(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108108342 B
(45)授权公告日 2021.09.03

(21)申请号 201711086041.7

(22)申请日 2017.11.07

(65)同一申请的已公布的文献号
申请公布号 CN 108108342 A

(43)申请公布日 2018.06.01

(73)专利权人 汉王科技股份有限公司
地址 100193 北京市海淀区东北旺西路8号汉王大厦

(72)发明人 虞文明 葛洋 陈峻峰

(74)专利代理机构 北京博雅睿泉专利代理事务所(特殊普通合伙) 11442
代理人 余西西 马佑平

(51)Int.Cl.
G06F 40/18(2020.01)
G06F 40/279(2020.01)
G06F 40/205(2020.01)

(56)对比文件
CN 104750801 A,2015.07.01
CN 105260727 A,2016.01.20
CN 105630817 A,2016.06.01
CN 102043861 A,2011.05.04
CN 106250830 A,2016.12.21
US 2016055376 A1,2016.02.25
万里鹏.非结构化到结构化数据转换的研究与实现.《中国优秀硕士学位论文全文数据库 信息科技辑》.2013,(第11期),

审查员 吴姝泓

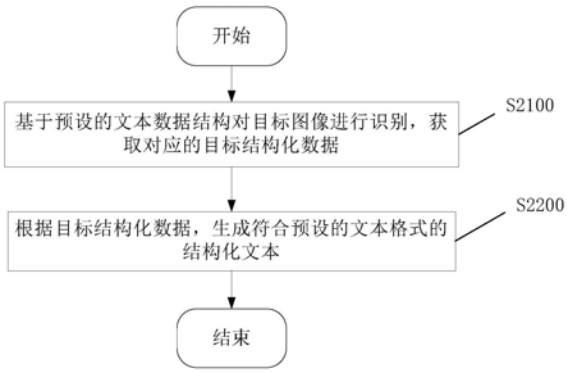
权利要求书2页 说明书14页 附图7页

(54)发明名称

结构化文本的生成方法、检索方法及装置

(57)摘要

本发明公开了一种结构化文本的生成方法、检索方法及装置。该生成方法包括:基于预设的文本数据结构对目标图像进行识别,获取对应的目标结构化数据;根据目标结构化数据,生成符合预设的文本格式的结构化文本。根据本发明,可以在数据交换过程中,对不同的文本格式或者数据平台,都能根据对应的结构化数据,有效还原文本版式,尤其适用于数据交换频繁、涉及大量数据分析的大数据应用场景。



CN 108108342 B

1.一种结构化文本的生成方法,其特征在于,所述方法包括:

基于预设的文本数据结构,对目标图像进行识别,获得对应的目标结构化数据,其中,所述预设的文本数据结构用于支持以结构化信息描述文本版式,所述文本版式是指文本内容的排版样式;

根据所述目标结构化数据,生成符合预设的文本格式的结构化文本。

2.根据权利要求1所述的方法,其特征在于,所述方法还包括生成文本数据结构作为预设的文本数据结构,用以根据预设的文本数据结构描述对目标图像的识别结果。

3.根据权利要求1或2所述的方法,其特征在于,

所述预设的文本数据结构至少包括识别出的文本单元的数目、每个文本单元对应的结构化信息,其中,所述结构化信息包含相应的文本单元的类型、单元位置信息、文字内容以及每个文字的文字位置信息。

4.根据权利要求3所述的方法,其特征在于,

所述文本单元的类型至少包括文本块;

当所述文本单元是文本块时,所述文本单元的数目包括所述文本块的数目,每个所述文本块的单元位置信息至少包括对应的所述文本块的块序号以及文本行信息。

5.根据权利要求3所述的方法,其特征在于,

所述文本单元类型至少包括表格;

当所述文本单元是表格时,所述文本单元的数目包括所述表格的数目,每个所述表格的单元位置信息至少包括对应的所述表格的块序号、表格行数目、表格列数目以及单元格信息;

其中,所述单元格信息包括对应的单元格所在的表格位置信息、单元格跨行数目、单元格跨行列数目、单元格包含的文本行信息。

6.根据权利要求3所述的方法,其特征在于,

所述文字位置信息包括与文字对应的符号图像相对于所述目标图像的左上角顶点作为原点得到的坐标位置信息。

7.一种文本检索方法,其特征在于,

接收文本检索请求,并基于所述文本检索请求获取用于文本检索的检索关键信息,其中,所述检索关键信息至少包括描述文本版式的结构化信息,所述文本版式是指文本内容的排版样式;

采用获得的所述检索关键信息,在若干结构化文本中进行检索,获得与所述文本检索请求对应的文本内容,生成对应的检索结果。

8.根据权利要求7所述的方法,其特征在于,

所述检索关键信息包括文本单元的结构化信息的至少部分,所述结构化信息包括文本单元的类型、单元位置信息、文字内容以及每个文字的文字位置信息;

和/或,

所述结构化文本包括对应的目标结构化数据的符合预设的文本格式的文本,所述目标结构化数据符合预设的文本数据结构,所述文本数据结构至少包括对应文本的文本单元的数目、每个所述文本单元的所述结构化信息。

9.一种结构化文本的生成装置,其特征在于,包括:

数据识别单元，基于预设的文本数据结构对目标图像进行识别，获取对应的目标结构化数据；其中，所述预设的文本数据结构用于支持以结构化信息描述文本版式，所述文本版式是指文本内容的排版样式；

文本生成单元，用于根据所述目标结构化数据，生成符合预设的文本格式的结构化文本。

10.一种结构化文本的检索装置，其特征在于，包括：

信息获取单元，用于接收文本检索请求，并基于所述文本检索请求获取用于文本检索的检索关键信息，其中，所述检索关键信息至少包括描述文本版式的结构化信息，所述文本版式是指文本内容的排版样式；

检索执行单元，用于采用获得的所述检索关键信息，在若干结构化文本中进行检索，获得与所述文本检索请求对应的文本内容，生成对应的检索结果。

结构化文本的生成方法、检索方法及装置

技术领域

[0001] 本发明涉及图像处理技术领域，更具体地，涉及一种结构化文本的生成方法、检索方法及装置。

背景技术

[0002] 光学字符识别(Optical Character Recognition,OCR)技术是指对图像中包含的字符图像进行处理，将字符图像转换为字符计算机内码，从而得到可编辑的文字编码字符流的信息处理技术。

[0003] 目前OCR技术广泛应用于数字图书、文档管理等领域，然而，现有技术中，采用OCR技术识别字符图像转换得到的识别文本，普遍是市面上通用的文本编辑软件支持编辑的非结构化文本，例如Microsoft Office、WPS等等文本编辑软件可支持编辑的word文本、Windows操作系统自带的txt文本等。

[0004] 发明人发现，非结构化文本包含的非结构化数据没有固定的数据结构，在基于非结构化文本进行数据交换时，会因为交换过程中涉及的非结构化文本之间的文本格式差异、涉及的数据平台支持的数据结构差异等原因，导致数据交换后的数据无法有效还原原始文本版式，而且，在进行文本检索时，若检索范围涵盖了非结构化文本，则只能基于检索词对文本进行全篇检索，从而导致文本检索效率低下。

[0005] 尤其针对目前互联网中数据交换的数据量规模呈指数级爆发，以及数据交换频繁、涉及大量数据分析的大数据应用场景，非结构化数据带来的数据交换时，无法有效还原原始文本版式，以及数据检索效率低的问题尤为突出。

[0006] 有鉴于此，有必要针对上述现有技术中存在的问题至少之一进行改进。

发明内容

[0007] 本发明的一个目的是提供用于生成结构化文本的方法。

[0008] 根据本发明的第一方面，提供了一种结构化文本的生成方法，包括：

[0009] 基于预设的文本数据结构对目标图像进行识别，获取对应的目标结构化数据，

[0010] 其中，所述预设的文本数据结构用于支持以结构化信息描述文本版式；

[0011] 根据所述目标结构化数据，生成符合预设的文本格式的结构化文本。

[0012] 可选地，所述方法还包括生成文本数据结构作为预设的文本数据结构，用以根据预设的文本数据结构描述对目标图像的识别结果。

[0013] 可选地，

[0014] 所述预设的文本数据结构至少包括识别出的文本单元的数目、每个文本单元对应的结构化信息，其中，所述结构化信息包含相应的文本单元的类型、单元位置信息、文字内容以及每个文字的文字位置信息。

[0015] 可选地，

[0016] 所述文本单元的类型至少包括文本块；

[0017] 当所述文本单元是文本块时，所述文本单元的数目包括所述文本块的数目，每个所述文本块的单元位置信息至少包括对应的所述文本块的块序号以及文本行信息。

[0018] 可选地，

[0019] 所述文本单元类型至少包括表格；

[0020] 当所述文本单元是表格时，所述文本单元的数目包括所述表格的数目，每个所述表格的单元位置信息至少包括对应的所述表格的块序号、表格行数目、表格列数目以及单元格信息；其中，所述单元格信息包括对应的单元格所在的表格位置信息、单元格跨行数目、单元格跨行列数目、单元格包含的文本行信息。

[0021] 可选地，所述文字位置信息包括与文字对应的符号图像相对于所述目标图像的左上角顶点作为原点得到的坐标位置信息。

[0022] 根据本发明的第二方面，提供一种结构化文本的生成装置，包括：

[0023] 数据识别单元，基于预设的文本数据结构对目标图像进行识别，获取对应的目标结构化数据；

[0024] 其中，所述预设的文本数据结构用于支持以结构化信息描述文本版式；

[0025] 文本生成单元，用于根据所述目标结构化数据，生成符合预设的文本格式的结构化文本。

[0026] 本发明的另一个目的是提供一种用于检索结构化文本的新技术方案。

[0027] 根据本发明的第三方面，提供一种结构化的文本检索方法，包括：

[0028] 接收文本检索请求，并基于所述文本检索请求获取用于文本检索的检索关键信息，

[0029] 其中，所述检索关键信息至少包括描述文本版式的结构化信息；

[0030] 采用获得的所述检索关键信息，在若干结构化文本中进行检索，获得与所述文本检索请求对应的文本内容，生成对应的检索结果。

[0031] 可选地，

[0032] 所述检索关键信息包括文本单元的结构化信息的至少部分，所述结构化信息包括文本单元的类型、单元位置信息、文字内容以及每个文字的文字位置信息；

[0033] 和/或

[0034] 所述结构化文本包括对应的目标结构化数据的符合预设的文本格式的文本，所述目标结构化数据符合预设的文本数据结构，所述文本数据结构至少包括对应文本的文本单元的数目、每个所述文本单元的所述结构化信息。

[0035] 根据本发明的第四方面，提供一种结构化的文本检索装置，包括：

[0036] 信息获取单元，用于接收文本检索请求，并基于所述文本检索请求获取用于文本检索的检索关键信息，

[0037] 其中，所述检索关键信息至少包括描述文本版式的结构化信息；检索执行单元，用于采用获得的所述检索关键信息，在若干结构化文本中进行检索，获得与所述文本检索请求对应的文本内容，生成对应的检索结果。

[0038] 本发明实施例中，通过结合预设的文本数据结构，对目标图像进行识别，可以得到符合预设的文本数据结构的目标结构化数据，进而，可以生成对应的符合预设的文本格式的结构化文本，使得在数据交换过程中，针对不同的文本格式或者数据平台，都能根据对应

的结构化数据，有效还原文本版式，克服非结构数据存在的在数据交换过程中，一旦因为变换文本格式或者跨越数据平台就难以还原文本版式的缺陷，而且，对于数据交换频繁、涉及大量数据分析的大数据应用场景，本发明实施例提供的方案尤其适用。

[0039] 通过以下参照附图对本发明的示例性实施例的详细描述，本发明的其它特征及其优点将会变得清楚。

附图说明

[0040] 被结合在说明书中并构成说明书的一部分的附图示出了本发明的实施例，并且连同其说明一起用于解释本发明的原理。

[0041] 图1示出了可用于实现本发明实施例的计算系统的硬件配置的框图。

[0042] 图2示出了本发明第一实施例的结构化文本的生成方法的流程图。

[0043] 图3示出了本发明第一实施例的结构化文本的生成装置的示意框图。

[0044] 图4示出了本发明第一实施例的结构化文本的生成方法的目标图像的示意图。

[0045] 图5示出了本发明第一实施例的结构化文本的生成方法的结果表达形式的示意图。

[0046] 图6示出了本发明第二实施例的结构化文本的检索方法的流程图。

[0047] 图7示出了本发明第二实施例的结构化文本的检索装置的示意框图。

[0048] 图8示出了本发明实施例的文本块的结构化数据的示意图。

[0049] 图9示出了本发明实施例的表格的结构化数据的示意图。

具体实施方式

[0050] 现在将参照附图来详细描述本发明的各种示例性实施例。应注意到：除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本发明的范围。

[0051] 以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本发明及其应用或使用的任何限制。

[0052] 对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为说明书的一部分。

[0053] 在这里示出和讨论的所有例子中，任何具体值应被解释为仅仅是示例性的，而不是作为限制。因此，示例性实施例的其它例子可以具有不同的值。

[0054] 应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

[0055] 〈硬件配置〉

[0056] 图1示出了可以实现本发明实施例的计算机系统1000的硬件配置框图。

[0057] 如图1所示，计算机系统1000包括计算机1110。在一个例子中，计算机1110可以是手机、平板电脑、掌上电脑、台式计算机或者笔记本电脑，或者，计算机1110可以是服务器，例如刀片服务器、云平台服务器等。计算机1110包括经由系统总线1121连接的处理器1120、存储器1130、固定非易失性存储器接口1140、移动非易失性存储器接口1150、用户输入接口1160、网络接口1170、视频接口1190和输出外围接口1195。

[0058] 存储器1130包括ROM(只读存储器)和RAM(随机存取存储器)。BIOS(基本输入输出系统)驻留在ROM内。操作系统、应用程序、其它程序模块和某些程序数据驻留在RAM内。

[0059] 诸如硬盘的固定非易失性存储器连接到固定非易失性存储器接口1140。固定非易失性存储器例如可以存储操作系统、应用程序、其它程序模块和某些程序数据。

[0060] 诸如软盘驱动器和CD-ROM驱动器的移动非易失性存储器连接到移动非易失性存储器接口1150。例如,软盘可以被插入到软盘驱动器中,以及CD(光盘)可以被插入到CD-ROM驱动器内。

[0061] 诸如鼠标和键盘的输入设备被连接到用户输入接口1160。

[0062] 计算机1110可以通过网络接口1170连接到远程计算机1180。例如,网络接口1170可以通过局域网连接到远程计算机。或者,网络接口1170可以连接到调制解调器(调制器—解调器),以及调制解调器经由广域网连接到远程计算机1180。

[0063] 远程计算机1180可以包括诸如硬盘的存储器,其可以存储远程应用程序。

[0064] 视频接口1190连接到监视器。

[0065] 输出外围接口1195连接到打印机和扬声器。

[0066] 图1所示的计算机系统仅仅是说明性的并且决不意味着对本发明、其应用或使用的任何限制。应用于本发明的实施例中,计算机1110的所述存储器1130用于存储指令,所述指令用于控制所述处理器1130进行操作以执行本发明实施例提供的任意一项结构化文本的生成方法或结构化文本的检索方法。

[0067] 本领域技术人员应当理解,尽管在图1中对计算机1110示出了多个装置,但是,本发明可以仅涉及其中的部分装置,例如,计算机1110只涉及处理器1120和存储装置1130。技术人员可以根据本发明所公开方案设计指令。指令如何控制处理器进行操作,这是本领域公知,故在此不再详细描述。

[0068] <第一实施例>

[0069] 本发明实施例中,提供一种结构化文本的生成方法,如图2所示,包括:

[0070] 步骤S2100,采用光学字符识别技术对目标图像进行识别,基于预设的文本数据结构获取对应的目标结构化数据。

[0071] 具体的,本发明实施例中,目标图像可以是任何包含至少一个字符的数字图像,例如,数码相机拍摄或扫描仪扫描得到的纸质文档图像,或者是含文字的广告图像,又或者是含文字的表格图像,等等。

[0072] 进一步,本发明实施例中,可以采用光学字符识别技术(Optical Character Recognition,OCR)对目标图像进行识别,所谓OCR技术,是指通过检测目标图像中字符图像的暗、亮的模式确定其形状,再基于字符识别方法将确定的形状翻译成计算机文字。

[0073] 现有技术下,通过OCR技术对目标图像进行识别,所获得的计算机文字的文本格式为通用的文本编辑软件所支持的非结构化文本格式,例如,Microsoft Office支持的word文本格式。由于,基于这类非结构化文本格式的非结构化文本,包含的都是非结构化数据,没有固定的数据结构。因此,在基于非结构化文本进行数据交换时,会因为交换过程中涉及的非结构化文本之间的文本格式差异、涉及的数据平台支持的数据结构差异等原因,导致数据交换后的数据无法有效还原原始文本版式。

[0074] 本发明实施例中,虽然依然采用OCR技术对目标图像进行识别,但由于在识别时是

基于预设的文本数据结构，该预设的文本数据结构用于支持以结构化信息描述文本版式，而文本版式是指文本内容的排版样式，因此，识别得到的目标结构化数据，具有固定的数据结构(该数据结构符合预设的文本数据结构)，不仅可以像现有技术中识别得到的非结构化数据一样，描述图像中显示的字符对应的文本内容，还可以描述图像中显示字符对应的文本版式。

[0075] 例如，通过OCR技术对目标图像进行识别时，当识别出目标图像中存在文字时，基于预设的文本数据结构，针对识别出的文字得到对应的文字数目、文字内容、文字位置等数据，或者当识别出目标图像中存在表格时(例如识别出表格的线条)，基于预设的文本数据结构，针对识别出的表格，得到对应的表格数目、表格中的文字内容、表格位置、表格中文字位置等数据。这些数据就是本发明实施例中对目标图像识别后得到的目标结构化数据，具有符合预设的文本数据结构的固定的数据结构，并且，数据内容中包括目标图像中的文字或表格相关的文字内容，以及包括文字或表格的数目、位置这些与文本版式相关的内容，可以用于描述目标图像的文本内容和文本版式。

[0076] 通过上述具有固定的数据结构的目标结构化数据，可以在后续步骤中生成对应的结构化文本。基于该结构化文本进行数据交换，即使数据交换时涉及不同的文本格式或者跨越不同的数据平台，但都可以从结构化文本中包含的结构化数据中，获取目标图像的文字内容和文本版式后，再保持原有的文本版式处理为对应的文本格式或者适配对应的数据平台，实现有效还原文本版式。因此，可以克服非结构化数据在数据交换过程中存在的，因变换文本格式或者跨越数据平台文本版式，而难以还原原始文本格式的缺陷，并尤其适用于数据交换频繁、涉及大量数据分析的大数据应用场景。

[0077] 本发明实施例中，提供的结构化文本的生成方法还包括生成文本数据结构作为预设的文本数据结构，用以根据预设的文本数据结构描述对目标图像的识别结果。具体地，可以针对根据具体的应用场景或者具体的应用需求，对应生成对应的文本数据结构。例如，在对具有相似的文本版式的大批量文本进行识别的应用场景，可以针对相似的文本版式，生成匹配对应的文本数据结构。

[0078] 具体地，预设的文本数据结构至少包括识别出的文本单元的数目、每个文本单元对应的结构化信息，其中，一个文本单元对应的结构化信息，至少包含该文本单元的类型、该文本单元的位置信息、该文本单元的文字内容，以及文字内容中各个文字的文字位置信息。

[0079] 文本单元是图像显示的文本中的一个图像区域，由多个字符图像构成。具体地，文本单元可以是文本段、表格、文本块等，其中，文本单元的大小和类型的划分可以根据具体的应用场景或者实际需求进行调整，在此不做限制。

[0080] 例如，文本单元的类型可以至少包括文本块；

[0081] 当文本单元是文本块时，所述文本单元的数目包括所述文本块的数目，每个所述文本块的单元位置信息至少包括对应的所述文本块的块序号以及文本行信息。

[0082] 文本块的块序号是对应的文本块在目标图像显示的文本中包括的所有文本单元(可以仅包括文本块，还可以包括其他类型的文本单元)中顺序编号得到的序号。

[0083] 文本块的文本行信息用于描述对应的文本块包括的文本行，可以包括文本块包括的文本行数目和对应的文本行序号。

[0084] 具体地，每一个文本块可以是图像显示的文本中的一个完整的文本段。对应的，文本块的数目就是图像显示的文本中包括的文本段落数目。每个文本块的单元位置信息包括对应文本段落对应的块序号、以及该文本段落所包括的文本行数目、对应的文本行序号。

[0085] 或者，又例如，文本单元类型可以至少包括表格；

[0086] 当所述文本单元是表格时，所述文本单元的数目包括所述表格的数目，每个所述表格的单元位置信息至少包括对应的所述表格的块序号、表格行数目、表格列数目以及单元格信息。

[0087] 表格的块序号是对应的表格在目标图像显示的文本中包括的所有文本单元(可以仅包括表格，还可以包括其他类型的文本单元)中顺序编号得到的序号。

[0088] 单元格信息用于描述对应的表格中每个表格的信息。在实际应用中，表格里的每个单元格的格式并不一定是完全相同的，可能存在跨表格行或表格列的单元格，每个单元格的位置、每个单元格包括的文本行数目、文本行的序号也各有不同。具体地，所述单元格信息包括对应的单元格所在的表格位置信息、单元格跨行数目、单元格跨行列数目、单元格包含的文本行信息。

[0089] 具体地，表格位置信息用于描述对应的单元格在表格中的位置，可以包括该单元格所在表格的表格行号、表格列号。单元格的文本行信息用于描述对应的表格包括的文本行，可以包括表格包括的文本行数目和对应的文本行序号。文本单元的结构化信息包括对应的所述文本单元中文字内容中每个文字的文字位置信息。该文字的文字位置信息用于描述对应的文字在目标图像显示的文本中的位置。

[0090] 具体地，所述文字位置信息包括与文字对应的符号图像相对于所述目标图像的左上角顶点作为原点得到的坐标位置信息。该坐标位置信息可以根据具体的应用需求设置，只要能确定该文字在目标图像显示的文本中的具体位置即可。例如，每个文字的坐标位置信息可以是该文字对应的符号图像相对于所述目标图像的左上角顶点作为原点，得到的字符图像(字符图像为矩形图像)的左上角点的X坐标值、Y坐标值，以及右下角点的X坐标值、Y坐标值。

[0091] 在一个例子中，每个所述文本单元的结构化信息还包括对应的所述文本单元中包含的每个文字的置信度。每个文字的置信度用于表征该文字识别的准确度，可以是在与该文字对应的字符图像被识别出文字后，与预先构建的字库进行比对得到。在一个例子中，每个文字的置信度可以分为多个等级，例如A级到E级，置信度依次降低。

[0092] 上述已经结合附图说明本发明实施例的步骤S2100，得到具有固定文本数据结构、可以描述目标图像显示的文本内容以及文本结构的目标结构化数据，之后进入S2200。

[0093] 步骤S2200，根据所述目标结构化数据，生成符合预设的文本格式的结构化文本。

[0094] 在本实施例中，预设的文本格式可以根据具体的应用场景设置，例如，针对数据量规模极大、数据交换频繁的大数据应用场景，可以选择更适用于数据交换的文本格式等等。

[0095] 在一个例子中，所述预设的文本格式是JSON、XML、Protobuf中的一种。

[0096] JSON(Java Script Object Notation，JS对象标记)是一种轻量级的数据交换格式，采用完全独立于编程语言的文本格式来存储和表示数据，易于人阅读和编写，同时也易于机器解析和生成，并有效地提升网络传输效率。

[0097] XML(Extensible Markup Language，可扩展标记语言)，是一种用于标记电子文件

使其具有结构性的标记语言可以用来标记数据、定义数据类型，并允许用户对自己的标记语言进行定义，提供统一的方法来描述和交换独立于应用程序或供应商的结构化数据，适合数据的交换、传输。

[0098] Protobuf(protocol buffer的缩写)，是Google公司提供的一种独立于语言和平台的数据交换的格式可以用于分布式应用之间的数据通信或者异构环境下的数据交换，具有较好的兼容性和较高的传输效率。

[0099] 选择JSON、XML、Protobuf中的一种作为预设的文本格式，以根据目标结构化数据生成结构化文本，可以更好地支持跨平台的数据交换，提升数据交换效率，尤其适用于数据交换频繁、涉及大量数据分析的大数据应用场景。

[0100] 在本发明实施例中，还提供一种结构化文本的生成装置3000，如图3所示，包括数据识别单元3100、文本生成单元3200，用于实施本实施例中提供的任一项结构化文本的生成方法，在此不再赘述。

[0101] 该生成装置3000，包括：

[0102] 数据识别单元3100，基于预设的文本数据结构对目标图像进行识别，获取的对应的目标结构化数据，其中，所述文本数据结构用于支持以结构化信息描述文本版式；

[0103] 具体地，所述预设的文本数据结构至少包括识别出的文本单元的数目、每个所述文本单元的结构化信息，所述结构化信息包括对应的所述文本单元的类型、单元位置信息、文字内容以及每个文字的文字位置信息；

[0104] 文本生成单元3200，用于根据所述目标结构化数据，生成符合预设的文本格式的结构化文本。

[0105] 具体地，所述文本单元的类型至少包括文本块；

[0106] 当所述文本单元是文本块时，所述文本单元的数目包括所述文本块的数目，每个所述文本块的单元位置信息至少包括对应的所述文本块的块序号以及文本行信息。

[0107] 或者，所述文本单元类型至少包括表格；

[0108] 当所述文本单元是表格时，所述文本单元的数目包括所述表格的数目，每个所述表格的单元位置信息至少包括对应的所述表格的块序号、表格行数目、表格列数目以及单元格信息。

[0109] 具体地，所述单元格信息包括对应的单元格所在的表格位置信息、单元格跨行数目、单元格跨行列数目、单元格包含的文本行信息。

[0110] 在一个例子中，所述结构化信息还包括对应的所述文本单元中包含的每个文字的置信度。

[0111] 具体地，所述文字位置信息包括与文字对应的符号图像相对于所述目标图像的左上角顶点作为原点得到的坐标位置信息。

[0112] 具体地，所述预设的文本格式是JSON、XML、Protobuf中的一种。

[0113] 本领域技术人员应当明白，可以通过各种方式来实现生成装置3000。例如，可以通过指令配置处理器来实现生成装置3000。例如，可以将指令存储在ROM中，并且当启动设备时，将指令从ROM读取到可编程器件中来实现生成装置3000。例如，可以将生成装置3000固化到专用器件(例如ASIC)中。可以将生成装置3000分成相互独立的单元，或者可以将它们合并在一起实现。生成装置3000可以通过上述各种实现方式中的一种来实现，或者可以通

过上述各种实现方式中的两种或更多种方式的组合来实现。

[0114] 在本实施例中，生成装置3000的实体设备形式可以如图1所示的计算机1100，具体地，可以是云平台服务器。

[0115] 〈例子〉

[0116] 以下将结合如图4、图5所示的例子进一步说明通过本发明实施例提供的结构化文本的生成装置，实施的本实施例提供的任意一项结构化文本的生成方法。

[0117] 在本例中，目标图像如图4所示。

[0118] 在预设的文本数据结构中，文本单元的类型包括文本块以及表格，文本块是一个完整的文本段，对应的，文本数据结构中包括：文本块的数目、每个文本块包含的文本行数目、文本行序号、文字内容、每个文字的文字位置信息、每个文字的置信度；表格的数目、每个表格里包括的表格行数目、表格列数目、每个单元格在表格里的表格行号、表格列号、跨行数目、跨列数据、单元格包含的文字内容、每个文字的文字位置信息(每个文字对应的字符图像相对于目标图像左上角顶点为原点的坐标信息)、每个文字的置信度等。预设的文本格式可以为JSON格式。

[0119] 具体地，该文本数据结构，可以如表1所示。

[0120] 表1文本数据结构表

[0121]

字段	父字段	描述
code	root	返回码
result	root	返回码所对应的信息

[0122]

block_count	root	文本块的总数
table_count	root	表格的总数
block_list	root	文本块的集合
table_list	root	表格的集合
block_list 结构如下		
order	block_list	在全篇文本的块顺序号，这个号是与表格进行全文总排序
line_list	block_list	一个文本块里行的集合
line_num	line_list	一个文本块的行号
text	line_list	识别出来的文字内容，第一候选
confidence	line_list	每个文字的置信度,从 A 到 E，置信度一次降低
pos	line_list	每个文字的位置，文字间用|分隔，位置之间用-分隔
table_list 结构如下		
order	table_list	一个表格在全篇文本的块顺序号，这个号是包含文本块的全文总排序
tr_count	table_list	一个表格行数
td_count	table_list	一个表格列数
table	table_list	表示一个表格
tr	table	一个表格里的一行
td	tr	某一行里的一列
colspan	td	单元格可横跨的列数
rowspan	td	单元格可纵跨的行数
line_list	td	一个单元格里行的集合

[0123]

line_num	line_list	一个文本块的行号
text	line_list	识别出来的文字内容，第一候选
confidence	line_list	每个文字的置信度,从 A 到 E，置信度一次降低
pos	line_list	每个文字的位置，文字间用|分隔，位置之间用-分隔

[0124] 例如,假设预设的文本格式为JSON格式,以图4所示目标图像的第一个文本块(“甲乙丙丁”建设专项课题项目申请表)为例,基于上述所示的文本数据结构,根据本发明实施例提供的结构化文本的生成方法,得到对应的JSON格式的目标结构化文本中,对应的结构化数据如图8所示。

[0125] 同样地,以图4所示目标图像的表格第一行(“课题名称”所在行)为例,基于上述所示的文本数据结构,根据本发明实施例提供的结构化文本的生成方法,得到对应的JSON格式的目标结构化文本中,对应的结构化数据如图9所示。

[0126] 对于图4所示的图像显示的文本中其他的文本块以及表格对应的结构化数据与上述所示的类似,只是基于文本数据结构对应的内容不同,在此不一一列举。

[0127] 通过上述处理,可以得到图4所示图像显示文本对应的JSON格式的结构化文本,在通过该结构化文本进行数据交换时,可以对不同的文本格式或者数据平台,都能根据对应的结构化数据,有效还原文本版式。

[0128] 例如,对于某些数据交换的场景,可以解析JSON格式的结构化文本,以如图5所示的EXCEL表格形式,将图4所示图像显示的文本表达出来,并且还保留原有图像显示文本的版式信息,使得用户可以通过符合自身应用需求的表达形式,获取图像所示的文本内容以及文本版式。

[0129] 以上已经结合符合附图和例子,说明了本实施例中提供的结构化文本的生成方法及装置,根据对目标图像识别得到符合预设的文本数据结构的目标结构化数据,生成对应的符合预设的文本格式的结构化文本,使得在数据交换过程中,可以对不同的文本格式或者数据平台,都能根据对应的结构化数据,有效还原文本版式,克服非结构数据存在的在数据交换过程中一旦变换文本格式或者跨越数据平台就难以还原文本版式的缺陷。尤其适用于数据交换频繁、涉及大量数据分析的大数据应用场景。

[0130] 〈第二实施例〉

[0131] 在本发明的第二实施例中,提供一种结构化文本的检索方法,如图6所示,包括:

[0132] 步骤S4100,接收文本检索请求,获取对应的用于文本检索的检索关键信息。

[0133] 其中,所述检索关键信息至少包括描述文本版式的结构化信息;

[0134] 具体地,所述检索关键信息包括文本单元的结构化信息的至少部分,所述结构化信息包括文本单元的类型、单元位置信息、文字内容以及每个文字的文字位置信息。

[0135] 可选地,所述文本单元的类型至少包括文本块;

[0136] 当所述文本单元是文本块时,所述文本单元的数目包括所述文本块的数目,每个所述文本块的单元位置信息至少包括对应的所述文本块的块序号以及文本行信息。

[0137] 或者,所述文本单元类型至少包括表格;

[0138] 当所述文本单元是表格时，所述文本单元的数目包括所述表格的数目，每个所述表格的单元位置信息至少包括对应的所述表格的块序号、表格行数目、表格列数目以及单元格信息。

[0139] 进一步可选地，所述单元格信息包括对应的单元格所在的表格位置信息、单元格跨行数目、单元格跨行列数目、单元格包含的文本行信息。

[0140] 可选地，所述结构化信息还包括对应的所述文本单元中包含的每个文字的置信度。

[0141] 可选地，所述文字位置信息包括与文字对应的符号图像相对于所述目标图像的左上角顶点作为原点得到的坐标位置信息。

[0142] 上述文本单元的结构化信息在第一实施例中已经详细描述，在此不再赘述。

[0143] 在步骤S4100中，支持用户进行检索文本时，输入的检索请求中包含的不是关键词信息，而是文本单元的结构化信息，例如，具体检索的文本单元的类型、文本块的文本行序号、或者单元格的表格行序号、表格列序号等，使得可以通过后续的步骤S4200，以文本单元的结构化信息对结构化文本进行检索，而不是通过关键词进行全文检索，提升检索效率。

[0144] 步骤S4200，采用获得的所述检索关键信息，在若干结构化文本中进行检索，获得与所述文本检索请求对应的文本内容，生成对应的检索结果，

[0145] 其中，所述结构化文本是包括对应的目标结构化数据的符合预设的文本格式的文本，所述目标结构化数据符合预设的文本数据结构，所述文本数据结构至少包括对应文本的文本单元的数目、每个所述文本单元的所述结构化信息。

[0146] 具体地，所述预设的文本格式是JSON、XML、Protobuf中的一种。上述三种文本格式在第一实施例中已经描述，在此不再赘述。

[0147] 在步骤S4200中，可以根据检索关键信息中文本单元的至少部分，例如，具体检索的文本单元的类型、文本块的文本行序号、或者单元格的表格行序号、表格列序号等，检索多个结构化文本。检索时可以直接定位到匹配的文本单元中，获取对应的文本内容，生成对应的检索结果，无需通过关键词进行全文检索，提升检索效率。尤其适用于涉及大量数据分析的大数据应用场景。

[0148] 在本实施例中，还提供一种结构化文本的检索装置5000，如图7所示，包括信息获取单元5100、检索执行单元5200，用于实施本实施例中提供的结构化文本的检索方法，在此不再赘述。

[0149] 该检索装置5000，包括：

[0150] 信息获取单元5100，用于接收文本检索请求，并基于所述文本检索请求获取用于文本检索的检索关键信息，

[0151] 其中，所述检索关键信息至少包括描述文本版式的结构化信息；

[0152] 具体地，所述检索关键信息包括文本单元的结构化信息的至少部分，所述结构化信息包括文本单元的类型、单元位置信息、文字内容以及每个文字的文字位置信息；

[0153] 检索执行单元5200，用于采用获得的所述检索关键信息，在若干结构化文本中进行检索，获得与所述文本检索请求对应的文本内容，生成对应的检索结果。

[0154] 具体地，所述结构化文本是包括对应的目标结构化数据的符合预设的文本格式的文本，所述目标结构化数据符合预设的文本数据结构，所述文本数据结构至少包括对应文

本的文本单元的数目、每个所述文本单元的所述结构化信息。

[0155] 具体地，所述文本单元的类型至少包括文本块；

[0156] 当所述文本单元是文本块时，所述文本单元的数目包括所述文本块的数目，每个所述文本块的单元位置信息至少包括对应的所述文本块的块序号以及文本行信息。

[0157] 或者，所述文本单元类型至少包括表格；

[0158] 当所述文本单元是表格时，所述文本单元的数目包括所述表格的数目，每个所述表格的单元位置信息至少包括对应的所述表格的块序号、表格行数目、表格列数目以及单元格信息。

[0159] 具体地，所述单元格信息包括对应的单元格所在的表格位置信息、单元格跨行数目、单元格跨行列数目、单元格包含的文本行信息。

[0160] 在一个例子中，所述结构化信息还包括对应的所述文本单元中包含的每个文字的置信度。

[0161] 具体地，所述文字位置信息包括与文字对应的符号图像相对于所述目标图像的左上角顶点作为原点得到的坐标位置信息。

[0162] 具体地，所述预设的文本格式是JSON、XML、Protobuf中的一种。

[0163] 本领域技术人员应当明白，可以通过各种方式来实现检索装置5000。例如，可以通过指令配置处理器来实现检索装置5000。例如，可以将指令存储在ROM中，并且当启动设备时，将指令从ROM读取到可编程器件中来实现检索装置5000。例如，可以将检索装置5000固化到专用器件(例如ASIC)中。可以将检索装置5000分成相互独立的单元，或者可以将它们合并在一起实现。检索装置5000可以通过上述各种实现方式中的一种来实现，或者可以通过上述各种实现方式中的两种或更多种方式的组合来实现。

[0164] 在本实施例中，检索装置5000的实体设备形式可以如图1所示的计算机1100，具体地，可以是云平台服务器。

[0165] 〈例子〉

[0166] 本发明实施例中，可以通过检索装置5000，实施本实施例中提供的结构化文本的检索方法，可以应用于检索征信报告、审计报告这类具有相对固定的文本结构的文本。

[0167] 具体地，可以对摄像机、扫描仪获取纸质的征信报告的图像，通过第一实施例中提供的结构化文本的生成方法，得到对应的结构化文本。对于多个同类型的征信报告，具有相对固定的文本结构，例如，第一个表格(对应的块序号为2)显示的是用户整体的征信情况，其中表格行序号2、表格列序号为2的单元格内容是用户的公积金贷款数目，可以在检索请求中设置检索关键信息是块序号为2、表格行序号2、表格列序号2的单元格，检索时定位多个征信报告对应的结构化文本中的第一个表格，获取表格行序号为2、表格列序号2的内容，从而得到多个用户的公积金贷款数目。检索效率较高。

[0168] 应用于检索审计报告的例子也类似，在此不再举例赘述。

[0169] 以上已经结合附图、例子描述了本发明实施例中提供的结构化文本的检索方法及装置，可以通过文本单元的至少部分结构化信息，对结构化文本进行检索，提升检索效果。尤其适用于检索数据规模巨大的大数据应用场景。

[0170] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0171] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0172] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令,并转发该计算机可读程序指令,以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0173] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等,以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA),该电子电路可以执行计算机可读程序指令,从而实现本发明的各个方面。

[0174] 这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0175] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0176] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它

设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0177] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。对于本领域技术人员来说公知的是，通过硬件方式实现、通过软件方式实现以及通过软件和硬件结合的方式实现都是等价的。

[0178] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。本发明的范围由所附权利要求来限定。

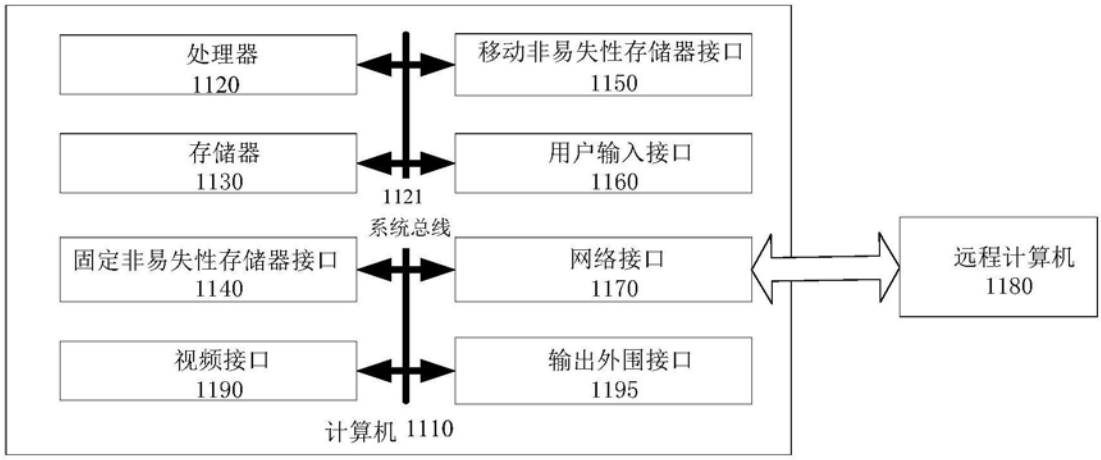

处理器
1120
移动非易失性存储器接口
1150
存储器
1130
用户输入接口
1160
1121
系统总线
固定非易失性存储器接口
1140
网络接口
1170
远程计算机
1180
视频接口
1190
输出外围接口
1195
计算机 1110

图1

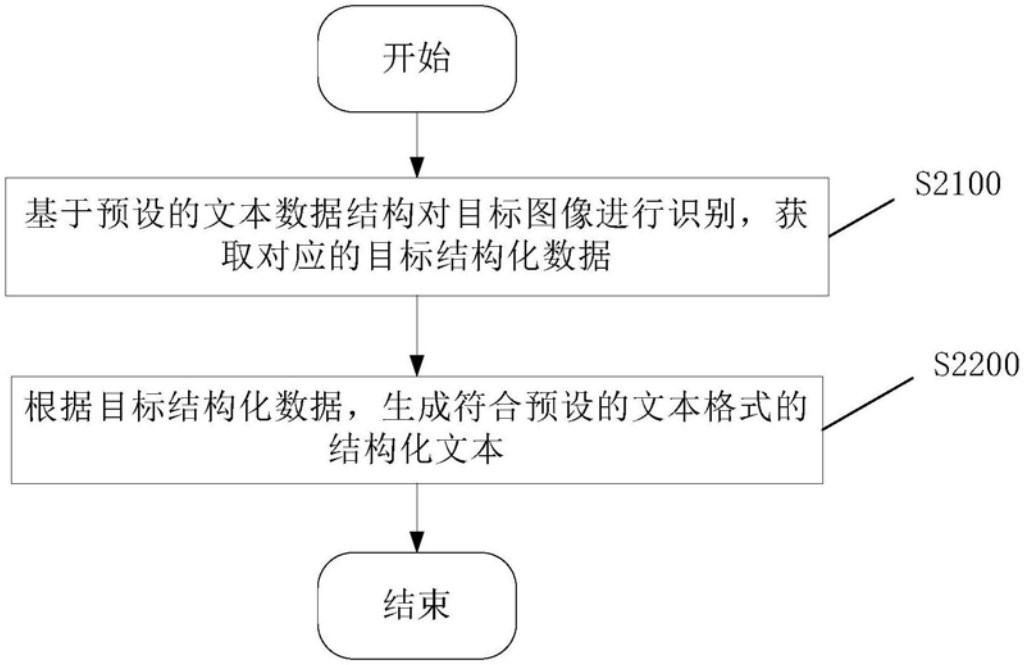

开始
基于预设的文本数据结构对目标图像进行识别，获取对应的目标结构化数据
S2100
根据目标结构化数据，生成符合预设的文本格式的结构化文本
S2200
结束

图2

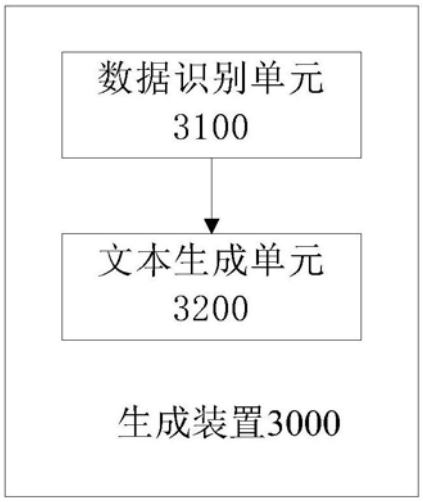


图3

“甲乙丙丁”建设专项课题项目申请表

2017年5月14日，王某某在“甲乙丙丁”区域合作高峰合作论坛上指出“推进‘甲乙丙丁’”建设，要聚焦发展这个根本性问题，释放各区发展潜力，实现经济大融合、发展大联动、成果大贡献”。随着“甲乙丙丁”建设的不断推进，鼓励和支持区内各个人员参与"甲乙丙丁"建设，将为我区各个人员带来新的发展机遇和广阔的发展空间。

<table>
<tr><td>课题名称</td><td colspan="6"></td></tr>
<tr><td>承担单位</td><td colspan="6"></td></tr>
<tr><td rowspan="3">课题负责人</td><td>姓 名</td><td></td><td>性 别</td><td></td><td>年龄</td><td></td></tr>
<tr><td>专 业</td><td colspan="2"></td><td>职务和职称</td><td colspan="2"></td></tr>
<tr><td>工作单位</td><td colspan="2"></td><td>联系电话</td><td colspan="2"></td></tr>
<tr><td rowspan="2">项目主要参加人员</td><td>姓 名</td><td colspan="3">工 作 单 位</td><td colspan="2">职务和职称</td></tr>
<tr><td></td><td colspan="3"></td><td colspan="2"></td></tr>
<tr><td>课题研究提纲</td><td colspan="6"></td></tr>
</table>

根据专项行动的精神，甲公司信息化部将与乙区贸促会建立工作机制，整合双方服务资源，创新服务模式，形成工作合力。甲公司信息化部作为区域促进各个人员发展工作的牵头部门，负责拟订促进各个人员对外经济合作的经济措施，推荐建立与各个区域和各个企业之间的各个人员之间的交流联系。

图4

块号	表格行号	表格列号	序号	识别结果 置信度低的文字加粗	左上X	左上Y	右下X	右下Y
1			0	“甲乙丙丁”建设专项课题项目申请表	127	17	257	24
2			0	2017年j月14日， 王某某 在“甲乙丙丁”区域合作高峰论坛上指出“推进‘甲	35	39	340	41
2			1	乙丙丁’建设，要聚焦发展这个根本**眭问**题，释放各区发展潜力，实现经济大融合、发展	22	48	340	54
2			2	大联动、成果大共享”，随着“甲乙丙丁”建设的不断推进，鼓**励**和支持 区内各个人员参	21	59	340	65
2			3	与“甲乙丙丁”建设，将为我区各个人员带来新的发展机遇和广阔的发展空间，	21	69	302	75
3	0	0	0	课题名称	25	98	55	104
3	1	0	0	承担单位	25	123	55	129
3	2	0	0	课题	37	161	52	168
3	2	0	1	负责人	34	172	56	178
3	2	1	0	姓 **名**	73	145	103	151
3	2	3	0	**性**别	185	145	206	151
3	2	5	0	年龄	283	145	298	151
3	3	0	0	专 **业**	73	167	104	173
3	3	2	0	职务和职称	223	167	261	173
3	4	0	0	工作单位	73	188	104	194
3	4	2	0	联系电话	223	188	253	194
3	5	0	0	项目主要参	26	223	64	230
3	5	0	1	加人员	34	234	55	240
3	5	1	0	姓 **名**	73	210	103	216
3	5	2	0	工作单位	174	211	217	216
3	5	3	0	职务和职称	285	211	324	216
3	7	0	0	课题	37	281	52	287
3	7	0	1	研究提纲	29	291	60	297
4			0	根据专项行动的精神，甲公司 信**邑**化部将与乙区 贸促会建立工作机制，整合**取**万服务	35	329	340	335
4			1	资源，创新服务模式，形成工作合力，甲公司信**邑**化部作为区域促进各个人员发展工作的	22	339	340	345
4			2	牵头部门，负责拟订促进各个人员对外经济合作的经**策**措施，推动建立与各个区域和各个企业之	21	349	340	355
4			3	间的 各个人员之间的交流联系	21	359	137	365

图5

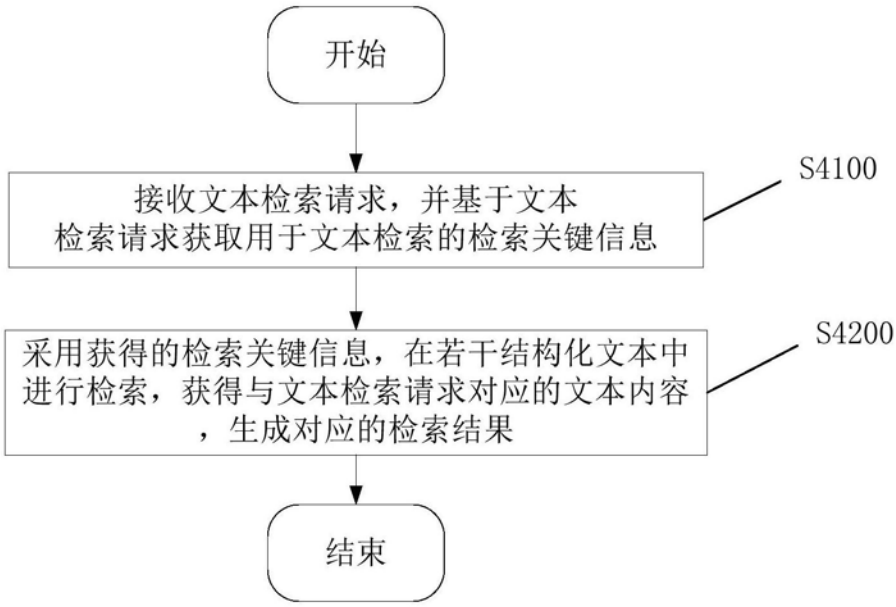

开始
接收文本检索请求，并基于文本
检索请求获取用于文本检索的检索关键信息
S4100
采用获得的检索关键信息，在若干结构化文本中
进行检索，获得与文本检索请求对应的文本内容
，生成对应的检索结果
S4200
结束

图6

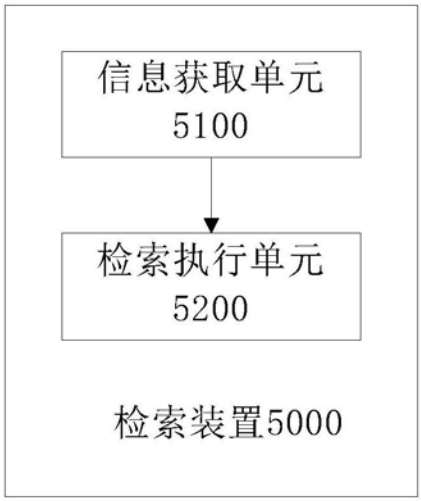

信息获取单元
5100
检索执行单元
5200
检索装置5000

图7

```
{
  "code": "0", //返回码
  "result": "ESC_SUCCESS", //返回码所对应的信息:处理成功
  "block_count": "3", //文本块的总数：3
  "table_count": "1", //表格的总数:1
  "block_list": //文本块的集合
  [
   {
     "order": "1", //块顺序号:1
     "line_list": //文本块里行的集合
      [
       {
        "line_num": "0", //行顺序号：0
        "text": "“甲乙丙丁”建设专项课题项目申请表", // 文字内容
        "confidence": "AAAAAEAAAAAAAAAAA", //文字置信度
        "pos":    //文字位置
            "307-43-313-47|318-48-334-49|338-42-353-58|356-48-372-49|37
            5-42-391-58|396-43-402-47|413-42-429-58|432-42-448-58|451-42
            -467-58|470-42-486-58|489-42-505-57|509-42-524-57|527-42-543
            -58|549-43-559-57|567-42-579-58|584-42-600-58|603-42-619-58"
                    }
```


图8

```
"table_list":
 [
  {
    "order": "3",   //块顺序号:1
    "tr_count": "8",   //表格行数:8
    "td_count": "7",   //表格列数:7
    "table":   //表格
     [
      {
       "tr": //表格里的一行
           [
            {
              "td": //一行里的一列
                  {
                    "colspan": "1", //单元格可横跨的列数：1
                    "rowspan": "1", //单元格可纵跨的行数:1
                    "line_list":
                      [
                       {
                        "line_num": "0", //文本块的行号:0
                        "text": "课题名称", //文本内容
                        "confidence": "AAAA", //文字置信度
                        "pos":     //文字位置
                              "61-236-77-251|81-236-96-251|100-236-
                              113-251|118-236-134-251"
                                    }
                            ]
                       }……
```


图9