(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 106776552 B
(45)授权公告日 2018.06.22

(21)申请号 201611113449.4

(22)申请日 2016.12.06

(65)同一申请的已公布的文献号
申请公布号 CN 106776552 A

(43)申请公布日 2017.05.31

(73)专利权人 掌阅科技股份有限公司
地址 100124 北京市朝阳区四惠东通惠河畔四惠大厦2层2029E

(72)发明人 胡元琪

(74)专利代理机构 北京市浩天知识产权代理事务所(普通合伙) 11276
代理人 宋菲 陈翠

(51)Int.Cl.
G06F 17/27(2006.01)
G06F 17/21(2006.01)

(56)对比文件
王爽 等.教学PPT中文字元素的精益化设计研究.《中小学电教》.2009,第1-2页.
苏俊霞.网页风格变换—《CSS》教学案例.《中小学电教》.2014,第64-67页.

审查员 王爽

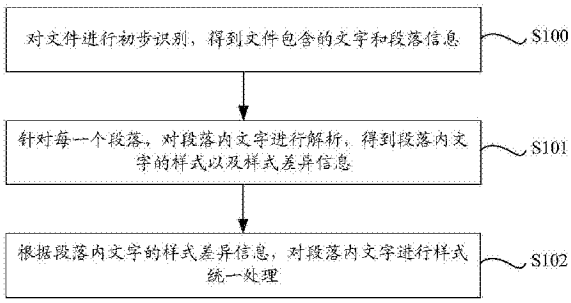
权利要求书3页 说明书10页 附图3页

(54)发明名称

文件识别方法、装置、服务器和计算机存储介质

(57)摘要

本发明公开了一种文件识别方法、装置、服务器和计算机存储介质。其中,方法包括:对文件进行初步识别,得到文件包含的文字和段落信息;针对每一个段落,对段落内文字进行解析,得到段落内文字的样式以及样式差异信息;根据段落内文字的样式差异信息,对段落内文字进行样式统一处理。利用本发明的方案,将段落内文字的样式统一处理成一种样式,解决了文件中文字样式繁多而造成的样式膨胀问题,而且还减少了样式信息的存储量,节省了存储空间,降低了资源占用率。



CN 106776552 B

1.一种文件识别方法,其包括:

对文件进行初步识别,得到所述文件包含的文字和段落信息;

针对每一个段落,对段落内文字进行解析,得到段落内文字的样式以及样式差异信息,其中,文字的样式包括:文字的字号和字体;

根据段落内文字的样式差异信息,对段落内文字进行样式统一处理。

2.根据权利要求1所述的方法,其中,所述对段落内文字进行解析,得到段落内文字的样式以及样式差异信息进一步包括:对段落内文字进行解析,得到段落内文字的字号以及字号差异度;

所述根据段落内文字的样式差异信息,对段落内文字进行样式统一处理进一步包括:

判断段落内文字的字号差异度是否小于或等于预设字号阈值;

若是,则根据段落内文字的字号确定设定字号,将段落内文字的字号统一处理为设定字号。

3.根据权利要求1所述的方法,其中,所述方法还包括:预先设置多个字体集,每个字体集内的字体差异度在预设范围内;

所述对段落内文字进行解析,得到段落内文字的样式以及样式差异信息进一步包括:对段落内文字进行解析,得到段落内文字的字体以及段落内文字的字体所形成的字体集合;

所述根据段落内文字的样式差异信息,对段落内文字进行样式统一处理进一步包括:

判断段落内文字的字体所形成的字体集合是否为预先设置的任意一个字体集的子集;

若是,则根据段落内文字的字体确定设定字体,将段落内文字的字体统一处理为设定字体。

4.根据权利要求2所述的方法,其中,所述根据段落内文字的字号确定设定字号进一步包括:

判断段落内文字的字号与其它段落的设定字号的字号差异度是否小于或等于预设字号阈值,若是,则确定该段落的设定字号与其它段落的设定字号相同。

5.根据权利要求3所述的方法,其中,所述根据段落内文字的字体确定设定字体进一步包括:

判断段落内文字的字体所形成的字体集合与其它段落的设定字体是否为预先设置的任意一个字体集的子集,若是,则确定该段落的设定字体与其它段落的设定字体相同。

6.根据权利要求3所述的方法,其中,在所述根据段落内文字的样式差异信息,对段落内文字进行样式统一处理之前,所述方法还包括:判断段落内文字的字体数量是否大于或等于预设值;

所述根据段落内文字的样式差异信息,对段落内文字进行样式统一处理具体为:若段落内文字的字体数量大于或等于预设值,则根据段落内文字的样式差异信息,对段落内文字进行样式统一处理。

7.根据权利要求1-6中任一项所述的方法,其中,所述对段落内文字进行解析具体为:对段落内除了角标以外的文字进行解析。

8.根据权利要求1-6中任一项所述的方法,其中,在所述对段落内文字进行样式统一处理之后,所述方法还包括:

选取样式相同的至少一个段落，对所述至少一个段落的样式进行统一调整。

9.一种文件识别装置，其包括：

识别模块，适于对文件进行初步识别，得到所述文件包含的文字和段落信息；

解析模块，适于针对每一个段落，对段落内文字进行解析，得到段落内文字的样式以及样式差异信息，其中，文字的样式包括：文字的字号和字体；

处理模块，适于根据段落内文字的样式差异信息，对段落内文字进行样式统一处理。

10.根据权利要求9所述的装置，其中，所述解析模块进一步适于：对段落内文字进行解析，得到段落内文字的字号以及字号差异度；

所述处理模块进一步包括：判断单元，适于判断段落内文字的字号差异度是否小于或等于预设字号阈值；

处理单元，适于判断出段落内文字的字号差异度小于或等于预设字号阈值的情况下，根据段落内文字的字号确定设定字号，将段落内文字的字号统一处理为设定字号。

11.根据权利要求9所述的装置，其中，所述装置还包括：设置模块，适于预先设置多个字体集，每个字体集内的字体差异度在预设范围内；

所述解析模块进一步适于：对段落内文字进行解析，得到段落内文字的字体以及段落内文字的字体所形成的字体集合；

所述处理模块进一步包括：判断单元，适于判断段落内文字的字体所形成的字体集合是否为预先设置的任意一个字体集的子集；

处理单元，适于判断出段落内文字的字体所形成的字体集合为预先设置的任意一个字体集的子集的情况下，根据段落内文字的字体确定设定字体，将段落内文字的字体统一处理为设定字体。

12.根据权利要求10所述的装置，其中，所述处理单元进一步适于：判断段落内文字的字号与其它段落的设定字号的字号差异度是否小于或等于预设字号阈值，若是，则确定该段落的设定字号与其它段落的设定字号相同。

13.根据权利要求11所述的装置，其中，所述处理单元进一步适于：判断段落内文字的字体所形成的字体集合与其它段落的设定字体是否为预先设置的任意一个字体集的子集，若是，则确定该段落的设定字体与其它段落的设定字体相同。

14.根据权利要求11所述的装置，其中，所述装置还包括：判断模块，适于判断段落内文字的字体数量是否大于或等于预设值；

所述处理模块具体适于：在判断出段落内文字的字体数量大于或等于预设值的情况下，根据段落内文字的样式差异信息，对段落内文字进行样式统一处理。

15.根据权利要求9-14中任一项所述的装置，其中，所述解析模块具体适于：对段落内除了角标以外的文字进行解析。

16.根据权利要求9-14中任一项所述的装置，其中，所述装置还包括：调整模块，适于选取样式相同的至少一个段落，对所述至少一个段落的样式进行统一调整。

17.一种服务器，包括：处理器、存储器、通信接口和通信总线，所述处理器、所述存储器和所述通信接口通过所述通信总线完成相互间的通信；

所述存储器用于存放至少一可执行指令，所述可执行指令使所述处理器执行如权利要求1-8中任一项所述的文件识别方法对应的操作。

18.一种计算机存储介质,所述存储介质中存储有至少一可执行指令,所述可执行指令使处理器执行如权利要求1-8中任一项所述的文件识别方法对应的操作。

文件识别方法、装置、服务器和计算机存储介质

技术领域

[0001] 本发明涉及互联网技术领域，具体涉及一种文件识别方法、装置、服务 器和计算机存储介质。

背景技术

[0002] 随着网络技术的发展，人们可以通过不同的设备、不同的途径获得各种 各样的电子文件，这些电子文件极大地丰富了人们的工作和生活内容。

[0003] 然而，随着技术发展，盗版越来越猖獗，为了防止盗版，很多文件在排 版时，会对文字的样式进行不同的设置，例如，一段文字中，设置多种字号 和字体，以增加文件再次排版的难度，然而，很多正规渠道得到的文件也是 类似文件，这样就会造成样式信息膨胀，而且还需要更大的存储空间来存储 这些样式信息，占用资源。

发明内容

[0004] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分 地解决上述问题的文件识别方法、文件识别装置、服务器和计算机存储介质。

[0005] 根据本发明的一个方面，提供了一种文件识别方法，其包括：

[0006] 对文件进行初步识别，得到文件包含的文字和段落信息；

[0007] 针对每一个段落，对段落内文字进行解析，得到段落内文字的样式以及 样式差异信息；

[0008] 根据段落内文字的样式差异信息，对段落内文字进行样式统一处理。

[0009] 根据本发明的另一方面，提供了一种文件识别装置，其包括：

[0010] 识别模块，适于对文件进行初步识别，得到文件包含的文字和段落信息；

[0011] 解析模块，适于针对每一个段落，对段落内文字进行解析，得到段落内 文字的样式以及样式差异信息；

[0012] 处理模块，适于根据段落内文字的样式差异信息，对段落内文字进行样 式统一处理。

[0013] 根据本发明的又一方面，提供了一种服务器，包括：处理器、存储器、 通信接口和通信总线，处理器、存储器和通信接口通过通信总线完成相互间 的通信；

[0014] 存储器用于存放至少一可执行指令，可执行指令使处理器执行上述文件 识别方法对应的操作。

[0015] 根据本发明的再一方面，提供了一种计算机存储介质，存储介质中存储 有至少一可执行指令，可执行指令使处理器执行如上述文件识别方法对应的 操作。

[0016] 根据本发明提供的方案，对文件进行初步识别，得到文件包含的文字和 段落信息，针对每一个段落，对段落内文字进行解析，得到段落内文字的样 式以及样式差异信息，根据段落内文字的样式差异信息，对段落内文字进行 样式统一处理。利用本发明的方案，将段落内文字的样式统一处理成一种样 式，解决了文件中文字样式繁多而造成的样式膨

胀问题，而且还减少了样式 信息的存储量，节省了存储空间，降低了资源占用率。

[0017] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技 术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它 目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0018] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本 领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的， 而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示 相同的部件。在附图中：

[0019] 图1示出了根据本发明一个实施例的文件识别方法的流程示意图；

[0020] 图2示出了根据本发明另一个实施例的文件识别方法的流程示意图；

[0021] 图3示出了根据本发明一个实施例的文件识别装置的结构示意图；

[0022] 图4示出了根据本发明另一个实施例的文件识别装置的结构示意图；

[0023] 图5示出了根据本发明一个实施例的服务器的结构示意图。

具体实施方式

[0024] 下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示 了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不 应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地 理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0025] 图1示出了根据本发明一个实施例的文件识别方法的流程示意图。如图 1所示，该方法包括以下步骤：

[0026] 步骤S100，对文件进行初步识别，得到文件包含的文字和段落信息。

[0027] 对于任一文件，该文件的内容可以包含文字、图片和/或表格，该文件中 的文字又可以组成若干个段落，本发明实施例意在对文件包含的文字的样式 进行处理。在获取到一文件后，需要对文件进行初步识别，主要是为了得到 文件包含的文字和段落信息，其中，段落信息主要是用于区分各段落，能够 确定出文件中哪些文字属于一个段落，哪些文字属于另一个段落。

[0028] 对于既包含文字又包含图片和/或表格的文件，或者仅包含文字的文件但 文件本身被做了特殊处理的情况时，都需要对文件进行初步识别，从中识别 出文字和段落信息，具体地识别算法这里不做具体限定，本领域技术人员可 以根据实际需要进行选择。

[0029] 步骤S101，针对每一个段落，对段落内文字进行解析，得到段落内文字 的样式以及样式差异信息。

[0030] 在根据步骤S100得到文件包含的文字和段落信息后，可以确定出文件中 的各个段落，以及各个段落内的文字，然后，针对每一个段落，需要对段落 内的文字进行解析，这里对段落内的文字进行解析主要是为了确定段落内文 字与文字之间是否有差异，主要是指文字的样式是否存在差异，在对段落内 文字进行解析后，可以得到段落内文字的样式以及样式差异信息，其中，文 字的样式包括：文字的字号和字体。

[0031] 步骤S102，根据段落内文字的样式差异信息，对段落内文字进行样式统 一处理。

[0032] 在得到段落内文字的样式差异信息后，可以根据段落内文字的样式差异 信息，对段落内文字进行样式统一处理，也就是说，将该段落内文字的样式 处理成一种样式，实现文字的样式的统一，具体地，在对段落内文字进行样 式统一处理时，可以将段落内文字的样式统一处理成段落内某个文字的样式；当然，也可以将段落内文字的样式统一处理成一种不同于段落内文字的样式 其他的样式，这里不做具体限定，本领域技术人员可以根据实际需要将段落 内文字的样式统一处理成需要的样式。

[0033] 根据本发明上述实施例提供的方法，对文件进行初步识别，得到文件包 含的文字和段落信息，针对每一个段落，对段落内文字进行解析，得到段落 内文字的样式以及样式差异信息，根据段落内文字的样式差异信息，对段落 内文字进行样式统一处理。利用本发明的方案，将段落内文字的样式统一处 理成一种样式，解决了文件中文字样式繁多而造成的样式膨胀问题，而且还 减少了样式信息的存储量，节省了存储空间，降低了资源占用率。

[0034] 图2示出了根据本发明另一个实施例的文件识别方法的流程示意图。如 图2所示，该方法包括以下步骤：

[0035] 步骤S200，对文件进行初步识别，得到文件包含的文字和段落信息。

[0036] 对于任一文件，该文件的内容可以包含文字、图片和/或表格，该文件中 的文字又可以组成若干个段落，本发明实施例意在对文件包含的文字的样式 进行处理。在获取到一文件后，需要对文件进行初步识别，主要是为了得到 文件包含的文字和段落信息，其中，段落信息主要是用于区分各段落，能够 确定出文件中哪些文字属于一个段落，哪些文字属于另一个段落。

[0037] 这里以PDF文件为例，PDF文件中可以既包含文字又包含图片和/或表 格，或者仅包含文字但该PDF文件本身被做了特殊处理的情况时，都需要对 文件进行初步识别，从中识别出文字和段落信息，具体地识别算法这里不做 具体限定，本领域技术人员可以根据实际需要进行选择。

[0038] 在得到文件包含的文字和段落信息后，针对每一个段落，需要对段落内 文字进行解析，具体地，对段落内除了角标以外的文字进行解析，以确定段 落内文字与文字之间是否有差异，主要指文字的样式是否存在差异，其中， 文字的样式包括：文字的字号和字体，具体地，将分别针对文字的字号和字 体进行详细说明：

[0039] 步骤S201，针对每一个段落，对段落内文字进行解析，得到段落内文字 的字号以及字号差异度。

[0040] 具体地，在根据步骤S200得到文件包含的文字和段落信息后，可以确定 出文件中的各个段落，以及各个段落内的文字，然后，针对每一个段落，需 要对段落内的文字进行解析，这里对段落内的文字进行解析主要是为了确定 段落内文字的字号，并根据文字的字号确定字号差异度，其中字号差异度具 体可以是字号之间的差值，举例说明，在对段落内文字解析后，得到段落内 文字的字号分别为：11、11.1、11.2、11.3、11.5，字号差异度为字号之间的 差值：0.1、0.2、0.3、0.4、0.5。

[0041] 步骤S202，判断段落内文字的字号差异度是否小于或等于预设字号阈 值，若是，则执行步骤S203；若否，则方法结束。

[0042] 为了确定是否对段落内文字的字号进行统一处理，在得到段落内文字的 字号差异度后，需要判断一下段落内文字的字号差异度是否小于或等于预设 字号阈值，其中，预

设字号阈值可以是本领域技术人员根据实际需要进行设 定的，例如，可以设定预设字号阈值为1，若段落内文字的字号差异度小于 或等于1，则可以对段落内文字的字号进行统一处理；若段落内文字的字号 差异度大于1，则不对段落内文字的字号进行统一处理。

[0043] 步骤S203，根据段落内文字的字号确定设定字号，将段落内文字的字号 统一处理为设定字号。

[0044] 在判断出段落内文字的字号差异度小于或等于预设字号阈值的情况下， 需要对段落内文字的字号进行处理，将段落内文字的字号统一处理为一种字 号，具体地，可以根据段落内文字的字号来确定文字被统一处理后的字号， 即，设定字号，这里以步骤S201中得到的段落内文字的字号分别为：11、 11.1、11.2、11.3、11.5为例进行说明，在本步骤中，可以将设定字号确定为 11、11.1、11.2、11.3、11.5中的一个，例如，可以确定设定字号为11；当然 也可以将设定字号确定为其他的字号，这里不做具体限定，在根据段落内文 字的字号确定设定字号后，将段落内文字的字号统一处理为设定字号，在本 步骤中，主要是将一个段落内文字的字号统一处理为设定字号。

[0045] 一般情况下，任一文件都包含多个段落，在对文件内容进行排版时，可 能会对段落内的文字的字号进行不同的设置，因此，可能存在一些段落的文 字的字号相似或相同，而另一些段落的文字的字号较大或较小的情况。而本 发明实施例不仅能够实现段落内文字的字号的统一，还能够实现段落间文字 的字号进行统一处理，当然，只在满足相应的条件的情况下，才会将段落间 文字的字号统一处理成一种字号。具体地，需要判断段落内文字的字号与其 它段落的设定字号的字号差异度是否小于或等于预设字号阈值，若是，则确定该段落的设定字号与其它段落的设定字号相同；若否，则根据段落内文字 的字号确定设定字号，将段落内文字的字号统一处理为设定字号。

[0046] 举例说明，预设字号阈值为1，利用步骤S201-步骤S203将段落1内文 字的字号统一处理为11号，利用步骤S201得到段落2内文字的字号分别为： 8、8.1、8.2、8.3、8.5，字号差异度为字号之间的差值：0.1、0.2、0.3、0.4、 0.5，利用步骤S202判断出段落内文字的字号差异度小于或等于预设字号阈 值1，则需要判断段落2内文字的字号与段落1的设定字号的字号差异度是 否小于或等于预设字号阈值1，分别计算段落2内文字的字号：8、8.1、8.2、8.3、8.5与段落1的设定字号11的字号差异度分别为：3、2.9、2.8、2.7、2.5， 计算得到的字号差异度大于预设字号阈值1，则根据段落2内文字的字号确 定设定字号，例如8，将段落2内文字的字号统一处理为设定字号8。

[0047] 利用步骤S201得到段落3内文字的字号分别为：10、10.1、10.2、10.3、 10.5，字号差异度为字号之间的差值：0.1、0.2、0.3、0.4、0.5，利用步骤S202 判断出段落内文字的字号差异度小于或等于预设字号阈值1，则需要判断段 落3内文字的字号与段落1、2的设定字号的字号差异度是否小于或等于预设 字号阈值1，计算段落3内文字的字号与段落1的设定字号11的字号差异度 小于1，段落3内文字的字号与段落2的设定字号8的字号差异度大于1，则 确定该段落3的设定字号与段落1的设定字号相同，设定字号为11，然后将 段落3内文字的字号统一处理为设定字号11。

[0048] 本发明实施例在判断出段落内文字的字号与其它段落的设定字号的字号 差异度小于或等于预设字号阈值的情况下，确定该段落的设定字号与其它段 落的设定字号相同，将该段落的字号与其他段落的字号统一成一种字号，方 便后续对具有相同字号的段落的

统一处理，例如，统一调整字号。

[0049] 本发明不仅可以对字号进行统一处理，还可以对字体进行统一处理，具 体地，可以采用如下方法对段落内字体进行处理：

[0050] 步骤S204，预先设置多个字体集，每个字体集内的字体差异度在预设范 围内。

[0051] 在对段落内文字的字体进行处理之前，需要预先设置多个字体集，其中， 每个字体集中存储的字体具有一定的相似性，在查看这些字体对应的文字时， 并不能很明显地区分出这些文字的字体有何不同，也就是说，每个字体集内 的字体差异度在预设范围内，举例说明，预先设置了字体集，字体集1为{ 宋体，新宋体，仿宋，仿宋_GB2312，华文仿宋}、字体集2为{华文楷体、 楷体、楷体_GB2312}。

[0052] 步骤S205，针对每一个段落，对段落内文字进行解析，得到段落内文字 的字体以及段落内文字的字体所形成的字体集合。

[0053] 具体地，在根据步骤S200得到文件包含的文字和段落信息后，可以确定 出文件中的各个段落，以及各个段落内的文字，然后，针对每一个段落，这 里对段落内的文字进行解析主要是为了确定段落内文字的字体，并根据文字 的字体确定段落内文字的字体所形成的字体集合，举例说明，针对段落1， 对段落1内文字进行解析，得到段落内文字的字体分别为：宋体、新宋体、 仿宋，以及段落内文字的字体所形成的字体集合{宋体，新宋体，仿宋}。

[0054] 步骤S206，判断段落内文字的字体数量是否大于或等于预设值，若是， 则步骤S207；若否，则方法结束。

[0055] 在进行文字排版时，很可能存在由于特殊排版需要而对段落内文字的字 体进行不同的设置，例如为了突出显示等而将段落内部分文字的字体设置成 与其他文字的字体不同，对于这种情况就不需要对段落内文字的字体进行统 一处理，这种情况下，段落内文字的字体一般是两种，当然，也可能存储多 种的情况，因此，需要判断段落内文字的字体数量是否大于或等于预设值， 例如，预设值为3，也就是判断段落内文字的字体数量是否大于或等于3，若 段落内文字的字体数量大于或等于3，表明需要对段落内文字的字体进行统一处理；若段落内文字的字体数量小于3，表明不需要对段落内文字的字体 进行统一处理，方法结束。其中，预设值是根据实际需要进行设置的，这里 不做具体限定。

[0056] 步骤S207，判断段落内文字的字体所形成的字体集合是否为预先设置的 任意一个字体集的子集，若是，则步骤S208；若否，则方法结束。

[0057] 根据步骤S205得到段落内文字的字体所形成的字体集合为{宋体，新宋 体，仿宋}，根据步骤S206判断出段落内文字的字体数量等于3后，需要判 断段落内文字的字体所形成的字体集合是否为预先设置的任意一个字体集的 子集，这里判断段落内文字的字体所形成的字体集合是否为预先设置的任意 一个字体集的子集，主要是为了确定是否需要对段落内文字的字体进行统一 处理，若段落内文字的字体所形成的字体集合为预先设置的任意一个字体集 的子集，则表明需要对段落内文字的字体进行统一处理，若段落内文字的字 体所形成的字体集合不是预先设置的任意一个字体集的子集，则表明不需要 对段落内文字的字体进行统一处理。

[0058] 具体地，可以判断{宋体，新宋体，仿宋}是否为字体集1{宋体，新 宋体，仿宋，仿宋_GB2312，华文仿宋}或字体集2{华文楷体、楷体、楷 体_GB2312}的子集。

[0059] 步骤S208，根据段落内文字的字体确定设定字体，将段落内文字的字体 统一处理

为设定字体。

[0060] 在判断出段落内文字的字体所形成的字体集合为预先设置的任意一个字 体集的子集的情况下，需要对段落内文字的字体进行处理，将段落内文字的 字体统一处理为一种字体，具体地，可以根据段落内文字的字体来确定文字 被统一处理后的字体，即，设定字体，这里以步骤S205中得到的段落内文 字的字体分别为宋体、新宋体、仿宋为例进行说明，在本步骤中，可以将设 定字体确定为宋体、新宋体、仿宋中的一个，例如，可以确定设定字体为宋 体；当然也可以将设定字体确定为其他的字体，这里不做具体限定，在根据 段落内文字的字体确定设定字体后，将段落内文字的字体统一处理为设定字 体，在本步骤中，主要是将一个段落内文字的字体统一处理为设定字体。

[0061] 而本发明实施例不仅能够实现段落内文字的字体的统一，还能够实现段 落间文字的字体进行统一处理，当然，只在满足相应的条件的情况下，才会 将段落间文字的字体统一处理成一种字体。具体地，需要判断段落内文字的 字体所形成的字体集合与其它段落的设定字体是否为预先设置的任意一个字 体集的子集，若是，则确定该段落的设定字体与其它段落的设定字体相同； 若否，则根据段落内文字的字体确定设定字体，将段落内文字的字体统一处 理为设定字体。

[0062] 本发明实施例在判断出段落内文字的字体所形成的字体集合与其它段落 的设定字体为预先设置的任意一个字体集的子集的情况下，确定该段落的设 定字体与其它段落的设定字体相同，将该段落的字体与其他段落的字体统一 成一种字体，方便后续对具有相同字体的段落的统一处理，例如，统一调整 字体。

[0063] 在本发明实施例中，可以同时对段落内文字的字号和字体进行处理，也 可以先对段落内文字的字号进行处理，再对字体进行处理，或者先对段落内 文字的字体进行处理，再对字号进行处理，本实施例仅仅是举例说明，并未 具体限定处理顺序。

[0064] 步骤S209，选取样式相同的至少一个段落，对至少一个段落的样式进行 统一调整。

[0065] 在对段落内文字的字号和字体进行统一处理后，可以根据需求选取样式 相同的至少一个段落，对至少一个段落的样式进行统一调整。

[0066] 根据本发明上述实施例提供的方法，通过对段落内文字的字号和字体进 行统一处理，解决了文件中文字样式繁多而造成的样式膨胀问题，而且还减 少了样式信息的存储量，节省了存储空间，降低了资源占用率，通过将段落 内文字的字号和字体确定为与其他段落的字号和字体相同，实现了对具有相 同样式的段落的统一调整，提高了效率，避免重复操作，节省时间。

[0067] 图3示出了根据本发明一个实施例的文件识别装置的结构示意图。如图 3所示，该装置包括：识别模块300、解析模块310、处理模块320。

[0068] 识别模块300，适于对文件进行初步识别，得到文件包含的文字和段落 信息。

[0069] 解析模块310，适于针对每一个段落，对段落内文字进行解析，得到段 落内文字的样式以及样式差异信息。

[0070] 处理模块330，适于根据段落内文字的样式差异信息，对段落内文字进 行样式统一处理。

[0071] 根据本发明上述实施例提供的装置，对文件进行初步识别，得到文件包 含的文字

和段落信息，针对每一个段落，对段落内文字进行解析，得到段落 内文字的样式以及样式差异信息，根据段落内文字的样式差异信息，对段落 内文字进行样式统一处理。利用本发明的方案，将段落内文字的样式统一处 理成一种样式，解决了文件中文字样式繁多而造成的样式膨胀问题，而且还 减少了样式信息的存储量，节省了存储空间，降低了资源占用率。

[0072] 图4示出了根据本发明另一个实施例的文件识别装置的结构示意图。如 图4所示，该装置包括：识别模块400、解析模块410、处理模块420。

[0073] 识别模块400，适于对文件进行初步识别，得到文件包含的文字和段落 信息。

[0074] 解析模块410，适于对段落内文字进行解析，得到段落内文字的字号以 及字号差异度。

[0075] 处理模块420包括：判断单元421，适于判断段落内文字的字号差异度 是否小于或等于预设字号阈值。

[0076] 处理单元422，适于判断出段落内文字的字号差异度小于或等于预设字 号阈值的情况下，根据段落内文字的字号确定设定字号，将段落内文字的字 号统一处理为设定字号。

[0077] 此外，处理单元422还进一步适于：判断段落内文字的字号与其它段落 的设定字号的字号差异度是否小于或等于预设字号阈值，若是，则确定该段 落的设定字号与其它段落的设定字号相同。

[0078] 该装置还包括：设置模块430，适于预先设置多个字体集，每个字体集 内的字体差异度在预设范围内；

[0079] 解析模块410进一步适于：对段落内文字进行解析，得到段落内文字的 字体以及段落内文字的字体所形成的字体集合；

[0080] 判断单元421进一步适于：判断段落内文字的字体所形成的字体集合是 否为预先设置的任意一个字体集的子集；

[0081] 处理单元422进一步适于：判断出段落内文字的字体所形成的字体集合 为预先设置的任意一个字体集的子集的情况下，根据段落内文字的字体确定 设定字体，将段落内文字的字体统一处理为设定字体。

[0082] 此外，处理单元422进一步适于：判断段落内文字的字体所形成的字体 集合与其它段落的设定字体是否为预先设置的任意一个字体集的子集，若是， 则确定该段落的设定字体与其它段落的设定字体相同。

[0083] 该装置还包括：判断模块440，适于判断段落内文字的字体数量是否大 于或等于预设值；

[0084] 处理模块420具体适于：在判断出段落内文字的字体数量大于或等于预 设值的情况下，根据段落内文字的样式差异信息，对段落内文字进行样式统 一处理。

[0085] 解析模块410具体适于：对段落内除了角标以外的文字进行解析。

[0086] 装置还包括：调整模块450，适于选取样式相同的至少一个段落，对至 少一个段落的样式进行统一调整。

[0087] 根据本发明上述实施例提供的装置，通过对段落内文字的字号和字体进 行统一处理，解决了文件中文字样式繁多而造成的样式膨胀问题，而且还减 少了样式信息的存储量，节省了存储空间，降低了资源占用率，通过将段落 内文字的字号和字体确定为与其他

段落的字号和字体相同，实现了对具有相 同样式的段落的统一调整，提高了效率，避免重复操作，节省时间。

[0088] 本发明还提供了一种非易失性计算机存储介质，计算机存储介质存储有 至少一可执行指令，该计算机可执行指令可执行上述任意方法实施例中的文 件识别方法。

[0089] 图5示出了根据本发明一个实施例的服务器的结构示意图，本发明具体 实施例并不对服务器的具体实现做限定。

[0090] 如图5所示，该服务器可以包括：处理器(processor)502、通信接口(Communications Interface)504、存储器(memory)506、以及通信总线508。

[0091] 其中：

[0092] 处理器502、通信接口504、以及存储器506通过通信总线508完成相互 间的通信。

[0093] 通信接口504，用于与其它设备比如客户端或其它服务器等的网元通 信。

[0094] 处理器502，用于执行程序510，具体可以执行上述文件识别方法实施 例中的相关步骤。

[0095] 具体地，程序510可以包括程序代码，该程序代码包括计算机操作指 令。

[0096] 处理器502可能是中央处理器CPU，或者是特定集成电路ASIC (Application Specific Integrated Circuit)，或者是被配置成实施本发明实施例 的一个或多个集成电路。服务器包括的一个或多个处理器，可以是同一类型 的处理器，如一个或多个CPU；也可以是不同类型的处理器，如一个或多 个CPU以及一个或多个ASIC。

[0097] 存储器506，用于存放第一数据集合、第二数据集合以及程序510。存 储器506可能包含高速RAM存储器，也可能还包括非易失性存储器 (non-volatile memory)，例如至少一个磁盘存储器。

[0098] 程序510具体可以用于使得处理器502执行以下操作：对文件进行初步 识别，得到文件包含的文字和段落信息；针对每一个段落，对段落内文字进 行解析，得到段落内文字的样式以及样式差异信息；根据段落内文字的样式 差异信息，对段落内文字进行样式统一处理。

[0099] 在一种可选的实施方式中，程序510还用于使得处理器502在对段落内 文字进行解析，得到段落内文字的样式以及样式差异信息时：对段落内文字 进行解析，得到段落内文字的字号以及字号差异度；

[0100] 程序510还用于使得处理器502在根据段落内文字的样式差异信息，对 段落内文字进行样式统一处理时：判断段落内文字的字号差异度是否小于或 等于预设字号阈值；若是，则根据段落内文字的字号确定设定字号，将段落 内文字的字号统一处理为设定字号。

[0101] 在一种可选的实施方式中，程序510还用于使得处理器502预先设置多 个字体集，每个字体集内的字体差异度在预设范围内；

[0102] 程序510还用于使得处理器502在对段落内文字进行解析，得到段落内 文字的样式以及样式差异信息时：对段落内文字进行解析，得到段落内文字 的字体以及段落内文字的字体所形成的字体集合；

[0103] 程序510还用于使得处理器502在根据段落内文字的样式差异信息，对 段落内文字进行样式统一处理时：判断段落内文字的字体所形成的字体集合 是否为预先设置的任意一个字体集的子集；若是，则根据段落内文字的字体 确定设定字体，将段落内文字的字

体统一处理为设定字体。

[0104] 在一种可选的实施方式中，程序510还用于使得处理器502在根据段落 内文字的字号确定设定字号时：判断段落内文字的字号与其它段落的设定字 号的字号差异度是否小于或等于预设字号阈值，若是，则确定该段落的设定 字号与其它段落的设定字号相同。

[0105] 在一种可选的实施方式中，程序510还用于使得处理器502在根据段落 内文字的字体确定设定字体时：判断段落内文字的字体所形成的字体集合与 其它段落的设定字体是否为预先设置的任意一个字体集的子集，若是，则确 定该段落的设定字体与其它段落的设定字体相同。

[0106] 在一种可选的实施方式中，程序510还用于使得处理器502在根据段落 内文字的样式差异信息，对段落内文字进行样式统一处理之前，判断段落内 文字的字体数量是否大于或等于预设值；

[0107] 程序510还用于使得处理器502在根据段落内文字的样式差异信息，对 段落内文字进行样式统一处理时：若段落内文字的字体数量大于或等于预设 值，则根据段落内文字的样式差异信息，对段落内文字进行样式统一处理。

[0108] 在一种可选的实施方式中，程序510还用于使得处理器502在对段落内 文字进行解析时：对段落内除了角标以外的文字进行解析。

[0109] 在一种可选的实施方式中，程序510还用于使得处理器502在对段落内 文字进行样式统一处理之后，选取样式相同的至少一个段落，对至少一个段 落的样式进行统一调整。

[0110] 程序510中各步骤的具体实现可以参见上述业务对象数据处理实施例中 的相应步骤和单元中对应的描述，在此不赘述。所属领域的技术人员可以清 楚地了解到，为描述的方便和简洁，上述描述的设备和模块的具体工作过程， 可以参考前述方法实施例中的对应过程描述，在此不再赘述。

[0111] 可见，通过将段落内文字的样式统一处理成一种样式，解决了文件中文 字样式繁多而造成的样式膨胀问题，而且还减少了样式信息的存储量，节省 了存储空间，降低了资源占用率。

[0112] 在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固 有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描 述，构造这类系统所要求的结构是显而易见的。此外，本发明也不针对任何 特定编程语言。应当明白，可以利用各种编程语言实现在此描述的本发明的 内容，并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

[0113] 在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本 发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中，并未 详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

[0114] 类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个 或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时 被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开 的方法解释成反映如下意图：即所要求保护的本发明要求比在每个权利要求 中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映 的那样，发明方面在于少于前面公开的单个实施例的所有特征。

因此，遵循 具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

[0115] 本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自 适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以 把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可 以把它们分成多个子模块或子单元或子组件。除了这样的特征和/或过程或 者单元中的至少一些是相互排斥之外，可以采用任何组合对本说明书(包括 伴随的权利要求、摘要和附图)中公开的所有特征以及如此公开的任何方法 或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书(包括 伴随的权利要求、摘要和附图)中公开的每个特征可以由提供相同、等同或 相似目的的替代特征来代替。

[0116] 此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其 它实施例中所包括的某些特征而不是其它特征，但是不同实施例的特征的组 合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权 利要求书中，所要求保护的实施例的任意之一都可以以任意的组合方式来使 用。

[0117] 应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实 施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要 求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于 元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以 借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在 列举了若干装置的单元权利要求中，这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

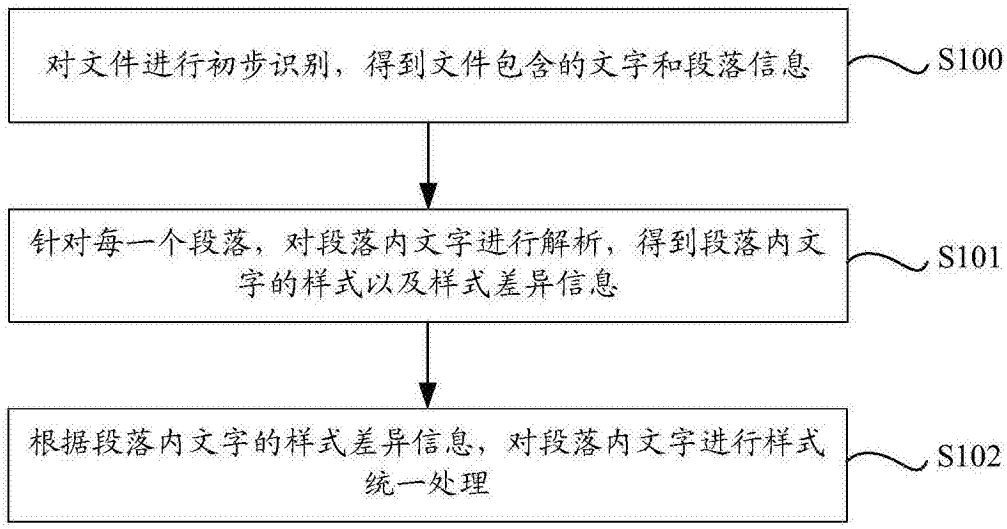

对文件进行初步识别，得到文件包含的文字和段落信息
S100
针对每一个段落，对段落内文字进行解析，得到段落内文字的样式以及样式差异信息
S101
根据段落内文字的样式差异信息，对段落内文字进行样式统一处理
S102

图1

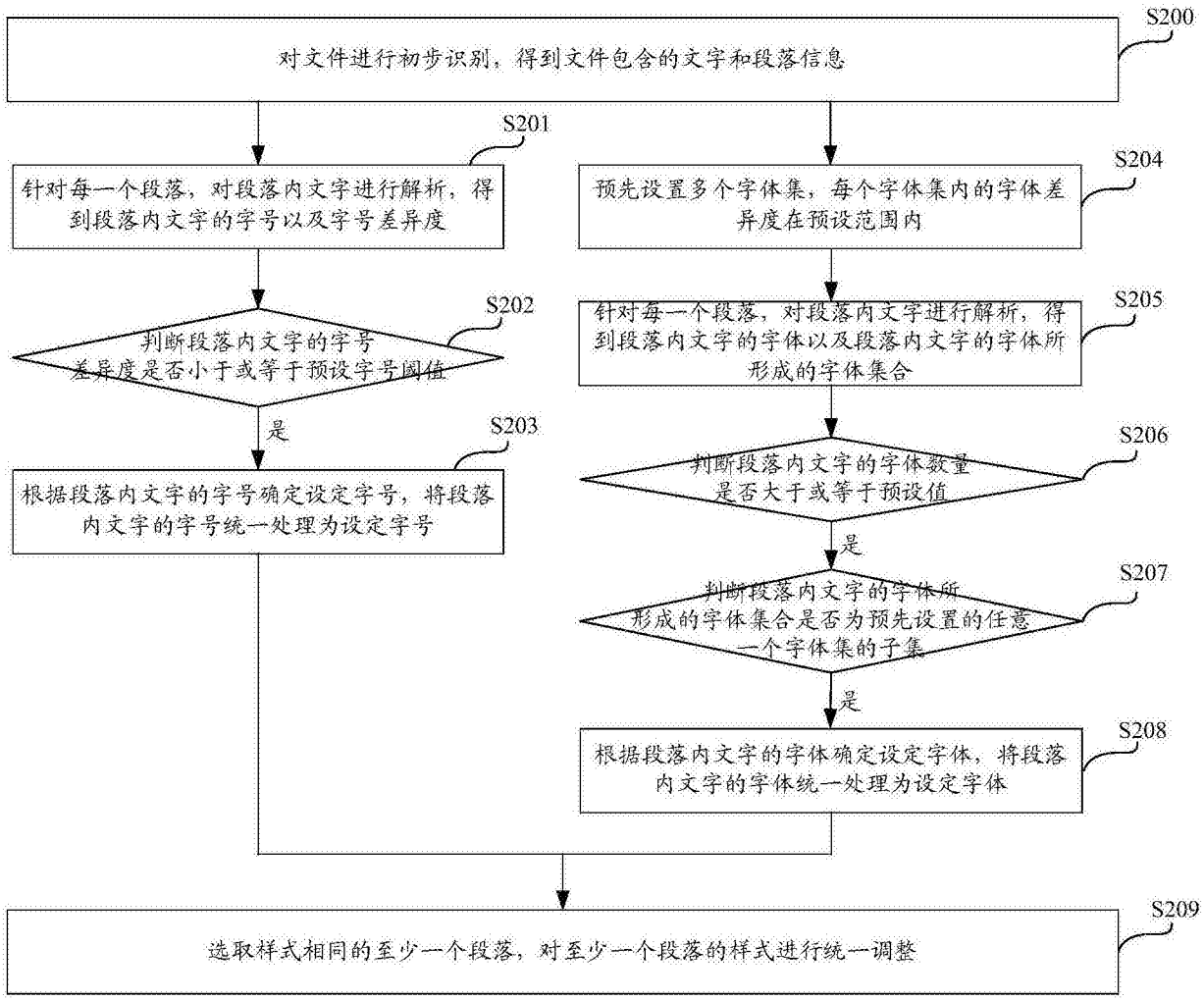

对文件进行初步识别，得到文件包含的文字和段落信息
S200
针对每一个段落，对段落内文字进行解析，得到段落内文字的字号以及字号差异度
S201
判断段落内文字的字号差异度是否小于或等于预设字号阈值
S202
是
根据段落内文字的字号确定设定字号，将段落内文字的字号统一处理为设定字号
S203
预先设置多个字体集，每个字体集内的字体差异度在预设范围内
S204
针对每一个段落，对段落内文字进行解析，得到段落内文字的字体以及段落内文字的字体所形成的字体集合
S205
判断段落内文字的字体数量是否大于或等于预设值
S206
是
判断段落内文字的字体所形成的字体集合是否为预先设置的任意一个字体集的子集
S207
是
根据段落内文字的字体确定设定字体，将段落内文字的字体统一处理为设定字体
S208
选取样式相同的至少一个段落，对至少一个段落的样式进行统一调整
S209

图2

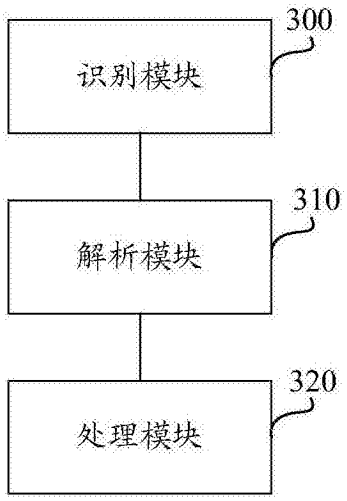

识别模块
300
解析模块
310
处理模块
320

图3

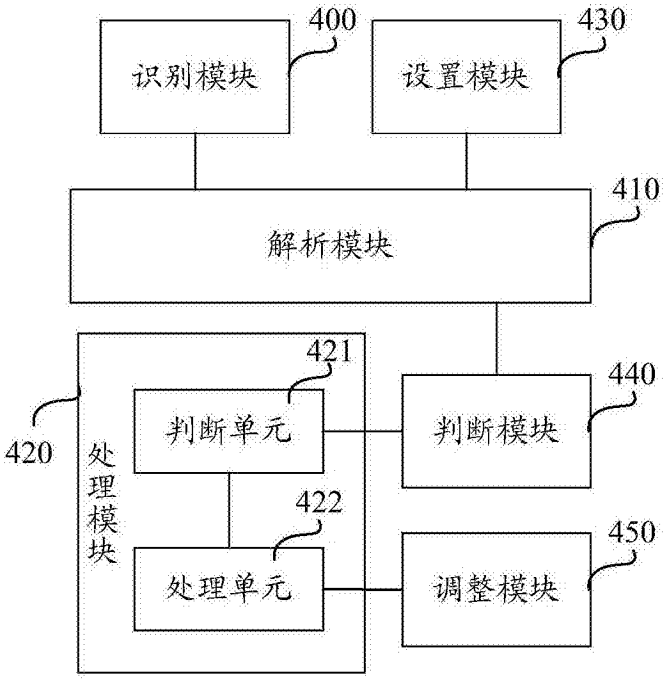

识别模块
400
设置模块
430
解析模块
410
420
处理模块
421
判断单元
422
处理单元
判断模块
440
调整模块
450

图4

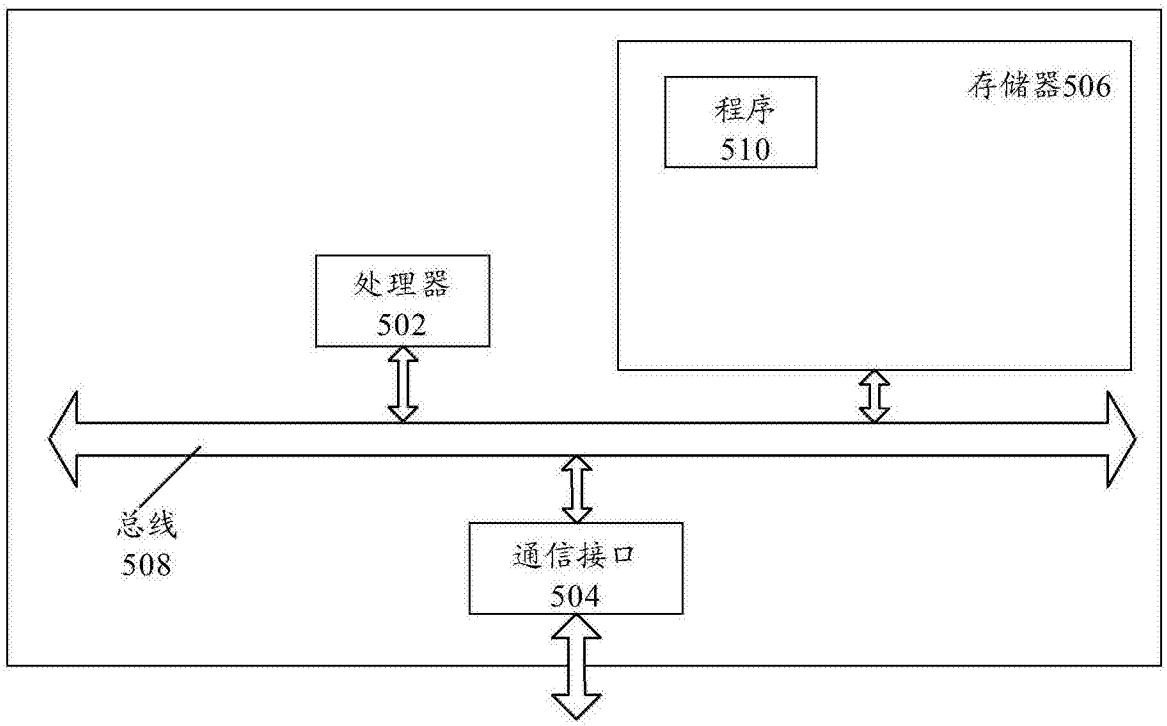


图5