(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 106897690 A
(43)申请公布日 2017.06.27

(21)申请号 201710095978.4

(22)申请日 2017.02.22

(71)申请人 南京述酷信息技术有限公司
地址 210012 江苏省南京市雨花台区玉兰路88号南部创业园2栋305室

(72)发明人 郑龙 夏磊

(74)专利代理机构 上海领洋专利代理事务所（普通合伙） 31292
代理人 刘秋兰

(51)Int.Cl.
G06K 9/00(2006.01)

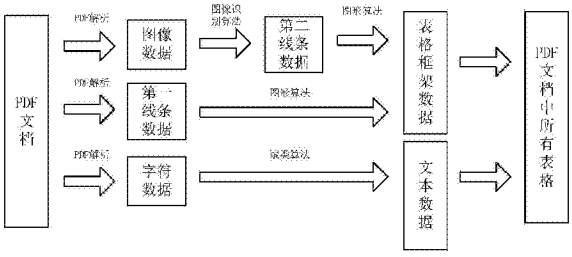
权利要求书4页 说明书11页 附图9页

(54)发明名称

PDF表格提取方法

(57)摘要

本发明技术方案公开了一种PDF表格提取方法，对PDF文档按页码进行解析，获取所有的图像数据、第一线条数据和字符数据，采用图像识别算法对图像数据按页码依次进行处理，从具有表格数据的图像数据中获得其表格数据对应的第二线条数据；采用图形算法对第一线条数据和第二线条数据按页码依次进行处理，获得具有表格行数据和列数据的表格框架数据；采用聚类算法对字符数据进行聚类处理，获得具有字符串集合的文本数据；经由最终所有表格框架和所有文本数据得到PDF文档中所有的表格数据。本发明对PDF文档中表格提取的方法提高了PDF文档中表格提取的准确率和效率，能得到更准确的表格数据，适用于对表格数据提取的准确率和效率要求更高的领域。



CN 106897690 A

1.一种PDF表格提取方法，其特征在于，该方法包括：

步骤A，对PDF文档进行解析，获取图像数据、第一线条数据和字符数据；

步骤B，采用图像识别算法对经由步骤A获取的图像数据进行处理，从具有表格数据的图像数据中获得其表格数据对应的第二线条数据；

步骤C，采用图形算法对经由步骤A获得的第一线条数据和经由步骤B获得的第二线条数据分别进行处理，获得具有表格行数据和列数据的表格框架数据；

步骤D，采用聚类算法对经由步骤A获得的字符数据进行聚类处理，获得具有字符串集合的文本数据；

步骤E，经由步骤C获得的表格框架数据中的表格行数据和列数据，得到对应的表格单元格，将表格单元格与步骤D获得的文本数据中的字符串集合相匹配，获得PDF文档中的表格数据。

2.如权利要求1所述的PDF表格提取方法，其特征在于，步骤C中所述图形算法处理包括对步骤A获得的第一线条数据和步骤B获得的第二线条数据对应的垂直线条进行垂直投影、水平线条进行水平投影、水平线条进行垂直投影及垂直线条进行水平投影，从而获得具有表格行数据和列数据的表格框架数据。

3.如权利要求1所述的PDF表格提取方法，其特征在于，步骤D中所述聚类算法处理包括对步骤A获得的字符数据中任两个连续字符数据的坐标数据依次进行垂直坐标Y阈值化处理、水平坐标X阈值化处理和/或垂直线条约束处理，将同类连续字符聚类到相应的字符串集合中，从而获得具有字符串集合的文本数据。

4.如权利要求1～3任一项所述的PDF表格提取方法，其特征在于，步骤A至步骤E中的各处理均是按页码依次进行处理；步骤A获取的图像数据、第一线条数据、字符数据及步骤B获取的第二线条数据均以页码为关联词存储到字典的PDF数据单元中，步骤C获得的表格框架数据以页码为关联词存储到字典的表格数据单元中，步骤D获得的文本数据以页码为关联词存储到字典的文本数据单元中。

5.如权利要求4所述的PDF表格提取方法，其特征在于，步骤C具体包括：

步骤C1，按页码依次从字典的PDF数据单元中获取当前页的第一线条数据和第二线条数据的垂直线条数据和水平线条数据；

步骤C2，对当前页的第一线条数据和第二线条数据的线条数据进行图形算法处理，获得当前页的表格数量及每个表格对应的上下边位置数据、左右边的位置数据、每行上下边位置数据和每列左右边的位置数据，即获得具有表格行数据和列数据的当前页的表格框架数据，并将其以页码为关联词存储到字典的表格数据单元中。

6.如权利要求5所述的PDF表格提取方法，其特征在于，所述步骤C2具体包括：

步骤C21，对当前页的第一线条数据和第二线条数据的垂直线条进行垂直投影，对所获得的垂直投影结果数据进行处理分析，获得当前页的表格数量及每个表格对应的上下边位置数据；

步骤C22，依次遍历当前页的每一个表格，对当前表格的第一线条数据或第二线条数据的水平线条进行水平投影，获得当前表格左右边的位置数据；对当前表格的第一线条数据或第二线条数据的水平线条进行垂直投影，获得当前表格每行上下边位置数据；对当前表格的第一线条数据或第二线条数据的垂直线条进行水平投影，获得当前表格每列左右边的

位置数据，从而获得具有表格行数据和列数据的当前表格的表格框架数据；

步骤C23，判断当前表格是否为当前页的最后一表格，如若当前表格不是当前页的最后一表格，则继续从步骤C22开始；如若当前表格是当前页的最后一表格，则以页码为关联词将具有表格行数据和列数据的当前页的表格框架数据存储到字典的表格数据单元中。

7.如权利要求6所述的PDF表格提取方法，其特征在于，还包括步骤C3：

步骤C3，判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则继续从步骤C1开始处理；如若当前页是最后一页，将存储有表格框架数据的字典存储到磁盘文件中。

8.如权利要求4所述的PDF表格提取方法，其特征在于，步骤D包括：

步骤D1，按页码依次从字典的PDF数据单元中获取当前页所有字符数据；

步骤D2，依次遍历当前页的两个连续字符数据，对获取的当前两个连续字符的坐标数据依次进行垂直坐标Y阈值化处理并判断拆分与否，水平坐标X阈值化处理并判断拆分与否，再对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否，将最后判定不拆分的两个连续字符合并并聚类到相应的字符串集合中。

9.如权利要求8所述的PDF表格提取方法，其特征在于还包括：

步骤D3，判断当前的两个连续字符是否为当前页的最后两个连续字符，如若不是，继续从步骤D2开始处理；若是当前页的最后两个连续字符，则以页码为关联词将当前页对应的所有字符串集合作为文本数据存储到字典的文本数据单元中，再继续进行步骤D4；

步骤D4，判断当前页是否为最后一页，如若当前页不是最后一页，则继续从步骤D1开始处理；如若当前页是最后一页，将存储有字符串集合的文本数据的字典存储到磁盘文件中。

10.如权利要求9所述的PDF表格提取方法，其特征在于，

步骤D2中的垂直坐标Y阈值化处理，即比较两个连续字符的垂直坐标Y数据之差的绝对值是否大于设定的垂直阈值，如若大于设定的垂直阈值，则将两个连续字符拆分，然后继续从步骤D3开始；如若不大于设定的垂直阈值，则继续进行水平坐标X阈值化处理；

步骤D2中的水平坐标X阈值化处理，即比较两个连续字符的水平坐标X数据之差的绝对值是否大于设定的水平阈值，如若大于设定的水平阈值，则将两个连续字符拆分，然后继续从步骤D3开始；如若不大于设定的水平阈值，则继续对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否；

步骤D2中的对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否是指：判断两个连续字符的水平坐标数据之间是否具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，如若两个连续字符的水平坐标数据之间具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，则将两个连续字符拆分，然后继续从步骤D3开始；如若两个连续字符的水平坐标数据之间不具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，则将最后判定不拆分的两个连续字符合并，并聚类到相应的字符串集合中，然后继续从步骤D3开始。

11.如权利要求4所述的PDF表格提取方法，其特征在于，步骤E包括：

步骤E1，按页码依次从字典的表格数据单元和文本数据单元中分别获取当前页的表格框架数据和文本数据；

步骤E2，依次遍历当前页的表格框架数据和文本数据，经由表格框架数据对应的当前

表格的上下边位置数据、左右边的位置数据、每行上下边位置数据和每列左右边的位置数据，获得表格框架的行和列，进而可得到当前表格的所有单元格的矩形坐标数据；

步骤E3，将单元格的矩形坐标数据和文本数据中的字符串集合对应的坐标数据范围相匹配，得到当前表格的表格数据；

步骤E4，判断当前表格是否为当前页的最后一个表格，若不是当前页的最后一个表格，则继续从步骤E2开始；若是当前页最后一个表格，则以页码为关联词将当前页的表格数据存储到字典中，并继续判断当前页是否为最后一页，如若不是最后一页，则继续从步骤E1开始处理；如若是最后一页，将提取出的表格数据存储到磁盘文件中，从而完成对PDF文档的表格数据的全部提取过程。

12.如权利要求1所述的PDF表格提取方法，其特征在于，步骤A的具体过程包括：

步骤A1，首先对PDF文档按页码依次进行解析处理，获得当前页对应的渲染及绘图指令；

步骤A2，经由当前页的渲染及绘图指令获得当前页的第一线条数据、字符数据和图像数据，并将第一线条数据、字符数据和图像数据以页码为关联词存储到字典的PDF数据单元中，完成对当前页的解析处理；

步骤A3，判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则继续从步骤A1开始处理；如若当前页是最后一页，则继续进行步骤A4；

步骤A4，将存储PDF数据的字典存储到磁盘文件中，以便按页码依次进行快速存取。

13.如权利要求12所述的PDF表格提取方法，其特征在于，所述的第一线条数据包括：线条的坐标、方向、长度、宽度和/或颜色；所述的字符数据包括：字符的坐标、编码、字体、字号和/或颜色；所述的图像数据包括图像的坐标、宽高和/或点阵数据。

14.如权利要求12所述的PDF表格提取方法，其特征在于，步骤B具体过程包括：

步骤B1，按页码依次从字典的PDF数据单元中提取当前页的所有图像数据；

步骤B2，对图像数据进行预处理，并得到图像数据预处理结果；

步骤B3，对图像数据预处理结果进行图像识别算法处理，获得具有表格数据的图像数据中表格数据对应的第二线条数据。

15.如权利要求14所述的PDF表格提取方法，其特征在于，步骤B2具体为，遍历当前页的每一组图像数据，对当前组图像数据依次采用去噪、增强和/或倾斜校正方式对其进行预处理，并得到当前组图像数据的图像预处理结果。

16.如权利要求15所述的PDF表格提取方法，其特征在于，步骤B3包括：

步骤B31，对当前图像预处理结果依次采用形态学、霍夫变换及图像投影方式进行处理，来判别当前组图像数据的图像预处理结果中是否具有表格数据，如若具有表格数据，则继续步骤B32；如若不具有表格数据，则继续步骤B33；

步骤B32，提取出当前组图像数据中具有的表格数据，对其依次进行自适应二值化处理和形态学处理，获得对应的水平线二值图和垂直线二值图；由其对应的水平线二值图和垂直线二值图，获得具有表格数据的当前组图像数据中的表格数据对应的叠加图和交点，经由叠加图和交点得到当前组图像数据中的表格数据对应的第二线条数据；

步骤B33，判断当前组图像数据是否为当前页的最后一组图像数据，如若不是，对当前页的下一组图像数据继续从步骤B2开始处理；如若当前图像数据是当前页的最后一组图像

数据，则将当前页的所有图像数据的第二线条数据以页码为关联词存储在字典的PDF数据单元中，并继续步骤B34；

步骤B34，继续判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则继续从步骤B1开始处理；如若当前页是最后一页，将存储有第二线条数据的字典存储到磁盘文件中。

17.如权利要求16所述的PDF表格提取方法，其特征在于，所述第二线条数据包括线条的坐标、方向、长度、宽度和/或颜色。

18.如权利要求1所述的PDF表格提取方法，其特征在于，步骤A、步骤B、步骤C、步骤D及步骤E可经由多线程编程技术独立且并发运行。

PDF表格提取方法

技术领域

[0001] 本发明涉及PDF文档数据挖掘与提取技术领域，具体涉及一种PDF表格提取方法。

背景技术

[0002] PDF(Portable Document Format)即便携式文档格式，由Adobe Systems用于进行文件交换所开发出的文件格式，其与应用程序、操作系统及其他硬件均无交互关系。PDF文档以PostScript语言图像模型为基础，保证PDF文档在任何一台打印机上都可具有精确的颜色和准确的打印效果，即PDF会如实地再现PDF文档中的每一个字符、颜色以及图像等内容。随着计算机及互联网技术的快速发展，PDF文档越来越广泛地用在经济、金融、教育、科研及学术等各种领域。由于PDF设计目的只是为了展示文档或用于打印文档，而没有与其他计算机程序进行通讯与交互的功能。因此，PDF文档中所包含的大量数据，特别其中包含的表格数据，难以被其他计算机程序直接使用。PDF文档主要由图像、表格及文本等数据组成。对于现有的PDF文档的提取技术，基本上都可以方便地提取出PDF文档中的文本数据，但对于PDF文档中的表格数据的提取，存在着提取准确率不高，对于提取准确率要求不高的普通领域而言，现有技术的PDF文档中文本数据及表格数据的提取技术基本可以满足要求；但是对于数据提取要求较高的金融等领域，却存在无法满足准确率要求的情况，因此难以实现大规模应用在金融等准确率要求高的领域。同时，现有的提取技术存在着提取效率不高，对于及时性要求较高的场合，难以保证提取数据的时效性。

发明内容

[0003] 本发明针对现有技术中对于PDF文档中表格数据提取的准确性及效率不足的缺陷，目的在于提供一种高准确性及高效率的提取PDF文档中表格数据的PDF表格提取方法。

[0004] 实现上述目的的技术方案是：

[0005] 本发明PDF表格提取方法，该方法包括：

[0006] 步骤A，对PDF文档进行解析，获取图像数据、第一线条数据和字符数据；

[0007] 步骤B，采用图像识别算法对经由步骤A获取的图像数据进行处理，从具有表格数据的图像数据中获得其表格数据对应的第二线条数据；

[0008] 步骤C，采用图形算法对经由步骤A获得的第一线条数据和经由步骤B获得的第二线条数据分别进行处理，获得具有表格行数据和列数据的表格框架数据；

[0009] 步骤D，采用聚类算法对经由步骤A获得的字符数据进行聚类处理，获得具有字符串集合的文本数据；

[0010] 步骤E，经由步骤C获得的表格框架数据中的表格行数据和列数据，得到对应的表格单元格，将表格单元格与步骤D获得的文本数据中的字符串集合相匹配，获得PDF文档中的表格数据。

[0011] 本发明的一实施例中，步骤C中所述图形算法处理包括对步骤A获得的第一线条数据和步骤B获得的第二线条数据对应的垂直线条进行垂直投影、水平线条进行水平投影、水

平线条进行垂直投影及垂直线条进行水平投影，从而获得具有表格行数据和列数据的表格框架数据。

[0012] 本发明的一实施例中，步骤D中所述聚类算法处理包括对步骤A获得的字符数据中任两个连续字符数据的坐标数据依次进行垂直坐标Y阈值化处理、水平坐标X阈值化处理和/或垂直线条约束处理，将同类连续字符聚类到相应的字符串集合中，从而获得具有字符串集合的文本数据。

[0013] 本发明的一实施例中，步骤A至步骤E中的各处理均是按页码依次进行处理；步骤A获取的图像数据、第一线条数据、字符数据及步骤B获取的第二线条数据均以页码为关联词存储到字典的PDF数据单元中，步骤C获得的表格框架数据以页码为关联词存储到字典的表格数据单元中，步骤D获得的文本数据以页码为关联词存储到字典的文本数据单元中。

[0014] 本发明的一实施例中，步骤C具体包括：

[0015] 步骤C1，按页码依次从字典的PDF数据单元中获取当前页的第一线条数据和第二线条数据的垂直线条数据和水平线条数据；

[0016] 步骤C2，对当前页的第一线条数据和第二线条数据的线条数据进行图形算法处理，获得当前页的表格数量及每个表格对应的上下边位置数据、左右边的位置数据、每行上下边位置数据和每列左右边的位置数据，即获得具有表格行数据和列数据的当前页的表格框架数据，并将其以页码为关联词存储到字典的表格数据单元中。

[0017] 本发明的一实施例中，所述步骤C2具体包括：

[0018] 步骤C21，对当前页的第一线条数据和第二线条数据的垂直线条进行垂直投影，对所获得的垂直投影结果数据进行处理分析，获得当前页的表格数量及每个表格对应的上下边位置数据；

[0019] 步骤C22，依次遍历当前页的每一个表格，对当前表格的第一线条数据或第二线条数据的水平线条进行水平投影，获得当前表格左右边的位置数据；对当前表格的第一线条数据或第二线条数据的水平线条进行垂直投影，获得当前表格每行上下边位置数据；对当前表格的第一线条数据或第二线条数据的垂直线条进行水平投影，获得当前表格每列左右边的位置数据，从而获得具有表格行数据和列数据的当前表格的表格框架数据；

[0020] 步骤C23，判断当前表格是否为当前页的最后一表格，如若当前表格不是当前页的最后一表格，则继续从步骤C22开始；如若当前表格是当前页的最后一表格，则以页码为关联词将具有表格行数据和列数据的当前页的表格框架数据存储到字典的表格数据单元中。

[0021] 本发明的一实施例中，还包括步骤C3：

[0022] 步骤C3，判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则继续从步骤C1开始处理；如若当前页是最后一页，将存储有表格框架数据的字典存储到磁盘文件中。

[0023] 本发明的一实施例中，步骤D包括：

[0024] 步骤D1，按页码依次从字典的PDF数据单元中获取当前页所有字符数据；

[0025] 步骤D2，依次遍历当前页的两个连续字符数据，对获取的当前两个连续字符的坐标数据依次进行垂直坐标Y阈值化处理并判断拆分与否，水平坐标X阈值化处理并判断拆分与否，再对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否，将最后判定不拆分的两个连续字符合并并聚类到相应的字符串集合中。

[0026] 本发明的一实施例中，还包括步骤D3：

[0027] 步骤D3，判断当前的两个连续字符是否为当前页的最后两个连续字符，如若不是，继续对当前页的下两个连续字符从步骤D2开始处理；若是当前页的最后两个连续字符，则以页码为关联词将当前页对应的所有字符串集合作为文本数据存储到字典的文本数据单元中，再继续进行步骤D4；

[0028] 步骤D4，判断当前页是否为最后一页，如若当前页不是最后一页，则继续从步骤D1开始处理；如若当前页是最后一页，将存储有文本数据的字典存储到磁盘文件中。

[0029] 本发明的一实施例中，步骤D2中的垂直坐标Y阈值化处理，即比较两个连续字符的垂直坐标Y数据之差的绝对值是否大于设定的垂直阈值，如若大于设定的垂直阈值，则将两个连续字符拆分，然后继续从步骤D3开始；如若不大于设定的垂直阈值，则继续进行水平坐标X阈值化处理；

[0030] 步骤D2中的水平坐标X阈值化处理，即比较两个连续字符的水平坐标X数据之差的绝对值是否大于设定的水平阈值，如若大于设定的水平阈值，则将两个连续字符拆分，然后继续从步骤D3开始；如若不大于设定的水平阈值，则继续对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否；

[0031] 步骤D2中的对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否是指：判断两个连续字符的水平坐标数据之间是否具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，如若两个连续字符的水平坐标数据之间具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，则将两个连续字符拆分，然后继续从步骤D3开始；如若两个连续字符的水平坐标数据之间不具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，则将不拆分的两个连续字符合并，并聚类到相应的字符串集合中，然后继续从步骤D3开始。

[0032] 本发明的一实施例中，步骤E包括：

[0033] 步骤E1，按页码依次从字典的表格数据单元和文本数据单元中分别获取当前页的表格框架数据和文本数据；

[0034] 步骤E2，依次遍历当前页的表格框架数据和文本数据，经由表格框架数据对应的当前表格的上下边位置数据、左右边的位置数据、每行上下边位置数据和每列左右边的位置数据，获得表格框架的行和列，进而可得到当前表格的所有单元格的矩形坐标数据；

[0035] 步骤E3，将单元格的矩形坐标数据和文本数据中的字符串集合对应的坐标数据范围相匹配，得到当前表格的表格数据；

[0036] 步骤E4，判断当前表格是否为当前页的最后一个表格，若不是当前页的最后一个表格，则继续从步骤E2开始；若是当前页最后一个表格，则以页码为关联词将当前页的表格数据存储到字典中，并继续判断当前页是否为最后一页，如若不是最后一页，则对当前页的下一页继续从步骤E1开始处理；如若是最后一页，将提取出的表格数据存储到磁盘文件中，从而完成对PDF文档的表格数据的全部提取过程。

[0037] 本发明的一实施例中，步骤A的具体过程包括：

[0038] 步骤A1，首先对PDF文档按页码依次进行解析处理，获得当前页对应的渲染及绘图指令；

[0039] 步骤A2，经由当前页的渲染及绘图指令获得当前页的第一线条数据、字符数据和

图像数据，并将第一线条数据、字符数据和图像数据以页码为关联词存储到字典的PDF数据单元中，完成对当前页的解析处理；

[0040] 步骤A3，判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则对当前页的下一页继续从步骤A1开始处理；如若当前页是最后一页，则继续进行步骤A4；

[0041] 步骤A4，将存储PDF数据的字典存储到磁盘文件中，以便按页码依次进行快速存取。

[0042] 本发明的一实施例中，所述的第一线条数据包括：线条的坐标、方向、长度、宽度和/或颜色；所述的字符数据包括：字符的坐标、编码、字体、字号和/或颜色；所述的图像数据包括图像的坐标、宽高和/或点阵数据。

[0043] 本发明的一实施例中，步骤B具体过程包括：

[0044] 步骤B1，按页码依次从字典的PDF数据单元中提取当前页的所有图像数据；

[0045] 步骤B2，对图像数据进行预处理，并得到图像数据预处理结果；

[0046] 步骤B3，对图像数据预处理结果进行图像识别算法处理，获得具有表格数据的图像数据中表格数据对应的第二线条数据。

[0047] 本发明的一实施例中，步骤B2具体为，遍历当前页的每一组图像数据，对当前组图像数据依次采用去噪、增强和/或倾斜校正方式对其进行预处理，并得到当前组图像数据的图像预处理结果。

[0048] 本发明的一实施例中，步骤B3包括：

[0049] 步骤B31，对当前图像预处理结果依次采用形态学、霍夫变换及图像投影方式进行处理，来判别当前组图像数据的图像预处理结果中是否具有表格数据，如若具有表格数据，则继续步骤B32；如若不具有表格数据，则继续步骤B33；

[0050] 步骤B32，提取出当前组图像数据中具有的表格数据，对其依次进行自适应二值化处理和形态学处理，获得对应的水平线二值图和垂直线二值图；由其对应的水平线二值图和垂直线二值图，获得具有表格数据的当前组图像数据中的表格数据对应的叠加图和交点，经由叠加图和交点得到当前组图像数据中的表格数据对应的第二线条数据；

[0051] 步骤B33，判断当前组图像数据是否为当前页的最后一组图像数据，如若不是，对当前页的下一组图像数据继续从步骤B2开始处理；如若当前图像数据是当前页的最后一组图像数据，则将当前页的所有图像数据的第二线条数据以页码为关联词存储在字典的PDF数据单元中，并继续步骤B34；

[0052] 步骤B34，继续判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则对当前页的下一页继续从步骤B1开始处理；如若当前页是最后一页，将存储有第二线条数据的字典存储到磁盘文件中。

[0053] 本发明的一实施例中，所述第二线条数据包括线条的坐标、方向、长度、宽度和/或颜色。

[0054] 本发明的一实施例中，步骤A、步骤B、步骤C、步骤D及步骤E可经由多线程编程技术独立且并发运行。

[0055] 本发明的积极进步效果在于：

[0056] 本发明PDF表格提取方法，首先对PDF文档进行PDF解析处理，得到PDF文档所有页对应的图像数据、第一线条数据、和字符数据；采用图像识别算法对图像数据进行处理，将

具有表格数据的图像数据的表格数据转换成表格数据对应的第二线条数据;采用图形算法对第一线条数据和第二线条数据进行处理,获得所有表格框架数据;采用聚类算法对字符数据进行处理,获得表格数据对应的文本数据;将获得的表格框架和文本数据合并,得到输出具有表格框架和文本数据的所有表格数据。主要的积极进步效果如下:

[0057] (1)在采用聚类算法对PDF文档解析后的字符数据进行处理过程中,使用线条数据对字符数据的处理进行约束,使得字符数据有效地合并与分离到对应的单元格中,提高了文本数据提取的准确率;

[0058] (2)PDF解析过程、图像识别算法对图像数据进行处理过程、图形算法对线条数据进行处理过程及聚类算法对字符数据进行处理过程四处理过程可以经由多线程编程技术并发运行,提高PDF文档表格数据提取的效率;

[0059] (3)采用的提取技术架构清晰明了,便于理解、容易实现、方便调试程序及后期维,护增强了提取技术的扩展性;

[0060] (4)降低了提取成本,且本发明的整个方案为后续的改进及扩展的基础,可进一步增强PDF文档表格数据提取的准确率和效率。

附图说明

[0061] 图1为本发明PDF表格提取方法的流程示意图;

[0062] 图2为本发明PDF表格提取方法的PDF解析过程流程图;

[0063] 图3为本发明PDF表格提取方法的图像识别算法图像处理过程流程图;

[0064] 图4为本发明PDF表格提取方法的图形算法线条处理过程流程图;

[0065] 图5为本发明PDF表格提取方法的图形算法线条处理的第一示意图;

[0066] 图6为本发明PDF表格提取方法的图形算法线条处理的第二示意图;

[0067] 图7为本发明PDF表格提取方法的图形算法线条处理的第三示意图;

[0068] 图8为本发明PDF表格提取方法的聚类算法过程流程图;

[0069] 图9为本发明PDF表格提取方法的聚类算法过程的第一示意图;

[0070] 图10为本发明PDF表格提取方法的聚类算法过程的第二示意图。

具体实施方式

[0071] 下面举出较佳实施例,并结合图1至图10来更清楚完整地说明本发明。

[0072] 本发明实施方式的PDF表格提取方法,包括:

[0073] 如图1所示,为本发明PDF表格提取方法的流程图,即PDF表格提取方法具体实现过程包括:先对PDF文档进行解析处理,获取图像数据、第一线条数据和字符数据。对经由PDF解析获得的图像数据采用图像识别算法进行处理,从具有表格数据的图像数据中获得其表格数据对应的第二线条数据。采用图形算法对经由PDF解析获得的第一线条数据和经由图像识别算法对图像数据进行处理获得的第二线条数据进行处理,获得PDF文档具有的所有表格行数据和列数据的表格框架数据。采用聚类算法对经由PDF解析获得的所有字符数据进行聚类处理,获得具有字符串集合的文本数据。经由获得的所有表格框架数据中的表格行数据和列数据,进而获得与表格框架对应的单元格,将表格单元格与经由聚类算法获得的文本数据中的字符串集合相匹配,获得PDF文档中的表格数据,即完成对PDF文档的表格

数据的全部提取过程。其中，对PDF文档进行解析处理、对图像数据采用图像识别算法进行处理、对第一线条数据和第二线条数据采用图形算法进行处理、对字符数据采用聚类算法进行处理及将表格框架数据与文本数据相匹配的处理均按页码依次进行处理，且可经由多线程编程技术独立且并发运行，提高对PDF文档中表格数据提取的效率。

[0074] 以下结合图2至图10，对本发明PDF表格提取方法的详细过程进行说明。

[0075] 如图2所示，首先对PDF文档按页码依次进行解析处理，获得每一页渲染及绘图指令。遍历PDF文档每一页，经由当前页对应的渲染及绘图指令获得当前页对应的图像数据、第一线条数据和字符数据。渲染及绘图指令包括Fillchar、Moveto、Lineto、Image等指令。图像数据、第一线条数据和字符数据统称为PDF数据，将获得的当前页对应的PDF数据以页码为关联词存储到字典的PDF数据单元中。判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则对当前页的下一页继续进行解析处理；如若当前页是最后一页，则将存储PDF数据的字典存储到磁盘文件中，以便按页码依次进行快速存取PDF数据。图像数据包括图像对应的坐标、宽高及点阵等数据，PDF文档中所有图像数据中的部分图像数据为包含表格数据的图像数据。第一线条数据包括线条的坐标、垂直、水平、长度、宽度及颜色等数据。其中，第一线条数据不包括包含表格数据的图像数据中的表格数据对应的线条数据。字符数据包括字符的坐标、编码、字体、字号和颜色等数据。

[0076] 如图3所示，按页码依次从字典的PDF数据单元中获取经由PDF解析处理得到的所有图像数据。图像识别算法处理过程包括对图像数据进行去噪、增强、倾斜校正、霍夫变换、投影、二值化、腐蚀、膨胀、边缘检测等处理过程。(去噪、增强、倾斜校正、霍夫变换、投影、二值化、腐蚀、膨胀、边缘检测等图像处理过程可以见参考书《数字图像处理(第三版)》，出版社：电子工业出版社，作者：美，Rafael C.Gonzalez，Richard E.Woods；译者：阮秋琦，阮宇智。其中，去噪处理见该参考书的第三章灰度变换与空间滤波，P88-97；增强处理见该参考书的第三章灰度变换与空间滤波，P97-105；倾斜校正处理见该参考书的第四章频率域滤波，P124-191；霍夫变换处理见该参考书第十章图像分割，P472-477；投影处理见该参考书的第五章图像复原与重建，P232-235；二值化处理见该参考书的第三章灰度变换与空间滤波，P64；形态学、腐蚀处理、膨胀处理均见该参考书的第九章形态学图像处理，P428-429；边缘检测处理见该参考书的第十章图像分割，P443-508)

[0077] 首先，遍历当前页的当前组图像数据，对当前组图像数据依次采用去噪、增强及倾斜校正方式进行预处理，得到当前组图像数据图像预处理结果。其中去噪处理是基于形态学实现对图像数据的过滤处理，进而经由颜色模型又称HSV(Hue Saturation Value)模型，将过滤处理后的图像数据对应的色调、饱和度及明度进行处理，得到更加清晰的图像；进而对颜色模型处理后的清晰图像基于傅里叶变换进行倾斜校正处理，使得图像的上下边或左右边与图像所在页面上下或左右边平行，进而得到当前组图像数据的图像预处理结果。对当前组图像数据的图像预处理结果采用形态学方法提取图像数据中对于表达和描绘表格形状有用处的图像数据，再将其经由霍夫变换从当前图像数据中分离出具有相同特征的若干直线数据，得到当前具有表格数据的图像数据中所有直线数据，即对当前组图像数据的图像预处理结果依次采用形态学、霍夫变换及图像投影方式进行处理，来判别当前组图像数据的图像预处理结果中是否具有表格数据，如若具有表格数据，则对获得的当前图像数据中的表格数据依次进行自适应二值化处理和形态学方法进行处理获得当前组图像数据

中的表格数据相应的水平线二值图和垂直线二值图，通过水平线二值图和垂直线二值图，即可得到叠加图、水平线及垂直线的交点，由叠加图和交点得到图像数据中的表格数据对应的第二线条数据。第二线条数据包括线条的坐标、方向、长度、宽度和/或颜色等。如若当前组图像数据的图像预处理结果中不具有表格数据，则继续判别当前组图像数据是否为当前页的最后一组图像数据，如若当前组图像数据不是当前页的最后一组图像数据，则继续对当前页的下一组图像数据进行图像识别算法处理过程。

[0078] 如若当前图像数据是当前页的最后一组图像数据，则将当前页的所有组图像数据的第二线条数据以页码为关联词存储在字典的PDF数据单元中；继续判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，继续对当前页的下一页的图像数据继续进行图像识别处理过程；如若当前页是最后一页，将存储有第二线条数据的字典存储到磁盘文件中。第二线条数据包括线条的坐标、方向、长度、宽度和/或颜色等。

[0079] 如图4至图7所示，按页码依次从字典的PDF数据单元中获取当前页的经由PDF解析过程获得的第一线条数据和经由图像识别算法获得的第二线条数据的垂直线条数据和水平线条数据，对当前页的第一线条数据和第二线条数据的垂直线条数据和水平线条数据分别进行图形算法处理，获得当前页的表格数量及每个表格对应的上下边位置数据、左右边的位置数据、每行上下边位置数据和每列左右边的位置数据，即获得具有表格行数据和列数据的当前页的表格框架数据，并将其以页码为关联词存储到字典的表格数据单元中。图形算法处理包括对第一线条数据和/或第二线条数据的垂直线条和水平线条的垂直线条进行垂直投影、水平线条进行水平投影、水平线条进行垂直投影及垂直线条进行水平投影。首先，对第一线条数据中的垂直线条数据和第二线条数据的中的垂直线条数据对应的垂直线条进行垂直投影，获得垂直投影结果数据，如图5所示，对所获得的垂直投影结果数据进行处理分析，得到当前页的表格数量及每个表格对应的上下边位置数据。依次遍历当前页的每一个表格，对当前表格的第一线条数据或第二线条数据的水平线条进行水平投影，如图6所示，获得当前表格左右边的位置数据；对当前表格的第一线条数据或第二线条数据的水平线条进行垂直投影，如图7所示，获得当前表格每行上下边位置数据；对当前表格的第一线条数据或第二线条数据的垂直线条进行水平投影，如图7所示，获得当前表格每列左右边的位置数据，从而获得具有表格行数据和列数据的当前表格的表格框架数据。

[0080] 判断当前表格是否为当前页的最后一表格，如若当前表格不是当前页的最后一表格，则对当前页的下一表格继续进行水平线条进行水平投影、水平线条进行垂直投影及垂直线条进行水平投影处理；如若当前表格是当前页的最后一表格，则以页码为关联词将具有表格行数据和列数据的当前页的表格框架数据存储到字典的表格数据单元中。继续判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则对当前页的下一页继续进行图形算法处理；如若当前页是最后一页，则将存储有表格框架数据的字典存储到磁盘文件中，从而完成对PDF文档中第一线条数据和第二线条数据的图形算法处理过程。

[0081] 如图8至图10所示，对经由PDF解析处理获得的PDF文档中所有的字符数据的坐标数据采用聚类算法进行处理，获得具有字符串集合的文本数据。聚类算法过程包括对经由PDF解析处理获得字符数据中任两个连续字符数据的坐标数据依次进行垂直坐标Y阈值化处理、水平坐标X阈值化处理和/或垂直线条约束处理而将同类连续字符聚类到相应的字符串集合中，从而获得具有字符串集合的文本数据。

[0082] 首先，按页码依次从字典的PDF数据单元中获取当前页的字符数据，依次遍历当前页的任意两个连续字符数据，对获取的当前两个连续字符的坐标数据依次进行垂直坐标Y阈值化处理(阈值根据获得的分布情况，取95％置信区间对应的值)，即比较两个连续字符的垂直坐标Y数据之差的绝对值是否大于设定的垂直阈值，如若大于设定的垂直阈值，则将两个连续字符拆分，然后继续判断当前两个连续字符是否为当前页的最后两个连续字符，如若当前两个连续字符不是当前页的最后两个连续字符，则对当前页的下两个连续字符的坐标数据继续进行垂直坐标Y阈值化处理、水平坐标X阈值化处理和/或垂直线条约束处理；如若当前两个连续字符是当前页的最后两个连续字符，则以页码为关联词将当前页对应的所有字符串集合作为文本数据存储到字典的文本数据单元中；再继续判断当前页是否为最后一页，如若当前页不是最后一页，则对当前页的下一页继续进行聚类算法处理过程；如若当前页是最后一页，则将存储有字符串集合的文本数据的字典存储到磁盘文件中。

[0083] 如若当前两个连续字符的坐标数据的垂直坐标Y之差不大于垂直阈值，则继续对当前两个连续字符的坐标数据进行水平坐标X阈值化处理，即比较两个连续字符的水平坐标X数据之差的绝对值是否大于设定的水平阈值，如若大于设定的水平阈值，则将两个连续字符拆分，然后继续判断当前两个连续字符是否为当前页的最后两个连续字符，如若当前两个连续字符不是当前页的最后两个连续字符，则对当前页的下两个连续字符的坐标数据继续进行垂直坐标Y阈值化处理、水平坐标X阈值化处理和/或垂直线条约束处理；如若当前两个连续字符是当前页的最后两个连续字符，则以页码为关联词将当前页对应的所有字符串集合作为文本数据存储到字典的文本数据单元中；再继续判断当前页是否为最后一页，如若当前页不是最后一页，则对当前页的下一页继续进行聚类算法处理过程；如若当前页是最后一页，则将字典中的文本数据单元中的具有字符串集合的文本数据存储到磁盘文件中。

[0084] 如若当前两个连续字符的坐标数据的水平坐标X之差不大于水平阈值，则继续对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束，即判断两个连续字符的水平坐标数据之间是否具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，如若两个连续字符的水平坐标数据之间具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，则将两个连续字符拆分，然后继续判断当前两个连续字符是否为当前页的最后两个连续字符，如若当前两个连续字符不是当前页的最后两个连续字符，则对当前页的下两个连续字符的坐标数据继续进行垂直坐标Y阈值化处理、水平坐标X阈值化处理和/或垂直线条约束处理；如若当前两个连续字符是当前页的最后两个连续字符，则以页码为关联词将当前页对应的所有字符串集合作为文本数据存储到字典的文本数据单元中；再继续判断当前页是否为最后一页，如若当前页不是最后一页，则对当前页的下一页继续进行聚类算法处理过程；如若当前页是最后一页，则将存储有字符串集合的文本数据的字典存储到磁盘文件中。

[0085] 如若两个连续字符的水平坐标数据之间不具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，则将两个连续字符合并，然后继续判断当前两个连续字符是否为当前页的最后两个连续字符，如若当前两个连续字符不是当前页的最后两个连续字符，则对当前页的下两个连续字符的坐标数据继续进行垂直坐标Y阈值化处理、水平坐标X阈值化处理和/或垂直线条约束处理；如若当前两个连续字符是当前页的最后两个连续字

符，则以页码为关联词将当前页对应的所有字符串集合作为文本数据存储到字典的文本数据单元中；再继续判断当前页是否为最后一页，如若当前页不是最后一页，则对当前页的下一页继续进行聚类算法处理过程；如若当前页是最后一页，则将存储有字符串集合的文本数据的字典存储到磁盘文件中。

[0086] 如图9所示，为一对字符数据采用聚类算法进行处理的说明实施例，其中任意两个连续字符数据垂直阈值化处理的拆分处理结果，如g-h，r-s之间的垂直阈值化拆分；任意两个连续字符数据水平阈值化处理的拆分处理，如d-e，k-l，m-n，p-q，u-v及2-3之间；任意两个连续字符数据的水平坐标数据之间设有第一线条数据或第二线条数据中的垂直线条约束，如y-z之间。连续字符数据合并后获得的字符串文本，如abcd，efg，hijk，…，vwxy，z12，34。

[0087] 按页码依次从字典的表格数据单元和文本数据单元中分别获取当前页的表格框架数据和文本数据。依次遍历当前页的表格框架数据和文本数据，经由表格框架数据对应的当前表格的上下边位置数据、左右边的位置数据、每行上下边位置数据和每列左右边的位置数据，获得表格框架的行和列，进而可得到当前表格的所有单元格的矩形坐标数据；每个单元格对应的字符串文本数量大于等于零。判断该表格是否为当前页的最后一个表格，若不是当前页的最后一个表格，则对当前页的下一表格继续进行获得其所有单元格矩形坐标数据处理；若是当前页最后一个表格，则以页码为关联词将当前页的表格数据以json格式按页码存储到字典中，并继续判断当前页是否为最后一页，如若不是最后一页，则对当前页的下一页继续获取其表格框架数据和文本数据，得到其对应的表格数据的处理过程；如若是最后一页，将提取出的表格数据存储到磁盘文件中，从而完成对PDF文档的表格数据的全部提取过程。

[0088] 如图10所示，按页码从磁盘文件中字典的表格数据单元获得的表格框架数据，由表格框架数据对应表格的上下边位置数据、左右边的位置数据、每行上下边位置数据和每列左右边的位置数据，获得表格框架的行和列，进而得到当前表格的所有单元格的矩形坐标数据，将单元格的矩形坐标数据和文本数据中的字符串集合对应的坐标数据范围相匹配，得到当前表格的表格数据。

[0089] 本发明PDF表格提取方法的具体实施步骤如下：

[0090] 对PDF文档进行PDF解析过程，具体过程如下：

[0091] 步骤S11，首先对PDF文档按页码依次进行解析处理，获得当前页对应的渲染及绘图指令；

[0092] 步骤S12，经由当前页的渲染及绘图指令获得当前页的第一线条数据、字符数据和图像数据，并将第一线条数据、字符数据和图像数据以页码为关联词存储到字典的PDF数据单元中，完成对当前页的解析处理；

[0093] 步骤S13，判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则对继续从步骤A2开始处理；如若当前页是最后一页，继续进行步骤A4；

[0094] 步骤S14，将存储有PDF数据的字典存储到磁盘文件中，以便按页码依次进行快速存取。

[0095] 采用图像识别算法对经由PDF解析获取的图像数据按页码依次进行处理，具体过程如下：

[0096] 步骤S21，按页码依次从字典的PDF数据单元中提取当前页的所有图像数据；

[0097] 步骤S22，遍历当前页的图像数据，对当前图像数据依次采用去噪、增强和/或倾斜校正方式对其进行预处理，并得到当前图像预处理结果；

[0098] 步骤S23，对当前图像预处理结果依次采用形态学、霍夫变换及图像投影方式进行处理，来判别当前图像预处理结果中是否具有表格数据，如若具有表格数据，继续步骤S24；如若不具有表格数据，继续步骤S25；

[0099] 步骤S24，提取出当前图像数据中具有的表格数据，对其依次进行自适应二值化处理和形态学处理，获得对应的水平线二值图和垂直线二值图；由其对应的水平线二值图和垂直线二值图，获得具有表格数据的图像数据中的表格数据对应的叠加图和交点，经由叠加图和交点得到当前图像数据中的表格数据对应的第二线条数据；

[0100] 步骤S25，判断当前图像数据是否为当前页的最后一组图像数据，如若不是，继续从步骤S22开始；如若当前图像数据是当前页的最后一组图像数据，则将当前页的所有图像数据的第二线条数据以页码为关联词存储在字典的PDF数据单元中，并继续步骤S26；

[0101] 步骤S26，继续判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则对当前页的下一页继续从步骤S21开始处理；如若当前页是最后一页，将存储有第二线条数据的字典存储到磁盘文件中。

[0102] 采用图形算法对经由PDF解析获得第一线条数据和经由图像识别算法对图像数据进行处理获得第二线条数据按页码依次进行处理，具体过程如下：

[0103] 步骤S31，按页码依次从字典的PDF数据单元中获取当前页的第一线条数据和第二线条数据的垂直线条数据和水平线条数据；

[0104] 步骤S32，对当前页的第一线条数据和第二线条数据的垂直线条进行垂直投影，对所获得的垂直投影结果数据进行处理分析，获得当前页的表格数量及每个表格对应的上下边位置数据；

[0105] 步骤S33，依次遍历当前页的每一个表格，对当前表格的第一线条数据或第二线条数据的水平线条进行水平投影，获得当前表格左右边的位置数据；对当前表格的第一线条数据或第二线条数据的水平线条进行垂直投影，获得当前表格每行上下边位置数据；对当前表格的第一线条数据或第二线条数据的垂直线条进行水平投影，获得当前表格每列左右边的位置数据，从而获得具有表格行数据和列数据的当前表格的表格框架数据；

[0106] 步骤S34，判断当前表格是否为当前页的最后一表格，如若当前表格不是当前页的最后一表格，则继续从步骤S33开始；如若当前表格是当前页的最后一表格，则以页码为关联词将当前页的所有表格框架数据存储到字典的表格数据单元中；

[0107] 步骤S35，判断当前页是否为PDF文档的最后一页，如若当前页不是最后一页，则继续从步骤S31开始；如若当前页是最后一页，则将存储有表格数据的字典存储到磁盘文件中。

[0108] 采用聚类算法对经由步骤A获得的字符数据进行聚类处理，获得具有字符串集合的文本数据过程包括：

[0109] 步骤S41，按页码依次从字典的PDF数据单元中获取当前页所有字符数据；

[0110] 步骤S42，依次遍历当前页的两个连续字符数据，对获取的当前两个连续字符的坐标数据依次进行垂直坐标Y阈值化处理并判断拆分与否，水平坐标X阈值化处理并判断拆分

与否，再对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否，将最后判定不拆分的两个连续字符合并，并聚类到相应的字符串集合中；

[0111] 垂直坐标Y阈值化处理，即比较两个连续字符的垂直坐标Y数据之差的绝对值是否大于设定的垂直阈值，如若大于设定的垂直阈值，则将两个连续字符拆分，然后继续从步骤S43开始；如若不大于设定的垂直阈值，则继续进行水平坐标X阈值化处理；

[0112] 水平坐标X阈值化处理，即比较两个连续字符的水平坐标X数据之差的绝对值是否大于设定的水平阈值，如若大于设定的水平阈值，则将两个连续字符拆分，然后继续从步骤S43开始；如若不大于设定的水平阈值，则继续对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否；

[0113] 对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束判断拆分与否是指：判断两个连续字符的水平坐标数据之间是否具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，如若两个连续字符的水平坐标数据之间具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，则将两个连续字符拆分，然后继续从步骤S43开始；如若两个连续字符的水平坐标数据之间不具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据，则将两个连续字符合并，然后继续从步骤S43开始。

[0114] 步骤S43，判断当前的两个连续字符是否为当前页的最后两个连续字符，如若不是，继续从步骤S42开始；若是当前页的最后两个连续字符，则以页码为关联词将当前页对应的所有字符串集合作为文本数据存储到字典的文本数据单元中，再继续进行步骤S44；

[0115] 步骤S44，判断当前页是否为最后一页，如若当前页不是最后一页，则继续从步骤S41开始；如若当前页是最后一页，则将存储有文本数据的字典存储到磁盘文件中。

[0116] 经由步骤C获得的表格框架数据中的表格行数据和列数据，得到对应的表格单元格，将表格单元格与步骤D获得的文本数据中的字符串集合相匹配，获得PDF文档中的表格数据，具体过程如下：

[0117] 步骤S51，按页码依次从字典的表格数据单元和文本数据单元中分别获取当前页的表格框架数据和文本数据；

[0118] 步骤S52，依次遍历当前页的表格框架数据和文本数据，经由表格框架数据对应的当前表格的上下边位置数据、左右边的位置数据、每行上下边位置数据和每列左右边的位置数据，获得表格框架的行和列，进而可得到当前表格的所有单元格的矩形坐标数据；

[0119] 步骤S53，将单元格的矩形坐标数据和文本数据中的字符串集合对应的坐标数据范围相匹配，得到当前表格数据；

[0120] 步骤S54，判断该表格是否为当前页的最后一个表格，若不是当前页的最后一个表格，则继续从步骤S52开始；若是当前页最后一个表格，则以页码为关联词将当前页的表格数据存储到字典中，并继续判断当前页是否为最后一页，如若不是最后一页，继续从步骤E1开始；如若是最后一页，将提取出的表格数据存储到磁盘文件中，从而完成对PDF文档的表格数据的全部提取过程。

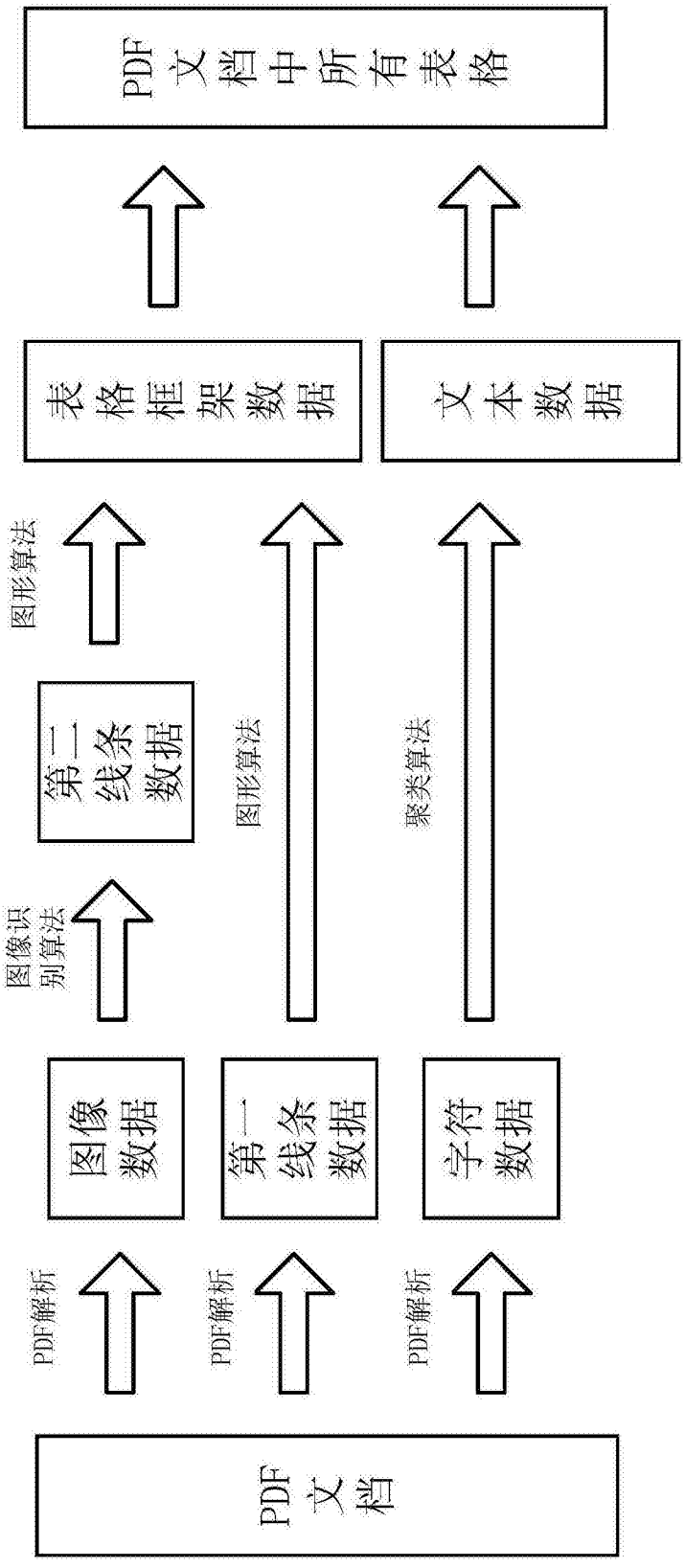

PDF 文档中所有表格
表格框架数据
文本数据
图形算法
第一线条数据
图形算法
聚类算法
图像识别算法
图像数据
第一线条数据
字符数据
PDF解析
PDF解析
PDF解析
PDF 文档

图1

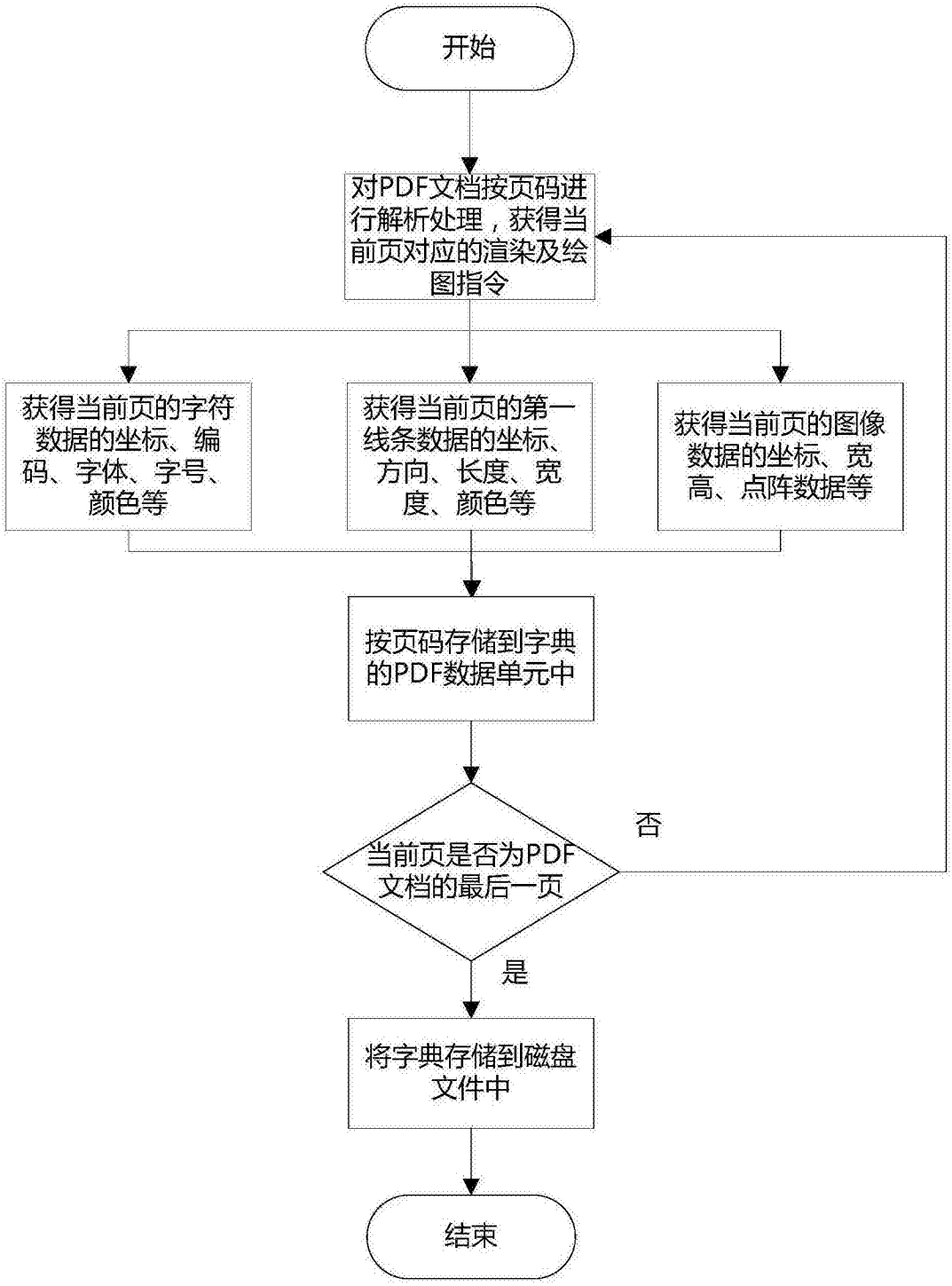

开始
对PDF文档按页码进行解析处理，获得当前页对应的渲染及绘图指令
获得当前页的字符数据的坐标、编码、字体、字号、颜色等
获得当前页的第一线条数据的坐标、方向、长度、宽度、颜色等
获得当前页的图像数据的坐标、宽高、点阵数据等
按页码存储到字典的PDF数据单元中
当前页是否为PDF文档的最后一页
否
是
将字典存储到磁盘文件中
结束

图2

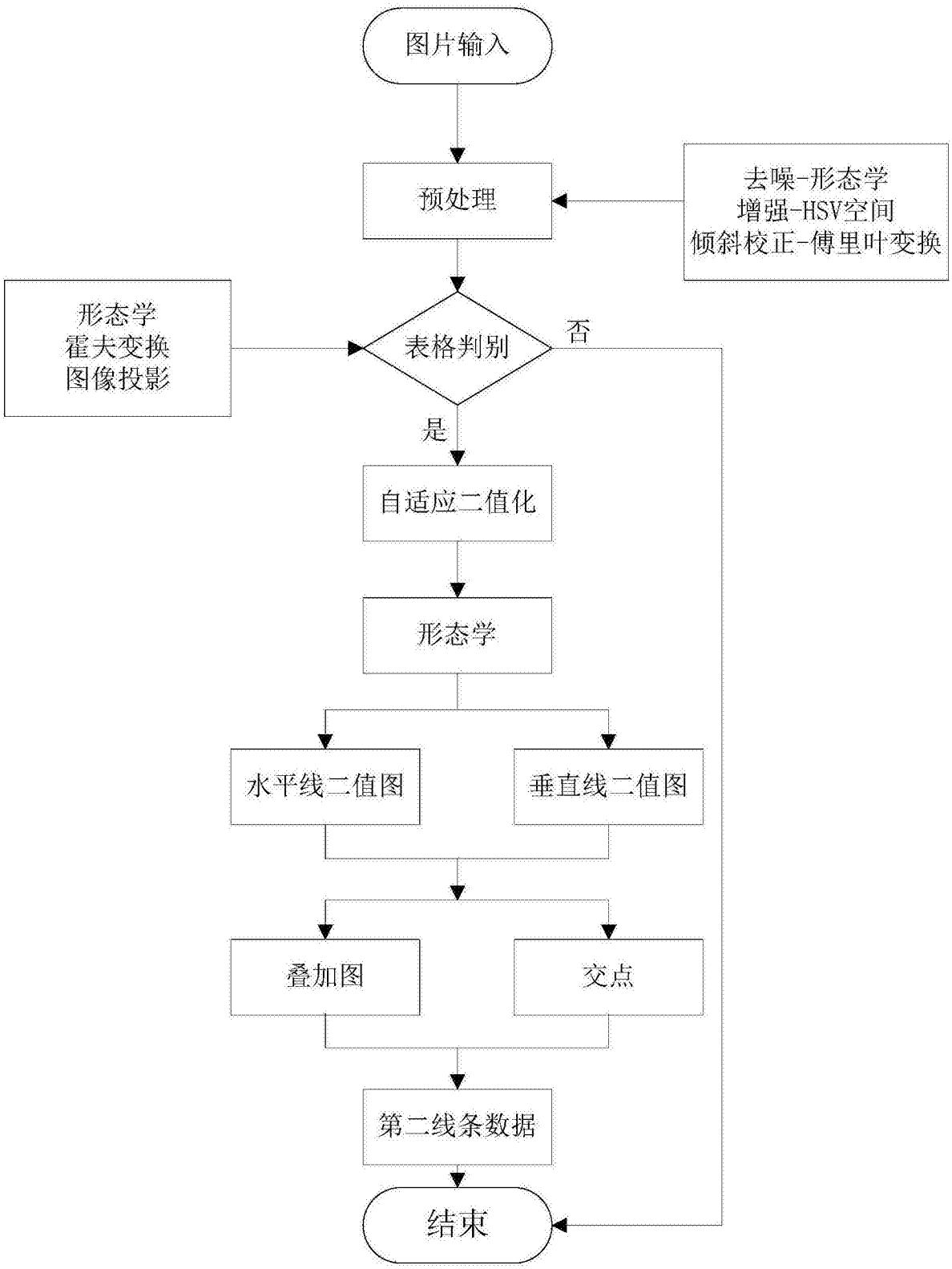

图片输入
预处理
去噪-形态学
增强-HSV空间
倾斜校正-傅里叶变换
形态学
霍夫变换
图像投影
表格判别
否
是
自适应二值化
形态学
水平线二值图
垂直线二值图
叠加图
交点
第二线条数据
结束

图3

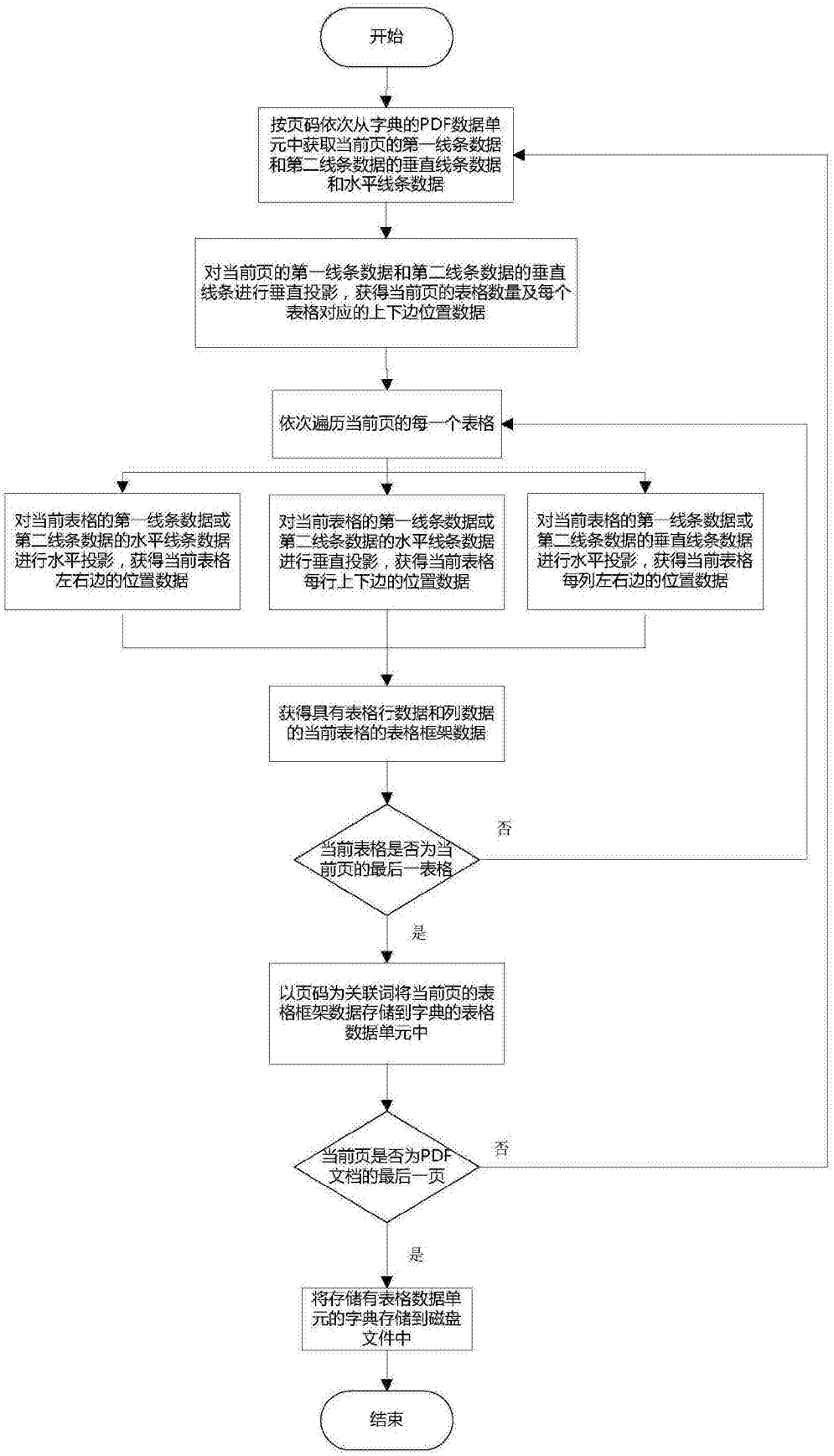

开始
按页码依次从字典的PDF数据单元中获取当前页的第一线条数据和第二线条数据的垂直线条数据和水平线条数据
对当前页的第一线条数据和第二线条数据的垂直线条进行垂直投影，获得当前页的表格数量及每个表格对应的上下边位置数据
依次遍历当前页的每一个表格
对当前表格的第一线条数据或第二线条数据的水平线条数据进行水平投影，获得当前表格左右边的位置数据
对当前表格的第一线条数据或第二线条数据的水平线条数据进行垂直投影，获得当前表格每行上下边的位置数据
对当前表格的第一线条数据或第二线条数据的垂直线条数据进行水平投影，获得当前表格每列左右边的位置数据
获得具有表格行数据和列数据的当前表格的表格框架数据
当前表格是否为当前页的最后一表格
否
是
以页码为关联词将当前页的表格框架数据存储到字典的表格数据单元中
当前页是否为PDF文档的最后一页
否
是
将存储有表格数据单元的字典存储到磁盘文件中
结束

图4

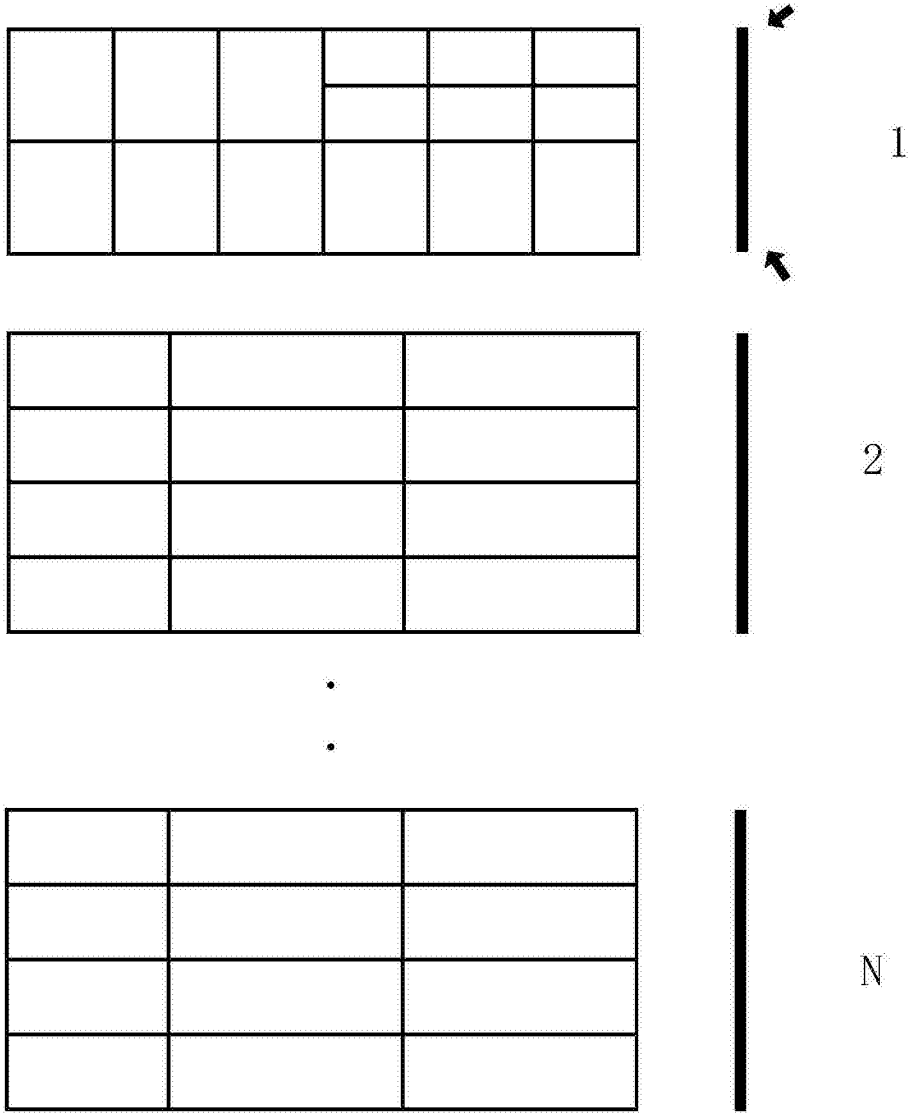


图5

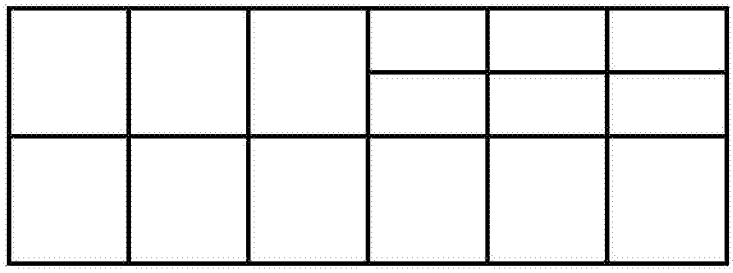


图6

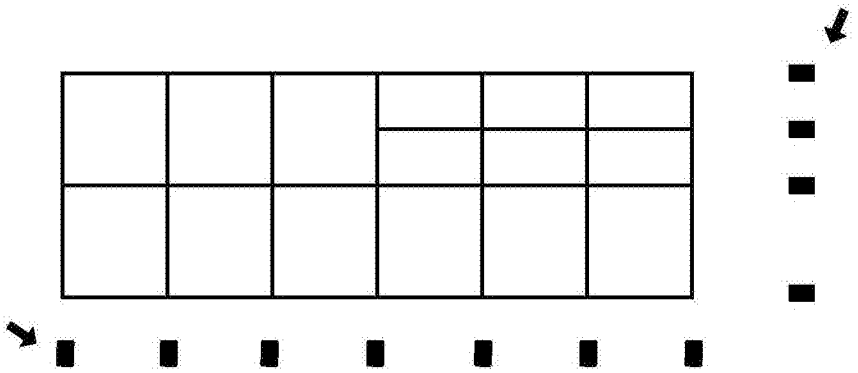


图7

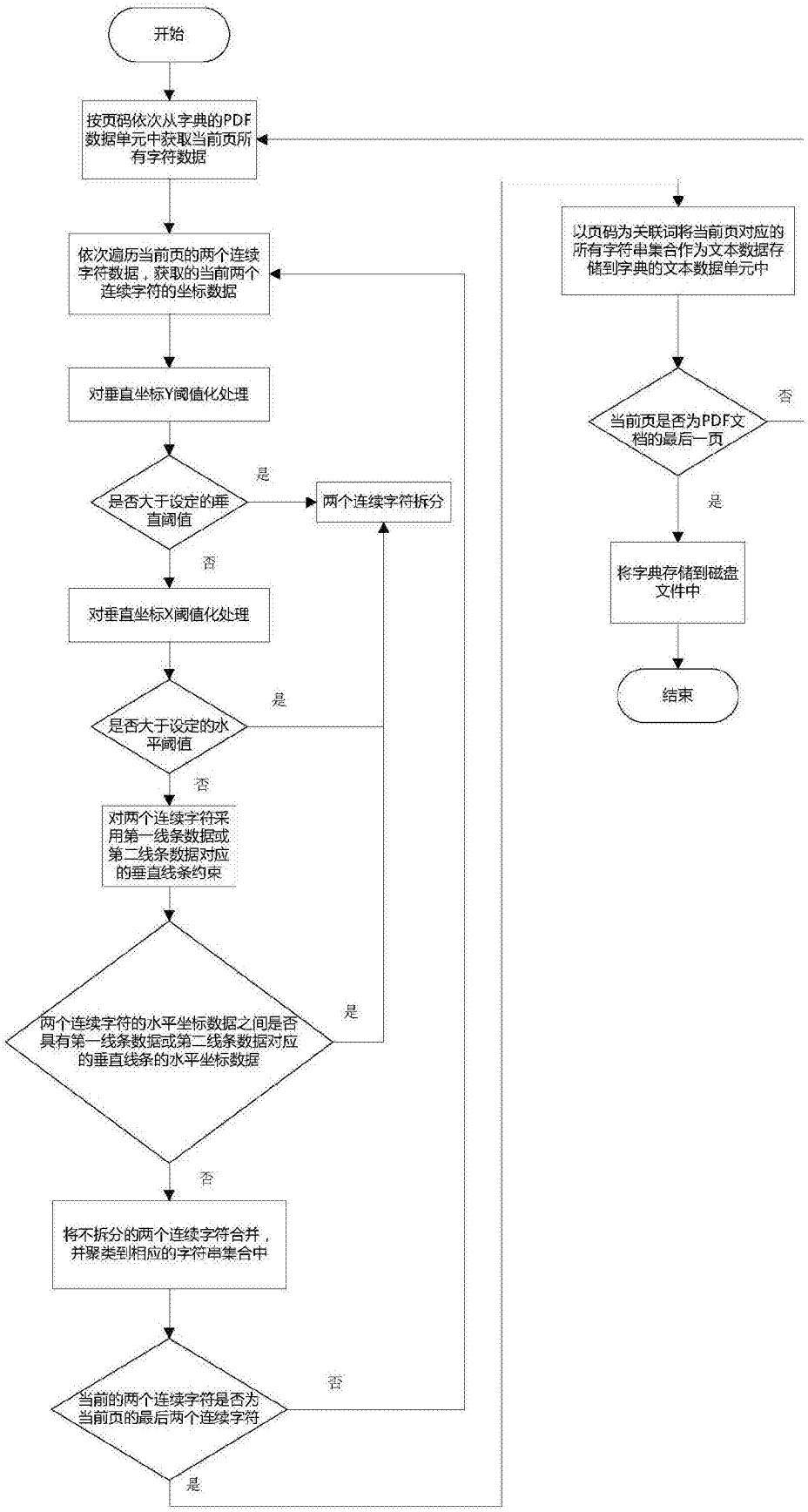

开始
按页码依次从字典的PDF数据单元中获取当前页所有字符数据
依次遍历当前页的两个连续字符数据，获取的当前两个连续字符的坐标数据
对垂直坐标Y阈值化处理
是否大于设定的垂直阈值
是
两个连续字符拆分
否
对垂直坐标X阈值化处理
是否大于设定的水平阈值
是
否
对两个连续字符采用第一线条数据或第二线条数据对应的垂直线条约束
两个连续字符的水平坐标数据之间是否具有第一线条数据或第二线条数据对应的垂直线条的水平坐标数据
是
否
将不拆分的两个连续字符合并，并聚类到相应的字符串集合中
当前的两个连续字符是否为当前页的最后两个连续字符
否
是
以页码为关联词将当前页对应的所有字符串集合作为文本数据存储到字典的文本数据单元中
当前页是否为PDF文档的最后一页
否
是
将字典存储到磁盘文件中
结束

图8

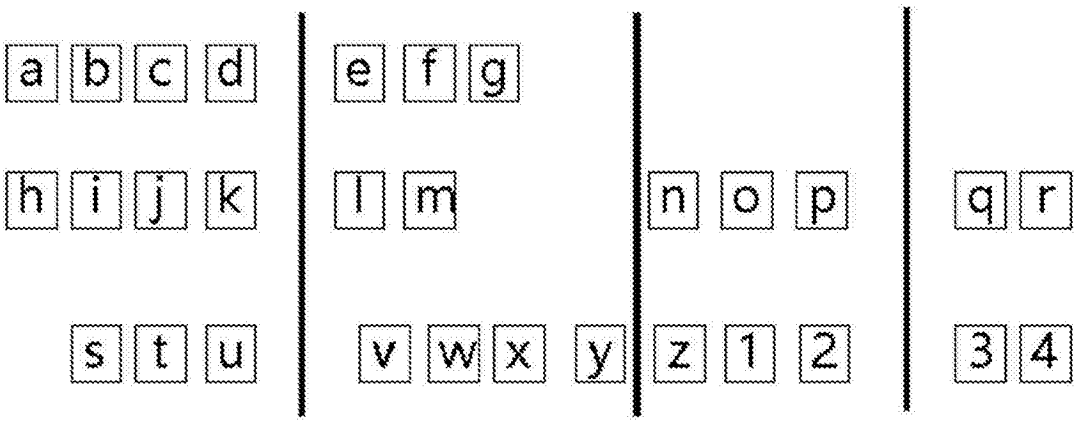


图9

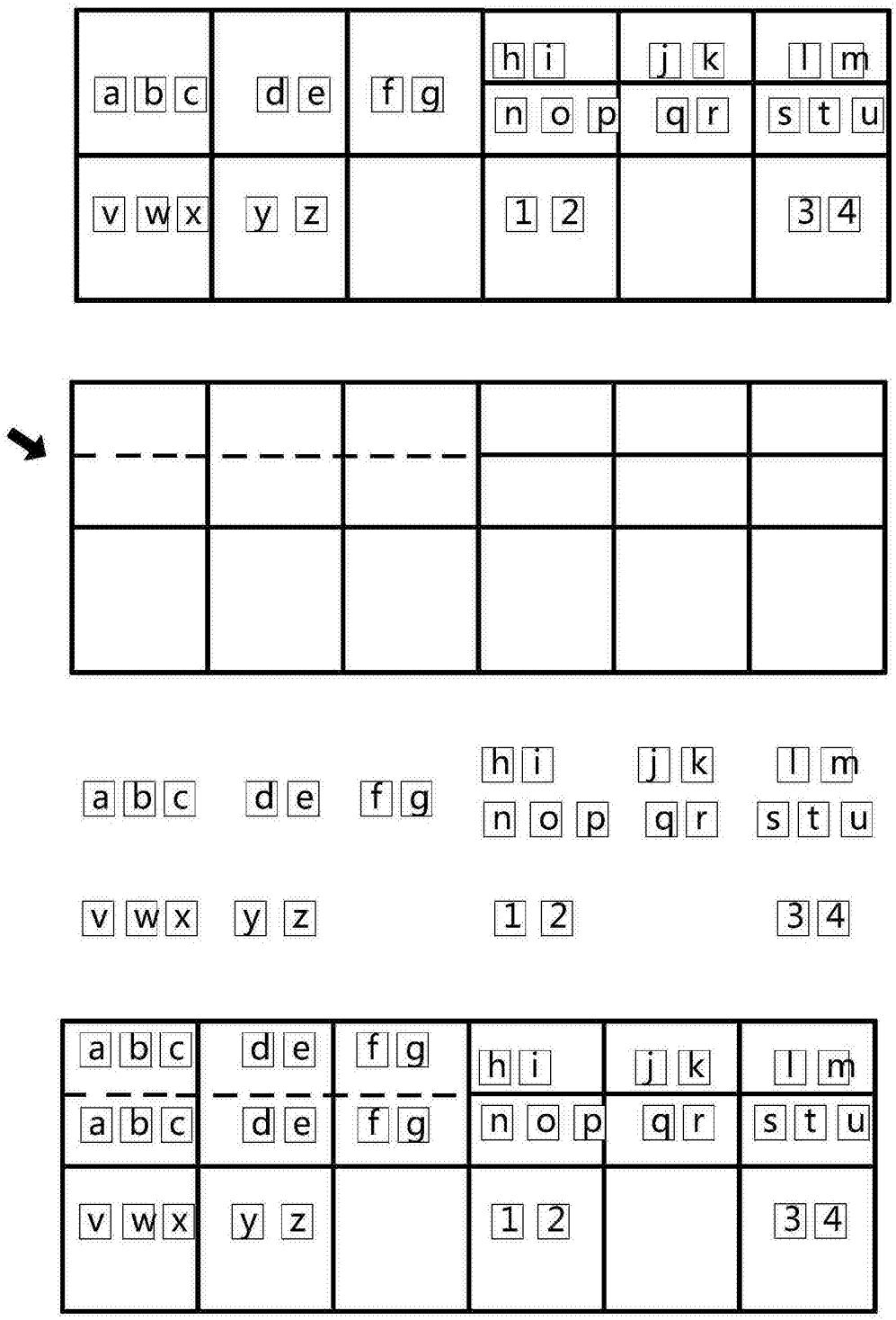


图10