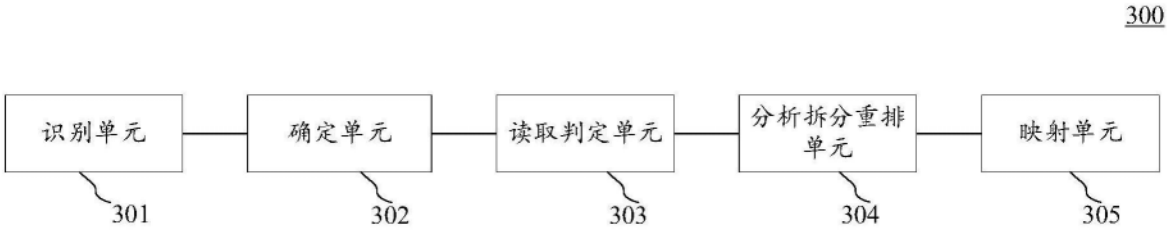

300
识别单元
301
确定单元
302
读取判定单元
303
分析拆分重排单元
304
映射单元
305

图3

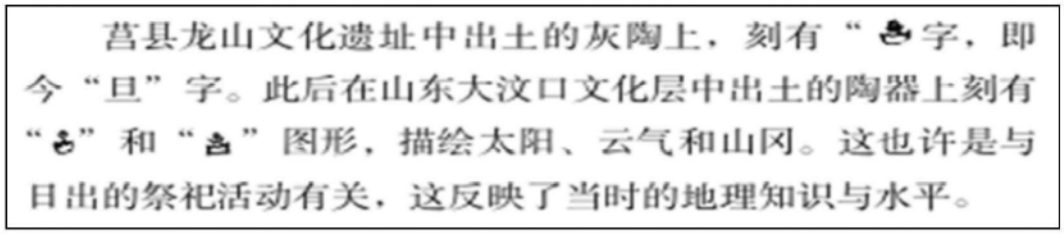

莒县龙山文化遗址中出土的灰陶上，刻有“ 字，即今“旦”字。此后在山东大汶口文化层中出土的陶器上刻有“ ”和“ ”图形，描绘太阳、云气和山冈。这也许是与日出的祭祀活动有关，这反映了当时的地理知识与水平。

图4

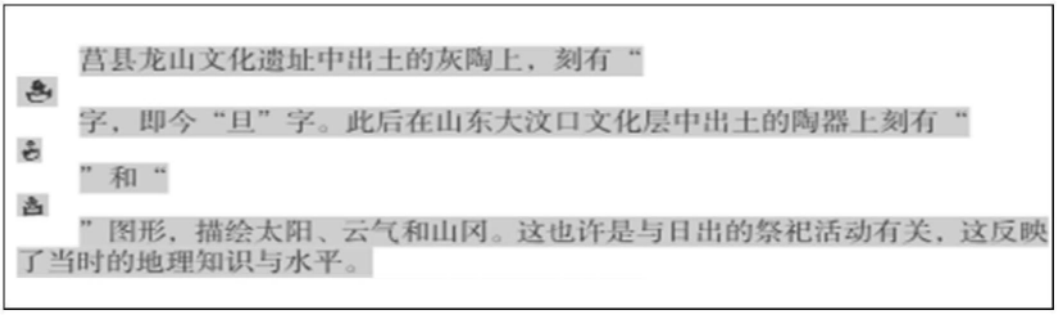

莒县龙山文化遗址中出土的灰陶上，刻有“
字，即今“旦”字。此后在山东大汶口文化层中出土的陶器上刻有“
”和“
”图形，描绘太阳、云气和山冈。这也许是与日出的祭祀活动有关，这反映了当时的地理知识与水平。

图5

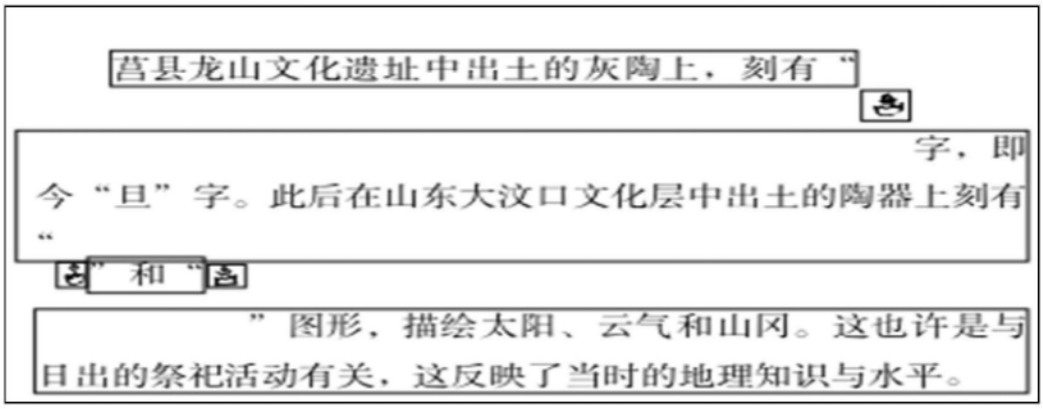

莒县龙山文化遗址中出土的灰陶上，刻有“
字，即今“旦”字。此后在山东大汶口文化层中出土的陶器上刻有“
”和“
”图形，描绘太阳、云气和山冈。这也许是与日出的祭祀活动有关，这反映了当时的地理知识与水平。

图6

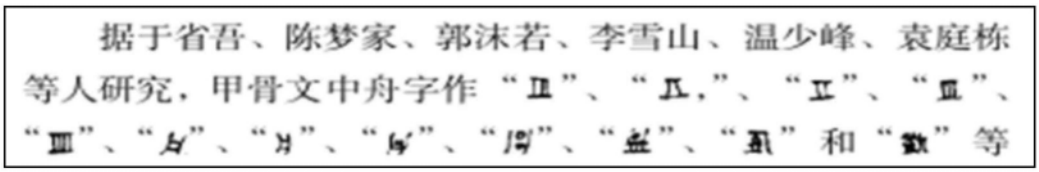

据于省吾、陈梦家、郭沫若、李雪山、温少峰、袁庭栋等人研究，甲骨文中舟字作“”、“”、“”、“”、“”、“”、“”、“”、“”、“”、“”和“”等

图7


据于省吾、陈梦家、郭沫若、李雪山、温少峰、袁庭栋等人研究，甲骨文中舟字作“
”、“
”、“
”、“
”、“
”、“
”、“
”、“
”、“
”、“
”和“
”等

图8

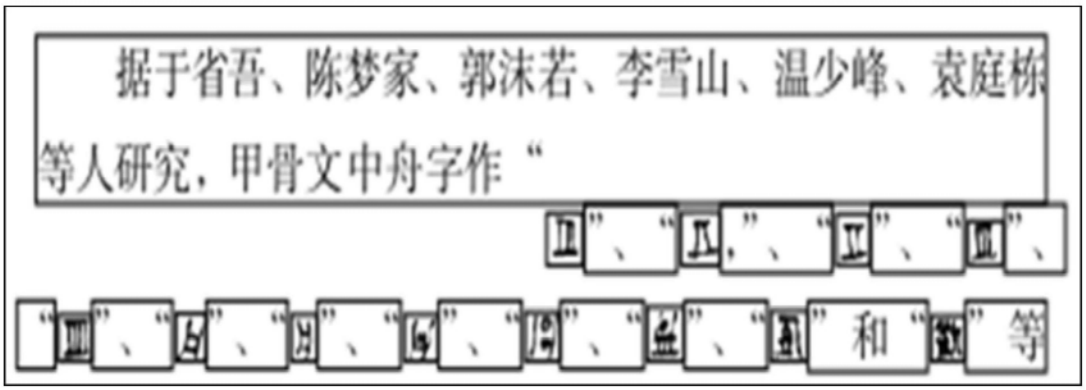

据于省吾、陈梦家、郭沫若、李雪山、温少峰、袁庭栋
等人研究，甲骨文中舟字作“
”、“
”、“
”、“
”、
”、“
”、“
”、“
”、“
”、“
”、“
”和“
”等

图9

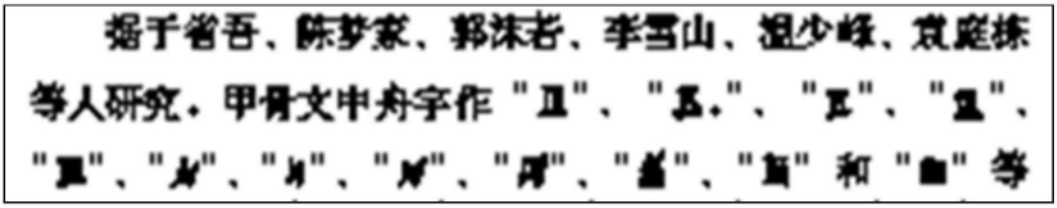

据于省吾、陈梦家、郭沫若、李雪山、温少峰、袁庭栋
等人研究，甲骨文中舟字作"
"、"
"、"
"、"
"、
"、"
"、"
"、"
"、"
"、"
"和"
"等

图10


项目策划／设计制作／紫图图书 ZITO

图11

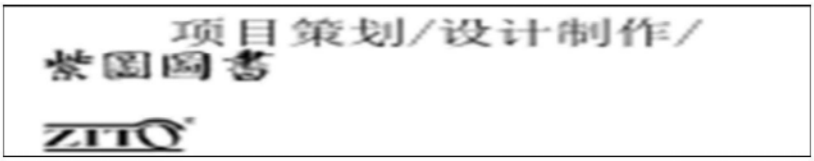

项目策划/设计制作/
紫图图书

图12

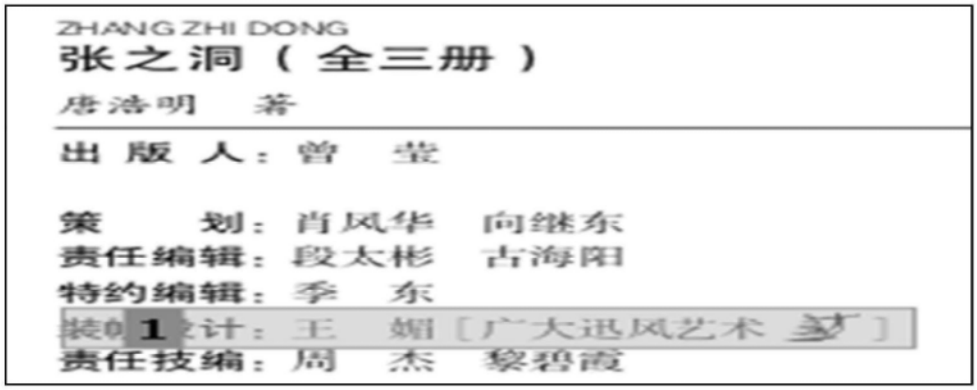

ZHANG ZHI DONG
张之洞（全三册）
唐浩明 著
出 版 人：曾 莹
策 划：肖凤华 向继东
责任编辑：段太彬 古海阳
特约编辑：季 东
装帧设计：王 媚［广大迅风艺术
责任技编：周 杰 黎碧霞

图13

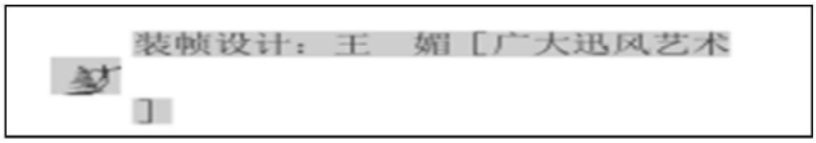

装帧设计：王 媚［广大迅风艺术
］

图14