(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111444331 A
(43)申请公布日 2020.07.24

(21)申请号 202010171994.9

(22)申请日 2020.03.12

(71)申请人 腾讯科技(深圳)有限公司
地址 518000 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 白冰 张峻旗 林也 白琨

(74)专利代理机构 北京志霖恒远知识产权代理事务所(普通合伙) 11435
代理人 郭栋梁

(51)Int.Cl.
G06F 16/335(2019.01)
G06F 16/35(2019.01)
G06F 16/9536(2019.01)
G06K 9/62(2006.01)
G06N 20/00(2019.01)

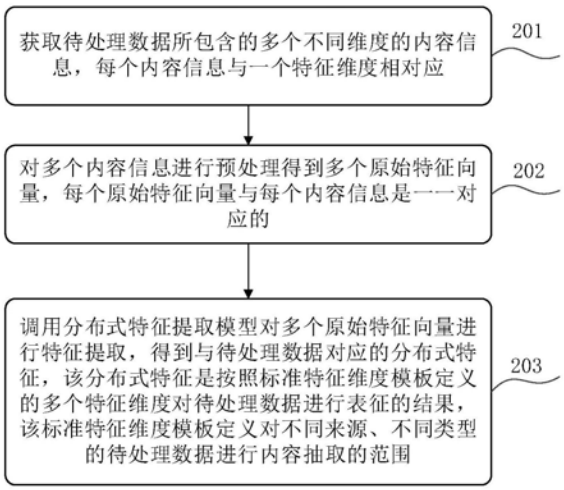
权利要求书2页 说明书16页 附图6页

(54)发明名称

基于内容的分布式特征提取方法、装置、设备及介质

(57)摘要

本申请公开了基于内容的分布式特征提取方法、装置、设备及介质。该方法通过获取待处理数据所包含的多个不同维度的内容信息,每个内容信息与一个特征维度相对应;对多个内容信息中的每一个分别进行预处理,得到多个与每个内容信息一一对应的原始特征向量;调用分布式特征提取模型对多个原始特征向量进行特征提取,得到与待处理数据对应的分布式特征,该分布式特征是按照标准特征维度模板定义的多个特征维度对待处理数据进行表征的结果,该标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。本申请实施例利用分布式特征提取模型将多个内容信息映射到同一个向量空间,以获得精准推荐的效果。



CN 111444331 A

1.一种基于内容的分布式特征提取方法,其特征在于,该方法包括:

获取待处理数据所包含的多个不同维度的内容信息,每个所述内容信息与一个特征维度相对应;

对所述多个内容信息进行预处理得到多个原始特征向量,每个所述原始特征向量与每个所述内容信息是一一对应的;

调用分布式特征提取模型对所述多个原始特征向量进行特征提取,得到与所述待处理数据对应的分布式特征,所述分布式特征是按照标准特征维度模板定义的多个特征维度对所述待处理数据进行表征的结果,所述标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。

2.根据权利要求1所述的方法,其特征在于,所述获取待处理数据所包含的多个不同维度的内容信息,包括:

调用所述标准特征维度模板对所述待处理数据进行内容抽取,得到所述待处理数据所包含的多个不同维度的内容信息。

3.根据权利要求1所述的方法,其特征在于,所述对所述多个内容信息进行预处理得到多个原始特征向量,包括:

调用与每个所述内容信息相对应的预处理策略对所述内容信息进行预处理,得到与所述内容信息相对应的原始特征向量。

4.根据权利要求3所述的方法,其特征在于,所述调用与每个所述内容信息相对应的预处理策略对所述内容信息进行预处理,包括:

确定所述内容信息的数据类型;

根据所述数据类型确定与所述数据类型对应的预处理策略;

利用所述预处理策略将所述内容信息转换成与所述内容信息对应的所述原始特征向量。

5.根据权利要求1所述的方法,其特征在于,所述调用分布式特征提取模型对所述多个原始特征向量进行特征提取,包括:

将多个所述原始特征向量进行特征拼接处理和加噪处理;

调用所述分布式特征提取模型对前述拼接处理和加噪处理后的结果进行特征提取,得到与所述待处理数据对应的分布式特征。

6.根据权利要求5所述的方法,其特征在于,所述调用所述分布式特征提取模型对前述拼接处理和加噪处理后的结果进行特征提取,包括:

将前述拼接处理和加噪处理后的结果与权重矩阵相乘,输出线性特征向量;

利用激活函数对所述线性特征向量进行非线性处理,得到所述分布式特征。

7.根据权利要求5所述的方法,其特征在于,在将多个所述原始特征向量进行特征拼接处理和加噪处理之前,所述调用分布式特征提取模型对所述多个原始特征向量进行特征提取还包括:

对所述多个原始特征向量中所包含的多热编码向量分别进行归一化处理。

8.根据权利要求1所述的方法,其特征在于,所述分布式特征提取模型是按照以下步骤训练得到:

获取训练样本集,所述训练样本集中每个样本包括多个不同的内容信息;

对每个所述样本进行预处理，得到与每个所述样本相对应的多个样本原始特征向量，每个所述样本原始特征向量与每个所述样本所包含的一个内容信息相对应；

将与每个所述样本对应的多个样本原始特征向量输入到待训练的分布式特征提取模型，输出与所述多个样本原始特征向量相对应的重构特征向量和与所述每个所述样本对应的分布式特征；

利用目标损失函数对所述待训练的分布式特征提取模型进行训练，直到所述目标损失函数达到最小值时训练完成，得到所述分布式特征提取模型，所述目标损失函数是根据所述样本原始特征向量、所述重构特征向量和所述分布式特征定义的。

9.根据权利要求8所述的方法，其特征在于，所述根据所述样本原始特征向量、所述重构特征向量和所述分布式特征定义目标损失函数包括以下步骤：

按照与所述样本原始特征向量对应的损失函数类型计算所述样本原始特征向量和与其对应的重构特征向量之间的损失误差，得到第一部分损失值；

计算所述分布式特征之间的相似度损失函数，得到第二部分损失值；

将所述第一部分损失值和所述第二部分损失值加权求和得到所述目标损失函数。

10.一种基于内容的分布式特征提取装置，其特征在于，该装置包括：

数据获取单元，用于获取待处理数据所包含的多个不同维度的内容信息，每个所述内容信息与一个特征维度相对应；

数据预处理单元，用于对所述多个内容信息进行预处理得到多个原始特征向量，每个所述原始特征向量与每个所述内容信息是一一对应的；

特征提取单元，用于调用分布式特征提取模型对所述多个原始特征向量进行特征提取，得到与所述待处理数据对应的分布式特征，所述分布式特征是按照标准特征维度模板定义的多个特征维度对所述待处理数据进行表征的结果，所述标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。

11.一种计算机设备，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，其特征在于，所述处理器执行所述程序时实现如权利要求1-9中任一项所述的方法。

12.一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如权利要求1-9中任一项所述的方法。

基于内容的分布式特征提取方法、装置、设备及介质

技术领域

[0001] 本申请一般涉及大数据技术领域，尤其涉及基于内容的分布式特征提取方法、装置、设备及介质。

背景技术

[0002] 随着电子设备的发展，越来越多的人选择在电子设备上阅读新闻资讯。基于人工智能的个性化新闻推荐系统，通常是通过机器学习算法，特别是神经网络来对新闻内容进行特征提取。例如，基于协同信息的表征提取，或者基于内容信息的表征提取，或者基于前两者的组合方式的表征提取。前者依赖用户的交互信息(例如点击、收藏等)，其依赖于用户的交互操作。后者，依赖新闻内容(包括标题、作者、文本、正文等)自身的信息。

[0003] 但是，现在新闻资讯的形式多样化，内容信息非常丰富。利用现有的表征提取方法，不能针对不同来源、不同类型的资讯数据进行统一的特征提取，从而导致推荐系统不能准确地理解不同来源、类型的资讯数据。

发明内容

[0004] 鉴于现有技术中的上述缺陷或不足，期望提供一种基于内容的分布式特征提取方法、装置、设备及介质，来获取待处理数据所包含的内容信息之间的相关性。

[0005] 一方面，本申请实施例提供了一种基于内容的分布式特征提取方法，该方法包括：

[0006] 获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应；

[0007] 对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的；

[0008] 调用分布式特征提取模型对多个原始特征向量进行特征提取，得到与待处理数据对应的分布式特征，分布式特征是按照标准特征维度模板定义的多个特征维度对待处理数据进行表征的结果，该标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。

[0009] 一方面，本申请实施例提供了一种基于内容的分布式特征提取装置，该装置包括：

[0010] 数据获取单元，用于获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应；

[0011] 数据预处理单元，用于对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的；

[0012] 特征提取单元，用于调用分布式特征提取模型对多个原始特征向量进行特征提取，得到与待处理数据对应的分布式特征，该分布式特征是按照标准特征维度模板定义的多个特征维度对待处理数据进行表征的结果，该标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。

[0013] 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本申请实施例描述的方法。

[0014] 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：

[0015] 该计算机程序被处理器执行时实现如本申请实施例描述的方法。

[0016] 本申请实施例提供的基于内容的分布式特征提取方法、装置、设备及介质，该方法通过获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应；再对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的；最后，调用分布式特征提取模型对多个原始特征向量进行特征提取，得到与待处理数据对应的分布式特征，该分布式特征是按照标准特征维度模板定义的多个特征维度对待处理数据进行表征的结果，该标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。本申请实施例通过调用分布式特征提取模型将待处理数据的多个内容信息映射到同一个向量空间，有效地解决了无法准确理解待处理数据所包含内容信息之间的相关性的问题，利用这些相关性有助于提升下游系统的性能，获取精准推荐的效果。

附图说明

[0017] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

[0018] 图1是本申请实施例提供的基于内容的分布式特征提取方法的实施环境架构图；

[0019] 图2示出了本申请实施例提供的基于内容的分布式特征提取方法的流程示意图；

[0020] 图3示出了本申请实施例提供的基于内容的分布式特征提取方法的流程示意图；

[0021] 图4示出了本申请实施例提供的构建分布式特征提取模型的方法步骤的流程示意图；

[0022] 图5示出了本申请实施例提供的分布式特征提取方法的完整流程示意图；

[0023] 图6示出了本申请实施例提供的分布式特征提取方法的原理示意图；

[0024] 图7示出了本申请实施例提供的构建分布式特征提取模型的原理示意图；

[0025] 图8示出了根据本申请实施例提供的基于内容的分布式特征提取装置的示例性结构框图；

[0026] 图9示出了适于用来实现本申请实施例的终端设备或服务器的计算机系统的结构示意图。

具体实施方式

[0027] 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关公开，而非对该公开的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与公开相关的部分。

[0028] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0029] 技术术语解释

[0030] 资讯数据，是能够表达特定事件的信息，资讯具有时效性，通常来说，用户希望查看的是最近时间段内发生的新闻事件。其类型可以包括文章、图文、小视频、短视频等。

[0031] 分布式特征用于从多个特征维度表达待处理数据的语义关系。

[0032] 图1是本申请实施例提供的基于内容的分布式特征提取方法的实施环境架构图。如图1所示，该实施环境架构包括：终端设备101和服务器102。

[0033] 在智能推荐系统中，用户通过终端设备接收服务器推送的资讯数据，并在终端设备上展示资讯数据。资讯数据的来源、类型有多种多样。智能推荐系统，可以是新闻推荐系统、广告推荐系统等。

[0034] 终端设备101用于从服务器接收并展示资讯数据。终端设备可以是智能电视、智能电视机顶盒等智能家居设备，也可以是智能手机、平板电脑以及电子书阅读器、智能眼镜、智能手表、智能音箱等移动设备，还可以是台式电脑等，但并不局限于此。

[0035] 在终端设备101安装有多个应用程序，应用程序可以是基于自然语言处理的应用程序，声音社交应用程序，即时通讯应用程序、用于系统管理的应用程序中的部分资讯子程序、基于购物的社交类应用程序、浏览器程序等。

[0036] 服务器102可以是独立的物理服务器，也可以是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、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。

[0037] 其中，服务器102向终端设备101提供资讯数据内容。终端101与服务器102之间通过有线或无线通信方式进行直接或间接地连接。可选地，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，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(Wide Area 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。

[0038] 本申请提供的基于内容的分布式特征提取方法可以由基于内容的分布式特征提取装置来实施。基于内容的分布式特征提取装置可以安装在服务器上，也可以安装在其他终端设备上。

[0039] 请参考图2，图2示出了本申请实施例提供的基于内容的分布式特征提取方法的流程示意图。如图2所示，该方法包括：

[0040] 步骤201，获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应。

[0041] 在上述步骤中，待处理数据可以是新入库的资讯数据、广告数据等。例如，资讯数据的类型可以不同，可以是新闻文章，包含图片的图文文章、小视频、图片、短视频等。

[0042] 获取待处理数据所包含的多个内容信息。内容信息是待处理数据所包含的内容。不同类型的待处理数据所包含的内容信息不同。例如，待处理数据为文章，文章可以包括类型、来源、标题、作者、直接来源标签、或者间接来源标签、一级类目、二级类目、正文、正文长度等内容信息。又例如待处理数据为小视频，则小视频可以包括类型、来源、标题、作者、直接来源标签、或者间接来源标签、一级类目、二级类目、封面图、视频分辨率、视频长度等内容信息。

[0043] 可选地，通过调用标准特征维度模板对待处理数据进行内容抽取，得到待处理数

据所包含的多个不同维度的内容信息。

[0044] 标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。不同的应用场景定义的标准特征维度模板可以是不同的。例如在资讯推荐场景中，标准特征维度模板包括资讯类型、来源、标题、作者、标签(Tag)、一级类目、二级类目、正文、正文长度、组图、图片数、封面图、视频分辨率、视频长度等。对于不同来源、不同类型的资讯均按照标准特征维度模板进行内容抽取，得到多个内容信息。若资讯数据中不包含标准特征维度模板定义的特征维度时，该特征维度对应的权重值至为0。

[0045] 步骤202，对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的。

[0046] 在上述步骤中，通过对多个内容信息进行预处理得到多个原始特征向量。例如，待处理数据为图文数据，按照标准特征维度模板从图文数据中抽取多个内容信息，如资讯类型、来源、标题、作者、标签、一级类目、二级类目、正文、正文长度、组图、图片数等。对于图文数据中未包含的特征维度，如封面图、视频分辨率、视频长度等内容特征则将其权重值设置为零。

[0047] 对多个内容信息进行预处理得到多个原始特征向量，可以包括以下步骤：

[0048] 调用与每个内容信息相对应的预处理策略对与内容信息进行预处理，得到与内容信息相对应的原始特征向量。

[0049] 其中，原始特征向量是将内容信息转换成特征向量的结果。

[0050] 预处理策略是指将内容信息转换成特征向量的处理算法。例如可以是单词化处理、向量化处理、分词处理、归一化处理、图像特征提取处理等。预处理策略还可以是多种处理的组合。例如，预处理策略包括以下组合方式：

[0051] 针对内容信息进行单词化处理和独热编码处理；

[0052] 针对内容信息进行分词处理、单词化处理和独热编码处理；

[0053] 针对内容信息进行图像特征提取处理和求平均处理；

[0054] 针对内容信息进行归一化处理。

[0055] 通过调用与每个内容信息相对应的预处理策略对内容信息进行预处理还可以包括：

[0056] 确定内容信息的数据类型；

[0057] 根据数据类型确定与数据类型对应的预处理策略；

[0058] 利用预处理策略将内容信息转换成与内容信息对应的原始特征向量。

[0059] 数据类型是指待处理数据的类型。例如，可以是文章、图文、小视频、短视频等。

[0060] 根据数据类型确定与数据类型对应的预处理策略，例如对文章所包含的来源信息进行单词化处理、向量化处理。单词化处理和向量化处理可以通过开源工具来完成。

[0061] 在单词化处理、向量化处理之后，向量化处理可以包括独热编码处理或多热编码处理。独热编码处理，即One-Hot编码，又称一位有效编码，其方法是使用N位状态寄存器来对N个状态进行编码，每个状态都由他独立的寄存器位，并且在任意时候，其中只有一位有效。例如来源信息为[直接来源，间接来源]。经过独热编码后对应的独热编码向量为[10，01]。

[0062] 又例如，文章包含正文信息，其对应的预处理策略是对正文进行分词处理、单词化

处理、独热编码处理。分词处理可以是基于规则的分词处理，即按照预先构建的词典，将正文信息中出现的词进行切分。然后，对切分的每个词进行单词化处理、向量化处理和独热编码处理。将与每个词对应的独热编码向量进行相或，得到与正文信息对应多热编码向量。单词化处理可以包括获取每个单词或ID对应的频次，将该频次与预设的过滤阈值进行比较，若该频次大于预设过滤阈值时，则过滤该单词或ID。若该频次小于等于预设过滤阈值时，则将该单词或ID标记为低频词或者低频ID。

[0063] 原始特征向量可以包括独热编码向量，多热编码向量，稠密向量，标量中的至少一种。例如，与来源信息对应的原始特征向量为独热编码向量，与正文信息对应的原始特征向量为多热编码向量，与正文长度信息对应的原始特征向量为标量。

[0064] 下面为了更清楚地描述对待处理数据进行预处理过程，结合表(1)进一步展开描述。

[0065]

多元特征类型	数据格式	处理方式	输出结果	涉及资讯类型
类型	ID	Token化、向量化	One-hot向量	文章、图文、小视频、短视频
来源	ID	Token化、向量化	One-hot向量	文章、图文、小视频、短视频
标题	文本（短）	分词、Token化、向量化	Many-hot向量	文章、图文、小视频、短视频
作者	词	Token化、向量化	One-hot向量	文章、图文、小视频、短视频
tag	词的集合	Token化、向量化	Many-hot向量	文章、图文、小视频、短视频
一级类目	词	Token化、向量化	One-hot向量	文章、图文、小视频、短视频
二级类目	词	Token化、向量化	One-hot向量	文章、图文、小视频、短视频
正文	文本（长）	分词、Token化、向量化	Many-hot向量	文章、图文
正文长度	整数型INT	归一化	标量	文章、图文
组图	图片的集合	图片特征提取、求平均	稠密向量	图文
图片数	整数型INT	归一化	标量	图文
封面图	图片	图片特征提取	稠密向量	小视频、短视频
视频分辨率	整数型INT	归一化	标量	小视频、短视频
视频长度	整数型INT	归一化	标量	小视频、短视频

[0066] 表(1)

[0067] 多元特征类型是标准特征维度模板定义的对不同来源、不同类型的待处理数据进行内容抽取的范围。数据格式是待处理数据所包含的内容信息的数据格式。例如，文本，词，整数型。处理方式，即预处理策略，是指对内容信息进行预处理，例如对于正文信息，其数据

格式为文本类型，且是长文本，对于文本的处理，需要先进行分词处理，然后单词化处理，然后在对每个单词进行独热编码，将独热编码的结果相加或者相或可以得到与正文信息对应的多热编码向量。

[0068] 对于正文长度、图片数、视频分辨率、视频长度等整数型的数据进行归一化处理。归一化处理可以包括获取内容信息的百分位点，利用该百分位点对内容信息进行归一化处理。这里的内容信息是与正文长度、图片数、视频分辨率、视频长度相对应的。

[0069] 对于包含图片的待处理数据，则调用图片特征提取模型对待处理数据所包含的图片进行特征提取处理，得到与图片对应的稠密向量。图片特征提取模型，例如是基于卷积神经网络的VGG16，取VGG16的全连接层FC7的特征作为提取得到的特征。图片特征提取模型的训练可以是基于ImageNet图片训练集按照深度学习算法模型训练规则进行训练得到，这里对于图片特征提取模型的具体实现结构不作限定。

[0070] 表中One-hot向量表示独热编码向量，Many-hot向量表示多热编码向量。多热编码向量是将多个词对应的独热编码向量进行相加或者相与的结果。稠密向量表示双精度浮点型数组。标量是具体的数值。

[0071] 获取待处理数据之后，根据如表1定义的标准特征维度模板对待处理数据进行内容抽取，可以包括：

[0072] 根据标准特征维度模板定义的数据类型逐个查找待处理数据中所包含的内容信息；

[0073] 确定内容信息对应的预处理策略；或者确定内容信息对应的数据格式；再根据数据格式确定预处理策略；

[0074] 按照预处理策略对内容信息进行预处理处理得到与内容信息对应的原始特征向量。

[0075] 本申请实施例中将属于同一条待处理数据的多个内容信息，按照预处理策略与之对应的内容信息进行预处理，将内容信息转换成原始特征向量。例如，内容信息为标题，标题的数据格式通常为短文本格式，将标题内容依次进行分词处理、单词化处理和向量化处理，将标题转换成多热编码向量。若待处理数据中不存在标准特征维度模板定义的内容信息，则置0，设置与内容信息对应的权重系数为零。

[0076] 步骤203，调用分布式特征提取模型对多个原始特征向量进行特征提取，得到与待处理数据对应的分布式特征，该分布式特征是按照标准特征维度模板定义的多个特征维度对待处理数据进行表征的结果，该标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。

[0077] 在上述步骤中，将与待处理数据所包含的多个内容信息一一对应的原始特征向量，输入到分布式特征提取模型，输出与待处理数据对应的分布式特征。

[0078] 分布式特征是按照标准特征维度模板定义的多个特征维度对待处理数据进行表征的结果。例如分布式特征向量为$[a_1,a_2,a_3,\ldots,a_N]$，N为标准特征维度模板定义的特征的数量。分布式特征是将与每个内容信息一一对应的原始特征向量融合到同一个向量空间的结果。

[0079] 其中，分布式特征提取模型是按照预先定义的目标损失函数对待训练的分布式特征提取模型进行训练得到的。该目标损失函数是根据样本原始特征向量、重构特征向量和

分布式特征定义的。在目标损失函数达到最小值时完成训练，得到最终的分布式特征提取模型。

[0080] 分布式特征提取模型的训练过程与上述方法步骤中描述的调用分布式特征提取模型对待处理数据进行特征提取得到分布式特征，是两个不同的阶段。训练分布式特征提取模型是线下预先训练完成的，调用分布式特征提取模型是在线预测阶段执行的。

[0081] 本申请实施例，通过定义标准特征维度模板，有效地解决了相关技术中针对不同来源、不同类型的资讯不能进行统一的特征提取的问题。

[0082] 在上述基础上，本申请实施例还提供了一种基于内容的分布式特征提取方法。请参考图3，图3示出了本申请实施例提供的基于内容的分布式特征提取方法的流程示意图。

[0083] 如图3所示，该方法包括：

[0084] 步骤301，获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应。

[0085] 步骤302，对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的；

[0086] 步骤303，对多个原始特征向量进行特征拼接处理和加噪处理；

[0087] 步骤304，调用分布式特征提取模型对经过拼接处理和加噪处理后的结果进行特征提取，得到与待处理数据对应的分布式特征。

[0088] 在上述实施例中，将多个原始特征向量进行特征拼接处理，对拼接处理的结果进行加噪处理。

[0089] 可选地，调用分布式特征提取模型对经过拼接处理和加噪处理后的结果进行特征提取，包括：

[0090] 将经过拼接处理和加噪处理后的结果与权重矩阵相乘，输出线性特征向量；

[0091] 利用激活函数对线性特征向量进行非线性处理，得到分布式特征。

[0092] 在上述步骤中，分布式特征提取模型可以在输入层接收将多个原始特征向量进行拼接处理和加噪处理后的结果。权重矩阵是初始化设置的，将经过拼接处理和加噪处理后的结果与权重矩阵相乘，然后利用激活函数对相乘的结果进行非线性处理，输出与待处理数据对应的分布式特征。分布式特征的维度数量与标准特征维度模板定义维度数量一致。

[0093] 可选地，在步骤303之前，还包括：

[0094] 对多个原始特征向量中所包含的多热编码向量分别进行归一化处理。归一化处理可以是按照L2归一化处理。即将多热编码向量中每个元素除以多热编码向量的L2范数。本申请实施例中对多热编码向量进行归一化处理，可以有效地抑制文章长度对分布式特征的质量的影响。

[0095] 下面结合图4进一步描述训练分布式特征提取模型的方法。请参考图4，图4示出了本申请实施例提供的构建分布式特征提取模型的方法步骤的流程示意图。如图4所示，该方法包括以下步骤：

[0096] 步骤401，获取训练样本集，该训练样本集中每个样本包括多个不同的内容信息；

[0097] 步骤402，对每个样本进行预处理，得到与每个样本相对应的多个样本原始特征向量，每个样本原始特征向量与每个样本所包含的一个内容信息相对应；

[0098] 步骤403，将与每个样本对应的多个样本原始特征向量输入到待训练的分布式特

征提取模型，输出与多个样本原始特征向量相对应的重构特征向量和与所述每个所述样本对应的分布式特征；

[0099] 步骤404，利用目标损失函数对待训练的分布式特征提取模型进行训练，直到目标损失函数达到最小值时训练完成，得到分布式特征提取模型，该目标损失函数是根据样本原始特征向量、重构特征向量和分布式特征定义的。

[0100] 在上述步骤中，将与每个样本对应的多个样本原始特征向量输入到待训练的分布式特征提取模型包括：

[0101] 对与每个样本对应的多个样本原始特征向量进行拼接处理和加噪处理；

[0102] 将拼接处理和加噪处理后的结果输入到待训练的分布式特征提取模型。

[0103] 分布式特征提取模型的输入层接收与每个样本对应的多个样本原始特征向量进行拼接处理和加噪处理后的结果；然后，将该结果与第一权重矩阵相乘，再利用第一激活函数进行非线性处理得与每个样本对应的分布式特征。将分布式特征再与第二权重矩阵相乘，再利用第二激活函数进行非线性处理，将处理结果分别进行特征重构，得到与每个原始特征向量一一对应的重构特征向量。将处理结果分别进行特征重构是将经过第二激活函数进行非线性处理的结果，再与第三权重矩阵相乘，加上相应的偏置值的处理过程。

[0104] 在上述步骤中，获取训练样本集，训练样本集包括多个样本，每个样本按照标准特征维度模板提取得到多个内容信息。

[0105] 将每个样本所包含的内容信息进行预处理，得到样本原始特征向量，将样本原始特征向量输入到待训练的分布式特征提取模型。待训练的分布式特征提取模型可以是深度学习神经网络，多层感知器。

[0106] 待训练的分布式特征提取模型包括输入层、多个隐藏层和输出层。其中多个隐藏层预先定义权重矩阵和激活函数。

[0107] 将多个样本原始特征向量中的多热编码向量经过归一化处理之后，经过特征拼接处理、加噪处理后，输入到共享隐含层，该共享隐含层包括两个串联的隐含层，与输入层紧邻的第一隐含层将经过特征拼接处理和加噪处理后的特征向量与该隐含层定义的权重矩阵相乘之后，通过该隐含层定义的激活函数将线性结果进行非线性处理。该隐含层定义的激活函数可以使用双曲正切(Tanh)函数、整流线性单元(Rectified Linear Unit，缩写为ReLU)函数、渗漏整流线性单元(Leaky Rectified Linear Unit，缩写为Leaky ReLU)函数、带参数整流线性单元(Parametric Rectified Linear Unit，缩写为PReLU)函数等。

[0108] 第一隐含层的输出结果即为分布式特征。将第一隐含层的输出结果输入到第二隐含层进行处理，将第一隐含层的输出结果与第二隐含层定义的权重矩阵相乘之后，通过第二隐含层定义的激活函数将线性结果进行非线性处理。

[0109] 在完成共享隐含层的处理之后，将共享隐含层的输出结果按照与每个内容信息相对应的重构方式进行处理得到与每个内容信息对应的重构特征向量。即将共享隐含层的输出结果与第三权重矩阵相乘之后，再对相乘处理后的每个元素进行误差调整。

[0110] 在上述步骤中，根据样本原始特征向量、重构特征向量和分布式特征定义目标损失函数包括以下步骤：

[0111] 按照与样本原始特征向量对应的损失函数类型计算样本原始特征向量和与其对应的重构特征向量之间的损失误差，得到第一部分损失值；

[0112] 计算分布式特征之间的相似度损失函数，得到第二部分损失值；

[0113] 将第一部分损失值和第二部分损失值加权求和得到目标损失函数。

[0114] 其中，按照与样本原始特征向量对应的损失函数类型计算样本原始特征向量和与其对应的重构特征向量之间的损失误差，得到第一部分损失值包括：

[0115] 根据内容信息确定对应的损失函数类型；

[0116] 将该内容信息所对应的多个样本原始特征向量和其对应的重构特征向量按照损失函数类型计算样本原始特征向量对应的损失误差。

[0117] 损失函数类型可以是KL(即Kullback-Leibler divergence)散度损失函数、或者多项式损失函数、或者JS(即Jensen-Shannon)散度损失函数、或者交叉熵损失函数、或者均方误差(Mean-Square Error，缩写MSE)损失函数、平均绝对误差(Mean Absolute Error，缩写为MAE)损失函数等。为了更清楚地说明根据内容信息确定对应的损失函数类型，下面结合表(2)进一步展开说明。

[0118]

输入向量类型	涉及多元特征类型	损失函数类型	备选损失函数
Many-hot向量	标题、tag、正文、	KL散度	多项式损失、JS散度
One-hot向量	类型、来源、作者、一级类目、二级类目	交叉熵	
标量	正文长度、图片数、视频分辨率、视频长度	MSE	MAE
稠密向量	组图、封面图	MSE	MAE

[0119] 表(2)

[0120] 确定多个原始特征向量中所包含的输入向量类型；确定与每个输入向量类型对应的损失函数类型；

[0121] 可以先确定内容信息对应的输入向量类型，输入向量类型如表(2)示出的输入向量类型，输入向量类型可以是多热编码向量(即Many-hot向量)、独热编码向量(One-hot向量)、标量、稠密向量。根据输入向量类型确定其对应的损失函数类型。例如，对于多热编码向量可以采用基于KL散度损失函数，或者多项式损失函数，JS散度损失函数。对于独热编码向量可以采用基于交叉熵损失函数，对于标量和稠密向量则可以利用均方误差损失函数。

[0122] 在确定与内容信息对应的损失函数类型之后，对训练样本集中每个样本对应的每个样本原始特征向量和与其对应的重构特征向量按照与每个样本原始特征向量一一对应的损失函数计算损失误差。

[0123] 例如，对于第i个样本，其对应的样本原始特征向量为N个；确定N个样本原始特征向量中每一个所对应的输入向量类型；根据输入向量类型确定与其对应的损失函数类型；在经过待训练的分布式特征提取模型进行处理之后，输出重构特征向量时，利用与输入向量类型对应的损失函数类型计算样本原始特征向量和其对应的重构特征向量之间的损失误差。对每个输入向量类型对应的损失函数类型再乘以与之对应的权重系数，然后得到第i个样本对应损失误差。

[0124] 计算训练样本集中所有样本对应的损失误差的和作为第一部分损失值。

[0125] 训练样本集中每个样本在经过待训练的分布式特征提取模型进行处理之后，分布式特征提取模型的中间隐藏层输出与样本对应的分布式特征；计算分布式特征中任意两个分布式特征向量之间的相似度损失误差，得到第二部分损失值。

[0126] 将第一部分损失值和第二部分损失值加权求和得到目标损失函数。其具体可以通过如下公式来表达：

[0127] 目标损失函数的表达式如下：

[0128] $$\theta=\arg\min_{w,w',b,b'}\sum_{x\in T}\sum_{i=1}^{M}\sum_{n=1}^{N}w_n L_n(z_{in},x_{in})+\lambda\phi(h_1,h_2,h_M) \qquad \text{公式（1）}$$

[0129] 其中，i表示待处理数据的个数，M表示待处理数据的总数量，T表示训练样本集。n表示内容信息对应的特征维度的序号，N是标准特征维度模板定义特征维度的总数量。其中，$L_n(z_{in},x_{in})$表示第i个待处理数据的第n个特征维度对应的原始特征向量与第n个特征维度对应的重构特征向量之间的损失误差，Ln的类型可以根据表(2)确定。z_{in}表示第i待处理数据的第n个特征维度对应的重构特征向量，x_{in}表示第i待处理数据的第n个特征维度对应的原始特征向量。w_n表示与第n个特征维度对应的损失函数的权重值。

[0130] 如果某个待处理数据中不包含某个特征维度的内容信息，则该某特征维度对应的损失函数的权重值设置为0。例如，对于小视频、短视频类型的待处理数据，不包括正文内容，则正文内容对应的损失函数的权重值为0。其中，损失函数Φ(h_1,h_2,h_M)表示分布式特征之间的相似度损失。

[0131] 最后通过梯度优化算法寻找公式(1)对应的最优的模型参数，在找到最优的模型参数之后，即完成训练。

[0132] 本申请实施例在目标损失函数中引入分布式特征之间的相似度损失误差，有效地保证了分布式特征的距离特性。

[0133] 本申请实施例还提供分布特征提取方法的完整处理过程。请参考图5，图5示出了本申请实施例提供的分布式特征提取方法的完整流程示意图。该方法包括：在线部分和离线部分。

[0134] 其中，在线部分的处理过程包括：

[0135] 步骤501，获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应。

[0136] 步骤502，调用预处理模型对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的；

[0137] 在该步骤中预处理模型的构建是在离线部分完成的。

[0138] 步骤503，将多个原始特征向量进行拼接处理和加噪处理；

[0139] 步骤504，调用分布式特征提取模型对经过拼接处理和加噪处理后的结果进行特征提取，得到与待处理数据对应的分布式特征。

[0140] 在该步骤中，分布式特征提取模型构建也是在离线部分完成的。

[0141] 离线部分的处理流程可以包括：

[0142] 步骤5021，获取训练样本集，该训练样本集中每个样本包括多个不同的内容信息；

[0143] 在上述步骤中，获取资讯样本数据，可以对不同来源、不同类型的资讯数据进行收集作为资讯样本集。

[0144] 步骤5022，按照标准特征维度模板对每个样本进行内容抽取，得到每个样本的多个内容信息。

[0145] 在上述步骤中，特征预处理阶段是对资讯样本数据进行预处理。标准特征维度模板是预先定义了用于表征资讯数据的多个特征维度。当资讯样本数据中不包含标准特征维度模板规定的内容信息时，该特征维度对应的数据设置为0，该特征维度对应的权重系数对应的设置为0。例如资讯样本数据包括来源、标题、正文、正文长度等，不包括小视频。则小视频对应的数据设置为0，其权重系数也设置为0.

[0146] 步骤5023，利用训练样本集对待训练的预处理模型进行训练，得到预处理模型，该预处理模型包括与每个特征维度对应的预处理策略。

[0147] 在该步骤中，标准特征维度模板还可以规定与内容信息对应的预处理策略。例如内容信息为来源，则对其进行单词化处理和独热编码处理得到与内容特征对应的独热编码向量，预处理策略还包括预先构建的图像特征提取模型对包含图像的资讯数据进行特征提取。预处理模型可以包括多个数据映射关系。例如，预处理策略包括调用预先建立的词典对内容信息进行分词处理；其中词典可以是人工建立词库，通过词典匹配方式来完成分词处理。预处理策略还包括：对每个分词的频次进行统计，将分词对应的频次与过滤阈值进行比较。若分词的频次大于过滤阈值，则过滤该分词，若分词小于等于过滤阈值，则标记该分词为低频词。预处理策略还可以包括调用预先训练得到的图像提特征提取模型，对资讯数据出现的图片数据进行特征提取。其中，图像特征提取模型优选为VGG16的FC7层输出的结果为提取的图像特征。

[0148] 对于训练样本集中每个样本，按照步骤5022的方法提取得到多个内容信息，确定每个内容信息对应的预处理策略，建立内容信息与预处理策略之间的关联关系。然后，利用预处理策略对内容信息进行预处理。

[0149] 训练预处理模型和训练分布式特征提取模型是两个独立的过程。在预处理模型构建之后，调用预处理模型对样本数据集中每个样本进行预处理后，利用预处理的结果来训练分布式特征提取模型，包括：

[0150] 步骤5041，调用预处理模型对每个样本进行预处理，得到与每个样本相对应的多个样本原始特征向量，每个样本原始特征向量与每个样本所包含的一个内容信息相对应；

[0151] 步骤5042，将与每个样本对应的多个样本原始特征向量输入到待训练的分布式特征提取模型，输出与多个样本原始特征向量一一对应的重构特征向量和与每个样本对应的分布式特征。

[0152] 步骤5043，利用目标损失函数对待训练的分布式特征提取模型进行训练，直到目标损失函数达到最小值时训练完成，得到分布式特征提取模型，该目标损失函数是根据样本原始特征向量、重构特征向量和分布式特征定义的。

[0153] 在上述步骤中，训练模型阶段，首先定义模型结构和模型损失函数，即选定待训练的神经网络模型，确定模型损失函数为公式(1)所示。

[0154] 在完成上述定义之后，利用特征预处理阶段的输出结果按照定义的模型损失函数进行训练，直到模型损失函数达到最小值时完成训练，训练完成的分布式特征提取模型。

[0155] 下面以资讯数据数据为例，结合图6本申请实施例提供的分布式特征提取方法的工作原理进行描述。图6示出了本申请实施例提供的分布式特征提取方法的原理示意图。

[0156] 如图6所示，从待处理数据中提取分布式特征的过程可以包括两个阶段，第一阶段是离线阶段，在离线阶段利用海量资讯样本来训练构建分布式特征提取模型，即按照图6中实线箭头示意的处理流程，将海量资讯样本数据输入到特征预处理模型进行数据的预处理，按照模型训练算法利用特征预处理模型的输出结果对待训练的分布式特征提取模型进行训练，直到定义的目标损失函数达到最小值时，训练完成得到分布式特征提取模型。

[0157] 离线阶段的训练过程可以结合图7的内容来理解。请参考图7，图7示出了本申请实施例提供的构建分布式特征提取模型的原理示意图。

[0158] 如图7所示，在训练过程中通过按照标准特征维度模板中对每一条资讯数据进行内容提取，得到与每一条资讯数据对应的多个内容信息。例如，标准特征维度模板定义资讯类型、来源、标题、作者、标签(Tag)、一级类目、二级类目、正文、正文长度、组图、图片数、封面图、视频分辨率、视频长度等多个特征维度。每个特征维度对应一个样本原始特征向量。如图7的第一层(至下而上的方向的第一层)显示的是同一个样本的多个特征维度的样本原始特征向量。

[0159] 然后，对多个样本原始特征向量中的多热编码向量进行L2归一化处理。例如，图7中示出的对与标题对应的样本原始特征向量进行L2归一化处理，对与正文内容对应的样本原始特征向量进行L2归一化处理。

[0160] 然后，将所有的样本原始特征向量经过特征拼接处理和加噪处理后。将特征拼接处理和加噪处理后的结果输入到第一隐含层，第一隐含层定义第一权重矩阵和第一PReLU激活函数，第一隐含层将拼接处理和加噪处理结果与第一权重矩阵相乘后，再利用第一PreLU激活函数进行非线性处理，则输出分布式特征。在训练模型阶段，将分布式特征再输入到第二隐含层，第二隐含层定义第二权重矩阵和第二PReLU激活函数，第二隐含层将分布式特征与第二权重矩阵相乘之后，再利用第二PreLU激活函数进行非线性处理。第一隐含层和第二隐含层构成共享隐含层，将经过共享隐含层处理后的特征向量按照内容信息对应的重构方式进行处理，得到与内容信息对应的重构特征向量。

[0161] 如图7所示，在顶层中包括与第一层中多个原始特征向量一一对应的重构特征向量，如顶层中“来源”对应的重构特征向量，是将共享隐含层的输出结果乘以与“来源”对应的权重矩阵之后，再通过偏置值进行调整，即每个元素加上一个偏置值。

[0162] 然后，利用与“来源”对应的损失函数类型，如交叉熵损失函数，利用交叉熵损失函数计算“来源”的原始特征向量与重构特征向量之间的损失误差。

[0163] 利用分布式特征提取层输出的多个分布式特征计算分布式特征的损失误差。可选地，将分布式特征按照相似度划分成相似的分布式特征对，和不相似的分布式特征对。计算相似的分布式特征对之间的损失误差和不相似的分布式特征之间损失误差作为分布式特征的损失误差。

[0164] 假设分别对三条资讯数据进行分布式特征提取，得到与资讯数据对应的分布式特征h_1,h_2,h_3。其中，h_1,h_2是来自相同类型的来源，相同的二级类目，例如均属于足球类新闻，h_1,h_3是来相同类型的来源，不同的一级类目，如h_1是足球类新闻，h_3是时政类新闻，则针对分布式特征的损失函数计算按照如下公式计算：

[0165] $$\Phi(h_1,h_2,h_3)=\log(1+\exp(h_1^{T}h_3-h_1^{T}h_2)) \quad \text{公式(2)}$$

[0166] 本申请实施例，在目标损失函数中引入分布式特征之间的损失误差，在尽可能保

留原始特征的信息的基础上，保证了分布式特征的距离特性。

[0167] 在完成模型训练之后，调用训练得到的模型可以对资讯数据进行分布式特征提取。如图6所示，在线阶段的处理过程为图中示出的虚线箭头指示的处理流程，该流程包括：

[0168] 接收到新入库资讯数据；

[0169] 调用特征预处理模型对新入库资讯数据进行预处理，得到与新入库资讯数据对应的多个原始特征向量；

[0170] 调用分布式特征提取模型对多个原始特征向量进行特征提取，得到新入库资讯数据对应的分布式特征。

[0171] 在提取到资讯数据对应的分布式特征之后，将其存储，再提供给下游任务处理。其中下游服务场景包括推荐服务中召回、去重等；或者是广告推荐系统的CTR预估(Click-Through Rate Prediction)场景，CTR预估是互联网计算广告中的关键环节，预估准确性直接影响公司广告收入。

[0172] 本申请实施例提供的基于内容的分布式特征提取方法，可以针对不同类型资讯、不同来源的资讯数据进行统一的特征提取，有助于提高数据的处理效率，其利用分布式特征提取模型提取的分布式特征不仅大量保留原始特征向量的信息，还保证了向量空间中分布式特征的距离特性，利用分布式特征的距离特性有助于提高推荐结果的精准度。

[0173] 应当注意，尽管在附图中以特定顺序描述了本公开方法的操作，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。

[0174] 本申请实施例还提供了一种基于内容的分布式特征提取装置。参考图8，图8示出了根据本申请实施例提供的基于内容的分布式特征提取装置的示例性结构框图。如图8所示，该装置可以预先安装在终端设备或者服务器内。该装置包括：

[0175] 数据获取单元701，用于获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应；

[0176] 数据预处理单元702，用于对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的；

[0177] 特征提取单元703，用于调用分布式特征提取模型对多个原始特征向量进行特征提取，得到与待处理数据对应的分布式特征，该分布式特征是按照标准特征维度模板定义的多个特征维度对待处理数据进行表征的结果，标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围。

[0178] 可选地，数据获取单元701还用于：

[0179] 调用标准特征维度模板对待处理数据进行内容抽取，得到待处理数据所包含的多个不同维度的内容信息。

[0180] 数据预处理单元702还用于：

[0181] 策略调用子模块，调用与每个内容信息相对应的预处理策略对与内容信息进行预处理，得到与内容信息相对应的原始特征向量。

[0182] 策略调用子模块还用于：

[0183] 确定内容信息的数据类型；

[0184] 根据数据类型确定与数据类型对应的预处理策略；

[0185] 按照预处理策略将内容信息转换成与内容信息对应的原始特征向量。

[0186] 特征提取单元703还包括：

[0187] 处理子单元，用于将多个原始特征向量进行特征拼接处理和加噪处理；

[0188] 特征提取子单元，用于调用分布式特征提取模型对前述拼接处理和加噪处理后的结果进行特征提取，得到与待处理数据对应的分布式特征。

[0189] 其中，特征提取子单元，还可以用于将前述拼接处理和加噪处理后的结果与权重矩阵相乘，输出线性特征向量；利用激活函数对线性特征向量进行非线性处理，得到分布式特征。

[0190] 特征提取单元703，还包括：

[0191] 归一化子单元，用于对多个原始特征向量中所包含的多热编码向量分别进行归一化处理。

[0192] 应当理解，上述装置中记载的诸单元或模块与参考图2描述的方法中的各个步骤相对应。由此，上文针对方法描述的操作和特征同样适用于装置及其中包含的单元，在此不再赘述。装置可以预先实现在电子设备的浏览器或其他安全应用中，也可以通过下载等方式而加载到电子设备的浏览器或其安全应用中。装置中的相应单元可以与电子设备中的单元相互配合以实现本申请实施例的方案。

[0193] 在上文详细描述中提及的若干模块或者单元，这种划分并非强制性的。实际上，根据本公开的实施方式，上文描述的两个或更多模块或者单元的特征和功能可以在一个模块或者单元中具体化。反之，上文描述的一个模块或者单元的特征和功能可以进一步划分为由多个模块或者单元来具体化。

[0194] 下面参考图9，图9示出了适于用来实现本申请实施例的终端设备或服务器的计算机系统的结构示意图。

[0195] 如图9所示，计算机系统包括中央处理单元(CPU)801，其可以根据存储在只读存储器(ROM)802中的程序或者从存储部分808加载到随机访问存储器(RAM)803中的程序而执行各种适当的动作和处理。在RAM 803中，还存储有系统操作所需的各种程序和数据。CPU 801、ROM 802以及RAM 803通过总线804彼此相连。输入/输出(I/O)接口805也连接至总线804。

[0196] 以下部件连接至I/O接口805：包括键盘、鼠标等的输入部分806；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807；包括硬盘等的存储部分808；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809。通信部分809经由诸如因特网的网络执行通信处理。驱动器810也根据需要连接至I/O接口805。可拆卸介质811，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810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808。

[0197] 特别地，根据本公开的实施例，上文参考流程图图2描述的过程可以被实现为计算机软件程序。例如，本公开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809从网络上被下载和安装，和/或从可拆卸介质811被安装。在该计算机程序被中央处理单元(CPU)801执行时，执行本申请的系统中限定的

上述功能。

[0198] 需要说明的是，本公开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。在本公开中，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。而在本公开中，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。

[0199] 附图中的流程图和框图，图示了按照本公开各种实施例的系统、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。

[0200] 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：一种处理器包括数据获取单元、数据预处理单元以及特征提取单元。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，数据获取单元还可以被描述为“用于获取获取待处理数据所包含的多个不同维度的内容信息的单元”。

[0201] 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人工智能推荐模型的基于内容的分布式特征提取方法。

[0202] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离前述公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的(但不限于)具有类似功

能的技术特征进行互相替换而形成的技术方案。

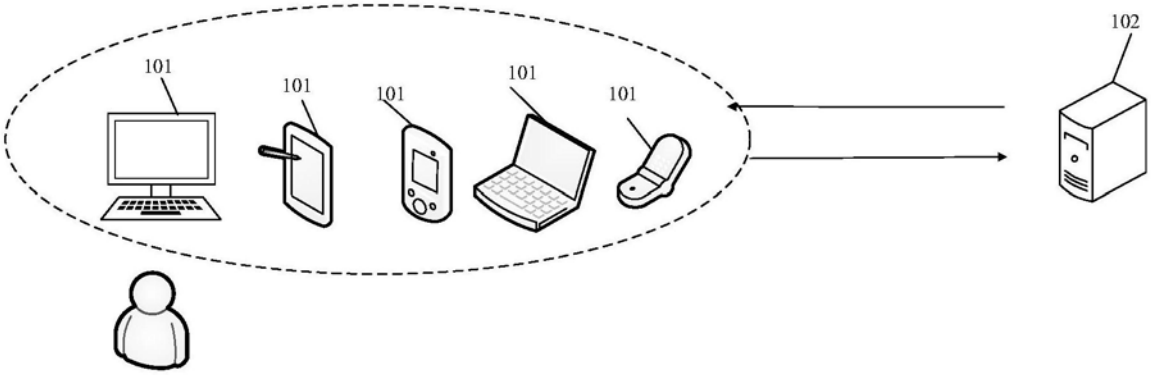

101
101
101
101
101
102

图1

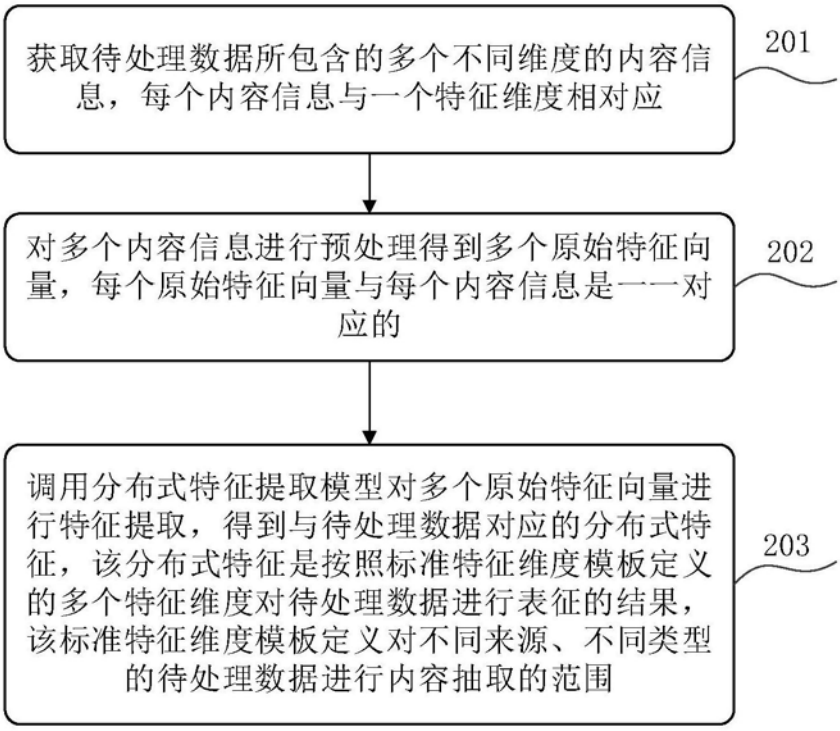

获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应
201
对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的
202
调用分布式特征提取模型对多个原始特征向量进行特征提取，得到与待处理数据对应的分布式特征，该分布式特征是按照标准特征维度模板定义的多个特征维度对待处理数据进行表征的结果，该标准特征维度模板定义对不同来源、不同类型的待处理数据进行内容抽取的范围
203

图2

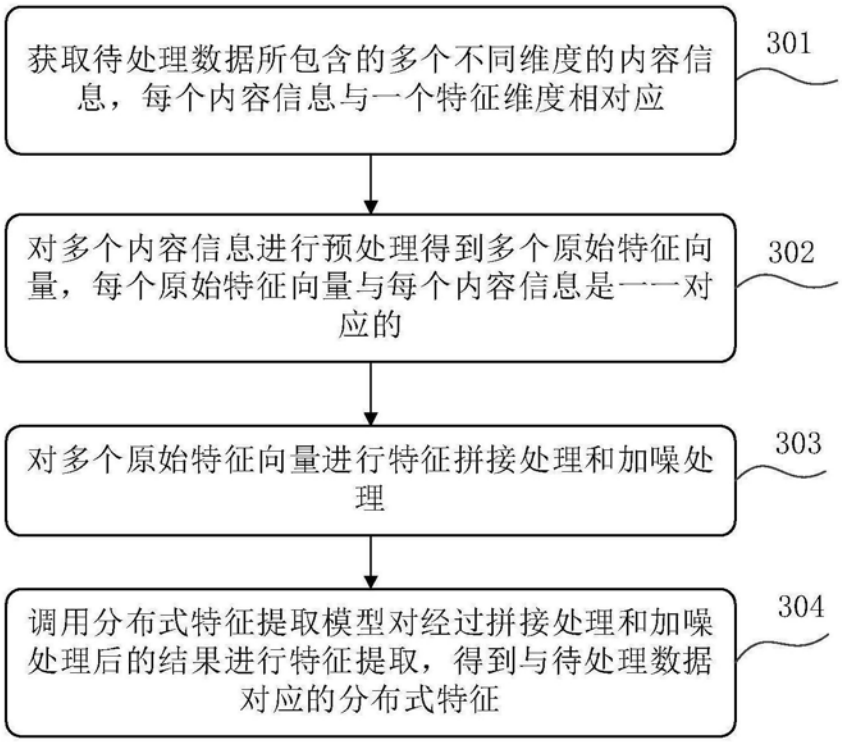

获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应
301
对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的
302
对多个原始特征向量进行特征拼接处理和加噪处理
303
调用分布式特征提取模型对经过拼接处理和加噪处理后的结果进行特征提取，得到与待处理数据对应的分布式特征
304

图3

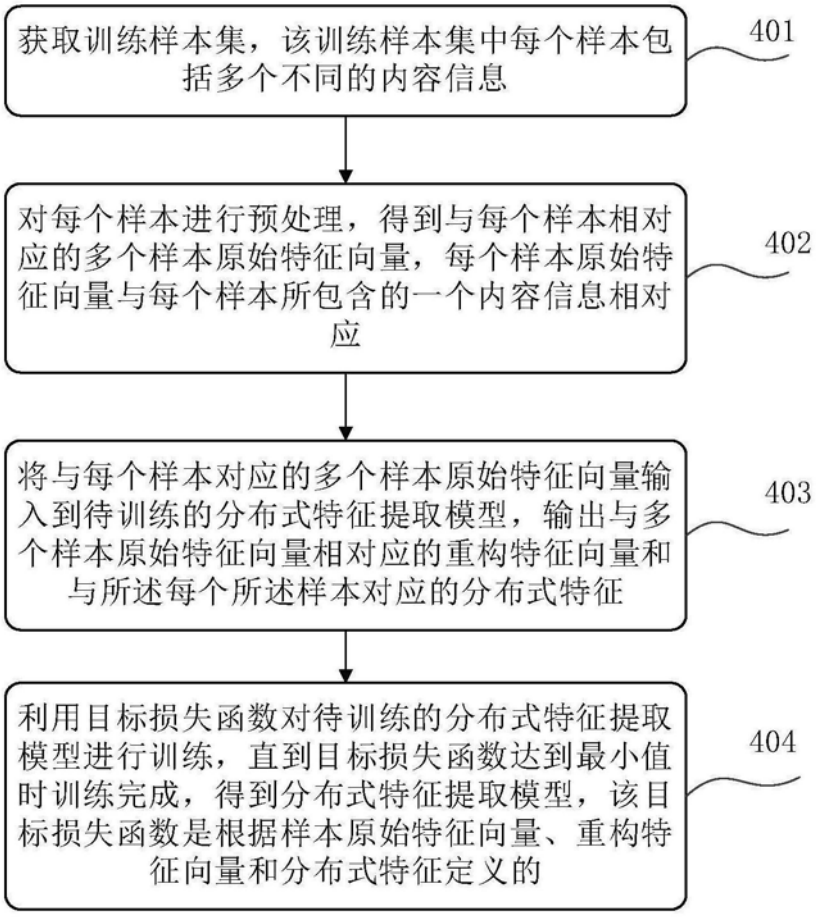

获取训练样本集，该训练样本集中每个样本包括多个不同的内容信息
401
对每个样本进行预处理，得到与每个样本相对应的多个样本原始特征向量，每个样本原始特征向量与每个样本所包含的一个内容信息相对应
402
将与每个样本对应的多个样本原始特征向量输入到待训练的分布式特征提取模型，输出与多个样本原始特征向量相对应的重构特征向量和与所述每个所述样本对应的分布式特征
403
利用目标损失函数对待训练的分布式特征提取模型进行训练，直到目标损失函数达到最小值时训练完成，得到分布式特征提取模型，该目标损失函数是根据样本原始特征向量、重构特征向量和分布式特征定义的
404

图4

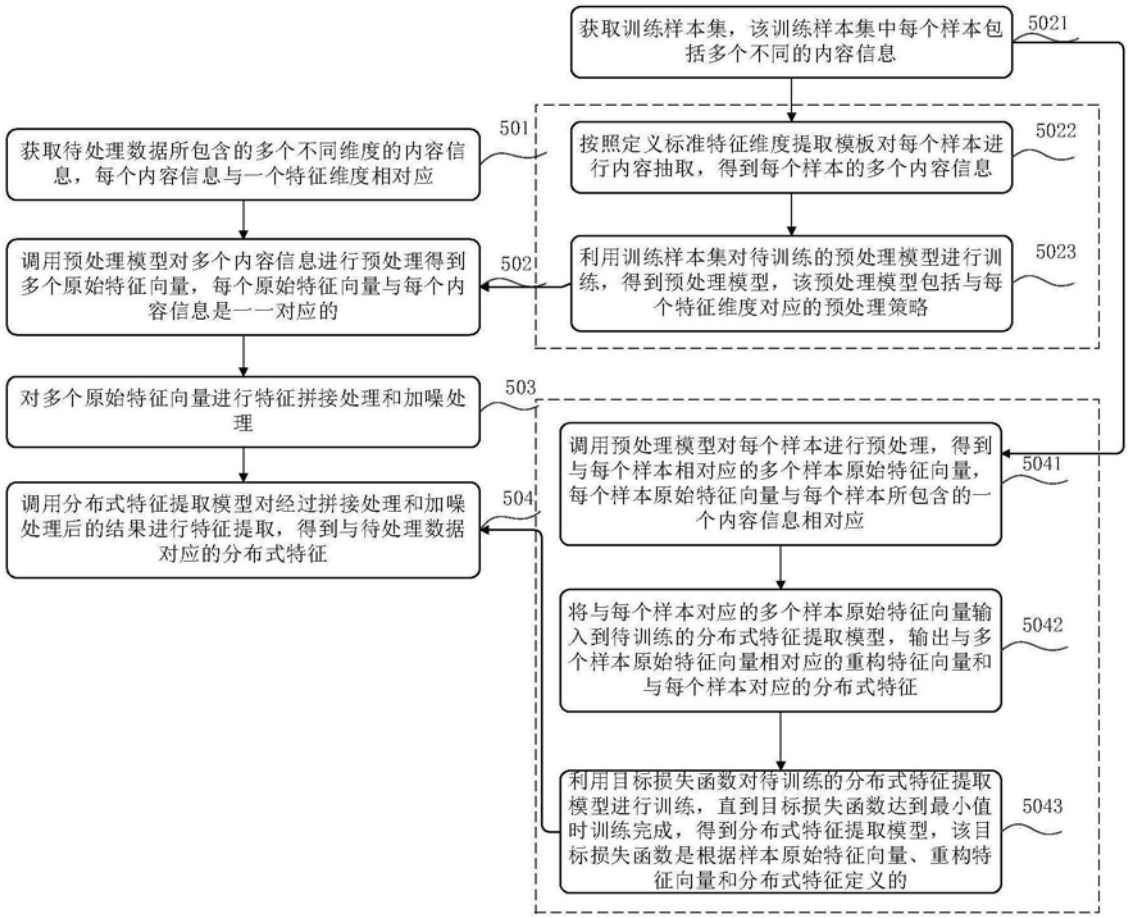

获取训练样本集，该训练样本集中每个样本包括多个不同的内容信息
5021
501
获取待处理数据所包含的多个不同维度的内容信息，每个内容信息与一个特征维度相对应
按照定义标准特征维度提取模板对每个样本进行内容抽取，得到每个样本的多个内容信息
5022
调用预处理模型对多个内容信息进行预处理得到多个原始特征向量，每个原始特征向量与每个内容信息是一一对应的
502
利用训练样本集对待训练的预处理模型进行训练，得到预处理模型，该预处理模型包括与每个特征维度对应的预处理策略
5023
对多个原始特征向量进行特征拼接处理和加噪处理
503
调用预处理模型对每个样本进行预处理，得到与每个样本相对应的多个样本原始特征向量，每个样本原始特征向量与每个样本所包含的一个内容信息相对应
5041
调用分布式特征提取模型对经过拼接处理和加噪处理后的结果进行特征提取，得到与待处理数据对应的分布式特征
504
将与每个样本对应的多个样本原始特征向量输入到待训练的分布式特征提取模型，输出与多个样本原始特征向量相对应的重构特征向量和与每个样本对应的分布式特征
5042
利用目标损失函数对待训练的分布式特征提取模型进行训练，直到目标损失函数达到最小值时训练完成，得到分布式特征提取模型，该目标损失函数是根据样本原始特征向量、重构特征向量和分布式特征定义的
5043

图5

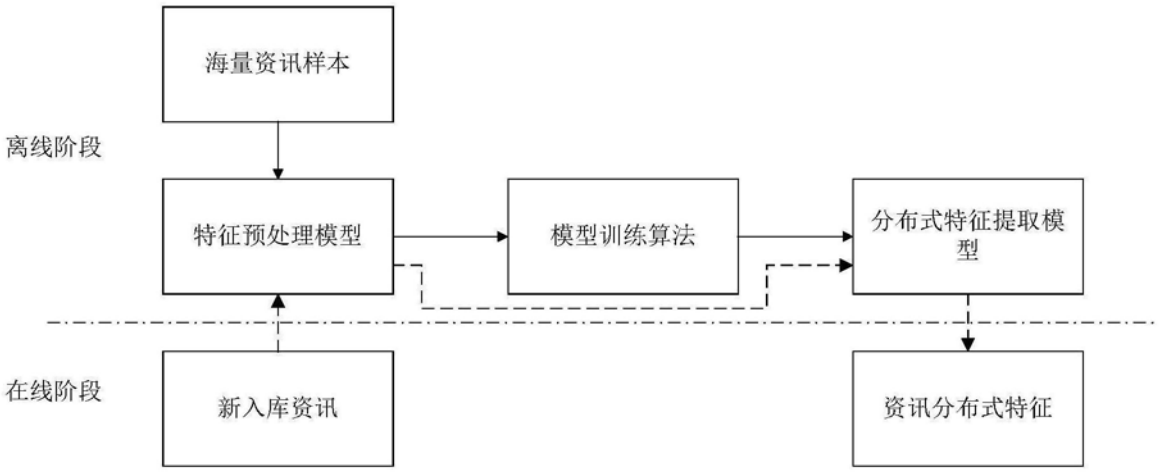

海量资讯样本
离线阶段
特征预处理模型
模型训练算法
分布式特征提取模型
在线阶段
新入库资讯
资讯分布式特征

图6

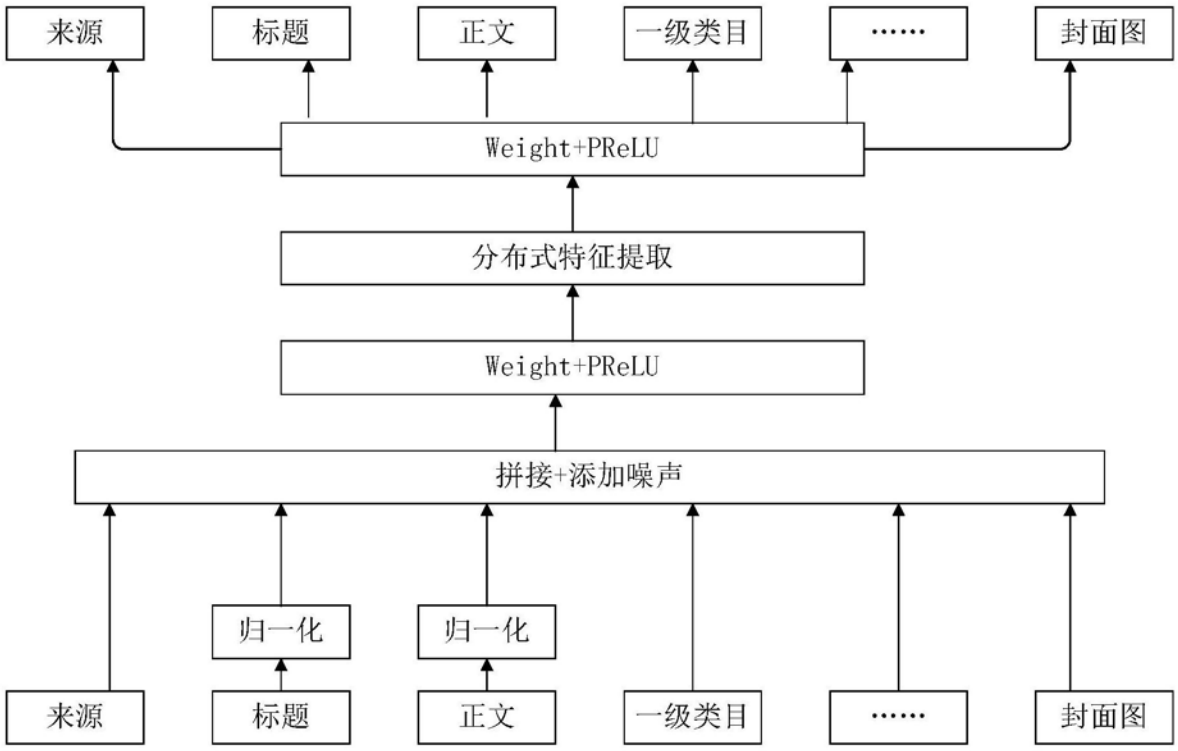

来源
标题
正文
一级类目
……
封面图
Weight+PReLU
分布式特征提取
Weight+PReLU
拼接+添加噪声
归一化
归一化
来源
标题
正文
一级类目
……
封面图

图7

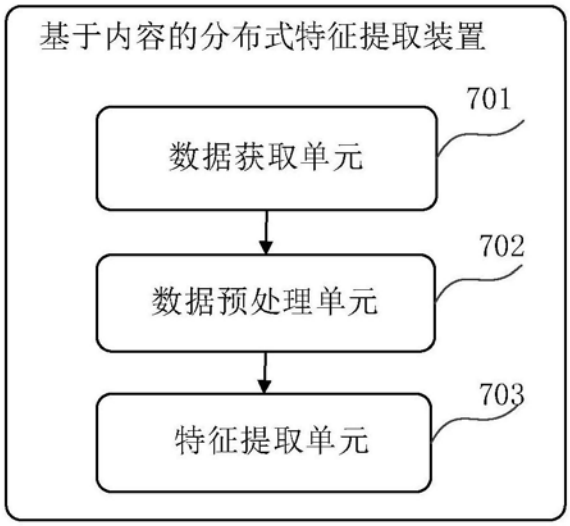

基于内容的分布式特征提取装置
701
数据获取单元
702
数据预处理单元
703
特征提取单元

图8

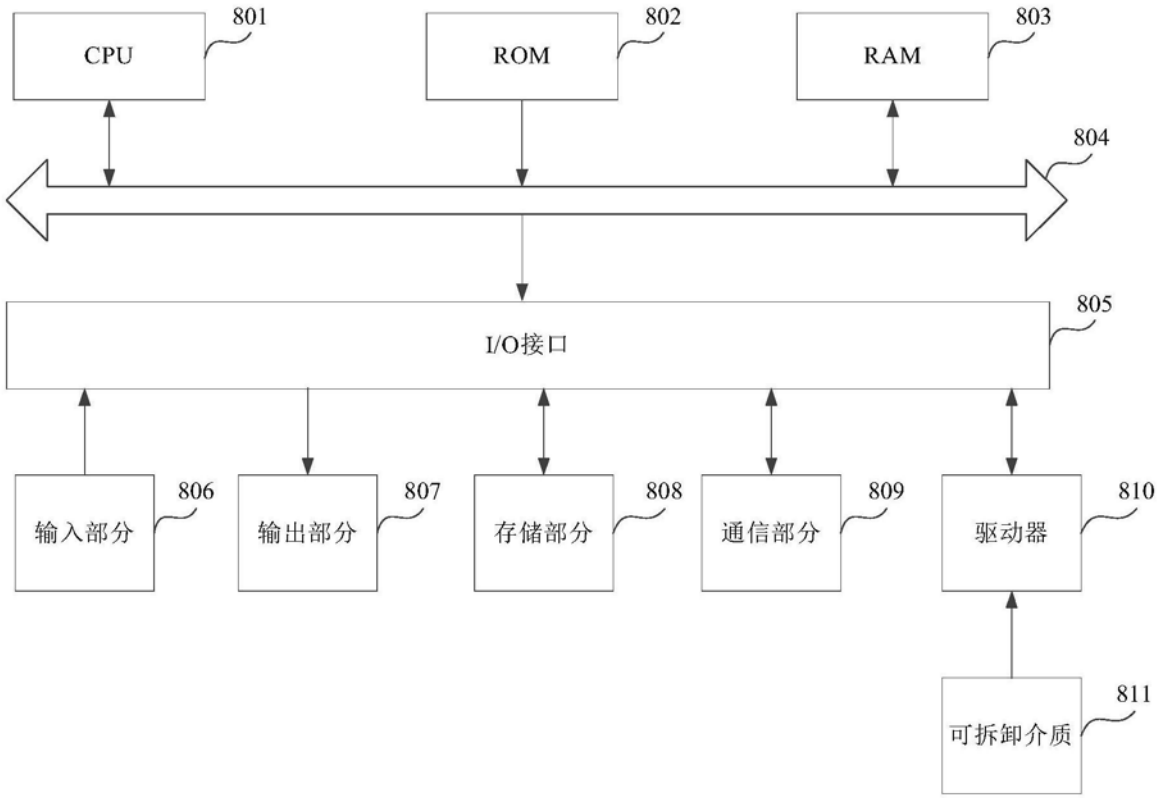

CPU
801
ROM
802
RAM
803
804
I/O接口
805
输入部分
806
输出部分
807
存储部分
808
通信部分
809
驱动器
810
可拆卸介质
811

图9