(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116663495 B
(45)授权公告日 2023.10.20

(21)申请号 202310946650.4
(22)申请日 2023.07.31
(65)同一申请的已公布的文献号
申请公布号 CN 116663495 A
(43)申请公布日 2023.08.29
(73)专利权人 中国电子技术标准化研究院
地址 100007 北京市东城区安定门东大街1号
专利权人 北京赛西科技发展有限责任公司
(72)发明人 崔静 吕千千 孔庆炜 王立玺 安淑荻 王一禾 魏梅 胡晨 高艳炫
(74)专利代理机构 北京智燃律师事务所 11864
专利代理师 柴琳琳
(51)Int.Cl.
G06F 40/103(2020.01)
G06F 40/205(2020.01)
G06F 40/258(2020.01)
G06F 16/35(2019.01)
G06F 16/36(2019.01)
(56)对比文件
CN 112905757 A,2021.06.04
CN 114153939 A,2022.03.08
CN 114706961 A,2022.07.05
CN 115098706 A,2022.09.23
JP 2002288175 A,2002.10.04
审查员 张慧雯

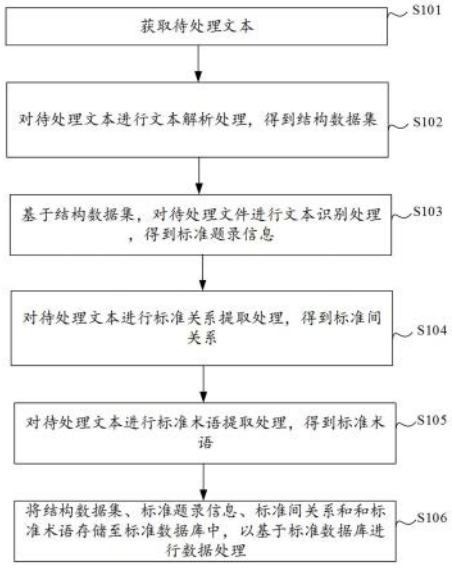
权利要求书2页 说明书12页 附图5页

(54)发明名称

文本标准化处理方法、装置、设备及介质

(57)摘要

本发明公开了文本标准化处理方法、装置、设备及介质，涉及数据处理技术领域，该方法包括：获取待处理文本；对所述待处理文本进行文本解析处理，得到结构数据集；基于所述结构数据集，对所述待处理文本进行文本识别处理，得到标准题录信息；对所述待处理文本进行标准关系提取处理，得到标准间关系；对所述待处理文本进行标准术语提取处理，得到标准术语；将所述结构数据集、所述标准题录信息、所述标准间关系和所述标准术语存储至标准数据库中，以基于所述标准数据库进行数据处理。该方案无需依赖人工经验，能够自动对待处理文本进行解析处理，精准地提取到结构数据集、标准题录信息、标准间关系和标准术语等信息，提高了标准结构化处理效率。



CN 116663495 B

1.一种文本标准化处理方法,其特征在于,该方法包括:

获取待处理文本;

对所述待处理文本进行文本解析处理,得到结构数据集;

基于所述结构数据集,对所述待处理文本进行文本识别处理,得到标准题录信息;

对所述待处理文本进行标准关系提取处理,得到标准间关系;

对所述待处理文本进行标准术语提取处理,得到标准术语;

将所述结构数据集、所述标准题录信息、所述标准间关系和所述标准术语存储至标准数据库中,以基于所述标准数据库进行数据处理;

其中,对所述待处理文本进行文本解析处理,得到结构数据集,包括:

对所述待处理文本进行特征标准类型识别处理,确定所述待处理文本的标准类型;

对所述待处理文本进行时间信息识别处理,确定所述待处理文本的时间信息;所述时间信息包括年代信息和版型信息;

基于所述待处理文本的标准类型和时间信息,对所述待处理文本进行标准要素识别和提取处理,得到标准要素;

对所述标准类型、所述时间信息和所述标准要素进行处理得到结构数据集;

所述对所述待处理文本进行标准关系提取处理,得到标准间关系,包括:

对所述待处理文本进行关系识别处理,获取标准关系;

对所述标准关系进行提取处理,并基于所述标准关系构建标准间关系图谱;

对所述标准间关系图谱进行分析处理,得到标准间关系;

对所述待处理文本进行标准术语提取处理,得到标准术语,包括:

对所述待处理文本进行标准术语识别处理,确定标准术语要素和章节位置;

根据所述标准术语要素和章节位置,对所述待处理文本进行抽取处理,得到标准术语。

2.根据权利要求1所述的方法,其特征在于,基于所述结构数据集,对所述待处理文本进行文本识别处理,得到标准题录信息,包括:

将所述待处理文本进行特征提取和文字检测处理,得到文本信息;

基于所述结构数据集中的所述标准要素、所述标准类型和所述时间信息,识别标准题录信息的位置信息;

基于所述位置信息,提取题录信息字段;

将所述题录信息字段的格式和内容进行校验和修改处理,得到标准题录信息。

3.根据权利要求1所述的方法,其特征在于,将所述结构数据集、所述标准题录信息、所述标准间关系和所述标准术语存储至标准数据库中,以进行数据处理,包括:

获取新标准和与所述新标准对应的新内容;

在所述标准数据库中根据标准题录信息查找原标准;

基于所述原标准,获取与所述原标准对应的待修改内容;

基于所述新标准,将所述原标准中的待修改内容修改为新内容。

4.根据权利要求1所述的方法,其特征在于,所述标准题录信息包括以下任意一项:分类信息、发布结构、发布实施日期、提出归口单位、起草单元、起草人;

所述标准间关系包括以下任意一项:代替关系、引用关系和采用关系;

所述标准术语包括以下任意一项:术语名称、术语定义、术语所在的标准信息、适用范

围、术语注释、术语符号、术语图例。

5.一种文本标准化处理装置，其特征在于，所述装置包括：

获取模块，用于获取待处理文本；

解析模块，用于对所述待处理文本进行文本解析处理，得到结构数据集；

题录信息识别模块，用于基于所述结构数据集，对所述待处理文本进行文本识别处理，得到标准题录信息；

标准间关系提取模块，用于对所述待处理文本进行标准关系提取处理，得到标准间关系；

标准术语提取模块，用于对所述待处理文本进行标准术语提取处理，得到标准术语；

处理模块，用于将所述结构数据集、所述标准题录信息、所述标准间关系和所述标准术语存储至标准数据库中，以基于所述标准数据库进行数据处理；

其中，所述解析模块，具体用于：

对所述待处理文本进行特征标准类型识别处理，确定所述待处理文本的标准类型；

对所述待处理文本进行时间信息识别处理，确定所述待处理文本的时间信息；所述时间信息包括年代信息和版型信息；

基于所述待处理文本的标准类型和时间信息，对所述待处理文本进行标准要素识别和提取处理，得到标准要素；

对所述标准类型、所述时间信息和所述标准要素进行处理得到结构数据集；

所述标准间关系提取模块，具体用于：

对所述待处理文本进行关系识别处理，获取标准关系；

对所述标准关系进行提取处理，并基于所述标准关系构建标准间关系图谱；

对所述标准间关系图谱进行分析处理，得到标准间关系；

所述标准术语提取模块，具体用于：

对所述待处理文本进行标准术语识别处理，确定标准术语要素和章节位置；

根据所述标准术语要素和章节位置，对所述待处理文本进行抽取处理，得到标准术语。

6.一种计算机设备，其特征在于，所述计算机设备包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，所述处理器用于执行所述程序时实现如权利要求1-4任一项所述的文本标准化处理方法。

7.一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序用于实现如权利要求1-4任一项所述的文本标准化处理方法。

文本标准化处理方法、装置、设备及介质

技术领域

[0001] 本发明涉及数据处理技术领域，尤其涉及文本标准化处理方法、装置、设备及介质。

背景技术

[0002] 随着信息技术的快速发展，文本标准化作为自然语言处理的重要环节，已经越来越多地应用到文本数据处理当中。其中，标准作为共同遵守的准则和依据，是对重复性事物和概念所做的统一规定，它以科学、技术和实践经验的综合为基础。为了使得文本数据更为规范化，对文本数据进行标准化处理显得尤为重要。

[0003] 目前，相关技术中对于传统标准文本是通过操作人员进行结构化处理，抽取标准条款、标准题录、标准间关系和标准术语，从而处理得到全文结构化的标准文本，然而该方案需要依赖大量的人工经验，费时费力，导致标准结构化处理效率较低。

发明内容

[0004] 有鉴于此，本发明提供一种文本标准化处理方法、装置、设备及介质，至少部分解决现有技术中存在的问题。

[0005] 根据本申请的另一方面，本申请实施例提供了一种文本标准化处理方法，该方法包括：

[0006] 获取待处理文本；

[0007] 对所述待处理文本进行文本解析处理，得到结构数据集；

[0008] 基于所述结构数据集，对所述待处理文本进行文本识别处理，得到标准题录信息；

[0009] 对所述待处理文本进行标准关系提取处理，得到标准间关系；

[0010] 对所述待处理文本进行标准术语提取处理，得到标准术语；

[0011] 将所述结构数据集、所述标准题录信息、所述标准间关系和所述标准术语存储至标准数据库中，以进行数据处理。

[0012] 在其中一个实施例中，对所述待处理文本进行文本解析处理，得到结构数据集，包括：

[0013] 对所述待处理文本进行特征标准类型识别处理，确定所述待处理文本的标准类型；

[0014] 对所述待处理文本进行时间信息识别处理，确定所述待处理文本的时间信息；所述时间信息包括年代信息和版型信息；

[0015] 基于所述待处理文本的标准类型和时间信息，对所述待处理文本进行标准要素识别和提取处理，得到标准要素；

[0016] 对所述标准类型、所述时间信息和所述标准要素进行处理得到结构数据集。

[0017] 在其中一个实施例中，基于所述结构数据集，对所述待处理文本进行文本识别处理，得到标准题录信息，包括：

[0018] 将所述待处理文本进行特征提取和文字检测处理，得到文本信息；

[0019] 基于所述结构数据集中的所述标准要素，所述标准类型和所述时间信息，识别标准题录信息的位置信息；

[0020] 基于所述位置信息，提取题录信息字段；

[0021] 将所述题录信息字段的格式和内容进行校验和修改处理，得到标准题录信息。

[0022] 在其中一个实施例中，对所述待处理文本进行标准关系提取处理，得到标准间关系，包括：

[0023] 对所述待处理文本进行关系识别处理，获取标准关系；

[0024] 对所述标准关系进行提取处理，并基于所述标准关系构建标准间关系图谱；

[0025] 对所述标准间关系图谱进行分析处理，得到标准间关系。

[0026] 在其中一个实施例中，对所述待处理文本进行标准术语提取处理，得到标准术语，包括：

[0027] 对所述待处理文本进行标准术语识别处理，确定标准术语要素和章节位置；

[0028] 根据所述标准术语要素和章节位置，对所述待处理文本进行抽取处理，得到标准术语。

[0029] 在其中一个实施例中，将所述结构数据集、所述标准题录信息、所述标准间关系和所述标准术语存储至标准数据库中，以进行数据处理，包括：

[0030] 获取新标准和与所述新标准对应的新内容；

[0031] 在所述标准数据库中根据标准题录信息查找原标准；

[0032] 基于所述原标准，获取与所述原标准对应的待修改内容；

[0033] 基于所述新标准，将所述原标准中的待修改内容修改为新内容。

[0034] 在其中一个实施例中，所述标准题录信息包括以下任意一项：分类信息、发布结构、发布实施日期、提出归口单位、起草单元、起草人；

[0035] 所述标准间关系包括以下任意一项：代替关系、引用关系和采用关系；

[0036] 所述标准术语包括以下任意一项：术语名称、术语定义、术语所在的标准信息、适用范围、术语注释、术语符号、术语图例。

[0037] 根据本申请的另一方面，本申请实施例提供了一种文本标准化处理装置，该装置包括：

[0038] 获取模块，用于获取待处理文本；

[0039] 解析模块，用于对所述待处理文本进行文本解析处理，得到结构数据集；

[0040] 题录信息识别模块，用于基于所述结构数据集，对所述待处理文本进行文本识别处理，得到标准题录信息；

[0041] 标准间关系提取模块，用于对所述待处理文本进行标准关系提取处理，得到标准间关系；

[0042] 标准术语提取模块，用于对所述待处理文本进行标准术语提取处理，得到标准术语；

[0043] 处理模块，用于将所述结构数据集、所述标准题录信息、所述标准间关系和所述标准术语存储至标准数据库中，以基于所述标准数据库进行数据处理。

[0044] 根据本申请的另一方面，本申请实施例提供了一种计算机设备，包括存储器、处理

器以及存储在存储器上并可在处理器上运行的计算机程序，该处理器执行该程序时实现如上述的文本标准化处理方法。

[0045] 根据本申请的另一方面，本申请实施例提供了一种计算机可读存储介质，其上存储有计算机程序，该计算机程序用于实现如上述的文本标准化处理方法。

[0046] 本申请实施例中提供的文本标准化处理方法、装置、设备及介质，通过获取待处理文本，并对待处理文本进行文本解析处理，得到结构数据集，基于结构数据集，对待处理文本进行文本识别处理，得到标准题录信息，并对待处理文本进行标准关系提取处理，得到标准间关系，然后对待处理文本进行标准术语提取处理，得到标准术语，并将结构数据集、标准题录信息、标准间关系和标准术语存储至标准数据库中，以进行数据处理。该技术方案无需依赖人工经验，能够自动对待处理文本进行解析处理，从而精准地提取到结构数据集、标准题录信息、标准间关系和标准术语等信息，并存储至数据库中，以会根据标准数据库进行数据处理，减少了人工干预和时间成本，提高了标准结构化处理效率和实施效果，极大地降低了标准的维护成本。

附图说明

[0047] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

[0048] 图1为本申请实施例提供的文本标准化处理方法的系统架构图；

[0049] 图2为本申请实施例提供的文本标准化处理方法的流程示意图；

[0050] 图3为本申请实施例提供的对待处理文本进行文本解析处理得到结构数据集方法的过程示意图；

[0051] 图4为本申请实施例提供的文本标准化处理装置的结构示意图；

[0052] 图5为本申请实施例示提供的一种计算机设备的结构示意图。

具体实施方式

[0053] 下面结合附图对本发明实施例进行详细描述。

[0054] 需说明的是，在不冲突的情况下，以下实施例及实施例中的特征可以相互组合；并且，基于本公开中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本公开保护的范围。

[0055] 需要说明的是，下文描述在所附权利要求书的范围内的实施例的各种方面。应显而易见，本文中所描述的方面可体现于广泛多种形式中，且本文中所描述的任何特定结构及/或功能仅为说明性的。基于本公开，所属领域的技术人员应了解，本文中所描述的一个方面可与任何其它方面独立地实施，且可以各种方式组合这些方面中的两者或两者以上。举例来说，可使用本文中所阐述的任何数目个方面来实施设备及/或实践方法。另外，可使用除了本文中所阐述的方面中的一或多者之外的其它结构及/或功能性实施此设备及/或实践此方法。为了便于理解，下面对本申请实施例涉及的一些技术术语进行解释：

[0056] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理

论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

[0057] 人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件主要包括计算机视觉、语音处理技术、自然语言技术以及机器学习/深度学习等几大方向。

[0058] 自然语言处理(Nature Language processing,NLP)是计算机科学领域与人工智能领域的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0059] 机器学习(Machine Learning,ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎么模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习使人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0060] 本申请实施例提供的方案涉及人工智能的自然语言处理以及机器学习等技术，具体通过下述实施例进行说明。

[0061] 如背景技术中提到的，在文本处理过程中，相关技术中可以通过操作人员对传统标准文本进行结构化处理，抽取标准条款、标准题录、标准间关系和标准术语，从而处理得到全文结构化的标准文本，然而该方案需要依赖大量的人工经验，费时费力，导致标准结构化处理效率较低。

[0062] 基于上述缺陷，本申请提供了一种文本标准化处理方法、装置、设备及介质，与现有技术相比，该技术方案无需依赖人工经验，能够自动对待处理文本进行解析处理，从而精准地提取到结构数据集、标准题录信息、标准间关系和标准术语等信息，并存储至数据库中，以根据标准数据库进行数据处理，减少了人工干预和时间成本，提高了标准结构化处理效率和实施效果，极大地降低了标准的维护成本。

[0063] 图1是本申请实施例提供的一种文本标准化处理方法的实施环境架构图。如图1所示，该实施环境架构包括：终端100和服务器200。

[0064] 终端100可以是各类AI应用场景中的终端设备。例如，终端100可以是智能电视、智能电视机顶盒等智能家居设备，或者终端100可以是智能手机、平板电脑以及电子书阅读器等移动式便携终端，或者，该终端100可以是智能眼镜、智能手表等智能可穿戴设备，本实施例对此不进行具体限定。

[0065] 其中，终端100中可安装有基于自然语言处理的AI应用。比如，该AI应用可以是智能搜索、智能问答等应用。

[0066] 服务器200可以是独立是物理服务器，也可以是由多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、

云通信、中间件服务、域名服务、安全服务、内容分发网络(content delivery network，CDN)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[0067] 其中，服务器200可以是为上述终端100中安装的AI应用提供后台服务的服务器设备。

[0068] 终端100与服务器200之间通过有线或无线网络建立通信连接。可选的，上述的无线网络或有线网络使用标准通信技术和/或协议。网络通常为因特网、但也可以是任何网络，包括但不限于局域网(Local Area Network，LAN)、城域网(Metropolitan Area Network，MAN)、广域网(WideArea Network，WAN)、移动、有线或者无线网络、专用网络或者虚拟专用网络的任何组合。

[0069] 为了便于理解和说明，下面通过图2至图5详细说明本申请实施例提供的文本标准化处理方法、装置、设备及存储介质。

[0070] 图2为本申请实施例提供的文本标准化处理方法的流程示意图，如图2所示，该方法可以应用于计算机设备，该计算机设备可以是服务器或终端，也可以是服务器与终端的组合，该方法包括：

[0071] S101、获取待处理文本。

[0072] 需要说明的是，上述待处理文本是指需要进行标准化处理的文本。

[0073] 可选的，该待处理文本可以是计算机设备获取的任意文本类型的文本，其中，可以是从用户指定的位置获取的待处理文本，也可以是通过其他外部设备导入的待处理文本，还可以是用户向计算机设备提交的待处理文本，本实施例对此不做限定。该待处理文本可以是一个，也可以是多个，每个待处理文本中可以包括至少一个词。

[0074] 示例性地，该待处理文本中可以是包括文章的一段内容，也可以包括文章的多段内容等文本数据。

[0075] S102、对待处理文本进行文本解析处理，得到结构数据集。

[0076] 上述结构数据集是指结构化解析处理后的文本，可以包括结构化数据库和结构化数字文件。结构化数据库是指用于存储结构数据集的存储位置，结构化数字文件是指包括结构化信息的文本。在获取到待处理文本后，可以对待处理文本进行全文结构化解析处理，得到结构数据集。

[0077] 需要说明的是，标准文本是按照固定的格式和结构进行编辑的，该标准文本是按照标准属性自定义设置的，目前已有的文本标准分为七个类型和六个年代版本。

[0078] 作为一种可选的实现方式，请参见图3所示，在对待处理文本进行解析处理得到结构数据集的过程中，该方法可以包括如下步骤：

[0079] S201、对待处理文本进行特征标准类型识别处理，确定待处理文本的标准类型。

[0080] S202、对待处理文本进行时间信息识别处理，确定待处理文本的时间信息，时间信息包括年代信息和版型信息。

[0081] S203、基于待处理文本的标准类型和时间信息，对待处理文本进行标准要素识别和提取处理，得到标准要素。

[0082] S204、对标准类型、时间信息和标准要素进行处理得到结构数据集。

[0083] 需要说明的是，上述待处理文本的标准类型用于表征待处理文本所属的类型标准，可以包括产品标准、基础标准、方法标准、安全标准、卫生标准、环保标准、数据标准等。

不同标准类型的文本信息所包含的标准要素也不同。待处理文本的时间信息用于表征待处理文本所属的年代信息和版型信息，年代信息和时间信息例如可以包括1981年以前、1981年至1987年、1988年至1993年、1994年至2000年、2000年至2009年以及2009年以后。标准要素用于表征待处理文本中所包含的要素信息，可以包括正文、附录章节段落、图片、表格、公式、参考文献及修改单等信息。

[0084] 在获取到待处理文本后，可以采用训练好的标准类型识别模型对待处理文本进行识别处理，得到该待处理文本属于哪种标准类型，该标准类型识别模型可以包括特征提取模块和分类模块，特征提取模块用于对待处理文本进行特征提取处理，得到特征信息，然后将特征信息通过分类模块进行处理，得到标准类型。

[0085] 可以理解的是，上述标准类型识别模型是一个输入为待处理文本，输出为标准类型的识别结果，且具有对待处理文本进行标准类型检测的能力，能够预测标准类型的神经网络模型。该标准类型识别模型用于负责建立待处理文本

[0086] 与标准类型之间的关系，其模型参数已处于最优的状态。该分类模块的可以包括但不限于全连接层和激活函数。全连接层可以包括一层，或者也可以包括多层。全连接层主要是用于对特征信息进行分类的作用。

[0087] 在得到特征信息之后，可以将特征信息通过全连接层进行处理，得到全连接相邻，并采用激活函数对全连接向量进行处理，得到待处理文本的预测结果，该预测结果可以是待处理文本属于多个不同标准类型的概率，对于每个标准类型，可以选择概率值的最大值作为待处理文本的预测结果。其中，上述激活函数可以是softmax函数，激活函数的作用是用来加入非线性因素，因为线性模型的表达能力不够，能够把输入的连续实值变换为0和1之间的输出。

[0088] 在确定出待处理文本的标准类型之后，可以通过特征模型对待处理文本进行时间信息识别处理，确定待处理文本的时间信息，该时间信息包括年代信息和版型信息。上述特征模型是一个输入为待处理文本，输出为时间信息的识别结果，且具有对待处理文本进行时间信息检测的能力，能够预测时间信息的神经网络模型。该特征模型用于负责建立待处理文本与时间信息之间的关系，其模型参数已处于最优的状态。

[0089] 由于不同的标准类型和时间信息对应的标准要素不同，在确定了标准类型和时间信息的基础上，可以根据标准类型和时间信息，进行标准要素的识别和提取，可以通过特征要素识别模型对待处理文本进行标准要素识别和提取处理，得到标准要素。特征要素识别模型是一个输入为待处理文本，输出为标准要素的识别结果，且具有对待处理文本进行标准要素检测的能力，能够预测标准要素的神经网络模型。该特征要素识别模型用于负责建立待处理文本与标准要素之间的关系，其模型参数已处于最优的状态。

[0090] 在获取到标准类型、时间信息和标准要素之后，可以对标准类型、时间信息和标准要素进行数字化处理，例如包括标准中图片、表格、公式的自动化截图和表格、公式的数字化处理，从而得到结构数据集。

[0091] 本实施例中，在对解析过程中，一方面，利用光学字符识别(Optical Character Recognition, OCR)特定产品容错库，提高识别容错率，增强系统鲁棒性；另一方面，通过基于自然语言处理(Natural Language Processing，NLP)方案实现批量自动化任务处理功能，提高加工效率，节约人力成本。

[0092] S103、基于结构数据集，对待处理文本进行文本识别处理，得到标准题录信息。

[0093] 需要说明的是，上述标准题录信息用于表征待处理文本的基本信息，可以包括以下任意一项：分类信息、发布结构、发布实施日期、提出归口单位、起草单元、起草人。其中，分类信息可以包括ICS、CCS 分类信息。标准题录信息能够为标准智能检索及系统中的其他功能奠定基础，使得标准包含的采用关系、引用文件关系分析以及与替代标准间的主要变化与更新查询、标准包含的修改单查询等应用有据可依。

[0094] 具体地，在确定出结构数据集之后，可以通过题录抽取功能模块从待处理文本中抽取到题录信息。目前获取到的待处理文本可以是通过扫描得到的图像文件，可以对待处理文本进行特征提取和文字检测处理，得到文本信息，然后基于结构数据集中的标准要素，标准类型和时间信息，识别标准题录信息的位置信息，然后基于位置信息，提取题录信息字段，并将题录信息字段的格式和内容进行校验和修改处理，得到标准题录信息。

[0095] 其中，获取到的待处理文本可以是通过扫描得到的图像文件之后，可以是将待处理文本的图像文件进行分析识别处理，获取文字及其版面信息的过程，将图像中的文字进行识别，得到文本信息。

[0096] 可以理解的是，标准文本的质量受扫描过程中每英寸点数(dpi)及操作收发等因素的影响，为保证文字识别的准确率，应通过多种主流的OCR产品进行加工来建立容错机制，有效降低识别和判定风险。在将待处理文本进行特征提取和文字检测处理，得到文本信息之后，可以根据结构数据集中的标准要素，标准类型和时间信息，识别标准题录信息的位置信息，例如可以是某一章节某几段，然后根据位置信息抽取题录信息字段，并将题录信息字段的格式整理为复合标准数据库中的存储格式，然后对题录信息字段的内容进行校验，判断其是否符合预设格式，如果符合预设格式则无需进行处理，如果不符合预设格式则需要对其进行修改处理，从而得到标准题录信息。最后，还可以通过人工进行辅助审核，将抽取的标准题录信息存储至标准数据库中，以便于后续检索和其它功能模板使用。

[0097] 其中，分辨率是扫描过程中最为重要的一个参数，代表了扫描仪在单位长度内扫描图像包含的取样点数或像素数，每英寸点数(dpi)表示。

[0098] 本实施例中通过基于所述结构数据集，对待处理文本进行文本识别处理，能够更精准地得到标准题录信息，便于后续检索和其它功能模板基于标准题录信息进行数据处理。

[0099] S104、对待处理文本进行标准关系提取处理，得到标准间关系。

[0100] 需要说明的是，上述标准间关系用于表征不同标准间之间的相互关系，可以包括以下任意一项：代替关系、引用关系和采用关系。因此在使用标准时，需要通过标准群或标准族的形式成体系的使用。基于规范性引用文件、参考文献、章条之间的相互引用关系，可以形成一个庞大的标准间关系网。标准间关系对分析某个标准的整体内容至关重要，在使用标准时也能够了解关联标准的相关信息。

[0101] 其中，在对待处理文本进行标准关系提取处理过程中，可以先对待处理文本进行关系识别处理，获取标准关系，然后对标准关系进行提取处理，并基于标准关系构建标准间关系图谱，对标准间关系图谱进行分析处理，得到标准间关系。该标准间关系图谱用于表征数据库级的各个标准之间的标准关系，可以直观地反映各个标准之间的关系。

[0102] 在得到标准题录信息之后，可以通过预先训练的特征提取模块进行关系识别处

理，获取标准关系，并进行提取处理，基于标准关系构建标准间关系图谱。其中，特征提取模型是一个输入为待处理文本，输出为标准关系的识别结果，且具有对待处理文本进行标准关系检测的能力，能够预测标准关系的神经网络模型。该特征提取模型用于负责建立待处理文本与标准关系之间的关系，其模型参数已处于最优的状态。

[0103] 其中，该特征提取模型可以包括但不限于卷积层、归一化层和激活函数，卷积层、归一化层和激活函数可以包括一层，或者也可以包括多层。卷积层用于对待处理文本的文本特征进行特征提取；归一化层用于对卷积层得到的文本特征进行归一化处理，例如可以将文本特征减去均值除以方差，得到均值为零，方差为一的正态分布，可以防止梯度爆炸和梯度消失；其中，上述激活函数可以是Sigmoid函数，也可以是Tanh函数，还可以是ReLU函数，通过将归一化处理的归一化特征经过激活函数处理，能够将其结果映射到0~1之间。

[0104] 本实施例中通过对待处理文本进行标准关系提取处理，能够精准地得到标准间关系，从而便于根据标准间关系更全面地进行数据标准分析和数据处理。

[0105] S105、对待处理文本进行标准术语提取处理，得到标准术语。

[0106] 需要说明的是，上述标准术语用于表征待处理文本中的术语属性信息，可以包括以下任意一项：术语名称、术语定义、术语所在的标准信息、适用范围、术语注释、术语符号、术语图例。

[0107] 可以理解的是，标准术语用于为标准起草人提供便利的名词术语查询渠道，在标准修订过程中，辅助规避术语的概念和编写不准确的问题，同时，通过统一的标准术语查询入口，帮助标准用户更好的理解标准中的内容。可以对标准术语进行抽取、加工、汇聚、建设形成标准名词术语子库。

[0108] 具体地，先对待处理文本进行标准术语识别处理，确定标准术语要素和章节位置，然后根据标准术语要素和章节位置，对待处理文本进行抽取处理，得到标准术语。其中，在对待处理文本进行抽取处理的过程中，可以分为两种情况，一种是将待处理文本中的标准术语按照标准类型进行部分提取，得到与标准类型对应的标准关系，例如，当标准类型包括产品标准时，可以分别得到产品标准中某一产品对应的术语集，也可以得到产品标准中对象的术语集。另一种是将待处理文本进行整体提取处理，得到标准关系，最终生成统一的术语子库对外提供服务。

[0109] 其中，标准全文的结构化加工是术语抽取功能的前提。对待处理文本进行标准术语识别处理，得到标准术语要素和章节位置，然后根据标准术语要素和章节位置，对待处理文本进行抽取处理，得到标准术语，该标准术语可以包括术语的中文名、英文名、定义、注释、相关说明标准、描述、符号、图例等信息。

[0110] 本实施例中通过对待处理文本进行标准术语提取处理，能够精准地确定出标准术语，从而便于用户更好地理解标准中的内容。

[0111] S106、将结构数据集、标准题录信息、标准间关系和标准术语存储至标准数据库中，以基于标准数据库进行数据处理。

[0112] 具体地，在获取到结构数据集、标准题录信息、标准间关系和标准术语之后，可以将标准题录信息、标准间关系和标准术语存储至标准数据库中，使得用户通过标准数据库可查询数字标准全文、标准题录信息、标准间关系和标准术语，从而实现数字标准的便捷共享。

[0113] 可选的，在基于标准数据库进行数据处理的过程中，可以先获取新标准和与新标准对应的新内容，在标准数据库中根据标准题录信息查找原标准，然后

[0114] 基于原标准，获取与原标准对应的待修改内容，并基于新标准，将原标准中的待修改内容修改为新内容。其中，待修改内容可以是原标准中的部分标准条款。

[0115] 本申请实施例中提供的文本标准化处理方法，通过获取待处理文本，并对待处理文本进行文本解析处理，得到结构数据集，基于结构数据集，对待处理文本进行文本识别处理，得到标准题录信息，并对待处理文本进行标准关系提取处理，得到标准间关系，然后对待处理文本进行标准术语提取处理，得到标准术语，并将结构数据集、标准题录信息、标准间关系和标准术语存储至标准数据库中，以进行数据处理。该技术方案无需依赖人工经验，能够自动对待处理文本进行解析处理，从而精准地提取到结构数据集、标准题录信息、标准间关系和标准术语等信息，并存储至数据库中，以会根据标准数据库进行数据处理，减少了人工干预和时间成本，提高了标准结构化处理效率和实施效果，极大地降低了标准的维护成本。

[0116] 应当注意，尽管在附图中以特定顺序描述了本发明方法的操作，但是，这并非要求或者暗示必须按照该特定顺序来执行这些操作，或是必须执行全部所示的操作才能实现期望的结果。相反，流程图中描绘的步骤可以改变执行顺序。附加地或备选地，可以省略某些步骤，将多个步骤合并为一个步骤执行，和/或将一个步骤分解为多个步骤执行。

[0117] 另一方面，图4为本申请实施例提供的一种文本标准化处理装置的结构示意图。该装置可以为终端设备或服务器内的装置，如图4所示，该装置700包括：

[0118] 获取模块710，用于获取待处理文本；

[0119] 解析模块720，用于对待处理文本进行文本解析处理，得到结构数据集；

[0120] 题录信息识别模块730，用于基于结构数据集，对待处理文本进行文本识别处理，得到标准题录信息；

[0121] 标准间关系提取模块740，用于对待处理文本进行标准关系提取处理，得到标准间关系；

[0122] 标准术语提取模块750，用于对待处理文本进行标准术语提取处理，得到标准术语；

[0123] 处理模块760，用于将结构数据集、标准题录信息、标准间关系和标准术语存储至标准数据库中，以基于标准数据库进行数据处理。

[0124] 在一些实施例中，解析模块720，具体用于：

[0125] 对待处理文本进行特征标准类型识别处理，确定待处理文本的标准类型；

[0126] 对待处理文本进行时间信息识别处理，确定待处理文本的时间信息；时间信息包括年代信息和版型信息；

[0127] 基于待处理文本的标准类型和时间信息，对待处理文本进行标准要素识别和提取处理，得到标准要素；

[0128] 对标准类型、时间信息和标准要素进行处理得到结构数据集。

[0129] 在一些实施例中，题录信息识别模块730，具体用于：

[0130] 将待处理文本进行特征提取和文字检测处理，得到文本信息；

[0131] 基于结构数据集中的标准要素，标准类型和时间信息，识别标准题录信息的位置

信息；

[0132] 基于位置信息，提取题录信息字段；

[0133] 将题录信息字段的格式和内容进行校验和修改处理，得到标准题录信息。

[0134] 在一些实施例中，标准间关系提取模块740，具体用于：

[0135] 对待处理文本进行关系识别处理，获取标准关系；

[0136] 对标准关系进行提取处理，并基于标准关系构建标准间关系图谱；

[0137] 对标准间关系图谱进行分析处理，得到标准间关系。

[0138] 在一些实施例中，标准术语提取模块750，具体用于：

[0139] 对待处理文本进行标准术语识别处理，确定标准术语要素和章节位置；

[0140] 根据标准术语要素和章节位置，对待处理文本进行抽取处理，得到标准术语。

[0141] 在一些实施例中，处理模块760，具体用于：

[0142] 获取新标准和与新标准对应的新内容；

[0143] 在标准数据库中根据标准题录信息查找原标准；

[0144] 基于原标准，获取与原标准对应的待修改内容；

[0145] 基于新标准，将原标准中的待修改内容修改为新内容。

[0146] 在一些实施例中，标准题录信息包括以下任意一项：分类信息、发布结构、发布实施日期、提出归口单位、起草单元、起草人；

[0147] 标准间关系包括以下任意一项：代替关系、引用关系和采用关系；

[0148] 标准术语包括以下任意一项：术语名称、术语定义、术语所在的标准信息、适用范围、术语注释、术语符号、术语图例。

[0149] 综上所述，本申请实施例中提供的文本标准化处理装置，该技术方案无需依赖人工经验，能够自动对待处理文本进行解析处理，从而精准地提取到结构数据集、标准题录信息、标准间关系和标准术语等信息，并存储至数据库中，以会根据标准数据库进行数据处理，减少了人工干预和时间成本，提高了标准结构化处理效率和实施效果，极大地降低了标准的维护成本。

[0150] 另一方面，本申请实施例提供的计算机设备，包括存储器、处理器以及存储在存储器上并可在处理器上运行的计算机程序，该处理器执行该程序时实现如上述的文本标准化处理方法。

[0151] 下面参考图5，图5为本申请实施例的服务器的计算机系统的结构示意图。

[0152] 如图5所示，计算机系统300包括中央处理单元(CPU)301，其可以根据存储在只读存储器(ROM)302中的程序或者从存储部分303加载到随机访问存储器(RAM)303中的程序而执行各种适当的动作和处理。在RAM 303中，还存储有系统300操作所需的各种程序和数据。CPU301、ROM 302以及RAM 303通过总线304彼此相连。输入/输出(I/O)接口305也连接至总线304。

[0153] 以下部件连接至I/O接口305：包括键盘、鼠标等的输入部分306；包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分307；包括硬盘等的存储部分308；以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分309。通信部分309经由诸如因特网的网络执行通信处理。驱动器310也根据需要连接至I/O接口305。可拆卸介质311，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器310上，以便于从其上读出

的计算机程序根据需要被安装入存储部分308。

[0154] 特别地，根据本申请的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本申请的实施例包括一种计算机程序产品，其包括承载在机器可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分303从网络上被下载和安装，和/或从可拆卸介质311被安装。在该计算机程序被中央处理单元(CPU)301执行时，执行本申请的系统中限定的上述功能。

[0155] 需要说明的是，本申请所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本申请中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本申请中，计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：无线、电线、光缆、RF等等，或者上述的任意合适的组合。

[0156] 附图中的流程图和框图，图示了按照本申请各种实施例的系统、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、或代码的一部分，前述模块、程序段、或代码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。也应当注意，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，两个接连地表示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操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。

[0157] 描述于本申请实施例中所涉及到的单元或模块可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元或模块也可以设置在处理器中，例如，可以描述为：一种处理器，包括：获取模块、解析模块、题录信息识别模块、标准间关系提取模块、标准术语提取模块及处理模块。其中，这些单元或模块的名称在某种情况下并不构成对该单元或模块本身的限定，例如，获取模块还可以被描述为“用于获取待处理文本”。

[0158] 作为另一方面，本申请还提供了一种计算机可读存储介质，该计算机可读存储介质可以是上述实施例中描述的电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中的。上述计算机可读存储介质存储有一个或者多个程序，当上述前述程序被一个

或者一个以上的处理器用来执行描述于本申请的文本标准化处理方法：

[0159] 获取待处理文本；

[0160] 对所述待处理文本进行文本解析处理，得到结构数据集；

[0161] 基于所述结构数据集，对所述待处理文本进行文本识别处理，得到标准题录信息；

[0162] 对所述待处理文本进行标准关系提取处理，得到标准间关系；

[0163] 对所述待处理文本进行标准术语提取处理，得到标准术语；

[0164] 将所述结构数据集、所述标准题录信息、所述标准间关系和所述标准术语存储至标准数据库中，以进行数据处理。

[0165] 综上所述，本申请实施例中提供的文本标准化处理方法、装置、设备及介质，通过获取待处理文本，并对待处理文本进行文本解析处理，得到结构数据集，基于结构数据集，对待处理文本进行文本识别处理，得到标准题录信息，并对待处理文本进行标准关系提取处理，得到标准间关系，然后对待处理文本进行标准术语提取处理，得到标准术语，并将结构数据集、标准题录信息、标准间关系和标准术语存储至标准数据库中，以进行数据处理。该技术方案无需依赖人工经验，能够自动对待处理文本进行解析处理，从而精准地提取到结构数据集、标准题录信息、标准间关系和标准术语等信息，并存储至数据库中，以会根据标准数据库进行数据处理，减少了人工干预和时间成本，提高了标准结构化处理效率和实施效果，极大地降低了标准的维护成本。

[0166] 以上所述，仅为本发明的具体实施方式，但本发明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发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

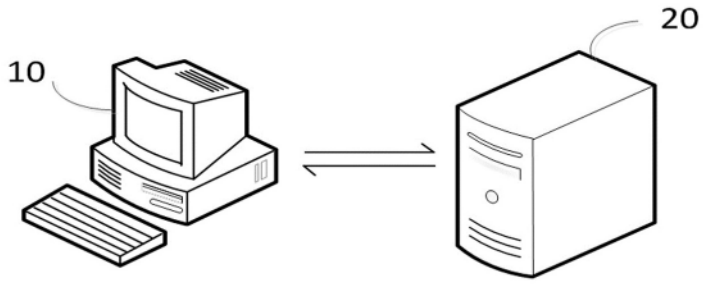


图 1

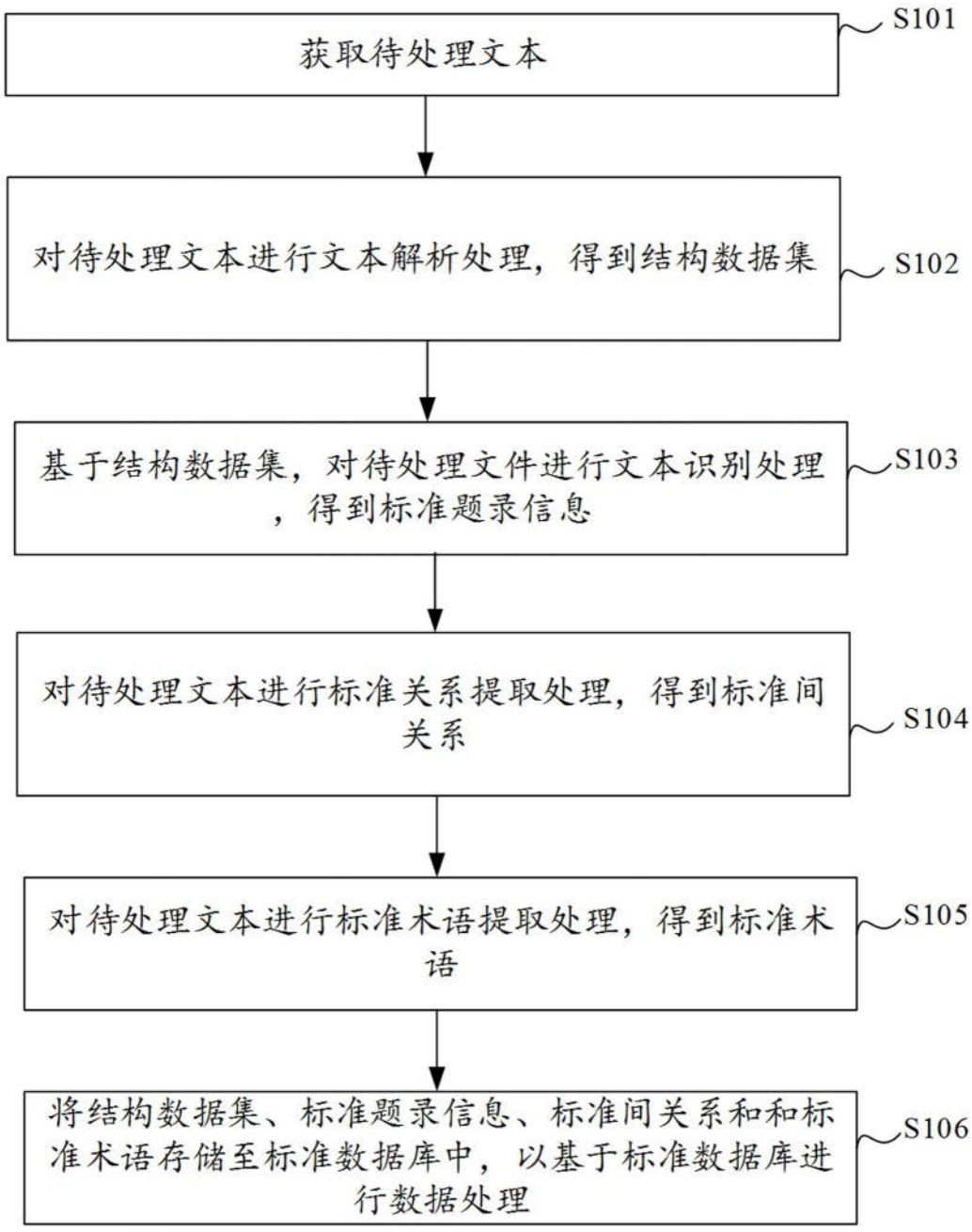

获取待处理文本
S101
对待处理文本进行文本解析处理，得到结构数据集
S102
基于结构数据集，对待处理文件进行文本识别处理，得到标准题录信息
S103
对待处理文本进行标准关系提取处理，得到标准间关系
S104
对待处理文本进行标准术语提取处理，得到标准术语
S105
将结构数据集、标准题录信息、标准间关系和和标准术语存储至标准数据库中，以基于标准数据库进行数据处理
S106

图 2

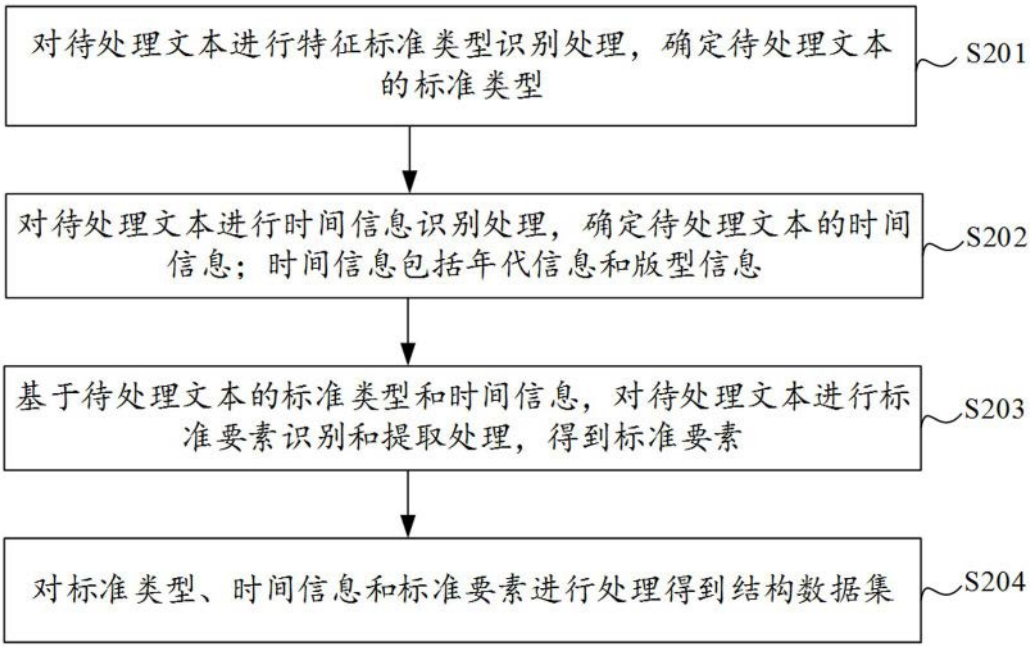

对待处理文本进行特征标准类型识别处理，确定待处理文本的标准类型
S201
对待处理文本进行时间信息识别处理，确定待处理文本的时间信息；时间信息包括年代信息和版型信息
S202
基于待处理文本的标准类型和时间信息，对待处理文本进行标准要素识别和提取处理，得到标准要素
S203
对标准类型、时间信息和标准要素进行处理得到结构数据集
S204

图 3

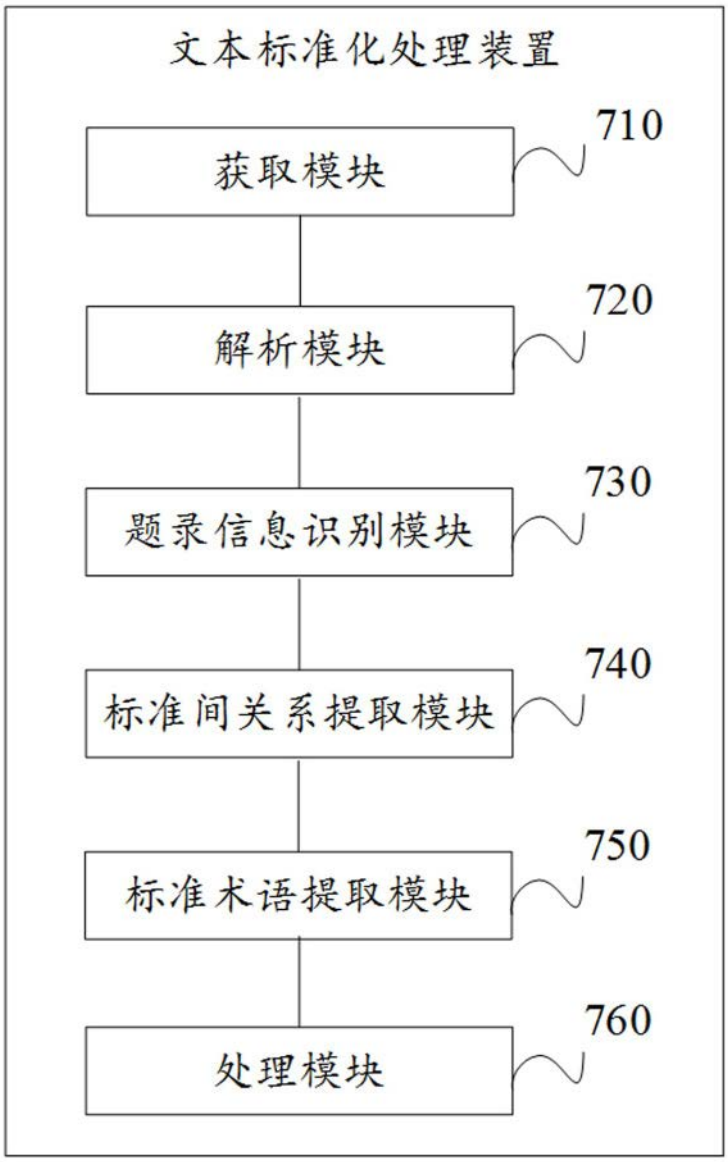

文本标准化处理装置
获取模块
710
解析模块
720
题录信息识别模块
730
标准间关系提取模块
740
标准术语提取模块
750
处理模块
760

图 4

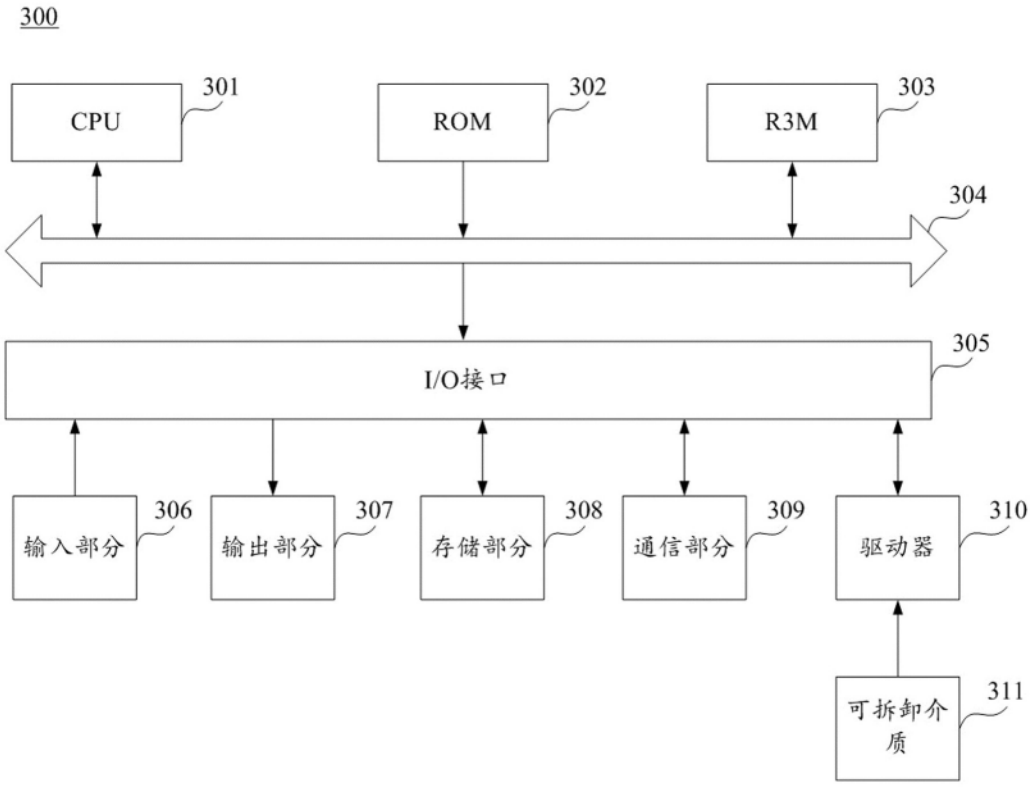

300
CPU
301
ROM
302
R3M
303
304
I/O接口
305
输入部分
306
输出部分
307
存储部分
308
通信部分
309
驱动器
310
可拆卸介质
311

图 5