



(12) 发明专利

(10) 授权公告号 CN 112424784 B
(45) 授权公告日 2025. 05. 16

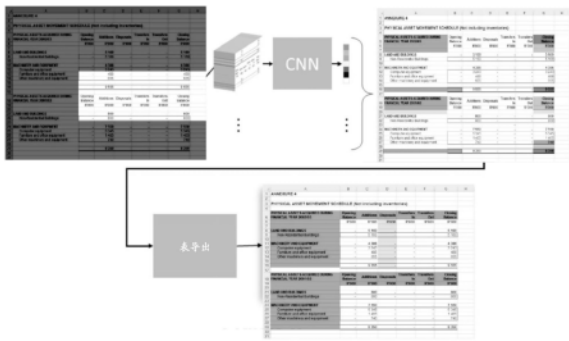
(21) 申请号 201980046843.6	(72) 发明人 B·G·佐恩 M·M·J·布罗克施密特 P·乔杜里 O·波罗佐夫 R·辛格 S·帕德希
(22) 申请日 2019.06.18	
(65) 同一申请的已公布的文献号 申请公布号 CN 112424784 A	
(43) 申请公布日 2021.02.26	(74) 专利代理机构 北京世辉律师事务所 16093 专利代理师 王俊
(30) 优先权数据 16/034,447 2018.07.13 US	(51) Int.Cl. G06F 40/18 (2020.01) G06F 16/338 (2019.01) G06N 20/00 (2019.01) G06N 3/0464 (2023.01) G06N 3/09 (2023.01)
(85) PCT国际申请进入国家阶段日 2021.01.12	
(86) PCT国际申请的申请数据 PCT/US2019/037579 2019.06.18	
(87) PCT国际申请的公布数据 W02020/013956 EN 2020.01.16	(56) 对比文件 W0 2012017056 A1, 2012.02.09
(73) 专利权人 微软技术许可有限责任公司 地址 美国华盛顿州	审查员 卢坤 权利要求书3页 说明书13页 附图10页

(54) 发明名称

用于使用神经网络的改进的表标识的系统、方法和计算机可读介质

(57) 摘要

公开了用于电子表格中的改进的表标识的系统、方法和计算机可读存储设备。一种方法包括：接收包括至少一个表的电子表格；针对所接收的电子表格的每个单元，使用机器学习标识多个类别中的一个或多个类别，其中多个类别包括拐角和非拐角；以及基于所接收的电子表格的每个单元的所标识的一个或多个类别，导出所接收的电子表格中的至少一个表。



1. 一种计算机实现的用于电子表格中的改进的表标识的方法,所述方法包括:
接收包括多个电子表格的数据集,其中所述多个电子表格中的至少一个电子表格包括至少一个表;
处理所述数据集以产生标注数据集,所述标注数据集包括具有标识的表、拐角和标题行的多个标注电子表格;以及
使用所述标注数据集训练神经网络模型,所述神经网络模型被配置为针对接收的电子表格的每个单元,输出多个类别中的一个或多个类别;
接收包括至少一个表的电子表格;
针对所接收的所述电子表格的每个单元,使用所述神经网络模型标识多个类别中的一个或多个类别,其中所述多个类别包括左上角、右上角、左下角、右下角、标题左下角、标题右下角和非拐角,所述标题左下角为在左表边界上的视觉表标题正下方的单元,并且所述标题右下角为在右表边界上的视觉表标题正下方的单元;以及
基于针对所接收的所述电子表格的每个单元所标识的所述一个或多个类别,导出所接收的所述电子表格中的至少一个表。
2. 根据权利要求1所述的方法,还包括:
从所述标注数据集的每个表中提取所述表的拐角单元和所述表的预定数量的非拐角单元;以及
基于每个表的所提取的所述拐角单元和每个表的所述预定数量的非拐角单元,将至少一种正则化技术应用于所述标注数据集。
3. 根据权利要求2所述的方法,其中所述表的预定数量的非拐角单元包括来自所述表中的预定位置的表的非拐角单元。
4. 根据权利要求3所述的方法,其中所述表的非拐角单元的所述预定位置包括以下中的一项或多项:所述表的表中心、所述表的边的中点、所述表的标题和数据之间的间隙、所述表的表拐角周围的单元、以及正态分布在所述表的目标拐角附近的随机单元。
5. 根据权利要求2所述的方法,其中所述至少一种正则化技术包括以下中的一项或多项:i) 基于在验证集上被优化的超参数,针对每个表对多个非拐角单元进行欠采样,ii) 在所述神经网络模型的架构中应用dropout,iii) 应用数据增强,以引入带有合成噪声的附加拐角示例,其中每个合成示例向不在表中的单元添加在拐角单元周围的小随机噪声,以及iv) 重新缩放所述标注数据集的权重,以提供与非拐角单元的总权重相匹配的拐角单元的总权重。
6. 根据权利要求5所述的方法,其中所述合成噪声遵循对称双峰正态分布。
7. 根据权利要求1所述的方法,其中经训练的所述神经网络模型包括卷积神经网络。
8. 根据权利要求1所述的方法,还包括:
构造包括多个神经元的所述神经网络模型,所述神经网络模型被配置为针对所接收的所述电子表格的每个单元输出所述多个类别中的所述一个或多个类别,所述多个神经元被布置在包括至少一个隐藏层的多个层中并且通过多个连接而被连接。
9. 根据权利要求1所述的方法,其中所接收的所述电子表格的每个单元包括预定数量的通道,所述预定数量的通道编码所述单元的视觉特征和其他特征中的一项或多项,以及
其中使用所述神经网络模型针对所接收的所述电子表格的每个单元标识所述多个类

别中的一个或多个类别包括:基于编码所述单元的视觉特征和其他特征中的所述一项或多项的所述预定数量的通道,对所接收的所述电子表格的每个单元分类。

10.根据权利要求9所述的方法,其中对所接收的所述电子表格的每个单元分类还基于所述单元的紧邻周围上下文。

11.根据权利要求1所述的方法,其中基于针对所接收的所述电子表格的每个单元的所标识的所述一个或多个类别导出所接收的所述电子表格中的所述至少一个表包括:使用概率图形模型导出所述至少一个表。

12.根据权利要求11所述的方法,其中使用所述概率图形模型导出所接收的所述电子表格中的所述至少一个表包括:

估计拐角出现在对应边界线的相交处的概率;

计算每个候选表的空间置信度得分;以及

当所述每个候选表的所述空间置信度得分大于预定阈值时,将所述每个候选表选择作为表。

13.根据权利要求12所述的方法,其中所述预定阈值是基于验证数据集而被优化的超参数。

14.根据权利要求1所述的方法,还包括:

接收测试数据集,所述测试数据集包括多个电子表格,每个电子表格具有至少一个表;以及

使用所接收的所述测试数据集评估经训练的所述神经网络。

15.一种用于电子表格中的改进的表标识的系统,所述系统包括:

数据存储设备,存储用于电子表格中的改进的表标识的指令;以及

处理器,被配置为执行所述指令以实施如下方法,所述方法包括:

接收包括多个电子表格的数据集,其中所述多个电子表格中的至少一个电子表格包括至少一个表;

处理所述数据集以产生标注数据集,所述标注数据集包括具有标识的表、拐角和标题行的多个标注电子表格;以及

使用所述标注数据集训练神经网络模型,所述神经网络模型被配置为针对接收的电子表格的每个单元输出多个类别中的一个或多个类别;

接收包括至少一个表的电子表格;

针对所接收的所述电子表格的每个单元,使用所述神经网络模型标识所述多个类别中的一个或多个类别,其中所述多个类别包括左上角、右上角、左下角、右下角、标题左下角、标题右下角和非拐角,所述标题左下角为在左表边界上的视觉表标题正下方的单元,并且所述标题右下角为在右表边界上的视觉表标题正下方的单元;以及

基于针对所接收的所述电子表格的每个单元所标识的所述一个或多个类别,导出所接收的所述电子表格中的至少一个表。

16.根据权利要求15所述的系统,其中针对基于所接收的所述电子表格的每个单元所标识的所述一个或多个类别,导出所接收的所述电子表格中的所述至少一个表包括:使用概率图形模型导出所述至少一个表,并且其中使用所述概率图形模型导出所接收的所述电子表格中的所述至少一个表包括:

估计拐角出现在对应边界线的相交处的概率；

计算每个候选表的空间置信度得分；以及

当所述每个候选表的所述空间置信度得分大于预定阈值时，将所述每个候选表选择作为表。

17. 一种存储指令的计算机可读存储设备，所述指令在由计算机执行时使所述计算机实施用于电子表格中的改进的表标识的方法，所述方法包括：

接收包括多个电子表格的数据集，其中所述多个电子表格中的至少一个电子表格包括至少一个表；

处理所述数据集以产生标注数据集，所述标注数据集包括具有标识的表、拐角和标题行的多个标注电子表格；以及

使用所述标注数据集训练神经网络模型，所述神经网络模型被配置为针对接收的电子表格的每个单元输出多个类别中的一个或多个类别；

接收包括至少一个表的电子表格；

针对所接收的所述电子表格的每个单元，使用所述神经网络模型标识所述多个类别中的一个或多个类别，其中所述多个类别包括左上角、右上角、左下角、右下角、标题左下角、标题右下角和非拐角，所述标题左下角为在左表边界上的视觉表标题正下方的单元，并且所述标题右下角为在右表边界上的视觉表标题正下方的单元；以及

基于针对所接收的所述电子表格的每个单元所标识的所述一个或多个类别，导出所接收的所述电子表格中的至少一个表。

用于使用神经网络的改进的表标识的系统、方法和计算机可读介质

技术领域

[0001] 本公开涉及电子表格 (spreadsheet) 中的表标识。具体地,本公开涉及使用神经网络的电子表格中的改进的表标识。

背景技术

[0002] 电子表格可以实现数据、计算和表示的灵活布置,以便为具有从新手到编程专家的经验的用户提供基于电子表格中的数据进行检查、计算和做出决策的能力。尽管数据的灵活布置提供了很多用途,但数据的灵活布置也可能会妨碍可以从格式良好的表结构得益的自动化工具表。

[0003] 存在自动标识电子表格中的表边界的问题。尽管标识表边界对个人来说可能是直观的,但是由于电子表格中的各种视觉信息和数据,用于标识表边界的基于规则的方法获取的成功很有限。

[0004] 尽管本公开具体地讨论了电子表格中的表标识,但是本公开的各方面不仅可以适用于电子表格,而且还可以适用于其他文件类型,包括表和/或灵活的数据布置。

发明内容

[0005] 根据某些实施例,公开了用于改进的表标识的系统、方法和计算机可读介质。

[0006] 根据某些实施例,公开了一种用于电子表格中的改进的表标识的计算机实现的方法。一种方法包括:接收包括至少一个表的电子表格;使用机器学习针对接收到的电子表格的每个单元标识多个类别中的一个或多个类别,其中多个类别包括拐角和非拐角;以及基于接收到的电子表格的每个单元的一个或多个标识出的类别在接收到的电子表格中导出至少一个表。

[0007] 根据某些实施例,公开了一种用于电子表格中的改进的表标识的系统。一种系统包括:存储用于电子表格中的改进的表标识的指令的数据存储设备;以及被配置为执行所述指令以执行一种方法的处理器,该方法包括:接收包括至少一个表的电子表格;使用机器学习针对接收到的电子表格的每个单元标识多个类别中的一个或多个类别,其中多个类别包括拐角和非拐角;以及基于接收到的电子表格的每个单元的一个或多个标识出的类别在接收到的电子表格中导出至少一个表。

[0008] 根据某些实施例,公开了一种存储指令的计算机可读存储设备,该指令在由计算机执行时引起计算机执行用于电子表格中的改进的表标识的方法。一种计算机可读存储设备的方法包括:接收包括至少一个表的电子表格;使用机器学习针对接收到的电子表格的每个单元标识多个类别中的一个或多个类别,其中多个类别包括拐角和非拐角;以及基于接收到的电子表格的每个单元的一个或多个标识出的类别在接收到的电子表格中导出至少一个表。

[0009] 所公开的实施例的其他目的和优点将在下面的描述中部分地阐述,并且部分地将

从该描述中变得很清楚,或者可以通过所公开的实施例的实践而获知。所公开的实施例的目的和优点将通过所附权利要求中特别指出的要素和组合来实现和获取。

[0010] 应当理解,前面的一般描述和下面的详细描述都只是示例性和说明性的,而不限制所要求保护的公开实施例。

附图说明

[0011] 在下面的详细描述过程中,将参考附图。附图示出了本公开的不同方面,并且在适当的情况下,不同图中示出相似结构、组件、材料和/或元素的附图标记被相似地标记。应当理解,除了具体示出的之外的其他结构、组件和/或元素的各种组合是可以预期的并且在本公开的范围之内。

[0012] 此外,本文中描述和示出了本公开的很多实施例。本公开既不限于任何单个方面或其实施例,也不限于这样的方面和/或实施例的任何组合和/或排列。此外,本公开和/或其实施例的每个方面可以单独使用,或者与本公开和/或其实施例的一个或多个其他方面结合使用。为了简洁起见,某些排列和组合在本文中未单独讨论和/或示出。

[0013] 图1描绘了根据本公开的实施例的具有两个垂直对齐的表的示例性电子表格;

[0014] 图2描绘了根据本公开的实施例的被分解为拐角标识和表导出的表标识的示例性高级概览;

[0015] 图3描绘了根据本公开的实施例的用于拐角标识的卷积神经网络的示例性架构;

[0016] 图4描绘了根据本公开的实施例的表的事件定位的示例性概率图形模型;

[0017] 图5描绘了根据本公开的实施例的具有至少一个表的另一示例性电子表格,该电子表格可以具有错误标识的表;

[0018] 图6描绘了根据本公开的实施例的拐角标识模型和消融的准确性的条形图;

[0019] 图7描绘了根据本公开的实施例的作为电子表格大小的函数的表标识阶段的运行时间的曲线图;

[0020] 图8描绘了根据本公开的实施例的用于使用神经网络的改进的表标识的方法;

[0021] 图9描绘了根据本公开的实施例的可以根据本文中公开的系统、方法和计算机可读介质来使用的示例性计算设备的高级图示;以及

[0022] 图10描绘了根据本公开的实施例的可以根据本文中公开的系统、方法和计算机可读介质来使用的示例性计算系统的高级图示。

[0023] 再次,这里描述和示出了很多实施例。本公开既不限于任何单个方面或其实施例,也不限于这样的方面和/或实施例的任何组合和/或排列。本公开和/或其实施例的每个方面可以单独使用,或者与本公开和/或其实施例的一个或多个其他方面结合使用。为了简洁起见,这些组合和排列中的很多在本文中未单独讨论。

具体实施方式

[0024] 本领域的技术人员将认识到,本公开的各种实现和实施例可以根据说明书来实践。所有这些实现和实施例旨在被包括在本公开的范围之内。

[0025] 如本文中使用的,术语“包括(comprises)”、“包括(comprising)”、“具有(have)”、“具有(having)”、“包括(include)”、“包括(including)”或其任何其他变型旨在涵盖非排

他性的包括,使得包括一系列要素的过程、方法、物品或装置不仅仅包括这些要素,还可以包括未明确列出或这样的过程、方法、物品或装置固有的其他要素。术语“示例性”在“示例”的意义上而不是“理想”的意义上使用。另外,术语“或”旨在表示包括性的“或”而不是排他性的“或”。也就是说,除非另有说明或从上下文中清楚得知,否则短语“X采用A或B”旨在表示任何自然性的包括性的排列。例如,短语“X采用A或B”在以下任何情况下均得到满足:X采用A;X采用B;或X同时采用A和B。此外,除非另有说明或从上下文中清楚得知指向单数形式,否则本申请和所附权利要求书中使用的冠词“一个(a)”和“一个(an)”通常应当被解释为表示“一个或多个”。

[0026] 为简洁起见,与系统和服务器有关的用于执行系统和服务器(以及系统的各个操作组件)的方法和其他功能方面的传统技术在本文中可能不会详细描述。此外,本文中包含的各个附图中所示的连接线旨在表示各个要素之间的示例性功能关系和/或物理耦合。应当注意,在主题的实施例中可以存在很多替代和/或附加的功能关系或物理连接。

[0027] 现在将详细参考本公开的示例性实施例,其示例在附图中示出。在所有附图中将尽可能使用相同附图标记来表示相同或相似组件。

[0028] 除其他事项外,本公开总体上涉及一种使用神经网络来改进电子表格中的表标识的方法。尽管本公开具体地讨论电子表格中的表标识,但是本公开的各方面不仅可以适用于电子表格,而且还可以适用于其他文件类型,包括表和/或灵活的数据布置。

[0029] 转到图1,图1描绘了根据本公开的实施例的具有两个垂直对齐的表的示例性电子表格。如图1所示,电子表格中可以存在两个表(即,第一表的左上角在单元A5处,并且右下角在单元G16处,表示为A5:G16,第二表的左上角在单元A18处,并且右下角在单元G29处,表示为A18:G29)。如图1所示,这两个表是垂直堆叠的,并且包括描述列和行内容的文本。由于电子表格中包括各种信息和数据,因此自动检测两个表可能具有挑战性。自动检测表较困难的一个可能原因是,空行(例如,第7行和第15行)可能会增加可读性,但是空行也可能混淆用于检测表边界的尝试。自动检测表较困难的另一可能原因是,表标题可能是不规则的(例如,第6行和第19行的A列中的单元为空),这可能会混淆将行分类为标题。自动检测表较困难的另一可能原因是,单元边框线的使用可能是不一致的。因此,各种各样的布局和视觉提示可能会使自动表标识技术变得脆弱且无效。

[0030] 从电子表格中进行自动表提取可能有多种用途。例如,从电子表格中提取个体表作为数据库关系可以允许将数据库查询应用于数据集,并且可能潜在地包括跨多个表的筛选(filter)和联接(join)、以及基于自然语言的查询。另外,基于电子表格中的一个或多个表的标题信息的数据类型分析可以用于类型推断和/或一致性验证。

[0031] 在本公开的实施例中,电子表格中的表可以是包括行和列的、具有数据单元的矩形区域,并且表可以在表的顶部包括一个或多个附加行,包括字符串(诸如文本)、以及有关列中的数据的描述性信息(可以称为表标题)。电子表格中的表还可以包括一个或多个小计(subtotal)行,小计行可以使用累加函数和/或百分比来汇总一系列中的数据。

[0032] 本公开的示例性实施例描述了一种表标识技术,该表标识技术使用卷积神经网络来标识一个或多个表的独立拐角,诸如基于块(patch)的卷积神经网络。然而,本公开可以不限于卷积神经网络,并且本公开的实施例可以使用其他类型的神经网络来用于表标识。在标识每个拐角的候选之后,可以将拐角缝合在一起以构建一个或多个表。

[0033] 如下面更详细地讨论的,本公开的实施例提供了一种表标识方法,该方法将卷积神经网络(“CNN”)应用于围绕单元的矩形上下文,以将单元分类为潜在拐角,然后根据图形模型来组合拐角预测结果以标识表。表标识方法是端到端数据驱动的,比完整的电子表格卷积和/或对象检测快了几个数量级,并且能够从有限数目的电子表格(诸如1,638个电子表格)生成大型训练数据集,如下所述。

[0034] 在本公开的实施例中,S可以是具有 $R \times C$ 个单元的电子表格,其中R是行数并且C是列数。电子表格S可以包括一个或多个表 $\mathcal{T}_1, \dots, \mathcal{T}_n$,这些表具有一个或多个视觉提示和其他特征,诸如边框、颜色、字体样式和/或间距约定。每个表可以是电子表格S的矩形区域或正方形区域,由六(6)个拐角单元来标识,包括左上角(“TL”)、右上角(“TR”)、左下角(“BL”)、右下角(“BR”)、标题左下角(“HBL”)和标题右下角(“HBR”)。在电子表格符号中,表可以被定义为区域TL:BR及其标题区域TL:HBR。

[0035] 返回图1,图1描绘了具有多个表的示例性电子表格。给定电子表格S,诸如图1的电子表格,表标识问题可以如下:(a)标识电子表格S是否包括任何表;以及(b)标识每个表在电子表格S中的位置以及相应的列标题。

[0036] 如图1所示,表标识问题由于多种原因而可能具有挑战性,这些原因包括:可以使用电子表格的一个或多个单元的各种视觉和其他属性来指定的表,这些属性可以以不同组合来标识表。构建基于各种视觉和其他属性的基于规则的表标识可能很脆弱且繁琐。另外,可以通过利用空行和/或空列来稀疏电子表格中的表,以在电子表格的表中布置内容,以便页面上看起来更清晰。因此,“围绕给定单元的最大密集矩形/正方形”(在工业电子表格软件中使用的,诸如Microsoft Excel和Google Sheets)的常见启发式方法在很多电子表格中可能无效。表标识应当尽可能精确,因为几行和/或列中的错误可能会忽略电子表格中的重要数据。

[0037] 为了训练神经网络,每个具有零个或更多个表的电子表格被用作数据集。以下讨论是本公开的实施例的示例性技术细节。用于训练神经网络的训练数据可以来自多个来源,包括但不限于:电子表格的专家标注语料库、电子表格的众包标注语料库标注、和/或不同方法的混合。

[0038] 例如,本公开的示例性实施例使用来自公开可用数据集的电子表格的集合来测试,该公开可用数据集包括来自不同领域和/或各种不同来源的电子表格。例如,表标识数据集可以是个人和/或商用电子表格的开源和/或封源集合。

[0039] 初步服务可以用于检查电子表格,以确定电子表格是否包括具有数据的表。例如,可以使用一组初步标注器来标识具有带有数据的表的电子表格。然后,可以在电子表格上执行预定数目的标注。在执行预定数目的标注之后,标注数据集中的其余电子表格可以包括至少一个表。

[0040] 在处理电子表格以确定电子表格是否包括具有数据的表之后,第二服务可以标注其余电子表格。例如,对于每个电子表格,标注者可以标识电子表格中的一个或多个表,并且标记每个表的拐角和表的任何标题行。再次参考图1,对于服务和/或标注者,表标识可能具有挑战性,并且服务和/或标注者可能在表、拐角和/或标题行上没有达成一致。因此,在适用时,在标注者之间可以使用多数共识。替代地和/或另外地,经验最多的标注者的标签

可以起决定性作用。

[0041] 根据本公开的实施例,用于训练神经网络的数据集可以在训练神经网络之前被处理以产生标注数据集。该处理可以由自动化服务、一个或多个标注者和/或标注众包的标注者来执行。众包标注可以建立关于表标识的人类性能基线表。

[0042] 通过服务自动标注的电子表格可以与通过其他方法(诸如专家标注者)标注的电子表格合并。组合的标注电子表格可以构成用于表标识的训练数据集。当使用组合数据集时,对于重叠文件,优先使用专家标注者所做的标注作为真实基准。

[0043] 如上所述,本公开的实施例使用经训练的卷积神经网络(“CNN”)来用于表标识。CNN可以克服处理电子表格的挑战,该电子表格可以包含数以万计的单元。例如,在消费级机器上,具有50以上个通道的10,000个宽全图像卷积可能慢到不切实际。此外,如本公开的实施例中使用的,CNN可以克服需要数十万个数据点来学习数据的高级特征的挑战。例如,这种量级的标注表标识数据集可能不存在。

[0044] 为了应对这两个挑战,表标识可以被分解为两个阶段:(a)拐角标识和(b)表导出。拐角标识可以应用CNN来检测电子表格中可能是表的标题或拐角的单元。表导出使用概率图形模型将拐角单元集合组合到候选表中。图2描绘了根据本公开的实施例的被分解为拐角标识和表导出的表标识的示例性高级概览。将表标识缩减为拐角标识可以允许按单元分类来重现该问题,从而允许使用更多的可用数据并且使本地块的卷积速度更快。

[0045] 表标识的拐角标识阶段可以解决以下问题:给定 $W \times W$ 个单元的上下文块,将块的中心单元分类为6种表拐角(即,TL、TR、BL、BR、HBL或HBR(如上所述)或非拐角(“NaC”)之一。为了避免多类分类问题,可以将HBL拐角或HBR拐角定义为分别出现在左表边界或右表边界上的视觉表标题正下方的单元。

[0046] 为了将电子表格的单元分类为表拐角,神经网络可以考虑单元的各种视觉和其他特征以及单元的紧邻周围上下文。块中的每个单元都可以用预定数量的通道来表示,这些通道编码视觉和其他单元特征,例如, $F=51$ 个特征包括字体样式、颜色、边框、对齐、数据类型格式化等。

[0047] 图3描绘了根据本公开的实施例的用于拐角标识的卷积神经网络的示例性架构。根据如图3所示的流程,用于拐角标识的CNN可以处理 $W \times W \times F$ 个输入量,以预测拐角类别 $C = \{TL, TR, BL, BR, HBL, HBR, NaC\}$ 上的7路softmax分布。CNN架构可以将瓶颈卷积应用于下至24个特征。然后,CNN架构可以上采样至40个特征。之后,CNN可以应用具有指数线性单位(“ELU”)激活函数的三(3)个卷积层的堆叠,然后是具有dropout的两个全连接层。可以训练卷积神经网络使其不过拟合。可以使用更深的、具有更多个全连接层的卷积神经网络。卷积神经网络模型可以被构造为包括多个神经元,并且可以被配置为针对接收的电子表格的每个单元来输出多个类别中的一个或多个类别。多个神经元可以布置在包括至少一个隐藏层的多个层之中,并且可以通过多个连接进行连接。尽管图3描绘了特定示例性CNN架构,但可以使用获取关于类别的概率分布的任何CNN架构。

[0048] 根据本公开的示例性实施例,为了训练神经网络以用于拐角标识,如上所述,可以从标注真实基准数据集中为每个表生成训练块。具体地,对于训练数据集中的每个真实基准表 T ,可以从表中的各个关键位置提取拐角块以及预定数量的NaC块样本,诸如 $N=30$ 个NaC块。表中NaC块的关键位置包括以下一项或多项:表中心、边的中点、标题与数据之间的

间隙、表拐角周围的单元、以及通常分布在目标拐角附近的随机单元样本。可以在预定数量的类别(例如,TL、TR、BL、BR、HBL、HBR和NaC这七个类别)上使用分类交叉熵目标来训练神经网络模型。

[0049] 训练过程中的一个关键挑战可能是类别不平衡,因为非拐角(“NaC”)单元可能比拐角单元多很多。例如,每个表仅提供每种拐角的单个示例,但是可以包括NaC单元的多个示例。为了纠正类别不平衡,可以应用一种或多种正则化技术。一种正则化技术包括对每个表的可能的NaC单元进行欠采样 $\eta \cdot N$,其中 η 可以是可以在验证集上被优化的超参数,如下所述。另一种正则化技术包括应用具有预定概率 p 的dropout,其中 p 可以是超参数。如图3的CNN模型所示,预定概率 p 可以是例如 $p=0.5$ 。另一正则化技术包括应用数据增强,以引入具有合成噪声的其他拐角示例,其中每个合成示例可以仅在表外单元中在拐角单元周围添加小的随机噪声。噪声可以符合对称双峰正态分布,其中峰值在输入块的边缘处并且 $\sigma=(W/2)^{1/4}$ 。因此,每个表的拐角示例的总数可以与NaC的示例数相匹配。再一种正则化技术包括针对每个表的训练示例的权重进行重新缩放,使得拐角示例的总权重与NaC示例的总权重相匹配。

[0050] 对于拐角标识,高召回率可能比高准确率更为重要。拐角标识的目的是,建议候选,以用于根据拐角的后续表导出。表导出对于误报的拐角可能是健壮的,因为不太可能所有建议的拐角都为假。但是,表导出可能对漏报很敏感。因此,上述一种或多种正则化技术可以增加神经网络模型标识所有真实拐角的可能性,同时弱化NaC分类。

[0051] 转到表标识的第二阶段,表导出可以使用由神经网络模型的拐角标识所建议的候选拐角,以导出出现在给定电子表格中的表最可能的表集合。表导出的主要挑战可能是不确定性。由CNN标识出的拐角候选可能是噪声的,并且包括误报。为了准确建模来自有噪声拐角候选的表导出问题,可以使用概率图形模型(“PGM”)。

[0052] 图4描绘了根据本公开的实施例的表的事件定位的示例性概率图形模型。 $T=t, B=b, L=1$ 并且 $R=r$ 可以是独立事件,分别表示目标表 T 的顶部、底部、左侧和右侧边界线等于 t, b, l 和 r 。 $f_{\theta}(x, y; \gamma)$ 可以是单元 $\langle x, y \rangle$ 属于拐角类别 $\gamma \in C$ 的概率,该概率由应用于以单元 $\langle x, y \rangle$ 为中心的块的拐角标识模型 f_{θ} 估计。根据 f_{θ} ,可以观察到在对应边界线的相交处的出现表拐角的估计概率,例如, $\Pr(T=t, L=1)=f_{\theta}(t, l; T_L)$ 。候选表 T 的空间置信度得分可以如下计算。

$$\Pr(T = \langle t, l \rangle; \langle b, r \rangle) := \sqrt{f_{\theta}(t, l; TL) \cdot f_{\theta}(t, r; TR) \cdot f_{\theta}(b, l; BL) \cdot f_{\theta}(b, r; BR)}$$

[0053] (1)

[0054] 如下所示的算法1提供了根据本公开的实施例的表导出算法。表导出算法可以开始于,如果通过 f_{θ} 预测的单元的概率至少为预定量(诸如0.5),则表为每个单元 $\langle x, y \rangle \in S$ 分配类别 $\gamma \in C$ 。然后,表导出算法可以根据以上公式(1)使用表的空间置信度来对每个表候选进行评分。表导出算法还可以抑制具有较低空间置信度得分的重叠表。为了确保表与真实世界语义相匹配,表导出算法可以垂直合并没有至少一个良好定义的标题的表。最后,表导出算法可以选择空间置信度得分高于预定阈值 τ 的表,该预定阈值 τ 可以在验证数据集上被优化的超参数,如上所述。

算法1: 表导出**输入:** 拐角候选**输出:** 标识的表

```

1: Let  $f_{\theta}(x, y) := \gamma$  iff  $f_{\theta}(x, y; \gamma) \geq 0.5, \gamma \in C$ 
2:  $\tilde{T} \leftarrow \{(t, l): \langle b, r \rangle | 1 \leq t, b \leq R \wedge 1 \leq l, r \leq C \wedge f_{\theta}(t, l) =$ 
    $TL \wedge \dots \wedge f_{\theta}(b, r) = BR\}$ 
[0055] 3: 从  $\tilde{T}$  中删除任何重叠的表,
        每个重叠分组中仅保留具有最大  $\Pr(T)$  的表
4: for  $T_1 = \langle t_1, l \rangle: \langle b_1, r \rangle, T_2 = \langle t_2, l \rangle: \langle b_2, r \rangle \in \tilde{T}$  do
        (assuming  $t_2 > t_1$ )
5:     if  $\nexists h > t_2$  s.t.  $f_{\theta}(h, l) = HBL, f_{\theta}(h, r) = HBR$ 
6:         Merge  $T_2$  into  $T_1$  in  $\tilde{T}$ 
7: return  $\{T \in \tilde{T} | \Pr(T) > \tau\}$ 

```

[0056] 如上所述,用于训练神经网络模型的数据集可以被分割成多个数据集。例如,数据集可以被分割和/或划分为训练数据集、验证数据集和测试数据集。在本公开的示例性实施例中,由初步服务标注的电子表格可以用于训练。在标注电子表格中,各种预定百分比可以用于不同数据集。例如,可以将50%委托为训练数据集,将30%委托为验证数据集,而将20%委托为测试数据集。可以删除较小表(诸如 2×2 单元或更少),因为这样的标注表可能会导致误报。

[0057] 为了评估本公开的示例性实施例的各个方面,可以使用不同性能指标。可以使用的一种指标是Jaccard准确率/召回率。平均Jaccard准确率/召回率可以用于对象检测。平均Jaccard准确率和召回率可以定义如下:准确率=真阳性面积/预测面积,召回率=真阳性面积/真实基准面积,并且面积可以是建议的表所覆盖的单元数。可以针对不同的表计算准确率、召回率和F值的平均值表。

[0058] 可以使用的另一指标是平均归一化互信息(“NMI”),它是一种常用于评估聚类技术的指标。如本文中讨论的,表标识问题可以被视为电子表格单元的聚类,其中表对应于聚类。形式上,对于电子表格 S 上的真实基准表 $T_1^*, \dots, T_n^*$ 和预测表 $T_1, \dots, T_m$,每个单元 $c \in S$ 可以与相应真实聚类 $tc(c) \in \{0, 1, \dots, n\}$ 和预测聚类 $pc(c) \in \{0, 1, \dots, m\}$ 相关联。NMI可以计算为 $NMI([tc(c_1), \dots, tc(c_{|S|})], [pc(c_1), \dots, pc(c_{|S|})])$ 。为了简单起见,在本公开的实施例中,可以仅考虑在所有预测和真实基准表周围的边界框中的单元。

[0059] 平均Jaccard准确率/召回率和平均归一化互信息都可以在不同场景中使用。准确率/召回率可以是针对每个表预测的通用评估指标。但是,如图1所示,电子表格可以包括多个表,并且对多个表消除歧义可以是表标识问题的挑战的一部分。NMI可以是针对每个电子

表格的一组预测表相对于真实基准的准确性的聚类指标。

[0060] 可以在各种类型的计算机硬件上(诸如通过使用两个NVIDIA GTX 1080Ti GPU),使用开源软件库(诸如TensorFlow)通过预定时间量(诸如16个纪元(epoch)(每个模型10-12小时))来训练拐角标识神经网络模型。如下所示,本公开的实施例的实验在具有32GB RAM的Intel Xeon 3.60GHz 6核CPU(用于特征提取和表导出)和单个NVIDIA GTX 1080Ti GPU(用于拐角标识)上运行。

[0061] 对照基线评估根据本公开的示例性实施例的使用神经网络的表标识技术包括Microsoft Excel中的内置表检测、基于规则的检测、以及经由随机森林的单元分类。使用精确的表标识、基于单元聚类的归一化互信息、以及基于重叠的Jaccard指标以及其他指标来评估性能。

[0062] 例如,使用来自各种数据集的电子表格的集合来测试本公开的示例性实施例。使用Jaccard指标,本公开的示例性实施例的总体准确率为74.9%,召回率为93.6%,F值为83.2%,其可以与人类性能的86.9%的准确率、93.2%的召回率和89.9%的F值相当。

[0063] 如本文中讨论的,数据驱动方法可以比基于规则的方法更准确地检测表。由于表标识问题的固有困难,可以对照两个基于规则的基线来评估使用神经网络的表标识。如上所述,第一基线是Microsoft Excel中的“选择表”特征,该特征选择给定单元周围的最大密度矩形。第二基线是表检测的基于规则的独立实现,该过程查找具有兼容数据类型签名(假定空单元与任何数据类型“兼容”)的最大连续行块,而忽略间隙。第一基线和第二基线两者都检测给定单元周围的表,该给定单元可以是用户选择的。可以通过选择真实基准表中的 $m=5$ 个随机单元并且对 m 求平均指标来模拟交互。但是,这样的仿真可能使得无法针对基于规则的基线来测量NMI。

[0064] 表1总结了使用本公开的示例性实施例的来自我们的超参数扫描和消融测试的结果,如下面详细讨论的。特别地, $\eta=1/4$, $\tau=3/4$,并且 $W=11$ 。

[0065] 如下所示,就Jaccard召回率和F值而言,本公开的示例性实施例可以优于基于规则的方法。由于是保守实现,基于规则的方法可能具有更高的准确率和更低的召回率,该方法仅涵盖真实世界表的某些模式。诸如本公开的实施例等的数据驱动的表标识可以能够覆盖稀疏表布局的更加多样化的范围。

[0066] 表1

	Microsoft Excel				基于规则的				本公开			
	P	R	F	I	P	R	F	I	P	R	F	I
[0067] 验证	86.9	63.0	73.1	-	69.2	66.7	67.9	-	76.0	90.4	82.6	88.2
测试	83.3	58.8	68.9	-	77.7	71.2	74.3	-	74.9	93.6	83.2	87.3

[0068] 如表1所示,P表示Jaccard准确率,R表示Jaccard召回率,F表示F值,I表示NMI。表1中示出的每个值以百分比(%)为单位。

[0069] DeExcelerator项目是另一种类型的数据驱动方法,其可以使用随机森林和支持向量机(“SVM”)将电子表格单元分类为各种数据/标题/元数据的类型,然后可以应用基于规则的算法来将由这些分类导出的几何区域合并成表。通过对照其他类型的数据驱动方法的声称结果,来评估本公开的在其数据集上的表标识,可以执行本公开的实施例与其他类

型的数据驱动方法的比较。对于这个评估,可以使用以下指标:

$$[0070] \quad \text{matchDE}(\text{真实基准表 } T^*, \text{预测表 } T) := \frac{\text{overlap}(T^*, T)}{\max\{|T^*|, |T|\}} \geq 0.8 \quad (2)$$

[0071] 如公式(2)所示,上述指标以及与其关联的数据可能与表的自由形式的真实世界定义不匹配。表的定义和相应标注相对于类似类型的区域是连续的,这适用于其他类型的基于规则的表导出算法。

[0072] 图5描绘了根据本公开的实施例的具有至少一个表的另一示例性电子表格,该表可以具有错误标识的表。具体地,根据其他类型的数据驱动方法,电子表格被错误标识为具有四(4)个表。本公开的实施例可以标识具有空行和空列的单个表。在这种有分歧的情况下,可以执行附加评估。可以测量我们的专家标注与原始标注之间的人类一致性,并且两个标注达成一致的准确率为72.3%,召回率为90.0%,F值为80.2%。

[0073] 表2总结了针对如上所述的原始数据集和重新标记数据集的、本公开的示例性实施例与如上所述的其他类型的数据驱动方法的声称结果就公式(2)提供的指标以及上述指标的比较。针对重新标记数据集,与原始数据集相比,本公开的示例性实施例可以具有更好的性能,并且对于自身来说,针对重新标记的数据集的性能要优于原始数据集。基于公式(2)的准确率和F值的提高可以表明,与浅分类方法相比,神经网络能够更好地利用电子表格的丰富视觉结构。

[0074] 表2

数据集	方法	P _{DE}	R _{DE}	F _{DE}	P	R	F	I
[0075] 原始	其他方法	65.0	63.0	63.9	-	-	-	-
	本公开	68.5	47.6	56.1	57.9	88.6	70.1	81.3
重新标记	本公开	70.9	61.5	65.9	68.7	89.5	77.8	84.9

[0076] 如表2所示,P_{DE}、R_{DE}和F_{DE}分别表示按照公式(2)记录的两个表之间具有匹配的情况下的准确率、召回率和F值,P表示Jaccard准确率,R表示Jaccard召回率,F表示F值,I表示NMI。表2中示出的每个值均以百分比(%)为单位。

[0077] 如上所述,根据本公开的实施例,正则化技术可以提高拐角标识的准确性。图6描绘了根据本公开的实施例的拐角标识模型和消融的准确性的条形图。可以将根据本公开的实施例的使用神经网络模型的拐角标识的结果评估为它们在验证数据集和测试数据集上的不同拐角类别上的预测准确性的F值。条形图的x轴上的组可以描绘超参数 η 的不同值,这可以控制非拐角训练示例的数量。块大小W的不同值可以通过条的散列来描绘。每个神经网络模型可以表示一致的高准确性,其中每个拐角类别的相对准确性都有所不同。值 $\eta=1/4$ 和W=11可以根据其在验证集上所有拐角类别上的平均性能来选择。

[0078] 基于本公开的示例性实施例,表标识对于在消费级硬件上的实时使用是可能的。图7描绘了根据本公开的实施例的作为电子表格大小的函数的表标识阶段的运行时间的图。如图7所示,与电子表格中的单元数相比标识表所需要的时间量是线性的,中位数运行时间为71毫秒。大部分时间可能花费在从电子表格和拐角标识中提取特征。多个块的拐角标识可以在GPU上并行运行,并且自动受益于任何进一步的并行化。

[0079] 虽然图1和2描绘了神经网络框架,但本领域技术人员将理解,神经网络可以针对

模型来进行,并且可以包括以下阶段:模型创建(神经网络训练)、模型验证(神经网络测试)和模型利用率(神经网络评估),尽管这些阶段可能不会相互排斥。根据本公开的实施例,神经网络可以通过训练、推断和评估阶段来实现。至少一个服务器可以执行本文中描述的表标识系统的机器学习组件。如本领域技术人员将理解的,机器学习可以关于模型来进行,并且可以包括多个阶段:模型创建、模型测试、模型验证和模型利用,尽管这些阶段可能不是相互排斥的。此外,模型创建、测试、验证和利用可以是机器学习的持续过程。

[0080] 在第一机器学习阶段,模型创建阶段可以涉及从训练数据集的电子表格中提取特征。如本领域技术人员将理解的,这些提取的特征和/或其他数据可以在基于模式随时间而收集的大量数据上从统计分析和/或机器学习技术中导出。基于对该监测的观察,机器学习组件可以创建用于从电子表格中标识拐角的模型(即,一组规则或试探法)。如上所述,可以训练神经网络以取消强调漏报。

[0081] 在第二机器学习阶段,可以测试和/或验证在模型创建阶段创建的模型的准确性。在该阶段,机器学习组件可以从测试数据集和/或验证数据集中接收电子表格,从测试数据集和/或验证数据集中提取特征,并且将这些提取的特征与由模型制作的预测标签进行比较。通过持续跟踪和比较该信息并且持续一段时间,机器学习组件可以确定模型是否准确地预测电子表格的哪些部分可能是拐角。这种测试和/或验证通常用准确性来表示:即,模型预测标签的时间百分比是多少。可以将关于由模型进行的预测的成功或失败的信息反馈给模型创建阶段以改善模型,从而改善模型的准确性。

[0082] 第三机器学习阶段可以基于被验证为达到预定阈值准确度的模型。例如,被确定为具有至少50%的准确率的模型可能适用于使用阶段。根据本公开的实施例,在该第三利用阶段,机器学习组件可以从模型提示存在拐角的电子表格中标识出拐角。在遇到拐角时,模型建议存在哪种类型的拐角并且可以存储数据。当然,基于对各个存储拐角的确认或拒绝的信息可以作为用于完善模型的数据被返回给先前阶段(测试、验证和创建)以提高模型的准确性。

[0083] 尽管本公开具体地讨论电子表格处理,但是本公开的各方面不仅可以适用于电子表格,而且还可以适用于灵活的数据布置和/或分类问题。

[0084] 如上文和下文所述,本公开的实施例允许降低计算复杂度和存储器需求,这也可以降低功耗。本公开的实施例可以在诸如智能电话、平板电脑和/或甚至可穿戴物品、智能扬声器、计算机、膝上型计算机、汽车娱乐系统等移动设备上实现。

[0085] 图8描绘了根据本公开的实施例的用于诸如通过使用包括神经网络的机器学习而进行的改进的表标识的方法800。方法800可以在步骤802开始,在步骤802中,可以接收包括多个电子表格的数据集,其中多个电子表格中的至少一个电子表格包括至少一个表。接收到的电子表格的每个单元可以包括对单元的视觉和其他特征进行编码的预定数量的通道,并且这些特征可以包括字体样式、颜色、边框、对齐、数据类型格式等。

[0086] 然后,在步骤804,可以处理数据集以产生标注数据集,该标注数据集包括具有标识的表、拐角和标题行的多个标注电子表格。在标注之前,可以由服务处理电子表格以确定每个电子表格是否包括具有数据的表。然后,第二服务可以标注其余电子表格。此外,对于每个电子表格,标注者可以标识电子表格中的一个或多个表,并且标记每个表的拐角和表的任何标题行。

[0087] 在步骤806,可以从标注数据集的每个表中提取表的拐角单元和表的预定数量的非拐角单元。然后,在步骤808,可以基于每个表的所提取的拐角单元和每个表的预定数量的非拐角单元将至少一种正则化技术应用于标注数据集。表的预定数量的非拐角单元可以包括来自表中的预定位置的表的非拐角单元。表的非拐角单元的预定位置可以包括以下中的一项或多项:表的表中心、表的边的中点、表标题与数据之间的间隙、表的表拐角周围的单元、以及通常分布在表的目标拐角附近的随机单元。

[0088] 至少一种正则化技术可以包括以下中的一项或多项:i) 基于可以在验证集上优化的超参数,针对每个表对多个非拐角单元进行欠采样,ii) 在神经网络模型的架构中应用 dropout,iii) 应用数据增强,以引入带有合成噪声的其他拐角示例,其中每个合成示例在拐角单元周围向不在表中的单元添加小的随机噪声,以及iv) 重新缩放标注数据集的权重,使得拐角单元的总权重与非拐角单元的总权重相匹配。合成噪声可以遵循对称双峰正态分布,其中峰值在输入单元的边缘处并且 $\sigma = (W/2)^{1/4}$ 。

[0089] 在步骤810,可以构造包括多个神经元的神经网络模型,该神经网络模型被配置为针对接收到的电子表格的每个单元输出多个类别中的一个或多个类别,多个神经元布置在包括至少一个隐藏层多个层中并且可以通过多个连接而被连接。神经网络模型的至少一个隐藏层可以包括至少一个长短期存储器层。当然,神经网络模型的构造可以在使用神经网络模型之前的任何时间发生,并且神经网络模型的构造可以不限于在至少上述步骤之前和/或之后发生。

[0090] 在步骤812,可以使用标注数据集训练神经网络模型,神经网络模型可以被配置为针对接收到的电子表格的每个单元输出多个类别中的一个或多个类别。此外,经训练的神经网络模型可以包括卷积神经网络。多个类别可以包括左上角、右上角、左下角、右下角、标题左下角、标题右下角和非拐角,如上所述。在训练神经网络模型之后,在步骤814,可以输出被配置为针对接收到的电子表格的每个单元输出多个类别中的一个或多个类别的经训练的神经网络模型。

[0091] 在训练神经网络模型之后,在步骤816,可以接收包括电子表格的测试数据集和/或验证数据集,每个电子表格具有零个或更多个表。然后在步骤818,可以使用接收的测试数据集来评估经训练的神经网络。替代地,步骤816和818可以省略,和/或可以在不同时间执行。一旦评估为通过预定阈值,就可以利用经训练的神经网络。另外,在本公开的某些实施例中,可以重复方法800的步骤以产生多个经训练的神经网络。然后将多个经训练的神经网络相互和/或与其他神经网络比较。

[0092] 在步骤820,可以接收包括至少一个表的电子表格,其中接收的电子表格的每个单元包括对单元的视觉和其他特征进行编码的预定数量的通道。然后在步骤822,可以使用神经网络模型针对接收到的电子表格的每个单元标识多个类别中的一个或多个类别,其中多个类别可以至少包括拐角和非拐角。标识可以包括基于对单元的视觉和其他特征编码的预定数量的通道对接收的电子表格的每个单元分类。另外地和/或替代地,接收的电子表格的每个单元的分类还可以基于该单元的紧邻周围上下文。

[0093] 然后,在步骤824,可以基于接收的电子表格的每个单元的所标识的一个或多个类别,导出所接收的电子表格中的至少一个表,其中导出可以通过使用概率图形模型来完成。使用概率图形模型来导出接收到的电子表格中的至少一个表可以包括估计出现在对应边

界线的相交处的拐角的概率,计算每个候选表的空间置信度得分,以及当该每个候选表的空间置信度得分大于预定阈值时,选择该每个候选表作为表。预定阈值可以是基于验证数据集而优化的超参数。

[0094] 图9描绘了根据本公开的实施例的可以根据本文中公开的系统、方法、模块和计算机可读介质来使用的示例性计算设备900的高级图示。例如,根据本公开的实施例,计算设备900可以在使用神经网络处理诸如电子表格等数据的系统中使用。计算设备900可以包括执行存储在存储器904中的指令的至少一个处理器902。该指令可以是例如用于实现被描述为由以上讨论的一个或多个组件执行的功能的指令或用于实现上述一种或多种方法的指令。处理器902可以通过系统总线906访问存储器904。除了存储可执行指令,存储器904还可以存储数据、电子表格、一个或多个神经网络等。

[0095] 计算设备900可以另外包括处理器902可以通过系统总线906访问的一个数据存储库(也称为数据库)908。数据存储库908可以包括可执行指令、数据、示例、特征等。计算设备900还可以包括允许外部设备与计算设备900通信的输入接口910。例如,输入接口910可以用于接收来自外部计算机设备、用户等的指令。计算设备900还可以包括将计算设备900与一个或多个外部设备接口的输出接口912。例如,计算设备900可以通过输出接口912显示文本、图像等。

[0096] 可以想到,经由输入接口910和输出接口912与计算设备900通信的外部设备可以被包括在提供用户可以与之交互的基本上任何类型的用户接口的环境中。用户接口类型的示例包括图形用户界面、自然用户接口等。例如,图形用户界面可以接受来自采用诸如键盘、鼠标、遥控器等输入设备的用户的输入,并且可以在诸如显示器等输出设备上提供输出。此外,自然用户接口可以使得用户能够以不受由诸如键盘、鼠标、遥控器等输入设备施加的约束的方式与计算设备900交互。而是,自然用户接口可以依赖于语音标识、触摸和手写笔标识、屏幕上俄与屏幕相邻的手势标识、空中手势、头部和眼睛跟踪、声音和语音、视觉、触摸、手势、机器智能等。

[0097] 另外,尽管示出为单个系统,但是应当理解,计算设备900可以是分布式系统。因此,例如,几个设备可以通过网络连接进行通信,并且可以共同执行被描述为由计算设备900执行的任务。

[0098] 转到图10,图10描绘了根据本公开实施例的可以根据本文中公开的系统、方法、模块和计算机可读介质来使用的示例性计算系统1000的高级图示。披露。例如,计算系统1000可以是或可以包括计算设备900。另外地和/或替代地,计算设备900可以是或可以包括计算系统1000。

[0099] 计算系统1000可以包括多个服务器计算设备,诸如服务器计算设备1002和服务器计算设备1004(统称为服务器计算设备1002-1004)。服务器计算设备1002可以包括至少一个处理器、和存储器;至少一个处理器执行存储在存储器中的指令。指令例如可以是用于实现被描述为由以上讨论的一个或多个组件执行的功能的指令或用于实现上述一种或多种方法的指令。与服务器计算设备1002相似,除了服务器计算设备1002,服务器计算设备1002-1004的至少子集每个可以分别包括至少一个处理器、和存储器。此外,服务器计算设备1002-1004的至少子集可以包括相应数据存储库。

[0100] 一个或多个服务器计算设备1002-1004的处理器可以是或可以包括诸如处理器

902等处理器。此外,一个或多个服务器计算设备1002-1004的一个或多个存储器可以是或可以包括诸如存储器904等存储器。此外,一个或多个服务器计算设备1002-1004的一个或多个数据存储库可以是或可以包括诸如数据存储库908等数据存储库。

[0101] 计算系统1000还可以包括在服务器计算设备1002-1004之间传输数据的各种网络节点806。此外,网络节点1006可以通过网络1008将数据从服务器计算设备1002-1004传输到外部节点(例如,在计算系统1000外部)。网络节点1002还可以通过网络1008将数据从外部节点传输到服务器计算设备1002-1004。网络1008例如可以是互联网、蜂窝网络等。网络节点1006可以包括交换机、路由器、负载均衡器等。

[0102] 计算系统1000的结构控制器1010可以管理服务器计算设备1002-1004的硬件资源(例如,服务器计算设备1002-1004的处理器、存储器、数据存储库等)。结构控制器1010还可以管理网络节点1006。此外,结构控制器1010可以管理在服务器计算设备1002-1004上实例化的被管理运行时环境的创建、供应、取消供应和监督。

[0103] 如本文中使用的,术语“组件”和“系统”旨在涵盖配置有计算机可执行指令的计算机可读数据存储装置,该计算机可执行指令在由处理器执行时引起某些功能被执行。该计算机可执行指令可以包括例程、函数等。还应当理解,组件或系统可以位于单个设备上或者分布在多个设备之间。

[0104] 本文中描述的各种功能可以以硬件、软件或其任何组合来实现。如果以软件实现,则功能可以作为一个或多个指令或代码存储在计算机可读介质上和/或通过计算机可读介质传输。计算机可读介质可以包括计算机可读存储介质。计算机可读存储介质可以是计算机可以访问的任何可用存储介质。作为示例而非限制,这样的计算机可读存储介质可以包括RAM、ROM、EEPROM、CD-ROM或其他光盘存储、磁盘存储或其他磁性存储设备、或者可以用于以指令或数据结构的形式存储期望的程序代码并且可以由计算机访问的任何其他硬件介质。如本文中使用的,磁盘和光盘可以包括压缩光盘(“CD”)、激光光盘、光盘、数字多功能光盘(“DVD”)、软盘和蓝光光盘(“BD”),其中磁盘通常是磁性地复制数据,而光盘通常使用激光光学地复制数据。此外,传播信号不包括在计算机可读存储介质的范围内。计算机可读介质还可以包括通信介质,包括有助于将计算机程序从一个地方转移到另一地方的任何介质。连接例如可以是通信介质。例如,如果使用同轴电缆、光纤电缆、双绞线、数字用户线(“DSL”)或无线技术(诸如红外、无线电和微波)从网站、服务器或其他远程源传输软件,则同轴电缆、光纤电缆、双绞线、DSL或无线技术(诸如红外、无线电和微波)被包括在通信介质的定义中。上述各项的组合也可以被包括在计算机可读介质的范围内。

[0105] 替代地和/或另外地,本文中描述的功能可以至少部分由一个或多个硬件逻辑组件执行。例如而非限制,可以使用的示例性类型的硬件逻辑组件包括现场可编程门阵列(“FPGA”)、专用集成电路(“ASIC”)、专用标准产品(“ASSP”)、片上系统(“SOC”)、复杂可编程逻辑器件(“CPLD”)等。

[0106] 以上描述的内容包括一个或多个实施例的示例。当然,不可能为了描述上述方面而描述上述设备或方法的每个可能的修改和变更,但是本领域的普通技术人员可以认识到,各个方面的很多进一步的修改和排列是可能的。因此,所描述的各方面旨在涵盖落入所附权利要求的范围内的所有这样的改变、修改和变化。

	A	B	C	D	E	F	G	H
1	ANNEXURE 4							
2								
3	PHYSICAL ASSET MOVEMENT SCHEDULE (Not including inventories)							
4								
5	PHYSICAL ASSETS ACQUIRED DURING	Opening	Additions	Disposals	Transfers	Transfers	Closing	
6	FINANCIAL YEAR 2002/03	Balance			In	Out	Balance	
7		R'000	R'000	R'000	R'000	R'000	R'000	
8	LAND AND BUILDINGS	-	5 169	-	-	-	5 169	
9	Non-Residential Buildings	-	5 169	-	-	-	5 169	
10								
11	MACHINERY AND EQUIPMENT	-	4 386	-	-	-	4 386	
12	Computer equipment	-	3 243	-	-	-	3 243	
13	Furniture and office equipment	-	488	-	-	-	488	
14	Other machinery and equipment	-	655	-	-	-	655	
15								
16		-	9 555	-	-	-	9 555	
17								
18	PHYSICAL ASSETS ACQUIRED DURING	Opening	Additions	Disposals	Transfers	Transfers	Closing	
19	FINANCIAL YEAR 2001/02	Balance			In	Out	Balance	
20		R'000	R'000	R'000	R'000	R'000	R'000	
21	LAND AND BUILDINGS	-	800	-	-	-	800	
22	Non-Residential Buildings	-	800	-	-	-	800	
23								
24	MACHINERY AND EQUIPMENT	-	7 550	-	-	-	7 550	
25	Computer equipment	-	5 345	-	-	-	5 345	
26	Furniture and office equipment	-	1 465	-	-	-	1 465	
27	Other machinery and equipment	-	740	-	-	-	740	
28								
29		-	8 350	-	-	-	8 350	
30								
31								

图1

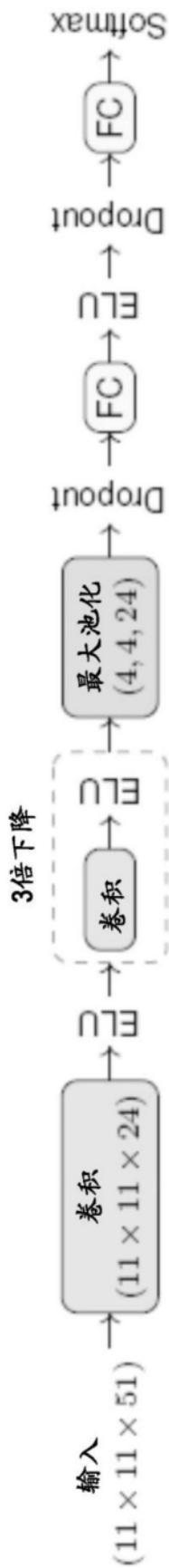


图3

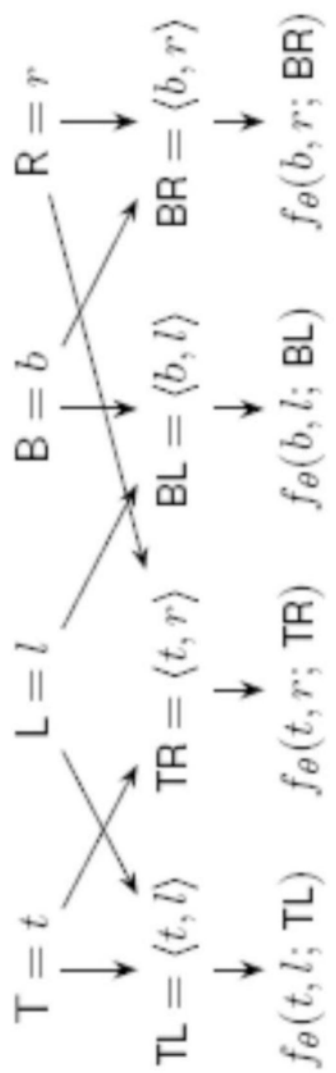


图4

	Total hours (in millions)	Annual hours per worker	Total number of workers (thousands)	GDP in 1995 prices (millions of Krones)	Share in GDP of			Capital service growth rate of			TFP growth
					Labour	IT-capital	Non-IT Capital	IT	NON-IT	All asset types	
1980	6,352	1,503	4,226	1,186,534	0.727	0.021	0.251	0.124	0.033	0.040	0.003
1981	6,306	1,495	4,219	1,180,946	0.724	0.021	0.255	0.129	0.031	0.038	-0.010
1982	6,355	1,508	4,213	1,194,517	0.716	0.020	0.263	0.117	0.026	0.032	-0.003
1983	6,402	1,518	4,218	1,218,906	0.693	0.022	0.284	0.144	0.026	0.035	0.004
1984	6,458	1,520	4,249	1,277,676	0.670	0.025	0.305	0.178	0.027	0.039	0.028

图5

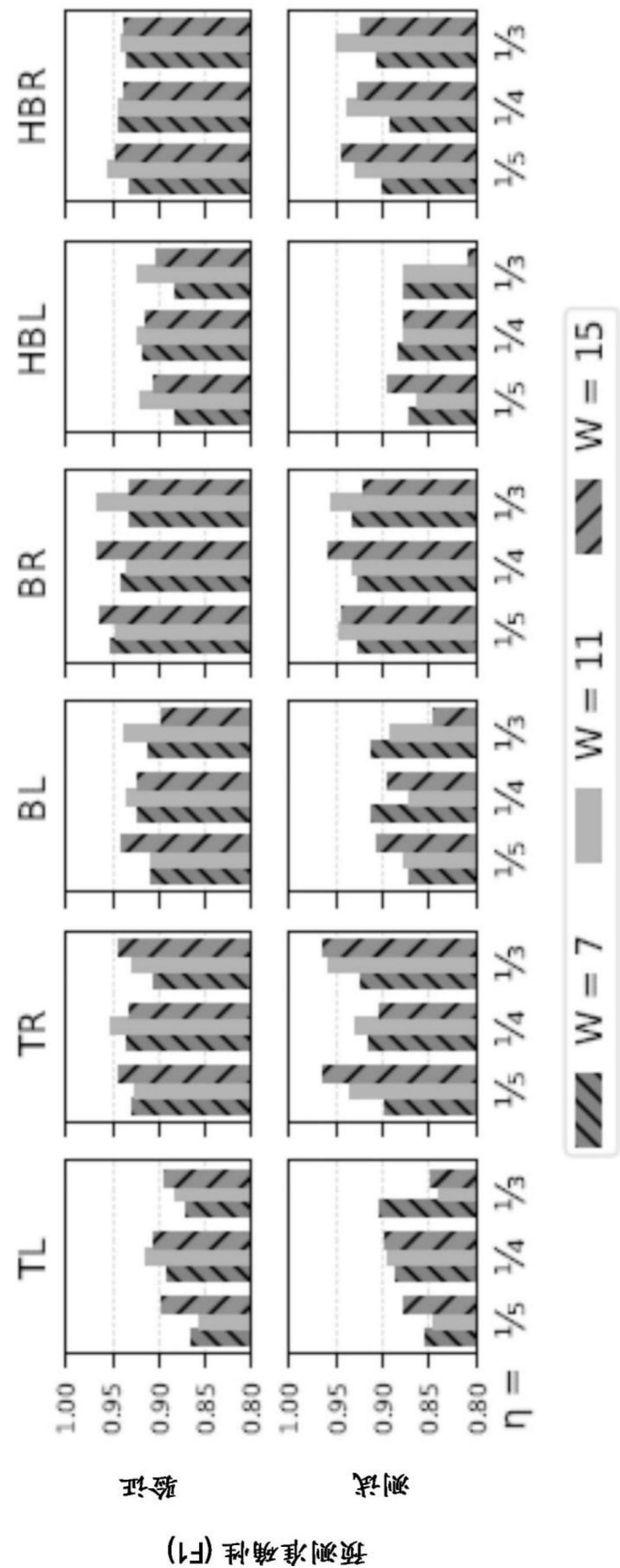


图6

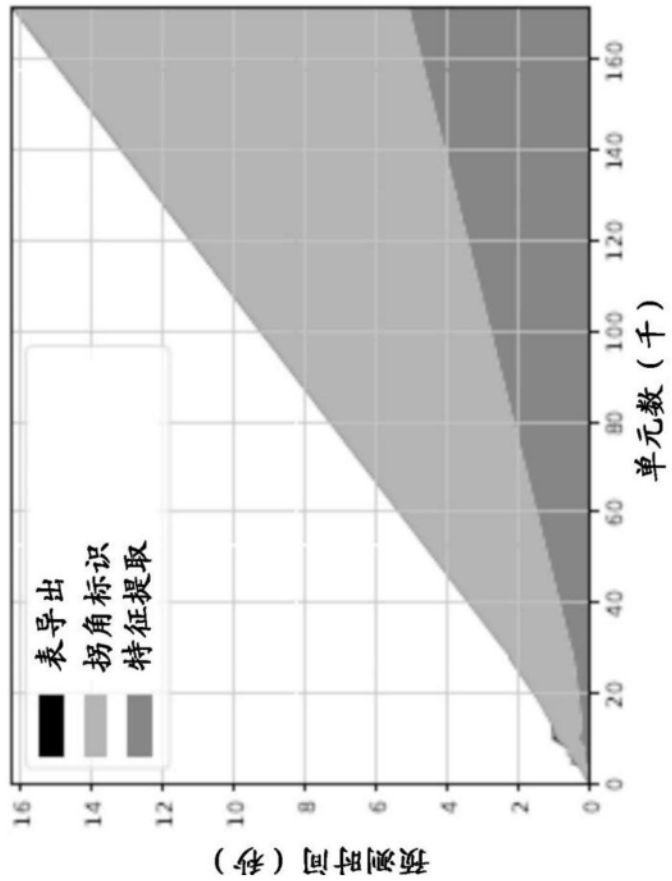


图7



图8

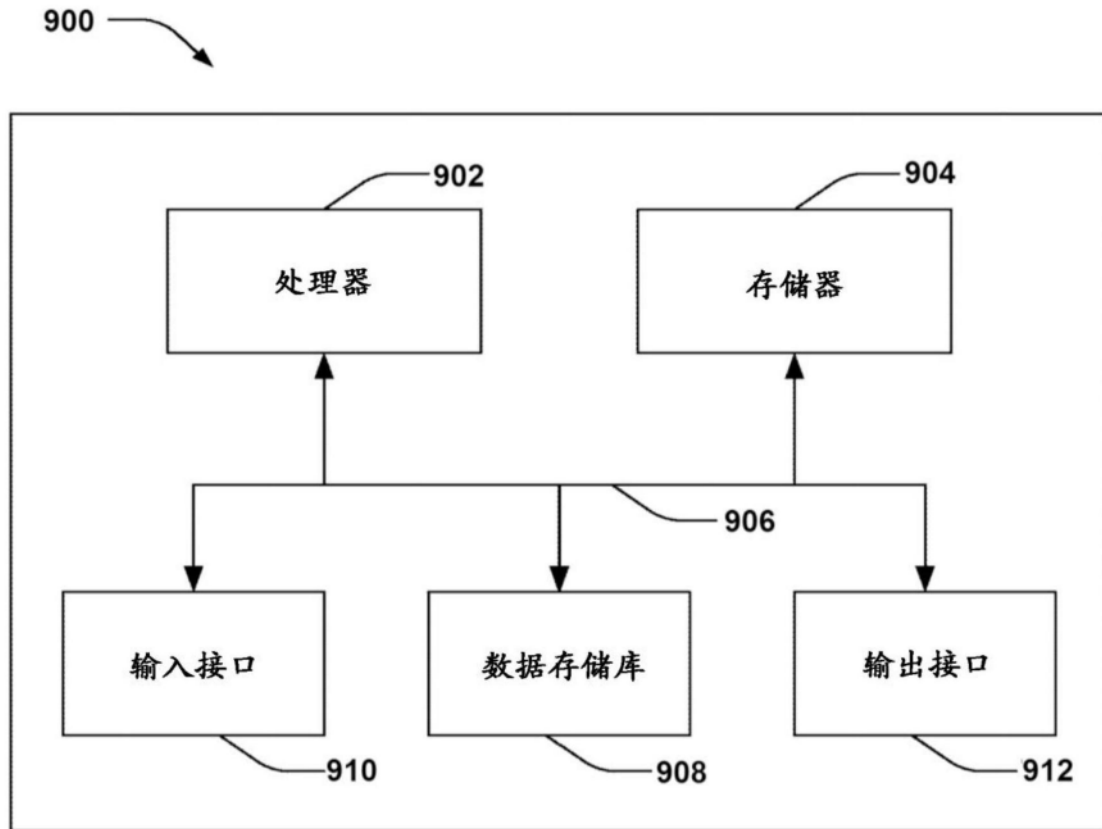


图9

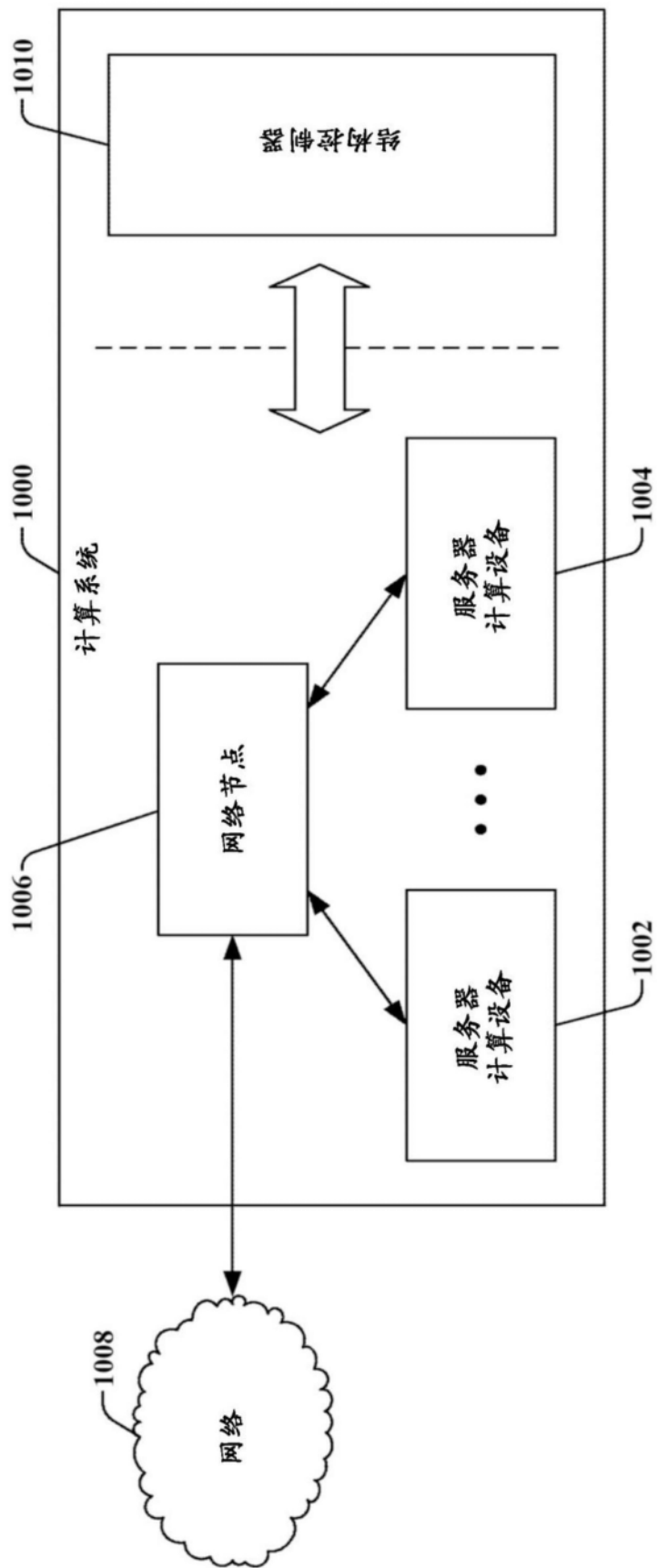


图10