



(21) 申请号 202111139725.5

(22) 申请日 2021.09.27

(65) 同一申请的已公布的文献号

申请公布号 CN 113849552 A

(43) 申请公布日 2021.12.28

(73) 专利权人 中国平安财产保险股份有限公司

地址 518000 广东省深圳市福田区益田路

5033号平安金融中心12、13、38、39、40

层

(72) 发明人 陈少琼

(74) 专利代理机构 深圳市沃德知识产权代理事

务所(普通合伙) 44347

专利代理师 高杰 于志光

(51) Int. Cl.

G06F 16/25 (2019.01)

G06F 16/583 (2019.01)

G06F 16/55 (2019.01)

G06N 3/0464 (2023.01)

G06N 3/045 (2023.01)

(56) 对比文件

CN 110738203 A, 2020.01.31

CN 112016273 A, 2020.12.01

CN 112949443 A, 2021.06.11

WO 2021147252 A1, 2021.07.29

审查员 肖均灵

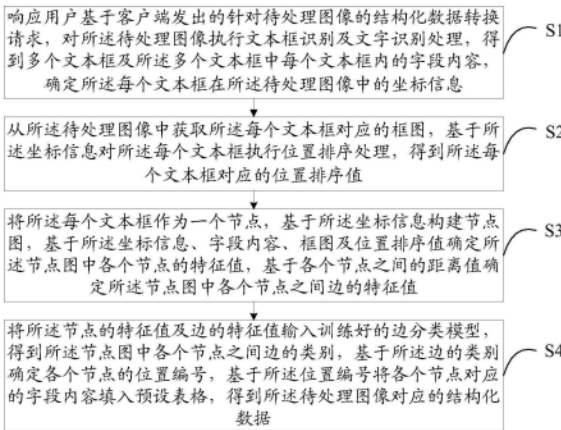
权利要求书2页 说明书12页 附图2页

(54) 发明名称

结构化数据转换方法、装置、电子设备及介
质

(57) 摘要

本发明涉及人工智能领域,揭露一种结构化数据转换方法,包括:识别待处理图像中的文本框及文本框内的字段内容,确定文本框的坐标信息;获取文本框对应的框图,对文本框位置排序,得到位置排序值;构建节点图,基于坐标信息、字段内容、框图及位置排序值确定节点图中各个节点的特征值,基于各个节点之间的距离值确定节点图中各个节点之间边的特征值;将节点的特征值及边的特征值输入训练好的边分类模型,得到节点图中各个节点之间边的类别,基于边的类别确定各个节点的位置编号,基于位置编号将各个节点对应的字段内容填入预设表格,得到结构化数据。本发明还提供一种结构化数据转换装置、电子设备及介质。本发明提高了数据转换准确度。



1. 一种结构化数据转换方法,其特征在于,所述方法包括:

响应用户基于客户端发出的针对待处理图像的结构化数据转换请求,对所述待处理图像执行文本框识别及文字识别处理,得到多个文本框及所述多个文本框中每个文本框内的字段内容,确定所述每个文本框在所述待处理图像中的坐标信息;

从所述待处理图像中获取所述每个文本框对应的框图,基于所述坐标信息对所述每个文本框执行位置排序处理,得到所述每个文本框对应的位置排序值;

将所述每个文本框作为一个节点,基于所述坐标信息构建节点图,基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,基于各个节点之间的距离值确定所述节点图中各个节点之间边的特征值;

将所述节点的特征值及边的特征值输入训练好的边分类模型,得到所述节点图中各个节点之间边的类别,基于所述边的类别确定各个节点的位置编号,基于所述位置编号将各个节点对应的字段内容填入预设表格,得到所述待处理图像对应的结构化数据;

其中,所述基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,包括:从所述节点图中选择一个节点,对选择的节点对应的字段内容执行编码处理及向量转换处理,得到所述选择的节点对应的语义特征;对所述选择的节点对应的框图执行特征提取处理,得到所述选择的节点对应的图像特征;基于所述坐标信息确定所述选择的节点对应的文本框的宽度值及高度值,将所述坐标信息、宽度值及高度值作为所述选择的节点对应的绝对位置特征;将所述位置排序值输入预设函数进行运算,得到所述选择的节点对应的相对位置特征;合并所述语义特征、图像特征、绝对位置特征及相对位置特征,得到所述选择的节点的特征值;

所述边的类别包括同行、同列、跨行、跨列及无关联关系,所述基于所述边的类别确定各个节点的位置编号,包括:将所述节点图中的节点两两组合,得到多个节点对,基于所述边的类别对所述节点对进行划分,得到每个类别对应的节点对集合;将所述节点对集合中属于同一行或同一列的节点作为一个节点组,得到多个行节点组及列节点组;基于所述行节点组及列节点组中节点的坐标信息确定所述节点图中每个节点对应的行号和列号,将所述行号和列号作为所述节点图中各个节点的位置编号;

所述基于所述行节点组及列节点组中节点的坐标信息确定所述节点图中每个节点对应的行号和列号,包括:将具有跨行或跨列关系的节点合并,得到更新后的行节点组及列节点组;将所述节点图中未被分配至更新后的行节点组的每个节点单独作为一个行节点组,得到行节点组集合;将所述节点图中未被分配至更新后的列节点组的每个节点单独作为一个列节点组,得到列节点组集合;计算所述行节点组集合中每个行节点组的节点的横坐标平均值,按照横坐标平均值从小到大的顺序执行第一排序,基于第一排序的结果为所述节点图中每个节点分配行号;计算所述列节点组集合中每个列节点组中节点的纵坐标平均值,按照纵坐标平均值从小到大的顺序执行第二排序,基于第二排序的结果为所述节点图中每个节点分配列号。

2. 如权利要求1所述的结构化数据转换方法,其特征在于,在所述得到多个行节点组及列节点组之后,所述方法还包括:

判断各个列节点组中节点对应的字段内容的数据格式是否相同;

若某一列节点组中节点对应的字段内容的数据格式不同,则拒绝所述结构化数据转换

请求,并向所述客户端发送预警信息。

3.如权利要求1所述的结构化数据转换方法,其特征在于,所述边分类模型的训练过程,包括:

从预设数据库获取框选了文本框并标注了文本框组合对关系类别的样本集,确定所述样本集中每个样本的各个文本框的节点特征及各个文本框组合对的边特征;

将所述节点特征和边特征输入初始边分类模型,得到各个文本框组合对的预测关系类别;

基于标注的文本框组合对关系类别确定各个文本框组合对的真实关系类别,通过最小化预测关系类别与真实关系类别之间的损失值确定所述初始边分类模型的结构参数,得到训练好的边分类模型。

4.如权利要求3所述的结构化数据转换方法,其特征在于,所述损失值的计算公式为:

$$\text{loss}(\mathbf{q}_{mn}, \mathbf{p}_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c \mathbf{p}_{mn} (\log \mathbf{q}_{mn})$$

其中, $\mathbf{q}_{mn}$ 为样本集中第m个样本的第n个文本框组合对的预测关系类别, $\mathbf{p}_{mn}$ 为样本集中第m个样本的第n个文本框组合对的真实关系类别, $\text{loss}(\mathbf{q}_{mn}, \mathbf{p}_{mn})$ 为样本集的预测关系类别和真实关系类别之间的损失值,c为样本集中样本的总数量,t为关系类别的总数量。

5.一种结构化数据转换装置,用于实现如权利要求1至4任一项所述的结构化数据转换方法,其特征在于,所述装置包括:

识别模块,用于响应用户基于客户端发出的针对待处理图像的结构化数据转换请求,对所述待处理图像执行文本框识别及文字识别处理,得到多个文本框及所述多个文本框中每个文本框内的字段内容,确定所述每个文本框在所述待处理图像中的坐标信息;

排序模块,用于从所述待处理图像中获取所述每个文本框对应的框图,基于所述坐标信息对所述每个文本框执行位置排序处理,得到所述每个文本框对应的位置排序值;

构建模块,用于将所述每个文本框作为一个节点,基于所述坐标信息构建节点图,基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,基于各个节点之间的距离值确定所述节点图中各个节点之间边的特征值;

分类模块,用于将所述节点的特征值及边的特征值输入训练好的边分类模型,得到所述节点图中各个节点之间边的类别,基于所述边的类别确定各个节点的位置编号,基于所述位置编号将各个节点对应的字段内容填入预设表格,得到所述待处理图像对应的结构化数据。

6.一种电子设备,其特征在于,所述电子设备包括:

至少一个处理器;以及,

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的结构化数据转换程序,所述结构化数据转换程序被所述至少一个处理器执行,以使所述至少一个处理器能够执行如权利要求1至4中任一项所述的结构化数据转换方法。

7.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质上存储有结构化数据转换程序,所述结构化数据转换程序可被一个或者多个处理器执行,以实现如权利要求1至4任一项所述的结构化数据转换方法。

结构化数据转换方法、装置、电子设备及介质

技术领域

[0001] 本发明涉及人工智能领域,尤其涉及一种结构化数据转换方法、装置、电子设备及介质。

背景技术

[0002] 随着数据时代的到来,数据种类、格式更加多样,为提高数据的处理效率,通常需要将非结构化数据转换为结构化数据。例如,将医疗清单图像中的数据转换为表格数据,以便于后续的核定损。

[0003] 当前,通常在识别出文本图像中的字段内容后,根据预先确定的字段位置关系将字段内容填入至对应的表格位置,然而,这种方式仅适用于格式固定的文本图像,对于格式多样、且无表格线的文本图像,容易出现位置填充错误,导致结构化数据转换准确度不高。因此,亟需一种结构化数据转换方法,以提高数据转换准确度。

发明内容

[0004] 鉴于以上内容,有必要提供一种结构化数据转换方法,旨在提高数据转换准确度。

[0005] 本发明提供的结构化数据转换方法,包括:

[0006] 响应用户基于客户端发出的针对待处理图像的结构化数据转换请求,对所述待处理图像执行文本框识别及文字识别处理,得到多个文本框及所述多个文本框中每个文本框内的字段内容,确定所述每个文本框在所述待处理图像中的坐标信息;

[0007] 从所述待处理图像中获取所述每个文本框对应的框图,基于所述坐标信息对所述每个文本框执行位置排序处理,得到所述每个文本框对应的位置排序值;

[0008] 将所述每个文本框作为一个节点,基于所述坐标信息构建节点图,基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,基于各个节点之间的距离值确定所述节点图中各个节点之间边的特征值;

[0009] 将所述节点的特征值及边的特征值输入训练好的边分类模型,得到所述节点图中各个节点之间边的类别,基于所述边的类别确定各个节点的位置编号,基于所述位置编号将各个节点对应的字段内容填入预设表格,得到所述待处理图像对应的结构化数据。

[0010] 可选的,所述基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,包括:

[0011] 从所述节点图中选择一个节点,对选择的节点对应的字段内容执行编码处理及向量转换处理,得到所述选择的节点对应的语义特征;

[0012] 对所述选择的节点对应的框图执行特征提取处理,得到所述选择的节点对应的图像特征;

[0013] 基于所述坐标信息确定所述选择的节点对应的文本框的宽度值及高度值,将所述坐标信息、宽度值及高度值作为所述选择的节点对应的绝对位置特征;

[0014] 将所述位置排序值输入预设函数进行运算,得到所述选择的节点对应的相对位置

特征;

[0015] 合并所述语义特征、图像特征、绝对位置特征及相对位置特征,得到所述选择的节点的特征值。

[0016] 可选的,所述边的类别包括同行、同列、跨行、跨列及无关联关系,所述基于所述边的类别确定各个节点的位置编号,包括:

[0017] 将所述节点图中的节点两两组合,得到多个节点对,基于所述边的类别对所述节点对进行划分,得到每个类别对应的节点对集合;

[0018] 将所述节点对集合中属于同一行或同一列的节点作为一个节点组,得到多个行节点组及列节点组;

[0019] 基于所述行节点组及列节点组中节点的坐标信息确定所述节点图中每个节点对应的行号和列号,将所述行号和列号作为所述节点图中各个节点的位置编号。

[0020] 可选的,所述基于所述行节点组及列节点组中节点的坐标信息确定所述节点图中每个节点对应的行号和列号,包括:

[0021] 将具有跨行或跨列关系的节点合并,得到更新后的行节点组及列节点组;

[0022] 将所述节点图中未被分配至更新后的行节点组的每个节点单独作为一个行节点组,得到行节点组集合;

[0023] 将所述节点图中未被分配至更新后的列节点组的每个节点单独作为一个列节点组,得到列节点组集合;

[0024] 计算所述行节点组集合中每个行节点组的节点的横坐标平均值,按照横坐标平均值从小到大的顺序执行第一排序,基于第一排序的结果为所述节点图中每个节点分配行号;

[0025] 计算所述列节点组集合中每个列节点组中节点的纵坐标平均值,按照纵坐标平均值从小到大的顺序执行第二排序,基于第二排序的结果为所述节点图中每个节点分配列号。

[0026] 可选的,在所述得到多个行节点组及列节点组之后,所述方法还包括:

[0027] 判断各个列节点组中节点对应的字段内容的数据格式是否相同;

[0028] 若某一列节点组中节点对应的字段内容的数据格式不同,则拒绝所述结构化数据转换请求,并向所述客户端发送预警信息。

[0029] 可选的,所述边分类模型的训练过程,包括:

[0030] 从预设数据库获取框选了文本框并标注了文本框组合对关系类别的样本集,确定所述样本集中每个样本的各个文本框的节点特征及各个文本框组合对的边特征;

[0031] 将所述节点特征和边特征输入初始边分类模型,得到各个文本框组合对的预测关系类别;

[0032] 基于标注的文本框组合对关系类别确定各个文本框组合对的真实关系类别,通过最小化预测关系类别与真实关系类别之间的损失值确定所述初始边分类模型的结构参数,得到训练好的边分类模型。

[0033] 可选的,所述损失值的计算公式为:

$$[0034] \quad \text{loss}(q_{mn}, p_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c p_{mn} (\log q_{mn})$$

[0035] 其中, q_{mn} 为样本集中第 m 个样本的第 n 个文本框组合对的预测关系类别, p_{mn} 为样本集中第 m 个样本的第 n 个文本框组合对的真实关系类别, $\text{loss}(q_{mn}, p_{mn})$ 为样本集的预测关系类别和真实关系类别之间的损失值, c 为样本集中样本的总数量, t 为关系类别的总数量。

[0036] 为了解决上述问题, 本发明还提供一种结构化数据转换装置, 所述装置包括:

[0037] 识别模块, 用于响应用户基于客户端发出的针对待处理图像的结构化数据转换请求, 对所述待处理图像执行文本框识别及文字识别处理, 得到多个文本框及所述多个文本框中每个文本框内的字段内容, 确定所述每个文本框在所述待处理图像中的坐标信息;

[0038] 排序模块, 用于从所述待处理图像中获取所述每个文本框对应的框图, 基于所述坐标信息对所述每个文本框执行位置排序处理, 得到所述每个文本框对应的位置排序值;

[0039] 构建模块, 用于将所述每个文本框作为一个节点, 基于所述坐标信息构建节点图, 基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值, 基于各个节点之间的距离值确定所述节点图中各个节点之间边的特征值;

[0040] 分类模块, 用于将所述节点的特征值及边的特征值输入训练好的边分类模型, 得到所述节点图中各个节点之间边的类别, 基于所述边的类别确定各个节点的位置编号, 基于所述位置编号将各个节点对应的字段内容填入预设表格, 得到所述待处理图像对应的结构化数据。

[0041] 为了解决上述问题, 本发明还提供一种电子设备, 所述电子设备包括:

[0042] 至少一个处理器; 以及,

[0043] 与所述至少一个处理器通信连接的存储器; 其中,

[0044] 所述存储器存储有可被所述至少一个处理器执行的结构化数据转换程序, 所述结构化数据转换程序被所述至少一个处理器执行, 以使所述至少一个处理器能够执行上述结构化数据转换方法。

[0045] 为了解决上述问题, 本发明还提供一种计算机可读存储介质, 所述计算机可读存储介质上存储有结构化数据转换程序, 所述结构化数据转换程序可被一个或者多个处理器执行, 以实现上述结构化数据转换方法。

[0046] 相较现有技术, 本发明首先识别得到待处理图像中的文本框及文本框内的字段内容, 确定文本框的坐标信息; 接着, 获取各个文本框对应的框图, 对文本框位置排序, 得到位置排序值; 然后, 构建节点图, 基于坐标信息、字段内容、框图及位置排序值确定节点图中各个节点的特征值, 基于各个节点之间的距离值确定节点图中各个节点之间边的特征值; 最后, 将节点的特征值及边的特征值输入训练好的边分类模型, 得到节点图中各个节点之间边的类别, 基于边的类别确定各个节点的位置编号, 基于位置编号将各个节点对应的字段内容填入预设表格, 得到待处理图像对应的结构化数据。本发明确定节点的特征值时, 融合了坐标信息对应的绝对位置特征、字段内容对应的语义特征、框图对应的图像特征及位置排序值对应的相对位置特征, 使得节点的特征值较为丰富, 分类准确度更高, 从而结构化数据转换准确度更高。因此, 本发明提高了数据转换准确度。

附图说明

[0047] 图1为本发明一实施例提供的结构化数据转换方法的流程示意图;

[0048] 图2为本发明一实施例提供的结构化数据转换装置的模块示意图;

[0049] 图3为本发明一实施例提供的实现结构化数据转换方法的电子设备的结构示意图;

[0050] 本发明目的的实现、功能特点及优点将结合实施例,参照附图做进一步说明。

具体实施方式

[0051] 本申请实施例可以基于人工智能技术对相关的数据进行获取和处理。其中,人工智能(Artificial Intelligence, AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

[0052] 人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、机器人技术、生物识别技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0053] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅用以解释本发明,并不用于限定本发明。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0054] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅用以解释本发明,并不用于限定本发明。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0055] 需要说明的是,在本发明中涉及“第一”、“第二”等的描述仅用于描述目的,而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此,限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外,各个实施例之间的技术方案可以相互结合,但是必须是以本领域普通技术人员能够实现为基础,当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在,也不在本发明要求的保护范围之内。

[0056] 本发明提供一种结构化数据转换方法。参照图1所示,为本发明一实施例提供的结构化数据转换方法的流程示意图。该方法可以由一个电子设备执行,该电子设备可以由软件和/或硬件实现。

[0057] 本实施例中,结构化数据转换方法包括:

[0058] S1、响应用户基于客户端发出的针对待处理图像的结构化数据转换请求,对所述待处理图像执行文本框识别及文字识别处理,得到多个文本框及所述多个文本框中每个文本框内的字段内容,确定所述每个文本框在所述待处理图像中的坐标信息。

[0059] 所述待处理图像为含有多个字段内容的图像,本方案的目的是将待处理图像中的字段内容以结构化数据形式(例如,表格形式)输出,例如,针对住院费用明细清单图像,一行数据通常表示一个项目明细,一行的项目明细包括项目类别(例如,校验、材料、床位等)、项目名称(例如,钙测定、血清白蛋白测定、输液接头、2人间等)、单位(例如,项、次、日等)、数量、单价、金额等,通过将住院费用明细清单图像中的字段内容以表格形式输出,可方便、

准确的抽取需要的项目明细(即行数据)或分析涉及到的项目类别及项目名称(即列数据)。

[0060] 本实施例中,采用OCR技术对待处理图像执行文本框识别及文字识别处理,以识别到待处理图像中的每个文本框(即待识别区)及每个文本框内的字段内容,OCR技术为现有技术,具体识别过程不再赘述。

[0061] 识别到待处理图像中的各个文本框后,以待处理图像中的预设位置(例如,待处理图像的左下角)为原点建立坐标系(例如,以待处理图像的下边框为x轴、左边框为y轴),以预设数值为刻度单位(例如,以1mm作为1个刻度),根据文本框与坐标轴的距离确定每个文本框的4个顶点的坐标值,得到每个文本框的坐标信息。

[0062] S2、从所述待处理图像中获取所述每个文本框对应的框图,基于所述坐标信息对所述每个文本框执行位置排序处理,得到所述每个文本框对应的位置排序值。

[0063] 将待处理图像中每个文本框所在位置的原始图像作为每个文本框对应的框图。

[0064] 本实施例中,根据预设顺序(例如,从上到下、从左到右的顺序)对各个文本框执行位置排序处理,并基于位置排序结果为每个文本框分配位置排序值,例如,排序第一的文本框的位置排序值可以是001,排序第二的文本框的位置排序值可以是002。

[0065] S3、将所述每个文本框作为一个节点,基于所述坐标信息构建节点图,基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,基于各个节点之间的距离值确定所述节点图中各个节点之间边的特征值。

[0066] 本实施例中,将一个文本框作为一个节点,得到按照坐标信息分布的多个节点,将各个节点连接起来,得到节点图。

[0067] 本方案根据文本框的坐标信息、字段内容、框图及位置排序值确定每个节点的特征值,坐标信息反映了文本框在待处理图像中的绝对位置特征、字段内容反映了文本框对应的语义特征、框图反映了文本框的图像特征、位置排序值反映了文本框的相对位置特征,从而节点的特征值融合了绝对位置特征、语义特征、图像特征及相对位置特征,表征较为丰富。

[0068] 根据节点之间的距离值可确定节点图中各个节点之间边的特征值,本实施例中,将节点对应的文本框的中心点坐标值的差值作为节点之间的距离值。

[0069] 所述基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,包括:

[0070] A11、从所述节点图中选择一个节点,对选择的节点对应的字段内容执行编码处理及向量转换处理,得到所述选择的节点对应的语义特征;

[0071] 编码处理用于将字段内容转换为数值型数据,本实施例中,可采用one hot编码对字段内容执行编码处理。

[0072] 本实施例中,通过将编码后的字段内容输入transformer网络进行向量转换处理,得到选择的节点对应的语义特征。

[0073] A12、对所述选择的节点对应的框图执行特征提取处理,得到所述选择的节点对应的图像特征;

[0074] 本实施例中,通过将选择的节点对应的框图输入卷积神经网络执行特征提取处理,得到选择的节点对应的图像特征(包括颜色、字体、图形、尺寸特征)。

[0075] A13、基于所述坐标信息确定所述选择的节点对应的文本框的宽度值及高度值,将

所述坐标信息、宽度值及高度值作为所述选择的节点对应的绝对位置特征；

[0076] 坐标信息包括文本框的四个顶点的坐标值,通过坐标值可计算得到文本框的宽度值和高度值,将坐标值、宽度值及高度值作为对应节点的绝对位置特征。

[0077] A14、将所述位置排序值输入预设函数进行运算,得到所述选择的节点对应的相对位置特征；

[0078] 本实施例中,所述预设函数为正余弦函数。

[0079] A15、合并所述语义特征、图像特征、绝对位置特征及相对位置特征,得到所述选择的节点的特征值。

[0080] 将上述语义特征、图像特征、绝对位置特征及相对位置特征合并,得到选择的节点的特征值,节点的特征值为一个多元数组。

[0081] 例如,若节点1对应的文本框的四个顶点的坐标值分别为 (x_1, y_1) 、 (x_2, y_2) 、 (x_3, y_3) 、 (x_4, y_4) ,宽度值为 w ,高度值为 h ,语义特征为 p ,图像特征为 q ,相对位置特征为 u ,则节点1的特征值为 $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4, w, h, p, q, u)$ 。

[0082] S4、将所述节点的特征值及边的特征值输入训练好的边分类模型,得到所述节点图中各个节点之间边的类别,基于所述边的类别确定各个节点的位置编号,基于所述位置编号将各个节点对应的字段内容填入预设表格,得到所述待处理图像对应的结构化数据。

[0083] 本实施例中,训练初始边分类模型可得到训练好的边分类模型,其中,初始边分类模型为图神经网络模型,图神经网络模型用于对节点分类和节点之间的边分类,图神经网络包括多层图卷积层,每个图卷积层仅处理一阶邻域信息,通过叠加多个图卷积层实现多阶邻域信息的传递。

[0084] 所述边的类别包括同行、同列、跨行、跨列及无关联关系,根据各个节点之间边的类别可确定节点图中各个节点的位置关系,根据该位置关系可确定各个节点在表格中的行号及列号,将各个节点对应的字段内容填入表格中行号及列号对应的位置,得到待处理图像对应的结构化数据。

[0085] 所述边分类模型的训练过程,包括:

[0086] B11、从预设数据库获取框选了文本框并标注了文本框组合对关系类别的样本集,确定所述样本集中每个样本的各个文本框的节点特征及各个文本框组合对的边特征；

[0087] 本实施例中,样本集中的样本为框选了文本框并标注了文本框组合对关系类别的图像,文本框组合对是由图像中文本框两两组合得到的,例如,若样本1中共10个文本框,则组合后得到 $C_{10}^2/2=45$ 个文本框组合对(因无方向,故需除以2)。

[0088] 文本框组合对关系类别体现了文本框组合对中两个文本框之间的关系类别,所述关系类别与节点之间边的类别相同,包括同行、同列、跨行、跨列及无关联关系。

[0089] 节点特征及边特征与步骤S2中节点的特征值及边的特征值的确定过程相同,在此不再赘述。

[0090] B12、将所述节点特征和边特征输入初始边分类模型,得到各个文本框组合对的预测关系类别；

[0091] 将节点特征和边特征输入图神经网络模型,可输出各个文本框组合对在每个关系类别的预测概率,将预测概率最大的关系类别作为预测关系类别。

[0092] B13、基于标注的文本框组合对关系类别确定各个文本框组合对的真实关系类别,

通过最小化预测关系类别与真实关系类别之间的损失值确定所述初始边分类模型的结构参数,得到训练好的边分类模型。

[0093] 所述损失值的计算公式为:

$$[0094] \quad \text{loss}(q_{mn}, p_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c p_{mn} (\log q_{mn})$$

[0095] 其中, q_{mn} 为样本集中第 m 个样本的第 n 个文本框组合对的预测关系类别, p_{mn} 为样本集中第 m 个样本的第 n 个文本框组合对的真实关系类别, $\text{loss}(q_{mn}, p_{mn})$ 为样本集的预测关系类别和真实关系类别之间的损失值, c 为样本集中样本的总数量, t 为关系类别的总数量。

[0096] 所述基于所述边的类别确定各个节点的位置编号, 包括:

[0097] C11、将所述节点图中的节点两两组合, 得到多个节点对, 基于所述边的类别对所述节点对进行划分, 得到每个类别对应的节点对集合;

[0098] 边的类别即为节点对中两个节点之间的关系类别, 按照边的类别对节点对进行划分, 可得到同行对应的节点对集合、同列对应的节点对集合、跨行对应的节点对集合、跨列对应的节点对集合及无关联关系的节点对集合。

[0099] C12、将所述节点对集合中属于同一行或同一列的节点作为一个节点组, 得到多个行节点组及列节点组;

[0100] 因仅同行对应的节点对集合及同列对应的节点对集合中的节点对具有同一行或同一列的关系, 故该步骤仅针对同行对应的节点对集合及同列对应的节点对集合进行分组。

[0101] 本实施例中, 采用深度优先搜索树算法将属于同一行或同一列的节点分配到同一组。

[0102] 例如, 若同行对应的节点对集合中包括{(节点1, 节点2), (节点1, 节点3), (节点4, 节点7), (节点5, 节点8), (节点6, 节点9), (节点5, 节点7)}, 将属于同一行的节点作为一个节点组, 则可得到三个行节点组, 分别为(节点1, 节点2, 节点3), (节点4, 节点5, 节点7, 节点8), (节点6, 节点9)。

[0103] C13、基于所述行节点组及列节点组中节点的坐标信息确定所述节点图中每个节点对应的行号和列号, 将所述行号和列号作为所述节点图中各个节点的位置编号。

[0104] 行节点组中的各节点对应的行号相同, 列节点组中各节点对应的列号相同, 根据各节点的坐标信息可为节点图中每个节点分配行号和列号。

[0105] 所述基于所述行节点组及列节点组中节点的坐标信息确定所述节点图中每个节点对应的行号和列号, 包括:

[0106] D11、将具有跨行或跨列关系的节点合并, 得到更新后的行节点组及列节点组;

[0107] 例如, 若节点1与节点10之间边的类别为跨行, 则对节点1所在的行节点组1进行更新, 更新后的行节点组1为{节点1-节点10, 节点2, 节点3}。

[0108] D12、将所述节点图中未被分配至更新后的行节点组的每个节点单独作为一个行节点组, 得到行节点组集合;

[0109] 该步骤得到的行节点组集合中包括节点图中所有的节点。

[0110] D13、将所述节点图中未被分配至更新后的列节点组的每个节点单独作为一个列节点组, 得到列节点组集合;

[0111] 该步骤得到的列节点组集合中包括节点图中所有的节点。

[0112] D14、计算所述行节点组集合中每个行节点组的节点的横坐标平均值,按照横坐标平均值从小到大的顺序执行第一排序,基于第一排序的结果为所述节点图中每个节点分配行号;

[0113] 本实施例中,节点的横坐标为节点对应的文本框的中心点横坐标。

[0114] 若一个行节点组中仅一个节点,则将这个节点的横坐标作为这个行节点组的横坐标平均值。

[0115] 本实施例中,可按照横坐标平均值排序结果依次为行节点组集合中的节点分配行号,例如,若排序结果为行节点组1、行节点组2、行节点组3,则行节点组1中每个节点的行号为1、行节点组2中每个节点的行号为2,该步骤实现了为节点图中每个节点分配行号。

[0116] D15、计算所述列节点组集合中每个列节点组中节点的纵坐标平均值,按照纵坐标平均值从小到大的顺序执行第二排序,基于第二排序的结果为所述节点图中每个节点分配列号。

[0117] 节点的纵坐标为节点对应的文本框的中心点纵坐标。

[0118] 本实施例中,按照纵坐标平均值排序结果依次为列节点组集合中的节点分配列号,例如,若排序结果为列节点组1、列节点组2,则列节点组1中每个节点的列号为1、列节点组2中每个节点的列号为2,该步骤实现了为节点图中每个节点分配列号。

[0119] 在所述得到多个行节点组及列节点组之后,所述方法还包括:

[0120] E11、判断各个列节点组中节点对应的字段内容的数据格式是否相同;

[0121] 正常情况下,一个列节点组中的节点对应的字段内容的数据格式相同,例如,对于数量列,其字段内容皆为自然数。

[0122] E12、若某一列节点组中节点对应的字段内容的数据格式不同,则拒绝所述结构化数据转换请求,并向所述客户端发送预警信息。

[0123] 例如,若一个列节点组中节点对应的字段内容既有自然数,又有文字,则很可能出现了边的类别识别错误,或分组错误,此时,可拒绝所述结构化数据转换请求。

[0124] 由上述实施例可知,本发明提出的结构化数据转换方法,首先,识别得到待处理图像中的文本框及文本框内的字段内容,确定文本框的坐标信息;接着,获取各个文本框对应的框图,对文本框位置排序,得到位置排序值;然后,构建节点图,基于坐标信息、字段内容、框图及位置排序值确定节点图中各个节点的特征值,基于各个节点之间的距离值确定节点图中各个节点之间边的特征值;最后,将节点的特征值及边的特征值输入训练好的边分类模型,得到节点图中各个节点之间边的类别,基于边的类别确定各个节点的位置编号,基于位置编号将各个节点对应的字段内容填入预设表格,得到待处理图像对应的结构化数据。本发明确定节点的特征值时,融合了坐标信息对应的绝对位置特征、字段内容对应的语义特征、框图对应的图像特征及位置排序值对应的相对位置特征,使得节点的特征值较为丰富,分类准确度更高,从而结构化数据转换准确度更高。因此,本发明提高了数据转换准确度。

[0125] 如图2所示,为本发明一实施例提供的结构化数据转换装置的模块示意图。

[0126] 本发明所述结构化数据转换装置100可以安装于电子设备中。根据实现的功能,所述结构化数据转换装置100可以包括识别模块110、排序模块120、构建模块130及分类模块

140。本发明所述模块也可以称之为单元,是指一种能够被电子设备处理器所执行,并且能够完成固定功能的一系列计算机程序段,其存储在电子设备的存储器中。

[0127] 在本实施例中,关于各模块/单元的功能如下:

[0128] 识别模块110,用于响应用户基于客户端发出的针对待处理图像的结构化数据转换请求,对所述待处理图像执行文本框识别及文字识别处理,得到多个文本框及所述多个文本框中每个文本框内的字段内容,确定所述每个文本框在所述待处理图像中的坐标信息。

[0129] 排序模块120,用于从所述待处理图像中获取所述每个文本框对应的框图,基于所述坐标信息对所述每个文本框执行位置排序处理,得到所述每个文本框对应的位置排序值。

[0130] 构建模块130,用于将所述每个文本框作为一个节点,基于所述坐标信息构建节点图,基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,基于各个节点之间的距离值确定所述节点图中各个节点之间边的特征值。

[0131] 所述基于所述坐标信息、字段内容、框图及位置排序值确定所述节点图中各个节点的特征值,包括:

[0132] A21、从所述节点图中选择一个节点,对选择的节点对应的字段内容执行编码处理及向量转换处理,得到所述选择的节点对应的语义特征;

[0133] A22、对所述选择的节点对应的框图执行特征提取处理,得到所述选择的节点对应的图像特征;

[0134] A23、基于所述坐标信息确定所述选择的节点对应的文本框的宽度值及高度值,将所述坐标信息、宽度值及高度值作为所述选择的节点对应的绝对位置特征;

[0135] A24、将所述位置排序值输入预设函数进行运算,得到所述选择的节点对应的相对位置特征;

[0136] A25、合并所述语义特征、图像特征、绝对位置特征及相对位置特征,得到所述选择的节点的特征值。

[0137] 分类模块140,用于将所述节点的特征值及边的特征值输入训练好的边分类模型,得到所述节点图中各个节点之间边的类别,基于所述边的类别确定各个节点的位置编号,基于所述位置编号将各个节点对应的字段内容填入预设表格,得到所述待处理图像对应的结构化数据。

[0138] 所述边分类模型的训练过程,包括:

[0139] B21、从预设数据库获取框选了文本框并标注了文本框组合对关系类别的样本集,确定所述样本集中每个样本的各个文本框的节点特征及各个文本框组合对的边特征;

[0140] B22、将所述节点特征和边特征输入初始边分类模型,得到各个文本框组合对的预测关系类别;

[0141] B23、基于标注的文本框组合对关系类别确定各个文本框组合对的真实关系类别,通过最小化预测关系类别与真实关系类别之间的损失值确定所述初始边分类模型的结构参数,得到训练好的边分类模型。

[0142] 所述损失值的计算公式为:

$$[0143] \quad \text{loss}(q_{mn}, p_{mn}) = -\frac{1}{c} \sum_n^t \sum_m^c p_{mn} (\log q_{mn})$$

[0144] 其中, q_{mn} 为样本集中第 m 个样本的第 n 个文本框组合对的预测关系类别, p_{mn} 为样本集中第 m 个样本的第 n 个文本框组合对的真实关系类别, $\text{loss}(q_{mn}, p_{mn})$ 为样本集的预测关系类别和真实关系类别之间的损失值, c 为样本集中样本的总数量, t 为关系类别的总数量。

[0145] 所述边的类别包括同行、同列、跨行、跨列及无关联关系, 所述基于所述边的类别确定各个节点的位置编号, 包括:

[0146] C21、将所述节点图中的节点两两组合, 得到多个节点对, 基于所述边的类别对所述节点对进行划分, 得到每个类别对应的节点对集合;

[0147] C22、将所述节点对集合中属于同一行或同一列的节点作为一个节点组, 得到多个行节点组及列节点组;

[0148] C23、基于所述行节点组及列节点组中节点的坐标信息确定所述节点图中每个节点对应的行号和列号, 将所述行号和列号作为所述节点图中各个节点的位置编号。

[0149] 所述基于所述行节点组及列节点组中节点的坐标信息确定所述节点图中每个节点对应的行号和列号, 包括:

[0150] D21、将具有跨行或跨列关系的节点合并, 得到更新后的行节点组及列节点组;

[0151] D22、将所述节点图中未被分配至更新后的行节点组的每个节点单独作为一个行节点组, 得到行节点组集合;

[0152] D23、将所述节点图中未被分配至更新后的列节点组的每个节点单独作为一个列节点组, 得到列节点组集合;

[0153] D24、计算所述行节点组集合中每个行节点组的节点的横坐标平均值, 按照横坐标平均值从小到大的顺序执行第一排序, 基于第一排序的结果为所述节点图中每个节点分配行号;

[0154] D25、计算所述列节点组集合中每个列节点组中节点的纵坐标平均值, 按照纵坐标平均值从小到大的顺序执行第二排序, 基于第二排序的结果为所述节点图中每个节点分配列号。

[0155] 在所述得到多个行节点组及列节点组之后, 所述分类模块140还用于:

[0156] E21、判断各个列节点组中节点对应的字段内容的数据格式是否相同;

[0157] E22、若某一列节点组中节点对应的字段内容的数据格式不同, 则拒绝所述结构化数据转换请求, 并向所述客户端发送预警信息。

[0158] 如图3所示, 为本发明一实施例提供的实现结构化数据转换方法的电子设备的结构示意图。

[0159] 所述电子设备1是一种能够按照事先设定或者存储的指令, 自动进行数值计算和/或信息处理的设备。所述电子设备1可以是计算机、也可以是单个网络服务器、多个网络服务器组成的服务器组或者基于云计算的由大量主机或者网络服务器构成的云, 其中云计算是分布式计算的一种, 由一群松散耦合的计算机集组成的一个超级虚拟计算机。

[0160] 在本实施例中, 电子设备1包括, 但不限于, 可通过系统总线相互通信连接的存储器11、处理器12、网络接口13, 该存储器11中存储有结构化数据转换程序10, 所述结构化数据转换程序10可被所述处理器12执行。图3仅示出了具有组件11-13以及结构化数据转换

程序10的电子设备1,本领域技术人员可以理解的是,图3示出的结构并不构成对电子设备1的限定,可以包括比图示更少或者更多的部件,或者组合某些部件,或者不同的部件布置。

[0161] 其中,存储器11包括内存及至少一种类型的可读存储介质。内存为电子设备1的运行提供缓存;可读存储介质可为如闪存、硬盘、多媒体卡、卡型存储器(例如,SD或DX存储器等)、随机访问存储器(RAM)、静态随机访问存储器(SRAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM)、磁性存储器、磁盘、光盘等的非易失性存储介质。在一些实施例中,可读存储介质可以是电子设备1的内部存储单元,例如该电子设备1的硬盘;在另一些实施例中,该非易失性存储介质也可以是电子设备1的外部存储设备,例如电子设备1上配备的插接式硬盘,智能存储卡(Smart Media Card, SMC),安全数字(Secure Digital, SD)卡,闪存卡(Flash Card)等。本实施例中,存储器11的可读存储介质通常用于存储安装于电子设备1的操作系统和各类应用软件,例如存储本发明一实施例中的结构化数据转换程序10的代码等。此外,存储器11还可以用于暂时地存储已经输出或者将要输出的各类数据。

[0162] 处理器12在一些实施例中可以是中央处理器(Central Processing Unit, CPU)、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器12通常用于控制所述电子设备1的总体操作,例如执行与其他设备进行数据交互或者通信相关的控制和处理等。本实施例中,所述处理器12用于运行所述存储器11中存储的程序代码或者处理数据,例如运行结构化数据转换程序10等。

[0163] 网络接口13可包括无线网络接口或有线网络接口,该网络接口13用于在所述电子设备1与客户端(图中未画出)之间建立通信连接。

[0164] 可选的,所述电子设备1还可以包括用户接口,用户接口可以包括显示器(Display)、输入单元比如键盘(Keyboard),可选的用户接口还可以包括标准的有线接口、无线接口。可选的,在一些实施例中,显示器可以是LED显示器、液晶显示器、触控式液晶显示器以及OLED(Organic Light-Emitting Diode,有机发光二极管)触摸器等。其中,显示器也可以适当的称为显示屏或显示单元,用于显示在电子设备1中处理的信息以及用于显示可视化的用户界面。

[0165] 应该了解,所述实施例仅为说明之用,在专利申请范围上并不受此结构的限制。

[0166] 所述电子设备1中的所述存储器11存储的结构化数据转换程序10是多个指令的组合,在所述处理器12中运行时,可以上述结构化数据转换方法。

[0167] 具体地,所述处理器12对上述结构化数据转换程序10的具体实现方法可参考图1对应实施例中相关步骤的描述,在此不赘述。

[0168] 进一步地,所述电子设备1集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读存储介质中。所述计算机可读存储介质可以是非易失性的,也可以是非易失性的。所述计算机可读存储介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM, Read-Only Memory)。

[0169] 所述计算机可读存储介质上存储有结构化数据转换程序10,所述结构化数据转换程序10可被一个或者多个处理器执行,以实现上述结构化数据转换方法。

[0170] 在本发明所提供的几个实施例中,应该理解到,所揭露的设备、装置和方法,可以

通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述模块的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式。

[0171] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的,作为模块显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0172] 另外,在本发明各个实施例中的各功能模块可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用硬件加软件功能模块的形式实现。

[0173] 对于本领域技术人员而言,显然本发明不限于上述示范性实施例的细节,而且在不背离本发明的精神或基本特征的情况下,能够以其他的具体形式实现本发明。

[0174] 因此,无论从哪一点来看,均应将实施例看作是示范性的,而且是非限制性的,本发明的范围由所附权利要求而不是上述说明限定,因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

[0175] 本发明所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain),本质上是一个去中心化的数据库,是一串使用密码学方法相关联产生的数据块,每一个数据块中包含了一批次网络交易的信息,用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

[0176] 此外,显然“包括”一词不排除其他单元或步骤,单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第二等词语用来表示名称,而并不表示任何特定的顺序。

[0177] 最后应说明的是,以上实施例仅用以说明本发明的技术方案而非限制,尽管参照较佳实施例对本发明进行了详细说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或等同替换,而不脱离本发明技术方案的精神和范围。

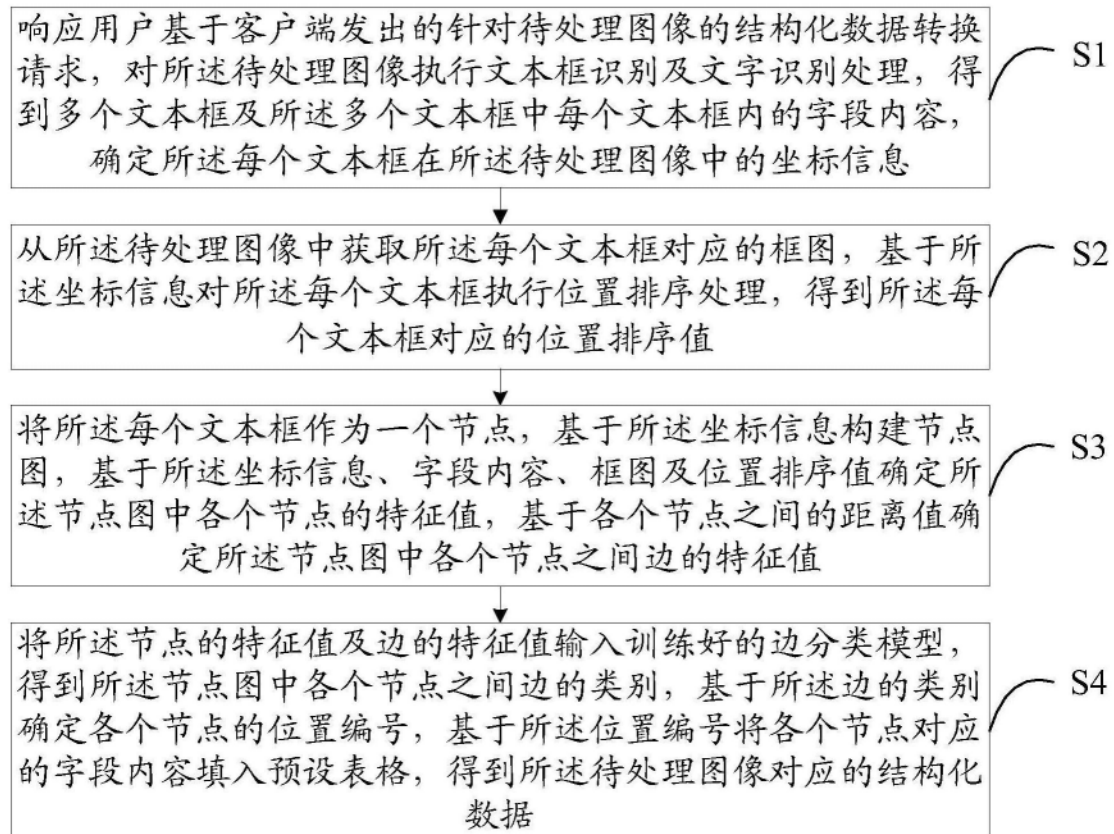


图1



图2

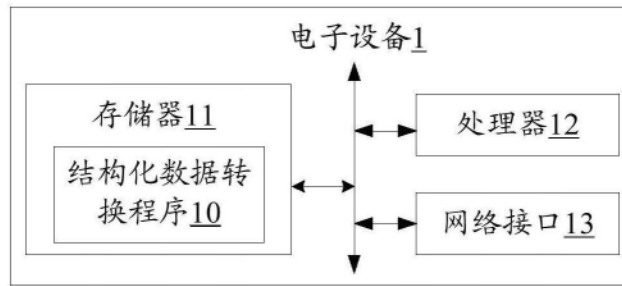


图3