(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119365905 A
(43)申请公布日 2025.01.24

(21)申请号 202380045218.6
(22)申请日 2023.06.06
(30)优先权数据
22177366.6 2022.06.06 EP
(85)PCT国际申请进入国家阶段日
2024.12.05
(86)PCT国际申请的申请数据
PCT/EP2023/065079 2023.06.06
(87)PCT国际申请的公布数据
WO2023/237532 EN 2023.12.14
(71)申请人 蓝色棱镜有限公司
地址 英国沃灵顿
(72)发明人 约翰·爱德华·里德
米格尔·安德烈亚·阿罗卡奥莱特
佩塔尔·马林诺夫·彼得罗夫
(74)专利代理机构 北京华夏正合知识产权代理事务所(普通合伙) 11017
专利代理师 韩登营 徐洲
(51)Int.Cl.
G06V 30/19(2006.01)
G06V 30/412(2006.01)
G06V 30/418(2006.01)

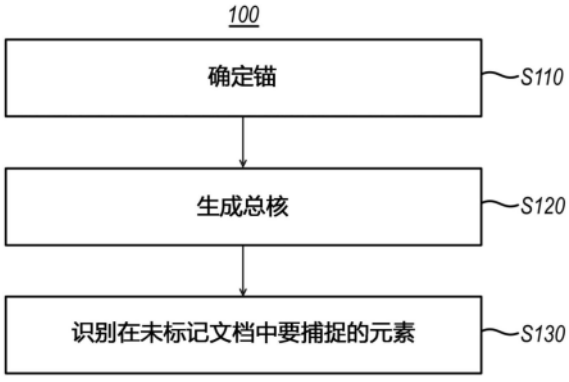
权利要求书2页 说明书19页 附图13页

(54)发明名称

用于执行数据捕捉的方法和系统

(57)摘要

提供了一种用于基于已标记文档在未标记文档中执行数据捕捉的方法(100),其中已标记文档包含一个或多个元素,每个元素关于字段被标记。所述方法包括:确定(S110)一个或多个锚,其中,锚包括在所述已标记文档与所述未标记文档之间匹配的点对;使用所确定的一个或多个锚来生成(S120)总核;以及在核机器中使用所述总核,识别(S130)在所述未标记文档中要捕捉的一个或多个元素,其中,所述要捕捉的一个或多个元素在字段方面对应于所述已标记文档中包含的所述一个或多个已标记的元素。



CN 119365905 A

1.一种用于基于已标记文档在未标记文档中执行数据捕捉的计算机实现的方法(100),其中所述已标记文档包含一个或多个元素,每个元素关于字段被标记,所述方法包括:

确定(S110)一个或多个锚,其中,锚包括在所述已标记文档与所述未标记文档之间匹配的点对;

使用所确定的一个或多个锚来生成(S120)总核;以及

在核机器中使用所述总核,识别(S130)在所述未标记文档中要捕捉的一个或多个元素,其中,所述要捕捉的一个或多个元素在字段方面对应于所述标记文档包含的所述一个或多个已标记的元素。

2.根据权利要求1所述的方法,其中生成(S120)所述总核包括:

为所述一个或多个锚中的每一个生成卷积核;以及

通过对所述一个或多个卷积核求和来生成所述总核。

3.根据权利要求2所述的方法,其中,每个卷积核测量在所述已标记文档和所述未标记文档内和之间的单词的相似性,其中,所述相似性是相应的单词与所述锚的相应点的距离以及相对于所述锚的相应点的相对位置的相似性的函数。

4.根据前述权利要求中任一项所述的方法,其中,所述已标记文档还包括由一个或多个标签指示的一个或多个字段。

5.根据前述权利要求中任一项所述的方法,还包括为所述已识别的一个或多个元素中的每个元素确定相应元素所对应的字段,其中所述确定在核机器中使用所述总核。

6.根据权利要求5所述的方法,其中,在所述未标记文档中识别多个元素,并且其中,确定所述已识别的一个或多个元素中的每个元素所对应的字段包括将所述多个已识别的元素分组为一个或多个组,每个组对应于一个字段。

7.根据权利要求6所述的方法,其中,将所述多个已识别的元素分组包括:

生成所述未标记文档中的所述多个已识别的元素的线性文本序列;

确定所生成的线性文本序列中的每个元素的argmax分类;以及

将所生成的线性文本序列中的共享argmax分类的各个元素的每个最大邻接子序列分组到相应的组中。

8.根据权利要求4至7中任一项所述的方法,其中,所述已标记文档中的所述一个或多个字段和一个或多个元素是以信息结构呈现的,并且所述方法还包括以相同的信息结构输出所述已识别的一个或多个元素及其对应的字段。

9.根据权利要求4至8中任一项所述的方法,还包括在确定(S110)一个或多个锚之前,获取模式,所述模式定义包含在所述已标记文档中的所述一个或多个字段以及与所述已标记文档中的每个字段相对应的元素是否以表格呈现。

10.根据前述权利要求中任一项所述的方法,其中,确定(S110)锚包括:

提取所述已标记文档和所述未标记文档中的所有单词;

通过将包含在所述标记文档中的单词排序成线性文本序列并移除对应于已标记的元素的任何单词来生成第一线性文本序列;

通过将包含在所述未标记文档中的单词排序成线性文本序列来生成第二线性文本序列;

识别所述第一线性文本序列与所述第二线性文本序列之间的匹配;以及

将所匹配的单词指定为所述锚。

11.根据前述权利要求中任一项所述的方法,其中,确定(S110)锚包括:

识别存在于所述已标记文档和所述未标记文档两者中的视觉特征;以及

将所识别的视觉特征指定为所述锚。

12.根据前述权利要求中任一项所述的方法,其中,确定(S110)锚包括:

围绕所述未标记文档的边界放置多个点和围绕所述已标记文档的边界放置相同数量的点;

识别所述已标记文档中的点与所述未标记文档中的点之间的匹配;以及

将所匹配的点指定为所述锚。

13.根据前述权利要求中的任一项所述的方法,还包括在确定(S110)一个或多个锚之前:

获取多个候选已标记文档;以及

基于所述多个候选已标记文档中的每一个与所述未标记文档之间的相似性度量,从所获取的多个候选已标记文档中确定所述已标记文档。

14.一种用于基于已标记文档在未标记文档中执行数据捕捉的系统(200),其中所述已标记文档包含一个或多个元素,每个元素关于字段被标记,所述系统包括:

确定单元(210),其配置为确定一个或多个锚,其中锚包括在所述已标记文档与所述未标记文档之间匹配的对应点对;

生成单元(220),其配置为使用所确定的一个或多个锚来生成总核;以及

识别单元(230),其配置为在核机器中使用所述总核来识别在所述未标记文档中要捕捉的包含要捕捉的信息的一个或多个元素,其中,所述要捕捉的一个或多个元素在字段方面对应于包含在所述已标记文档中的所述一个或多个已标记的元素。

15.一种包括指令的计算机可读存储介质,所述指令在由计算机执行时使所述计算机执行权利要求1至13中任一项所述的方法的步骤。

用于执行数据捕捉的方法和系统

技术领域

[0001] 本公开涉及用于执行数据捕捉的方法和系统。更具体地,本公开涉及用于基于一个或多个已标记文档在一个或多个未标记文档中执行数据捕捉的系统和计算机实现的方法。

背景技术

[0002] 各种类型的文档被广泛地用于收集和记录许多目的和领域(诸如医学、商业、教育和政府等)的信息。由于当今计算机和通信网络的广泛使用,这些文档现在通常被电子地创建和提供,使得它们能够被数字地生成和共享。这些文档通常包含结构化或半结构化格式的数据。结构化文档可以具有嵌入代码,所述嵌入代码使得信息能够以指定格式排列。半结构化文档可以是诸如账单、银行报表、公用事业账单、护照等的文档,其不遵循结构化文档所遵循的严格布局,并且它们不像结构化文档那样被绑定到指定数据字段。

[0003] 为了利用来自这些电子文档的信息来执行期望的操作(例如,企业相关操作),需要适当地提取信息。然而,至少对于半结构化电子文档,可能难以可靠地解析和提取所需信息,因为这些文档通常不具有明确定义的布局。

发明内容

[0004] 数据捕捉是从文档中提取相关信息的任务。本公开中描述的一次性数据捕捉(OSDC)技术设计为从结构化和半结构化文档中提取信息。要捕捉的数据可以包括文档中的键值对的值,例如账单编号、总金额和/或文档中的表格中的值。在本公开的上下文中,键值对可以被称为“字段-元素对”,其中键可以被称为“字段”并且值可以被称为“元素”。在其最简单的形式中,OSDC技术可以对文档对起作用:一个是已标记的,一个是未标记的。在这种情况下,任务是从未标记的文档中提取与在已标记文档中已标记信息相同类型的信息。

[0005] 本文描述的实施例针对用于基于已标记文档在未标记文档中执行数据捕捉的方法和系统。

[0006] 根据本公开的一个方面,提供了一种用于基于已标记文档在未标记文档中执行数据捕捉的计算机实现的方法,其中已标记文档包含一个或多个元素,每个元素关于字段被标记。所述方法包括:确定一个或多个锚,其中锚包括在所述已标记文档与所述未标记文档之间匹配的点对;使用所确定的一个或多个锚来生成总核;以及在核机器中使用所述总核,识别在所述未标记文档中要捕捉的一个或多个元素,其中所述要捕捉的一个或多个元素在字段方面对应于所述已标记文档中包含的所述一个或多个已标记的元素。

[0007] 根据本公开的另一方面,提供了一种用于基于已标记文档在未标记文档中执行数据捕捉的系统,其中已标记文档包含一个或多个元素,每个元素关于字段被标记。所述系统包括:确定单元,其配置为确定一个或多个锚,其中锚包括在所述已标记文档与所述未标记文档之间匹配的对应点对;生成单元,其配置为使用所确定的一个或多个锚来生成总核;以及识别单元,其配置为在核机器中使用所述总核识别在所述未标记文档中要捕捉的包含要

捕捉的信息的一个或多个元素,其中,所述要捕捉的一个或多个元素在字段方面对应于所述已标记文档包含的所述一个或多个已标记的元素。

[0008] 根据本公开的另一方面,提供了一种包括指令的计算机可读存储介质,所述指令在由计算机执行时使计算机执行如本公开中描述的方法的步骤。

[0009] 在所附从属权利要求中阐述了可选特征。

[0010] 当结合以下描述和附图考虑时,将更好地认识和理解本公开的这些和其它方面。以下描述虽然指示本公开的各种实施例及其许多具体细节,但是以说明而非限制的方式给出。在本公开的范围内可以进行许多替换、修改、添加或重新布置,并且本公开包括所有这样的替换、修改、添加或重新布置。

附图说明

[0011] 现在将参考附图仅通过示例的方式描述本公开的实施例,在附图中:

图1是根据本公开的实施例的用于执行数据捕捉的计算机实现的方法的流程图;

图2是根据本公开的实施例的用于执行数据捕捉的系统的框图;

图3A和图3B是示出根据本公开的实施例的示例性已标记文档和示例性未标记文档的图;

图4是示出确定图3A和3B的已标记文档和未标记文档之间的锚的示例性过程的图;

图5是示出根据图4所示的示例性过程确定的锚的图;

图6是示出根据本公开的实施例的示例性已标记文档和示例性未标记文档的图;

图7A是示出基于图6的已标记文档和未标记文档确定的锚的图;

图7B是示出基于图6的已标记文档和未标记文档确定的另一锚的图;

图8是示出根据图6的文档和另一未标记文档确定的锚的图;以及

图9是示出根据本公开的实施例的用于执行数据捕捉的拓扑的图。

具体实施方式

[0012] 在以下描述中,阐述了许多具体细节以便提供对实施例的透彻理解。然而,对于本领域普通技术人员来说,显然不需要采用具体细节来实践实施例。在其它示例中,未详细描述众所周知的材料或方法以便避免混淆本公开。

[0013] 图1是根据本公开的实施例的用于执行数据捕捉的计算机实现的方法100的流程图。更详细地,所示方法100用于基于已标记文档在未标记文档中执行数据捕捉。已标记文档包含一个或多个元素,每个元素关于字段被标记。每个元素可以由单词、单词串、或数字、或其组合来表示。在一些实施例中,已标记文档还可以包括由一个或多个标签指示的一个或多个字段。所述已标记文档和/或所述未标记文档可以是PDF文档或图像。在一些实现中,方法可由客户端计算设备、网络、服务器或其组合,诸如图9中所示的客户端计算机设备905、网络910和服务器920,来执行。

[0014] 方法100在步骤S110处开始,在所述步骤处确定一个或多个锚。锚包括在已标记文档与未标记文档之间匹配的点对。在一些实现中,为每个新的文档对(其中所述对包括未标记文档和已标记文档)确定一个或多个锚的新集合。

[0015] 在一些实施例中,在步骤S110处确定锚可以包括接收选择未标记文档和已标记文档中的点对的用户输入。或者,在一些实施例中,在步骤S110处确定锚可包括以下步骤:

-(例如,使用OCR)提取所述已标记文档和所述未标记文档中的所有单词;

-通过将包含在标记文档中的单词排序成线性文本序列(并且可选地,移除对应于已标记的元素的任何单词)来生成第一线性文本序列;

-通过将包含在未标记文档中的单词排序成线性文本序列来生成第二线性文本序列;

-识别第一和第二线性文本序列之间的匹配(如果在这个阶段处对应于已标记的元素的单词没有被移除,则在匹配识别步骤中忽略这些单词);以及

-将所匹配的单词指定为所述锚。

[0016] 在这种情况下,匹配涉及两个单词的固有相似性,以及它们的对齐(即,在各自序列中的各自单词的线性顺序方面的相似性)。在这些实施例中,第一和第二线性文本序列之间的匹配可以是在固有相似性(例如视觉特征)和对齐(即,它们在各自文本序列中的各自线性顺序方面是多相似)两者方面超过预定阈值的匹配。换句话说,匹配可以是看起来相同或相似并且在各自文本序列中具有相同或相似的线性顺序的两个单词。作为视觉相似性的示例,单词“bam”和“barn”可以被确定为由于其视觉相似性而超过预定视觉相似性阈值(例如,其中阈值考虑了诸如“m”和“rn”之类的光学字符识别(OCR)错误)。作为对齐的示例,第一文本序列中的第8个单词与第二文本序列中的第8个单词对齐的程度将高于与序列文本序列中的第个7单词对齐的程度。在一些实施例中,匹配的识别(更具体地,两个单词是否超过对齐阈值的确定)可以基于对齐算法,诸如类似于或等同于尼德曼-翁施(Needleman-Wunsch)算法的算法。

[0017] 替代地或附加地,在这些实施例中,可以使用阅读顺序算法(即,试图如人阅读单词那样对单词排序的任何算法)来执行对包含在已标记和未标记文档中的单词排序。

[0018] 替代地或附加地,在这些实施例中,所述一个或多个锚中的每一个可以被定义为在已标记文档中的相应匹配单词的中心与未标记文档中的相应匹配单词的中心之间的对应关系。

[0019] 在一些实施例中,在步骤S110处确定锚可以包括(例如,使用图像对齐算法)识别存在于已标记文档和未标记文档两者中的视觉特征,并将所识别的视觉特征指定为锚。图像对齐算法可以基于以下中的至少一个的算法:关键点检测器(例如,高斯差、哈里斯(Harris)、要跟踪的良好特征(GFTT)、局部不变描述符(例如,尺度不变特征变换(SIFT)、加速鲁棒特征(SURF)、以及定向FAST和旋转BRIEF(ORB)等)以及关键点匹配过程(例如,随机采样一致性(RANSAC))。

[0020] 在一些实施例中,在步骤S110处确定锚可以包括围绕未标记文档的边界放置多个点和围绕已标记文档的边界放置相同数量的点,识别已标记文档中的点与未标记文档中的点之间的匹配,以及将所匹配的点指定为锚。

[0021] 尽管在本实施例中描述了为文档对(即,已标记和未标记文档)确定所述一个或多个锚,但是应当理解,上述用于确定锚的过程可以应用到多于两个文档(例如,多个已标记文档和/或多个未标记文档)。例如,锚可以包括匹配(即,超过视觉相似性阈值和对齐阈值)的多个点,所述多个点中的每个点在不同的文档中。

[0022] 此外,尽管在本实施例中描述了锚可以包括匹配的单个单词(其可以或可以不由数字组成或包括数字),但是在其他实施例中,锚可以包括匹配的短语,作为单词的补充或替代,例如,单词“total amount(总金额)”作为单个短语而不是两个单独的单词而被处理,因此,相应的锚对应于在文档对中匹配的短语“total amount”。

[0023] 返回图1,在步骤S120处,使用所确定的一个或多个锚来生成总核。在一些实施例中,在步骤S120处生成总核可以包括为在步骤S110处确定的所述一个或多个锚的每个生成卷积核,并通过对所述一个或多个卷积核求和来生成总核。总核可以是所述一个或多个卷积核的加权和。

[0024] 每个卷积核可以涉及在已标记文档和未标记文档内和之间的所有单词对。在存在多于两个文档的其它实施例中,在所述多于两个文档中执行数据捕捉,每个卷积核可以涉及在两个或更多文档内的所有单词对以及跨所有两个或更多文档的所有单词对。每个卷积核可以测量在已标记文档和未标记文档内和之间(或跨在其中涉及多于两个文档的所有文档)的单词的相似性,相似性是相应单词离锚的相应点的距离和相对于锚的相应点的相对位置的相似性的函数。

[0025] 为了更详细地解释卷积核,我们将文档d定义为单词的序列,$w_d = w_{d,1},\cdots,w_{d,n_d}$,其中$n_d$是文档中的单词的数量。

[0026] 如果文档被标记,则还可以提供用于每个单词的标签,$l_d = l_1,\cdots,l_{n_d}$,其中标签可以属于K个类别(每个类别对应于一个字段)中的任何一个:

$l_i \in \{1,\cdots,K\}, 1 \leqslant i \leqslant n_d$

[0027] 此外,每个单词可以具有在文档的页面上的位置。第i个单词在第d个文档中的位置$w_{d,i}$被指定为$x_{d,i}$。

[0028] 跨两个文档d_1和d_2的锚a可以由位置对来表示,在每个文档中一个位置,$a=(a_1, a_2)$,其中a_1是在d_1中的位置,并且a_2是在d_2中的位置。已标记文档可以被定义为$d=d_1$,并且未标记文档可以被定义为$d=d_2$。

[0029] 对单词对$(w_{d,i},w_{d',j})$进行操作的卷积核k可以相对于锚a定义为其它径向基函数(RBF)核k_{off}和k_{rel}的乘积。到核的键输入是单词到锚的相对位置,定义为:

$z=w_{d,i}-a_d$

$z'=w_{d',j}-a_{d'}$

核可以被定义为:

$k(w_{d,i},w_{d',j};a)=k_{off}(z)k_{off}(z')k_{rel}(z,z')$

$k_{off}(z)=k_{RBF}(||zA_{off}||_2)$

$k_{rel}(z,z')=k_{RBF}(||(z-z')A_{rel}||_2)$

$$k_{RBF}(r)=e^{-\frac{r^2}{2}}$$

在上述方程中,r是距离(即,向量||(x-x')||的L2-范数)。

[0030] 由于卷积核可以作用于任何单词对,而不管它们是否在同一文档中,因此注意,对于$d=d'$或$d\neq d'$是可能的。

[0031] 线性变换(矩阵)A_{off}和A_{rel}用作逆长度尺度。在一个实现中,这些参数被对角地参数化,其中σ是长度尺度:

$$A_{off} = \begin{pmatrix} \frac{1}{\sigma_{x,off}} & 0 \\ 0 & \frac{1}{\sigma_{y,off}} \end{pmatrix}$$

$$A_{rel} = \begin{pmatrix} \frac{1}{\sigma_{x,rel}} & 0 \\ 0 & \frac{1}{\sigma_{y,rel}} \end{pmatrix}$$

这允许核在文档中的每个方向上具有不同的影响。例如,σ的值可以被设置为使得在文档中对宽度的影响比对高度的影响更灵活。

[0032] 上述核方程的原因是偏移核$_{koff}$随着单词离锚越远而衰减。也就是说,锚仅具有局部效果。相反,相对核$_{krel}$量化锚和每个单词之间的空间关系之间的相似性。也就是说,当各个单词相对于锚位于相同位置时,它们是相似的。可以取核的乘积,使得仅当各个单词都靠近锚并且它们相对于锚的位置相似时,才认为它们相似。

[0033] 尽管上面描述了每个卷积核可以涉及单词对,但是在替代实施例中,代替单个单词对,或者除了单个单词对之外,卷积核可以涉及短语对(例如,每个短语包括多个单词)。

[0034] 如上所述,在一些实施例中,用于确定锚的过程可以应用到两个以上文档(例如,一个以上已标记文档和一个未标记文档),其中每个锚可以包含匹配的多个点,所述多个点中的每个存在于不同文档中。在这些实施例中,锚可以被定义为a＝(a_1,a_2,...,a_N),,其中N-1是已经被选择用于数据捕捉的已标记文档的数量(因此N是文档的总数)。此外,在这些实施例中,每个卷积核可以涉及在所有文档内和跨所有文档的所有单词对,并且总核是所有卷积核的总和。这个更大的总核(当与使用仅两个文档的情况相比时)可以在用于识别要捕捉的一个或多个元素的步骤S130处向核机器提供更多信息。关于文档的集合定义的每个卷积核可以表示为对称的正半定矩阵,其具有对应于相应的文档的集合中的所有文档中的单词的数量的维度。因此,在这种情况下,矩阵中的每个矩阵条目对应于文档的集合中的各个文档中的单词中的特定单词对之间的相应相似性。类似地,由于通过对卷积核求和来生成总核,因此总核也可以表示为对称的正半定矩阵,其中矩阵中的每个矩阵条目对应于文档的集合中的各个文档中的单词中的特定单词对之间的相应相似性。

[0035] 返回图1,在步骤S130处,在核机器中使用总核来识别在未标记文档中要捕捉的一个或多个元素。要捕捉的一个或多个元素在字段方面对应于包含在已标记文档中的所述一个或多个已标记的元素。核机器可以包括以下中的至少一个:支持向量机(SVM)、正则化核判别分析(RKDA)和高斯过程(GP)。可以基于由已标记文档中的相应标签所指示的一个或多个元素和一个或多个字段中的至少一个来训练核机器。在一些情况下,可以基于由已标记文档中的相应标签所指示的一个或多个元素和一个或多个字段中的全部来训练核机器,以便最大化用于训练的样本的数量。在一些实施例中,可以使用用于总核的矩阵条目子集来训练核机器,其中用于总核的矩阵条目子集中的每个矩阵条目对应于已标记文档内的单词对。已识别的元素可以是对应于具有由核机器确定的高(预测)相似性/置信度得分的单词的元素。在这些实施例中,方法还可以包括输出由核机器确定的相似性/置信度得分。

[0036] 尽管图1中未示出,但是在一些实施例中,方法100还可以包括在步骤S130处识别

所述一个或多个元素之后,接收批准或拒绝识别的结果的用户输入和接收修改识别的结果的用户输入中的至少一个。在这些实施例中,如果接收到批准结果的用户输入,则方法100还可以包括(例如,经由显示器)输出结果,或执行结果的进一步处理(例如,以如下所述的方式进行字段确定和/或分组)。如果接收到修改识别的结果的用户输入,则方法100还可以包括基于用户输入执行修改和/或将修改的结果存储为新的已标记文档的至少一部分。

[0037] 在一些实施例中,方法还可以包括为所识别的一个或多个元素中的每一个元素,确定相应元素所对应的字段。这个确定操作也可以在核机器中使用总核来执行。将理解,在一些实施例中,在步骤S130处识别一个或多个元素的操作和为所识别的一个或多个元素中的每一个元素确定字段的操作可以作为核机器中的单个步骤/过程同时执行。

[0038] 在存在多个已识别的元素的一些实施例中,为所识别的元素进行字段确定的操作可以包括将多个已识别的元素分组为一个或多个组中,其中每个组对应于一个字段。这种分组可以基于单词分组算法。更具体地,将所述多个已识别的元素分组可以包括:

-(例如,通过未标记文档中的所述多个已识别的元素的使用阅读顺序算法的阅读顺序)生成所述多个已识别的元素的线性文本序列;

-(例如,使用核机器)确定所生成的线性文本序列中的每个元素的分类;以及

-将所生成的线性文本序列中的共享分类的各个元素的每个最大邻接子序列分组到相应的组中。

[0039] 为每个元素确定的分类可以是argmax分类。确定分类可以基于多个可能的分类,其中每个可能的分类对应于已标记文档中的字段或对应于元素不与已标记文档中的任何字段相关联的指示。

[0040] 在将所述多个已识别的元素分组(例如,使用如上所述的分组算法)的实施例中,方法还可以包括基于元素在组中的相对顺序将相应的行标识符分配给组中的每个元素。这允许属于表格格式的元素被如此输出(即,具有相同行标识符的元素被放置在相同的行中)。

[0041] 在一些实施例中,已标记文档中的所述一个或多个字段和一个或多个元素可以是以信息结构呈现的。例如,在账单中,所述一个或多个字段可以呈现在表格的标题中,并且所述一个或多个元素可以呈现在表格的标题下,每列对应于一个字段,每行对应于一个购买的商品。在这些实施例中,方法100还可以包括以相同(或相似)信息结构输出已识别的一个或多个元素及其对应字段。或者,方法100还可以包括仅输出所述一个或多个元素(无论字段和/或元素是否以相同/相似或不同的信息结构呈现)。

[0042] 方法还可以包括向已识别的一个或多个元素中的每一个分配新的标签,以指示相应元素所对应的字段。

[0043] 在一些实施例中,所述方法还可以包括将所述一个或多个已识别的元素中的每一个放置成归一化形式。例如,对应于日期信息的任何元素可以以DD/MM/YYYY的格式返回。这个操作可以基于字段不可知的规则(即,用于所有字段的一个规则),诸如用于移除已识别的元素的相应文本串中的任何在前或结尾标点符号的规则。或者,操作可以基于一个或多个字段特定规则,其中字段特定规则可以基于已标记文档中的标签的元数据。

[0044] 应当理解,在一些实施例中,在已标记文档中使用标签足以指示哪个(些)元素是感兴趣的元素,并且因此还指示应当识别和捕捉的未标记文档中的哪个(些)对应元素。然

而,在一些实施例中,所述方法还可包括,在步骤S110处确定一个或多个锚之前,获取模式,所述模式定义包含在已标记文档中的所述一个或多个字段以及与已标记文档中的每个字段相对应的元素是否以表格呈现。模式可以提供关于应从未标记文档捕捉什么数据的更明确的信息。这个模式可以用于例如在方法100的S110处确定锚,使得由模式定义的一个或多个字段每一个可以用作锚的点。

[0045] 尽管图1中未示出,但是在一些实施例中,方法100还可以包括在步骤S110处确定一个或多个锚之前,接收将一个或多个标签分配给已标记文档的元素的用户输入。在一些实施方式中,接收分配标签的用户输入可以包括:接收用户对已标记文档中的字段的选择,以及接收对与所选字段相对应的一个或多个元素的一个或多个用户选择。作为示例,元素的用户选择可以通过允许用户例如经由诸如触摸屏之类的用户接口围绕文档上的元素绘制矩形框来实现。与标签(其可以由矩形框表示)相对应的字段可以借助于视觉指示符来指示。例如,每个矩形框可以具有指示元素的对应字段的颜色。在这种情况下,对应于相同第一字段的元素的矩形框将具有相同的第一颜色,并且对应于相同第二字段的元素的矩形框具有相同的第二颜色,等等。此外,在一些实施例中,每个标签可以包含元数据。例如,这可以是用于指示字段是否呈数字的形式的元数据和/或用于指示与字段相对应的元素是否以表格格式呈现的元数据等。

[0046] 在一些实施例中,方法100还可以包括在确定一个或多个锚之前,获取多个候选已标记文档,以及接收从所述多个候选已标记文档中选择已标记文档的用户输入。替代地,方法100还可以包括,在确定一个或多个锚之前,获取多个候选已标记文档,以及基于所述多个候选已标记文档中的每一个与未标记文档之间的相似性度量,从所获取的多个候选已标记文档中确定已标记文档。在一些实施例中,从多个候选已标记文档中确定已标记文档可以基于相似性函数和最佳匹配算法。相似性函数可以是能够处理文档对(即,在此情况下,相应的候选已标记文档和未标记文档)并基于处理返回相似性度量的函数。相似性度量的示例可以是指示两个文档的相似度的分数。

[0047] 在一些实施例中,相似性函数可以基于单词袋模型,单词袋模型通过跨文档的公共单词的数量(其可以包括或可以不包括重复单词)来比较文档对。例如,如果文档由多个标记“a a a b c”组成,并且另一文档由多个标记“a a c c d”组成,则相似性函数可以返回相似性得分3,因为文档对共享三个标记:两个‘a’和一个‘c’。

[0048] 一旦所有候选已标记文档对与未标记文档对之间的所有相似性得分都已以这个方式确定,则最佳匹配算法可以随后被用来基于所述多个候选已标记文档中的每一个与未标记文档之间的相似性度量来选择“最佳匹配”。例如,最佳匹配算法可以用于选择具有最高相似性得分的文档对。

[0049] 在一些情况下,相似性得分阈值可以被用于确定素数多个候选标记文档中没有一个适合用于根据本文描述的方法执行数据捕捉。例如,如果相似性得分阈值被设置为4,并且如果文档对(每对包含候选已标记文档和未标记文档)的各个相似性得分中没有一个等于或超过4,则所述方法还可以包括提供输出以指示各个(当前)候选已标记文档中没有一个能够被用于数据捕捉。方法还可以包括获取多个新候选已标记文档的集合,并基于所述多个新候选已标记文档中的每一个与未标记文档之间的相似性度量来确定已标记文档。

[0050] 替代地或附加地,在一些情况下,相似性得分阈值可以用于确定多个候选标记文

档适合用于根据本文描述的方法执行数据捕捉。例如,如果相似性得分阈值被设置为4,并且如果多于一个文档对(每对包含候选已标记文档和未标记文档)具有等于或超过4的相似性得分,则在步骤S110处确定锚的步骤可以基于多个已标记文档(每个已标记文档对应于高于阈值的相似性得分)和未标记文档。如上所述,在这种情况下,每个锚可以包括匹配的多个点,所述多个点中的每个点都在不同的文档中。此外,在这种情况下,在步骤S120处生成总核的步骤可以基于所述多个已标记文档和未标记文档。换句话说,每个卷积核可以涉及所述多个已标记文档(每个标记文档对应于高于阈值的相似性分数)和未标记文档。

[0051] 图2是根据本公开的实施例的用于执行数据捕捉的系统的框图。更详细地,系统200用于基于已标记文档在未标记文档中执行数据捕捉。已标记文档包含一个或多个元素,每个元素关于字段被标记。每个元素可以由单词、单词串、或数字、或其组合来表示。在一些实施例中,已标记文档还可以包括由一个或多个标签指示的一个或多个字段。已标记文档和/或未标记文档可以是PDF文档或图像。

[0052] 如图2所示,系统200包括确定单元210、生成单元220和识别单元230。确定单元210配置为确定一个或多个锚。如上所述,锚包括在已标记文档与未标记文档之间匹配的点对。在一些实现中,确定单元210可以配置成为每个新的文档对(其中,所述对包括未标记文档和已标记文档)确定一个或多个锚的新集合。

[0053] 在一些实施例中,确定单元210可以配置为通过接收选择未标记文档和已标记文档中的点对的用户输入来确定锚。或者,在一些实施例中,确定单元210可以配置为通过以下步骤确定锚:

-(例如,使用OCR)提取已标记文档和未标记文档中的所有单词;

-通过将包含在已标记文档中的单词排序成线性文本序列(并且可选地,移除对应于已标记的元素的任何单词)来生成第一线性文本序列;

-通过将包含在未标记文档中的单词排序成线性文本序列来生成第二线性文本序列;

-识别第一和第二线性文本序列之间的匹配(如果在这个阶段对应于已标记的元素的单词没有被移除,则在匹配识别步骤中忽略这些单词);以及

-将所匹配的单词指定为锚。

[0054] 本上下文中的匹配(例如,在视觉上)涉及两个单词的固有相似性,以及在相应序列中的相应单词的线性顺序方面的相似性或匹配。在这些实施例中,第一和第二线性文本序列之间的匹配可以是就固有相似性和对齐相似性两者而言超过预定阈值的匹配。此外,确定单元210可以配置为基于对齐算法(例如,类似于或等同于尼德曼-翁施算法的算法)来识别匹配(更具体地,确定两个单词是否超过对齐相似性阈值)。

[0055] 在这些实施例中,确定单元210可以配置为使用阅读顺序算法对包含在已标记和未标记文档中的单词进行排序。此外,在这些实施例中,所述一个或多个锚中的每一个可以被定义为已标记文档中的相应匹配单词的中心与未标记文档中的相应匹配单词的中心之间的对应关系。

[0056] 在一些实施例中,确定单元210可以配置为通过(例如,使用图像对齐算法)识别在已标记文档和未标记文档两者中存在的视觉特征并且将所识别的视觉特征指定为锚来确定锚。图像对齐算法可以是基于以下中的至少一个的算法:关键点检测器(例如,高斯差、哈

里斯(Harris)、要跟踪的良好特征(GFTT))、局部不变描述符(例如,尺度不变特征变换(SIFT)、加速鲁棒特征(SURF)、以及定向FAST和旋转BRIEF(ORB)等)以及关键点匹配过程(例如,随机采样一致性(RANSAC))。

[0057] 在一些实施例中,确定单元210可以配置为通过以下操作来确定锚:围绕未标记文档的边界放置多个点且围绕已标记文档的边界放置相同数量的点;识别已标记文档中的点与未标记文档中的点之间的匹配;以及将所匹配的点指定为锚。

[0058] 尽管在本实施例中描述了为文档对(即,已标记和未标记文档)确定一个或多个锚,但是应当理解,上述用于确定单元210确定锚的过程能够应用到多于两个文档(例如,多个已标记文档和/或多个未标记文档)。此外,尽管在本实施例中描述了锚可以包括匹配的单个单词,但是在其他实施例中,锚可以包括匹配的短语而不是单词,例如,单词“total amount(总金额)”作为单个短语而不是两个单独的单词而被处理。

[0059] 生成单元220配置为使用所确定的一个或多个锚来生成总核。在一些实施例中,生成单元210可以配置为通过以下步骤生成总核:为所述一个或多个锚中的每一个生成卷积核,以及通过对所述一个或多个卷积核求和来生成总核。总核可以是所述一个或多个卷积核的加权和。

[0060] 每个卷积核可以涉及在已标记文档和未标记文档内和之间的所有单词对。在存在数据捕捉所基于的多于两个文档的一些实施例中,每个卷积核可以涉及在所有文档内和跨所有文档的所有单词对。每个卷积核可以测量在已标记文档和未标记文档内和之间(或跨涉及多于两个文档的所有文档)的单词的相似性,相似性是相应单词离锚的相应点的距离和相对于锚的相应点的相对位置的相似性的函数。卷积核的数学细节在上面参考图1提供,因此为了简洁起见,这里不再重复,应当理解,这些数学方程可应用于本实施例。此外,如以上参照图1所解释的,卷积核和关于文档的集合定义的总核每一个可以被表示为对称的半正定矩阵,其具有与文档的集合中的所有文档中的单词的数量相对应的维度,并且矩阵中的每个矩阵条目对应于文档的集合中的文档中的单词中的特定单词对之间的相应相似性。

[0061] 尽管上面描述了每个卷积核可以涉及单词对,但是在替代实施例中,代替单个单词的对或除了单个单词的对之外,卷积核可以涉及短语对(例如,每个短语对由多个单词组成)。

[0062] 识别单元230配置为在核机器中使用总核来识别在未标记文档中要捕捉的一个或一个以上元素。要捕捉的一个或多个元素在字段方面对应于包含在已标记文档中的所述一个或多个已标记的元素。核机器可以包括以下中的至少一个:支持向量机(SVM)、正则化核判别分析(RKDA)和高斯过程(GP)。可以基于由已标记文档中的相应标签所指示的一个或多个元素和一个或多个字段来训练核机器。在一些情况下,可以基于由已标记文档中的相应的(一个或多个)标签指示的一个或多个元素和一个或多个字段中的全部来训练核机器,以便最大化用于训练的样本的数量。在一些实施例中,可以使用用于总核的矩阵条目的子集来训练核机器,其中用于总核的矩阵条目的子集中的每个矩阵条目对应于已标记文档内的单词对。已识别的元素可以是对应于具有由核机器确定的高(预测)相似性/置信度得分的单词的元素。在这些实施例中,系统200可以包括输出单元,其配置为输出由核机器确定的(一个或多个)相似性/置信度得分。

[0063] 尽管图2中未示出,但在一些实施例中,系统200还可包括接收单元,其配置成在识

别单元230识别所述一个或多个元素之后,接收批准或拒绝识别的结果的用户输入和修改识别的结果的用户输入中的至少一个。此外,系统200可以包括输出单元,其配置为如果接收到批准结果的用户输入则(例如,经由显示器)输出结果,和/或执行单元,其配置为执行结果的进一步处理(例如,字段确定和/或分组,因此执行单元可以是确定单元)。如果接收到修改识别的结果的用户输入,则执行单元可以配置为基于用户输入执行修改,和/或系统200中的存储单元可以配置为将修改的结果存储为新的已标记文档的至少一部分。

[0064] 在一些实施例中,确定单元210还可以配置成为所识别的一个或多个元素中的每一个元素,确定相应元素所对应的字段。可以在核机器(例如由识别单元230用于识别(多个)元素的核机器)中使用总核来执行这个确定。将理解,在一些实施例中,确定单元210可以配置为作为核机器中的单个步骤/过程同时执行所述一个或多个元素的识别和为所识别的一个或多个元素中的每一个的字段确定。

[0065] 尽管图2中未示出,但是在存在多个已识别的元素的一些实施例中,确定单元210可以配置为通过将所述多个已识别的元素分组为一个或多个组来为所所述多个已识别的元素确定字段。在这些实施例中,每个组可以对应于一个字段。分组可以基于单词分组算法。更具体地,确定单元210可以配置为通过以下操作将所述多个已识别的元素分组:

-生成未标记文档中的所述多个已识别的元素的线性文本序列;

-(例如,使用核机器)确定所生成的线性文本序列中的每个元素的分类;以及

-将所生成的线性文本序列中的共享分类的各个元素的每个最大邻接子序列分组到相应的组中。

[0066] 为每个元素确定的分类可以是argmax分类。确定单元210确定分类可以基于多个可能的分类,其中每个可能的分类对应于已标记文档中的字段或者对应于元素不与已标记文档中的任何字段相关联的指示。

[0067] 此外,在这些实施例中,系统200还可以包括分配单元,其配置为基于元素在组中的相应顺序将相应的行标识符分配给组中的每个元素。

[0068] 在一些实施例中,已标记文档中的所述一个或多个字段和一个或多个元素可以是以信息结构呈现的。在这些实施例中,系统200还可以包括输出单元,其配置为以相同(或相似)信息结构输出所识别的一个或多个元素及其对应字段。或者,输出单元可以配置为仅输出所述一个或多个元素(无论字段和/或元素是否以相同/相似或不同的信息结构呈现)。

[0069] 系统200还可以包括分配单元,其配置为向所识别的一个或多个元素中的每一个分配新的标签,以指示相应元素所对应的字段。

[0070] 在一些实施例中,确定单元210还可以配置为将所述一个或多个已识别的元素中的每一个放置成归一化形式。这个操作可以基于字段不可知的规则或者基于一个或多个字段特定的规则。

[0071] 应当理解,在一些实施例中,在已标记文档中使用标签足以指示哪个(些)元素是感兴趣的元素,并且因此还指示应当识别和捕捉未标记文档中的哪个(些)对应元素。然而,在一些实施例中,方法还可以包括获取单元,其配置为在确定单元210确定一个或多个锚之前,获取模式,所述模式定义了包含在已标记文档中的所述一个或多个字段以及与已标记文档中的每个字段相对应的元素是否以表格呈现。模式可以提供关于应从未标记文档捕捉什么数据的更明显信息。这个模式可例如由确定单元210在其确定锚时使用,使得由模式定

义的所述一个或多个字段每一个可以被用作锚的点。

[0072] 在一些实施例中,系统200还可以包括接收单元,其配置为在确定单元210确定一个或多个锚之前,接收将一个或多个标签分配给已标记文档的元素的用户输入。在一些实现中,接收单元可以配置为通过以下操作来接收分配标签的用户输入:接收对已标记文档中的字段的用户选择,以及接收对与所选字段相对应的一个或多个元素的一个或多个用户选择。作为示例,对元素的用户选择可以通过允许用户例如经由用户界面围绕文档上的元素绘制矩形框来实现。与标签(其可以由矩形框表示)相对应的字段可以借助于视觉指示符来指示。例如,每个矩形框可以具有指示元素的对应字段的颜色。此外,在一些实施例中,每个标签可以包含元数据(例如,用于指示字段是否呈数字的形式的元数据、用于指示与字段相对应的元素是否呈表格格式的元数据等)。

[0073] 在一些实施例中,系统200还可以包括获取单元,其配置为获取多个候选标记文档。在这些实施例中,系统200还可以包括接收单元,其配置成接收从所述多个候选已标记文档中选择已标记文档的用户输入。替代地,在这些实施例中,确定单元210还可以配置成基于所述多个候选已标记文档中的每个文档与未标记文档之间的相似性度量,从所获取的多个候选已标记文档中确定已标记文档。更具体地,确定单元210可以配置为基于相似性函数和最佳匹配算法来确定已标记文档。以上参考图1详细描述和解释了相似性函数和最佳匹配算法,因此这些组件的技术功能性将不在此重复,而是将被理解为也等同地应用。这些实施例中描述的获取单元、接收单元和确定单元的操作可以在确定单元210确定所述一个或多个锚之前执行。

[0074] 图3A和图3B是示出根据本公开的实施例的示例性已标记文档和示例性未标记文档的图。如图3A和3B所示,提供了已标记文档310和未标记文档320。在这个实施例中,已标记文档310和未标记文档320是账单(即,半结构化文档)。

[0075] 在已标记文档310中,信息以字段-元素对格式或以表格格式呈现。例如,“账单#”和“00123456”作为字段-元素对呈现,“账单日期”和“2021年9月2日”作为字段-元素对呈现,并且“总计”和“4596”作为字段-元素对呈现。在这些示例中,“账单#”、“账单日期”和“总计”被认为是字段,而“00123456”、“2021年9月2日”和“4596”被认为是相应的元素。

[0076] 在这个示例中,其余信息以表格格式呈现,其中元素对应于字段“描述”、“数量”、“单位价格”和“金额”中的一个。更详细地,元素“佳能5D”、“存储卡”和“镜头”对应于“描述”,这些是购买的商品的描述。元素“1”、“2”和“1”对应于“数量”,这些数字指示每个购买商品的数量。元素“2499”、“149”和“1799”对应于“单位价格”,这些数字指示每个购买的商品的单位价格。元素“2499”、“298”和“1799”对应于“金额”,这些数字指示每个购买的商品的总金额(价格)。

[0077] 如图3A中可以看到的,所有元素都被标记(在附图中通过围绕每个元素的框来指示,这些框具有相应的附图标记311、312、313A-C、314A-C、315A-C、316A-C和317)。这可以通过接收指示文档中的哪些单词和/或短语(或图像像素的哪些部分)要被视为元素的用户输入来实现。

[0078] 标签可以指示元素对应的字段。在这个示例中,用于“00123456”的标签(即标签311)可以指示相应元素对应于字段“账单#”。类似地,用于“2021年9月2日”的标签(即,标签312)可以指示相应的元素对应于字段“账单日期”。标签313A-C可以指示元素“佳能5D”、“存

储卡”和“镜头”对应于字段“描述”,标签314A-C可以指示元素“1”、“2”和“1”对应于字段“数量”,标签315A-C可以指示元素“2499”、“149”和“1799”对应于字段“金额”,标签317可以指示元素“4596”对应于字段“总计”。

[0079] 现在参考图3A中的未标记文档320,可以看出账单遵循与已标记文档310类似的信息结构。具体地,未标记文档320包括相同的字段:“账单#”、“账单日期”、“描述”、“数量”、“单位价格”、“金额”和“总计”。通过应用如上参考图1所述的计算机实现的方法100或使用如上参考图2所述的系统200,能够识别并输出未标记文档中的元素。

[0080] 图3B中示出了这个输出的示例,其中未标记文档320现在被提供有标签321、322、323、324、325、326和327,每一个对应于已识别的元素。在这种情况下,各个已识别的元素是“00123457”、“2021年9月3日”、“索尼ZV-1”、“1”、“699”、“699”和“699”。这些已识别的元素在它们的字段方面与已经在已标记文档310中标记的元素相对应。具体地,元素“00123457”在其字段(“账单#”)方面对应于元素“00123456”,元素“2021年9月3日”在字段(“账单日期”)方面对应于元素“2021年9月2日”,元素“索尼ZV-1”在其字段(“描述”)方面对应于元素“佳能5D”、“存储卡”和“镜头”,元素“1”在其字段(“数量”)方面对应于元素“1”、“2”和“1”(标记为314A-C),单位价格“699”(标记为325)在其字段(“单位价格”)方面对应于元素“2499”、“149”、“1799”(标记为315A-C),元素“699”(标记为326)在其字段(“金额”)方面对应于元素“2499”、“298”和“1799”(标记为316A-C),以及元素“699”(标记为326)其字段(“总计”)方面对应于元素“4596”。

[0081] 图4是示出确定图3A和3B的已标记文档和未标记文档之间的锚的示例性过程的图。图4示出了上部线性文本序列和下部线性文本序列。可以通过提取已标记文档310中的所有单词并将所提取的单词(例如,通过其阅读顺序)排序成线性文本序列来生成上部线性文本序列。类似地,可以通过提取未标记文档320中的所有单词并将所提取的单词(例如,通过其阅读顺序)排序成线性文本序列来生成下部线性文本序列。

[0082] 在这个示例中,在确定锚的过程中忽略上部线性文本序列中的对应于已识别的元素(例如,“00123456”、“2021年9月2日”、“佳能5D”等)的任何单词。或者,在其他实施例或示例中,在确定锚的过程中,可以从线性文本序列中移除与已识别的元素相对应的单词。

[0083] 一旦已经生成了对应于已标记和未标记文档310、320的线性文本序列,就可以在这两个线性文本序列之间识别匹配。如上部和下部线性文本序列之间的箭头所示,在这个示例中已经识别十个匹配(从左到右):“账单”、“#”、“账单”、“日期”、“描述”、“数量”、“单位”、“价格”、“金额”和“总计”。注意,至少在这个示例中,每个单个文本块可以被认为是“单词”。因此,“账单”和“#”被认为是两个(单独的)单词,并且类似地,“单位”和“价格”被认为是两个单词,即使“单位价格”被共同认为是一个字段。

[0084] 然后,将已经识别的每个匹配指定为锚,从而在这个示例中提供分别对应于单词“账单”、“#”、“账单”、“日期”、“描述”、“数量”、“单位”、“价格”、“金额”和“总计”的十个锚。尽管在这个示例中识别的匹配都是精确匹配,但是将理解,在一些其他示例或实施例中,匹配可能(例如,由于OCR错误)不是100%匹配。

[0085] 图5是示出根据图4所示的示例性过程确定的锚的图。为了示出本公开的一个方面,图5仅如指向已标记文档510(其与图3A和3B的已标记文档310相同)和未标记文档520(其与图3A和3B的已标记文档320相同)中的单词“数量”的上拱形箭头所示那样聚焦在锚

上。这个锚是经由以上参考图4所述的过程识别的各个匹配中的一个。

[0086] 如上参考图1和2所述,在一些实施例中,总核的生成包括为所述一个或多个锚中的每一个生成卷积核,每个卷积核涉及在已标记文档和未标记文档内和之间的所有单词和/或短语对。更具体地,每个卷积核可以测量在已标记文档和未标记文档内和之间的单词和/或短语的相似性,相似性是相应单词和/或短语离锚的相应点的距离和相对于锚的相应点的相对位置的相似性的函数。因此,在本示例中,锚“数量”的卷积核可以测量在已标记文档510内、在未标记文档520内以及在已标记文档510和未标记文档520之间的单词和/或短语相对于锚的相似性。

[0087] 作为如何测量单词的相似性的示例,图5聚焦于在已标记文档510中的“1”和未标记文档520中的“1”之间的相似性(如指向两个单词的下部箭头所示)。这个相似性是以下的函数:在已标记文档510中的“1”和“数量”之间(如由文档中的短箭头所指示的)的距离、在未标记文档520中的“1”和“数量”之间(如由文档中的短箭头所指示的)的距离、以及在每个文档中单词“1”相对于单词“数量”的相对位置的相似性。

[0088] 如上参考多个数学方程所述,至少在一些实现中,仅当单词和/或短语都靠近锚并且它们相对于锚的位置相似时(通过取偏移核和相对核的乘积),才认为它们相似。在这种情况下,由于已标记文档510中的“1”和“数量”的接近性、未标记文档520中的“1”和“数量”的接近性、以及在每个文档中单词“1”相对于单词“数量”的相对位置的相似性(如两个“1”位于单词“数量”下方),由卷积核测量的相似性将是高的。相反,由于它们都更远离相应文档中的单词“数量”并且它们相对于“数量”的相对位置不同的事实,相对于相同锚的、在已标记文档510中的单词“4596”和未标记文档520中的单词“00123457”之间的所测量的相似性将更低。

[0089] 可以以类似的方式为其它锚生成卷积核,其中,每个卷积核涉及相对于相应锚在已标记文档和未标记文档内和之间的所有单词和/或字段对。一旦所有卷积核都已经生成,它们就可以被求和以生成为了识别来自未标记文档520的元素的目的的总核。

[0090] 图6是示出根据本公开的实施例的示例性已标记文档和示例性未标记文档的图。如图6所示,提供了已标记文档610和未标记文档620。在这个实施例中,已标记文档和未标记文档是账单(即,半结构化文档)。

[0091] 在已标记文档610中,一些信息以字段-元素对格式呈现。例如,将“账单编号”和“237481709”呈现为字段-元素对,将“应付总额”和“£1901.98”呈现为字段-元素对,并且将“到期日”和“95年5月16日”呈现为字段-元素对。在这些示例中,“账单编号”、“应付总额”和“到期日”被认为是字段,而“237481709”、“£1901.98”和“95年5月16日”被认为是相应的元素。

[0092] 在已标记文档610中的一些其他信息以表格格式呈现,其中元素对应于字段“描述”、“价格”、“折扣”、“增值税”和“金额”中的一个。例如,元素“能量棒”、“100%果汁苹果葡萄”和“天然矿泉水”等对应于“描述”,这些是对购买的商品的描述。

[0093] 此外,在这个示例中,一些信息被呈现为(没有明确的对应字段的)元素。例如,元素“伊尔克利LS29 9EU,布里奇巷56号,桥内饰公司”是地址,因此其被理解为对应于“地址”字段。尽管这个字段没有明确地呈现在已标记文档中,但是分配给这个元素的标签612可以指示元素对应于字段“地址”。

[0094] 如图6中可见,仅标记(在附图中以相应的附图标记611、612、613、614和615通过围绕每个元素的框来指示)已标记文档610中的各个元素中的一些,这可以通过接收指示文档中的哪些单词和/或短语(或图像像素的哪些部分)将被视为元素的用户输入来实现。在这个示例中,并非已标记文档610中的所有元素都被分配了标签,这可能是因为一些元素(未标记)不是感兴趣的,因此不期望或不必捕捉在字段方面与这些元素相对应的元素。例如,即使“将购买订单”和“855496”呈现为字段-元素对,也没有给“855496”分配标签。作为另一示例,在已标记文档610中没有标记与“价格”相对应的元素。

[0095] 在这个示例中,标签可以指示元素对应的字段。例如,如上所述,标签612可以指示相应的元素对应于字段“地址”。类似地,“237481709”的标签(即标签611)可以指示相应元素对应于字段“账单编号”;“£1901.98”的标签(即标签613)可以指示相应元素对应于字段“应付总额”;“95年5月16日”的标签(即标签614)可以指示相应元素对应于字段“到期日”;并且位于“描述”之下的各个元素的标签615可以指示这些元素对应于字段“描述”。

[0096] 现在参考图6中的未标记文档620,可以看出账单遵循与已标记文档类似的信息结构。具体地,未标记文档620包括相同的字段,例如“账单编号”、“应付总额”、“到期日”和“描述”。通过应用如上参考图1所述的计算机实现的方法100或使用如上参考图2所述的系统200,可以识别并输出未标记文档中的元素。

[0097] 作为数据捕捉过程的锚确定方面的示例,图7A示出了基于图6的已标记文档和未标记文档确定的文本锚。图7A中的已标记文档710对应于图6的已标记文档610,而未标记文档720对应于图6的未标记文档620。本示例中的文本锚由跨两个文档的箭头指示,该箭头指向已标记文档710中的单词“描述”711和未标记文档721中的单词“描述”721。这个锚可以通过多个步骤来确定:

(1)(例如,使用OCR)提取已标记文档710和未标记文档720中的所有单词;

(2)通过将包含在已标记文档710中的单词排序成线性文本序列来生成第一线性文本序列;

(3)通过将包含在未标记文档720中的单词排序成线性文本序列来生成第二线性文本序列;

(4)识别第一和第二线性文本序列之间的匹配(忽略对应于已标记元素的单词,例如“能量棒”);以及

(5)将所匹配的单词指定为锚。

[0098] 一旦确定了这个锚(“描述”),就可以为锚生成卷积核,其中卷积核涉及在已标记文档710和未标记文档720内和之间的所有单词和/或短语对。卷积核测量在已标记文档和未标记文档内和之间的单词和/或短语的相似性,相似性是相应单词离锚的相应点的距离以及相对于锚的相应点的相对位置的相似性的函数。例如,对于这两个文档,已标记文档中的短语“能量棒”712和单词“酸面包粉”722将具有高的测量相似度(相对于锚“描述”),这是由于它们都在它们各自的文档中与单词“描述”紧密接近,并且它们相对于单词“描述”的相对位置是相似的(即,在各自的文档中“能量棒”和“酸面包粉”都在“描述”正下方)。

[0099] 可以以类似的方式为其它锚生成卷积核,其中,每个卷积核涉及相对于相应锚在已标记文档和未标记文档内和之间的所有单词和/或短语对。一旦所有卷积核都已经生成,它们就可以被求和以生成为了识别来自未标记文档720的元素的目的的总核。

[0100] 作为数据捕捉过程的锚确定方面的另一示例,图7B示出了基于已标记文档710和未标记文档720确定的视觉锚。本示例中的视觉锚由跨两个文档的箭头指示,所述箭头指向已标记文档710中的水平分隔线713和未标记文档720中的水平分隔线723。这个锚可以通过识别存在于已标记文档和未标记文档两者中的视觉特征并将所识别的视觉特征指定为锚来确定。在这个示例中,已识别的视觉特征是两个文档中的水平分隔线。视觉特征的识别可以基于图像对齐算法,诸如基于SIFT的算法。

[0101] 一旦确定了这个视觉锚,就可以为锚生成卷积核,其中,卷积核涉及在已标记文档710和未标记文档720内或之间的所有单词和/或短语对。卷积核测量在已标记文档和未标记文档内和之间的单词和/或短语的相似性,所述相似性是相应单词离锚的相应点的距离以及相对于锚的相应点的相对位置的相似性的函数。

[0102] 例如,在已标记文档710中的短语“100%果汁苹果葡萄”和在未标记文档720中的短语“蛋白质加棒饼干&奶油味”将具有(相对于视觉锚)高的测量相似度,这是由于这些短语相对紧密地接近它们各自文档中的水平分隔线,并且它们相对于水平分隔线的相对位置是相似的。相反,由于至少短语“胡椒蛋黄酱”位于更远离水平分隔线和这些短语(“胡椒蛋黄酱”对“洋葱圈”)的相对位置基本上不同的事实,已标记文档710中的短语“胡椒蛋黄酱”和未标记文档中的短语“洋葱圈”具有(相对于视觉锚)较低的测量相似度。

[0103] 图8示出了已标记文档810、第一未标记文档820和第二未标记文档830。已标记文档810对应于图6所示的已标记文档610,且第一未标记文档820对应于图6所示的未标记文档620。第二未标记文档630也是采用与其它两个文档类似的信息结构的账单。在这个示例中,已经为已标记文档810和未标记文档820确定锚,其由指向(已标记文档810中的)框816和(第一已标记文档820中的)框826的箭头指示且由这两个文档中的锚符号指示。当为其在此情况下是第二未标记文档830的新未标记文档(例如,属于与820相同批次未标记文档的文档)确定锚时,这个锚可以用作参考。在图8中可见,可以以此方式确定由指向(在已标记文档810中的)框816和(在第二未标记文档830中的)框836的箭头指示且由两个文档中的锚符号指示的新锚。

[0104] 图9是示出根据本公开的实施例的用于执行数据捕捉的拓扑的图。拓扑900可以包括客户端计算设备905和配置为通过网络910通信地耦合的服务器920。

[0105] 网络910可以是有线或无线网络,诸如因特网、内联网、局域网(LAN)、广域网(WAN)、近场通信(NFC)网络、蓝牙、红外、射频、蜂窝网络或另一类型的网络。将理解,网络910可以是多个不同种类的有线或无线网络的组合。

[0106] 每个客户端计算设备905可以是智能电话、平板电脑、笔记本电脑、个人数据助理或具有配置为处理指令并且连接到网络910的一个或多个部分的硬件处理器的任何其他类型的移动设备。每个客户端计算设备905可以具有配置为允许用户与客户端计算设备905的处理器交互的图形用户界面。

[0107] 服务器920可以包括驻留在特定位置的物理计算设备,或者可以部署在云计算网络环境中。在本公开中,“云计算”可以被定义为用于使得能够无处不在的、方便的、按需网络访问可配置计算资源(例如,网络、服务器、存储、应用和服务)的共享池的模型,所述可配置计算资源可以经由虚拟化被快速供应并且以最小的管理工作或服务提供商交互被释放,并且然后被相应地缩放。云模型可以由各种特性(例如,按需自助服务、广域网络访问、资源

池、快速弹性、测量服务等)、服务模型(例如,软件即服务(SaaS)、平台即服务(PaaS)、基础设施即服务(IaaS)、以及部署模型(例如,私有云、社区云、公共云、混合云等)组成。

[0108] 应当理解,所述说明仅作为示例;在不偏离权利要求书中限定的本公开的范围的情况下,可以对所描述的实施例进行变更和修改。

[0109] 尽管为了说明的目的,基于当前被认为是最实际和优选的实现方式,已经详细描述了本技术,但是应当理解,这样的细节仅仅是为了该目的,并且本技术不限于所公开的实现方式,而是相反,旨在覆盖在所附权利要求的范围内的修改和等同布置。例如,应当理解,本技术预期,在可能的程度上,任何实施方式的一个或多个特征可以与任何其他实施方式的一个或多个特征组合。

[0110] 在整个说明书中,对“一个实施例”、“实施例”或“一个示例”或“示例”的引用意味着结合实施例或示例描述的特定特征、结构或特性被包括在本公开的至少一个实施例中。因此,在本说明书中各处出现的短语“在一个实施例中”、“在实施例中”、“一个示例”或“示例”不一定都指相同的实施例或示例。此外,在一个或多个实施例或示例中,特定特征、结构或特性可以以任何合适的组合和/或子组合进行组合。此外,应当理解,这里提供的附图是为了向本领域普通技术人员进行解释的目的,并且附图不一定按比例绘制。

[0111] 根据本公开的实施例可以被实现为设备、方法或计算机程序产品。例如,在一些实施例中,可以提供一种包括指令的计算机程序产品,当由处理器执行程序时,所述指令使处理器执行关于图1描述的方法。作为另一个示例,在一些实施例中,可以提供一种包括指令的计算机可读存储介质,当由计算机执行时,所述指令使计算机执行图1中描述的方法的步骤。因此,本发明的实施例可以采取完全硬件实施例、完全软件实施例(包括固件、驻留软件、微代码等)或组合了软件和硬件方面的实施例的形式,其在这里可以被统称为“模块”或“系统”。此外,本公开的实施例可以采取在任何有形表达介质中体现的计算机程序产品的形式,该有形表达介质具有在该介质中体现的计算机可用程序代码。

[0112] 尽管结合示例性计算系统环境进行了描述,但是本公开的实施例可用于众多其它通用或专用计算系统环境或配置。可以适用于本公开的各方面的公知的计算系统、环境和/或配置的示例包括但不限于移动计算装置、个人计算机(例如,台式计算机)、服务器计算机、手持式或膝上型装置、多处理器系统、游戏控制台、基于微处理器的系统、机顶盒、可编程消费电子产品、移动电话、网络PC、小型机、大型计算机、包括任何上述系统或装置的分布式计算环境等。

[0113] 可以利用一个或多个计算机可用或计算机可读介质的任何组合。例如,计算机可读介质可以包括便携式计算机磁盘、硬盘、随机存取存储器(RAM)装置、只读存储器(ROM)装置、可擦除可编程只读存储器(EPROM或闪存)装置、便携式光盘只读存储器(CDROM)、光存储设备和磁存储装置中的一个或多个。用于执行本公开的实施例的操作的计算机程序代码可以以一种或多种编程语言的任何组合来编写。

[0114] 流程图和框图示出了根据本公开的各种实施例的系统、方法和计算机程序产品的可能实现的架构、功能和操作。在这点上,流程图或框图中的每个框可以表示代码的模块、段或部分,其包括用于实现指定的逻辑功能的一个或多个可执行指令。还将注意,框图和/或流程图的每个框以及框图和/或流程图中的框的组合可以由执行指定功能或动作的基于专用硬件的系统或专用硬件和计算机指令的组合来实现。这些计算机程序指令也可以存储

在计算机可读介质中,其可以引导计算机或其他可编程数据处理装置以特定方式工作,使得存储在计算机可读介质中的指令产生包括实现流程图和/或框图中指定的功能/动作的指令装置的制造物品。

以下是本公开的实施例的列表:

1.一种用于基于已标记文档在未标记文档中执行数据捕捉的计算机实现的方法,其中所述已标记文档包含一个或多个元素,每个元素关于字段被标记,所述方法包括:

确定一个或多个锚,其中所述锚包括在所述已标记文档与所述未标记文档之间匹配的点对;

使用所确定的一个或多个锚来生成总核;以及

在核机器中使用所述总核识别在所述未标记文档中要捕捉的一个或多个元素,其中所述要捕捉的一个或多个元素在字段方面对应于在所述已标记文档中包含的所述一个或多个已标记的元素。

2.根据实施例1所述的方法,其中生成所述总核包括:

为所述一个或多个锚中的每一个生成卷积核;以及

通过对所述一个或多个卷积核求和来生成所述总核。

3.根据实施例2所述的方法,其中,所述总核是所述一个或多个卷积核的加权和。

4.根据实施例2或实施例3所述的方法,其中,每个卷积核涉及在所述已标记文档和所述未标记文档内和之间的所有单词对。

5.根据实施例2至4中任一项所述的方法,其中,每个卷积核测量在所述已标记文档和所述未标记文档内和之间的单词的相似性,其中所述相似性是相应单词距所述锚的相应点的距离以及相对于所述锚的相应点的相对位置的相似性的函数。

6.根据实施例2至5中任一项所述的方法,其中,所述已标记文档被定义为$d=d_1$,所述未标记文档被定义为$d=d_2$,d是文档,并且每个文档被定义为单词序列$w_d = w_{d,1},\cdots,w_{d,n_d}$,$n_d$是相应文档中的单词数量,以及

其中,每个卷积核被定义为$k(w_{d,i},w_{d',j};a)=k_{off}(z)k_{off}(z')k_{rel}(z,z')$,每个单词对被表示为$(w_{d,i},w_{d',j})$,以及

其中a是相应的锚,并且被定义为$a=(a_1,a_2)$,a_1是d_1中的锚的对应点的位置,并且a_2是d_2中的锚的对应点的位置,并且

其中$k_{off}(z)=k_{RBF}(||zA_{off}||_2)$,$k_{rel}(z,z')=k_{RBF}(||(z-z')A_{rel}||_2)$,以及$k_{RBF}(r) = e^{-\frac{r^2}{2}}$,

其中$A_{off} = \begin{pmatrix} \frac{1}{\sigma_{x,off}} & 0 \\ 0 & \frac{1}{\sigma_{y,off}} \end{pmatrix}$并且$A_{rel} = \begin{pmatrix} \frac{1}{\sigma_{x,rel}} & 0 \\ 0 & \frac{1}{\sigma_{y,rel}} \end{pmatrix}$,

并且其中r是距离,σ是长度尺度。

7.根据实施例6所述的方法,其中,$d=d'$和/或$d\neq d'$。

8.根据前述实施例中任一项所述的方法,还包括在识别所述未标记文档中的所述一个或多个元素之后,进行以下中的至少一个:

接收批准或拒绝所述识别的结果的用户输入;以及

接收修改所述识别的结果的用户输入。

9.根据前述实施例中任一项所述的方法,其中,所述已标记文档还包括由一个或多个标签指示的一个或多个字段。

10.根据前述实施例中任一项所述的方法,还包括为所述已识别的一个或多个元素中的每个元素确定相应元素所对应的字段,其中所述确定在核机器中使用所述总核。

11.根据实施例10所述的方法,其中,在所述未标记文档中识别多个元素,并且其中确定所述已识别的一个或多个元素中的每个元素所对应的字段包括将所述多个已识别的元素分组为一个或多个组,每个组对应一个字段。

12.根据实施例11所述的方法,其中,将所述多个已识别的元素分组基于单词分组算法。

13.根据实施例11或实施例12所述的方法,其中将所述多个已识别的元素分组包括:

生成所述未标记文档中的所述多个已识别的元素的线性文本序列;

确定所生成的线性文本序列中的每个元素的argmax分类;以及

将所生成的线性文本序列中的共享argmax分类的各个元素的每个最大邻接子序列分组到相应的组中。

14.根据实施例11至13中任一项所述的方法,还包括基于元素在组中的相对顺序将相应的行标识符分配给组中的每个元素。

15.根据实施例9至14中任一项所述的方法,其中,所述已标记文档中的所述一个或多个字段和一个或多个元素是以信息结构呈现的,并且所述方法还包括以相同的信息结构输出所述已识别的一个或多个元素及其对应的字段。

16.根据实施例9至15中任一项的方法,还包括在确定一个或多个锚之前,获取模式,其定义包含在所述已标记文档中的所述一个或多个字段以及与所述已标记文档中的每个字段相对应的元素是否以表格呈现。

17.根据前述实施例中任一项所述的方法,其中,确定锚包括:

提取所述已标记文档和所述未标记文档中的所有单词;

通过将包含在所述已标记文档中的单词排序成线性文本序列并移除对应于已标记的元素的任何单词来生成第一线性文本序列;

通过将包含在所述未标记文档中的单词排序成线性文本序列来生成第二线性文本序列;

识别所述第一与第二线性文本序列之间的匹配;以及

将所匹配的单词指定为所述锚。

18.按照实施例17所述的方法,其中,所述第一和第二线性文本序列之间的匹配是超过预定相似性阈值和预定对齐阈值的匹配。

19.根据实施例17至实施例18所述的方法,其中,提取所述已标记文档和所述未标记文档中的单词是基于光学字符识别OCR的。

20.根据实施例17至19中任一项所述的方法,其中,将包含在所述已标记和未标记文档中的单词排序是使用阅读顺序算法来执行的。

21.根据实施例17至20中任一项所述的方法,其中,所述一个或多个锚中的每一个

被定义为所述已标记文档中的相应匹配单词的中心与所述未标记文档中的相应匹配单词的中心之间的对应关系。

22.根据实施例17至21中任一项所述的方法,其中,识别所述第一与第二线性文本序列之间的匹配是基于对齐算法的。

23.根据实施方案22的方法,其中,对齐算法类似于或等同于尼德曼-翁施算法。

24.根据前述实施例中任一项所述的方法,其中,确定锚包括:

识别存在于所述已标记文档和所述未标记文档两者中的视觉特征;以及

将所识别的视觉特征指定为所述锚。

25.根据实施例24所述的方法,其中,识别视觉特征是基于图像对齐算法的。

26.根据实施例25所述的方法,其中,图像对齐算法基于以下中的至少一项:关键点检测器、局部不变描述符和关键点匹配过程。

27.根据前述实施例中任一项所述的方法,其中,确定锚包括:

围绕所述未标记文档的边界放置多个点和围绕所述已标记文档的边界放置相同数量的点;

识别所述已标记文档中的点与所述未标记文档中的点之间的匹配;以及

将所匹配的点指定为所述锚。

28.根据前述实施方案中任一项所述的方法,还包括在确定一个或多个锚之前,接收将所述一个或多个标签分配给所述已标记文档的元素的用户输入。

29.根据前述实施方式中的任一项所述的方法,还包括在确定一个或多个锚之前:

获取多个候选已标记文档;以及

基于所述多个候选已标记文档中的每一个与所述未标记文档之间的相似性度量,从所获取的多个候选已标记文档中确定所述已标记文档。

30.根据前述实施例中任一项所述的方法,其中,所述核机器包括以下中的至少一个:支持向量机SVM,正则化核判别分析RKDA,以及高斯过程GP。

31.根据前述实施方案中任一项所述的方法,其中,所述已标记文档呈PDF文档或图像的形式。

32.一种用于基于已标记文档在未标记文档中执行数据捕捉的系统,其中所述已标记文档包含一个或多个元素,每个元素关于字段被标记,所述系统包括:

确定单元,其配置为确定一个或多个锚,其中锚包括在所述已标记文档与所述未标记文档之间匹配的对应点对;

生成单元,其配置为使用所确定的一个或多个锚来生成总核;以及

识别单元,其配置为在核机器中使用所述总核,识别在所述未标记文档中要捕捉的包含要捕捉的信息的一个或多个元素,其中,所述要捕捉的一个或多个元素在字段方面对应于包含在所述已标记文档中的所述一个或多个已标记的元素。

33.一种计算机可读存储介质,包括指令,所述指令在由计算机执行时使所述计算机执行实施例1至31中任一项所述的方法的步骤。

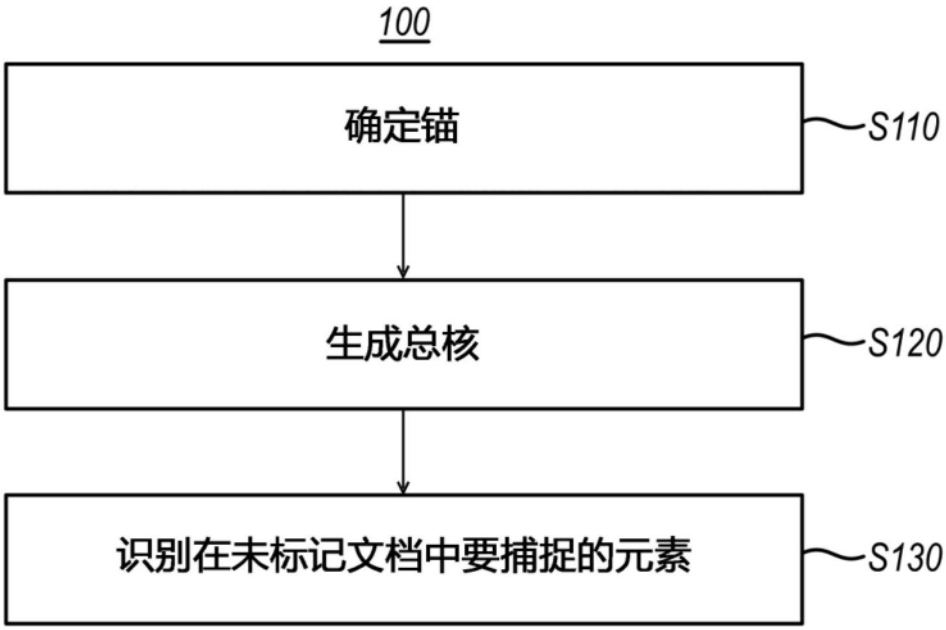

100
确定锚
S110
生成总核
S120
识别在未标记文档中要捕捉的元素
S130

图1

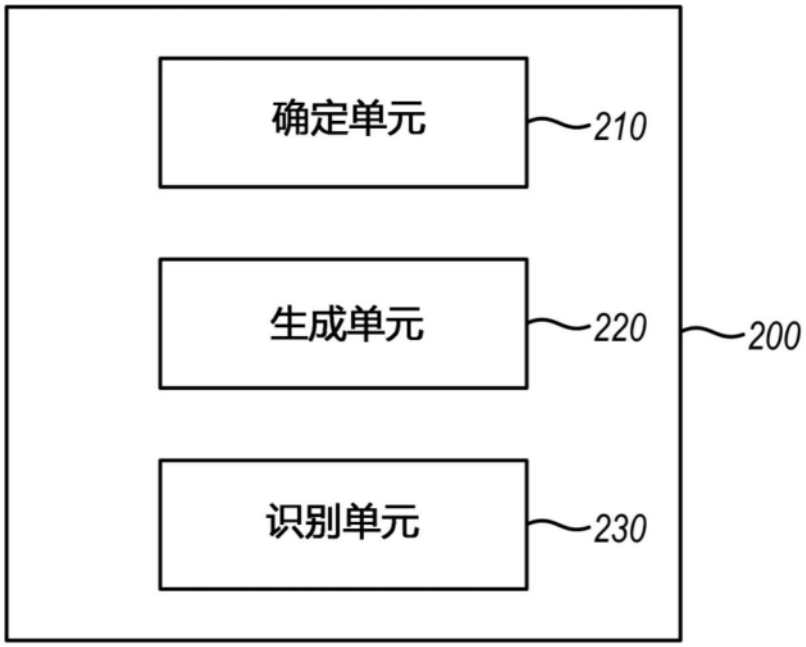

确定单元
210
生成单元
220
200
识别单元
230

图2

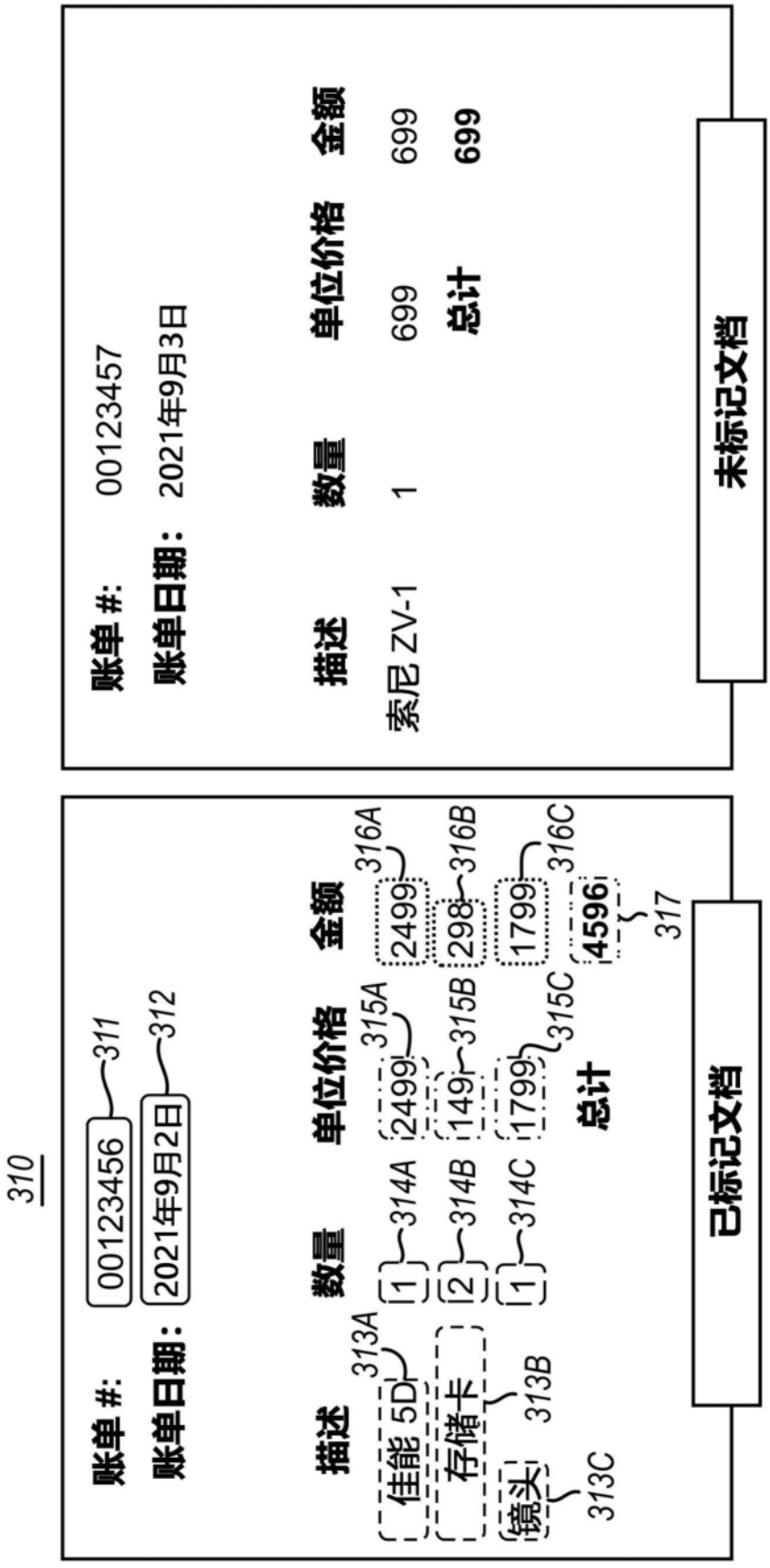

310
账单 #: 00123456
311
账单日期: 2021年9月2日
312
描述
数量
单位价格
金额
313A
佳能 5D
1
314A
2499
315A
2499
316A
存储卡
2
314B
149
315B
298
316B
镜头
313B
1
314C
1799
315C
1799
316C
313C
总计
4596
317
已标记文档
账单 #: 00123457
账单日期: 2021年9月3日
描述
数量
单位价格
金额
索尼 ZV-1
1
699
699
总计
699
未标记文档

图3A

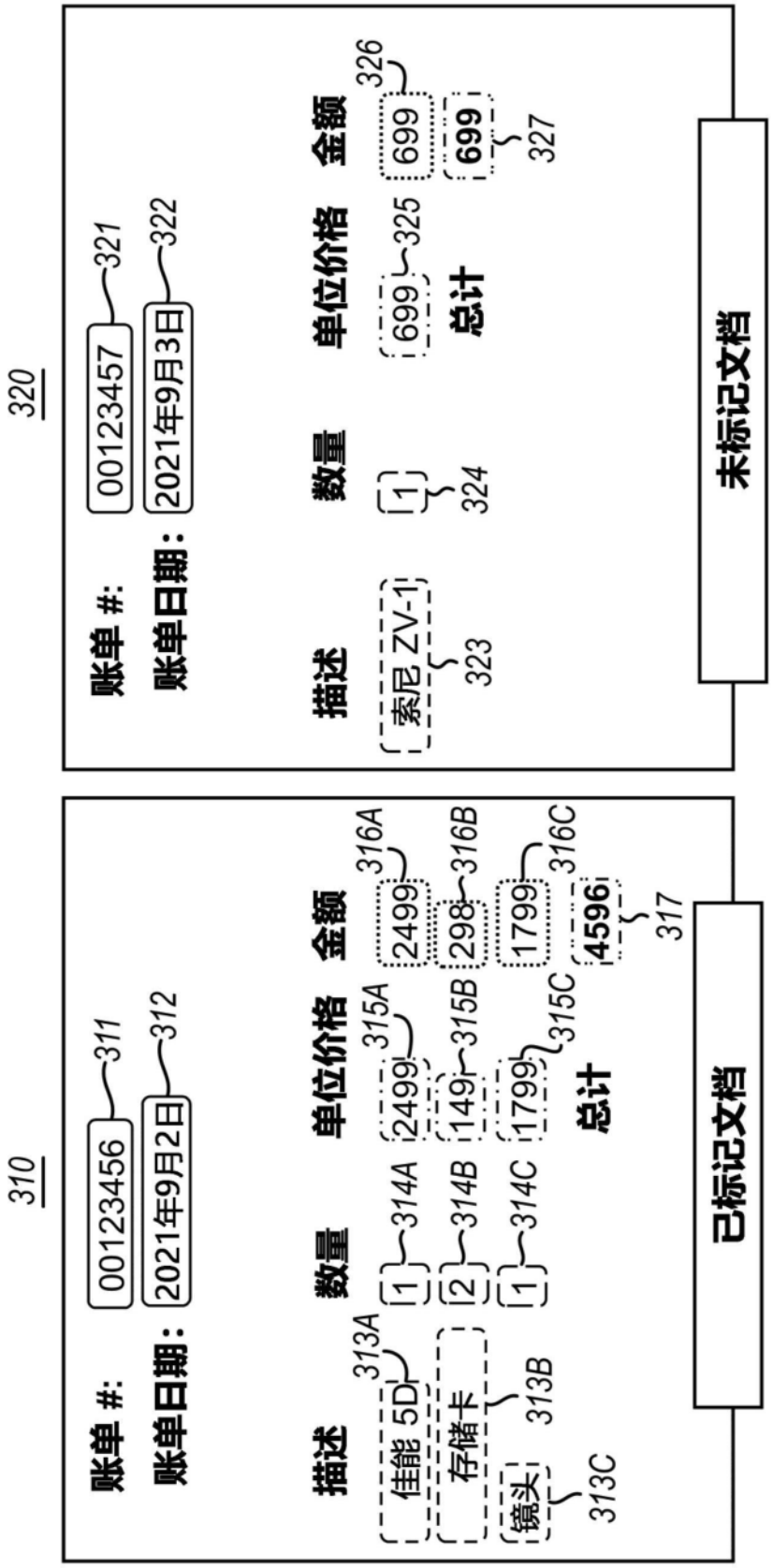

310
账单 #: 00123456 311
账单日期: 2021年9月2日 312
描述 数量 单位价格 金额
佳能 5D 313A 1 314A 2499 315A 2499 316A
存储卡 2 314B 149 315B 298 316B
镜头 313B 1 314C 1799 315C 1799 316C
313C
总计 4596 317
已标记文档
320
账单 #: 00123457 321
账单日期: 2021年9月3日 322
描述 数量 单位价格 金额
索尼 ZV-1 323 1 324 699 325 699 326
总计 699 327
未标记文档

图3B

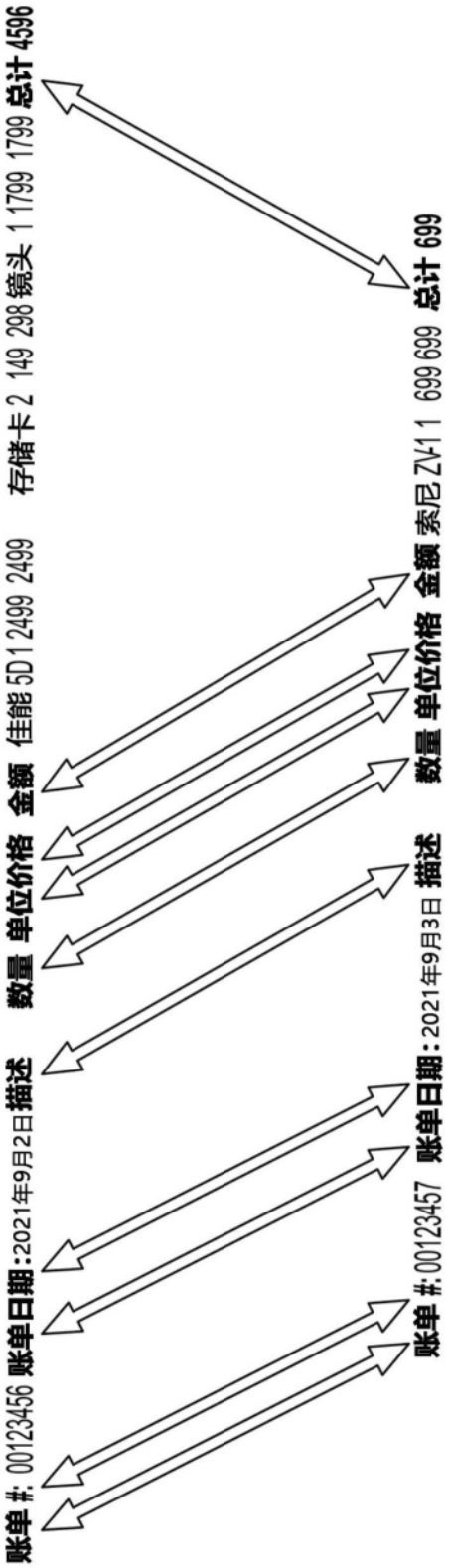

账单 #: 00123456 账单日期: 2021年9月2日 描述 数量 单位价格 金额
佳能 5D 1 2499 2499
存储卡 2 149 298
镜头 1 1799 1799
总计 4596
账单 #: 00123457 账单日期: 2021年9月3日 描述 数量 单位价格 金额
索尼 ZV-1 1 699 699
总计 699

图4

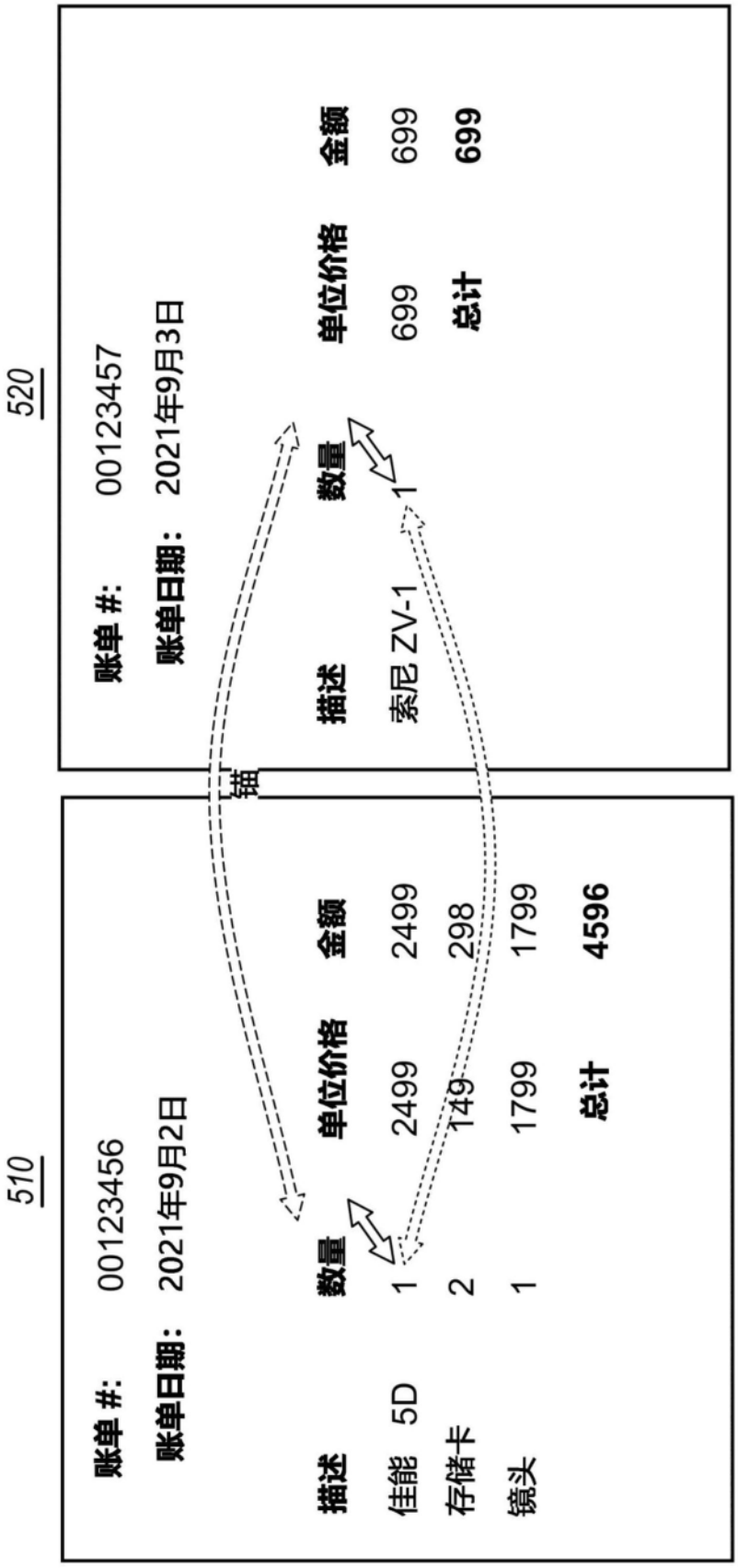

520
账单 #: 00123457
账单日期: 2021年9月3日
描述 数量 单位价格 金额
索尼 ZV-1 1 699 699
总计 699
锚
510
账单 #: 00123456
账单日期: 2021年9月2日
描述 数量 单位价格 金额
佳能 5D 1 2499 2499
存储卡 2 149 298
镜头 1 1799 1799
总计 4596

图5

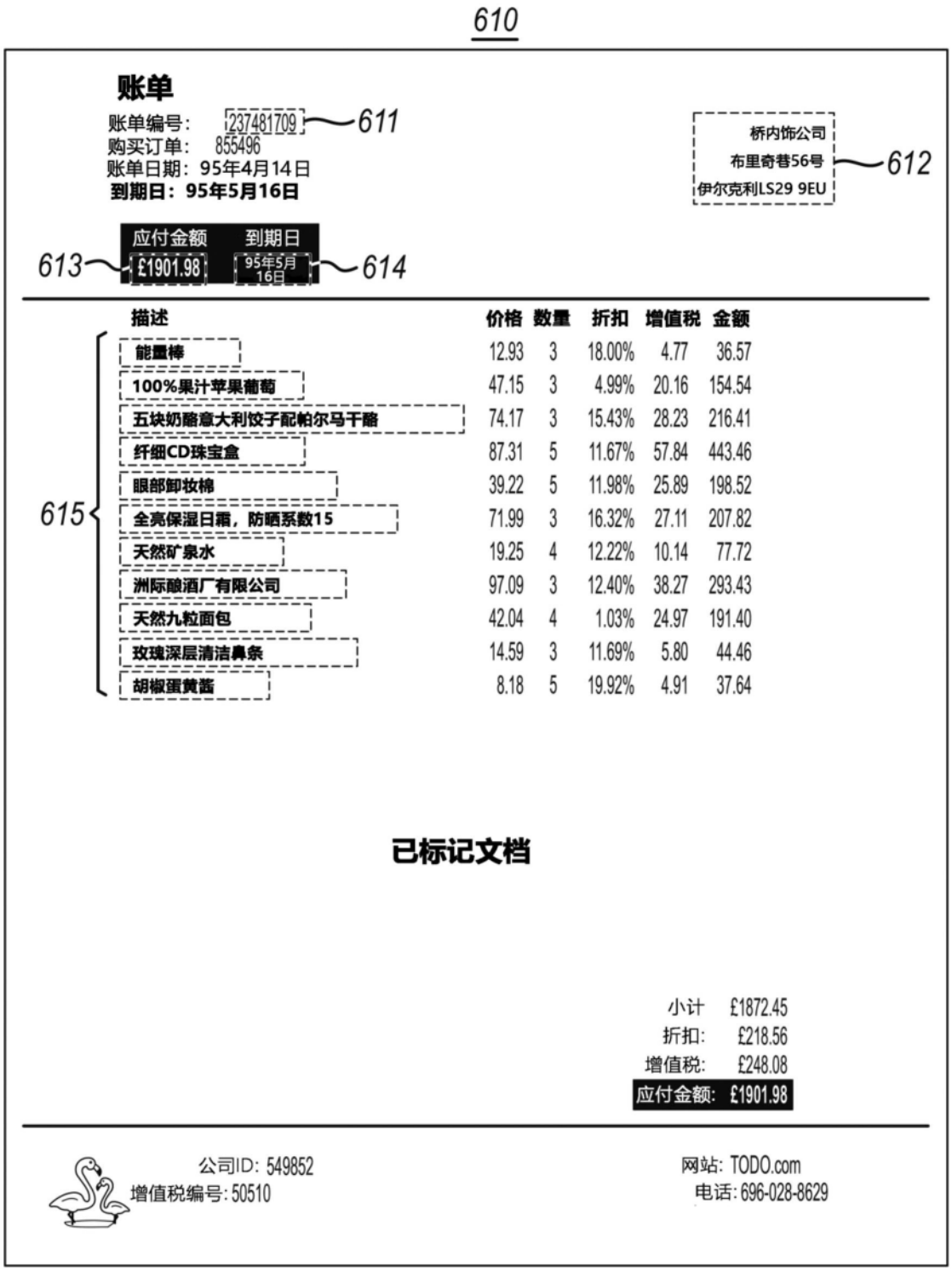

610
账单
账单编号: 237481709
611
购买订单: 855496
账单日期: 95年4月14日
到期日: 95年5月16日
桥内饰公司
布里奇巷56号
伊尔克利LS29 9EU
612
应付金额
到期日
613
£1901.98
95年5月16日
614
描述 价格 数量 折扣 增值税 金额
615
能量棒 12.93 3 18.00% 4.77 36.57
100%果汁苹果葡萄 47.15 3 4.99% 20.16 154.54
五块奶酪意大利饺子配帕尔马干酪 74.17 3 15.43% 28.23 216.41
纤细CD珠宝盒 87.31 5 11.67% 57.84 443.46
眼部卸妆棉 39.22 5 11.98% 25.89 198.52
全亮保湿日霜，防晒系数15 71.99 3 16.32% 27.11 207.82
天然矿泉水 19.25 4 12.22% 10.14 77.72
洲际酿酒厂有限公司 97.09 3 12.40% 38.27 293.43
天然九粒面包 42.04 4 1.03% 24.97 191.40
玫瑰深层清洁鼻条 14.59 3 11.69% 5.80 44.46
胡椒蛋黄酱 8.18 5 19.92% 4.91 37.64
已标记文档
小计 £1872.45
折扣: £218.56
增值税: £248.08
应付金额: £1901.98
公司ID: 549852
增值税编号: 50510
网站: TODO.com
电话: 696-028-8629

图6

620

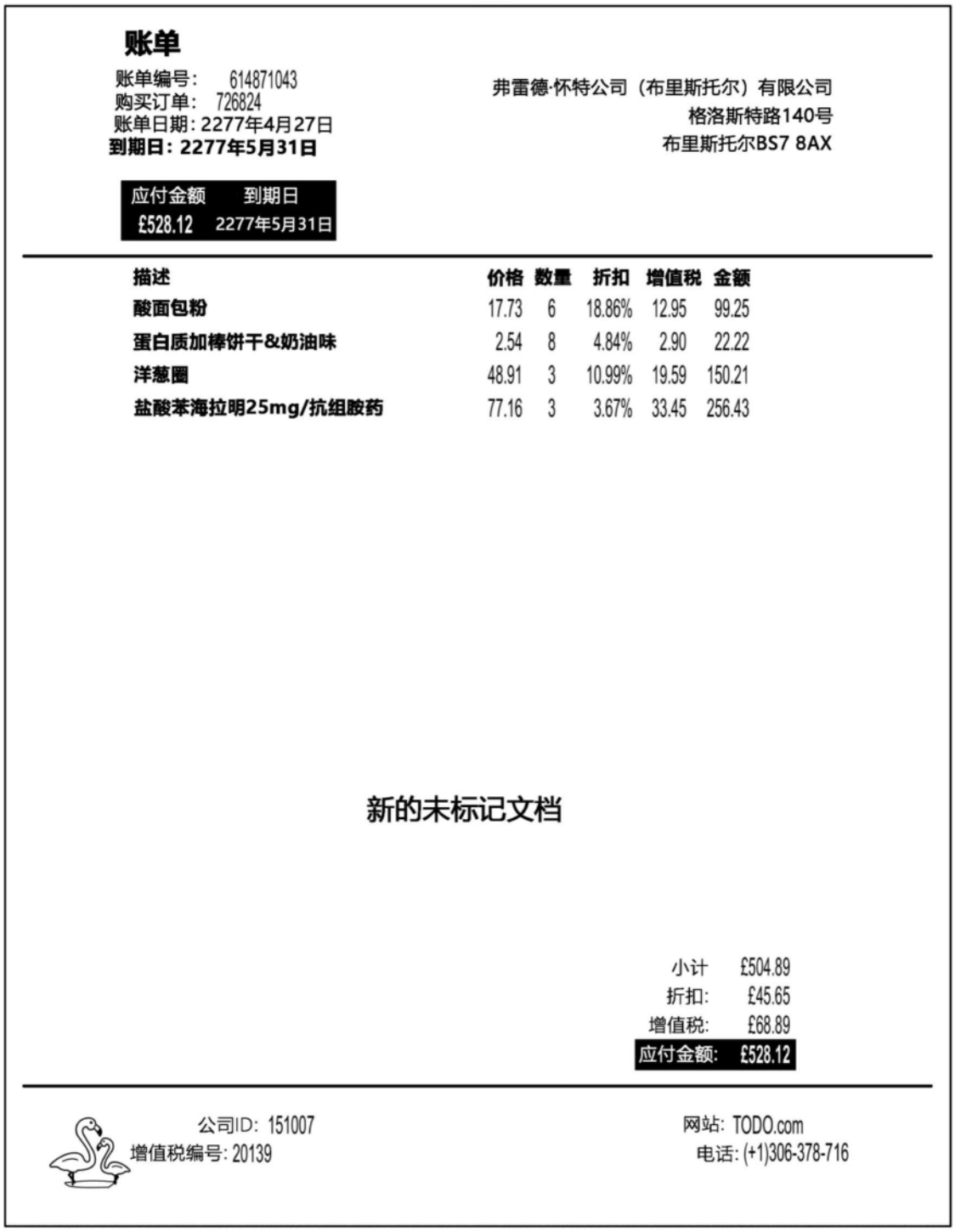

账单
账单编号: 614871043
购买订单: 726824
账单日期: 2277年4月27日
到期日: 2277年5月31日
弗雷德·怀特公司（布里斯托尔）有限公司
格洛斯特路140号
布里斯托尔BS7 8AX
应付金额 到期日
£528.12 2277年5月31日
描述 价格 数量 折扣 增值税 金额
酸面包粉 17.73 6 18.86% 12.95 99.25
蛋白质加棒饼干&奶油味 2.54 8 4.84% 2.90 22.22
洋葱圈 48.91 3 10.99% 19.59 150.21
盐酸苯海拉明25mg/抗组胺药 77.16 3 3.67% 33.45 256.43
新的未标记文档
小计 £504.89
折扣: £45.65
增值税: £68.89
应付金额: £528.12
公司ID: 151007
增值税编号: 20139
网站: TODO.com
电话: (+1)306-378-716

图6(继续)

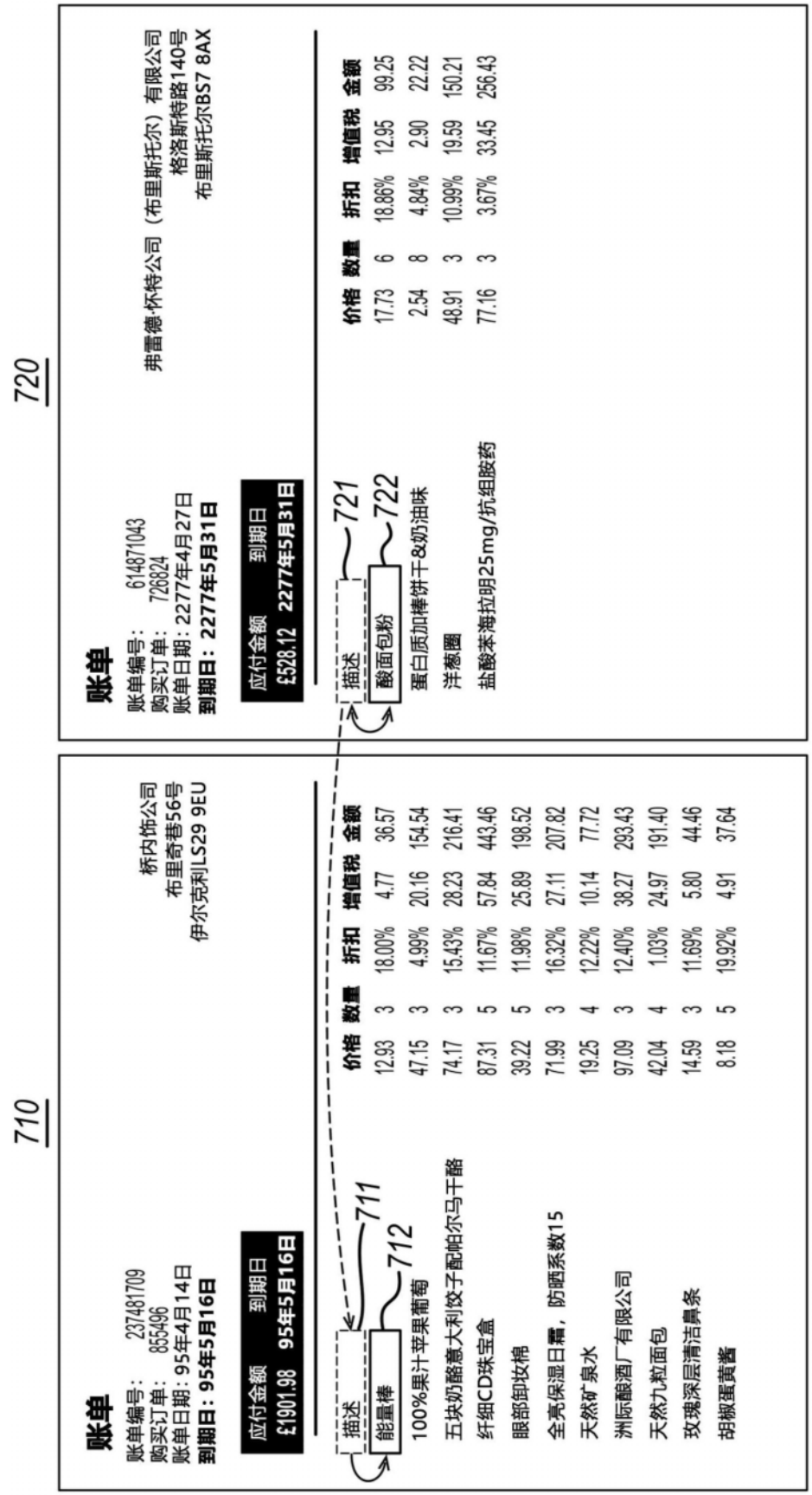

710
账单
账单编号: 237481709
购买订单: 855496
账单日期: 95年4月14日
到期日: 95年5月16日
桥内饰公司
布里奇巷56号
伊尔克利LS29 9EU
应付金额 £1901.98
到期日 95年5月16日
描述
711
能量棒
712
价格 数量 折扣 增值税 金额
12.93 3 18.00% 4.77 36.57
100%果汁苹果葡萄 47.15 3 4.99% 20.16 154.54
五块奶酪意大利饺子配帕尔马干酪 74.17 3 15.43% 28.23 216.41
纤细CD珠宝盒 87.31 5 11.67% 57.84 443.46
眼部卸妆棉 39.22 5 11.98% 25.89 198.52
全亮保湿日霜，防晒系数15 71.99 3 16.32% 27.11 207.82
天然矿泉水 19.25 4 12.22% 10.14 77.72
洲际酿酒厂有限公司 97.09 3 12.40% 38.27 293.43
天然九粒面包 42.04 4 1.03% 24.97 191.40
玫瑰深层清洁鼻条 14.59 3 11.69% 5.80 44.46
胡椒蛋黄酱 8.18 5 19.92% 4.91 37.64
720
账单
账单编号: 614871043
购买订单: 726824
账单日期: 2277年4月27日
到期日: 2277年5月31日
弗雷德·怀特公司（布里斯托尔）有限公司
格洛斯特路140号
布里斯托尔BS7 8AX
应付金额 £528.12
到期日 2277年5月31日
描述
721
酸面包粉
722
价格 数量 折扣 增值税 金额
17.73 6 18.86% 12.95 99.25
蛋白质加棒饼干&奶油味 2.54 8 4.84% 2.90 22.22
洋葱圈 48.91 3 10.99% 19.59 150.21
盐酸苯海拉明25mg/抗组胺药 77.16 3 3.67% 33.45 256.43

图7A

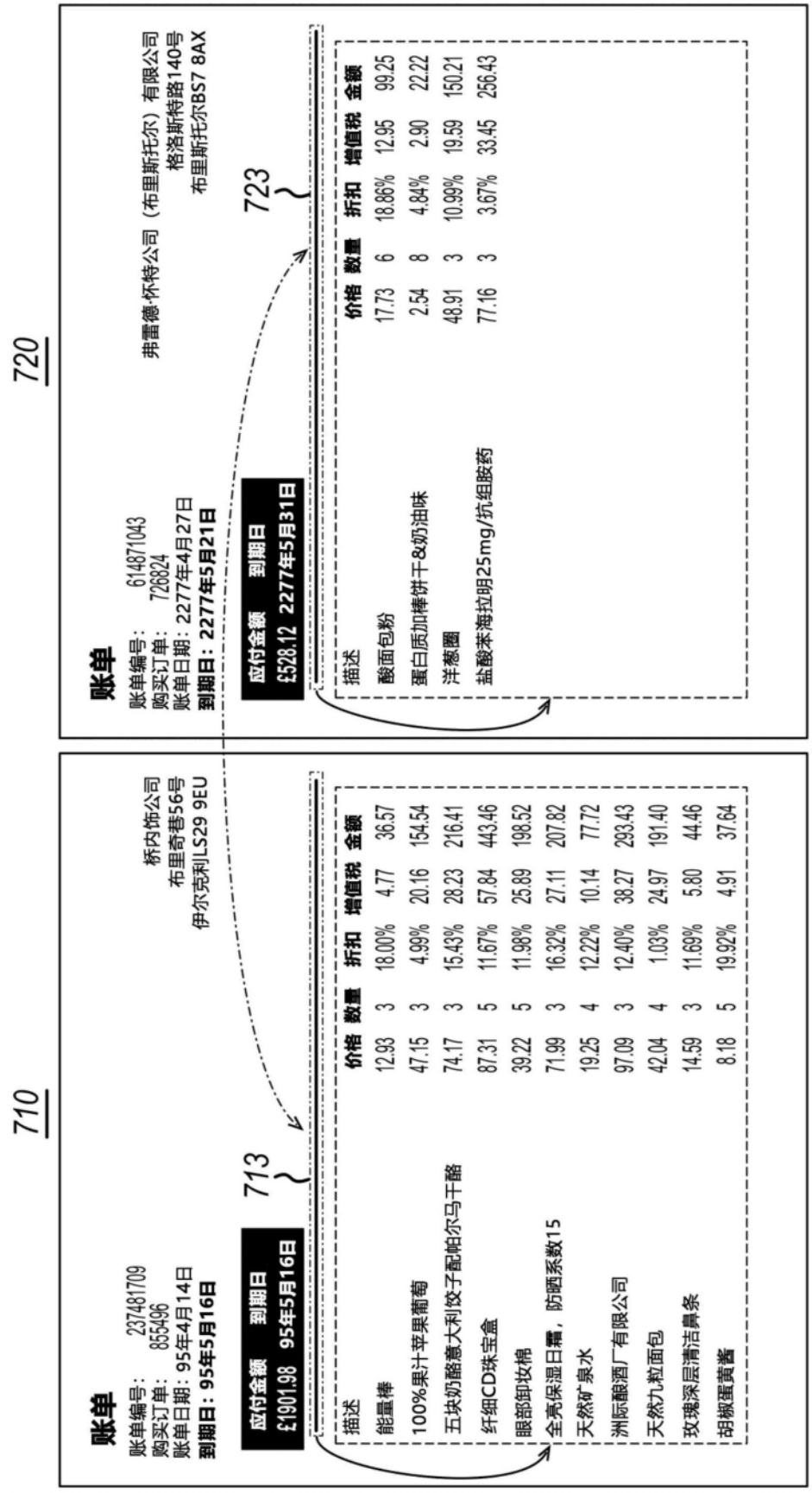

710
账单
账单编号: 237481709
购买订单: 855496
账单日期: 95年4月14日
到期日: 95年5月16日
桥内饰公司
布里奇巷56号
伊尔克利LS29 9EU
应付金额 到期日
£1901.98 95年5月16日
713
描述 价格 数量 折扣 增值税 金额
能量棒 12.93 3 18.00% 4.77 36.57
100%果汁苹果葡萄 47.15 3 4.99% 20.16 154.54
五块奶酪意大利饺子配帕尔马干酪 74.17 3 15.43% 28.23 216.41
纤细CD珠宝盒 87.31 5 11.67% 57.84 443.46
眼部卸妆棉 39.22 5 11.98% 25.89 198.52
全亮保湿日霜，防晒系数15 71.99 3 16.32% 27.11 207.82
天然矿泉水 19.25 4 12.22% 10.14 77.72
洲际酿酒厂有限公司 97.09 3 12.40% 38.27 293.43
天然九粒面包 42.04 4 1.03% 24.97 191.40
玫瑰深层清洁鼻条 14.59 3 11.69% 5.80 44.46
胡椒蛋黄酱 8.18 5 19.92% 4.91 37.64
720
账单
账单编号: 614871043
购买订单: 726824
账单日期: 2277年4月27日
到期日: 2277年5月21日
弗雷德·怀特公司（布里斯托尔）有限公司
格洛斯特路140号
布里斯托尔BS7 8AX
应付金额 到期日
£528.12 2277年5月31日
723
描述 价格 数量 折扣 增值税 金额
酸面包粉 17.73 6 18.86% 12.95 99.25
蛋白质加棒饼干&奶油味 2.54 8 4.84% 2.90 22.22
洋葱圈 48.91 3 10.99% 19.59 150.21
盐酸苯海拉明25mg/抗组胺药 77.16 3 3.67% 33.45 256.43

图7B

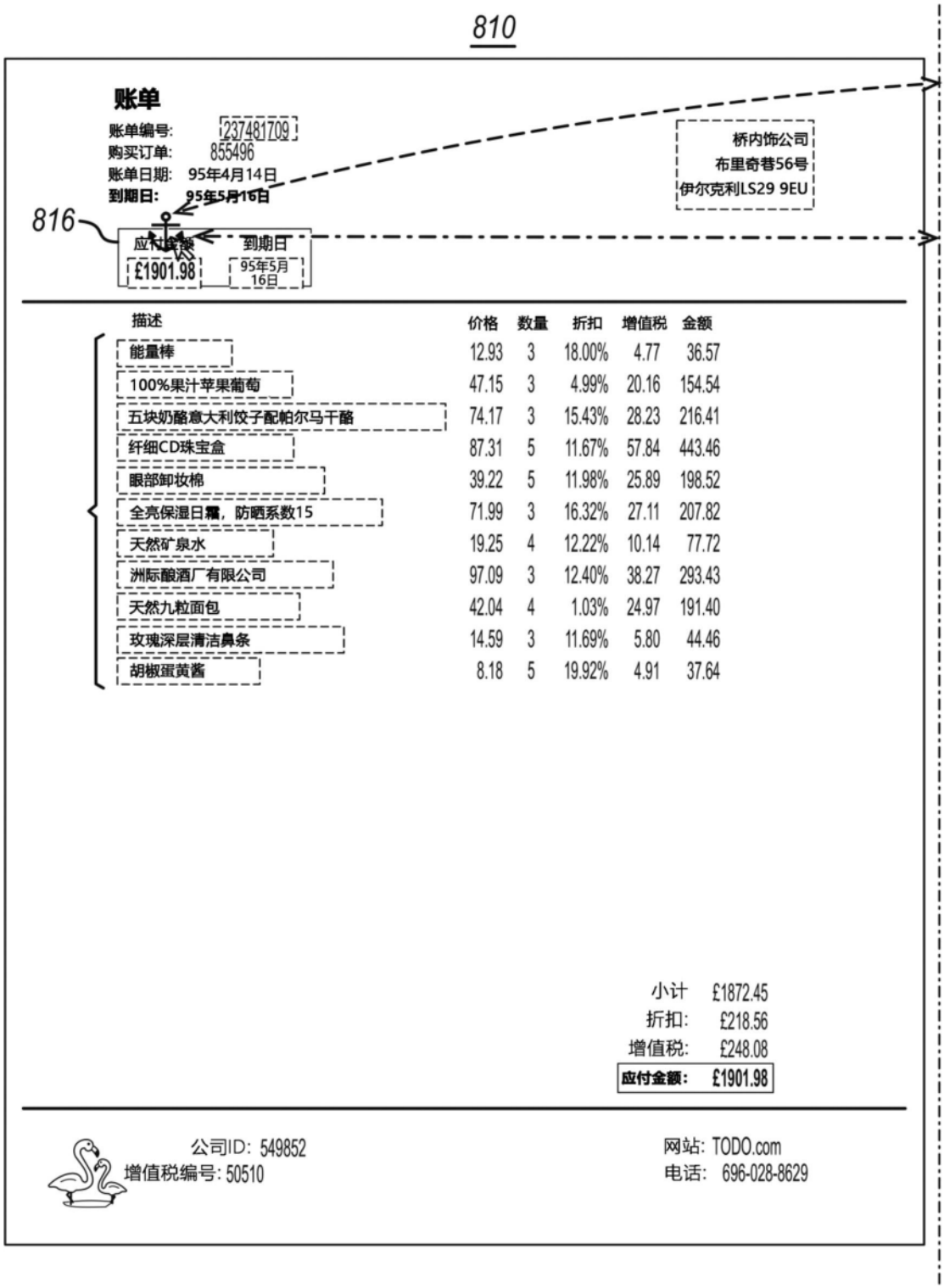

810
账单
账单编号: 237481709
购买订单: 855496
账单日期: 95年4月14日
到期日: 95年5月16日
桥内饰公司
布里奇巷56号
伊尔克利LS29 9EU
816
应付金额
£1901.98
到期日
95年5月16日
描述 价格 数量 折扣 增值税 金额
能量棒 12.93 3 18.00% 4.77 36.57
100%果汁苹果葡萄 47.15 3 4.99% 20.16 154.54
五块奶酪意大利饺子配帕尔马干酪 74.17 3 15.43% 28.23 216.41
纤细CD珠宝盒 87.31 5 11.67% 57.84 443.46
眼部卸妆棉 39.22 5 11.98% 25.89 198.52
全亮保湿日霜，防晒系数15 71.99 3 16.32% 27.11 207.82
天然矿泉水 19.25 4 12.22% 10.14 77.72
洲际酿酒厂有限公司 97.09 3 12.40% 38.27 293.43
天然九粒面包 42.04 4 1.03% 24.97 191.40
玫瑰深层清洁鼻条 14.59 3 11.69% 5.80 44.46
胡椒蛋黄酱 8.18 5 19.92% 4.91 37.64
小计 £1872.45
折扣: £218.56
增值税: £248.08
应付金额: £1901.98
公司ID: 549852
增值税编号: 50510
网站: TODO.com
电话: 696-028-8629

图8

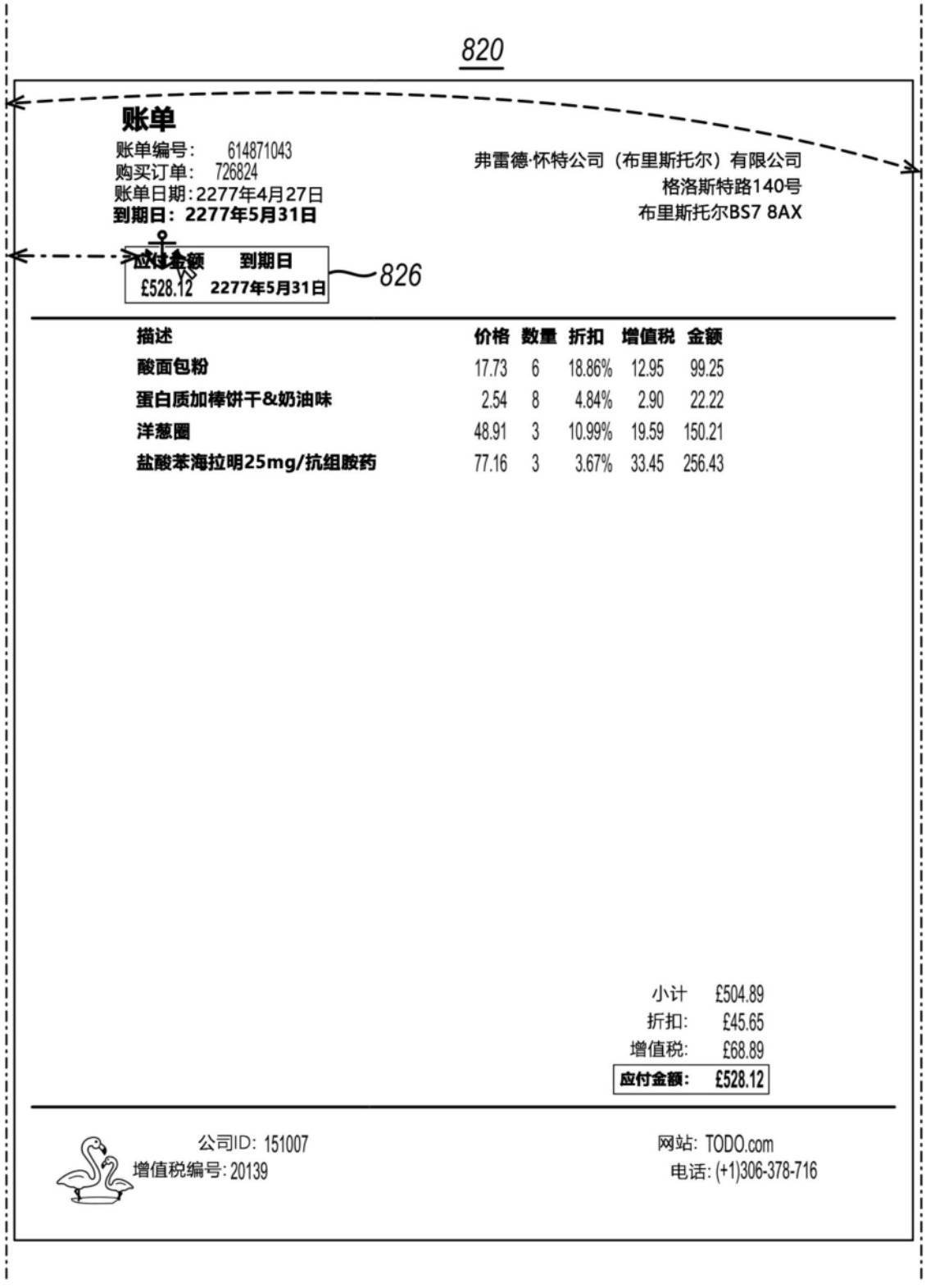

820
账单
账单编号: 614871043
购买订单: 726824
账单日期:2277年4月27日
到期日: 2277年5月31日
弗雷德·怀特公司（布里斯托尔）有限公司
格洛斯特路140号
布里斯托尔BS7 8AX
应付金额 到期日
£528.12 2277年5月31日
826
描述 价格 数量 折扣 增值税 金额
酸面包粉 17.73 6 18.86% 12.95 99.25
蛋白质加棒饼干&奶油味 2.54 8 4.84% 2.90 22.22
洋葱圈 48.91 3 10.99% 19.59 150.21
盐酸苯海拉明25mg/抗组胺药 77.16 3 3.67% 33.45 256.43
小计 £504.89
折扣: £45.65
增值税: £68.89
应付金额: £528.12
公司ID: 151007
增值税编号: 20139
网站: TODO.com
电话: (+1)306-378-716

图8(继续)

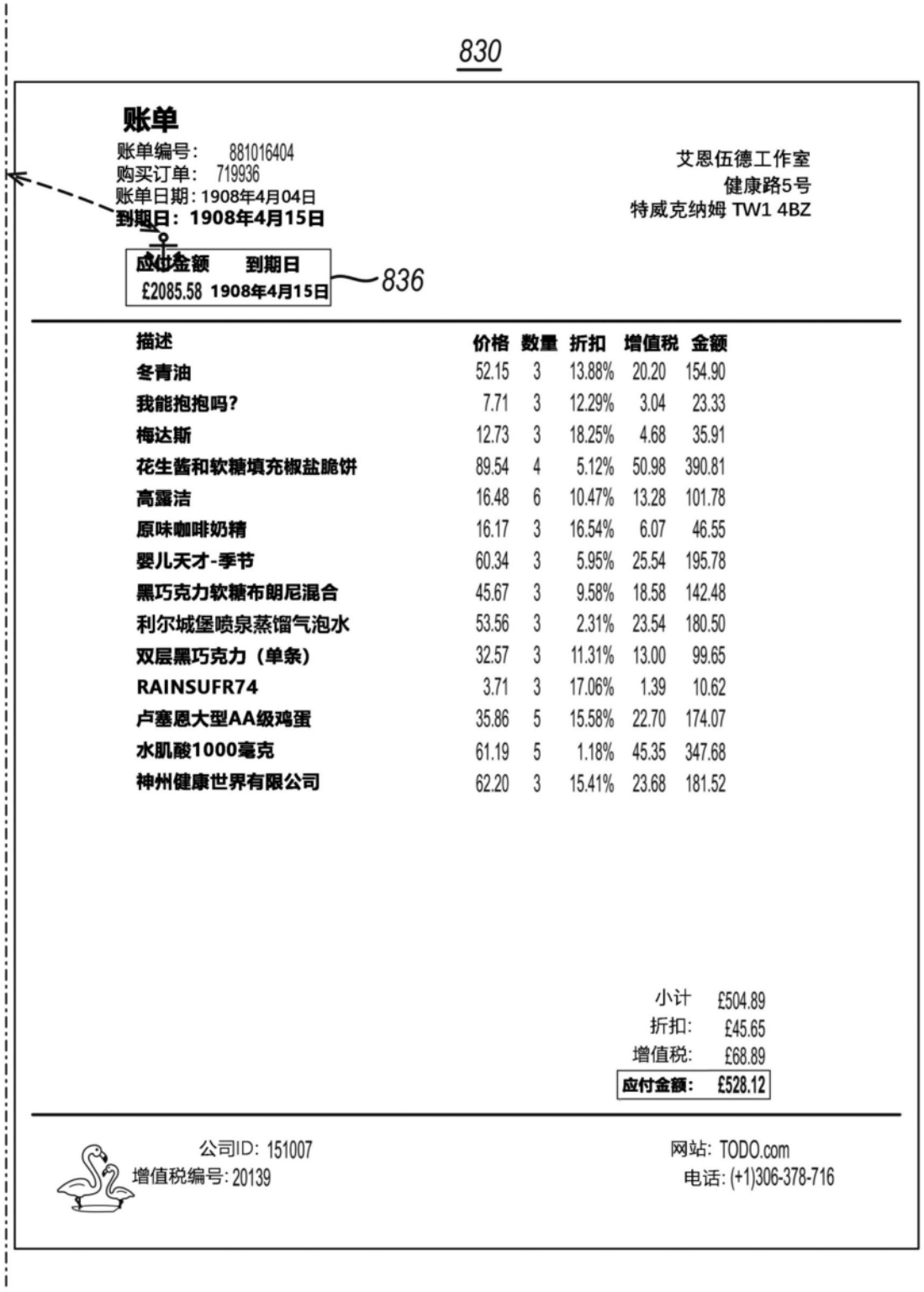

830
账单
账单编号: 881016404
购买订单: 719936
账单日期: 1908年4月04日
到期日: 1908年4月15日
艾恩伍德工作室
健康路5号
特威克纳姆 TW1 4BZ
应付金额 到期日
£2085.58 1908年4月15日
836
描述 价格 数量 折扣 增值税 金额
冬青油 52.15 3 13.88% 20.20 154.90
我能抱抱吗? 7.71 3 12.29% 3.04 23.33
梅达斯 12.73 3 18.25% 4.68 35.91
花生酱和软糖填充椒盐脆饼 89.54 4 5.12% 50.98 390.81
高露洁 16.48 6 10.47% 13.28 101.78
原味咖啡奶精 16.17 3 16.54% 6.07 46.55
婴儿天才-季节 60.34 3 5.95% 25.54 195.78
黑巧克力软糖布朗尼混合 45.67 3 9.58% 18.58 142.48
利尔城堡喷泉蒸馏气泡水 53.56 3 2.31% 23.54 180.50
双层黑巧克力（单条） 32.57 3 11.31% 13.00 99.65
RAINSUFR74 3.71 3 17.06% 1.39 10.62
卢塞恩大型AA级鸡蛋 35.86 5 15.58% 22.70 174.07
水肌酸1000毫克 61.19 5 1.18% 45.35 347.68
神州健康世界有限公司 62.20 3 15.41% 23.68 181.52
小计 £504.89
折扣: £45.65
增值税: £68.89
应付金额: £528.12
公司ID: 151007
增值税编号: 20139
网站: TODO.com
电话: (+1)306-378-716

图8(继续)

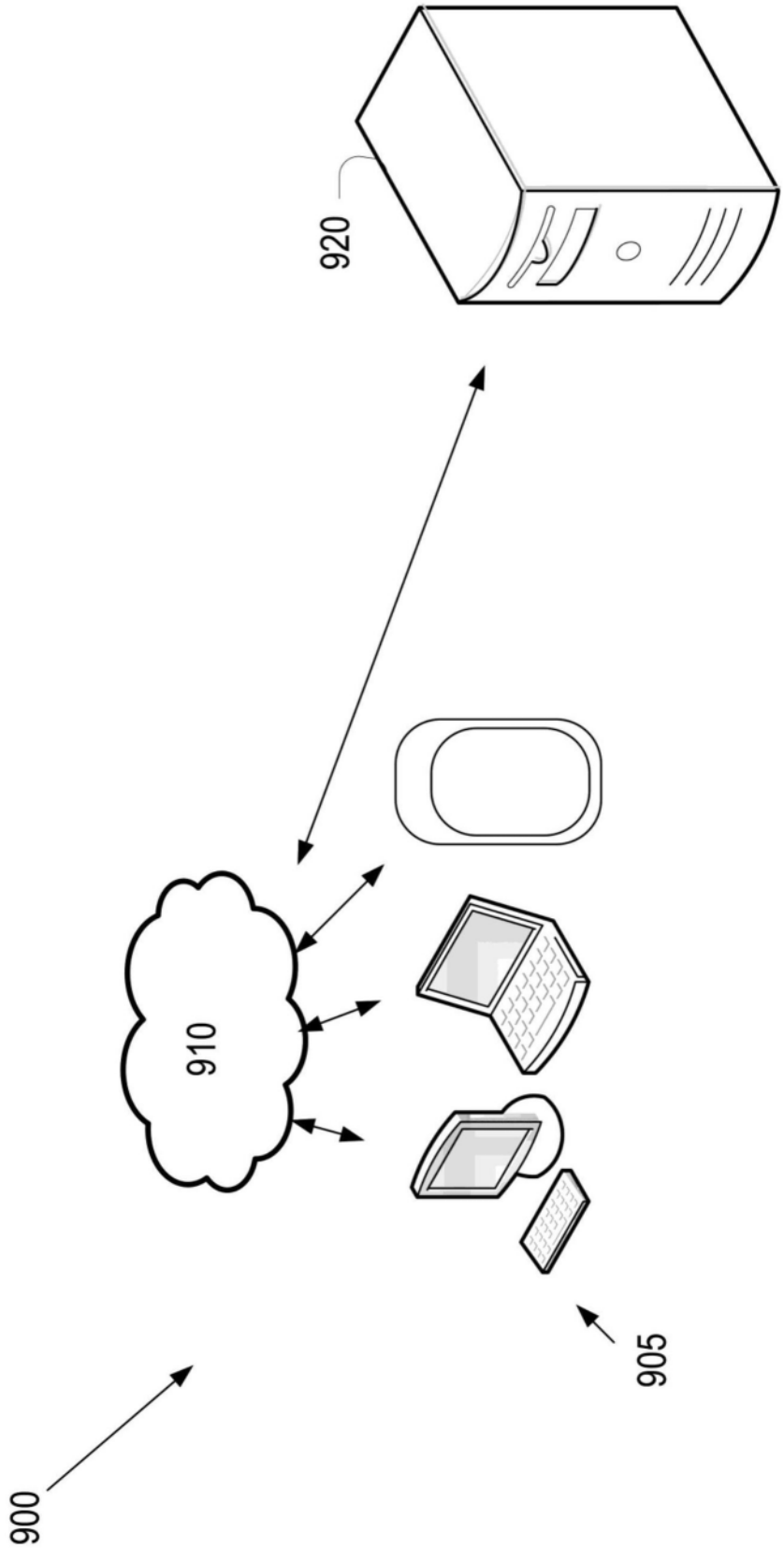


图9