(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117688123 A
(43)申请公布日 2024.03.12

(21)申请号 202211039454.0

(22)申请日 2022.08.29

(71)申请人 华为云计算技术有限公司
地址 550025 贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心

(72)发明人 李泽昌 顾迎捷 段新宇 王喆锋 怀宝兴

(74)专利代理机构 北京龙双利达知识产权代理有限公司 11329
专利代理师 左颖 时林

(51)Int.Cl.
G06F 16/31(2019.01)
G06F 16/33(2019.01)
G06F 40/30(2020.01)
G06N 3/04(2023.01)
G06N 3/08(2023.01)

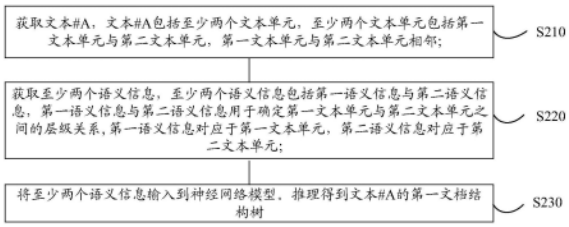
权利要求书2页 说明书11页 附图5页

(54)发明名称

文档结构树的生成方法以及装置

(57)摘要

本申请提供了一种文档结构树的生成方法以及装置,该方法包括:获取文本#A,该文本#A包括至少两个文本单元,至少两个文本单元包括第一文本单元与第二文本单元,第一文本单元与第二文本单元相邻;获取至少两个语义信息,至少两个语义信息包括第一语义信息与第二语义信息,第一语义信息与第二语义信息用于确定第一文本单元与第二文本单元之间的层级关系,第一文本单元对应于第一语义信息,第二文本单元对应于第二语义信息;将至少两个语义信息输入到神经网络模型,推理得到文本#A的第一文档结构树。通过上述方法,本申请实施例可以实现为绝大多数的文档生成相应的文档结构树,且不受文档的版式、字符的属性信息等因素的限制。



CN 117688123 A

1.一种文档结构树的生成方法,其特征在于,包括:

获取文本,所述文本包括至少两个文本单元,所述至少两个文本单元包括第一文本单元与第二文本单元,所述第一文本单元与所述第二文本单元相邻;

获取至少两个语义信息,所述至少两个语义信息包括第一语义信息与第二语义信息,所述第一语义信息与所述第二语义信息用于确定所述第一文本单元与所述第二文本单元之间的层级关系,所述第一文本单元对应于所述第一语义信息,所述第二文本单元对应于所述第二语义信息;

将所述至少两个语义信息输入到神经网络模型,推理得到所述文本的第一文档结构树。

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:

获取第一数据,所述第一数据是用户对所述第一文档结构树进行校验而确定的数据;

根据所述第一数据更新所述神经网络模型。

3.根据权利要求1所述的方法,其特征在于,所述方法还包括:

根据文档结构模板对所述第一文档结构树进行更新,得到所述文本的第二文档结构树。

4.根据权利要求3所述的方法,其特征在于,所述方法还包括:

获取第二数据,所述第二数据是用户对所述第二文档结构树进行校验而确定的数据;

根据所述第二数据更新所述神经网络模型。

5.根据权利要求2或4所述的方法,其特征在于,所述方法还包括:

将校验后的所述第一文档结构树或者所述第二文档结构树存储至文档模板库。

6.根据权利要求1至5中任一项所述的方法,其特征在于,所述文本单元包括以下至少一项:

语句,或者,段落。

7.一种文档结构树的生成装置,其特征在于,包括:

获取模块,用于获取文本,所述文本包括至少两个文本单元,所述至少两个文本单元包括第一文本单元与第二文本单元,所述第一文本单元与所述第二文本单元相邻;

所述获取模块,还用于获取至少两个语义信息,所述至少两个语义信息包括第一语义信息与第二语义信息,所述第一语义信息与所述第二语义信息用于确定所述第一文本单元与所述第二文本单元之间的层级关系,所述第一文本单元对应于所述第一语义信息,所述第二文本单元对应于所述第二语义信息;

处理模块,用于将所述至少两个语义信息输入到神经网络模型,推理得到所述文本的第一文档结构树。

8.根据权利要求7所述的装置,其特征在于,所述装置还包括:

校验模块,用于获取第一数据,所述第一数据是用户对所述第一文档结构树进行校验而确定的数据;

所述处理模块,还用于根据所述第一数据更新所述神经网络模型。

9.根据权利要求7所述的装置,其特征在于,所述处理模块,还用于根据文档结构模板对所述第一文档结构树进行更新,得到所述文本的第二文档结构树。

10.根据权利要求9所述的装置,其特征在于,所述装置还包括:

校验模块,用于获取第二数据,所述第二数据是用户对所述第二文档结构树进行校验而确定的数据;

所述处理模块,还用于根据所述第二数据更新所述神经网络模型。

11.根据权利要求8或10所述的装置,其特征在于,所述装置还包括:

存储模块,用于将校验后的所述第一文档结构树或者所述第二文档结构树存储至文档模板库。

12.根据权利要求7至11中任一项所述的装置,其特征在于,所述文本单元包括以下至少一项:

语句,或者,段落。

13.一种计算设备集群,其特征在于,包括至少一个计算设备,每个计算设备包括处理器和存储器;

所述至少一个计算设备的处理器用于执行所述至少一个计算设备的存储器中存储的指令,以使得所述计算设备集群执行如权利要求1至6中任一所述的方法。

14.一种计算机可读存储介质,存储有指令,当所述指令在计算机上运行时,使得所述计算机执行权利要求1至6中任一项所述的数据处理方法。

15.一种计算机设备,其特征在于,所述计算机设备包括处理器和存储器;

所述存储器,用于存储计算机程序指令;

所述处理器执行调用所述存储器中的计算机程序指令执行如权利要求1至6中任一项所述的方法。

文档结构树的生成方法以及装置

技术领域

[0001] 本申请涉及文档结构技术领域,更为具体地,涉及一种文档结构树的生成方法以及装置。

背景技术

[0002] 文档结构树可以用于指示文档的目录结构,用户可通过文档结构树认知文档的整体结构。用户还可以通过文档结构树从长文档中筛选出用户所需要的文本。

[0003] 然而,现有的文档结构树的生成方法会受到诸多因素的干扰,譬如,文档的版式、文档本身的特性,如字号、字体、行间距等,导致其泛化性较差,且生成的文档结构树的准确性不高。尤其对于层次较深的文档而言,现有的文档结构树的生成方法所生成的文档结构树很容易发生错误。

发明内容

[0004] 本申请提供一种文档结构树的生成方法以及装置,能够为绝大多数的文档生成相应的文档结构树,且不受文档的版式、字符的属性信息等因素的限制。

[0005] 第一方面,提供了一种文档结构树的生成方法,包括:获取文本,该文本包括至少两个文本单元,至少两个文本单元包括第一文本单元与第二文本单元,第一文本单元与第二文本单元相邻;获取至少两个语义信息,至少两个语义信息包括第一语义信息与第二语义信息,第一语义信息与第二语义信息用于确定所述第一文本单元与所述第二文本单元之间的层级关系,第一文本单元对应于第一语义信息,第二文本单元对应于第二语义信息;将至少两个语义信息输入到神经网络模型,推理得到文本的第一文档结构树。

[0006] 通过基于文档的语义信息的推理来生成文档的文档结构树,如此,就能够为绝大多数的文档生成相应的文档结构树,且不受文档的版式、字符的属性信息等因素的限制。

[0007] 结合第一方面,在第一方面的某些实现方式中,文本单元包括以下至少一项:语句,或者,段落。

[0008] 具体来说,当文本单元是语句时,本申请实施例的文档结构树的生成装置可以获取相邻语句的语义信息,并基于该语义信息确定相邻语句之间的层级关系,并基于该层级关系来生成文档的文档结构树。当文本单元是段落时,本申请实施例的文档结构树的生成装置可以获取相邻段落的语义信息,并基于该语义信息确定相邻段落之间的层级关系,并基于该层级关系来生成文档的文档结构树。如此,能够为绝大多数的文档生成相应的文档结构树,且不受文档的版式、字符的属性信息等因素的限制。

[0009] 结合第一方面,在第一方面的某些实现方式中,该方法还包括:获取第一数据,该第一数据是用户对第一文档结构树进行校验而确定的数据;根据该第一数据更新神经网络模型。

[0010] 通过基于用户对所生成的文档结构树的校验数据来更新神经网络模型,本申请实施例可以支持通过逐次迭代的方式来优化生成的文档结构树,以及更为准确地指示文档的

结构,且后续生成的文档结构树更能够符合用户的认知。

[0011] 结合第一方面,在第一方面的某些实现方式中,该方法还包括:根据文档结构模板对第一文档结构树进行更新,得到该文本的第二文档结构树。

[0012] 如此,可以使得所生成的第二文档结构树更准确,且更符合用户的要求。

[0013] 结合第一方面,在第一方面的某些实现方式中,该方法还包括:获取第二数据,该第二数据是用户对第二文档结构树进行校验而确定的数据;根据该第二数据更新该神经网络模型。

[0014] 通过基于用户对所生成的文档结构树的校验数据来更新神经网络模型,本申请实施例可以支持通过逐次迭代的方式来优化生成的文档结构树,以及更为准确地指示文档的结构,且后续生成的文档结构树更能够符合用户的认知。

[0015] 结合第一方面,在第一方面的某些实现方式中,该方法还包括:将校验后的第一文档结构树或者第二文档结构树存储至文档模板库。

[0016] 如此,可以便于后续生成更符合用户认知的文档结构树,能够生成更准确的文档结构树。

[0017] 第二方面,提供了一种文档结构树的生成装置,包括:获取模块,用于获取文本,该文本包括至少两个文本单元,至少两个文本单元包括第一文本单元与第二文本单元,第一文本单元与第二文本单元相邻;获取模块,还用于获取至少两个语义信息,至少两个语义信息包括第一语义信息与第二语义信息,第一语义信息与第二语义信息用于确定第一文本单元与第二文本单元之间的层级关系,第一文本单元对应于第一语义信息,第二文本单元对应于第二语义信息;处理模块,用于将至少两个语义信息输入到神经网络模型,推理得到该文本的第一文档结构树。

[0018] 结合第二方面,在第二方面的某些实现方式中,文本单元包括以下至少一项:语句,或者,段落。

[0019] 结合第二方面,在第二方面的某些实现方式中,该装置还包括校验模块,该校验模块用于获取第一数据,该第一数据是用户对第一文档结构树进行校验而确定的数据;该处理模块,还用于根据第一数据更新神经网络模型。

[0020] 结合第二方面,在第二方面的某些实现方式中,处理模块,还用于根据文档结构模板对第一文档结构树进行更新,得到该文本的第二文档结构树。

[0021] 结合第二方面,在第二方面的某些实现方式中,该装置还包括校验模块,该校验模块用于获取第二数据,该第二数据是用户对第二文档结构树进行校验而确定的数据;该处理模块,还用于根据第二数据更新神经网络模型。

[0022] 结合第二方面,在第二方面的某些实现方式中,该装置还包括:存储模块,用于将校验后的第一文档结构树或者第二文档结构树存储至文档模板库。

[0023] 第三方面,提供了一种计算设备集群,包括至少一个计算设备,每个计算设备包括处理器和存储器;至少一个计算设备的处理器用于执行至少一个计算设备的存储器中存储的指令,以使得计算设备集群执行如第一方面以及第一方面的任一种可能实现方式中所述的方法。

[0024] 第四方面,提供了一种计算机可读存储介质,存储有指令,当所述指令在计算机上运行时,使得所述计算机执行如第一方面以及第一方面的任一种可能实现方式中任一项所

述的数据处理方法。

[0025] 第五方面,提供了一种计算机设备,所述计算机设备包括处理器和存储器;所述存储器,用于存储计算机程序指令;所述处理器执行调用所述存储器中的计算机程序指令执行如第一方面以及第一方面的任一种可能实现方式中所述的方法。

附图说明

[0026] 图1是本申请实施例的适用应用场景的示意图。

[0027] 图2是本申请实施例的文档结构树的生成方法200的流程示意图。

[0028] 图3是本申请实施例的文档结构树的初始校正的示意图。

[0029] 图4是本申请实施例的文档结构树的生成方法400的流程示意图。

[0030] 图5是本申请实施例的文档结构树的用户校验的示意图。

[0031] 图6是本申请实施例的更新神经网络模型与文档模板库的示意图。

[0032] 图7是本申请实施例的文档结构树的生成装置700的结构示意图。

[0033] 图8是本申请实施例的计算设备集群的一种结构示意图。

[0034] 图9是本申请实施例的计算设备集群的又一种结构示意图。

[0035] 图10是本申请实施例的计算设备集群的再一种结构示意图。

具体实施方式

[0036] 下面将结合附图,对本申请实施例中的技术方案进行描述。

[0037] 文档结构树可以指示文档的目录结构。用户可以通过文档结构树获取文档的各级目录标题,以及标题对应的文本内容。用户还可以通过文档结构树查看文档的整体结构,也可通过文档结构树来定位文档中的具体文本。

[0038] 然而,只有少数文档标记了目录结构。对于未标记目录结构的文档而言,往往需要通过人工的方式来标注目录结构,但这需要耗费大量的财力与人力。因此,文档结构树的自动生成技术已经成为一个热门的研究方向。

[0039] 简单来说,文档结构树的自动生成技术可以基于给定文档的自身特征,实现文档结构树(也可以为目录结构)的自动生成。其中,文档结构树的应用领域极其广泛。具体可以参看图1。

[0040] 图1是本申请实施例的适用应用场景的示意图。如图1所示,文档结构树可以应用于长文档信息抽取、企业文档搜索、文本定位、文档管理,以及,长文档处理等诸多领域。

[0041] 目前,现有的文档结构树的生成方法会受到诸多因素的干扰,例如:

[0042] 1)生成的文档结构树的准确性较弱,尤其当文档的层次较深时,现有方法所生成的文档结构树易存在错误;

[0043] 2)泛化性较差,仅适用于特定类型的文档,无法适用于更广泛的文档类型;

[0044] 3)依赖文档本身的特性,如字号,字体,行间距,等等,对于无此类特性的纯文本文件而言,现有方法无法生成合适的文档结构树。

[0045] 譬如,现有的文档结构树的生成方法可以是基于字符的属性信息,其适用于存在字符的属性信息的文档,例如,PDF文档。具体地,通过获取PDF文档的每个字符的属性信息,属性信息可以包括字符的横纵坐标位置、字体样式、字号大小、行间距等信息,然后对所获

得的所有字符的属性信息进行统计,进而区分标题文本与正文文本,并在此基础之上生成文档结构树。

[0046] 该方法虽然能够很好地适用于能够提供字符的属性信息的PDF文档,但无法适用于不能提供字符的属性信息的文档,例如,纯文本的文档。

[0047] 鉴于上述技术问题,本申请提供了一种文档结构树的生成方法以及装置,能够为绝大多数的文档生成相应的文档结构树,且不受文档的版式、字符的属性信息等因素的限制。

[0048] 下文将结合附图对本申请实施例的文档结构树的生成方法以及装置进行描述。

[0049] 图2是本申请实施例的文档结构树的生成方法200的流程示意图。其中,文档结构树的生成方法200的执行主体为文档结构树的生成装置。如图2所示,文档结构树的生成方法200包括:

[0050] S210、获取文本#A,文本#A包括至少两个文本单元,至少两个文本单元包括第一文本单元与第二文本单元,第一文本单元与第二文本单元相邻。

[0051] 具体地,文档结构树的生成装置(下文简称为“第一装置”)可以获取用户上传的文本#A,该文本#A可以包括至少两个文本单元。示例性地,至少两个文本单元包括第一文本单元与第二文本单元,其中,第一文本单元与第二文本单元相邻。

[0052] 具体来说,文本单元是文本#A的基本组成。譬如,文本单元可以是语句,也可以是段落。换句话说,一个文本可以是由多个语句组成,或者,是由多个段落组成,本申请实施例不限定。

[0053] 示例性地,语句可以为一个简短的句子,例如:“今日下雨。”;或者,“今日的天气是阴天。”等等。段落也可以包括多个语句。例如,“今日的天气是阴天,所有的课外活动都被取消了。大家因此变得不那么高兴了”。上述描述仅作为示例,不作为限定。

[0054] 上述的至少两个文本单元中的“至少两个”仅为泛指,不作为具体数量的限定。

[0055] S220、获取至少两个语义信息,至少两个语义信息包括第一语义信息与第二语义信息,第一语义信息与第二语义信息用于确定第一文本单元与第二文本单元之间的层级关系,第一语义信息对应于第一文本单元,第二语义信息对应于第二文本单元。

[0056] 具体来说,第一装置获取的语义信息愈多,第一装置基于语义信息确定的语义关系也愈为丰富,从而能够生成更准确的文档结构树。

[0057] 具体地,语义信息与文本单元是一一对应的,即:一个语义信息可以对应于一个文本单元。示例性地,文本单元为“背景技术”,其可以对应于一个语义信息,该语义信息用于体现该文本单元的语义或者类似的信息。换句说,语义信息可以是文本单元的固有属性信息,其可以体现文本单元的语义。

[0058] 更为具体地说,第一装置可以获取文本#A的至少两个语义信息,至少两个语义信息中的相邻语义信息可以用于确定与该相邻语义信息对应的相邻文本单元之间的层级关系。具体地,相邻语义信息可以用于确定与该相邻语义信息对应的相邻文本单元的语义关系,该语义关系可以用于确定相邻文本单元之间的层级关系。

[0059] 示例性地,上述的至少两个语义信息中可以包括第一语义信息与第二语义信息,第一语义信息与第二语义信息是两个相邻的语义信息。其中,第一语义信息对应于第一文本单元,第二语义信息对应于第二文本单元。其中,第一文本单元与第二文本单元是相邻的

两个文本单元,换句说,第一文本单元可以是第二文本单元的相邻文本单元。第一语义信息与第二语义信息可以确定第一文本单元与第二文本单元之间的语义关系,该语义关系可以用于确定第一文本单元与第二文本单元之间的层级关系。

[0060] 为便于描述,下文便以语句为文本单元为例进行描述。

[0061] 应理解,当文本单元是语句时,本申请实施例的文档结构树的生成装置可以获取相邻语句的语义信息,并基于该语义信息确定相邻语句之间的层级关系,并基于该层级关系来生成文档的文档结构树。当文本单元是段落时,本申请实施例的文档结构树的生成装置可以获取相邻段落的语义信息,并基于该语义信息确定相邻段落之间的层级关系,并基于该层级关系来生成文档的文档结构树。如此,能够为绝大多数的文档生成相应的文档结构树,且不受文档的版式、字符的属性信息等因素的限制。

[0062] 在S220中,第一装置可以基于获取的语义信息来确定相邻语句之间的层级关系。例如,第一语句的语义信息为“附图说明”,第二语句的语义信息为“图1是应用场景示意图。”,第一装置可以根据第一语句的语义信息与第二语句的语义信息确定第一语句是第二语句的上文描述或者概括描述。由此,第一装置可以根据第一语句的语义信息与第二语句的语义信息确定第一语句与第二语句之间的语义关系,即:第一语句是第二语句的概括或者上位描述,继而可以确定第二语句从属于第一语句,即:第一语句的层级高于第二语句的层级。相应地,第一装置可以确定第一语句是第二语句的概括或者上位描述,因此,第一语句的层级高于第二语句的层级。

[0063] 又示例性地,第三语句的语义信息为“图2是实现方式的流程图。”,第一装置可以根据第二语句的语义信息与第三语句语义信息确定第二语句和第三语句之间的语义关系,即:第二语句的层级与第三语句的层级一致。

[0064] 进一步地,第一装置可以根据第一语句与第二语句的语义关系以及第二语句与第三语句之间的语义关系,确定第一语句与第三语句之间的语义关系,即:第一语句是第三语句的概括或者上位描述。

[0065] 通过上述描述,第一装置可以确定相邻文本单元之间的层级关系。

[0066] S230、将至少两个语义信息输入到神经网络模型中,推理得到文本#A的第一文档结构树。

[0067] 具体地,第一装置在前述获得的至少两个语义信息的基础之上,对这些语义信息进行推理,可以生成文本#A的第一文档结构树。

[0068] 第一装置对至少两个语义信息的推理可以是基于神经网络模型(也可以称为预训练模型)来实现的。其中,神经网络模型是近年来在自然语言处理(natural language processing,NLP)领域被广泛采用的技术。通过大量语料对神经网络模型进行预训练,可以使该神经网络模型学习到大量的语义知识,之后再通过下游任务(在深度学习中,下游任务指的是具体的情感分类任务以及命名实体识别任务,等等,本申请实施例中的生成文档结构树也是一个下游任务)进行微调,可以极大程度地提升下游任务的训练效果。第一装置可以基于神经网络模型实现对至少两个语义信息的推理,生成文本#A的第一文档结构树。

[0069] NLP是指利用人类交流所使用的自然语言与机器进行交互通讯的技术。通过人为的对自然语言的处理,使得计算机对其能够可读并理解。NLP的相关研究始于人类对机器翻译的探索。虽然自然语言处理涉及语音、语法、语义、语用等多维度的操作,但简单而言,NLP

的基本任务是基于本体词典、词频统计、上下文语义分析等方式对待处理语料进行分词,形成以最小词性为单位,且富含语义的词项单元。

[0070] NLP是以语言为对象,利用计算机技术来分析、理解和处理自然语言的一门学科,即把计算机作为语言研究的强大工具,在计算机的支持下对语言信息进行定量化的研究,并提供可供人与计算机之间能共同使用的语言描写。包括自然语言理解(natural language understanding,NLU)和自然语言生成(natural language generation,NLG)两部分。它是典型边缘交叉学科,涉及到语言科学、计算机科学、数学、认知学、逻辑学等,关注计算机和自然语言之间的相互作用的领域。

[0071] 示例性地,在本申请实施例中,可以首先对文本#A的目录结构进行标注,并获取所有相邻文本单元之间的语义关系,然后对神经网络模型进行推理。其中,神经网络模型的底层可以使用预训练模型,其输入为至少两个文本单元,其输出为语义关系的分类。关于具体的推理过程可以参看现有的算法或者流程,在此不再赘述。

[0072] 通过基于文档的语义信息的推理来生成文档的文档结构树,如此,就能够为绝大多数的文档生成相应的文档结构树,且不受文档的版式、字符的属性信息等因素的限制。

[0073] 一个可能的实现方式,该方法还可以包括:根据文档结构模板对第一文档结构树进行更新,得到文本#A的第二文档结构树。

[0074] 具体来说,第一装置可以根据文档结构模板库中已存储的文档结构模板对上述的第一文档结构树进行更新,来生成第二文档结构树。如此,可以使得第一装置所生成的文档结构树更准确,且更符合用户的要求。

[0075] 具体来说,第一装置基于神经网络模型对至少两个语义信息的推理,可以得到第一文档结构树。进一步地,第一装置可以将其生成的第一文档结构树与文档结构模板库进行比对,并根据文档模板库中所存储的模板来校正第一文档结构树,最终可以生成第二文档结构树。具体过程可以见图3。

[0076] 图3是本申请实施例的文档结构树的初始校正的示意图。如图3所示,第一装置可以先基于神经网络模型对文档#A中的至少两个语义信息的推理生成第一文档结构树。其中,第一文档结构树中的斜体部分可以显示为预测错误或者不能确认的内容。此时,第一装置可以将第一文档结构树与文档结构模板库中的第一文档结构模板进行比对,从而校正第一文档结构树中的斜体部分的内容,如此,可以使得所生成的第二文档结构树更准确,且可以更符合用户的要求。

[0077] 示例性地,第一装置确认“一.A市信用质量分析”为一级标题。此时,第一装置可以获取第一文档结构模板中的以“一.”开头的一级标题下的二级标题,发现“(二).A市财政实力”与“一.”下的二级标题“(二)”匹配,于是确认“(二).A市财政实力”应为二级标题,而非一级标题,下面的文本“近年来A市…”非一级文本,应为二级文本。如此,可以通过结合文档结构模版的方式增强第二文档结构树的准确性,可以使得所生成的第二文档结构树更准确,且可以更符合用户的要求。

[0078] 通过基于文档的语义信息的推理来生成文档的文档结构树,如此,就能够为绝大多数的文档生成相应的文档结构树,且不受文档的版式、字符的属性信息等因素的限制。

[0079] 可以理解的是,本申请还支持通过光学字符识别(optical character recognition,OCR)技术将其他的非纯文本文档变成前述的文本#A。

[0080] 下文将结合其他附图对图2所示的文档结构树的生成方法200做进一步的描述。

[0081] 图4是本申请实施例的文档结构树的生成方法400的流程示意图。其中,文档结构树的生成方法400的执行主体为文档结构树的生成装置。如图4所示,文档结构树的生成方法400包括:

[0082] S410、获取文本#A,文本#A包括至少两个文本单元,至少两个文本单元包括第一文本单元与第二文本单元,第一文本单元与第二文本单元相邻。

[0083] 具体内容可以参看S210的描述,在此不再赘述。

[0084] S420、获取至少两个语义信息,至少两个语义信息包括第一语义信息与第二语义信息,第一语义信息与第二语义信息用于确定第一文本单元与第二文本单元之间的层级关系,第一语义信息对应于第一文本单元,第二语义信息对应于第二文本单元。

[0085] 具体内容可以参看S220的描述,在此不再赘述。

[0086] S430、将至少两个语义信息输入到神经网络模型中,推理得到文本#A的第一文档结构树。

[0087] 具体内容可以参看S230的描述,在此不再赘述。

[0088] S440、获取第一数据,该第一数据是用户对第一文档结构树进行校验而确定的数据。

[0089] 具体来说,第一装置对用户上传的文本#A生成第一文档结构树之后,可以将第一文档结构树的界面呈现给用户。用户可以查阅第一装置所生成的第一文档结构树,并依据用户的经验或者知识对文本#A的第一文档结构树进行校验。具体可以参看图5。

[0090] 图5是本申请实施例的文档结构树的用户校验的示意图。如图5所示,用户可以对第一装置所生成的文本#A的第一文档结构树进行校验。第一装置支持用户自行定义文本#A的目录结构层级。其中,斜体部分为错误分类的内容。用户对文本#A的第一文档结构树(也可以理解为目录结构树)进行校验。例如,用户可以选定一段文本,并定义其目录结构的层次与类型,即标题或文本。示例性地,用户选定“(三).A市土地实力”,并将其从一级标题改为二级标题,并将“一段期间,A市”由一级文本改为二级文本。

[0091] 如此,可以支持用户能够校正第一装置所生成的文档结构树。

[0092] S450、根据第一数据更新神经网络模型。

[0093] 具体来说,第一装置可以基于获取的用户对第一文档结构树的校验数据对神经网络模型进行更新。譬如,第一装置可以根据第一数据,重新生成训练数据,在神经网络模型原有的参数基础上进行微调训练,并更新神经网络模块的部分参数,得到更新后的神经网络模型。

[0094] 由于更新后的神经网络模型是第一装置基于用户对第一文档结构树进行校验得到的第一数据而迭代得到的,更新后的神经网络模型因此可以优化第一装置生成的文档结构树,从而可以更为准确地指示文档的结构。

[0095] 通过基于用户对所生成的文档结构树的校验数据来更新神经网络模型,本申请实施例可以支持通过逐次迭代的方式优化生成的文档结构树,以及更为准确地指示文档的结构,且后续生成的文档结构树更能够符合用户的认知。

[0096] 可选地,方法400还可以包括:

[0097] S460、将校验后的第一文档结构树存储至文档结构模板库。

[0098] 具体来说,第一装置可以将经用户校验后的第一文档结构树存储至文档结构模板库,如此,可以便于后续生成更符合用户认知的文档结构树,能够生成更准确的文档结构树。具体可以参看图6。

[0099] 图6是本申请实施例的更新神经网络模型与文档结构模板库的示意图。如图6所示,第一装置可以归纳校验后的第一文档结构树的结构特征信息,并将其保存至文档结构模板库中。例如,第一装置可以获取第一文档结构树的一级标题的前若干字符,并存储到文档结构模板库中的一级目录中。再通过归纳总结第一文档结构树的二级标题的前若干字符,并加上一级目录作为前缀,添加到文档结构模板库中的二级目录中。在进行更新神经网络模型时,第一装置可以将训练数据分为输入和输出两部分,输入为两个相邻的文本单元,输出为两个相邻文本单元所对应的语义关系分类(可以用标签进标识)。对训练集合中的所有文本进行预处理,可得到若干个相邻文本单元。

[0100] 示例性地,在图6中,语句1与语句2是两个相邻语句,标签1用于指示语句1与语句2之间的语义关系的类别,语句2与语句3是两个相邻语句,标签2用于指示语句2与语句3之间的语义关系的类别,语句3与语句4是两个相邻语句,标签3用于指示语句3与语句4之间的语义关系的类别。在获取到标签1至标签3之后,可以使用这些信息更新底层模型,从而更新神经网络模型。

[0101] 可选地,本申请实施例的文档结构树的生成方法可以适用于长文档信息抽取领域。具体地,用户可根据按照上述的方法所生成的文档结构树的目录内容,对该字段进行定位。假如用户根据一定的先验知识,了解到生产总值只可能出现在一级目录B市政府信用质量分析中,或者二级目录B市经济实力中,就可以大幅缩小抽取的范围,节约服务调用的时间。

[0102] 可选地,本申请实施例的文档结构树的生成方法可以适用于企业文档搜索领域。具体地,用户可根据已有的先验知识,从上述的方法所生成的文档结构树中筛选特定目录层级下的文本,再对这部分文本进行模糊匹配,检索相关数据。

[0103] 总的来说,本申请实施例的文档结构树的生成方法不需要依赖pdf文件与人工撰写的规则,这可以节省大量的人力与时间成本,且迁移效果好,对于目录结构层次较深的复杂文档也能取得较好的效果。且提供了校验界面,用户可修正生成的结果,不断迭代优化底层的算法模型(如更新神经网络模型),并更新对应的文档结构模板库。

[0104] 一个可能的实现方式中,当第一装置向用户展示文本#A的第二文档结构树时,该第一装置还可获取用户对第二文档结构树进行校验而确定的第二数据。

[0105] 可选地,第一装置可以根据第二数据来更新神经网络模型,具体描述可以参看关于第一数据的描述,在此就不再赘述了。

[0106] 以上描述了本申请实施例的方法实施例,下面对相应的装置实施例进行介绍。

[0107] 图7本申请实施例的文档结构树的生成装置700示意性框图。装置700包括获取模块710与处理模块720。

[0108] 可选地,装置700还包括校验模块730以及存储模块740。

[0109] 存储器740包括但不限于是随机存储记忆体(random access memory,RAM)、只读存储器(read-only memory,ROM)、可擦除可编程只读存储器(erasable programmable read only memory,EPROM)、或便携式只读存储器(compact disc read-only memory,CD-

ROM),存储器440用于相关指令及数据。

[0110] 处理模块720可以是一个或多个中央处理器(central processing unit,CPU),在处理器720是一个CPU的情况下,该CPU可以是单核CPU,也可以是多核CPU。

[0111] 其中,获取模块710,用于执行以下操作:获取文本#A,文本#A包括至少两个文本单元,至少两个文本单元包括第一文本单元与第二文本单元,第一文本单元与第二文本单元相邻;获取至少两个语义信息,至少两个语义信息包括第一语义信息与第二语义信息,第一语义信息与第二语义信息用于确定第一文本单元与第二文本单元之间的层级关系,第一文本单元对应于第一语义信息,第二文本单元对应于第二语义信息。

[0112] 处理模块720,用于执行以下操作:将至少两个语义信息输入到神经网络模型,推理得到文本#A的第一文档结构树。

[0113] 上述所述内容仅作为示例性描述。装置700用于负责执行前述方法实施例中相关的方法或者步骤。

[0114] 可选地,校验模块730,用于执行以下操作:获取第一数据,该第一数据是用户对第一文档结构树进行校验而确定的数据;或者,获取第二数据,该第二数据是用户对第二文档结构树进行校验而确定的数据。

[0115] 可选地,存储模块740,用于执行以下操作:将校验后的第一文档结构树或者第二文档结构树存储至文档结构模板库。

[0116] 可选地,处理模块720,还可以用于执行以下操作:根据第一数据或者第二数据更新该神经网络模型。

[0117] 上述所述内容仅作为示例性描述。

[0118] 上述描述仅是示例性描述。具体内容可以参见上述方法实施例所示的内容。另外,图7中的各个操作的实现还可以对应参照图2至图6所示的方法实施例的相应描述。

[0119] 本申请实施例还提供一种计算设备集群,该计算设备集群的结构示意图具体如图8所示,所述计算设备集群包括至少一个计算设备800。计算设备集群中的一个或多个计算设备800中的存储器806中可以存有相同的文档结构树的生成装置700用于执行文档结构树的生成方法200的指令。

[0120] 需要说明的是,计算设备集群中的不同的计算设备800中的存储器806可以存储不同的指令,用于执行文档结构树的生成装置700的部分功能。

[0121] 图9是本申请实施例的计算设备集群的又一种结构示意框图。如图9所示,两个计算设备900A和900B通过通信接口908实现连接。计算设备900A中的存储器上存有用于执行交互单元202和处理单元206的功能的指令。计算设备900B中的存储器上存有用于执行存储单元204的功能的指令。换言之,计算设备900A和900B的存储器906共同存储了文档结构树的生成装置700用于执行前述实施例中所述的文档结构树的生成方法的指令。

[0122] 图9所示的计算设备集群之间的连接方式可以是考虑到本申请提供的前述实施例中所述的文档结构树的生成方法需要获取至少两个文本单元的语义信息。因此,考虑将存储功能交由计算设备900B执行。

[0123] 应理解,图9中示出的计算设备900A的功能也可以由多个计算设备800完成。同样,计算设备900B的功能也可以由多个计算设备800完成。

[0124] 在一些可能的实现方式中,计算设备集群中的一个或多个计算设备可以通过网络

连接。其中,所述网络可以是广域网或局域网等等。

[0125] 图10是本申请实施例的计算设备集群的再一种结构示意框图。如图10所示,两个计算设备1000A和1000B之间通过网络进行连接。具体地,通过各个计算设备中的通信接口与所述网络进行连接。在这一类可能的实现方式中,计算设备1000A中的存储器1006中存有执行交互单元202的指令。同时,计算设备1000B中的存储器1006中存有执行存储单元204和处理单元206的指令。

[0126] 本申请实施例还提供一种计算机可读存储介质,其上存储有用于实现上述方法实施例中所述方法的计算机指令。

[0127] 例如,该计算机程序被计算机执行时,使得该计算机可以实现上述方法实施例中所述方法。

[0128] 本申请实施例还提供一种包含指令的计算机程序产品,该指令被计算机执行时使得该计算机实现上述方法实施例中所述方法。

[0129] 本申请将围绕包括多个设备、组件、模块等的系统来呈现各个方面、实施例或特征。应当理解和明白的是,各个系统可以包括另外的设备、组件、模块等,并且/或者可以并不包括结合附图讨论的所有设备、组件、模块等。此外,还可以使用这些方案的组合。

[0130] 另外,在本申请实施例中,“示例的”、“例如”等词用于表示作例子、例证或说明。本申请中被描述为“示例”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言,使用示例的一词旨在以具体方式呈现概念。

[0131] 本申请实施例中,“相应的(corresponding,relevant)”和“对应的(corresponding)”有时可以混用,应当指出的是,在不强调其区别时,其所要表达的含义是一致的。

[0132] 本申请实施例描述的网络架构以及业务场景是为了更加清楚地说明本申请实施例的技术方案,并不构成对于本申请实施例提供的技术方案的限定,本领域普通技术人员可知,随着网络架构的演变和新业务场景的出现,本申请实施例提供的技术方案对于类似的技术问题,同样适用。

[0133] 在本说明书中描述的参考“一个实施例”或“一些实施例”等意味着在本申请的一个或多个实施例中包括结合该实施例描述的特定特征、结构或特点。由此,在本说明书中的不同之处出现的语句“在一个实施例中”、“在一些实施例中”、“在其他一些实施例中”、“在另外一些实施例中”等不是必然都参考相同的实施例,而是意味着“一个或多个但不是所有的实施例”,除非是以其他方式另外特别强调。术语“包括”、“包含”、“具有”及它们的变形都意味着“包括但不限于”,除非是以其他方式另外特别强调。

[0134] 本申请中,“至少一个”是指一个或者多个,“多个”是指两个或两个以上。“和/或”,描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:包括单独存在A,同时存在A和B,以及单独存在B的情况,其中A,B可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系。“以下至少一项(个)”或其类似表达,是指的这些项中的任意组合,包括单项(个)或复数项(个)的任意组合。例如,a,b,或c中的至少一项(个),可以表示:a,b,c,a-b,a-c,b-c,或a-b-c,其中a,b,c可以是单个,也可以是多个。

[0135] 以上所述,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到变化或替换,都应涵

盖在本申请的保护范围之内。因此,本申请的保护范围应以所述权利要求的保护范围为准

[0136] 本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0137] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0138] 在本申请所提供的几个实施例中,应该理解到,所揭露的系统、装置和方法,可以通过其它的方式实现。例如,以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,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,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0139] 所述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。

[0140] 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。

[0141] 所述功能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(可以是个人计算机,服务器,或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(read-only memory,ROM)、随机存取存储器(random access memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0142] 以上所述,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以所述权利要求的保护范围为准。

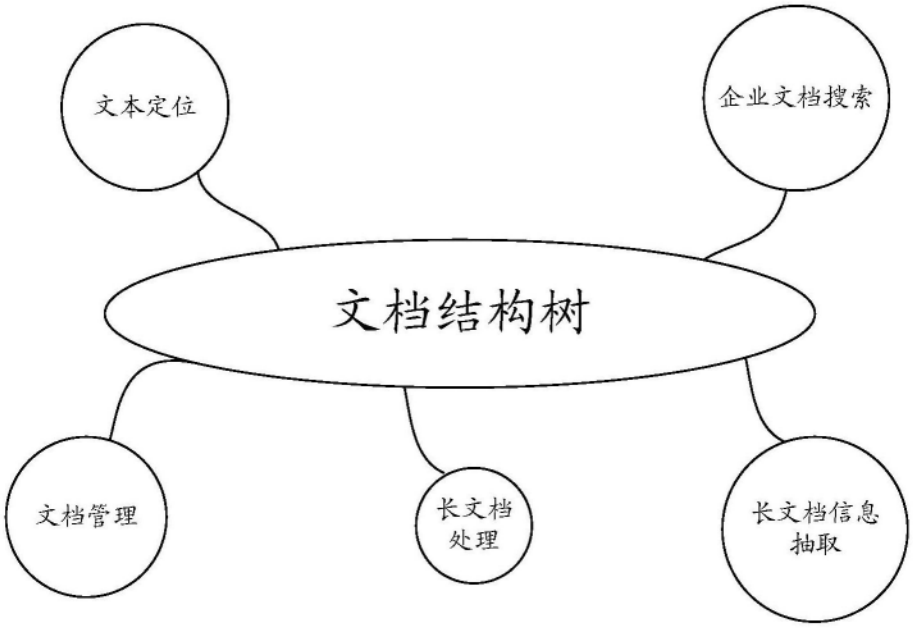

文本定位
企业文档搜索
文档结构树
文档管理
长文档处理
长文档信息抽取

图1

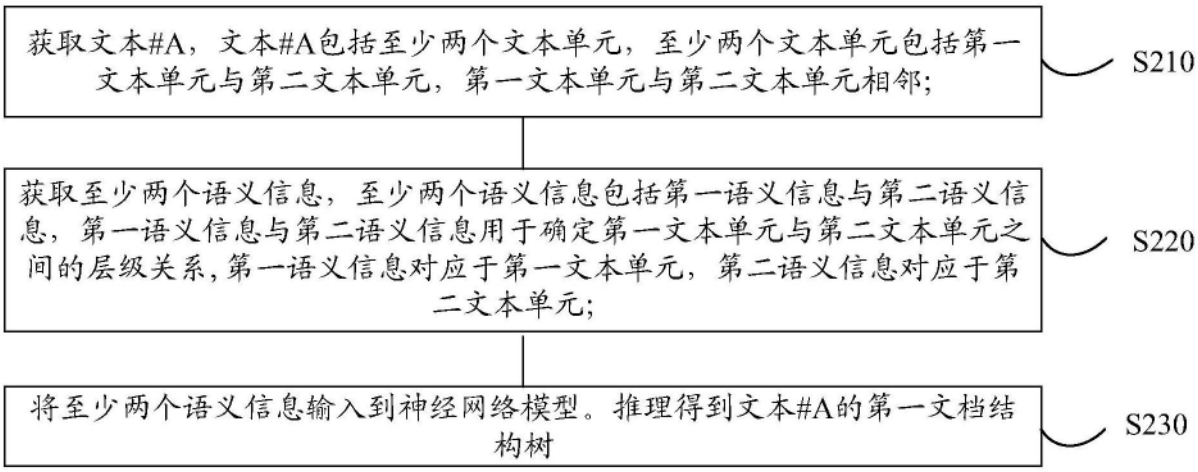

获取文本#A，文本#A包括至少两个文本单元，至少两个文本单元包括第一文本单元与第二文本单元，第一文本单元与第二文本单元相邻;
S210
获取至少两个语义信息，至少两个语义信息包括第一语义信息与第二语义信息，第一语义信息与第二语义信息用于确定第一文本单元与第二文本单元之间的层级关系,第一语义信息对应于第一文本单元，第二语义信息对应于第二文本单元;
S220
将至少两个语义信息输入到神经网络模型。推理得到文本#A的第一文档结构树
S230

图2

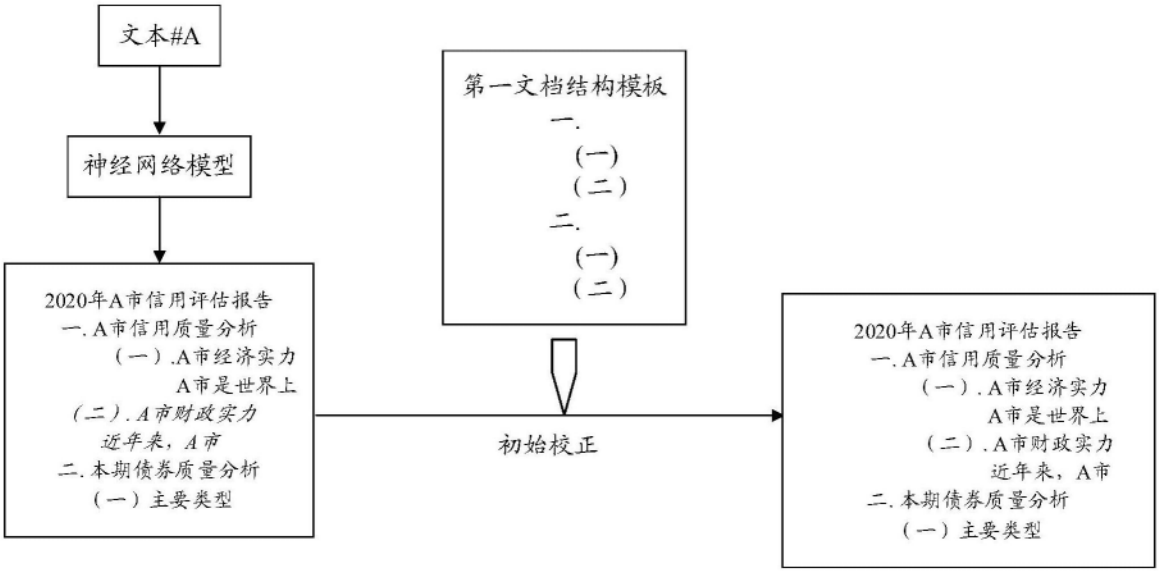

文本#A
神经网络模型
2020年A市信用评估报告
一.A市信用质量分析
（一）.A市经济实力
A市是世界上
（二）.A市财政实力
近年来，A市
二.本期债券质量分析
（一）主要类型
第一文档结构模板
一.
(一)
(二)
二.
(一)
(二)
初始校正
2020年A市信用评估报告
一.A市信用质量分析
（一）.A市经济实力
A市是世界上
（二）.A市财政实力
近年来，A市
二.本期债券质量分析
（一）主要类型

图3

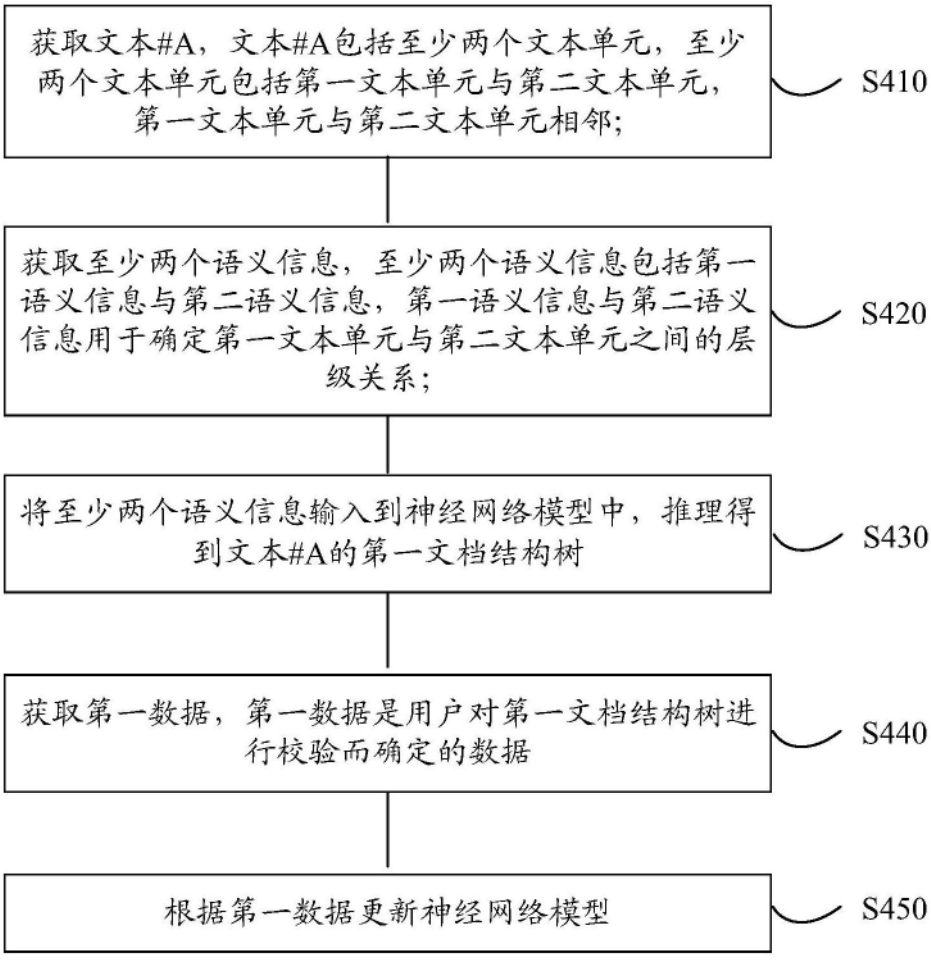

获取文本#A，文本#A包括至少两个文本单元，至少两个文本单元包括第一文本单元与第二文本单元，第一文本单元与第二文本单元相邻；
S410
获取至少两个语义信息，至少两个语义信息包括第一语义信息与第二语义信息，第一语义信息与第二语义信息用于确定第一文本单元与第二文本单元之间的层级关系；
S420
将至少两个语义信息输入到神经网络模型中，推理得到文本#A的第一文档结构树
S430
获取第一数据，第一数据是用户对第一文档结构树进行校验而确定的数据
S440
根据第一数据更新神经网络模型
S450

图4


2020年A市信用评估报告
一.A市信用质量分析
（一）.A市经济实力
A市是世界上
（二）.A市财政实力
近年来，A市
（三）.A市土地实力
一段期间，A市
二.本期债券质量分析
（一）主要类型
用户校验
2020年A市信用评估报告
一.A市信用质量分析
（一）.A市经济实力
A市是世界上
（二）.A市财政实力
近年来，A市
（三）.A市土地实力
一段期间，A市
二.本期债券质量分析
（一）主要类型

图5

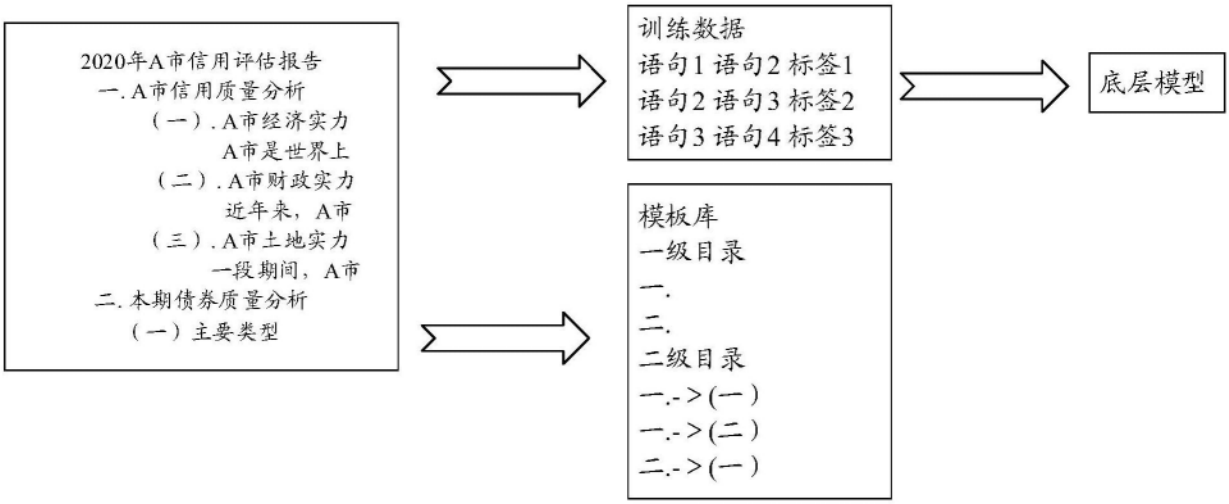

2020年A市信用评估报告
一.A市信用质量分析
（一）.A市经济实力
A市是世界上
（二）.A市财政实力
近年来，A市
（三）.A市土地实力
一段期间，A市
二.本期债券质量分析
（一）主要类型
训练数据
语句1 语句2 标签1
语句2 语句3 标签2
语句3 语句4 标签3
底层模型
模板库
一级目录
一.
二.
二级目录
一.->(一)
一.->(二)
二.->(一)

图6

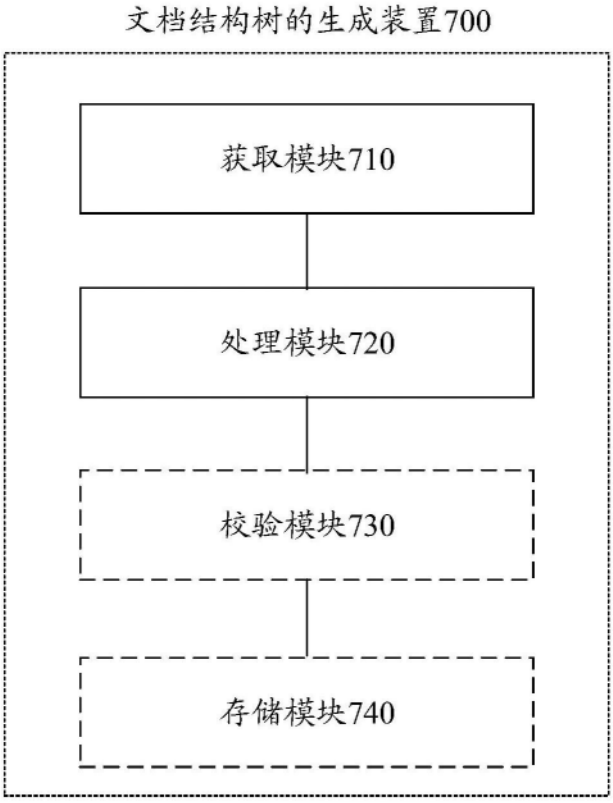

文档结构树的生成装置700
获取模块710
处理模块720
校验模块730
存储模块740

图7

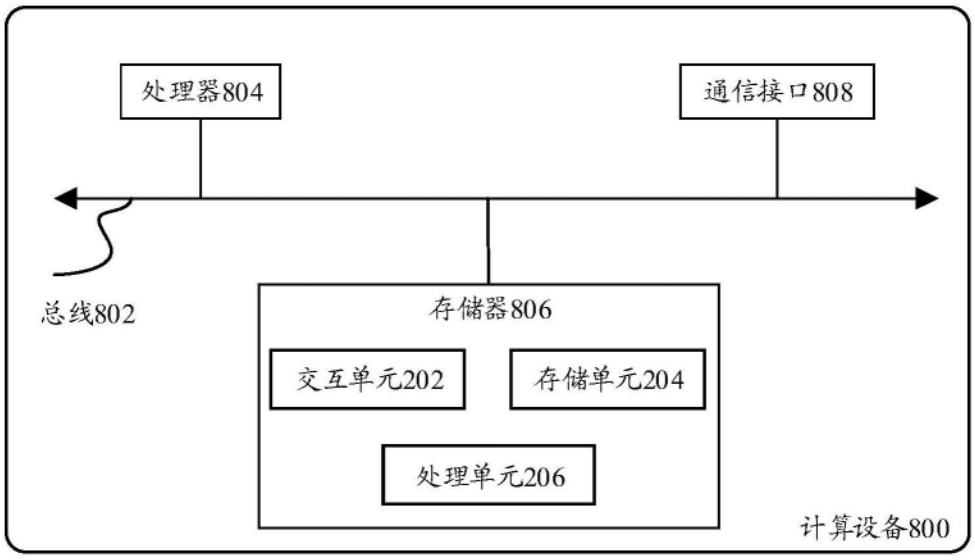

处理器804
通信接口808
总线802
存储器806
交互单元202
存储单元204
处理单元206
计算设备800

图8

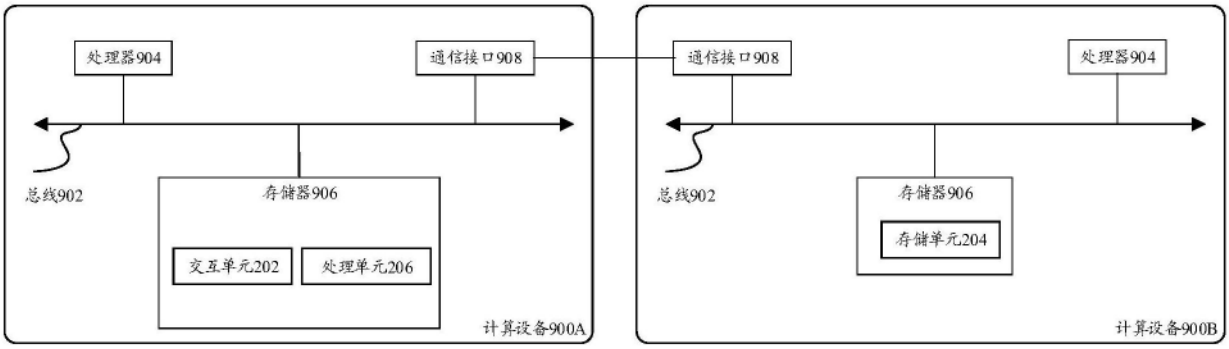

处理器904
通信接口908
总线902
存储器906
交互单元202
处理单元206
计算设备900A
通信接口908
处理器904
总线902
存储器906
存储单元204
计算设备900B

图9

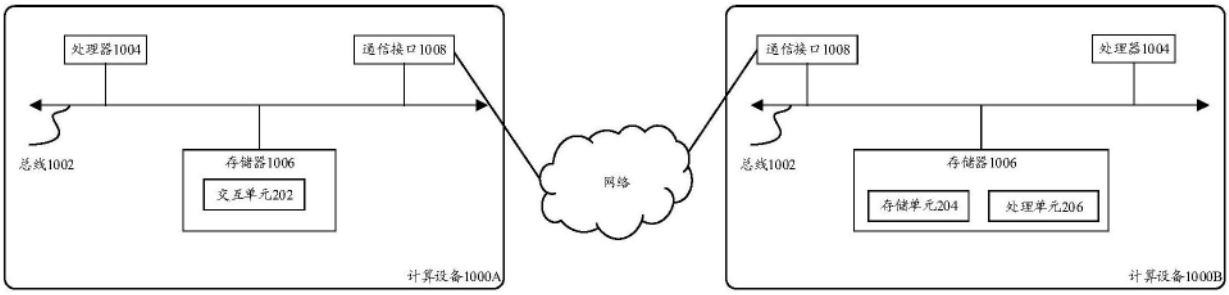

处理器1004
通信接口1008
总线1002
存储器1006
交互单元202
计算设备1000A
网络
通信接口1008
处理器1004
总线1002
存储器1006
存储单元204
处理单元206
计算设备1000B

图10