(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110334346 A
(43)申请公布日 2019.10.15

(21)申请号 201910560227.4

(22)申请日 2019.06.26

(71)申请人 京东数字科技控股有限公司
地址 100176 北京市北京经济技术开发区科创十一街18号C座2层221室

(72)发明人 郑宇宇

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 李阳 赵迪

(51)Int.Cl.
G06F 17/27(2006.01)
G06K 9/00(2006.01)

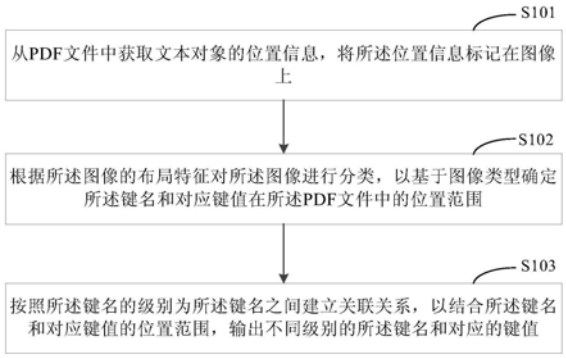
权利要求书2页 说明书12页 附图7页

(54)发明名称

一种PDF文件的信息抽取方法和装置

(57)摘要

本发明公开了一种PDF文件的信息抽取方法和装置，涉及计算机技术领域。该方法的一具体实施方式包括：从PDF文件中获取文本对象的位置信息，将位置信息标记在图像上；其中，文本对象包括至少一个键名和对应的键值；根据图像的布局特征对图像进行分类，以基于图像类型确定键名和对应键值在PDF文件中的位置范围；按照键名的级别为键名之间建立关联关系，以结合键名和对应键值的位置范围，输出不同级别的键名和对应的键值。该方法通过将PDF文件中文本对象的位置标记在图像上，并根据布局特征对图像分类后，按照图像类型确定键名、相应键值的位置，建立各级键名之间的关联关系，进而结合位置和关联关系结构化输出键名和对应键值，提高了信息抽取性能。



CN 110334346 A

1.一种PDF文件的信息抽取方法,其特征在于,包括:

从PDF文件中获取文本对象的位置信息,将所述位置信息标记在图像上;其中,所述文本对象包括至少一个键名和对应的键值;

根据所述图像的布局特征对所述图像进行分类,以基于图像类型确定所述键名和对应键值在所述PDF文件中的位置范围;

按照所述键名的级别为所述键名之间建立关联关系,以结合所述键名和对应键值的位置范围,输出不同级别的所述键名和对应的键值。

2.根据权利要求1所述的方法,其特征在于,所述将所述位置信息标记在图像上,包括:

根据多个所述文本对象之间横坐标的异同,以及纵坐标间隔与预设第一阈值的差值,判断多个所述文本对象是否支持被抽象为归属于同一条线段的点;

若多个所述文本对象支持被抽象为归属于同一条线段的点,则分别获取多个所述文本对象的横坐标的最值和纵坐标的最值,将所述最值对应的线段显示在至少一幅图像上。

3.根据权利要求1所述的方法,其特征在于,对于左右布局的图像类型,所述确定所述键名和对应键值在所述PDF文件中的位置范围,包括:

以字符为基本单元,将所述PDF文件的原始信息拆分为至少一个元素,将纵坐标相同、横坐标差值小于预设第二阈值的元素组合成元素集;

按照所述元素集的横坐标对所述元素集进行排序,以确定边界横坐标和位于所述边界横坐标之间的分界横坐标,

将位于左边界横坐标和所述分界横坐标之间,且行间距大于预设第三阈值的元素集作为所述键名,确定所述键名在所述PDF文件的坐标区间;

根据相邻两个所述键名的坐标区间,确定与其中一个所述键名对应的键值在所述PDF文件的坐标区间。

4.根据权利要求1所述的方法,其特征在于,对于上下布局的图像类型,所述确定所述键名和对应键值在所述PDF文件中的位置范围,包括:

以字符为基本单元,将所述PDF文件的原始信息拆分为至少一个元素,将位于同一行的元素组合成元素集;

将起始横坐标位于所述PDF文件左侧,行间距大于预设第五阈值,和/或以冒号字符结尾的元素集作为所述键名,确定所述键名在所述PDF文件的坐标区间;

根据相邻两个所述键名的坐标区间,确定与其中一个所述键名对应的键值在所述PDF文件的坐标区间。

5.根据权利要求1所述的方法,其特征在于,按照所述键名的级别为所述键名之间建立关联关系,包括:

将同级的键名并联,将上下级的键名串联,采用树形结构为所述键名之间建立关联关系;

所述输出不同级别的所述键名和对应的键值,包括:

采用先序遍历所述树形结构的方式,顺序输出不同级别的所述键名和对应的键值。

6.根据权利要求1所述的方法,其特征在于,所述方法还包括:

提取原始PDF文件的设定参考对象的基本信息,以根据所述基本信息确定所述原始PDF文件的非正文区域;

过滤所述原始PDF文件的非正文区域，将过滤结果作为所述PDF文件。

7.根据权利要求6所述的方法，其特征在于，所述参考对象包括下列任意一项或者多项：边框、线段、图片和文本，所述基本信息包括下列任意一项或者多项：字体、线段粗细、高度、宽度、横纵坐标和文本内容；所述非正文区域包括：目录、表格和注释；

所述根据所述基本信息确定所述原始PDF文件的非正文区域，包括：

根据所述目录的特征确定所述目录的提取维度，按照所述提取维度确定所述目录的上下边界；

获取所述原始PDF文件的最小字体，将首字符的字体等于最小字体，且位于所述原始PDF文件底部的区域作为所述注释的上下边界；

确定所述表格的边界线坐标，以及所述表格的单元格，以得到所述表格的形状和上下边界。

8.根据权利要求1-7的任一项所述的方法，其特征在于，在所述PDF文件包括多页的情况下，所述方法还包括：

将相邻页面的后一页的第一横纵坐标更新为前一页的第二横纵坐标；

计算所述前一页的第二横坐标与所述后一页的第二横坐标的和，将计算出的第一和值与所述后一页的第一横坐标做差得到第一差值，更新所述后一页的第二横纵坐标为所述第一差值；

计算所述前一页的第二纵坐标与所述后一页的第二纵坐标的和，将计算出的第二和值与所述后一页的第一纵坐标做差得到第二差值，更新所述后一页的第二纵坐标为所述第二差值。

9.一种PDF文件的信息抽取装置，其特征在于，包括：

获取标记模块，用于从PDF文件中获取文本对象的位置信息，将所述位置信息标记在图像上；其中，所述文本对象包括至少一个键名和对应的键值；

分类确定模块，用于根据所述图像的布局特征对所述图像进行分类，以基于图像类型确定所述键名和对应键值在所述PDF文件中的位置范围；

建立输出模块，用于按照所述键名的级别为所述键名之间建立关联关系，以结合所述键名和对应键值的位置范围，输出不同级别的所述键名和对应的键值。

10.一种电子设备，其特征在于，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-8中任一所述的方法。

11.一种计算机可读介质，其上存储有计算机程序，其特征在于，所述程序被处理器执行时实现如权利要求1-8中任一所述的方法。

一种PDF文件的信息抽取方法和装置

技术领域

[0001] 本发明涉及计算机领域，尤其涉及一种PDF文件的信息抽取方法和装置。

背景技术

[0002] 为了方便用户从PDF文件中获取感兴趣的内容，需要对PDF文件的内容进行结构化处理，识别出每个标题对应的父子标题，内容片断，图表内容等信息，并将其有序组织起来。现有技术中，针对PDF文件的信息抽取，主要是通过工具包抽取纯文本和纯表格。抽取纯文本是指从整个PDF文件中抽取出所有的文本信息，抽取纯表格是指从整个PDF文件中抽取出与表格相关的文本信息。

[0003] 在实现本发明过程中，发明人发现现有技术中至少存在如下问题：

[0004] 无法确定各级标题的归属关系，标题与相应内容片断的位置对应关系，表格与相关文本的位置对应关系；无法过滤目录、注释等干扰信息。

发明内容

[0005] 有鉴于此，本发明实施例提供一种PDF文件的信息抽取方法和装置，通过将PDF文件中文本对象的位置标记在图像上，并根据布局特征对图像分类后，按照图像类型确定键名、相应键值的位置，建立各级键名之间的关联关系，进而结合位置和关联关系结构化输出键名和对应键值，提高了信息抽取性能。

[0006] 为实现上述目的，根据本发明实施例的一个方面，提供了一种PDF文件的信息抽取方法。

[0007] 本发明实施例的一种PDF文件的信息抽取方法，包括：从PDF文件中获取文本对象的位置信息，将所述位置信息标记在图像上；其中，所述文本对象包括至少一个键名和对应的键值；根据所述图像的布局特征对所述图像进行分类，以基于图像类型确定所述键名和对应键值在所述PDF文件中的位置范围；按照所述键名的级别为所述键名之间建立关联关系，以结合所述键名和对应键值的位置范围，输出不同级别的所述键名和对应的键值。

[0008] 可选地，所述将所述位置信息标记在图像上，包括：根据多个所述文本对象之间横坐标的异同，以及纵坐标间隔与预设第一阈值的差值，判断多个所述文本对象是否支持被抽象为归属于同一条线段的点；若多个所述文本对象支持被抽象为归属于同一条线段的点，则分别获取多个所述文本对象的横坐标的最值和纵坐标的最值，将所述最值对应的线段显示在至少一幅图像上。

[0009] 可选地，对于左右布局的图像类型，所述确定所述键名和对应键值在所述PDF文件中的位置范围，包括：以字符为基本单元，将所述PDF文件的原始信息拆分为至少一个元素，将横坐标相同、纵坐标差值在预设第二阈值的元素组合成元素集；按照所述元素集的横坐标对所述元素集进行排序，以确定边界横坐标和位于所述边界横坐标之间的分界横坐标，将位于左边界横坐标和所述分界横坐标之间，且行间距大于预设第三阈值的元素集作为所述键名，确定所述键名在所述PDF文件的坐标区间；根据相邻两个所述键名的坐标区间，确

定与其中一个所述键名对应的键值在所述PDF文件的坐标区间。

[0010] 可选地，对于上下布局的图像类型，所述确定所述键名和对应键值在所述PDF文件中的位置范围，包括：以字符为基本单元，将所述PDF文件的原始信息拆分为至少一个元素，将位于同一行的元素组合成元素集；将起始横坐标位于所述PDF文件左侧，行间距大于预设第五阈值，和/或以冒号字符结尾的元素集作为所述键名，确定所述键名在所述PDF文件的坐标区间；根据相邻两个所述键名的坐标区间，确定与其中一个所述键名对应的键值在所述PDF文件的坐标区间。

[0011] 可选地，按照所述键名的级别为所述键名之间建立关联关系，包括：将同级的键名并联，将上下级的键名串联，采用树形结构为所述键名之间建立关联关系；所述输出不同级别的所述键名和对应的键值，包括：采用先序遍历所述树形结构的方式，顺序输出不同级别的所述键名和对应的键值。

[0012] 可选地，所述方法还包括：提取原始PDF文件的设定参考对象的基本信息，以根据所述基本信息确定所述原始PDF文件的非正文区域；过滤所述原始PDF文件的非正文区域，将过滤结果作为所述PDF文件。

[0013] 可选地，所述参考对象包括下列任意一项或者多项：边框、线段、图片和文本，所述基本信息包括下列任意一项或者多项：字体、线段粗细、高度、宽度、横纵坐标和文本内容；所述非正文区域包括：目录、表格和注释；所述根据所述基本信息确定所述原始PDF文件的非正文区域，包括：根据所述目录的特征确定所述目录的提取维度，按照所述提取维度确定所述目录的上下边界；获取所述原始PDF文件的最小字体，将首字符的字体等于最小字体，且位于所述原始PDF文件底部的区域作为所述注释的上下边界；确定所述表格的边界线坐标，以及所述表格的单元格，以得到所述表格的形状和上下边界。

[0014] 可选地，在所述PDF文件包括多页的情况下，所述方法还包括：将相邻页面的后一页的第一横纵坐标更新为前一页的第二横纵坐标；计算所述前一页的第二横坐标与所述后一页的第二横坐标的和，将计算出的第一和值与所述后一页的第一横坐标做差得到第一差值，更新所述后一页的第二横纵坐标为所述第一差值；计算所述前一页的第二纵坐标与所述后一页的第二纵坐标的和，将计算出的第二和值与所述后一页的第一纵坐标做差得到第二差值，更新所述后一页的第二纵坐标为所述第二差值。

[0015] 为实现上述目的，根据本发明实施例的另一方面，提供了一种PDF文件的信息抽取装置。

[0016] 本发明实施例的一种PDF文件的信息抽取装置，包括：获取标记模块，用于从PDF文件中获取文本对象的位置信息，将所述位置信息标记在图像上；其中，所述文本对象包括至少一个键名和对应的键值；分类确定模块，用于根据所述图像的布局特征对所述图像进行分类，以基于图像类型确定所述键名和对应键值在所述PDF文件中的位置范围；建立输出模块，用于按照所述键名的级别为所述键名之间建立关联关系，以结合所述键名和对应键值的位置范围，输出不同级别的所述键名和对应的键值。

[0017] 可选地，所述获取标记模块，还用于：根据多个所述文本对象之间横坐标的异同，以及纵坐标间隔与预设第一阈值的差值，判断多个所述文本对象是否支持被抽象为归属于同一条线段的点；以及若多个所述文本对象支持被抽象为归属于同一条线段的点，则分别获取多个所述文本对象的横坐标的最值和纵坐标的最值，将所述最值对应的线段显示在至

少一幅图像上。

[0018] 可选地，对于左右布局的图像类型，所述分类确定模块，还用于：以字符为基本单元，将所述PDF文件的原始信息拆分为至少一个元素，将横坐标相同、纵坐标差值在预设第二阈值的元素组合成元素集；按照所述元素集的横坐标对所述元素集进行排序，以确定边界横坐标和位于所述边界横坐标之间的分界横坐标，将位于左边界横坐标和所述分界横坐标之间，且行间距大于预设第三阈值的元素集作为所述键名，确定所述键名在所述PDF文件的坐标区间；以及根据相邻两个所述键名的坐标区间，确定与其中一个所述键名对应的键值在所述PDF文件的坐标区间。

[0019] 可选地，对于上下布局的图像类型，所述确定所述键名和对应键值在所述PDF文件中的位置范围，所述分类确定模块，还用于：以字符为基本单元，将所述PDF文件的原始信息拆分为至少一个元素，将位于同一行的元素组合成元素集；将起始横坐标位于所述PDF文件左侧，行间距大于预设第五阈值，和/或以冒号字符结尾的元素集作为所述键名，确定所述键名在所述PDF文件的坐标区间；以及根据相邻两个所述键名的坐标区间，确定与其中一个所述键名对应的键值在所述PDF文件的坐标区间。

[0020] 可选地，所述建立输出模块，还用于：将同级的键名并联，将上下级的键名串联，采用树形结构为所述键名之间建立关联关系；以及采用先序遍历所述树形结构的方式，顺序输出不同级别的所述键名和对应的键值。

[0021] 可选地，所述装置还包括：提取过滤模块，用于提取原始PDF文件的设定参考对象的基本信息，以根据所述基本信息确定所述原始PDF文件的非正文区域；以及过滤所述原始PDF文件的非正文区域，将过滤结果作为所述PDF文件。

[0022] 可选地，在所述PDF文件包括多页的情况下，所述装置还包括：坐标更新模块，用于：将相邻页面的后一页的第一横纵坐标更新为前一页的第二横纵坐标；计算所述前一页的第二横坐标与所述后一页的第二横坐标的和，将计算出的第一和值与所述后一页的第一横坐标做差得到第一差值，更新所述后一页的第二横纵坐标为所述第一差值；以及计算所述前一页的第二纵坐标与所述后一页的第二纵坐标的和，将计算出的第二和值与所述后一页的第一纵坐标做差得到第二差值，更新所述后一页的第二纵坐标为所述第二差值。

[0023] 为实现上述目的，根据本发明实施例的再一方面，提供了一种电子设备。

[0024] 本发明实施例的一种电子设备，包括：一个或多个处理器；存储装置，用于存储一个或多个程序，当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现本发明实施例的一种PDF文件的信息抽取方法。

[0025] 为实现上述目的，根据本发明实施例的再一方面，提供了一种计算机可读介质。

[0026] 本发明实施例的一种计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现本发明实施例的一种PDF文件的信息抽取方法。

[0027] 上述发明中的一个实施例具有如下优点或有益效果：通过将PDF文件中文本对象的位置标记在图像上，并根据布局特征对图像分类后，按照图像类型确定键名、相应键值的位置，建立各级键名之间的关联关系，进而结合位置和关联关系结构化输出键名和对应键值，提高了信息抽取性能，同时实现了PDF文件信息的自动化抽取，降低人工成本；将文本对象抽象为点，并将能够归属于同一线段的点连接成线段后显示在图像上，使得可以将文本对象的位置标记在图像中；对于不同图像类型，基于其各自的特点确定键名和键值所在的

坐标区间，进一步提高了信息抽取性能；采用树形结构和先序遍历的方式，实现了各级键名和对应的键值的结构化存储和顺序输出；基于非正文区域的特点，确定非正文区域在原始PDF文件中的位置后进行过滤，合理的去除了干扰信息；采用递归算法修改PDF文件的坐标，将所有页面的PDF文本拉伸到同一个页面内进行处理，可以一次性获得整个PDF文件的抽取信息，抽取效率高。

[0028] 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具体实施方式加以说明。

附图说明

[0029] 附图用于更好地理解本发明，不构成对本发明的不当限定。其中：

[0030] 图1是根据本发明实施例的PDF文件的信息抽取方法的主要步骤的示意图；

[0031] 图2是根据本发明实施例的PDF文件的信息抽取方法的主要流程示意图；

[0032] 图3是本发明实施例的PDF文件的目录示意图；

[0033] 图4是本发明实施例的PDF文件的信息抽取方法确定出的表格形状；

[0034] 图5是本发明实施例的PDF文件的信息抽取方法的左右布局的图像样本示意图；

[0035] 图6是本发明实施例的PDF文件的信息抽取方法的上下布局的图像样本示意图；

[0036] 图7是本发明实施例中的树形结构示意图；

[0037] 图8是根据本发明实施例的PDF文件的信息抽取装置的主要模块的示意图；

[0038] 图9是本发明实施例可以应用于其中的示例性系统架构图；

[0039] 图10是适用于来实现本发明实施例的电子设备的计算机装置的结构示意图。

具体实施方式

[0040] 以下结合附图对本发明的示范性实施例做出说明，其中包括本发明实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本发明的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0041] 图1是根据本发明实施例的PDF文件的信息抽取方法的主要步骤的示意图。如图1所示，本发明实施例的PDF文件的信息抽取方法，主要包括如下步骤：

[0042] 步骤S101：从PDF文件中获取文本对象的位置信息，将所述位置信息标记在图像上；其中，所述文本对象包括至少一个键名和对应的键值。提取PDF文件的文本对象的坐标，根据多个文本对象的横坐标的异同，以及纵坐标间隔与预设第一阈值的差值，判断其是否支持被抽象为归属于同一条线段的点；若多个文本对象支持被抽象为归属于同一条线段的点，则获取多个文本对象的横坐标的最小/最大值和纵坐标的最小/最大值，将最小/最大值对应的线段显示在至少一幅图像上，即可将文本对象的位置标记在图像上。

[0043] 步骤S102：根据所述图像的布局特征对所述图像进行分类，以基于图像类型确定所述键名和对应键值在所述PDF文件中的位置范围。采用深度学习方法，根据图像的布局特征将其分类为左右布局或者上下布局的图像类型。键名为PDF文件中类似标题的文本，键值为标题对应的内容片断。不同的图像类型对应不同的键名、键值界定组合方式。对于左右布局的图像类型，一般键名的起始横坐标位于文件左侧，以数字符号开始，长度较短，不会超

过文件中间位置，键名之间有较大的坐标间隔。对于上下布局的图像类型，一般键名的起始横坐标位于文件左侧，以数字符号开始，且结尾没有冒号以外的其它符号。键值一般位于相邻两个键名的坐标之间。故可根据不同图像类型的图像中键名和键值的分布特点，确定键名和对应键值在PDF文件的坐标区间。

[0044] 步骤S103：按照所述键名的级别为所述键名之间建立关联关系，以结合所述键名和对应键值的位置范围，输出不同级别的所述键名和对应的键值。在确定键名和对应键值在PDF文件的坐标区间后，可以将同级的键名并联，将上下级的键名串联，采用树形结构为多个键名之间建立关联关系；之后可以采用先序遍历该树形结构的方式，顺序输出不同级别的键名和对应的键值。通过上述步骤实现了对PDF文件的自动信息抽取，且输出的键名和键值相对应，降低了人工参与成本，提高了信息抽取性能。

[0045] 图2是根据本发明实施例的PDF文件的信息抽取方法的主要流程示意图。如图2所示，本发明实施例的PDF文件的信息抽取方法，主要包括如下步骤：

[0046] 步骤S201：提取原始PDF文件的设定参考对象的基本信息，以根据基本信息确定原始PDF文件的非正文区域。使用信息抽取工具从原始PDF文件中提取设定参考对象的基本信息，其中，参考对象是对于自动化解析PDF文件、结构化抽取信息有价值的对象，比如可以为边框、线段、图片和文本中的一种或者多种；基本信息可以为字体、线段粗细、高度、宽度、横纵坐标和文本内容中的一种或者多种。实施例中，信息抽取工具可以为Java编程语言中的Apache(网页服务器软件)，或者Python编程语言中的Pdfminer(是一个旨在帮助提取或者分析PDF文件的文本数据套件)。

[0047] 非正文区域即原始PDF文件中除去正文文本之外的区域，包括目录、注释、表格、页面页脚等。确定原始PDF文件中非正文区域相应包括确定目录、注释、表格、页面页脚等在原始PDF文件中的位置。下面分别进行说明。

[0048] (1)确定目录在原始PDF文件中的位置：

[0049] 根据目录的特征确定目录的提取维度，按照提取维度确定目录的上下边界。其中，提取维度可以包括：符号标识符，比如省略号、短划线；目录的边框位置；目录中每行文本信息的特征，包括数字频率、数字位置、数字文本的组合关系；目录起始和结束的大标题特征，比如字体类型、字体大小、大标题在页面的位置。目录的边框位置是指包围目录的边框的四个顶点的坐标。数字频率是指目录中每行的数字单元个数(包括中文数字)。数字文本的组合关系是指目录中数字和文本的对应关系，比如图3中所示的目录格式，该组合关系即目录中每行是否数字在最左边、数字后面有至少一个汉字。

[0050] 图3是本发明实施例的PDF文件的目录示意图。如图3所示，确定该目录在PDF文件中的位置时，提取维度可以是：目录的边框位置、目录中每行文本信息的特征，以及目录起始和结束的大标题特征。从以上提取维度即可以确定目录的上下边界。

[0051] (2)确定注释在原始PDF文件中的位置：

[0052] 注释一般是整个原始PDF文件中字符最小且位于底部的文本，因此在确定其位置时，需要获取原始PDF文件的最小字体，首字符的字体等于最小字体，且位于原始PDF文件底部的区域即为注释的上下边界。其中，首字符是指每行的第一个字符。

[0053] (3)确定表格在原始PDF文件中的位置：

[0054] 确定表格的边界线坐标，以及表格的各个单元格，以得到表格的形状和上下边界。

在确定表格的边界线(即表格外边框的四条线段)坐标时,需要利用聚类算法将线段端点相近的线段归为一组;之后将每组线段的坐标分别进行归一化处理,将非水平、竖直线段处理为水平、竖直线段;之后利用递归算法,走格思想,从某个点开始(比如最低点),上下左右4个方向进行移动,到达方向端点则90度转向,继续移动,如果可以走回原点,则证明为矩形,保留矩形顶点,按照上述方式反复抽取矩形顶点。其中,线段的坐标是线段两个端点的横纵坐标,比如[左端点坐标[1,2],右端点坐标[5,6]]。

[0055] 实施例中,采用KNN聚类算法对表格的线段进行聚类。此时,KNN算法的输入为:所有线段的端点坐标;处理过程:遍历所有点坐标,若两个点的横纵坐标的差值的绝对值在预设阈值(比如1)内,则将这两个点归为一组;遍历完毕后,取每组坐标的均值作为该组的中心点;将所有点与中心点进行比对,重复上述处理过程,直至满足聚类终止条件。KNN算法的输出为:分组后的所有端点坐标。

[0056] 下面对将非水平、竖直线段处理为水平、竖直线段进行说明:实施例中的非水平、竖直线段是指由文档提取工具不完美,所造成的略微倾斜的线段,其本质是水平、竖直线段。如果线段的两个纵坐标相同说明线段水平,如果线段的两个横坐标相同说明线段竖直。比如,坐标为[(1,2),(1,6)]的线段是竖直线段,坐标为[(2,4),(8,4)]的线段是水平线段。在聚类分组后,将同一组的端点坐标归一化为中心点坐标,即可将非水平、竖直线段处理为水平、竖直线段。比如,将[(1.1,2.2),(1.2,2.3),(1.1,2.4)]归一化为(1,2)。

[0057] 另外,在递归算法中,在上下左右4个方向进行移动之前,需先将所有点按照横纵坐标的大小进行排序。往上下方向移动,即找横坐标与当前点横坐标相同的点;往左右方向移动即找纵坐标与当前点纵坐标相同的点。

[0058] 图4是本发明实施例的PDF文件的信息抽取方法确定出的表格形状。由图4可知,通过本发明实施例的聚类算法、归一化、递归算法等对PDF文件的表格进行处理后,可以完美的确定PDF文件的表格形状。之后可以过滤表格内的内容,区分正文信息与表格文本信息。

[0059] 步骤S202:过滤原始PDF文件的非正文区域,得到PDF文件。确定目录、注释、表格、页面页脚等的上下边界后,即可精准定位非正文区域在原始PDF文件的位置,精准区分非正文区域和正文区域,过滤非正文区域的内容,去除干扰信息。

[0060] 步骤S203:从PDF文件中获取文本对象的位置信息,将位置信息标记在图像上。该步骤中首先提取PDF文件的文本对象、空格、标点符号的坐标。由于每个坐标有4个值:横坐标的最小/最大值和纵坐标的最大/最小值,实施例中统一取一组即可。比如,用[横坐标最小值,纵坐标最小值]作为文本对象、空格、标点符号的坐标。

[0061] 之后,对于文本对象、空格、标点符号,根据其横坐标的异同,以及纵坐标间隔与预设第一阈值的差值,判断其是否能够被抽象为归属于同一条线段的点。其中,第一阈值比如为10mm。

[0062] 最后,对于能够被抽象为归属于同一条线段的点对应的文本对象、空格、标点符号,获取其横坐标的最小/最大值和纵坐标的最小/最大值,将最小/最大值对应的线段显示在至少一幅图像上,即可将文本对象的位置标记在图像上。

[0063] 步骤S204:根据图像的布局特征对图像进行分类,以基于图像类型确定键名和对应键值在PDF文件中的位置范围。实施例中使用深度学习方法对图像进行分类,比如使用神经网络卷积模型CNN。使用图像提取工具获取图像中的像素点,之后通过卷积、池化处理,可

以从中概括提取图像的形状。CNN模型预先经有监督的训练，可以学会不同的图像形状，以此达到识别图像类型的效果。CNN模型的输出结果就是图像类型为左右布局或者上下布局。

[0064] 以保险领域的PDF文件为例，其合同、条款包含诸多的(名词，名词解释)、(条款，条款说明)、(短语，短语释义)，即(问题，答案)形式的PDF文件。比如，问题：投保年龄，答案：指您投保时被保险人的年龄，投保年龄以周岁计算。则此PDF文件的键名可以是：名词、条款、短语，相应的键值为：名词解释、条款说明、短语释义。

[0065] 图5是本发明实施例的PDF文件的信息抽取方法的左右布局的图像样本示意图。图6是本发明实施例的PDF文件的信息抽取方法的上下布局的图像样本示意图。如图5和图6所示，图5的图像样本具有明显的左右布局特征，图6的图像样本具有明显的上下布局特征，将上述图像样本分别输入CNN模型，经CNN模型处理后，即可输出图5的图像样本为左右布局的图像类型，图6的图像样本为上下布局的图像类型这一分类结果。

[0066] 不同的图像类型对应不同的键名、键值界定组合方式。对于左右布局的图像类型，在确定键名和对应键值在PDF文件中的位置范围时，首先以字符为基本单元，将PDF文件的原始信息拆分为至少一个元素，将纵坐标相同、横坐标差值小于预设第二阈值的元素组合成元素集；之后按照元素集的横坐标对元素集进行排序，以确定边界横坐标和位于左右边界横坐标之间的分界横坐标；假设键名的起始横坐标都在分界横坐标的左侧，键值都在分界横坐标的右侧，则之后将位于左边界横坐标和分界横坐标之间，且行间距大于预设第三阈值的元素集作为键名，确定键名在PDF文件的坐标区间；最后根据相邻两个键名的坐标区间，确定与前一个键名对应的键值在PDF文件的坐标区间。实施例中，预设第二阈值比如为10mm。将PDF文件的原始信息拆分为单个元素的目的是基于字符间距和字符位置，将各元素重新组合成一行行独立的文本(即重新组合成元素集)。

[0067] 在一优选的实施例中，同一键名可能被分为多行，此时需将行间距小于预设阈值的文本视为同一键名，进行文本组合。对于起始没有数字符号标记，但根据行间距可以独立形成文本单元的文本信息，也可视为键名。

[0068] 对于上下布局的图像类型，一般将起始横坐标位于整个文件左侧，以数字符号开始，且结尾没有冒号以外的其它符号的文本视为键名；还可以将起始横坐标位于整个文件左侧，与上下句之间行间距高于一定阈值，以冒号结尾的文本视为键名。在确定键名和对应键值在PDF文件中的位置范围时，首先以字符为基本单元，将PDF文件的原始信息拆分为至少一个元素，将位于同一行的元素组合成元素集；之后将起始横坐标位于PDF文件左侧，行间距大于预设第五阈值，且以冒号字符结尾的元素集作为键名，确定键名在PDF文件的坐标区间；最后根据相邻两个键名的坐标区间，确定与前一个键名对应的键值在PDF文件的坐标区间。实施例中，将纵坐标相同、横坐标差值在预设第四阈值的元素视为同一行的元素。

[0069] 在一优选的实施例中，在得到键名以及其对应的键值后，遍历键名、键值，先根据横坐标进行排序，之后再将同一行的文本根据纵坐标进行排序，组成有序的文本信息。获取键名和对应键值的位置后，后续以键名和键值组成的键值对为基本数据单元，将不同级别的键值对串联为父子级关系。

[0070] 步骤S205：按照键名的级别为多个键名之间建立关联关系，以结合键名和对应键值的位置范围，输出不同级别的键名和对应的键值。将同级的键名并联，将上下级的键名串联，采用树形结构为多个键名之间建立关联关系；之后采用先序遍历树形结构的方式，顺序

输出不同级别的键名和对应的键值。其中，先序遍历是指按照根左右的顺序沿一定路径经过路径上所有的节点。通过树形结构存储键名、采用先序遍历算法顺序输出当前级键名、对应的键值以及所述由上级键名，自动实现了对PDF文件中键名和键值相匹配的信息抽取。

[0071] 图7是本发明实施例中的树形结构示意图。如图7所示，树形结构包括根节点和三级子节点。根节点为PDF文档，第一级子节点为所有的一级键名，第二级子节点为隶属于对应一级键名的二级键名，第三级为隶属于对应二级键名的三级键名。

[0072] 在一优选的实施例中，在步骤S201之前，还可以根据每页的有效坐标范围，逐页累加坐标，进而将多页PDF文件拉伸到同一页面。具体实现为：将相邻页面的后一页的第一横纵坐标更新为前一页的第二横纵坐标；计算前一页的第二横坐标与后一页的第二横坐标的和，将计算出的第一和值与后一页的第一横坐标做差得到第一差值，更新后一页的第二横纵坐标为第一差值；计算前一页的第二纵坐标与后一页的第二纵坐标的和，将计算出的第二和值与后一页的第一纵坐标做差得到第二差值，更新后一页的第二纵坐标为第二差值。

[0073] 假设相邻两页的前一页的有效坐标范围为：[(x1,y1),(x11,y11)]，下一页的有效坐标范围为：[(x2,y2),(x21,y21)]，则按照上述方式处理后，下一页的坐标范围变更为：[(x11,y11),(x11+x21-x2,

[0074] y11+y21-y2)]。比如3页的PDF文件，第1至第3页的有效坐标范围为：[(100,50),(500,700)][(100,50),(500,600)][(100,50),(500,600)]，按照上述方式处理后，坐标范围变更为：[(100,50),(500,700)][(500,700),(900,1150)][(900,1150),(1300,1700)]。

[0075] 经过测试，对于保险领域的PDF文件中，使用本发明实施例的信息抽取得到可以达到96％的完整信息提取率，98％的信息降噪率。

[0076] 通过本发明实施例的PDF文件的信息抽取方法可以看出，通过将PDF文件中文本对象的位置标记在图像上，并根据布局特征对图像分类后，按照图像类型确定键名、相应键值的位置，建立各级键名之间的关联关系，进而结合位置和关联关系结构化输出键名和对应键值，提高了信息抽取性能，同时实现了PDF文件信息的自动化抽取，降低人工成本；将文本对象抽象为点，并将能够归属于同一线段的点连接成线段后显示在图像上，使得可以将文本对象的位置标记在图像中；对于不同图像类型，基于其各自的特点确定键名和键值所在的坐标区间，进一步提高了信息抽取性能；采用树形结构和先序遍历的方式，实现了各级键名和对应的键值的结构化存储和顺序输出；基于非正文区域的特点，确定非正文区域在原始PDF文件中的位置后进行过滤，合理的去除了干扰信息；采用递归算法修改PDF文件的坐标，将所有页面的PDF文本拉伸到同一个页面内进行处理，可以一次性获得整个PDF文件的抽取信息，抽取效率高。

[0077] 图8是根据本发明实施例的PDF文件的信息抽取装置的主要模块的示意图。如图8所示，本发明实施例的PDF文件的信息抽取装置800，主要包括：

[0078] 获取标记模块801，用于从PDF文件中获取文本对象的位置信息，将所述位置信息标记在图像上；其中，所述文本对象包括至少一个键名和对应的键值。提取PDF文件的文本对象的坐标，根据多个文本对象的横坐标的异同，以及纵坐标间隔与预设第一阈值的差值，判断其是否支持被抽象为归属于同一条线段的点；若多个文本对象支持被抽象为归属于同一条线段的点，则获取多个文本对象的横坐标的最小/最大值和纵坐标的最小/最大值，将最小/最大值对应的线段显示在至少一幅图像上，即可将文本对象的位置标记在图像上。

[0079] 分类确定模块802，用于根据所述图像的布局特征对所述图像进行分类，以基于图像类型确定所述键名和对应键值在所述PDF文件中的位置范围。采用深度学习方法，根据图像的布局特征将其分类为左右布局或者上下布局的图像类型。键名为PDF文件中类似标题的文本，键值为标题对应的内容片断。不同的图像类型对应不同的键名、键值界定组合方式。对于左右布局的图像类型，一般键名的起始横坐标位于文件左侧，以数字符号开始，长度较短，不会超过文件中间位置，键名之间有较大的坐标间隔。对于上下布局的图像类型，一般键名的起始横坐标位于文件左侧，以数字符号开始，且结尾没有冒号以外的其它符号。键值一般位于相邻两个键名的坐标之间。故可根据不同图像类型的图像中键名和键值的分布特点，确定键名和对应键值在PDF文件的坐标区间。

[0080] 建立输出模块803，用于按照所述键名的级别为所述键名之间建立关联关系，以结合所述键名和对应键值的位置范围，输出不同级别的所述键名和对应的键值。在确定键名和对应键值在PDF文件的坐标区间后，可以将同级的键名并联，将上下级的键名串联，采用树形结构为多个键名之间建立关联关系；之后可以采用先序遍历该树形结构的方式，顺序输出不同级别的键名和对应的键值。通过上述步骤实现了对PDF文件的自动信息抽取，且输出的键名和键值相对应，降低了人工参与成本，提高了信息抽取性能。

[0081] 另外，本发明实施例的PDF文件的信息抽取装置800还可以包括：提取过滤模块和坐标更新模块(图8中未示出)。其中，提取过滤模块用于提取原始PDF文件的设定参考对象的基本信息，以根据所述基本信息确定所述原始PDF文件的非正文区域；以及过滤所述原始PDF文件的非正文区域，将过滤结果作为所述PDF文件。坐标更新模块，用于：在所述PDF文件包括多页的情况下，将相邻页面的后一页的第一横纵坐标更新为前一页的第二横纵坐标；计算所述前一页的第二横坐标与所述后一页的第二横坐标的和，将计算出的第一和值与所述后一页的第一横坐标做差得到第一差值，更新所述后一页的第二横纵坐标为所述第一差值；以及计算所述前一页的第二纵坐标与所述后一页的第二纵坐标的和，将计算出的第二和值与所述后一页的第一纵坐标做差得到第二差值，更新所述后一页的第二纵坐标为所述第二差值。

[0082] 从以上描述可以看出，通过将PDF文件中文本对象的位置标记在图像上，并根据布局特征对图像分类后，按照图像类型确定键名、相应键值的位置，建立各级键名之间的关联关系，进而结合位置和关联关系结构化输出键名和对应键值，提高了信息抽取性能，同时实现了PDF文件信息的自动化抽取，降低人工成本。

[0083] 图9示出了可以应用本发明实施例的PDF文件的信息抽取方法或PDF文件的信息抽取装置的示例性系统架构900。

[0084] 如图9所示，系统架构900可以包括终端设备901、902、903，网络904和服务器905。网络904用以在终端设备901、902、903和服务器905之间提供通信链路的介质。网络904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。

[0085] 用户可以使用终端设备901、902、903通过网络904与服务器905交互，以接收或发送消息等。终端设备901、902、903上可以安装有各种通讯客户端应用，例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客户端、社交平台软件等。

[0086] 终端设备901、902、903可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0087] 服务器905可以是提供各种服务的服务器，例如对外部输入或者存储的PDF文件进行处理的后台管理服务器。后台管理服务器可以对PDF文件进行文本对象获取、非正文区域过滤、PDF分类等处理，并将处理结果(例如结构化数据)反馈给终端设备。

[0088] 需要说明的是，本申请实施例所提供的PDF文件的信息抽取方法一般由终端设备901、902、903或者服务器905执行，相应地，PDF文件的信息抽取装置一般设置于终端设备901、902、903或者服务器905中。

[0089] 应该理解，图9中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

[0090] 根据本发明的实施例，本发明还提供了一种电子设备和一种计算机可读介质。

[0091] 本发明的电子设备包括：一个或多个处理器；存储装置，用于存储一个或多个程序，当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现本发明实施例的一种PDF文件的信息抽取方法。

[0092] 本发明的计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现本发明实施例的一种PDF文件的信息抽取方法。

[0093] 下面参考图10，其示出了适用于来实现本发明实施例的电子设备的计算机系统1000的结构示意图。图10示出的电子设备仅仅是一个示例，不应对本发明实施例的功能和使用范围带来任何限制。

[0094] 如图10所示，计算机系统1000包括中央处理单元(CPU)1001，其可以根据存储在只读存储器(ROM)1002中的程序或者从存储部分1008加载到随机访问存储器(RAM)1003中的程序而执行各种适当的动作和处理。在RAM 1003中，还存储有计算机系统1000操作所需的各种程序和数据。CPU 1001、ROM 1002以及RAM 1003通过总线1004彼此相连。输入/输出(I/O)接口1005也连接至总线1004。

[0095] 以下部件连接至I/O接口1005：包括键盘、鼠标等的输入部分1006；包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分1007；包括硬盘等的存储部分1008；以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分1009。通信部分1009经由诸如因特网的网络执行通信处理。驱动器1100也根据需要连接至I/O接口1005。可拆卸介质1101，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器1100上，以便于从其上读出的计算机程序根据需要被安装入存储部分1008。

[0096] 特别地，根据本发明公开的实施例，上文主要步骤图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行主要步骤图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分1009从网络上被下载和安装，和/或从可拆卸介质1101被安装。在该计算机程序被中央处理单元(CPU)1001执行时，执行本发明的系统中限定的上述功能。

[0097] 需要说明的是，本发明所示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。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储

器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本发明中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本发明中,计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输,包括但不限于:无线、电线、光缆、RF等等,或者上述的任意合适的组合。

[0098] 附图中的流程图和框图,图示了按照本发明各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0099] 描述于本发明实施例中所涉及到的模块可以通过软件的方式实现,也可以通过硬件的方式来实现。所描述的模块也可以设置在处理器中,例如,可以描述为:一种处理器包括获取标记模块、分类确定模块和建立输出模块。其中,这些模块的名称在某种情况下并不构成对该模块本身的限定,例如,获取标记模块还可以被描述为“从PDF文件中获取文本对象的位置信息,将所述位置信息标记在图像上的模块”。

[0100] 作为另一方面,本发明还提供了一种计算机可读介质,该计算机可读介质可以是上述实施例中描述的设备中所包含的;也可以是单独存在,而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序,当上述一个或者多个程序被一个该设备执行时,使得该设备包括:从PDF文件中获取文本对象的位置信息,将所述位置信息标记在图像上;其中,所述文本对象包括至少一个键名和对应的键值;根据所述图像的布局特征对所述图像进行分类,以基于图像类型确定所述键名和对应键值在所述PDF文件中的位置范围;按照所述键名的级别为所述键名之间建立关联关系,以结合所述键名和对应键值的位置范围,输出不同级别的所述键名和对应的键值。

[0101] 从以上描述可以看出,通过将PDF文件中文本对象的位置标记在图像上,并根据布局特征对图像分类后,按照图像类型确定键名、相应键值的位置,建立各级键名之间的关联关系,进而结合位置和关联关系结构化输出键名和对应键值,提高了信息抽取性能,同时实现了PDF文件信息的自动化抽取,降低人工成本。

[0102] 上述产品可执行本发明实施例所提供的方法,具备执行方法相应的功能模块和有益效果。未在本实施例中详尽描述的技术细节,可参见本发明实施例所提供的方法。

[0103] 上述具体实施方式,并不构成对本发明保护范围的限制。本领域技术人员应该明

白的是，取决于设计要求和其他因素，可以发生各种各样的修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等，均应包含在本发明保护范围之内。

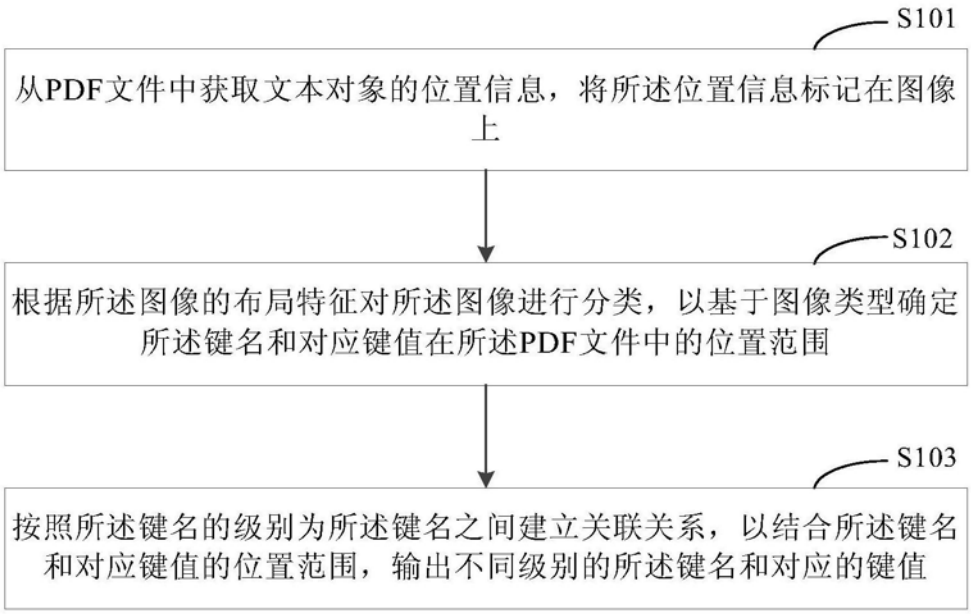

S101
从PDF文件中获取文本对象的位置信息，将所述位置信息标记在图像上
S102
根据所述图像的布局特征对所述图像进行分类，以基于图像类型确定所述键名和对应键值在所述PDF文件中的位置范围
S103
按照所述键名的级别为所述键名之间建立关联关系，以结合所述键名和对应键值的位置范围，输出不同级别的所述键名和对应的键值

图1

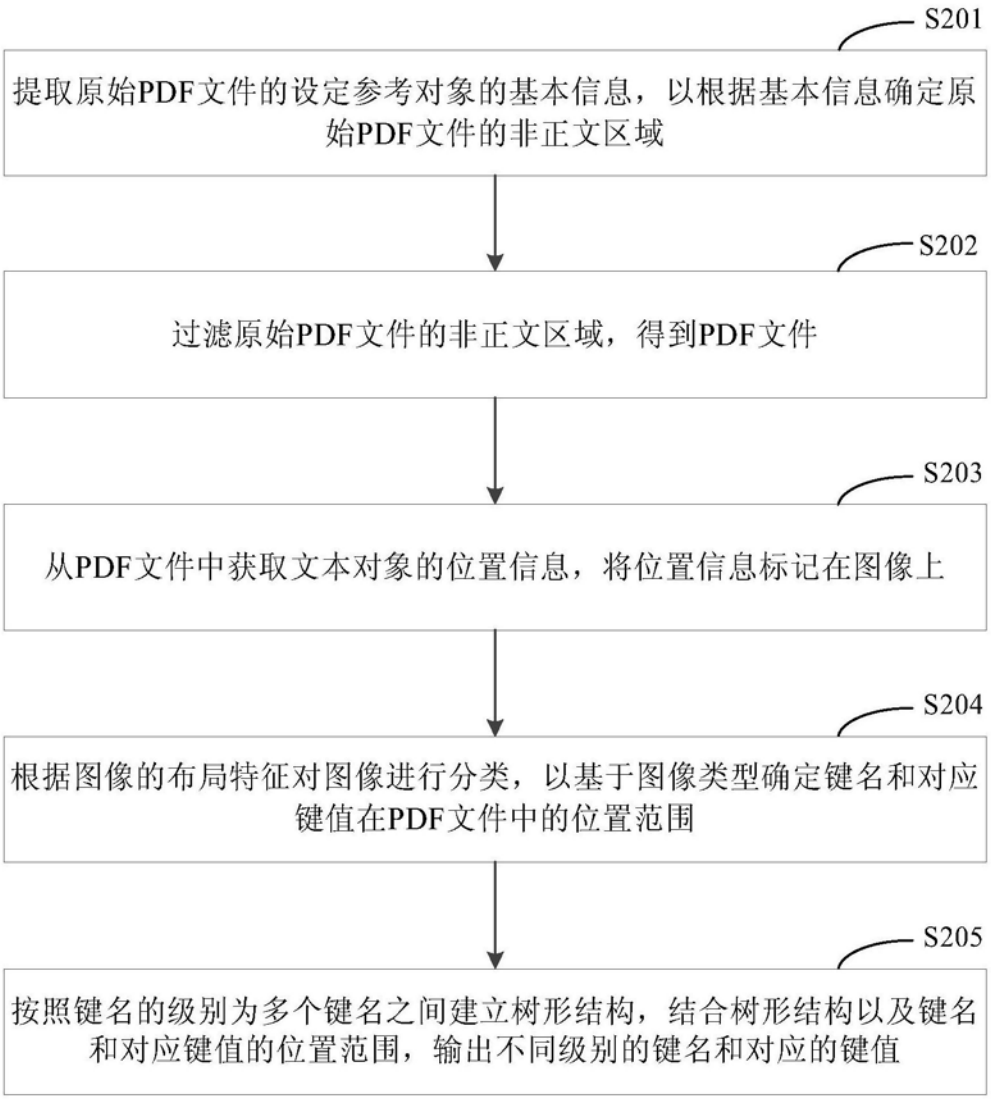

S201
提取原始PDF文件的设定参考对象的基本信息，以根据基本信息确定原始PDF文件的非正文区域
S202
过滤原始PDF文件的非正文区域，得到PDF文件
S203
从PDF文件中获取文本对象的位置信息，将位置信息标记在图像上
S204
根据图像的布局特征对图像进行分类，以基于图像类型确定键名和对应键值在PDF文件中的位置范围
S205
按照键名的级别为多个键名之间建立树形结构，结合树形结构以及键名和对应键值的位置范围，输出不同级别的键名和对应的键值

图2

目录

图3

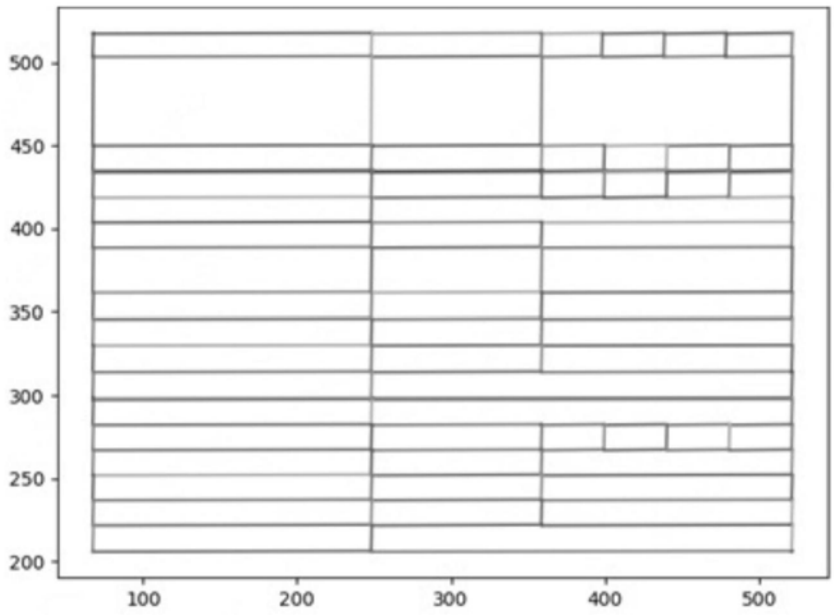


图4

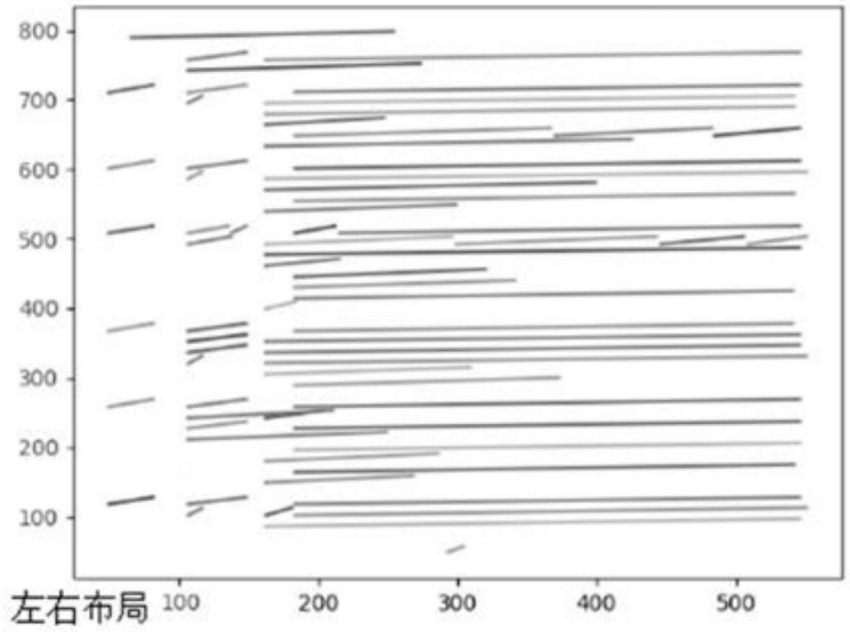

800
700
600
500
400
300
200
100
100
200
300
400
500
左右布局

图5

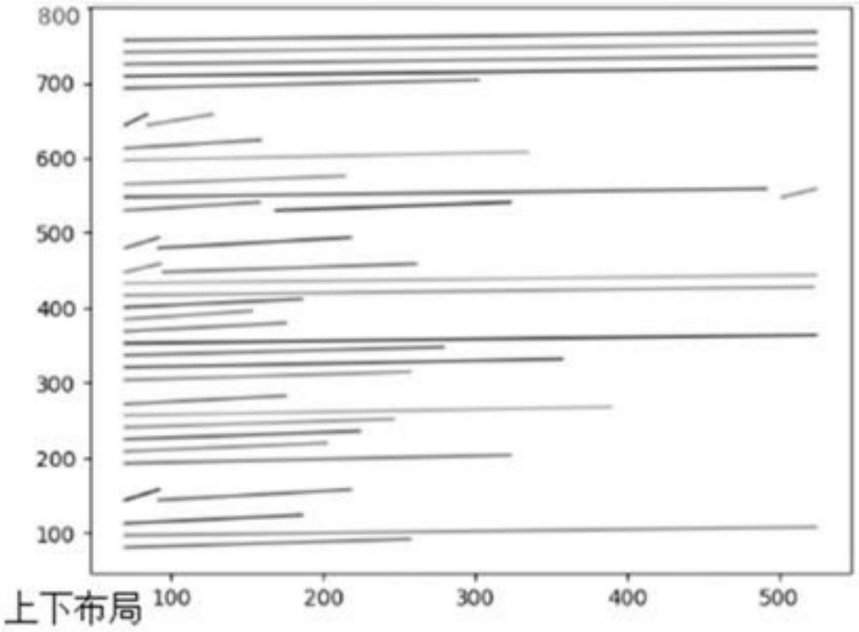

800
700
600
500
400
300
200
100
100
200
300
400
500
上下布局

图6

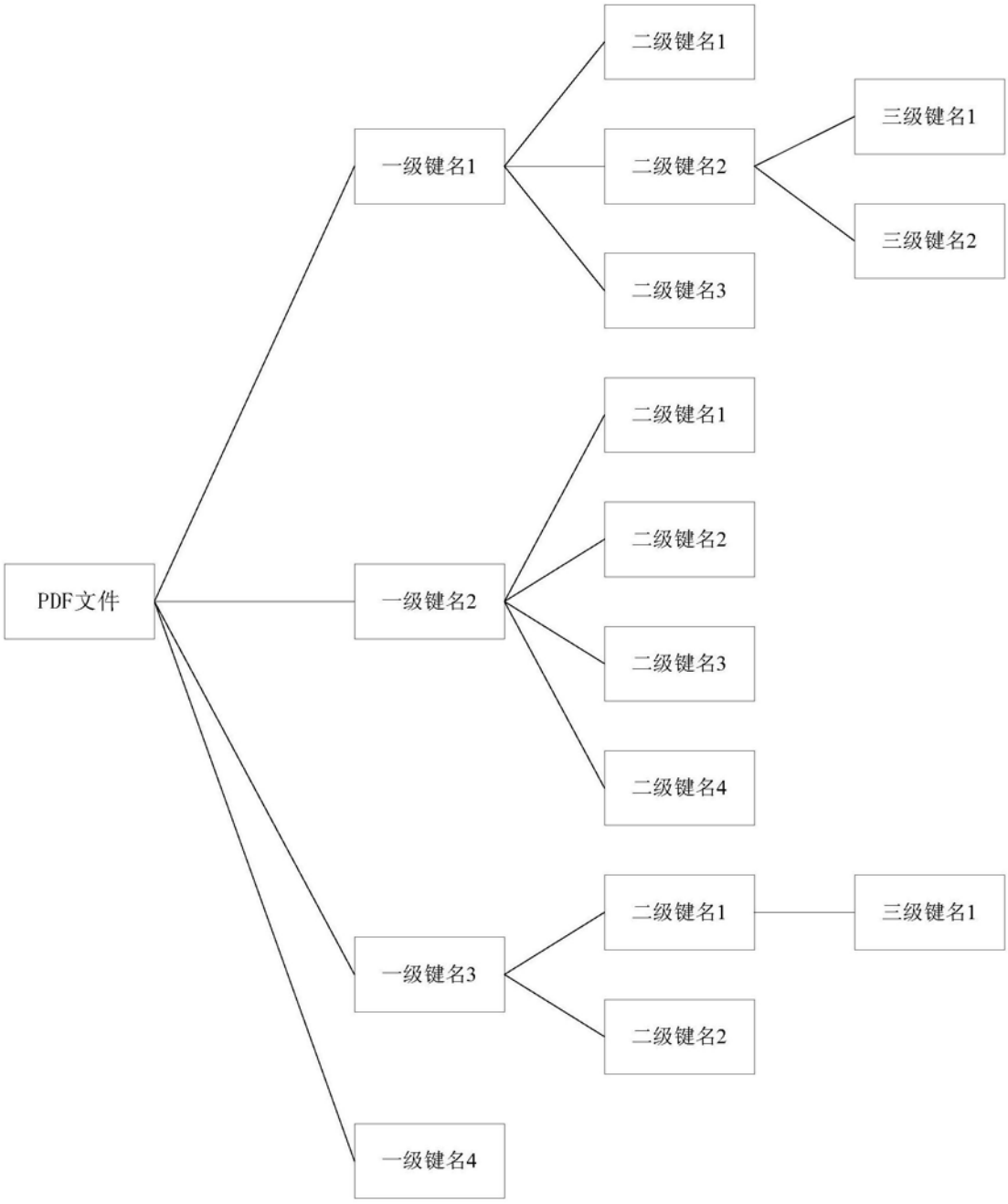

PDF文件
一级键名1
二级键名1
二级键名2
三级键名1
三级键名2
二级键名3
一级键名2
二级键名1
二级键名2
二级键名3
二级键名4
一级键名3
二级键名1
三级键名1
二级键名2
一级键名4

图7

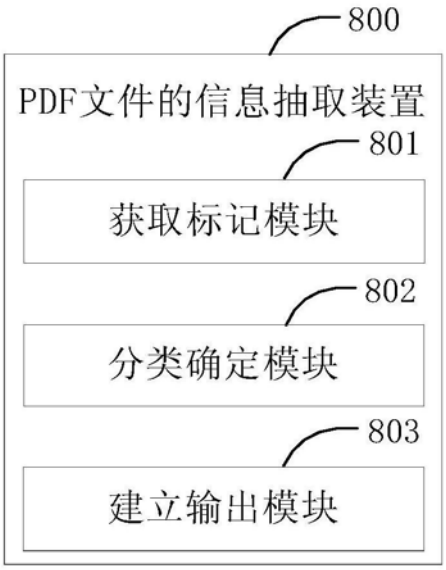

800
PDF文件的信息抽取装置
801
获取标记模块
802
分类确定模块
803
建立输出模块

图8

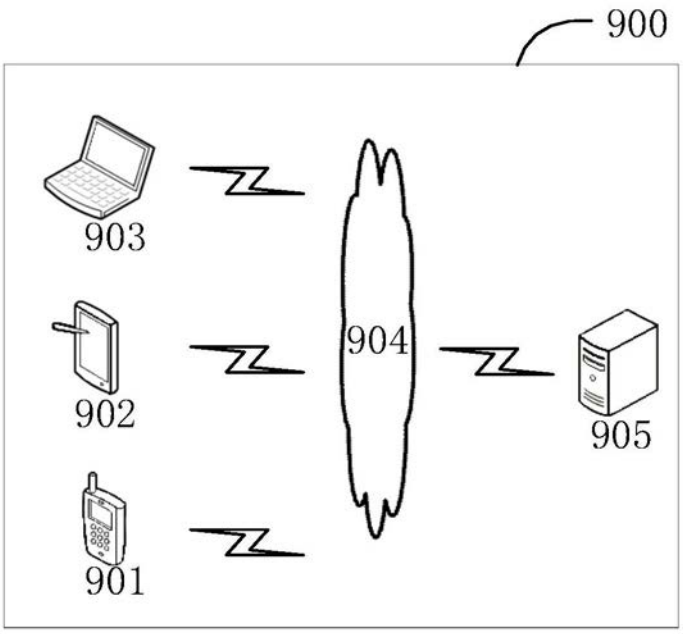

900
903
902
901
904
905

图9

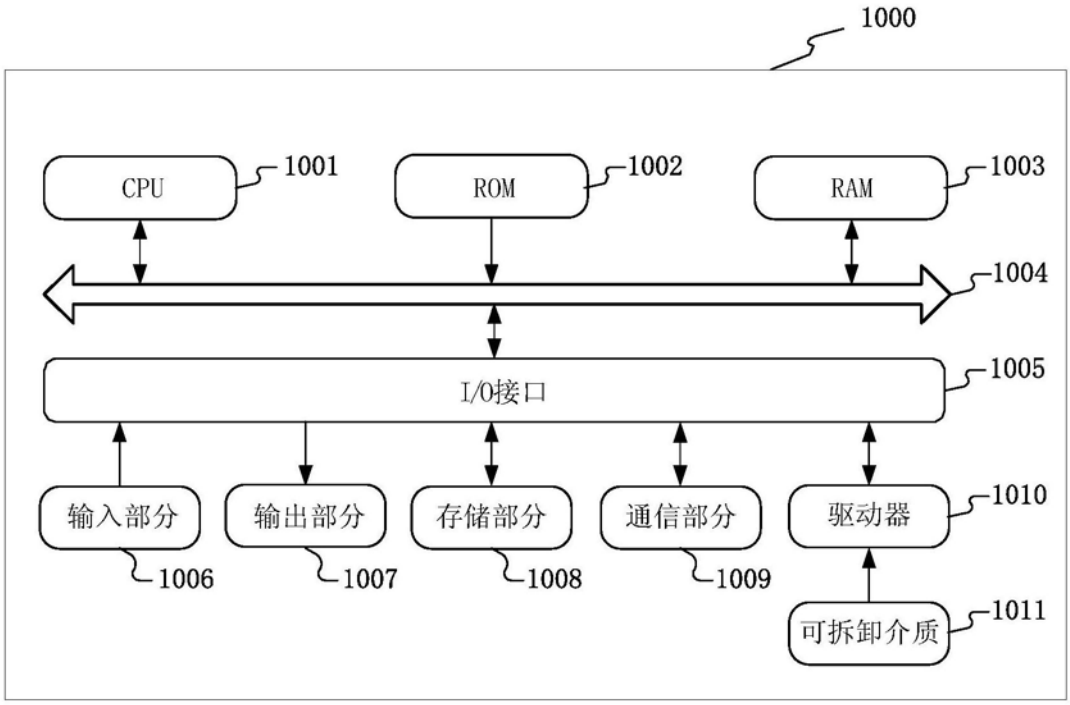

1000
CPU
1001
ROM
1002
RAM
1003
1004
I/O接口
1005
输入部分
1006
输出部分
1007
存储部分
1008
通信部分
1009
驱动器
1010
可拆卸介质
1011

图10