(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117540802 A
(43)申请公布日 2024.02.09

(21)申请号 202311205946.7

(22)申请日 2023.09.18

(71)申请人 杭州阿里云飞天信息技术有限公司
地址 311121 浙江省杭州市余杭区五常街道文一西路969号3幢5层553室

(72)发明人 陈湘楠 肖谦 李俊成 董铎 林君 刘晓钟 汤斯亮

(74)专利代理机构 北京博浩百睿知识产权代理有限责任公司 11134
专利代理师 李静茹

(51)Int.Cl.
G06N 5/025(2023.01)
G06F 40/30(2020.01)

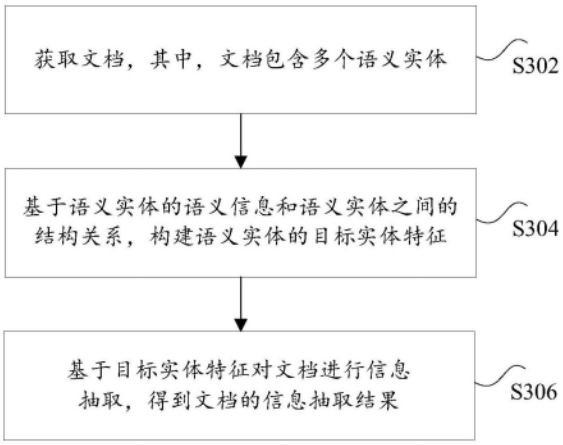
权利要求书2页 说明书21页 附图6页

(54)发明名称

文档的信息抽取方法、系统、电子设备和存储介质

(57)摘要

本申请公开了一种文档的信息抽取方法、系统、电子设备和存储介质，涉及大模型技术、文档处理领域。其中，该方法包括：获取文档，其中，文档包含多个语义实体；基于语义实体的语义信息和语义实体之间的结构关系，构建语义实体的目标实体特征；基于目标实体特征对文档进行信息抽取，得到文档的信息抽取结果。本申请解决了相关技术中对文档的信息抽取效果差的技术问题。



CN 117540802 A

1.一种文档的信息抽取方法,其特征在于,包括:

获取文档,其中,所述文档包含多个语义实体;

基于所述语义实体的语义信息和所述语义实体之间的结构关系,构建所述语义实体的目标实体特征;

基于所述目标实体特征对所述文档进行信息抽取,得到所述文档的信息抽取结果。

2.根据权利要求1所述的方法,其特征在于,基于所述语义实体的语义信息和所述语义实体之间的结构关系,构建所述语义实体的目标实体特征,包括:

利用文档处理模型基于所述语义信息和所述结构关系,构建所述目标实体特征。

3.根据权利要求2所述的方法,其特征在于,所述文档解析模型包括:特征提取组件、关系生成组件和信息挖掘组件,利用文档处理模型基于所述语义信息和所述结构关系,构建所述目标实体特征,包括:

利用所述特征提取组件对所述语义实体进行特征提取,得到所述语义实体的初始实体特征,其中,所述初始实体特征用于表征所述语义实体的所述语义信息;

利用所述关系生成组件对所述初始实体特征进行映射,得到多个所述语义实体的关系特征,其中,所述关系特征用于表征多个所述语义实体之间的语义关系;

利用所述信息挖掘组件对所述关系特征和所述结构关系进行注意力处理,得到全局结构信息;

将所述初始实体特征和所述全局结构信息进行融合,得到所述目标实体特征。

4.根据权利要求3所述的方法,其特征在于,所述特征提取组件包括:预训练文档模型和两个不同的线性层,利用所述特征提取组件对所述语义实体进行特征提取,得到所述语义实体的初始实体特征,包括:

利用所述预训练文档模型对所述语义实体进行编码,得到所述语义实体的编码向量;

利用所述两个不同的线性层分别对所述编码向量进行处理,生成所述语义实体的初始键特征和初始值特征,其中,所述初始键特征用于表征所述语义实体中位于头部的实体,所述初始值特征用于表征所述语义实体中位于尾部的实体;

基于所述初始键特征和所述初始值特征,得到所述初始实体特征。

5.根据权利要求3所述的方法,其特征在于,所述关系生成组件包括:关系确定模块和关系生成模块,利用所述关系生成组件对所述初始实体特征进行映射,得到多个所述语义实体的关系特征,包括:

利用所述关系确定模块基于所述初始实体特征,确定所述语义实体之间的语义关系;

利用所述关系生成模块基于所述语义关系和所述初始实体特征,生成所述关系特征。

6.根据权利要求3所述的方法,其特征在于,所述信息挖掘组件包括:局部注意力层、全局交互层、以及与所述局部注意力层和所述全局交互层连接的池化层,利用所述信息挖掘组件对所述关系特征和所述结构关系进行注意力处理,得到全局结构信息,包括:

利用所述局部注意力层对所述关系特征进行自注意力处理,得到第一注意力特征;

利用所述全局交互层对所述关系特征和所述结构关系进行交叉注意力处理,得到第二注意力特征;

对所述第一注意力特征和所述第二注意力特征进行叠加,得到叠加特征;

对所述叠加特征进行池化处理,得到所述全局结构信息。

7.根据权利要求3所述的方法,其特征在于,在所述文档解析模型包括:多个所述关系生成组件和多个所述信息挖掘组件的情况下,多个所述关系生成组件的数量与多个所述信息挖掘组件的数量相同,且多个所述关系生成组件和多个所述信息挖掘组件依次迭代运行。

8.根据权利要求1所述的方法,其特征在于,所述方法还包括:

对所述文档进行光学识别,确定所述语义实体所属边界框在所述文档中的空间信息;

基于所述空间信息,确定所述语义实体之间的距离和角度;

基于所述距离和所述角度,生成所述结构关系。

9.根据权利要求8所述的方法,其特征在于,所述空间信息包括:多个端点的子空间信息,基于所述空间信息,确定所述语义实体之间的距离和角度,包括:

基于所述多个端点中相同端点的子空间信息,确定所述语义实体之间的多个子距离和多个子角度;

对所述多个子距离进行拼接,得到所述语义实体之间的所述距离;

对所述多个子角度进行拼接,得到所述语义实体之间的所述角度。

10.根据权利要求2所述的方法,其特征在于,所述方法还包括:

在交互界面中输出所述文档和所述信息抽取结果;

响应作用于所述交互界面中的修改指令,确定所述修改指令对应的目标抽取结果;

基于所述文档和所述目标抽取结果,对所述文档处理模型的模型参数进行更新。

11.一种文档的信息抽取方法,其特征在于,包括:

响应交互界面中接收到的文档,其中,所述文档包含多个语义实体;

基于所述语义实体的语义信息和所述语义实体之间的结构关系,构建所述语义实体的目标实体特征;

基于所述目标实体特征对所述文档进行信息抽取,得到所述文档的信息抽取结果;

在所述交互界面中显示所述信息抽取结果。

12.一种文档的信息抽取系统,其特征在于,包括:

接收模块,用于显示交互界面,并捕获在所述交互界面中输入的文档,其中,所述文档包含多个语义实体;

处理装置,用于基于所述语义实体的语义信息和所述语义实体之间的结构关系,构建所述语义实体的目标实体特征,并基于所述目标实体特征对所述文档进行信息抽取,得到所述文档的信息抽取结果;

发送模块,用于将所述信息抽取结果发送至前端客户端,其中,所述信息抽取结果由所述前端客户端显示在所述交互界面中。

13.一种电子设备,其特征在于,包括:

存储器,存储有可执行程序;

处理器,用于运行所述程序,其中,所述程序运行时执行权利要求1至11中任意一项所述的方法。

14.一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包括存储的可执行程序,其中,在所述可执行程序运行时控制所述计算机可读存储介质所在设备执行权利要求1至11中任意一项所述的方法。

文档的信息抽取方法、系统、电子设备和存储介质

技术领域

[0001] 本申请涉及大模型技术、文档处理领域，具体而言，涉及一种文档的信息抽取方法、系统、电子设备和存储介质。

背景技术

[0002] 目前，企业数据大多以视觉文档的形式存在，例如以文本、图片、扫描件、电子表格、在线文档、邮件等形式，但是这些形式的文档通常难以流通和处理，大部分企业对视觉文档进行处理的方式是采用图神经网络，从视觉文档中抽取出较为关键的结构信息，然后再对该结构信息进行处理，以提高对视觉文档进行处理的效率。但是图神经网络大多是以NLP模型(Neuro-Linguistic Programming，逻辑思维层次模型)为基座进行训练的，对视觉文档中存在的结构化知识的捕获能力较弱，无法预测长距离的语义实体之间的依赖关系，对视觉文档进行信息抽取的效果较差。

[0003] 针对上述的问题，目前尚未提出有效的解决方案。

发明内容

[0004] 本申请实施例提供了一种文档的信息抽取方法、系统、电子设备和存储介质，以至少解决相关技术中对文档的信息抽取效果差的技术问题。

[0005] 根据本申请实施例的一个方面，提供了一种文档的信息抽取方法，包括：获取文档，其中，文档包含多个语义实体；基于语义实体的语义信息和语义实体之间的结构关系，构建语义实体的目标实体特征；基于目标实体特征对文档进行信息抽取，得到文档的信息抽取结果。

[0006] 根据本申请实施例的另一方面，还提供了一种文档的信息抽取方法，包括：响应交互界面中接收到的文档，其中，文档包含多个语义实体；基于语义实体的语义信息和语义实体之间的结构关系，构建语义实体的目标实体特征；基于目标实体特征对文档进行信息抽取，得到文档的信息抽取结果；在交互界面中显示信息抽取结果。

[0007] 根据本申请实施例的另一方面，还提供了一种文档的信息抽取方法，包括：通过调用第一接口获取文档，其中，第一接口包括第一参数，第一参数的参数值为文档，文档包含多个语义实体；基于语义实体的语义信息和语义实体之间的结构关系，构建语义实体的目标实体特征；基于目标实体特征对文档进行信息抽取，得到文档的信息抽取结果；通过调用第二接口输出信息抽取结果，其中，第二接口包括第二参数，第二参数的参数值为信息抽取结果。

[0008] 根据本申请实施例的另一方面，还提供了一种文档的信息抽取系统，其特征在于，包括：接收模块，用于显示交互界面，并捕获在交互界面中输入的文档，其中，文档包含多个语义实体；处理装置，用于基于语义实体的语义信息和语义实体之间的结构关系，构建语义实体的目标实体特征，并基于目标实体特征对文档进行信息抽取，得到文档的信息抽取结果；发送模块，用于将所述信息抽取结果发送至所述前端客户端，其中，所述信息抽取结果

由所述前端客户端显示在所述交互界面中。

[0009] 根据本申请实施例的另一方面,还提供了一种电子设备,包括:存储器,存储有可执行程序;处理器,用于运行程序,其中,程序运行时执行上述任意一项的方法。

[0010] 根据本申请实施例的另一方面,还提供了一种计算机可读存储介质,计算机可读存储介质包括存储的可执行程序,其中,在可执行程序运行时控制计算机可读存储介质所在设备执行上述任意一项的方法。

[0011] 在本申请实施例中,采用获取文档,其中,文档包含多个语义实体;基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果的方式,通过将语义实体的语义信息作为文档的局部实体表示,并根据语义实体之间的结构关系构建文档的全局结构信息,然后再根据该语义信息和结构信息,构建出包含文档的局部实体表示和全局结构信息的目标实体特征,以保证目标实体特征能够与文档全文相关联,具有较高的鲁棒性,最后再根据该目标实体特征对文档进行信息抽取,可以进一步地保证得到的信息抽取结果与文档的内容相符,提高信息抽取结果的准确度,从而实现了提高对文档进行信息抽取得到的结果的准确度的技术效果,进而解决了相关技术中对文档的信息抽取效果差的技术问题。

[0012] 容易注意到的是,上面的通用描述和后面的详细描述仅仅是为了对本申请进行举例和解释,并不构成对本申请的限定。

附图说明

[0013] 此处所说明的附图用来提供对本申请的进一步理解,构成本申请的一部分,本申请的示意性实施例及其说明用于解释本申请,并不构成对本申请的不当限定。在附图中:

[0014] 图1是根据本申请实施例1的一种文档的信息抽取方法的虚拟现实设备的硬件环境的示意图;

[0015] 图2是根据本申请实施例1的一种文档的信息抽取方法的计算环境的结构框图;

[0016] 图3是根据本申请实施例1的一种文档的信息抽取方法的流程图;

[0017] 图4是根据本申请实施例1的一种文档的信息抽取过程的示意图;

[0018] 图5是根据本申请实施例1的一种语义实体的结构关系确定过程的示意图;

[0019] 图6是根据本申请实施例2的一种文档的信息抽取方法的流程图;

[0020] 图7是根据本申请实施例3的文档的信息抽取方法的流程图;

[0021] 图8是根据本申请实施例4的一种文档的信息抽取装置的结构框图;

[0022] 图9是根据本申请实施例5的一种文档的信息抽取装置的结构框图;

[0023] 图10是根据本申请实施例6的一种文档的信息抽取装置的结构框图;

[0024] 图11是根据本申请实施例7的一种文档的信息抽取系统的结构框图;

[0025] 图12是根据本申请实施例8的一种电子设备的结构框图。

具体实施方式

[0026] 为了使本技术领域的人员更好地理解本申请方案,下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本申请一部分的实施例,而不是全部的实施例。基于本申请中的实施例,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,都应当属于本申请保护的范围。

[0027] 需要说明的是,本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0028] 本申请提供的技术方案主要采用大模型技术实现,此处的大模型是指具有大规模模型参数的深度学习模型,通常可以包含上亿、上百亿、上千亿、上万亿甚至十万亿以上的模型参数。大模型又可以称为基石模型/基础模型(Foundat ion Model),通过大规模无标注的语料进行大模型的预训练,产出亿级以上参数的预训练模型,这种模型能适应广泛的下游任务,模型具有较好的泛化能力,例如大规模语言模型(Large Language Model,LLM)、多模态预训练模型(mult i-modal pre-training model)等。

[0029] 需要说明的是,大模型在实际应用时,可以通过少量样本对预训练模型进行微调,使得大模型可以应用于不同的任务中。例如,大模型可以广泛应用于自然语言处理(Natural Language Process ing,简称NLP)、计算机视觉、语音处理等领域,具体可以应用于如视觉问答(Visual Quest ion Answer ing,简称VQA)、图像描述(Image Capt ion,简称IC)、图像生成等计算机视觉领域任务,也可以广泛应用于基于文本的情感分类、文本摘要生成、机器翻译等自然语言处理领域任务。因此,大模型主要的应用场景包括但不限于数字助理、智能机器人、搜索、在线教育、办公软件、电子商务、智能设计等。在本申请实施例中,以利用大语言模型对文档进行信息抽取为例进行解释说明。

[0030] 首先,在对本申请实施例进行描述的过程中出现的部分名词或术语适用于如下解释:

[0031] 视觉文档:以文本、图片、扫描件、电子表格、在线文档、邮件等形式存储信息的文档,用户可以通过阅读视觉文档获取自身需求的信息。

[0032] 文档解析:以非结构化的文档为输入,依托文档智能预训练技术和产品对文档进行信息抽取,输出处理后的结构化数据的过程。

[0033] 实施例1

[0034] 根据本申请实施例,提供了一种文档的信息抽取方法,需要说明的是,在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行,并且,虽然在流程图中示出了逻辑顺序,但是在某些情况下,可以以不同于此处的顺序执行所示出或描述的步骤。

[0035] 考虑到大模型的模型参数量庞大,且移动终端的运算资源有限,本申请实施例提供的上述文档的信息抽取方法可以应用于如图1所示的应用场景,但不仅限于此。图1是根据本申请实施例1的一种文档的信息抽取方法的硬件环境的示意图。在如图1所示的应用场景中,大模型部署在服务器10中,服务器10可以通过局域网连接、广域网连接、因特网连接,或者其他类型的数据网络,连接一个或多个客户端设备20,此处的客户端设备20可以包括

但不限于:智能手机、平板电脑、笔记本电脑、掌上电脑、个人计算机、智能家居设备、车载设备等。客户端设备20可以通过图形用户界面与用户进行交互,实现对大模型的调用,进而实现本申请实施例所提供的方法。

[0036] 在本申请实施例中,客户端设备和服务器构成的系统可以执行如下步骤:获取文档,其中,文档包含多个语义实体;基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果,从而解决了相关技术中对文档的信息抽取效果差的技术问题,达到了提高从文档中抽取出的信息的准确度的目的。需要说明的是,在客户端设备的运行资源能够满足大模型的部署和运行条件的情况下,本申请实施例可以在客户端设备中进行。

[0037] 图2是根据本申请实施例1的一种文档的信息抽取方法的计算环境的结构框图,如图2所示,计算环境201包括运行在分布式网络上的多个(图中采用210-1,210-2,…,来示出)计算节点(如服务器)。不同计算节点都包含本地处理和内存资源,终端用户202可以在计算环境201中远程运行应用程序或存储数据。应用程序可以作为计算环境201中的多个服务220-1,220-2,220-3和220-4进行提供,分别代表服务“A”,“D”,“E”和“H”。

[0038] 终端用户202可以通过客户端上的web浏览器或其他软件应用程序提供和访问服务,在一些实施例中,可以将终端用户202的供应和/或请求提供给入口网关230。入口网关230可以包括一个相应的代理来处理针对服务(计算环境201中提供的一个或多个服务)的供应和/或请求。

[0039] 服务是根据计算环境201支持的各种虚拟化技术来提供或部署的。在一些实施例中,可以根据基于虚拟机(Virtual Machine,VM)的虚拟化、基于容器的虚拟化和/或类似的方式提供服务。基于虚拟机的虚拟化可以是通过初始化虚拟机来模拟真实的计算机,在不直接接触任何实际硬件资源的情况下执行程序和应用程序。在虚拟机虚拟化机器的同时,根据基于容器的虚拟化,可以启动容器来虚拟化整个操作系统(Operating System,OS),以便多个工作负载可以在单个操作系统实例上运行。

[0040] 在基于容器虚拟化的一个实施例中,服务的若干容器可以被组装成一个Pod(例如,Kubernetes Pod)。举例来说,如图2所示,服务220-2可以配备一个或多个Pod 240-1,240-2,…,240-N(统称为Pod)。Pod可以包括代理245和一个或多个容器242-1,242-2,…,242-M(统称为容器)。Pod中一个或多个容器处理与服务的一个或多个相应功能相关的请求,代理245通常控制与服务相关的网络功能,如路由、负载均衡等。其他服务也可以配备类似的Pod。

[0041] 在操作过程中,执行来自终端用户202的用户请求可能需要调用计算环境201中的一个或多个服务,执行一个服务的一个或多个功能需要调用另一个服务的一个或多个功能。如图2所示,服务“A”220-1从入口网关230接收终端用户202的用户请求,服务“A”220-1可以调用服务“D”220-2,服务“D”220-2可以请求服务“E”220-3执行一个或多个功能。

[0042] 上述的计算环境可以是云计算环境,资源的分配由云服务提供上管理,允许功能的开发无需考虑实现、调整或扩展服务器。该计算环境允许开发人员在不构建或维护复杂基础设施的情况下执行响应事件的代码。服务可以被分割完成一组可以自动独立伸缩的功能,而不是扩展单个硬件设备来处理潜在的负载。

[0043] 在上述运行环境下,本申请提供了如图3所示的文档的信息抽取方法。需要说明的

是,该实施例的文档的信息抽取方法可以由图1所示实施例的移动终端执行。图3是根据本申请实施例1的一种文档的信息抽取方法的流程图。如图3所示,该方法可以包括如下步骤:

[0044] 步骤S302,获取文档,其中,文档包含多个语义实体。

[0045] 上述文档可以是指需要用户通过肉眼观察的文档,例如视觉文档,可以包括但不限于:文本、图片、扫描件、电子表格、在线文档、邮件等,为便于理解,下文以文档中包含的文本进行说明。上述语义实体可以是指文档中具有特定意义的实体,可以是指用户在阅读文档时,能够根据文档的背景或术语解释等内容理解的数据,例如可以是具体的人物、地点、商品、颜色,也可以是抽象出的事件、日期、数字或符号等。为便于理解,若文档是一件商品的介绍文本,则对应的语义实体可以是指该介绍文本中介绍的该商品的商品名称,用户可以通过该商品名称快速的联想到该商品。一般的,一个文档中可以包含多个语义实体。

[0046] 一般的,在文档的数据量较小,例如文档中仅包含的一个语义实体的情况下,信息抽取系统可以直接对该文档进行信息抽取,抽取出与该语义实体相关的关键词、句,用户可以根据该关键词、句,快速地获取文档所包含的内容,不需要过多的考虑抽取出的关键词、句是否能够准确的反映出文档的整体内容,但是在文档的数据量较大,例如文档中包含较多的语义实体的情况下,由于文档包含的内容较多,即与语义实体相关的信息较多,信息抽取系统对文档进行信息抽取时,抽取出的多个关键词、句等信息可能与语义实体不匹配,导致多个关键词、句等信息与文档实际包含的内容不符,进而导致用户根据该关键词、句等信息理解到错误的文档内容,因此,为了提高从文档中抽取出的信息的准确度,即抽取出的信息能够准确的反映出文档所包含的内容,信息抽取系统可以首先获取需要用户理解的文档,该文档中包含多个语义实体,以便于后续使用语义理解的模型辅助用户理解文档。

[0047] 步骤S304,基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征。

[0048] 上述语义信息可以是指文档中语义实体的语义,可以以特征向量的形式表示。上述结构关系可以是指不同语义实体之间的空间几何信息,例如不同语义文本之间的坐标位置关系、大小关系等,可以用来构建文档的全局结构信息。上述目标实体特征可以是指能够反映出文档的局部实体表示和全局结构信息的特征,其中,局部实体表示可以是指文档中的语义实体自身所包含信息,例如上述语义实体的语义信息,全局结构信息可以是指文档中不同语义实体之间的关联关系信息,例如语义实体的结构关系。

[0049] 在本实施例一种可选方案中,考虑到语义实体自身所包含的信息较为单一,不同语义实体之间可能不存在明确的指向关系,信息抽取系统在根据文档中不同语义实体的语义信息对文档进行信息抽取时,可能会出现将不同语义实体匹配错误的情况,例如,若文档中包括语义实体:“鞋子”、“展示台”、“白色”和“黑色”,其中,“鞋子”与“白色”对应,“展示台”与“黑色”对应,若仅根据语义实体的语义信息对该文档进行信息抽取,可能会出现将“鞋子”与“黑色”进行匹配、将“展示台”与“白色”进行匹配的情况,并且若相互匹配的语义实体之间的距离越长,则仅根据语义实体的语义信息将两者相互匹配的概率越低,而错误的匹配会导致对文档进行信息抽取得到的结果与文档实际所包含的内容不符,进而会对用户理解该文档产生误导,因此,为了保证对文档进行信息抽取得到的结果的准确度,信息抽取系统需要准确地建立不同语义实体之间的匹配关系。而考虑到用户或者机器在撰写文档时,通常会根据自身的撰写习惯或者文档的格式要求来撰写,而撰写习惯和格式要求通常

是不会发生改变的,对应的文档最终呈现出的不同语义实体之间的空间几何信息,即上述的结构关系,通常也不会发生变化,例如,若用户A喜欢将主语放在宾语前,则用户A在撰写前述的鞋子和展示台时,通常会写成“展示台的颜色为黑色,鞋子的颜色为白色”,对应的与语义实体“鞋子”相匹配的语义实体为后面的“白色”,而非前面的“黑色”,这也就意味着可以利用文档中语义实体的语义信息和语义实体之间的结构关系,来确定语义实体和语义信息的匹配情况,从而提高建立出的不同语义实体之间的匹配关系的准确度。其中,可以根据撰写该文档的用户或机器的撰写习惯、格式要求等信息来确定上述的结构关系。

[0050] 因此,为了保证对文档进行抽取得到的结果的准确度,即保证该结果与文档实际所包含的内容相符,信息抽取系统可以在对文档进行信息抽取的过程中,引入文档中语义实体之间的结构关系,来提高对世界文档进行信息抽取的准确度。而考虑到相比于语义实体之间的结构关系,文档的全局结构信息能够更好的体现文档中不同语义实体之间的依赖关系,提高对文档进行信息抽取的效果,因此信息抽取系统还可以进一步地根据不同语义实体之间的结构关系,构建出能够反映文档中长距离的语义实体之间的依赖关系的全局结构信息,并在对文档进行信息抽取时,将文档中的局部实体表示和全局结构信息相结合,即根据不同语义实体各自的语义信息,以及不同语义实体之间的结构关系,构建出上述的目标实体特征,使构建出的目标实体特征能够与文档整体相关联,具有较高的鲁棒性。

[0051] 步骤S306,基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果。

[0052] 在构建出文档中语义实体对应的目标实体特征之后,信息抽取系统便可以根据该目标实体特征对文档进行信息抽取,从而保证抽取得到的信息抽取结果的准确度。

[0053] 在本申请实施例中,采用获取文档,其中,文档包含多个语义实体;基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果的方式,通过将语义实体的语义信息作为文档的局部实体表示,将并根据语义实体之间的结构关系作为构建文档的全局结构信息,并然后再根据该语义信息和结构信息,构建出包含文档的局部实体表示和全局结构信息语义实体的目标实体特征,以保证目标实体特征能够与文档全文相关联,具有较高的鲁棒性,同时,然后最后再根据该目标实体特征对文档进行信息抽取,可以进一步地保证得到的信息抽取结果与文档的内容相符,提高信息抽取结果的准确度,从而实现了提高对文档进行信息抽取得到的结果的准确度的技术效果,进而解决了相关技术中对文档的信息抽取效果差的技术问题。

[0054] 在本申请实施例中,基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征,包括:利用文档处理模型基于语义信息和结构关系,构建目标实体特征。

[0055] 上述文档处理模型可以是指将文档的局部实体表示和全局结构信息进行融合解析的模型。

[0056] 在本实施例一种可选方案中,可以利用上述的文档处理模型对文档中语义实体的语义信息,以及该语义实体对应的结构关系进行融合,得到上述的目标实体特征。

[0057] 需要说明的是,在对文档进行信息抽取的过程中使用到的模型,均可以是大模型,例如上述的文档处理模型可以是大语言模型,利用大语言模型基于上述的语义信息和结构关系,构建上述的目标实体特征,可以保证构建出的目标实体特征的准确度。

[0058] 在本实施例一种可选方案中,上述文档处理模型中可以至少包括:用于获取语义实体的语义信息的组件,用于获取语义实体之间的关系,例如结构关系、语义关系的组件,以及用于根据结构关系构建文档的全局结构信息的组件,从而使信息抽取系统能够准确的获取到文档中的局部实体表示和全局结构信息,进而提高根据语义信息和结构信息构建出的目标实体特征的准确度。

[0059] 在本申请实施例中,文档解析模型包括:特征提取组件、关系生成组件和信息挖掘组件,利用文档处理模型基于语义信息和结构关系,构建目标实体特征,包括:利用特征提取组件对语义实体进行特征提取,得到语义实体的初始实体特征,其中,初始实体特征用于表征语义实体的语义信息;利用关系生成组件对初始实体特征进行映射,得到多个语义实体的关系特征,其中,关系特征用于表征多个语义实体之间的语义关系;利用信息挖掘组件对关系特征和结构关系进行注意力处理,得到全局结构信息;将初始实体特征和全局结构信息进行融合,得到目标实体特征。

[0060] 上述特征提取组件可以是指对语义实体进行特征提取,得到语义实体的语义信息的组件。上述初始实体特征可以是指语义实体的特征向量,可以用来表示语义实体的语义信息。上述关系生成组件可以是指用来初步确定不同语义实体之间的关联关系的组件。上述关系特征可以是指能够反映不同语义实体之间的关联关系,例如语义关系的特征。上述信息挖掘组件可以是指用来根据语义实体的结构关系确定文档的全局结构信息的组件,可以用来挖掘不同语义实体之间的依赖关系。为了提高构建出的目标实体特征的准确度,上述的文档解析模型可以至少包括:特征特区组件、关系生成组件和信息挖掘组件。

[0061] 在本实施例一种可选方案中,信息抽取系统在根据语义实体的局部实体表示和文档的全局结构信息,构建目标实体特征时,可以首先利用上述的特征提取组件来确定该局部实体表示,即可以利用特征提取组件对文档中的语义实体进行特征提取,对语义实体进行特征提取,得到语义实体对应的初始实体特征。

[0062] 在本实施例一种可选方案中,考虑到通过利用语义实体对应的初始实体特征,能够初步的确定不同语义实体之间的语义关系,而该语义关系可以初步地反映出不同语义实体之间的关联程度,因此,在得到语义实体对应的初始实体特征之后,信息抽取系统可以进一步地利用上述的关系生成组件对不同语义实体对应的初始实体特征进行映射处理,得到不同语义实体之间的关系特征。

[0063] 在本实施例一种可选方案中,考虑到不同语义实体之间的关联程度是不同的,例如在前文中提到的“鞋子”和“白色”的关联程度较高,“鞋子”和“展示台”的关联程度次之,“鞋子”和“黑色”的关联程度较低,而仅根据语义实体的结构关系可能无法准确的表示出不同语义实体之间的关联程度,而上述确定出的语义实体的关系特征可以反映出该关联程度,因此,信息抽取系统在根据语义实体的结构信息确定文档的全局结构信息时,可以在不同语义实体的结构关系的基础上,结合不同语义实体之间的关系特征,来确定上述的全局结构信息,即可以利用上述的信息挖掘组件对语义实体的关系特征和结构关系进行处理,以得到文档的全局结构信息,从而保证全局结构信息的准确度,进一步地,考虑到不同语义实体之间的依赖关系的计算复杂度较大,例如一篇文档中若包含有N个语义实体,则不同语义实体之间的关系特征、结构关系的数量就会有N^2个,对应的计算复杂度为$O(N^4)$,因此,可以利用自注意力机制,对上述的关系特征和结构关系进行注意力处理,将计算复杂度降低

至$O(N^2)$,从而提高得到全局结构信息的效率。

[0064] 在得到文档的全局结构信息之后,信息抽取系统便可以将上述的初始实体特征和该全局结构信息进行融合处理,从而得到上述的目标实体特征。

[0065] 在本申请实施例中,特征提取组件包括:预训练文档模型和两个不同的线性层,利用特征提取组件对语义实体进行特征提取,得到语义实体的初始实体特征,包括:利用预训练文档模型对语义实体进行编码,得到语义实体的编码向量;利用两个不同的线性层分别对编码向量进行处理,生成语义实体的初始键特征和初始值特征,其中,初始键特征用于表征语义实体中位于头部的实体,初始值特征用于表征语义实体中位于尾部的实体;基于初始键特征和初始值特征,得到初始实体特征。

[0066] 上述预训练文档模型可以是指用于对语义实体进行编码,得到语义实体对应的向量的模型,例如LayoutLM,XYLayoutLM等模型。上述初始键特征可以是指语义实体中位于头部的实体的实体特征,上述初始值特征可以是指语义实体中位于尾部的实体的实体特征。通过确定不同语义实体之间的初始键特征和初始值特征之间的匹配程度,可以初步地确定出不同语义实体之间是否存在关联关系。

[0067] 在本实施例一种可选方案中,为了保证得到的初始实体特征的准确度,信息抽取系统可以首先利用上述的预训练文档模型对语义实体进行编码,得到上述的编码向量,然后再利用两个不同的线性层对编码向量进行处理,即将上述的编码向量输入两个不同的线性层中,生成该语义实体的初始键特征和初始值特征,然后根据该初始键特征和初始值特征,便可以组合得到上述的初始实体特征。在本实施例一种可选方案中,上述预训练文档模型在特征提取组件中的形式为即插即用,用户可以根据需求随时选择任意的预训练文档模型,来完成对语义实体进行编码的目的,不会给确定目标实体特征的过程产生额外的计算负担。

[0068] 在本申请实施例中,关系生成组件包括:关系确定模块和关系生成模块,利用关系生成组件对初始实体特征进行映射,得到多个语义实体的关系特征,包括:利用关系确定模块基于初始实体特征,确定语义实体之间的语义关系;利用关系生成模块基于语义关系和初始实体特征,生成关系特征。

[0069] 在本实施例一种可选方案中,关系生成组件可以至少包括:关系确定模块和关系生成模块,其中,关系确定模块可以用于确定不同语义实体之间的语义关系,可以是指分类器,关系生成模块可以用于生成不同语义实体之间的关系特征,可以是指前向网络。

[0070] 在利用关系生成组件对初始实体特征进行映射时,可以首先利用该上述的关系确定模块,对不同语义实体的初始实体特征进行映射处理,得到不同语义实体之间的语义关系,然后再利用上述的关系生成模块,根据得到的语义关系和上述的初始实体特征,来生成上述的关系特征。

[0071] 在本申请实施例中,利用关系确定模块基于初始实体特征,确定语义实体之间的语义关系,包括:从第一语义实体的初始实体特征中读取第一语义实体的初始键特征,并从第二语义实体的初始实体特征中读取第二语义实体的初始值特征,其中,第一语义实体和第二语义实体用于表征多个语义实体中的任意两个语义实体;基于初始键特征和初始值特征之间的相似度,确定语义关系。

[0072] 在本实施例一种可选方案中,如前文所示,考虑到不同语义实体的初始键特征和

初始值特征之间的匹配程度，可以确定出不同语义实体之间是否存在关联关系，例如上述的语义关系，因此，在利用关系确定模块对初始语义实体进行处理时，可以首先从多个语义实体中筛选出任意的两个语义实体，即上述的第一语义实体和第二语义实体，然后从该第一语义实体对应的初始实体特征中读取该第一语义实体的初始键特征，并从该第二语义实体对应的初始实体特征中读取该第二语义实体的初始值特征，然后将该初始键特征和初始值特征进行匹配，确定两者之间的相似度，并根据该相似度确定第一语义实体和第二语义实体之间的语义关系。

[0073] 在本申请实施例中，信息挖掘组件包括：局部注意力层、全局交互层、以及与局部注意力层和全局交互层连接的池化层，利用信息挖掘组件对关系特征和结构关系进行注意力处理，得到全局结构信息，包括：利用局部注意力层对关系特征进行自注意力处理，得到第一注意力特征；利用全局交互层对关系特征和结构关系进行交叉注意力处理，得到第二注意力特征；对第一注意力特征和第二注意力特征进行叠加，得到叠加特征；对叠加特征进行池化处理，得到全局结构信息。

[0074] 在本实施例一种可选方案中，信息挖掘组件可以至少包括：局部注意力层、全局交互层和池化层，其中，池化层预局部注意力层和全局交互层连接。在利用信息挖掘组件对关系特征和结构关系进行注意力处理时，可以首先利用局部注意力层对关系特征进行自注意力处理，得到上述的第一注意力特征，同时利用全局交互层对关系特征和结构关系进行较差注意力处理，得到上述的第二注意力特征，然后将该第一注意力特征和第二注意力特征进行叠加处理，便可以得到一个叠加特征，最后将该叠加特征输入至池化层中进行处理，便可以得到上述的全局结构信息。

[0075] 在本申请实施例中，将初始实体特征和全局结构信息进行融合，得到目标实体特征，包括：基于门控机制将初始实体特征和全局结构信息进行融合，得到目标实体特征。

[0076] 在本实施例一种可选方案中，在得到上述的初始实体特征和全局结构信息之后，便可以将该初始实体特征和全局结构信息进行融合处理，以得到上述的目标实体特征。为了保证得到的目标实体特征的准确度，可以利用门控机制进行上述的融合过程。

[0077] 在本实施例一种可选方案中，如前文所示，考虑到根据用户在撰写文档时，可能会出现不同语义实体的撰写结果与用户的撰写习惯不符的情况，例如对于一些数据有指定的撰写格式，而该撰写格式可能与用户的撰写习惯不符，此时会出现获取到的语义实体的结构关系不准确的情况，导致确定出的全局结构信息不准确，进而导致构建出的目标实体特征可能存在误差，因此，在得到目标实体特征之后，信息抽取系统还可以进一步地将该目标实体特征重新输入至前述关系生成组件中，来重新确定上述的关系特征，然后再根据新的关系特征和结构关系，生成新的目标实体特征。按照该过程重复迭代多次后，通过提高生成的关系特征的精度来减小因结构关系不准确带来的误差，便可以得到一个误差较小的目标实体特征，从而提高得到的目标实体特征的准确度。

[0078] 在本申请实施例中，在文档解析模型包括：多个关系生成组件和多个信息挖掘组件的情况下，多个关系生成组件的数量与多个信息挖掘组件的数量相同，且多个关系生成组件和多个信息挖掘组件依次迭代运行。

[0079] 在本实施例一种可选方案中，为了保证利用文档解析模型得到目标实体特征的准确度，除了将生成的目标实体特征重新输入至关系生成组件外，还可以在文档解析模型中

设置多个关系生成组件和多个信息挖掘组件,并保证关系生成组件和信息挖掘组件的数量相同,两者的迭代顺序也相同,即先利用关系生成组件,再利用信息挖掘组件,来不断的挖掘语义实体的全局结构信息,并逐步增强语义实体的局部实体表示,从而提高构建出的目标实体特征的准确度。

[0080] 在本申请实施例中,该方法还包括:对文档进行光学识别,确定语义实体所属边界框在文档中的空间信息;基于空间信息,确定语义实体之间的距离和角度;基于距离和角度,生成结构关系。

[0081] 在本实施例一种可选方案中,在确定语义实体的结构关系时,信息抽取系统可以首先对文档进行光学识别,利用OCR技术(Optical Character Recognition,光学字符识别技术)确定文档中的语义实体所属边界框在文档中的空间信息,例如边界框的顶点在文档中的位置、边界框的面积大小等信息,然后再根据该空间信息,确定出不同语义实体之间的距离和角度,根据该距离和角度,便可以确定出不同语义实体之间的结构关系。其中,距离可以是指欧氏距离。

[0082] 在本申请实施例中,空间信息包括:多个端点的子空间信息,基于空间信息,确定语义实体之间的距离和角度,包括:基于多个端点中相同端点的子空间信息,确定语义实体之间的多个子距离和多个子角度;对多个子距离进行拼接,得到语义实体之间的距离;对多个子角度进行拼接,得到语义实体之间的角度。

[0083] 在本实施例一种可选方案中,语义实体所属边界框在文档中的空间信息可以至少包括:边界框上多个端点的子空间信息,例如端点坐标、端点数量、端点在边界框中的位置等信息。在确定语义实体之间的距离和角度时,可以首先不同语义实体所属边界框的多个端点中相同端点,例如在边界框中的标号相同、位置相同的端点,确定出语义实体之间的多个子距离和多个子角度,然后对多个子距离进行拼接,可以得到不同语义实体之间的距离,对多个子角度进行拼接,可以得到不同语义实体之间的角度。

[0084] 在本实施例一种可选方案中,上述多个端点可以是指边界框的坐上端点、右下端点和中心点。

[0085] 在本申请实施例中,该方法还包括:在交互界面中输出文档和信息抽取结果;响应作用于交互界面中的修改指令,确定修改指令对应的目标抽取结果;基于文档和目标抽取结果,对文档处理模型的模型参数进行更新。

[0086] 在本实施例一种可选方案中,为便于用户查看对文档进行信息抽取的过程和结果,信息抽取系统还可以在预设的交互界面中输出该文档和对应的信息抽取结果,用户可以根据该信息抽取结果确定当前的文档处理模型是否满足信息抽取需求,若不满足,则用户可以根据信息抽取结果对文档处理模型进行调整,即用户可以在交互界面中输入对信息抽取结果进行修改的修改指令,信息抽取系统可以根据该修改指令对信息抽取结果进行调整,得到与文档所包含的内容更相符的目标抽取结果,然后利用该文档和目标抽取结果对文档处理模型的模型参数进行更新,以提供文档处理模型的精度。

[0087] 在本申请实施例中,基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果,包括:利用文档识别模块对目标实体特征进行识别,得到语义实体的识别结果集合;利用归一化层确定识别结果集合包含的识别结果的概率;基于识别结果的概率,从识别结果集合中确定信息抽取结果。

[0088] 上述文档识别模块可以是指的分类器。

[0089] 在本实施例一种可选方案中,在根据目标实体特征对文档进行信息抽取时,可以首先利用上述的文档识别模块对目标实体特征进行识别,从文档中确定出该语义实体对应的识别结果集合,然后利用归一化层对该识别结果集合进行归一化处理,确定识别结果集合包含识别结果的概率,最后根据该识别结果的概率,可以从识别结果集合中确定出上述的信息抽取结果。

[0090] 为了便于理解上述过程,图4是根据本申请实施例1的一种文档的信息抽取过程的示意图,如图4所示,可以将待抽取的文档输入至文档处理模型中,经过文档处理模型处理后可以得到一个目标实体特征,信息抽取系统可以根据该目标实体特征对文档进行信息抽取。文档处理模型可以分为三个组件,其中,组件一代表的是特征提取组件,组件二代表的是关系生成组件,组件三代表的是信息挖掘组件。在组件一中至少包括一个预训练文档模型A和两个不同的线性层,线性层A和线性层B,在组件一中可以利用预训练文档模型A对文档中的语义实体进行编码处理,得到编码向量a,然后将该编码向量a输入至上述两个线性层中进行处理,得到语义实体的初始键特征b和初始值特征c,根据该初始键特征b和初始值特征c可以生成语义实体的初始实体特征H。在组件二中至少包括一个关系确定模块B和关系生成模块C,在组件二中可以利用关系确定模块对不同语义实体对应的初始键特征和初始值特征进行处理,确定不同语义实体之间的语义关系S,例如将语义实体1的初始键特征b1,和语义实体2的初始值特征c2进行匹配,确定两者的相似度,并根据该相似度确定语义关系S。然后再利用关系生成模块对该语义关系S和语义实体的初始实体特征H进行处理,以生成语义实体对应的关系特征R(t)。

[0091] 其中,关系特征R(t)的公式可以表示为:

[0092] $$R^{(t)} = \mathrm{softmax}(R^{(t)}W_q[SW_k^s; R^{(t)}W_k]^T)[SW_v^s; R^{(t)}W_v]^T。$$

[0093] 在组件三中至少包括局部注意力层D、全局交互层E和池化层F,在组件三中可以首先利用局部注意力层D对关系特征R(t)进行自注意力处理,以得到一个第一注意力特征d,同时利用全局交互层E对关系特征R(t)和语义实体的结构关系e进行交叉注意力处理,得到第二注意力特征f,然后可以对第一注意力特征d和第二注意力特征f进行叠加处理,得到叠加特征m,并利用池化层F对该叠加特征m进行池化处理,便可以得到语义实体的全局结构信息。

[0094] 最后按照门控机制对前述的初始实体特征和该全局结构信息进行融合,便可以得到目标实体特征H’。

[0095] 其中,目标实体特征H’的公式可以表示为:

[0096] $$g = \mathrm{sigmoid}(W_g\left[H_{k/v}^{(t)}; G_{k/v}^{(t)}\right] + b_g);$$

[0097] $$H_{k/v}^{(t+1)} = H_{k/v}^{(t)} + g \cdot G_{k/v}^{(t)}。$$

[0098] 其中,g为门控机制对应的控制参数,$H_{k/v}^{(t+1)}$代表的是融合后的目标实体特征H’,$H_{k/v}^{(t)}$代表的是初始实体特征H,$G_{k/v}^{(t)}$代表的是上述池化层F输出的全局结构信息。

[0099] 如图4所示,为了提高确定出的目标实体特征的准确度,还可以将上述初始确定出的目标实体特征作为输入,重新利用组件二确定新的关系特征,然后再重复上述过程,对目

标实体特征进行多次迭代处理。

[0100] 在得到目标实体特征之后,可以进一步地利用组件二中的分类器对目标实体特征进行处理,从文档中确定出一个识别结果集合,并利用归一化层对应的softmax函数从识别结果集合中确定出信息抽取结果。

[0101] 图5是根据本申请实施例1的一种语义实体的结构关系确定过程的示意图,如图5所示,在确定语义实体之间的结构关系时,可以首先对语义实体1和语义实体2进行OCR识别,确定语义实体所属边界框的空间信息,然后从该空间信息中确定出边界框上多个端点的子空间信息,根据该子空间信息,可以确定多个端点之间的多个距离和多个角度,对这多个距离进行拼接,可以得到语义实体之间的目标距离,对着多个角度进行拼接,可以得到语义实体之间的目标角度,通过该目标距离和目标角度,便可以确定出语义实体之间的结构关系。

[0102] 以语义实体1边界框上端点的坐标为(x_i,y_i),语义实体2边界框上端点的坐标为(x_j,y_j)为例,语义实体1和语义实体2之间的目标距离的公式可以是:

[0103] $$d(i,j)=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}。$$

[0104] 语义实体1和语义实体2之间的目标角度的公式可以是:

[0105] $$\theta(i,j)=\arctan\frac{y_j-y_i}{x_j-x_i}。$$

[0106] 对应的结构关系的公式可以表示为:

[0107] $g_{i,j}=[W_\theta\theta(i,j);W_d d(i,j)]$。

[0108] 需要说明的是,对于前述的各方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本申请并不受所描述的动作顺序的限制,因为依据本申请,某些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作和模块并不一定是本申请所必须的。

[0109] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中,包括若干指令用以使得一台终端设备(可以是手机,计算机,服务器,或者网络设备等)执行本申请各个实施例的方法。

[0110] 实施例2

[0111] 根据本申请实施例,还提供了一种文档的信息抽取方法,需要说明的是,在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行,并且,虽然在流程图中示出了逻辑顺序,但是在某些情况下,可以以不同于此处的顺序执行所示出或描述的步骤。

[0112] 图6是根据本申请实施例2的一种文档的信息抽取方法的流程图,如图6所示,该方法可以包括如下步骤:

[0113] 步骤S602,响应交互界面中接收到的文档,其中,文档包含多个语义实体。

[0114] 在本实施例一种可选方案中,信息抽取系统可以首先获取需要用户理解的文档,该文档中包含多个语义实体,可以是视觉文档。

[0115] 步骤S604,基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征。

[0116] 在本实施例一种可选方案中,信息抽取系统可以在对文档进行信息抽取的过程中引入文档中的全局结构信息,将文档中的局部实体表示和全局结构信息相结合,即根据不同语义实体各自的语义信息,以及不同语义实体之间的结构关系,构建出上述的目标实体特征,并以该目标实体特征为基础对文档进行信息抽取。

[0117] 步骤S606,基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果。

[0118] 在构建出文档中语义实体对应的目标实体特征之后,信息抽取系统便可以根据该目标实体特征对文档进行信息抽取,从而保证抽取得到的信息抽取结果的准确度。

[0119] 步骤S608,在交互界面中显示信息抽取结果。

[0120] 在抽取得到文档中的关键信息,即得到信息抽取结果之后,信息抽取系统可以在预设的交互界面中显示该信息抽取结果,以方便用户查阅。

[0121] 需要说明的是,本申请上述实施例中涉及到的优选实施方案与实施例1提供的方案以及应用场景、实施过程相同,但不仅限于实施例1所提供的方案。

[0122] 实施例3

[0123] 根据本申请实施例,还提供了文档的信息抽取方法,需要说明的是,在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行,并且,虽然在流程图中示出了逻辑顺序,但是在某些情况下,可以以不同于此处的顺序执行所示出或描述的步骤。

[0124] 图7是根据本申请实施例3的文档的信息抽取方法的流程图,如图7所示,该方法可以包括如下步骤:

[0125] 步骤S702,通过调用第一接口获取文档。

[0126] 其中,第一接口包括第一参数,第一参数的参数值为文档,文档包含多个语义实体。

[0127] 在本实施例一种可选方案中,在对文档进行信息抽取时,信息抽取系统可以利用上述的第一接口调用第一参数,即调用获取到的文档,该文档中包含多个语义实体,可以是视觉文档。

[0128] 步骤S704,基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征。

[0129] 在本实施例一种可选方案中,信息抽取系统可以在对文档进行信息抽取的过程中引入文档中的全局结构信息,将文档中的局部实体表示和全局结构信息相结合,即根据不同语义实体各自的语义信息,以及不同语义实体之间的结构关系,构建出上述的目标实体特征,并以该目标实体特征为基础对文档进行信息抽取。

[0130] 步骤S706,基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果。

[0131] 在构建出文档中语义实体对应的目标实体特征之后,信息抽取系统便可以根据该目标实体特征对文档进行信息抽取,从而保证抽取得到的信息抽取结果的准确度。

[0132] 步骤S708,通过调用第二接口输出信息抽取结果。

[0133] 其中,第二接口包括第二参数,第二参数的参数值为信息抽取结果。

[0134] 在本实施例一种可选方案中,在得到信息抽取结果之后,信息抽取系统可以通过调用第二接口的第二参数,即上述的信息抽取结果,用户可以根据需求查看调用出的信息抽取结果。

[0135] 在本实施例的一种可选方案中,可以通过调用第二接口输出确定出的风险指标,以使工作人员能够更直观的监测区域对应的风险指标。

[0136] 实施例4

[0137] 根据本申请实施例,还提供了一种用于实施上述文档的信息抽取方法的装置,该装置可以部署在目标客户端中。图8是根据本申请实施例4的一种文档的信息抽取装置的结构框图,如图8所示,该装置800包括:文档获取模块802,特征构建模块804,信息抽取模块806。

[0138] 其中,文档获取模块802用于获取文档,其中,文档包含多个语义实体;特征构建模块804用于基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;信息抽取模块806用于基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果。

[0139] 在本申请上述实施例中,特征构建模块804包括:特征构建单元,用于利用文档处理模型基于语义信息和结构关系,构建目标实体特征。

[0140] 在本申请上述实施例中,文档解析模型包括:特征提取组件、关系生成组件和信息挖掘组件,特征构建单元还用于:利用特征提取组件对语义实体进行特征提取,得到语义实体的初始实体特征,其中,初始实体特征用于表征语义实体的语义信息;利用关系生成组件对初始实体特征进行映射,得到多个语义实体的关系特征,其中,关系特征用于表征多个语义实体之间的语义关系;利用信息挖掘组件对关系特征和结构关系进行注意力处理,得到全局结构信息;将初始实体特征和全局结构信息进行融合,得到目标实体特征。

[0141] 在本申请上述实施例中,特征提取组件包括:预训练文档模型和两个不同的线性层,特征构建单元还用于:利用预训练文档模型对语义实体进行编码,得到语义实体的编码向量;利用两个不同的线性层分别对编码向量进行处理,生成语义实体的初始键特征和初始值特征,其中,初始键特征用于表征语义实体中位于头部的实体,初始值特征用于表征语义实体中位于尾部的实体;基于初始键特征和初始值特征,得到初始实体特征。

[0142] 在本申请上述实施例中,关系生成组件包括:关系确定模块和关系生成模块,特征构建单元还用于:利用关系确定模块基于初始实体特征,确定语义实体之间的语义关系;利用关系生成模块基于语义关系和初始实体特征,生成关系特征。

[0143] 在本申请上述实施例中,信息挖掘组件包括:局部注意力层、全局交互层、以及与局部注意力层和全局交互层连接的池化层,特征构建单元还用于:利用局部注意力层对关系特征进行自注意力处理,得到第一注意力特征;利用全局交互层对关系特征和结构关系进行交叉注意力处理,得到第二注意力特征;对第一注意力特征和第二注意力特征进行叠加,得到叠加特征;对叠加特征进行池化处理,得到全局结构信息。

[0144] 在本申请上述实施例中,特征构建单元还用于:基于门控机制将初始实体特征和全局结构信息进行融合,得到目标实体特征。

[0145] 在本申请上述实施例中,在文档解析模型包括:多个关系生成组件和多个信息挖

掘组件的情况下,多个关系生成组件的数量与多个信息挖掘组件的数量相同,且多个关系生成组件和多个信息挖掘组件依次迭代运行。

[0146] 在本申请上述实施例中,该装置还包括:信息识别模块,用于对文档进行光学识别,确定语义实体所属边界框在文档中的空间信息;参数确定模块,用于基于空间信息,确定语义实体之间的距离和角度;关系生成模块,用于基于距离和角度,生成结构关系。

[0147] 在本申请上述实施例中,空间信息包括:多个端点的子空间信息,参数确定模块包括:子参数确定单元,用于基于多个端点中相同端点的子空间信息,确定语义实体之间的多个子距离和多个子角度;距离确定单元,用于对多个子距离进行拼接,得到语义实体之间的距离;角度确定单元,用于对多个子角度进行拼接,得到语义实体之间的角度。

[0148] 在本申请上述实施例中,该装置还包括:信息输出模块,用于在交互界面中输出文档和信息抽取结果;结构修改模块,用于响应作用于交互界面中的修改指令,确定修改指令对应的目标抽取结果;参数更新模块,用于基于文档和目标抽取结果,对文档处理模型的模型参数进行更新。

[0149] 在本申请上述实施例中,信息抽取模块806包括:特征识别单元,用于利用文档识别模块对目标实体特征进行识别,得到语义实体的识别结果集合;概率确定单元,用于利用归一化层确定识别结果集合包含的识别结果的概率;结果确定单元,用于基于识别结果的概率,从识别结果集合中确定信息抽取结果。

[0150] 此处需要说明的是,上述的文档获取模块802,特征构建模块804,信息抽取模块806,对应于实施例1中的步骤S302至步骤S306,三个模块与对应的步骤所实现的实例和应用场景相同,但不限于上述实施例1所公开的内容。需要说明的是,上述模块或单元可以是存储在存储器中并由一个或多个处理器处理的硬件组件或软件组件,上述模块也可以作为装置的一部分可以运行在实施例1提供的AR/VR设备中。

[0151] 需要说明的是,本申请上述实施例中涉及到的优选实施方案与实施例1提供的方案以及应用场景、实施过程相同,但不仅限于实施例1所提供的方案。

[0152] 实施例5

[0153] 根据本申请实施例,还提供了一种用于实施上述文档的信息抽取的装置,该装置可以部署在目标客户端中。图9是根据本申请实施例5的一种文档的信息抽取装置的结构框图,如图9所示,该装置900包括:文档接收模块902,特征构建模块904,信息抽取模块906和结果显示模块908。

[0154] 其中,文档接收模块902用于响应交互界面中接收到的文档,其中,文档包含多个语义实体;特征构建模块904用于基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;信息抽取模块906用于基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果;结果显示模块908用于在交互界面中显示信息抽取结果。

[0155] 此处需要说明的是,上述的文档接收模块902,特征构建模块904,信息抽取模块906和结果显示模块908,对应于实施例2中的步骤S602至步骤S608,四个模块与对应的步骤所实现的实例和应用场景相同,但不限于上述实施例1所公开的内容。需要说明的是,上述模块或单元可以是存储在存储器中并由一个或多个处理器处理的硬件组件或软件组件,上述模块也可以作为装置的一部分可以运行在实施例1提供的AR/VR设备中。

[0156] 需要说明的是,本申请上述实施例中涉及到的优选实施方案与实施例1提供的方

案以及应用场景、实施过程相同,但不仅限于实施例1所提供的方案。

[0157] 实施例6

[0158] 根据本申请实施例,还提供了一种用于实施上述文档的信息抽取方法的装置,该装置可以部署在目标客户端中。图10是根据本申请实施例6的一种文档的信息抽取装置的结构框图,如图10所示,该装置1000包括:第一调用模块1002、特征构建模块1004,信息抽取模块1006和第二调用模块1008。

[0159] 其中,第一调用模块1002用于通过调用第一接口获取文档,其中,第一接口包括第一参数,第一参数的参数值为文档,文档包含多个语义实体;特征构建模块1004用于基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;信息抽取模块1006用于基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果;第二调用模块1008用于通过调用第二接口输出信息抽取结果,其中,第二接口包括第二参数,第二参数的参数值为信息抽取结果。

[0160] 此处需要说明的是,上述的第一调用模块1002、特征构建模块1004,信息抽取模块1006和第二调用模块1008,对应于实施例3中的步骤S702至步骤S708,四个模块与对应的步骤所实现的实例和应用场景相同,但不限于上述实施例1所公开的内容。需要说明的是,上述模块或单元可以是存储在存储器中并由一个或多个处理器处理的硬件组件或软件组件,上述模块也可以作为装置的一部分可以运行在实施例1提供的AR/VR设备中。

[0161] 需要说明的是,本申请上述实施例中涉及到的优选实施方案与实施例1提供的方案以及应用场景、实施过程相同,但不仅限于实施例1所提供的方案。

[0162] 实施例7

[0163] 根据本申请实施例,还提供了一种用于实施上述文档的信息抽取方法的系统,该系统可以部署在目标客户端中。图11是根据本申请实施例7的一种文档的信息抽取系统的结构框图,如图11所示,该系统1100包括:接收模块1102、处理装置1104和发送模块1106。

[0164] 其中,接收模块1102用于显示交互界面,并捕获在交互界面中输入的文档,其中,文档包含多个语义实体;处理装置1104用于基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征,并基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果;发送模块1106,用于将所述信息抽取结果发送至所述前端客户端,其中,所述信息抽取结果由所述前端客户端显示在所述交互界面中。

[0165] 需要说明的是,本申请上述实施例中涉及到的优选实施方案与实施例1提供的方案以及应用场景、实施过程相同,但不仅限于实施例1所提供的方案。

[0166] 实施例8

[0167] 本申请的实施例可以提供一种电子设备,该电子设备可以是电子设备群中的任意一个电子设备。可选地,在本实施例中,上述电子设备也可以替换为移动终端等终端设备。

[0168] 可选地,在本实施例中,上述电子设备可以位于计算机网格的多个网格设备中的至少一个网格设备。

[0169] 在本实施例中,上述电子设备可以执行文档的信息抽取方法中以下步骤的程序代码:获取文档,其中,文档包含多个语义实体;基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果。

[0170] 可选地,图12是根据本申请实施例8的一种电子设备的结构框图。如图所示,该电子设备A可以包括:处理器1202和存储器1204,其中,存储有可执行程序;处理器,用于运行程序,其中,程序运行时执行实施例1中所示的文档的信息抽取方法。

[0171] 可选地,如图12所示,在电子设备A中还可以包括:存储控制器和外设接口,其中,外设接口与射频模块、音频模块和显示器连接。

[0172] 其中,存储器可用于存储软件程序以及模块,如本申请实施例中的图像处理方法和装置对应的程序指令/模块,处理器通过运行存储在存储器内的软件程序以及模块,从而执行各种功能应用以及数据处理,即实现上述的图像分割方法。存储器可包括高速随机存储器,还可以包括非易失性存储器,如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中,存储器可进一步包括相对于处理器远程设置的存储器,这些远程存储器可以通过网格连接至终端A。上述网格的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0173] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征,包括:利用文档处理模型基于语义信息和结构关系,构建目标实体特征。

[0174] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:文档解析模型包括:特征提取组件、关系生成组件和信息挖掘组件,利用文档处理模型基于语义信息和结构关系,构建目标实体特征,包括:利用特征提取组件对语义实体进行特征提取,得到语义实体的初始实体特征,其中,初始实体特征用于表征语义实体的语义信息;利用关系生成组件对初始实体特征进行映射,得到多个语义实体的关系特征,其中,关系特征用于表征多个语义实体之间的语义关系;利用信息挖掘组件对关系特征和结构关系进行注意力处理,得到全局结构信息;将初始实体特征和全局结构信息进行融合,得到目标实体特征。

[0175] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:特征提取组件包括:预训练文档模型和两个不同的线性层,利用特征提取组件对语义实体进行特征提取,得到语义实体的初始实体特征,包括:利用预训练文档模型对语义实体进行编码,得到语义实体的编码向量;利用两个不同的线性层分别对编码向量进行处理,生成语义实体的初始键特征和初始值特征,其中,初始键特征用于表征语义实体中位于头部的实体,初始值特征用于表征语义实体中位于尾部的实体;基于初始键特征和初始值特征,得到初始实体特征。

[0176] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:关系生成组件包括:关系确定模块和关系生成模块,利用关系生成组件对初始实体特征进行映射,得到多个语义实体的关系特征,包括:利用关系确定模块基于初始实体特征,确定语义实体之间的语义关系;利用关系生成模块基于语义关系和初始实体特征,生成关系特征。

[0177] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:利用关系确定模块基于初始实体特征,确定语义实体之间的语义关系,包括:从第一语义实体的初始实体特征中读取第一语义实体的初始键特征,并从第二语义实体的初始实体特征中读取第二语义实体的初始值特征,其中,第一语义实体和第二语义实体用于表征多个语

义实体中的任意两个语义实体;基于初始键特征和初始值特征之间的相似度,确定语义关系。

[0178] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:信息挖掘组件包括:局部注意力层、全局交互层、以及与局部注意力层和全局交互层连接的池化层,利用信息挖掘组件对关系特征和结构关系进行注意力处理,得到全局结构信息,包括:利用局部注意力层对关系特征进行自注意力处理,得到第一注意力特征;利用全局交互层对关系特征和结构关系进行交叉注意力处理,得到第二注意力特征;对第一注意力特征和第二注意力特征进行叠加,得到叠加特征;对叠加特征进行池化处理,得到全局结构信息。

[0179] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:将初始实体特征和全局结构信息进行融合,得到目标实体特征,包括:基于门控机制将初始实体特征和全局结构信息进行融合,得到目标实体特征。

[0180] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:在文档解析模型包括:多个关系生成组件和多个信息挖掘组件的情况下,多个关系生成组件的数量与多个信息挖掘组件的数量相同,且多个关系生成组件和多个信息挖掘组件依次迭代运行。

[0181] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:该方法还包括:对文档进行光学识别,确定语义实体所属边界框在文档中的空间信息;基于空间信息,确定语义实体之间的距离和角度;基于距离和角度,生成结构关系。

[0182] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:空间信息包括:多个端点的子空间信息,基于空间信息,确定语义实体之间的距离和角度,包括:基于多个端点中相同端点的子空间信息,确定语义实体之间的多个子距离和多个子角度;对多个子距离进行拼接,得到语义实体之间的距离;对多个子角度进行拼接,得到语义实体之间的角度。

[0183] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:该方法还包括:在交互界面中输出文档和信息抽取结果;响应作用于交互界面中的修改指令,确定修改指令对应的目标抽取结果;基于文档和目标抽取结果,对文档处理模型的模型参数进行更新。

[0184] 处理器可以通过传输装置调用存储器存储的信息及应用程序,以执行下述步骤:基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果,包括:利用文档识别模块对目标实体特征进行识别,得到语义实体的识别结果集合;利用归一化层确定识别结果集合包含的识别结果的概率;基于识别结果的概率,从识别结果集合中确定信息抽取结果。

[0185] 在本申请实施例中,采用获取文档,其中,文档包含多个语义实体;基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果的方式,通过将语义实体的语义信息作为文档的局部实体表示,并根据语义实体之间的结构关系构建文档的全局结构信息,然后再根据该语义信息和结构信息,构建出包含文档的局部实体表示和全局结构信息的目标实体特征,以保证目标实体特征能够与文档全文相关联,具有较高的鲁棒性,最后再根据该目标实体特征对文档进行信息抽取,可以进一步地保证得到的信息抽取结果与文档的内容相

符,提高信息抽取结果的准确度,从而实现了提高对文档进行信息抽取得到的结果的准确度的技术效果,进而解决了相关技术中对文档的信息抽取效果差的技术问题。

[0186] 本领域普通技术人员可以理解,图12所示的结构仅为示意,计算机终端也可以是智能手机(如Android手机、iOS手机等)、平板电脑、掌上电脑以及移动互联网设备(Mobile Internet Devices,MID)、PAD等终端设备。图12其并不对上述电子装置的结构造成限定。例如,计算机终端A还可包括比图12中所示更多或者更少的组件(如网格接口、显示装置等),或者具有与图12所示不同的配置。

[0187] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令终端设备相关的硬件来完成,该程序可以存储于一计算机可读存储介质中,存储介质可以包括:闪存盘、只读存储器(Read-Only Memory,ROM)、随机存取器(Random Access Memory,RAM)、磁盘或光盘等。

[0188] 实施例10

[0189] 本申请的实施例还提供了一种计算机可读存储介质。可选地,在本实施例中,上述计算机可读存储介质可以用于保存上述实施例1所提供的文档的信息抽取方法所执行的程序代码。

[0190] 在本实施例中,上述计算机可读存储介质可以位于AR/VR设备网络中AR/VR设备终端群中的任意一个计算机终端中,或者位于移动终端群中的任意一个移动终端中。

[0191] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:获取文档,其中,文档包含多个语义实体;基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果。

[0192] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征,包括:利用文档处理模型基于语义信息和结构关系,构建目标实体特征。

[0193] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:文档解析模型包括:特征提取组件、关系生成组件和信息挖掘组件,利用文档处理模型基于语义信息和结构关系,构建目标实体特征,包括:利用特征提取组件对语义实体进行特征提取,得到语义实体的初始实体特征,其中,初始实体特征用于表征语义实体的语义信息;利用关系生成组件对初始实体特征进行映射,得到多个语义实体的关系特征,其中,关系特征用于表征多个语义实体之间的语义关系;利用信息挖掘组件对关系特征和结构关系进行注意力处理,得到全局结构信息;将初始实体特征和全局结构信息进行融合,得到目标实体特征。

[0194] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:特征提取组件包括:预训练文档模型和两个不同的线性层,利用特征提取组件对语义实体进行特征提取,得到语义实体的初始实体特征,包括:利用预训练文档模型对语义实体进行编码,得到语义实体的编码向量;利用两个不同的线性层分别对编码向量进行处理,生成语义实体的初始键特征和初始值特征,其中,初始键特征用于表征语义实体中位于头部的实体,初始值特征用于表征语义实体中位于尾部的实体;基于初始键特征和初始值特征,得到初始实体特征。

[0195] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:关系

生成组件包括:关系确定模块和关系生成模块,利用关系生成组件对初始实体特征进行映射,得到多个语义实体的关系特征,包括:利用关系确定模块基于初始实体特征,确定语义实体之间的语义关系;利用关系生成模块基于语义关系和初始实体特征,生成关系特征。

[0196] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:利用关系确定模块基于初始实体特征,确定语义实体之间的语义关系,包括:从第一语义实体的初始实体特征中读取第一语义实体的初始键特征,并从第二语义实体的初始实体特征中读取第二语义实体的初始值特征,其中,第一语义实体和第二语义实体用于表征多个语义实体中的任意两个语义实体;基于初始键特征和初始值特征之间的相似度,确定语义关系。

[0197] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:信息挖掘组件包括:局部注意力层、全局交互层、以及与局部注意力层和全局交互层连接的池化层,利用信息挖掘组件对关系特征和结构关系进行注意力处理,得到全局结构信息,包括:利用局部注意力层对关系特征进行自注意力处理,得到第一注意力特征;利用全局交互层对关系特征和结构关系进行交叉注意力处理,得到第二注意力特征;对第一注意力特征和第二注意力特征进行叠加,得到叠加特征;对叠加特征进行池化处理,得到全局结构信息。

[0198] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:将初始实体特征和全局结构信息进行融合,得到目标实体特征,包括:基于门控机制将初始实体特征和全局结构信息进行融合,得到目标实体特征。

[0199] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:在文档解析模型包括:多个关系生成组件和多个信息挖掘组件的情况下,多个关系生成组件的数量与多个信息挖掘组件的数量相同,且多个关系生成组件和多个信息挖掘组件依次迭代运行。

[0200] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:该方法还包括:对文档进行光学识别,确定语义实体所属边界框在文档中的空间信息;基于空间信息,确定语义实体之间的距离和角度;基于距离和角度,生成结构关系。

[0201] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:空间信息包括:多个端点的子空间信息,基于空间信息,确定语义实体之间的距离和角度,包括:基于多个端点中相同端点的子空间信息,确定语义实体之间的多个子距离和多个子角度;对多个子距离进行拼接,得到语义实体之间的距离;对多个子角度进行拼接,得到语义实体之间的角度。

[0202] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:该方法还包括:在交互界面中输出文档和信息抽取结果;响应作用于交互界面中的修改指令,确定修改指令对应的目标抽取结果;基于文档和目标抽取结果,对文档处理模型的模型参数进行更新。

[0203] 在本申请实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:基于目标实体特征对文档进行信息抽取,得到文档的信息抽取结果,包括:利用文档识别模块对目标实体特征进行识别,得到语义实体的识别结果集合;利用归一化层确定识别结果集合包含的识别结果的概率;基于识别结果的概率,从识别结果集合中确定信息抽取结果。

[0204] 在本申请实施例中,采用获取文档,其中,文档包含多个语义实体;基于语义实体的语义信息和语义实体之间的结构关系,构建语义实体的目标实体特征;基于目标实体特

征对文档进行信息抽取,得到文档的信息抽取结果的方式,通过将语义实体的语义信息作为文档的局部实体表示,并根据语义实体之间的结构关系构建文档的全局结构信息,然后再根据该语义信息和结构信息,构建出包含文档的局部实体表示和全局结构信息的目标实体特征,以保证目标实体特征能够与文档全文相关联,具有较高的鲁棒性,最后再根据该目标实体特征对文档进行信息抽取,可以进一步地保证得到的信息抽取结果与文档的内容相符,提高信息抽取结果的准确度,从而实现了提高对文档进行信息抽取得到的结果的准确度的技术效果,进而解决了相关技术中对文档的信息抽取效果差的技术问题。

[0205] 上述本申请实施例序号仅仅为了描述,不代表实施例的优劣。

[0206] 在本申请的上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详述的部分,可以参见其他实施例的相关描述。

[0207] 在本申请所提供的几个实施例中,应该理解到,所揭露的技术内容,可通过其它的方式实现。其中,以上所描述的装置实施例仅仅是示意性的,例如单元的划分,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,例如多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口,单元或模块的间接耦合或通信连接,可以是电性或其它的形式。

[0208] 作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元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单元来实现本实施例方案的目的。

[0209] 另外,在本申请各个实施例中的各功能单元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,也可以采用软件功能单元的形式实现。

[0210] 集成的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时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可为个人计算机、服务器或者网络设备等)执行本申请各个实施例方法的全部或部分步骤。而前述的存储介质包括:U盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0211] 以上仅是本申请的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本申请原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本申请的保护范围。

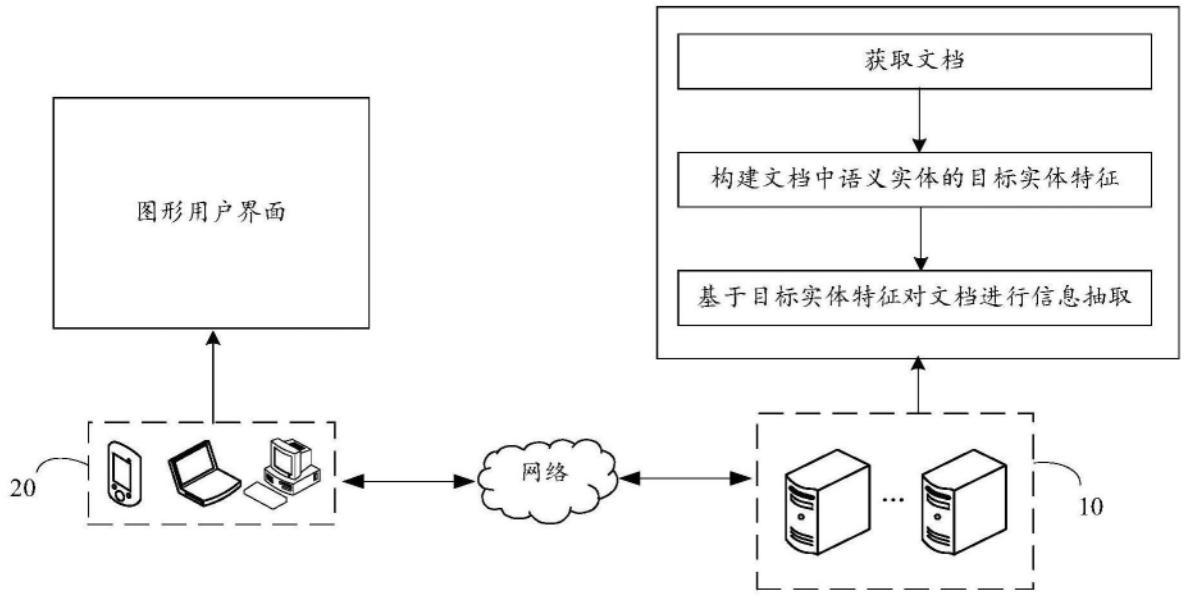

获取文档
构建文档中语义实体的目标实体特征
基于目标实体特征对文档进行信息抽取
图形用户界面
20
网络
10

图1

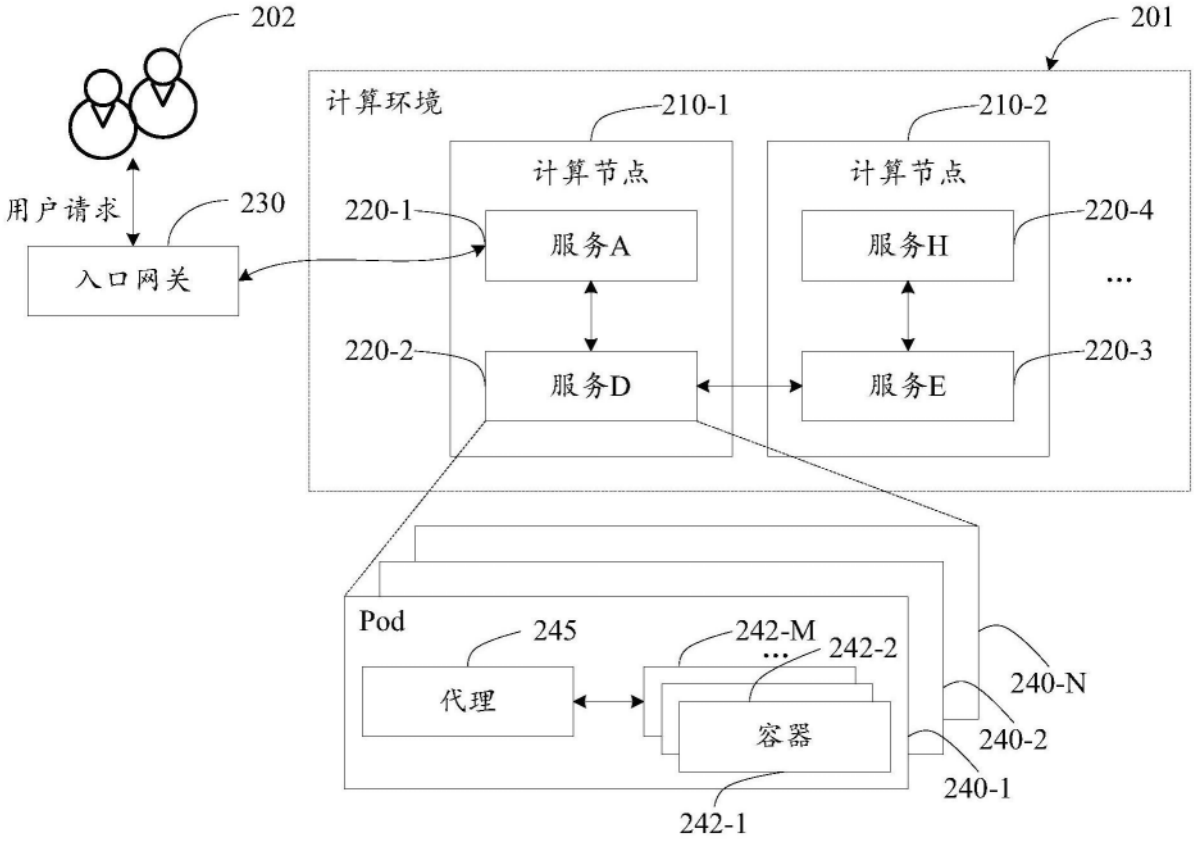

202
201
计算环境
210-1
210-2
计算节点
计算节点
用户请求
230
220-1
服务A
服务H
220-4
入口网关
220-2
服务D
服务E
220-3
Pod
245
242-M
242-2
240-N
代理
容器
240-2
240-1
242-1

图2

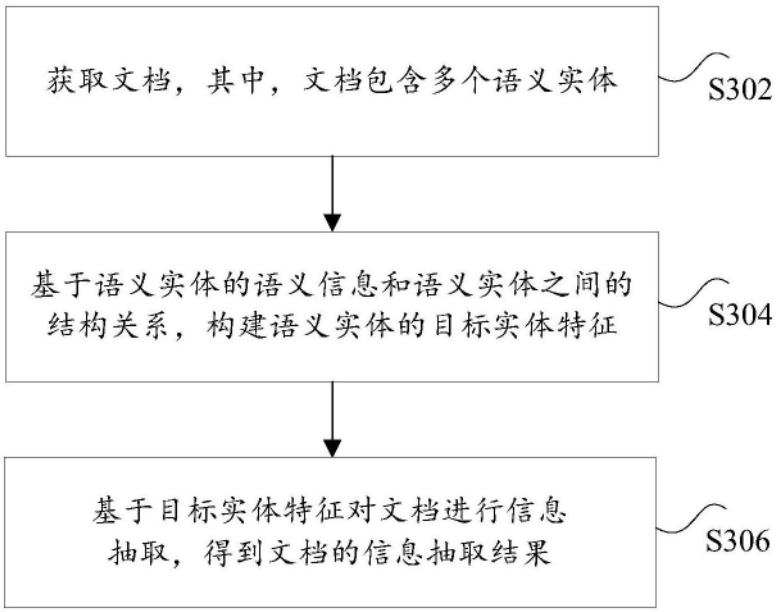

获取文档，其中，文档包含多个语义实体
S302
基于语义实体的语义信息和语义实体之间的结构关系，构建语义实体的目标实体特征
S304
基于目标实体特征对文档进行信息抽取，得到文档的信息抽取结果
S306

图3

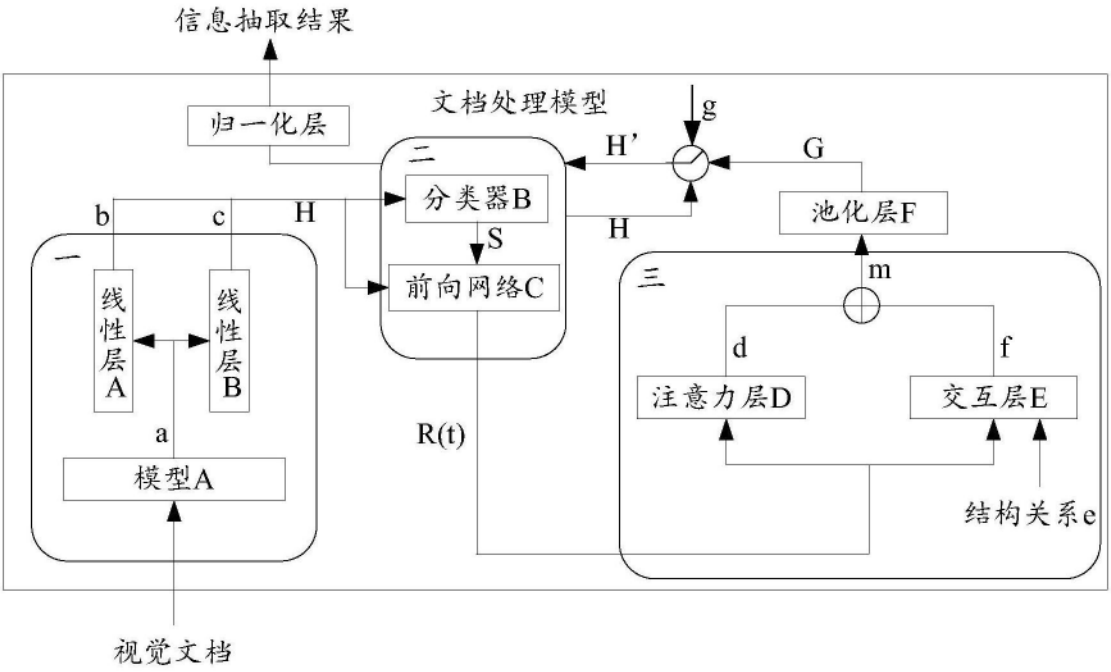

信息抽取结果
归一化层
文档处理模型
g
二
H'
G
分类器B
池化层F
b
c
H
H
S
一
三
m
线性层A
线性层B
前向网络C
d
f
注意力层D
交互层E
a
R(t)
模型A
结构关系e
视觉文档

图4

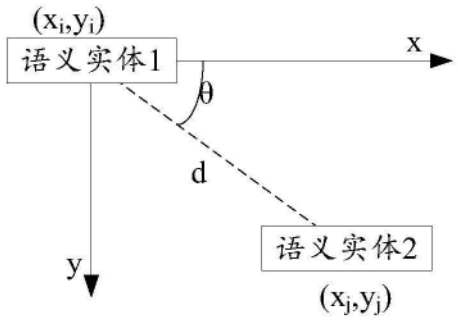

(xi,yi)
语义实体1
x
θ
d
y
语义实体2
(xj,yj)

图5

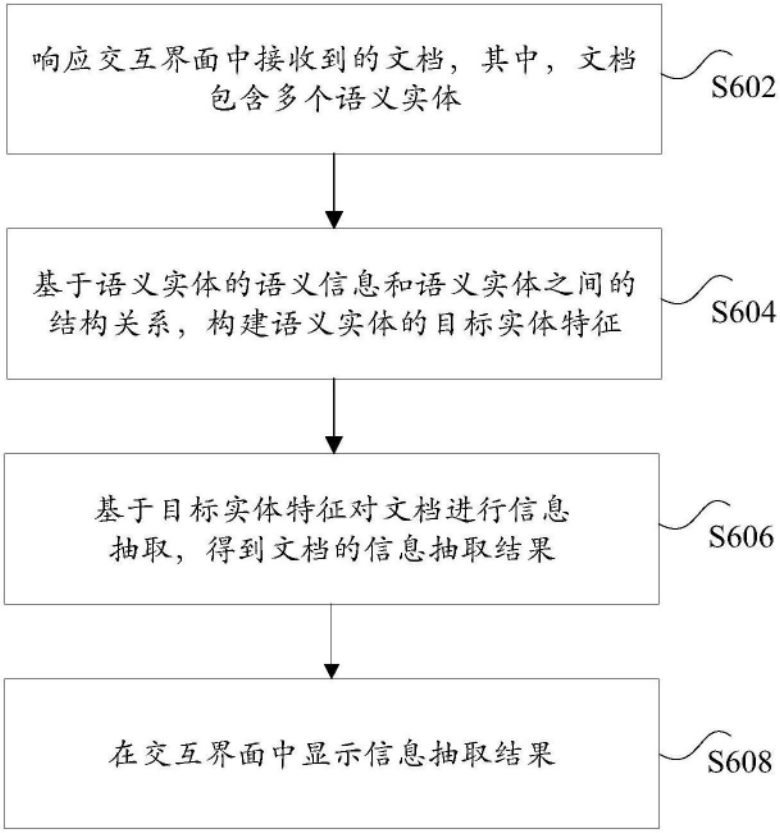

响应交互界面中接收到的文档，其中，文档包含多个语义实体
S602
基于语义实体的语义信息和语义实体之间的结构关系，构建语义实体的目标实体特征
S604
基于目标实体特征对文档进行信息抽取，得到文档的信息抽取结果
S606
在交互界面中显示信息抽取结果
S608

图6

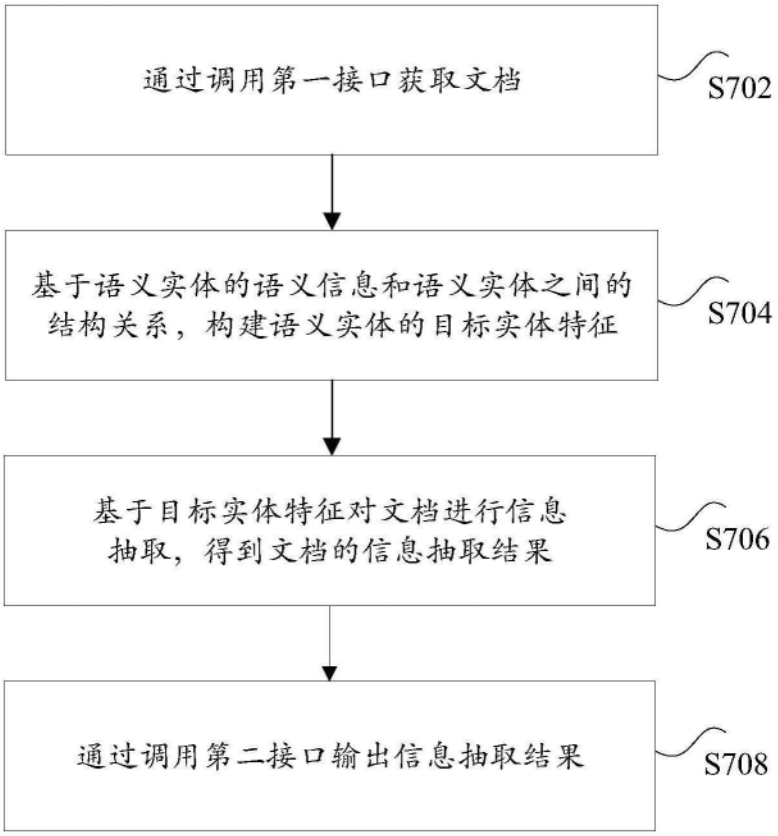

通过调用第一接口获取文档
S702
基于语义实体的语义信息和语义实体之间的结构关系，构建语义实体的目标实体特征
S704
基于目标实体特征对文档进行信息抽取，得到文档的信息抽取结果
S706
通过调用第二接口输出信息抽取结果
S708

图7

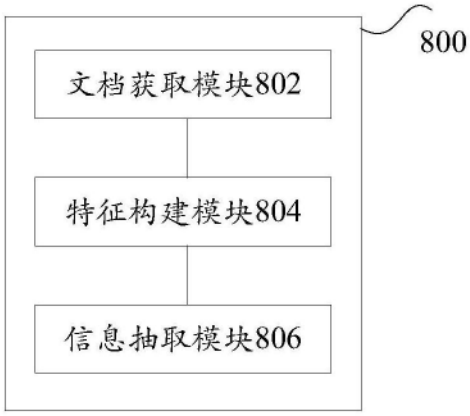

800
文档获取模块802
特征构建模块804
信息抽取模块806

图8

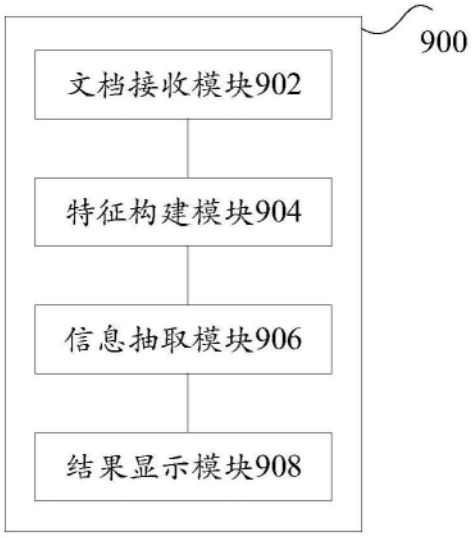

900
文档接收模块902
特征构建模块904
信息抽取模块906
结果显示模块908

图9

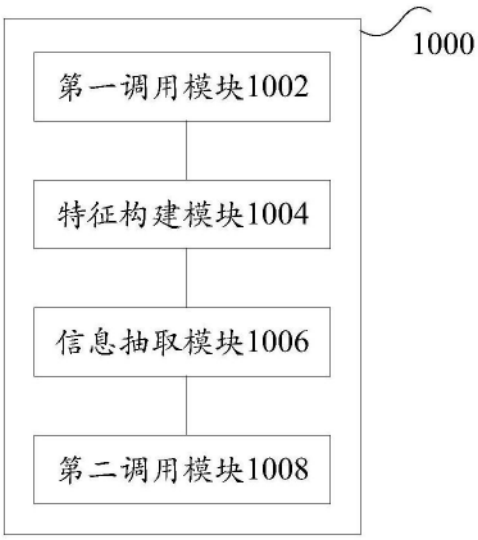

1000
第一调用模块1002
特征构建模块1004
信息抽取模块1006
第二调用模块1008

图10

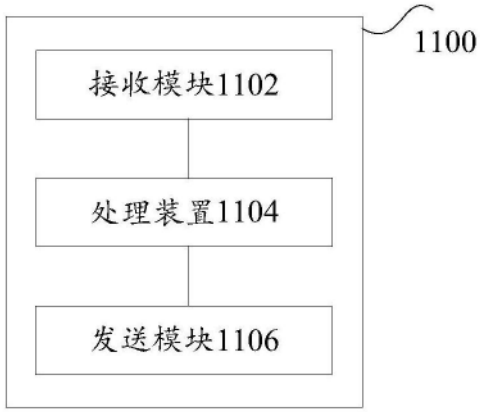

1100
接收模块1102
处理装置1104
发送模块1106

图11

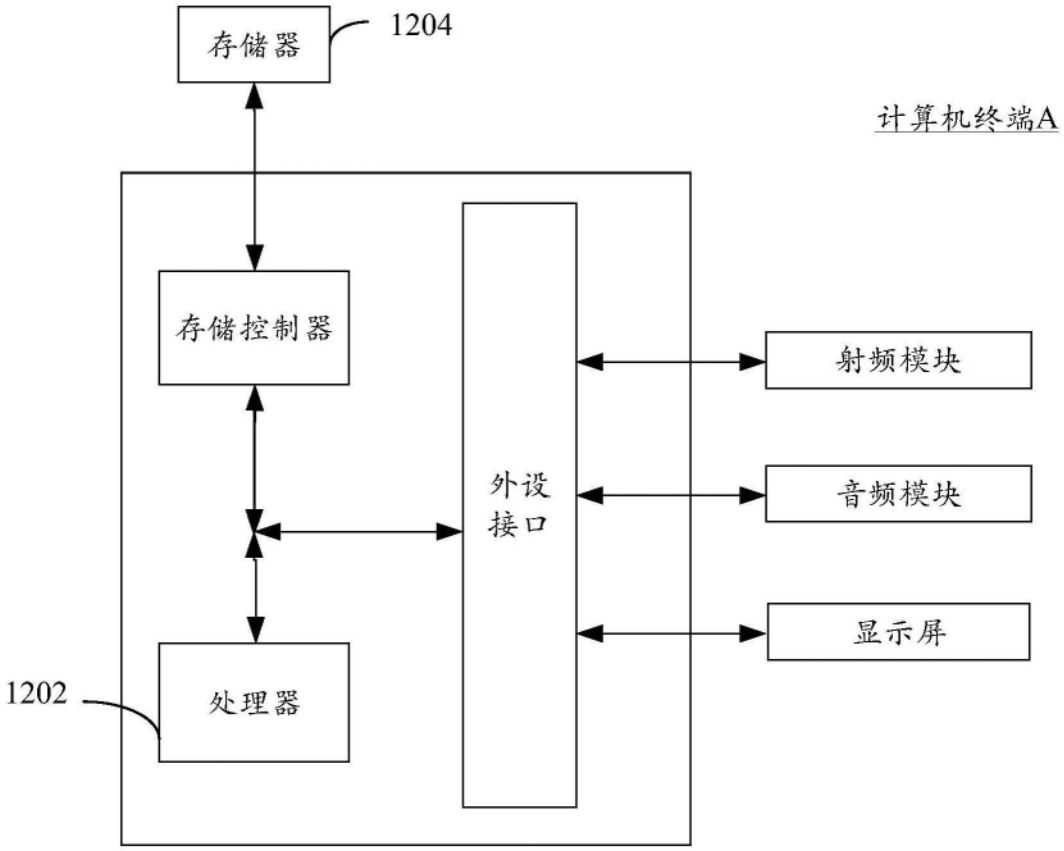

存储器
1204
计算机终端A
存储控制器
外设
接口
射频模块
音频模块
显示屏
1202
处理器

图12