



(12) 发明专利申请

(10) 申请公布号 CN 118786420 A

(43) 申请公布日 2024. 10. 15

(21) 申请号 202380024039.4

(22) 申请日 2023.02.27

(30) 优先权数据

17/685,328 2022.03.02 US

(85) PCT国际申请进入国家阶段日

2024.08.28

(86) PCT国际申请的申请数据

PCT/US2023/013970 2023.02.27

(87) PCT国际申请的公布数据

W02023/167824 EN 2023.09.07

(71) 申请人 奥特瑞克斯股份有限公司

地址 美国加利福尼亚州

(72) 发明人 J·D·拉德 A·布莱克

(74) 专利代理机构 永新专利商标代理有限公司
72002

专利代理师 潘冠伊

(51) Int.Cl.

G06F 16/00 (2006.01)

G06V 30/10 (2006.01)

G06V 30/00 (2006.01)

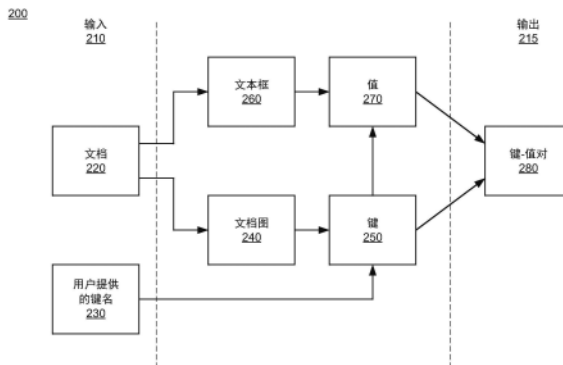
权利要求书3页 说明书11页 附图12页

(54) 发明名称

自动键-值对提取

(57) 摘要

接收要分析的文档和要从文档中提取的键名集合。识别包含在文档内的字符串集合和每个字符串的位置。此外,生成文档的文档图。文档图包括节点集合和边集合。节点集合中的每个节点对应于字符串集合中的字符串。边集合中的每个边将两个或更多个节点连接在一起。另外,基于文档图和所接收的键名集合,识别键集合。此外,从文档中提取值集合,并且基于所识别的键集合和所提取的值集合来生成键-值对集合。



1. 一种计算机实现的方法,包括:
接收要分析的文档;
接收要从所述文档中提取的键名集合;
识别所述文档内包含的多个字符串,以及所述多个字符串中的每个字符串的位置;
生成文档图,所述文档图包括:
节点集合,所述节点集合中的每个节点对应于来自所述多个字符串的字符串,以及
边集合,所述边集合中的每个边将两个或更多个节点连接在一起;
基于所述文档图和所接收的键名集合来识别键集合;
从所述文档中提取值集合;以及
基于所识别的键集合和所提取的值集合来生成键-值对集合。
2. 根据权利要求1所述的计算机实现的方法,其中,识别所述多个字符串包括对所接收的文档执行光学字符识别(OCR)。
3. 根据权利要求1所述的计算机实现的方法,其中,生成所述文档图包括:
响应于确定所述节点集合中的第二节点在距所述节点集合中的第一节点的阈值距离内,利用所述边集合中的边将所述第一节点连接到所述第二节点。
4. 根据权利要求3所述的计算机实现的方法,所述第一节点还是响应于确定所述第二节点在距所述第一节点的阈值角度内而被连接到所述第二节点的。
5. 根据权利要求1所述的计算机实现的方法,其中,基于所述文档图和所接收的键名集合来识别所述键集合包括:
基于键名的第一令牌来从所述文档图的所述节点集合中识别第一节点,对应于所述第一节点的字符串与所述键名的所述第一令牌匹配;
基于所述键名的第二令牌来从所述文档图的所述节点集合中识别第二节点,所述第二节点通过所述文档图的所述边集合中的边连接到所述第一节点,对应于所述第二节点的字符串与所述键名的所述第二令牌匹配。
6. 根据权利要求1所述的计算机实现的方法,其中,从所述文档中提取值集合包括:
对所述文档进行去线以生成去线文档;
将游程平滑算法(RLSA)应用于所述去线文档,以识别所述去线文档中的前景像素集合和背景像素集合;
基于所述去线文档中的所述前景像素集合来生成边界框集合;以及
基于所生成的边界框集合来提取所述值集合。
7. 根据权利要求6所述的计算机实现的方法,其中,基于所生成的边界框集合来提取所述值集合包括:
针对所述边界框集合中的每个边界框,确定所述键集合中位于所述边界框内的键数量;
响应于所述键数量大于一,拆分所述边界框以生成拆分边界框集合,每个拆分边界框具有所述键集合中位于所述拆分边界框内的一个键。
8. 一种存储可执行计算机程序指令的非暂时性计算机可读存储器,所述指令可执行以执行包括以下各项的操作:
接收要分析的文档;

接收要从所述文档中提取的键名集合；

识别所述文档内包含的多个字符串,以及所述多个字符串中的每个字符串的位置；

生成文档图,所述文档图包括：

节点集合,所述节点集合中的每个节点对应于来自所述多个字符串的字符串,以及

边集合,所述边集合中的每个边将两个或更多个节点连接在一起；

基于所述文档图和所接收的键名集合来识别键集合；

从所述文档中提取值集合；以及

基于所识别的键集合和所提取的值集合来生成键-值对集合。

9. 根据权利要求8所述的非暂时性计算机可读存储器,其中,识别所述多个字符串包括对所接收的文档执行光学字符识别(OCR)。

10. 根据权利要求8所述的非暂时性计算机可读存储器,其中,生成所述文档图包括：

响应于确定所述节点集合中的第二节点在距所述节点集合中的第一节点的阈值距离内,利用所述边集合中的边将所述第一节点连接到所述第二节点。

11. 根据权利要求10所述的非暂时性计算机可读存储器,所述第一节点还是响应于确定所述第二节点在距所述第一节点的阈值角度内而被连接到所述第二节点的。

12. 根据权利要求8所述的非暂时性计算机可读存储器,其中,基于所述文档图和所接收的键名集合来识别所述键集合包括：

基于键名的第一令牌来从所述文档图的所述节点集合中识别第一节点,对应于所述第一节点的字符串与所述键名的所述第一令牌匹配；

基于所述键名的第二令牌来从所述文档图的所述节点集合中识别第二节点,所述第二节点通过所述文档图的所述边集合中的边连接到所述第一节点,对应于所述第二节点的字符串与所述键名的所述第二令牌匹配。

13. 根据权利要求8所述的非暂时性计算机可读存储器,其中,从所述文档中提取值集合包括：

对所述文档进行去线以生成去线文档；

将游程平滑算法(RLSA)应用于所述去线文档,以识别所述去线文档中的前景像素集合和背景像素集合；

基于所述去线文档中的所述前景像素集合来生成边界框集合；以及

基于所生成的边界框集合来提取所述值集合。

14. 根据权利要求13所述的非暂时性计算机可读存储器,其中,基于所生成的边界框集合来提取所述值集合包括：

针对所述边界框集合中的每个边界框,确定所述键集合中位于该边界框内的键数量；

响应于所述键数量大于一,拆分所述边界框以生成拆分边界框集合,每个拆分边界框具有所述键集合中位于所述拆分边界框内的一个键。

15. 一种计算机系统,包括：

计算机处理器,其用于执行计算机程序指令；以及

非暂时性计算机可读存储器,其存储可由所述计算机处理器执行以执行操作的计算机程序指令,所述操作包括：

接收要分析的文档；

接收要从所述文档中提取的键名集合；
识别所述文档内包含的多个字符串,以及所述多个字符串中的每个字符串的位置；
生成文档图,所述文档图包括：
节点集合,所述节点集合中的每个节点对应于来自所述多个字符串的字符串,以及
边集合,所述边集合中的每个边将两个或更多个节点连接在一起；
基于所述文档图和所接收的键名集合来识别键集合；
从所述文档中提取值集合；以及
基于所识别的键集合和所提取的值集合来生成键-值对集合。

16. 根据权利要求15所述的计算机系统,其中,识别所述多个字符串包括对所接收的文档执行光学字符识别(OCR)。

17. 根据权利要求15所述的计算机系统,其中,生成所述文档图包括：

响应于确定所述节点集合中的第二节点在距所述节点集合中的第一节节点的阈值距离内以及所述第二节点在距所述第一节节点的阈值角度内,利用所述边集合中的边将所述第一节点连接到所述第二节点。

18. 根据权利要求15所述的计算机系统,其中,基于所述文档图和所接收的键名集合来识别所述键集合包括：

基于键名的第一令牌来从所述文档图的所述节点集合中识别第一节节点,对应于所述第一节节点的字符串与所述键名的所述第一令牌匹配；

基于所述键名的第二令牌来从所述文档图的所述节点集合中识别第二节节点,所述第二节节点通过所述文档图的所述边集合中的边连接到第一节节点,对应于所述第二节节点的字符串与所述键名的所述第二令牌匹配。

19. 根据权利要求15所述的计算机系统,其中,从所述文档中提取值集合包括：

对所述文档进行去线以生成去线文档；

将游程平滑算法(RLSA)应用于所述去线文档,以识别所述去线文档中的前景像素集合和背景像素集合；

基于所述去线文档中的所述前景像素集合来生成边界框集合；以及

基于所生成的边界框集合来提取所述值集合。

20. 根据权利要求19所述的计算机系统,其中,基于所生成的边界框集合来提取所述值集合包括：

针对所述边界框集合中的每个边界框,确定所述键集合中位于该边界框内的键数量；

响应于所述键数量大于一,拆分所述边界框以生成拆分边界框集合,每个拆分边界框具有所述键集合中位于所述拆分边界框内的一个键。

自动键-值对提取

技术领域

[0001] 概括而言,所描述的实施例涉及数据分析,具体而言,涉及从文档中提取键-值对。

背景技术

[0002] 数据分析平台(诸如大数据分析)的增长已经将数据处理扩展到用于处理大量数据以提取具有商业价值的信息的工具中。例如,小型企业可以利用采用专用计算和人力资源的第三方数据分析环境来收集、处理和分析来自各种源(诸如外部数据提供者、内部数据源(例如,本地计算机上的文件)、大数据存储和基于云的数据(例如,社交媒体信息))的大量数据。如数据分析中所使用的,以提取有用的定量和定性信息的方式处理这样的大数据集通常需要在强大的计算机设备上实现的复杂软件工具。

[0003] 在这样的数据分析平台中使用的数据可以来自文档集合。许多文档包含键和值(例如,税务表单上的“名字”字段)之间的关联。计算机要提取这些键-值对可能相当困难。例如,基本OCR提取可能导致复杂且费力的后处理和文本挖掘。因此,需要一种能够通过仅提供用户希望提取的键的名称来从文档中提取键-值对的工具。

发明内容

[0004] 通过用于从文档中提取键-值对的计算机实现的方法、计算机系统和计算机可读存储介质来解决上述和其它问题。计算机实现的方法的实施例包括接收要分析的文档和要从文档中提取的第一键集合。该方法还包括识别包含在文档内的字符串集合和每个字符串的位置。该方法还包括生成文档图。文档图包括节点集合和边集合。节点集合中的每个节点对应于字符串集合中的字符串。边集合中的每个边将两个或更多个节点连接在一起。该方法还包括基于文档图和所接收的第一键集合来识别第二键集合。此外,该方法包括从文档中提取值集合。最后,该方法包括基于所识别的第二键集合和所提取的值集合来生成键-值对集合。

[0005] 计算机系统的实施例包括用于执行计算机程序指令的计算机处理器。该系统还包括存储计算机程序指令的非暂时性计算机可读存储介质,所述计算机程序指令在被执行时使计算机处理器执行操作。操作包括接收要分析的文档和要从文档中提取的第一键集合。操作还包括识别包含在文档内的字符串集合和每个字符串的位置。操作还包括生成文档图。文档图包括节点集合和边集合。节点集合中的每个节点对应于字符串集合中的字符串。边集合中的每个边将两个或更多个节点连接在一起。操作还包括基于文档图和所接收的第一键集合来识别第二键集合。此外,操作包括从文档中提取值集合。最后,操作包括基于所识别的第二键集合和所提取的值集合来生成键-值对集合。

[0006] 非暂时性计算机可读存储介质的实施例存储可执行的计算机程序指令。指令可执行以执行操作。操作包括接收要分析的文档和要从文档中提取的第一键集合。操作还包括识别包含在文档内的字符串集合和每个字符串的位置。操作还包括生成文档图。文档图包括节点集合和边集合。节点集合中的每个节点对应于字符串集合中的字符串。边集合中的每

个边将两个或更多个节点连接在一起。操作还包括基于文档图和所接收的第一键集合来识别第二键集合。此外,操作包括从文档中提取值集合。最后,操作包括基于所识别的第二键集合和所提取的值集合来生成键-值对集合。

附图说明

- [0007] 图1是示出根据一个或多个实施例的包括数据分析系统的数据分析环境的框图。
- [0008] 图2是示出根据一个或多个实施例的从文档提取键-值对的工作流的框图。
- [0009] 图3是示出根据一个或多个实施例的数据分析应用的框图。
- [0010] 图4A示出了根据一个或多个实施例的示出OCR模块的输出的文档的标记图像。
- [0011] 图4B示出了根据一个或多个实施例的示出文档图的表示的文档的标记图像。
- [0012] 图4C示出了根据一个或多个实施例的示出键识别模块的输出的文档的标记图像。
- [0013] 图4D示出了根据一个或多个实施例的与通过将内核应用于文档的图像而生成的水平线和垂直线相对应的图像。
- [0014] 图4E示出了根据一个或多个实施例的通过从文档的图像中减去与文档中的水平线和垂直线相对应的图像而生成的去线(deline)文档。
- [0015] 图4F示出了根据一个或多个实施例的示出由值识别模块生成的边界框的文档的标记图像。
- [0016] 图4G示出了根据一个或多个实施例的示出具有所识别的键和值的边界框的文档的标记图像。
- [0017] 图5是示出根据一个或多个实施例的用于通过使用分析工作流程来提取键-值对的过程的流程图。
- [0018] 图6是示出根据一个实施例的用作图1的数据分析系统的典型计算机系统的功能视图的高级框图。
- [0019] 附图仅出于说明的目的描绘了各种实施例。本领域技术人员将容易地从以下讨论中认识到,在不脱离本文描述的实施例的原理的情况下,可以采用本文所示的结构和方法的替代实施例。各个附图中相同的附图标记和标号表示相同的元件。

具体实施方式

- [0020] 图1是示出根据一个实施例的包括数据分析系统110的数据分析环境100的框图。环境100还包括经由网络130连接到数据分析系统110的多个数据源120。尽管所示环境100仅包含耦合到多个数据源120的一个数据分析系统110,但是实施例可以具有多个数据分析系统和单个数据源。此外,即使数据源120被示为经由网络130连接到数据分析系统110,一个或多个数据源也可以本地存储在数据分析系统处。替代地,用户可以直接(通过网络130或者通过数据分析系统的本地接口)向数据分析系统提供数据。
- [0021] 数据分析系统110是用于处理大量数据的基于计算机的系统。在一些实施例中,经由网络130从多个数据源120收集、采集或以其它方式访问数据。数据分析系统110可以实现来自广泛的多种多样的数据源的数据处理任务中采用的可扩展软件工具和硬件资源。数据处理任务(也称为任务或数据分析任务)是用于处理数据的计算操作。数据处理任务包括例如访问数据、准备数据、混合数据、打包数据、分析数据、对数据的其它类型的操作或其某

种组合。数据分析系统110可以并行执行多个任务。当数据分析系统110具有可用的计算机资源(例如,存储器)时,数据分析系统110可以从准备好执行的多个任务中挑选要执行的任务。

[0022] 在图1的实施例中,数据分析系统110包括存储器140和数据分析应用150。存储器140存储在数据分析操作中使用的数据并且提供对数据存储区域的访问。在一些实施例中,存储器140包括随机存取存储器(RAM)、硬盘驱动器(HDD)、固态驱动器(SSD)、其它形式的计算机存储器或其某种组合。在一些实施例中,存储器140包括提供对相对小的数据存储区域的相对快速的访问的主存储器140和提供对相对大的数据存储区域的相对慢的访问的辅助存储器150。

[0023] 数据分析应用150是通过使用分析工作流(也称为“工作流”)(诸如图2所示的工作流)来支持数据分析任务的执行的软件应用。工作流是数据分析工具(也称为“工具”)或数据分析模块(也称为“模块”)的序列,数据分析任务通过该序列从发起(例如,数据文件的输入)到完成(例如,任务的结果的输出)进行传递。序列中的每个工具执行特定的处理操作或数据分析功能,诸如输入、输出、准备、联接、预测、空间、调查、解析、变换等。对工作流的执行包括对工作流中的工具的操作,并且生成表示任务的结果的结果项。在一个示例中,结果项是提供任务的结果的视觉表示的图表。

[0024] 数据分析应用150提供促进工作流的创建的环境。例如,数据分析应用150提供由数据分析应用150的GUI支持的视觉工作流环境。视觉工作流环境实现了一组拖放工具,其能够消除对提供软件代码来设计工作流的需要,并且消除了识别要由工作流实现的复杂公式的需要。在另一实施例中,工作流200是在诸如可扩展标记语言(XML)文档之类的文档中创建和描述的。

[0025] 环境还可以促进创建逻辑控制的工作流,诸如图2中所示的工作流。逻辑控制的工作流是包括与一个或多个逻辑控制器集成的工具序列的工作流。逻辑控制器具有逻辑功能并且适于基于输入来确定是否触发(和/或不触发)工具或模块。数据分析应用150使用工具序列来处理数据文件,并且使用逻辑控制器来在数据文件的处理期间控制序列中的工具的操作。在一些实施例中,逻辑控制器与工作流中的至少两个工具相关联:第一工具和在第一工具之后的第二工具。数据分析应用150将记录功能应用于第一工具,以记录在数据文件的处理中在执行第一工具的数据分析功能期间发生的事件。由第一工具生成的日志数据描述由第一工具在其对数据文件的处理期间执行的活动和/或检测到的事件。事件可以是例如在执行工具的数据分析功能期间检测到的错误、在执行工具的数据分析功能期间生成的警告、由工具处理的记录的数量、元数据文件的输出、对工具的数据分析功能的执行的完成等。第一工具还输出分析数据,例如,在元数据文件中。分析数据由第一工具从在数据文件的处理中执行其数据分析功能中生成。

[0026] 数据分析应用150使用由第一工具生成的日志数据和/或分析数据作为逻辑控制器的输入。例如,逻辑控制器可以分析日志数据以检测由日志数据描述的事件并且基于检测到的事件来确定是否触发第二工具。在一些实施例中,逻辑控制器是具有逻辑功能的数据分析工具。逻辑控制器中的工具对日志执行数据分析功能以检测由第一工具记录的事件。逻辑功能基于逻辑控制器中的工具的输出来输出命令。第二工具适于接收命令并且根据命令来处理数据。

[0027] 例如,响应于检测到第一工具的数据分析功能的成功完成,逻辑控制器触发第二工具操作。在一些实施例中,逻辑控制器从第一工具接收元数据文件,并且响应于确定触发第二工具而将元数据文件发送到第二工具。从逻辑控制器接收元数据文件的第二工具处理元数据文件。逻辑控制器可以基于由第一工具生成的日志来选择性地执行工作流中的其它工具。第二工具之后的工具(如果有的话)也操作以执行其数据分析功能并且完成对数据文件的处理。生成表示由工具序列对数据文件进行处理的结果的结果项。数据分析应用150提供结果项以供显示给用户。

[0028] 作为另一示例,逻辑控制器不响应于检测到在第一工具的操作期间发生的错误而触发第二工具的操作。在一些实施例中,响应于确定不触发第二工具的操作,逻辑控制器触发第三工具的操作。第三工具可以不在序列中。第三工具适于生成补充结果项。在一个示例中,补充结果项是指定在第一工具的操作期间发生的事件的消息,基于该事件,逻辑控制器已经确定不触发第二工具的数据分析功能。第三工具基于来自逻辑控制器的事件的信息来生成消息,并且可以将消息发送到与用户相关联的客户端设备。在一些实施例中,数据分析应用150从用户接收解决事件的动作。数据分析应用150可以基于用户的动作来解决事件。响应于解决用户的动作,数据分析应用150可以指示逻辑控制器触发第二工具来完成对数据文件的处理。

[0029] 示例是出于说明的目的来提供的。逻辑控制器可以检测其它类型的记录事件。此外,逻辑控制器可以控制其它工具的执行。例如,除了确定触发第二工具之外,逻辑控制器还可以确定触发(或不触发)一个或多个其它工具。一个或多个其它工具可以例如以与它们被布置在工具序列中的顺序相同的顺序连接到第二工具。类似地,除了确定不触发第二工具之外,逻辑控制器还可以确定触发(或不触发)一个或多个其它工具。此外,除了第一工具之外,逻辑控制器还可以接收由一个或多个其它工具生成的日志作为输入并且使用该输入。

[0030] 由数据分析应用150生成的逻辑控制的工作流是有利的,因为它允许用户监测和控制数据分析过程。通过使用这样的逻辑控制的工作流,数据分析过程可以在否定事件(即,逻辑控制器确定不触发第二工具所基于的事件)之后在第一工具处暂停。否定事件可以使得由第一工具生成的数据不适合用于进一步处理。例如,第一工具可以在错误发生时停止处理数据文件,使得由第一工具生成的数据不完整。因此,数据分析过程的暂停避免浪费计算资源来通过其它工具处理不正确的数据。此外,用户可以将第三工具添加到工作流,使得可以通过补充结果项向它们通知否定事件。这样,用户有机会解决否定事件并恢复数据分析过程。与常规数据分析技术相比,逻辑控制的工作流促进对数据分析过程的用户贡献,并且节省了否则会在处理不正确数据时浪费的计算资源。

[0031] 在一些情况下,数据分析应用150向多个终端用户提供支持对数据分析工具和宏的联网或基于云的访问的软件。作为示例,数据分析应用150支持以仪表板、web应用、移动应用等方式创建工作流。作为另一示例,数据分析应用150允许用户以类似于移动应用商店或其它基于云的服务的方式共享、浏览和消费分析。可以将分析数据、宏和工作流作为例如可以由数据分析系统110的其它用户访问的较小规模和可定制分析应用(即,app)来打包和执行。在一些情况下,对发布的分析应用的访问可以由数据分析系统110管理,即,准许或撤销访问,从而提供访问控制和安全能力。数据分析应用150可以执行与分析app相关联的

功能,诸如创建、部署、发布、迭代、更新等。

[0032] 另外,数据分析应用150可以支持在数据分析中涉及的各个阶段执行的功能,诸如访问、准备、混合、分析和输出分析结果的能力。在一些情况下,数据分析应用150可以访问各种数据源,例如在数据流中检索原始数据。由数据分析应用150收集的数据流可以包括原始数据的多个记录,其中原始数据具有不同的格式和结构。在接收到至少一个数据流之后,数据分析应用150可以执行操作以允许将数据流的记录用作数据分析操作的输入。此外,在记录的统计、定性或定量处理中涉及的分析功能(诸如预测分析(例如,预测建模、聚类、数据调查))可以由数据分析应用150实现。

[0033] 如上所述,数据源120向数据分析系统110提供电子数据。数据源120可以是计算机、数据库、网络API(应用编程接口)或云存储系统。数据源120还可以是可以从另一源检索数据的计算机系统。数据源120可以远离数据分析系统110并且经由网络130提供数据。另外,一些或所有数据源120可以直接耦合到数据分析系统并且在不通过网络130传递数据的情况下提供数据。由数据源120提供的数据通常被组织成数据记录,每个数据记录包括一个或多个值。例如,由数据源提供的数据记录可以包括一系列逗号分隔的值。该数据描述了与使用数据分析系统的企业的相关性信息。例如,来自数据源120的数据可以描述与在网站上可访问的内容和/或与社交媒体应用的基于计算机的交互(例如,点击跟踪数据)。

[0034] 网络130表示数据分析系统110与数据源120之间的通信路径。在一个实施例中,网络130是互联网并且使用标准通信技术和/或协议。在网络130上交换的数据可以使用包括超文本标记语言(HTML)、可扩展标记语言(XML)等的技术和/或格式来表示。在另一实施例中,代替以上描述的技术或者除了以上描述的技术之外,实体可以使用定制和/或专用数据通信技术。

[0035] 图2是示出根据一个实施例的使用数据分析系统110来处理数据而执行的工作流200的框图。工作流200由数据分析系统110的计算机设备执行。然而,在其它实施例中,工作流200被部署到可以经由网络(例如,网络130)通信地连接到数据分析系统110的另一计算机设备。

[0036] 如图所示,图2的工作流200包括接收输入集合210,并且生成和/或呈现输出集合215。在一些实施例中,使用输入工具集合(从数据分析系统110的用户、通过与数据分析系统110交互的另一过程、从存储设备等)接收输入集合。此外,可以通过浏览工具将输出集合呈现给数据分析系统110的用户。

[0037] 数据分析工具110接收文档220和用户提供的键名集合230作为输入。在一些实施例中,以指定的文档格式(例如,便携式文档格式(PDF))接收文档。然后,数据分析工具110基于用户提供的键名集合230来从文档220中提取键-值对集合280。

[0038] 在一些实施例中,为了提取键-值对集合280,数据分析系统110通过分析所接收的文档220来生成文档图240。文档图240可以包括节点集合和边集合。节点集合中的每个节点可以对应于在文档220中识别的字符串。在一些实施例中,每个节点存储其对应的字符串和该字符串在文档220内所位于的关联位置。此外,边集合中的每条边将两个或更多个节点彼此连接。

[0039] 另外,基于所生成的文档图240和用户提供的键名集合230,数据分析系统110识别文档内的键集合250。在一些实施例中,数据分析系统110搜索文档图240以定位文档图240

内的每个用户提供的键名230。数据分析工具可以另外从文档图240提取文档内的每个键250的位置。

[0040] 数据分析系统110识别文档220内的文本块集合260,并且基于所识别的文本块集合260和所识别的键集合250来从文档中提取值集合270。然后数据分析系统110将来自所识别的键集合250的键和来自所识别的值集合270的对应值进行匹配,以生成键-值对集合280。

[0041] 图3是示出根据一个或多个实施例的数据分析应用300的框图。数据分析应用300是图1中的数据分析应用150的实施例。在图3的实施例中,数据分析应用300包括数据接收器模块310、光学字符识别(OCR)模块320、图生成模块330、键识别模块340、值识别模块350和用户界面模块360。本领域技术人员将认识到,其它实施例可以具有与这里描述的组件不同的组件和/或其它组件,并且功能可以以不同的方式分布在组件之间。例如,可以在客户端设备或第三方服务器处执行数据分析应用300的一些或所有功能。

[0042] 数据接收器模块310从数据源(例如,数据源120)接收数据文件,并且将数据文件提供给数据分析应用300中的其它模块。在一个实施例中,数据接收器模块310用作数据分析应用中的模块之间的接口。在另一实施例中,数据接收器模块310不存在,并且数据接收器模块310的功能由数据分析应用的每个模块执行。

[0043] 在一些实施例中,数据接收器模块310被配置为接收要分析的文档220和用于分析文档的参数集合。用于分析文档的参数集合可以包括用于从文档220提取键-值对280的用户提供的键名集合230。在一些实施例中,数据接收模块310被配置为接收具有预定义格式(诸如PDF格式)的文档。数据接收器模块310可以被配置为识别文档220是否具有正确的格式,并且如果文档不具有预定义格式,则可以输出错误。

[0044] OCR模块320分析文档220并且识别包括在文档220中的字符串。在一些实施例中,OCR模块320另外确定所识别的字符串中的每个字符串的位置或坐标。字符串的位置或坐标可以对应于文档220内的字符串周围的边界框的形心。在一些实施例中,OCR模块320接收文档的图像(或对应于文档的页的一个或多个图像)作为输入。OCR模块320可以在增加正确地检测输入图像内包含的字符串的可能性之前预处理输入图像。例如,OCR模块可以将图像转换为预定分辨率(例如,以每英寸300个点(dpi)),将图像转换为灰度,执行自适应高斯阈值化、扩张、侵蚀、中值模糊和简单阈值化。

[0045] 图4A示出了根据一个或多个实施例的示出OCR模块的输出的文档的标记图像。如图4A所示,OCR模块识别包含字符串的文档的区域,并且识别字符串中的字符。在图4A的示例中,每个字符串被框包围。

[0046] 返回参考图3,图生成模块330从OCR模块320的输出生成文档220的文档图240。具体地,由图生成模块330基于OCR模块320的输出而生成的文档图240包括由边集合互连的节点集合。在一些实施例中,图生成模块330为由OCR模块320从文档识别的每个字符串添加节点。在一些实施例中,每个节点存储其对应的字符串和该字符串在文档内所位于的关联位置。

[0047] 在一些实施例中,对于文档图240中的每个节点,图生成模块330识别要连接到该节点的相邻节点集合。如果相邻节点在距节点的阈值距离内,则图生成模型330可以添加边,以将该节点与该相邻节点连接。在一些实施例中,如果相邻节点在距节点的阈值距离内

并且节点的位置和相邻节点的位置在彼此的角度范围内,则模型生成模型330添加边,以将该节点与该相邻节点连接。

[0048] 图4B示出了根据一个或多个实施例的示出图生成模块的输出的文档的标记图像。在图4B的示例中,由OCR模块320在图4A中所示的文档的图像中识别的每个字符串被指派给文档图中的节点421。此外,每个节点421通过边423连接到一个或多个节点。例如,对应于字符串“Work”的节点421A通过边423A连接到对应于字符串“Request”421B的节点421B。响应于确定节点421B在距节点421A的阈值距离内并且在距节点421A的阈值角度内,图生成模块330可以将节点421A连接到节点421B。类似地,节点421A通过边423B连接到对应于字符串“Requesting”的节点421D。然而,节点421A未连接到对应于字符串“NumberEWR-001-1001.”的节点421C。在图4B的示例中,节点421A未连接到节点421C,因为节点421C不在距节点421A的阈值角度内。

[0049] 返回参考图3,键识别模块340从文档220中识别键集合250以形成要从文档中提取的键-值对集合280。键识别模块340接收用户提供的键名集合230,并且遍历文档图240以在文档220内定位用户提供的键名230。

[0050] 在一些实施例中,对于每个用户提供的键名,键识别模块340确定用户提供的键名是否包括多个单词。如果用户提供的键名包括多个单词,则键识别模块将用户提供的键名拆分成单词集合(或令牌集合)。然后,键识别模块340可以通过以下方式来识别与用户提供的键名的单词集合相对应的一个或多个节点链:通过在文档图240内搜索用户提供的键名的第一单词来搜索用户提供的键名,以及遍历文档图以识别用户提供的键名的后续单词。也就是说,键识别模块340可以通过将由OCR模块320识别的每个节点串与用户提供的键名的第一单词进行比较来识别与用户提供的键名的第一单词匹配的节点集合。对于与用户提供的键的第一单词匹配的每个识别的节点,键识别模块340确定通过边连接到所识别的节点的第二节点是否匹配用户提供的键名的第二单词。如果所识别的节点未连接到与用户提供的键名的第二单词匹配的第二节点,则键识别模块340确定所识别的节点不对应于用户提供的键名。替代地,如果所识别的节点具有与用户提供的键名的第二单词匹配的第二节点,则键识别模块340将所识别的节点保持为与用户提供的键名相对应的候选键。如果用户提供的键名包括附加单词,则键识别模块340保持遍历文档图,直到找到了用户提供的键名的每个单词为止。例如,如果用户提供的键名包括第三单词,则键识别模块340确定通过边连接到第二节点的第三节点是否匹配用户提供的键名的第三单词,等等。

[0051] 在一些实施例中,节点链(即,通过边集合连接的节点集合)匹配用户提供的键名,键识别模块340可以将节点链中的节点合并到单个节点中。键识别模块340可以另外确定合并的节点的新位置。然后,键识别模块340可以用合并的节点替换节点链中的节点。

[0052] 在一些实施例中,键识别模块340使用模糊匹配来确定单词是否与文档图中的节点匹配。例如,键识别可以将单词与对应于节点的字符串进行比较,并且可以基于比较来确定Levenshtein比率(LR)。如果LR高于阈值(例如,90%),则键识别模块340确定单词与对应于节点的字符串匹配。替代地,键识别模块340选择具有最高LR值的最高设定数量的节点(例如,前10个节点)用于所考虑的单词。

[0053] 图4C示出了根据一个或多个实施例的示出键识别模块的输出的文档的标记图像。具体地,图4C示出了由键识别模块识别的键集合。由键识别模块识别的每个键被框包围。

[0054] 返回参考图3,值识别模块350从文档220中识别值集合270,以形成要从文档中提取的键-值对集合280。值识别模块350从文档220识别文本块260(即,文本的块),并且将文本块260与由键识别模块340识别的键250进行匹配。然后,值识别模块350从文本块260中提取值。

[0055] 在一些实施例中,值识别模块350对文档进行去线(deline)。例如,值识别模块350从文档的图像(或从对应于文档的每一页的图像)去除水平线和垂直线。为了对文档进行去线,将内核应用于文档的图像以生成对应于水平线和垂直线的图像。在一些实施例中,内核由文档的大小确定。此外,在一些实施例中,应用第一内核以生成对应于水平线的第一图像,并且应用第二内核以生成对应于垂直线的第二图像。从文档的图像中减去与水平线和垂直线相对应的所生成的一个(或多个)图像,以生成文档的去线图像。

[0056] 图4D示出了根据一个或多个实施例的对应于通过将内核应用于文档的图像而生成的水平线和垂直线的图像。图4E示出了根据一个或多个实施例的通过从文档的图像中减去与文档中的水平线和垂直线相对应的图像而生成的去线文档。

[0057] 然后处理文档的去线图像以生成边界框集合。在一些实施例中,将游程平滑算法(RLSA)应用于去线图像以生成边界框。在一些实施例中,RLSA识别前景像素集合(例如,填充或黑色像素)和背景像素集合(例如,空白或白色像素)。RLSA然后基于背景像素与前景像素的接近度来将背景像素转换为前景像素。在一些实施例中,如果背景像素在去线图像的水平轴上的两个方向上在距前景像素设定数量的像素内,则RLSA将背景像素转换为前景像素。

[0058] 在一些实施例中,RLSA转换具有低于阈值极限值的相邻背景像素数量的背景像素。在一些实施例中,RLSA将文档的背景像素表示为零,并且将文档的前景像素表示为一。然后,对于文档的表示中的每个零,如果相邻零数量小于阈值极限值,则RLSA用一替换零。例如,如果阈值极限值是3,则将具有1000111000001的表示的文档的一部分被转换为1111111000001的表示。在一些实施例中,在水平方向上应用RLSA。在其它实施例中,在水平方向和垂直方向两者上应用RLSA。

[0059] 在运行RLSA之后,值识别模块350执行轮廓检测算法以生成包围关联在一起的文本块的边界框。轮廓检测算法识别在应用RLSA之后被分类为前景像素的像素的边界框。在一些实施例中,值识别模块对边界框执行轮廓检测算法以将相邻框合并在一起。例如,值识别模块350可以将由边界框包围的所有像素转换为前景像素,并且重新应用轮廓检测算法以检测在应用第一轮廓检测算法之后被分类为前景像素的像素的边界框。

[0060] 图4F示出了根据一个或多个实施例的示出由值识别模块生成的边界框的文档的标记图像。具体地,图4F示出了由RLSA和边界框集合识别的前景像素集合,其中每个边界框包围前景像素的一个斑点。

[0061] 使用边界框,值识别模块350然后识别由键识别模块340识别的每个键的一个或多个值。在一些实施例中,对于由键识别模块340识别的每个键,值识别模块350通过基于键的位置识别边界框并且提取由所识别的边界框包围的文本来识别一个值。

[0062] 在一些实施例中,对于给定键,值识别模块350识别包围键的边界框。替代地,值识别模块350识别最接近键的边界框。在一些实施例中,值识别模块350遍历文档图,以识别将键与之相关联的边界框。在一些实施例中,值识别模块350确定任何给定的边界框是否是到

多个键的最接近的边界框。如果值识别模块350确定边界框是到多个键的最接近的边界框,则值识别模块350拆分边界框并且将每个键与边界框的一部分相关联。

[0063] 替代地,在一些实施例中,值识别模块350代替处理每个键以识别要与键相关联的边界框,值识别模块350处理每个边界框以识别是否将边界框与一个或多个键相关联。例如,对于每个边界框,值识别模块350可以识别由边界框包围的键。在该实施例中,值识别模块350可以考虑三种不同的场景。第一,边界框可以包含一个键和相关联的值。第二,边界框可以包含多个键和相关联的值。第三,边界框可以包含值而不包含键。

[0064] 如果值识别模块350确定边界框包含一个键,则值识别模块350从边界框中提取值,并且基于键和提取的值来生成键值对。例如,可以通过从由边界框包围的文本中移除键的文本来提取该值。

[0065] 此外,如果值识别模块350确定边界框包含多个键,则值识别模块350基于由边界框包围的键的数量来拆分边界框。在一些实施例中,值识别模块350通过尝试优化键和值之间的边界来拆分边界框。例如,如果边界框包括堆叠在列中的n个键,则值识别模块350将边界框水平地拆分成n个较小的边界框,其中每个边界框包含一个键和一个值。然后,值识别模块350将与边界框包含一个键和关联值的情况相对应的步骤应用于通过拆分包括多个键的边界框获得的每个边界框。

[0066] 另外,如果值识别模块350确定边界框不包含键,则值识别模块使用文档图来将边界框与键相关联。

[0067] 图4G示出了根据一个或多个实施例的示出具有所识别的键和值的边界框的文档的标记图像。在图4G的示例中,文档包括边界框集合471,其包括一个键和一个值。例如,边界框471A包括键“类型”和值“改变”。当处理边界框471A时,值识别模块确定仅一个键被边界框471A包围,并且从边界框中提取值以生成位于边界框内的键的键-值对。例如,值识别模块通过从位于边界框471A内的字符串中移除键来从边界框中提取值。此外,在图4G的示例中,文档包括边界框集合473,其包括值但不包括键。例如,边界框473A包括值“CARBURETOR”,但不包括键。这里,值识别模块可以识别位于最接近该值的键并且将所识别的键与由边界框界定的值相关联。在这种情况下,值识别模块识别键“Name”最接近边界框473A,并且通过将键“Name”与值“CARBURETOR”相关联来生成键-值对。

[0068] 图5是示出根据一个或多个实施例的用于通过使用分析工作流程来提取键-值对的过程的流程图。在一些实施例中,过程由数据分析应用150执行,但是在其它实施例中,过程中的一些或全部操作可以由其它实体执行。在一些实施例中,过程中的操作以不同的顺序执行,并且可以包括不同的和/或附加的步骤。

[0069] 在一些实施例中,数据分析应用150访问510分析工作流程。分析工作流程包括工具的序列。该序列中的每个工具适于执行数据分析功能,诸如输入、输出、准备、联接、预测、空间、调查、解析、变换等。分析工作流程是逻辑控制的工作流,并且还包括一个或多个逻辑控制器。

[0070] 数据分析应用150的数据接收模块310接收510要分析的文档。可以将要分析的文档作为数据文件(例如,由数据分析应用的用户上传)来接收。可以将数据文件作为支持的文档格式(例如,PDF格式)集合中的一种文档格式来接收。此外,数据分析应用150的数据接收模块310接收515要从文档提取的用户提供的键名集合。在一些实施例中,通过数据分析

应用的用户界面模块360提供的图形用户界面从数据分析应用150的用户接收键。

[0071] 数据分析应用150将文档提供给OCR模块320。数据分析应用150的OCR模块320识别520文档内的字符串(或文本串)集合。另外,OCR模块320可以识别每个识别的字符串的位置。在一些实施例中,OCR模块从与文档相对应的图像中提取字符串集合。此外,OCR模块320可以在从文档中提取字符串集合之前对文档或文档的图像执行预处理。

[0072] 数据分析应用150可以将OCR模块320的输出提供给图生成模块330。数据分析应用150的图生成模块330基于由OCR模块320提取的字符串集合来生成525文档的文档图。文档图可以包括节点集合(其中每个节点对应于由OCR模块320识别的字符串集合中的字符串)和边集合(其中每个边将两个或更多个节点连接在一起)。

[0073] 数据分析应用150然后可以将由图生成模块330生成的文档图和用户提供的键名提供给键识别模块340。数据分析应用150的键识别模块340识别530文档中的键集合。键识别模块340可以将每个用户提供的键名与文档图中的一个或多个节点或节点链(即,通过边集合连接的节点集合)匹配。

[0074] 数据分析应用150然后可以将由键识别模块340识别的文档和键提供给值识别模块350。在一些实施例中,数据分析应用150另外向值识别模块350提供文档图。数据分析应用150的值识别模块350识别535文档内的文本块集合。特别地,值识别模块350可以生成边界框集合,其中每个边界框包围文档内的一个或多个字符串。为了生成边界框,值识别模块350可以对文档进行去线,应用RLSA,并且应用轮廓检测算法。

[0075] 此外,数据分析应用150的值识别模块350从所识别的文本块中提取540每个键的值。在一些实施例中,为了提取值,值识别模块350处理所识别的边界框中的每个边界框,并且确定一个或多个键是否位于每个边界框内。如果多于一个键位于边界框内,则值识别模型350拆分边界框,使得仅一个键位于每个所得边界框内。对于仅包括一个键的每个边界框,值识别模块350从由边界框包围的文本中提取要与位于边界框内的键相关联的值。此外,如果键不位于边界框内,则值识别模块确定是否将边界框与键相关联。

[0076] 最后,基于由键识别模块340识别的键和由值识别模块350识别的值,生成545键-值对集合。在一些实施例中,数据分析应用150将所生成的键-值对输出到数据分析应用的用户。替代地,数据分析应用150可以将所生成的键-值对存储在文件中或存储在连接到数据分析应用的数据库中。

[0077] 图6是示出根据一个或多个实施例的用作图1的机器学习服务器110的典型计算机系统600的功能视图的高级框图。

[0078] 所示计算机系统包括耦合到芯片组604的至少一个处理器602。处理器602可以包括同一管芯上的多个处理器核。芯片组604包括存储器控制器集线器620和输入/输出(I/O)控制器集线器622。存储器606和图形适配器612耦合到存储器控制器集线器620,并且显示器618耦合到图形适配器612。存储设备608、键盘610、定点设备614和网络适配器616可以耦合到I/O控制器集线器622。在一些其它实施例中,计算机系统600可以具有附加的、更少的或不同的组件,并且组件可以以不同方式耦合。例如,计算机系统600的实施例可以缺少显示器和/或键盘。另外,在一些实施例中,计算机系统600可以被实例化为安装在机架上的刀片服务器或云服务器实例。

[0079] 存储器606保存由处理器602使用的指令和数据。在一些实施例中,存储器606是随

机存取存储器。存储设备608是非暂时性计算机可读存储介质。存储设备608可以是HDD、SSD或其它类型的非暂时性计算机可读存储介质。由机器学习服务器110处理和分析的数据可以存储在存储器606和/或存储设备608中。

[0080] 定点设备614可以是鼠标、轨迹球或其它类型的定点设备,并且与键盘610结合使用以将数据输入到计算机系统600中。图形适配器612在显示器618上显示图像和其它信息。在一些实施例中,显示器618包括用于接收用户输入和选择的触摸屏能力。网络适配器616将计算机系统600耦合到网络160。

[0081] 计算机系统600适于执行用于提供本文描述的功能的计算机模块。如本文所使用的,术语“模块”是指用于提供指定功能的计算机程序指令和其它逻辑。模块可以在硬件、固件和/或软件中实现。模块可以包括一个或多个过程,和/或仅由过程的一部分提供。模块通常存储在存储设备608上,加载到存储器606中,并且由处理器602执行。

[0082] 组件的特定命名、术语的大写、属性、数据结构或任何其它编程或结构方面不是强制性的或重要的,并且实现所描述的实施例的机制可以具有不同的名称、格式或协议。此外,系统可以经由硬件和软件的组合来实现,如所描述的,或者完全在硬件元件中实现。此外,本文描述的各种系统组件之间的功能的特定划分仅仅是示例性的,而不是强制性的;由单个系统组件执行的功能可以替代地由多个组件执行,并且由多个组件执行的功能可以替代地由单个组件执行。

[0083] 以上描述的一些部分根据对信息的操作的算法和符号表示来呈现特征。这些算法描述和表示是数据处理领域技术人员用来最有效地将其工作的实质传达给本领域技术人员的手段。在功能上或逻辑上描述的这些操作被理解为由计算机程序实现。此外,还证明有时方便地将这些操作布置称为模块或功能名称,而不失一般性。

[0084] 除非特别说明,否则如从以上讨论中显而易见的,应当理解,在整个说明书中,利用诸如“处理”或“计算”或“运算”或“确定”或“显示”等术语的讨论是指计算机系统或类似电子计算设备的动作和过程,其操纵和变换表示为计算机系统存储器或寄存器或其它此类信息存储、传输或显示设备内的物理(电子)量的数据。

[0085] 本文描述的某些实施例包括以算法的形式描述的过程步骤和指令。应当注意,实施例的过程步骤和指令可以体现在软件、固件或硬件中,并且当体现在软件中时,可以被下载以驻留在由实时网络操作系统使用的不同平台上并且从不同平台操作。

[0086] 最后,应当注意,说明书中使用的语言主要出于可读性和指导目的而被选择,并且可能尚未被选择来描绘或限定本发明的主题。因此,实施例的公开旨在是说明性的,而不是限制性的。

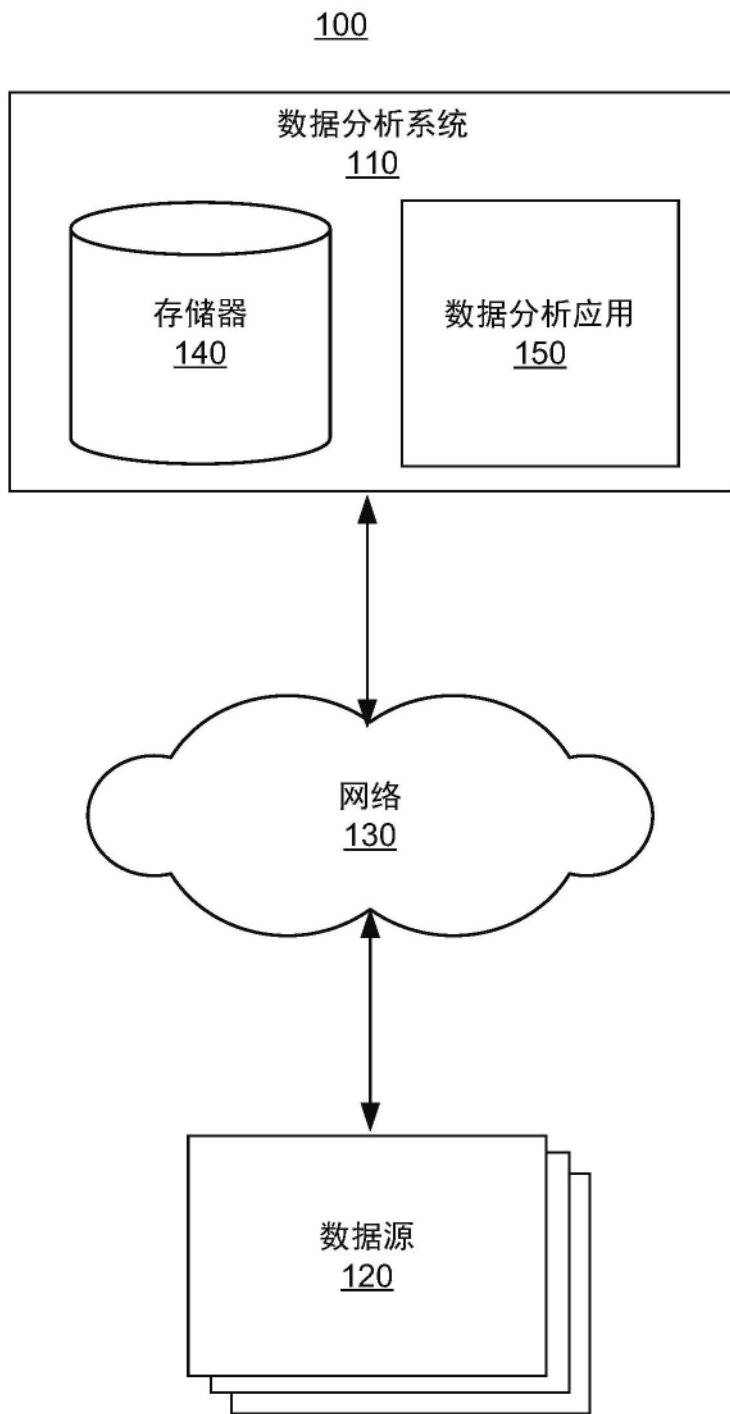


图1

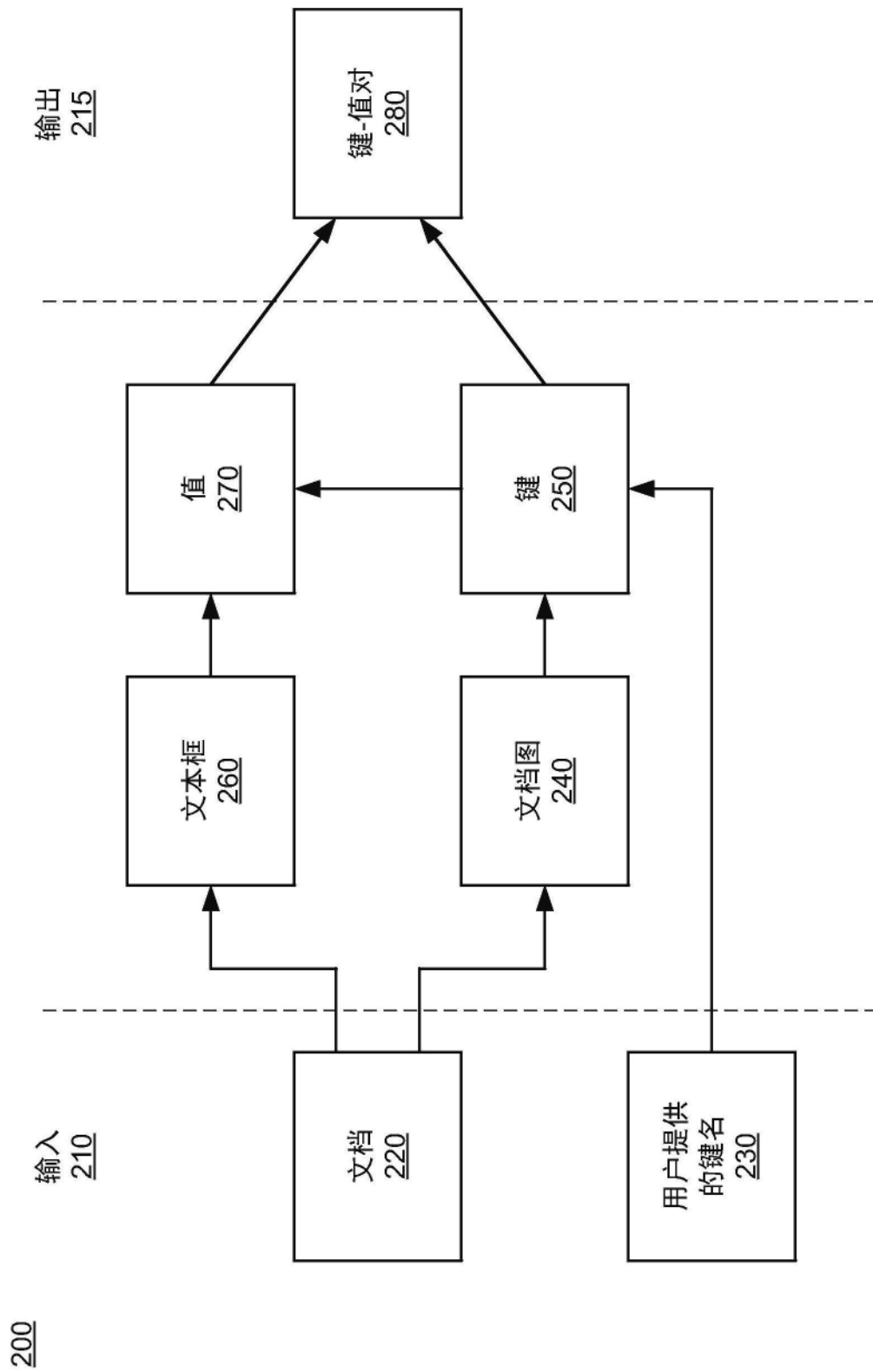


图2

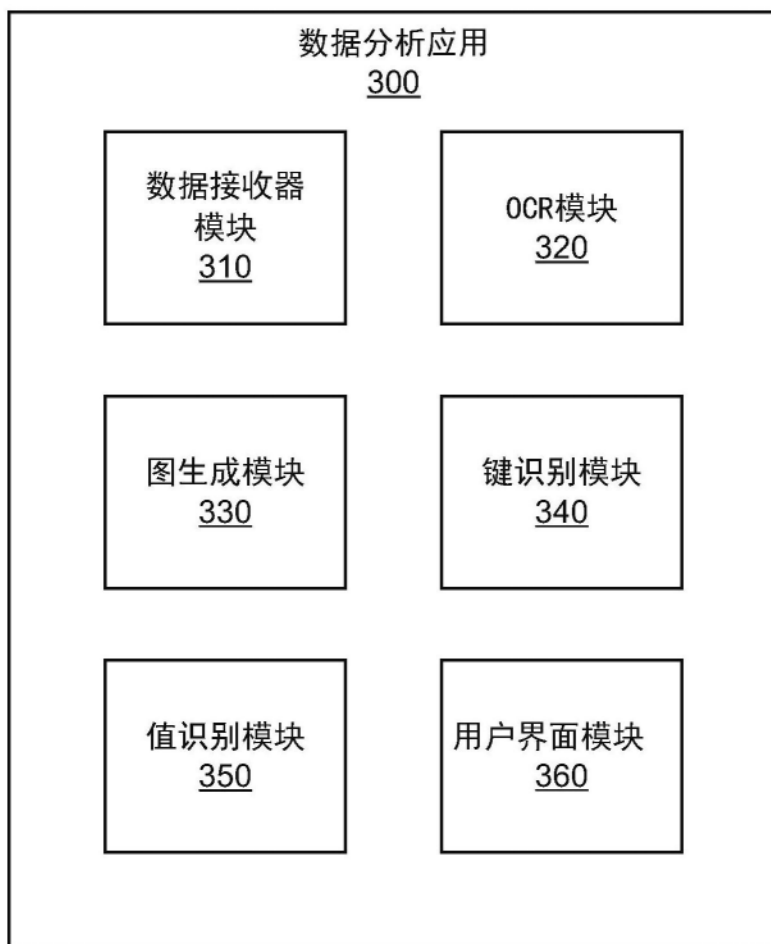


图3

Work Request

W/L

Weber und Less AG

15.12.2011

Date: 15.12.2011

Number EWR 991-1691

Revision: 1

Number EWR-001-1001

Revision: 1

Type: Change

Type: Change

Urgency: processing required

Urgency: processing required

Change kind: Functional Improvement

Change kind: Functional Improvement

Interchangeability: Upward only

Interchangeability: Upward only

Description: We need to reduce the weight of the Carburetor by 10%

Description: We need to reduce the weight of the Carburetor by 10%

Requesting person

Name: Hannah Keller

Name: Hannah Keller

Mail address: Hannah.Keller@weberundless.com

Mail address: Hannah.Keller@weberundless.com

Person in charge

Name: Alexander Koch

Name: Alexander Koch

Mail address: Alexander.Koch@weberundless.com

Mail address: Alexander.Koch@weberundless.com

Comments: Released

Comments: Rejected

Signature:

Signature:

Affected Objects

Type: Type

Type: Type

Number: 10035

Number: 10035

Item: 10035

Item: 10035

Version: 0

Version: 0

Revision: 0

Revision: 0

图4A

Work Request

Number: EWR-001-0001 Revision: 1 Type: Change Date: 15.12.2011

Description: We need to reduce the weight of the carburetor.

Change type: Functional improvement

Urgency: Processing required

Interchangeability: Upward only

Company: Weber und Leese AG

Phone Number: +49-6143-71462-223

Person in charge: Name: Alexander Kesh Mail address: Alexander.Kesh@weberundleese.com

Requesting person: Name: Hannah Kesh Mail address: Hannah.Kesh@weberundleese.com

Comments: Released Rejected

Affected Objects: Type: Number: 10036 Name: CARBURETOR

Version: 1 Revision: 0

图4B

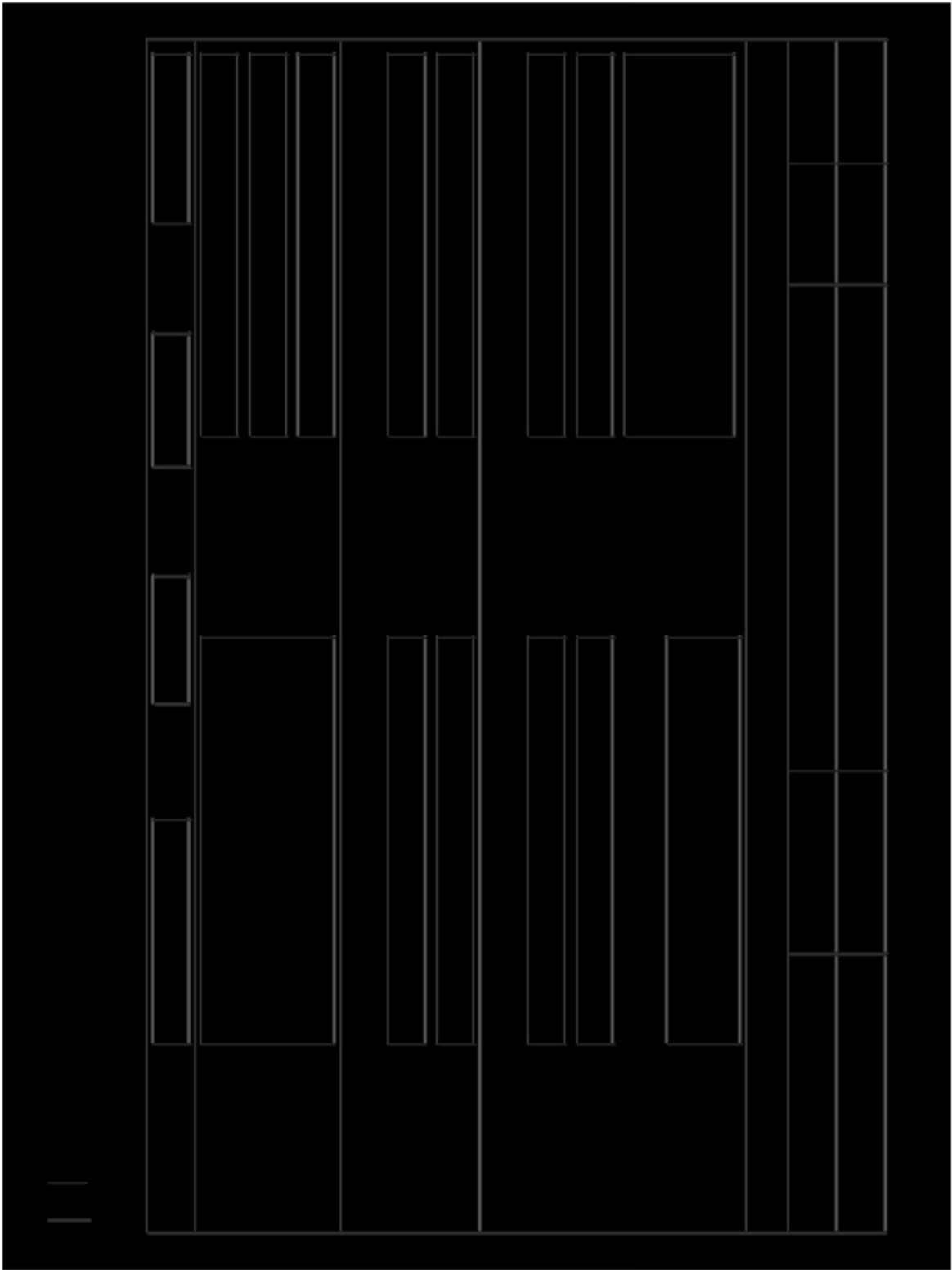


图4D

Work Request

NumberEWR-001-1001

Revision: 1

Description: We need to reduce the weight of the Carburetor by 10%

Type: Change

Date: 15.12.2011

Urgency: processing required

Change kind: Functional improvement

Interchangeability: Upward only

Requesting person

Name: Hannah Keller

Mail address: Hannah.Keller@weberundless.com

Person in charge

Name: Alexander Koch

Mail address: Alexander.Koch@weberundless.com

Comments:

Released

Rejected

Signature:

Company: Weber und Less AG

Phone Number: +49-6143-71462-223

Company: Weber und Less AG

Phone Number: +49-6143-71462-02

Affected Objects

Item	Type	Number	Name	Version	Revision
		10035	CARBURETOR	1	0



Weber und Less AG

图4E

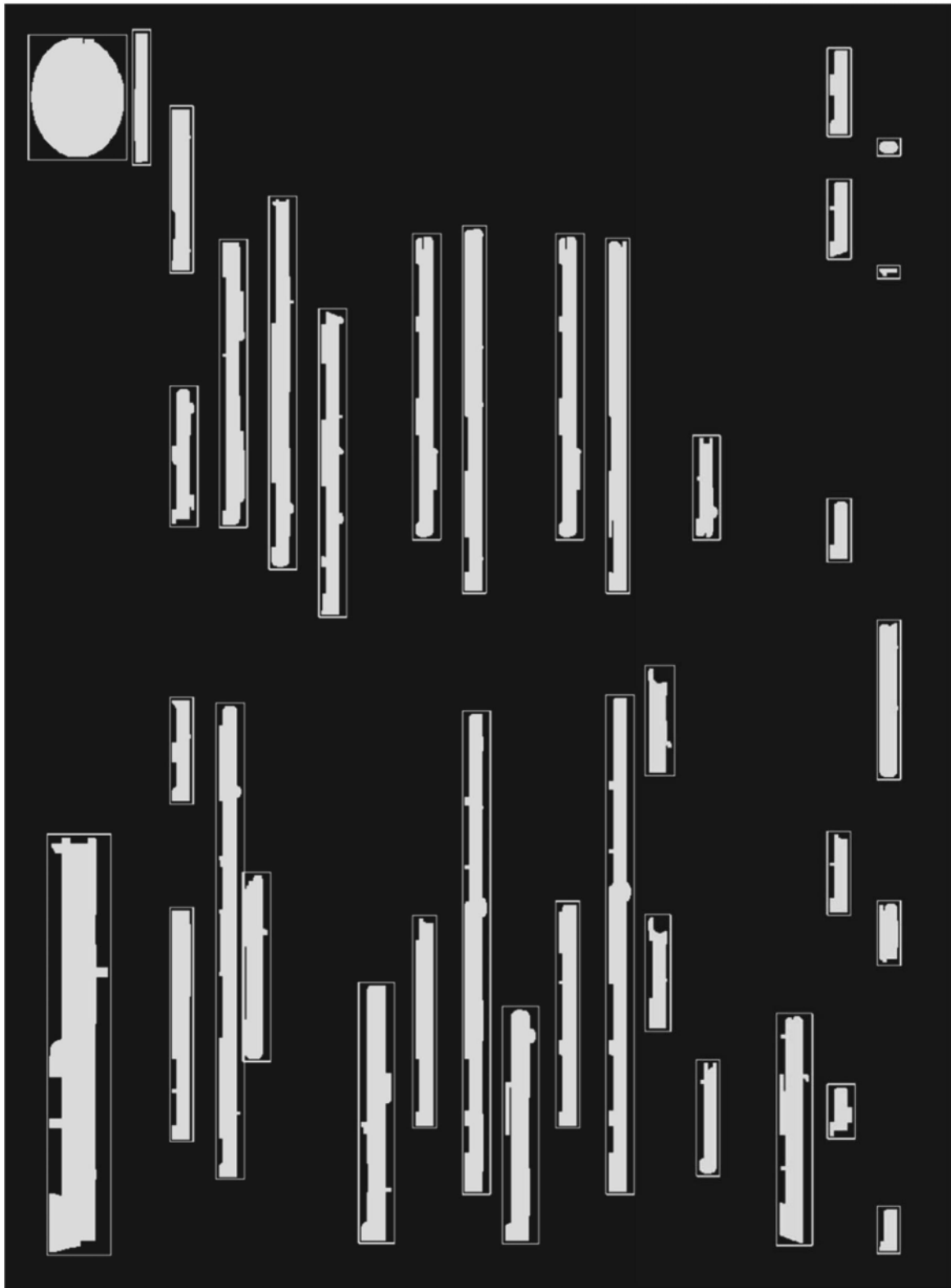


图4F

Work Request

Number: EWR-001-1001 Date: 15.12.2013

Description: We need to reduce the weight of the Carburetor by 10% Type: Change

Urgency: processing required 471A

Change kind: Functional improvement

Interchangeability: Upward only

Requesting person

Name: Hannah Keller Company: Weber und Less AG

Mail address: Hannah.Keller@weberundless.com Phone Number: +49-6143-71462-223

Person in charge

Name: Alexander Koch Company: Weber und Less AG

Mail address: Alexander.Koch@weberundless.com Phone Number: +49-6143-71462-02

Comments: Signature:

Affected Objects

Item	Type	Number	Name	Version	Revision
10035			CARBURETOR	1	0

473B

473A

图4G

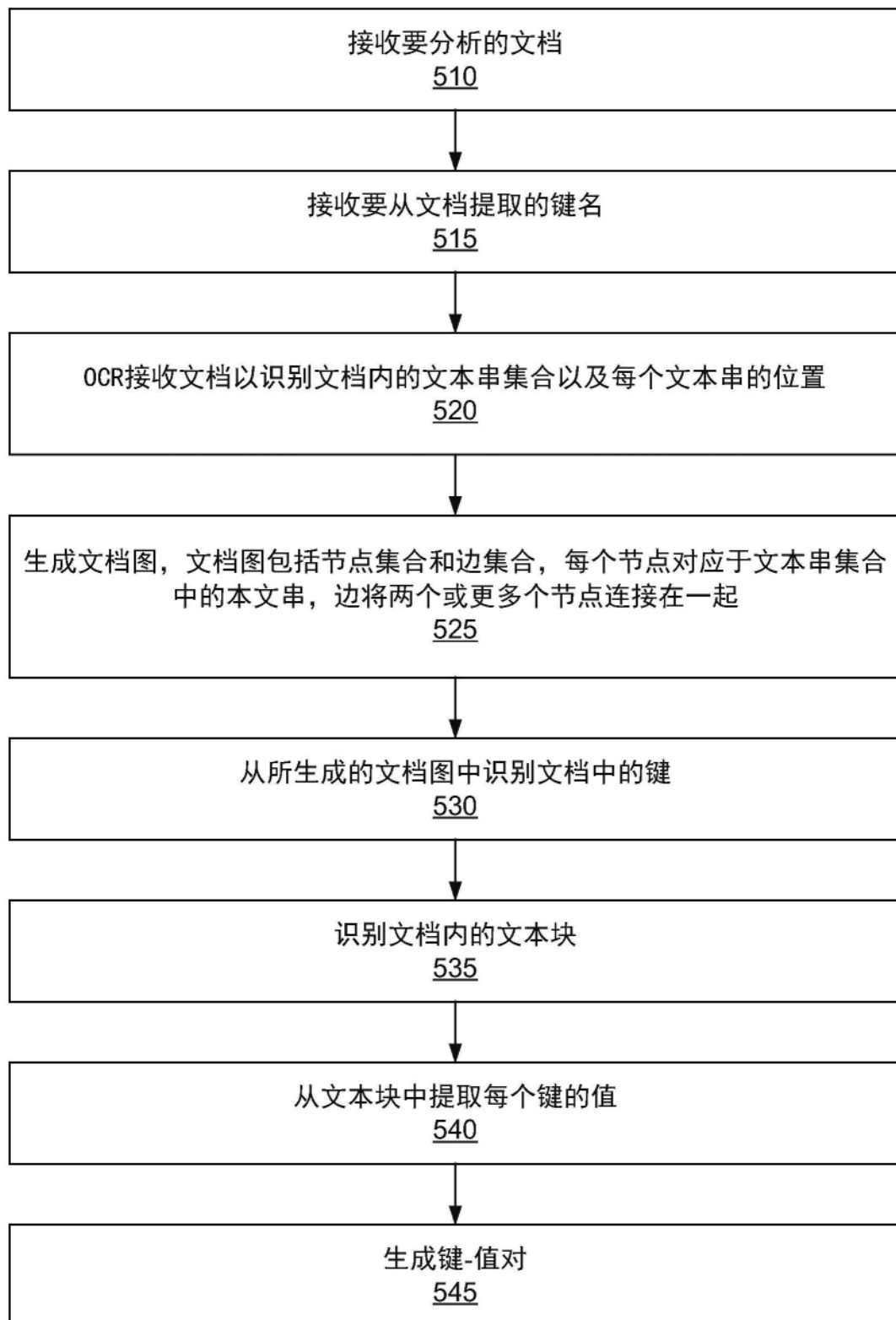


图5

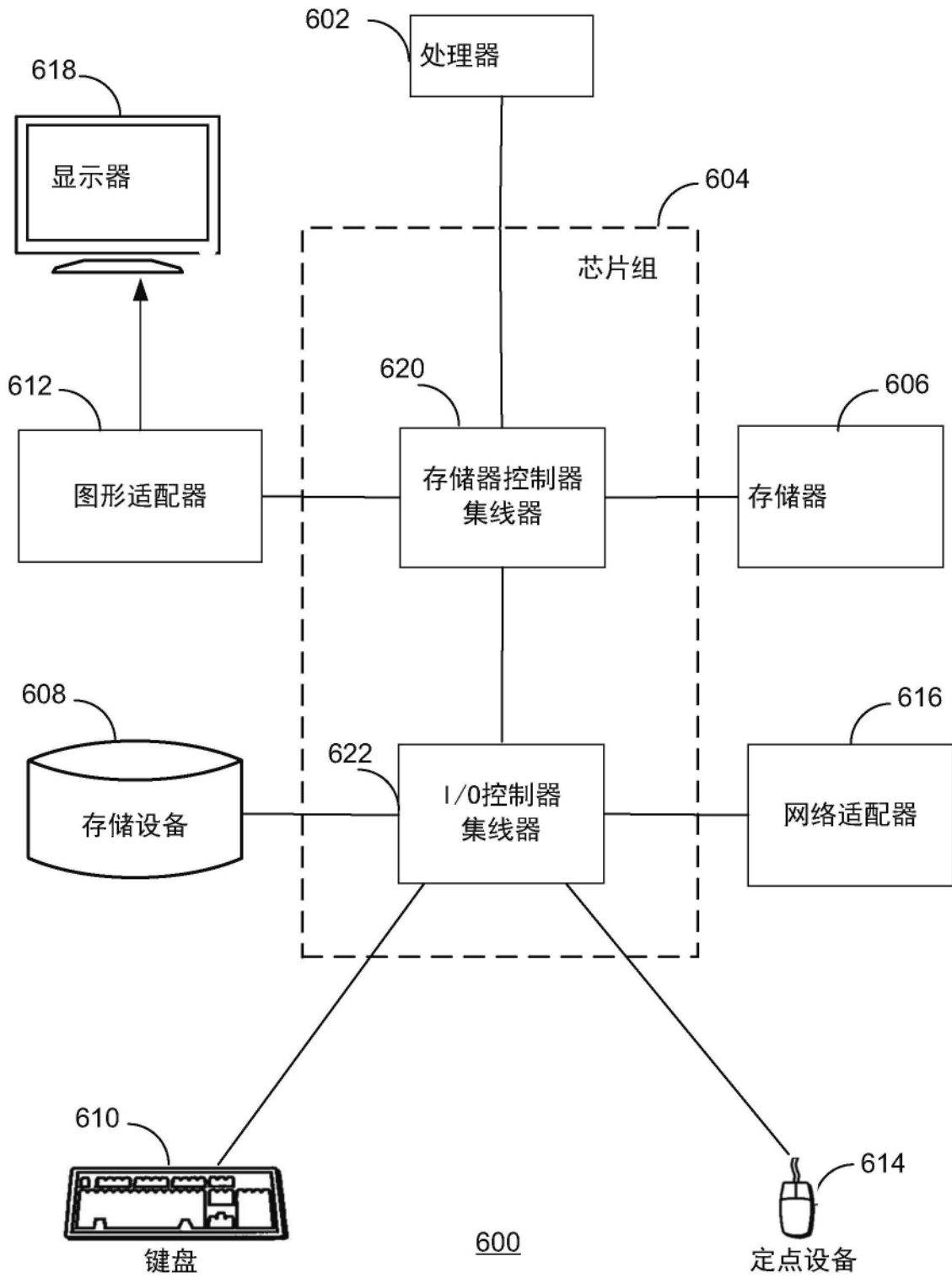


图6