



(12) 发明专利

(10) 授权公告号 CN 110717536 B

(45) 授权公告日 2025. 04. 29

(21) 申请号 201910944345.5

G06Q 30/0202 (2023.01)

(22) 申请日 2019.09.30

G06Q 30/0201 (2023.01)

G06Q 30/0601 (2023.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 110717536 A

(56) 对比文件

CN 107203518 A, 2017.09.26

CN 107424043 A, 2017.12.01

(43) 申请公布日 2020.01.21

(73) 专利权人 北京三快在线科技有限公司

审查员 徐颖

地址 100080 北京市海淀区北四环西路9号
2106-030

(72) 发明人 张立文 程佳 张涛 柳明海

(74) 专利代理机构 北京曼威知识产权代理有限公司 11709

专利代理师 方志炜

(51) Int. Cl.

G06F 18/214 (2023.01)

G06F 16/9535 (2019.01)

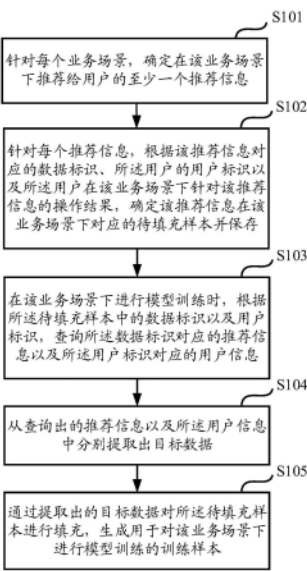
权利要求书2页 说明书11页 附图3页

(54) 发明名称

一种生成训练样本的方法及装置

(57) 摘要

本说明书公开了一种生成训练样本的方法及装置,在该方法中针对每个业务场景,确定推荐给用户的至少一个推荐信息,针对每个推荐信息,根据该推荐信息对应的数据标识、该用户的用户标识以及该用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存,在该业务场景下进行模型训练时,根据该待填充样本中的数据标识以及用户标识,查询所述数据标识对应的推荐信息以及该用户标识对应的用户信息,从查询出的推荐信息以及该用户信息中分别提取出目标数据,通过提取出的目标数据对该待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。这就给训练样本带来维护便利,提高训练样本生成效率。



1. 一种生成训练样本的方法,其特征在于,包括:

针对每个业务场景,确定在该业务场景下推荐给用户的至少一个推荐信息;

针对每个推荐信息,根据该推荐信息对应的数据标识、所述用户的用户标识以及所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存;

在该业务场景下进行模型训练时,根据所述待填充样本中的数据标识以及用户标识,查询所述数据标识对应的推荐信息以及所述用户标识对应的用户信息;

从查询出的推荐信息以及所述用户信息中分别提取出目标数据;

通过提取出的目标数据对所述待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

2. 如权利要求1所述的方法,其特征在于,根据该推荐信息对应的数据标识、所述用户的用户标识以及所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存,具体包括:

根据所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的样本标签;

确定所述用户的用户信息对应的数据类别作为第一数据类别,以及确定该推荐信息对应的数据类别作为第二数据类别;

根据所述样本标签、所述用户标识与所述第一数据类别的对应关系、所述数据标识与所述第二数据类别的对应关系,确定该推荐信息在该业务场景下对应的待填充样本并保存。

3. 如权利要求2所述的方法,其特征在于,从查询出的推荐信息以及所述用户信息中分别提取出目标数据,具体包括:

根据所述待填充样本中包含的第一数据类别,从所述用户信息中提取出所述第一数据类别所需的数据作为目标数据,根据所述待填充样本中包含的第二数据类别,从查询出的推荐信息中提取出所述第二数据类别所需的数据作为目标数据。

4. 如权利要求1或3所述的方法,其特征在于,通过提取出的目标数据对所述待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本,具体包括:

将从所述用户信息中提取出的目标数据作为所述用户标识对应的第一键值,将从查询出的推荐信息中提取出的目标数据作为所述数据标识对应的第二键值;

通过所述第一键值和所述第二键值,对该业务场景对应的待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

5. 如权利要求4所述的方法,其特征在于,所述方法还包括:

确定该业务场景下线上服务系统中设置的预测模型在该业务场景下线下测试系统中对应的镜像模型,作为待训练模型;

从所述训练样本所包含的目标数据中提取出所述待训练模型所需的特征数据;

根据所述特征数据以及所述训练样本中包含的样本标签,对所述待训练模型进行训练;

通过训练后的所述待训练模型,对所述预测模型进行更新。

6. 一种生成训练样本的装置,其特征在于,包括:

确定模块,用于针对每个业务场景,确定在该业务场景下推荐给用户的至少一个推荐信息;

保存模块,用于针对每个推荐信息,根据该推荐信息对应的数据标识、所述用户的用户标识以及所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存;

查询模块,用于在该业务场景下进行模型训练时,根据所述待填充样本中的数据标识以及用户标识,查询所述数据标识对应的推荐信息以及所述用户标识对应的用户信息;

提取模块,用于从查询出的推荐信息以及所述用户信息中分别提取出目标数据;

生成模块,用于通过提取出的目标数据对所述待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

7.如权利要求6所述的装置,其特征在于,所述保存模块具体用于,根据所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的样本标签;确定所述用户的用户信息对应的数据类别作为第一数据类别,以及确定该推荐信息对应的数据类别作为第二数据类别;根据所述样本标签、所述用户标识与所述第一数据类别的对应关系、所述数据标识与所述第二数据类别的对应关系,确定该推荐信息在该业务场景下对应的待填充样本并保存。

8.如权利要求7所述的装置,其特征在于,所述提取模块具体用于,根据所述待填充样本中包含的第一数据类别,从所述用户信息中提取出所述第一数据类别所需的数据作为目标数据,根据所述待填充样本中包含的第二数据类别,从查询出的推荐信息中提取出所述第二数据类别所需的数据作为目标数据。

9.一种计算机可读存储介质,其特征在于,所述存储介质存储有计算机程序,所述计算机程序被处理器执行时实现上述权利要求1~5任一项所述的方法。

10.一种电子设备,包括存储器、处理器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上述权利要求1~5任一项所述的方法。

一种生成训练样本的方法及装置

技术领域

[0001] 本说明书涉及计算机领域,尤其涉及一种生成训练样本的方法及装置。

背景技术

[0002] 为了能够给用户提供更好的业务体验,各业务平台可以基于用户的历史业务记录、历史浏览记录等信息,分析出用户的业务喜好和习惯,以向用户进行信息推荐。

[0003] 通常情况下,业务平台可以通过针对业务场景设置的预测模型,实现在该业务场景下的信息推荐。而为了能够使预测模型能够适应业务变化,业务平台需要基于业务日志,对预测模型进行训练,已实现对预测模型的更新。具体的,业务平台可以从业务日志中提取出特征数据,并将该特征数据输入到预测模型中,得到预测结果,并将通过该预测结果确定出的推荐信息推荐给用户,同时监测用户基于该推荐信息所执行的操作结果。业务平台后续可以基于该推荐信息以及该操作结果,对该预测模型进行训练,实现对该预测模型的更新。

[0004] 然而在现有技术中,不同业务场景的训练样本往往都是按照各自业务场景的训练样本生成方式来得到的,这就极大的增加了训练样本的维护困难。所以,如何能够降低训练样本的维护困难,则是一个亟待解决的问题。

发明内容

[0005] 本说明书提供一种生成训练样本的方法及装置,以部分的解决现有技术存在的上述问题。

[0006] 本说明书采用下述技术方案:

[0007] 本说明书提供了一种生成训练样本的方法,包括:

[0008] 针对每个业务场景,确定在该业务场景下推荐给用户的至少一个推荐信息;

[0009] 针对每个推荐信息,根据该推荐信息对应的数据标识、所述用户的用户标识以及所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存;

[0010] 在该业务场景下进行模型训练时,根据所述待填充样本中的数据标识以及用户标识,查询所述数据标识对应的推荐信息以及所述用户标识对应的用户信息;

[0011] 从查询出的推荐信息以及所述用户信息中分别提取出目标数据;

[0012] 通过提取出的目标数据对所述待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

[0013] 可选地,根据该推荐信息对应的数据标识、所述用户的用户标识以及所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存,具体包括:

[0014] 根据所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的样本标签;

[0015] 确定所述用户的用户信息对应的数据类别作为第一数据类别,以及确定该推荐信息对应的数据类别作为第二数据类别;

[0016] 根据所述样本标签、所述用户标识与所述第一数据类别的对应关系、所述数据标识与所述第二数据类别的对应关系,确定该推荐信息在该业务场景下对应的待填充样本并保存。

[0017] 可选地,从查询出的推荐信息以及所述用户信息中分别提取出目标数据,具体包括:

[0018] 根据所述待填充样本中包含的第一数据类别,从所述用户信息中提取出所述第一数据类别所需的数据作为目标数据,根据所述待填充样本中包含的第二数据类别,从查询出的推荐信息中提取出所述第二数据类别所需的数据作为目标数据。

[0019] 可选地,通过提取出的目标数据对所述待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本,具体包括:

[0020] 将从所述用户信息中提取出的目标数据作为所述用户标识对应的第一键值,将从查询出的推荐信息中提取出的目标数据作为所述数据标识对应的第二键值;

[0021] 通过所述第一键值和所述第二键值,对该业务场景对应的待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

[0022] 可选地,所述方法还包括:

[0023] 确定该业务场景下线上服务系统中设置的预测模型在该业务场景下线下测试系统中对应的镜像模型,作为待训练模型;

[0024] 从所述训练样本所包含的目标数据中提取出所述待训练模型所需的特征数据;

[0025] 根据所述特征数据以及所述训练样本中包含的样本标签,对所述待训练模型进行训练;

[0026] 通过训练后的所述待训练模型,对所述预测模型进行更新。

[0027] 本说明书提供了一种训练样本的实时生成装置,包括:

[0028] 确定模块,用于针对每个业务场景,确定在该业务场景下推荐给用户的至少一个推荐信息;

[0029] 保存模块,用于针对每个推荐信息,根据该推荐信息对应的数据标识、所述用户的用户标识以及所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存;

[0030] 查询模块,用于在该业务场景下进行模型训练时,根据所述待填充样本中的数据标识以及用户标识,查询所述数据标识对应的推荐信息以及所述用户标识对应的用户信息;

[0031] 提取模块,用于从查询出的推荐信息以及所述用户信息中分别提取出目标数据;

[0032] 生成模块,用于通过提取出的目标数据对所述待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

[0033] 可选地,所述保存模块具体用于,根据所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的样本标签;确定所述用户的用户信息对应的数据类别作为第一数据类别,以及确定该推荐信息对应的数据类别作为第二数据类别;根据所述样本标签、所述用户标识与所述第一数据类别的对应关系、所述数据标识与所

述第二数据类别的对应关系,确定该推荐信息在该业务场景下对应的待填充样本并保存。

[0034] 可选地,所述提取模块具体用于,根据所述待填充样本中包含的第一数据类别,从所述用户信息中提取出所述第一数据类别所需的数据作为目标数据,根据所述待填充样本中包含的第二数据类别,从查询出的推荐信息中提取出所述第二数据类别所需的数据作为目标数据。

[0035] 本说明书提供了一种计算机可读存储介质,所述存储介质存储有计算机程序,所述计算机程序被处理器执行时实现上述生成训练样本的方法。

[0036] 本说明书提供了一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,所述处理器执行所述程序时实现上述生成训练样本的方法。

[0037] 本说明书采用的上述至少一个技术方案能够达到以下有益效果:

[0038] 在本说明书提供生成训练样本的方法中,可以针对每个业务场景,确定推荐给用户的至少一个推荐信息,针对每个推荐信息,根据该推荐信息对应的数据标识、该用户的用户标识以及该用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存,在对该业务场景下进行模型训练时,根据该待填充样本中的数据标识以及用户标识,查询所述数据标识对应的推荐信息以及该用户标识对应的用户信息,从查询出的推荐信息以及该用户信息中分别提取出目标数据,通过提取出的目标数据对该待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

[0039] 从上述方法中可以看出,由于对于每个业务场景,可以根据推荐信息以及用户在该业务场景下针对该推荐信息的操作结果,确定出该推荐信息在该业务场景下对应的待填充样本并保存,后续需要进行模型训练时,可以对保存的各待填充样本进行填充,从而生成用于对该业务场景下模型训练的训练样本。这样就实现了按照统一的方式维护不同业务场景的训练样本,给训练样本的维护带来了极大的便利,并且,极大的提高了训练样本的生成效率。

附图说明

[0040] 此处所说明的附图用来提供对本说明书的进一步理解,构成本说明书的一部分,本说明书的示意性实施例及其说明用于解释本说明书,并不构成对本说明书的不当限定。在附图中:

[0041] 图1为本说明书提供的一种生成训练样本的流程示意图;

[0042] 图2为本说明书提供的确定待填充样本的示意图;

[0043] 图3为本说明书提供的服务器进行训练样本填充的过程示意图;

[0044] 图4为本说明书提供的一种生成训练样本的装置示意图;

[0045] 图5为本说明书提供的对应于图1的电子设备示意图。

具体实施方式

[0046] 为使本说明书的目的、技术方案和优点更加清楚,下面将结合本说明书具体实施例及相应的附图对本说明书中技术方案进行清楚、完整地描述。显然,所描述的实施例仅是本说明书一部分实施例,而不是全部的实施例。基于本说明书中的实施例,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,都属于本说明书保护的范

围。

[0047] 以下结合附图,详细说明本说明书各实施例提供的技术方案。

[0048] 图1为本说明书提供的一种生成训练样本的流程示意图,具体包括以下步骤:

[0049] S101:针对每个业务场景,确定在该业务场景下推荐给用户的至少一个推荐信息。

[0050] 在本说明书中,在生成用于进行模型训练的训练样本之前,可以确定出推荐给用户的至少一个推荐信息。其中,确定推荐信息的执行主体可以是服务器,也可以是诸如电脑等终端设备。为了便于描述,下面将仅以服务器是执行主体为例,对本说明书提供的生成训练样本的方法进行说明。

[0051] 服务器可以基于用户的用户标识,确定出需要向该用户推荐的至少一个推荐信息。其中,这里提到的基于用户的用户标识向用户进行信息推荐,可以是指用户服务器发送携带有该用户的用户标识的业务请求,服务器可以根据该用户标识,确定出能够反映出该用户的喜好、兴趣、购物习惯等信息的属性信息(如用户画像),进而可以基于该用户的属性信息,通过设置于服务器中的预测模型,确定出需要向该用户进行推荐的推荐信息。

[0052] 再例如,当用户启动终端中安装的应用(Application,App)时,终端可以基于用户在该App中登录的用户账号,向该App对应的服务器发送携带有用户标识的业务请求(该用户标识可以是指该用户账号,也可以是指通过该用户账号查询到的诸如用户的手机号、身份证号等标识信息)。相应的,服务器可以根据接收到的业务请求中携带的用户标识,查询出该用户的属性信息,进而基于该用户的属性信息,通过预测模型向该用户进行信息推荐。其他的具体形式在此就不一一举例说明了。

[0053] 需要说明的是,在实际应用中,同一信息可以作为不同业务场景的推荐信息推荐给用户。而由于不同业务场景下,用于向用户进行信息推荐所使用的模型不尽相同,所以,在本说明书中,对应同一推荐信息,需要针对每个业务场景,确定出适用于该业务场景下所使用的模型的训练样本。即,对于同一推荐信息来说,该推荐信息在不同业务场景下所对应的训练样本可能是不同的。

[0054] S102:针对每个推荐信息,根据该推荐信息对应的数据标识、所述用户的用户标识以及所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存。

[0055] 针对每个业务场景,服务器确定出向该用户发送的至少一个推荐信息后,可以进一步地获取到该用户在该业务场景下针对这些推荐信息的操作结果。其中,该操作结果可以反映出用户在该业务场景下针对这些推荐信息的具体操作情况。如,对于该用户获取到的每个推荐信息,该操作结果中记录了该用户是否对该推荐信息进行了诸如点击查看、具体的浏览时间等信息。

[0056] 服务器可以针对确定出的每个推荐信息,根据该推荐信息对应的数据标识、该用户的用户标识以及该用户针对该推荐信息的操作结果,确定出该推荐信息在该业务场景下对应的待填充样本。这里提到的该推荐信息对应的数据标识用于唯一标识该推荐信息,而用户针对该推荐信息的操作结果主要用于确定出该推荐信息所对应的样本标签。

[0057] 具体的,若是通过该操作结果,确定出用户通过执行诸如点击、按压等指定操作查看了该推荐信息,则可以确定出该推荐信息在该业务场景下对应的样本标签为正样本,否则,可以确定出该推荐信息在该业务场景下对应的样本标签为负样本。换句话说,服务器可

以通过该操作结果,对该推荐信息进行正负样本的标注。其中,这里提到的用户通过执行指定操作查看了该推荐信息,并不是指用户在推荐信息列表中查看了该推荐信息,而是指用户通过执行指定操作,打开了该推荐信息并查看到该推荐信息的具体内容。

[0058] 由于同一信息可以作为多个业务场景的推荐信息推荐给用户,而对于同一推荐信息来说,用户在不同业务场景下对该推荐信息的操作结果可能不尽相同。例如,对于一个运动鞋的推荐信息来说,如果在运动特卖会的业务场景下,用户查看该推荐信息的可能性较大,而在冬季服饰特卖会的业务场景下,用户查看该推荐信息的可能性将大大降低。

[0059] 基于此,由于用户在不同业务场景下对该推荐信息的操作结果不尽相同,所以,该推荐信息在不同业务场景下对应的样本标签也不尽相同,所以,服务器针对该推荐信息在不同业务场景下所生成的训练样本也不尽相同。

[0060] 在本说明书中,服务器可以按照预设的数据格式,根据该推荐信息对应的数据标识、该用户的用户标识,以及该用户在该业务场景下针对该推荐信息的操作结果,确定出该推荐信息在该业务场景下对应的待填充样本。具体的,服务器可以根据预先保存的各信息与各数据类别的对应关系,确定出该用户的用户信息对应的数据类别为第一数据类别,以及确定出该推荐信息对应的数据类别为第二数据类别。而后,根据确定出的该推荐信息在该业务场景下对应的样本标签,该用户的用户标识与该第一数据类别的对应关系(由于用户信息属于该第一数据类别,则该第一数据类别与该用户的用户标识相对应),以及该推荐信息的数据标识与该第二数据类别的对应关系(用于该推荐信息属于该第二数据类别,所以,该第二数据类别与该推荐信息的数据标识相对应),确定出该推荐信息在该业务场景下对应的待填充样本。

[0061] 其中,第二数据类别中包含有多种数据类别,例如,推荐信息为广告时,属于广告数据类别,推荐信息为上下文关联信息时,该推荐信息属于上下文数据类别。而这里提到的广告数据类别和上下文数据类别均可以称之为是第二数据类别。服务器确定待填充样本具体如图2所示。

[0062] 图2为本说明书提供的确定待填充样本的示意图。

[0063] 在图2中,服务器根据用户针对该推荐信息的操作结果,确定出该推荐信息在该业务场景下对应的训练样本为正样本,在可以在图2所示的待填充样本中positive的值置为1,即,positive的值为1时,表明该推荐信息在该业务场景下对应的训练样本为正样本。服务器可以根据预先保存的各信息与各数据类别的对应关系,确定出用户的用户信息所属的数据类别为数据类别11,则可以将该用户的用户标识添加在数据类别11,即,将用户标识: id=193475132添加在type=11处。同理,服务器确定出推荐信息为广告时,可以确定出该推荐信息所属的数据类别为数据类别13,则可以将该推荐信息的数据标识添加在数据类别13处,即,将数据标识: id=652742951添加在type=13处。而该推荐信息还属于上下文关联信息,与之关联的信息所遇上下文数据类别,其对应的数据类别为数据类别12,则服务器可以将与该推荐信息存在上下文关联的信息的数据标识添加在数据类别12处,即,将数据标识: id=456328943添加在type=12处。其中,这里提到的数据类别12和数据类别13均为第二数据类别。

[0064] 在图2中,待填充样本的前段50007用于唯一标识该待填充样本,而从图2中可以看出,由于该待填充样本中已经标注出了其所对应的训练样本为正样本还是负样本,并且,该

待填充样本中已经记录了该用户的用户标识以及推荐信息的数据标识,这样在后续生成训练样本的过程中,只需要按照该待填充样本中记录的用户标识和数据标识,查询出相应的数据,并完成待填充样本与查询出的数据之间的组合,即可得到该推荐信息在该业务场景下的训练样本。

[0065] 上述提到的上下文关联信息是指用户查看了一个推荐信息这一情况的发生是因为在查看了另一个推荐信息而发生的。例如,用户在查看推荐信息A的过程中,看到了推荐信息B,并对推荐信息B进行了查看,则推荐信息A和B属于上下文关联信息。

[0066] 服务器在确定出待填充样本后,可以将待填充样本进行保存。而由于待填充样本中并不是训练样本的全量数据,而是记录了用于查询这些数据的数据标识以及用户标识,这样就极大的节省了服务器的存储空间。

[0067] S103:在该业务场景下进行模型训练时,根据所述待填充样本中的数据标识以及用户标识,查询所述数据标识对应的推荐信息以及所述用户标识对应的用户信息。

[0068] S104:从查询出的推荐信息以及所述用户信息中分别提取出目标数据。

[0069] 为了保证能够通过预测模型在业务执行过程中向用户提供准确的推荐信息,需要对不同业务场景的预测模型进行训练。而针对每个业务场景来说,服务器在对该业务场景的预测模型进行训练时,可以根据预先保存的该业务场景的待填充样本,查询出待填充样本所需的信息,进而在后续过程中,生成用于对该业务场景的预测模型进行训练的训练样本。

[0070] 具体的,对于该业务场景下任意的待填充样本来说,服务器可以根据该待填充样本中记录的用户标识以及数据标识,查询出与该用户标识对应的用户的用户信息,以及与该数据标识对应的推荐信息。

[0071] 服务器中预先记录了各数据类别所需的数据,因此,由于用户信息与第一数据类别相对应,所以,服务器可以按照该第一数据类别,从该用户信息中提取出该第一数据类别所需的数据作为目标数据。同理,服务器可以按照待填充样本中的第二数据类别,从推荐信息中提取出该第二数据类别所需的数据作为目标数据。

[0072] 需要说明的是,服务器提取出的目标数据可以理解成:服务器将各业务场景的预测模型可能需要的数据全部提取出来,这样一来,即使后续预测模型所需的特征维度发生了一些变化,服务器也可以从该目标数据中提取出相应的特征数据,从而提高了特征数据的提取效率。并且,由于该目标数据为服务器提取出的预测模型可能需要的特征维度的数据,也就是说,服务器将用户信息中一些对于模型信息无用的信息进行了剔除,因此,可以进一步地降低生成的训练样本在服务器中所占用的存储空间,节省了服务器的存储资源。

[0073] S105:通过提取出的目标数据对所述待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

[0074] 在从用户信息以及推荐信息中提取出相应的目标数据后,服务器可以通过提取出的目标数据对该待填充样本进行填充,从而得到相应的训练样本。具体的,在本说明书中,服务器可以将从用户信息中提取出的目标数据作为该待填充样本中用户标识对应的第一键值,并且,将从查询出的推荐信息中提取出的目标数据作为待填充样本中数据标识对应的第二键值。而后,服务器可以根据该第一键值,在该待填充样本中用于存储用户信息的位置处进行填充,同时,可以根据该第二键值,在该待填充样本中用于存储推荐信息的位置处

进行填充,从而得到训练样本,如图3所示。

[0075] 图3为本说明书提供的服务器进行训练样本填充的过程示意图。

[0076] 服务器在对图3所示的待填充样本进行填充时,可以先根据该待填充样本中记录的用户标识以及数据标识,查询出相应的信息。其中,图3中示出的User Entity为用户信息实体,即表示服务器根据该待填充样本中记录的用户标识所查询出的用户信息。Ctx Poi Entity为上下文关联信息实体,即表示服务器根据待填充样本中记录的上下文关联信息的数据标识所查询出的推荐信息。AdEntity为广告信息实体,即表示服务器根据该待填充样本中记录的数据标识,所查询出的作为推荐信息的广告信息。

[0077] 服务器查询出用户信息以及各推荐信息后,可以分别从用户信息以及推荐信息中提取出目标数据,并将提取出的目标数据按照用户信息以及推荐信息所属的数据类别,填充在该待填充样本的相应位置上。其中,用户信息属于数据类别11,则服务器可以将从用户信息中提取出的目标数据填充在待填充样本中数据类别11所对应的位置上。同理,上下文关联信息和作为推荐信息的广告信息分别属于数据类别12和13,服务器可以将从上下文关联信息和广告信息中分别提取出的目标数据分别填充在这两个数据类别相应的位置上,从而得到训练样本。

[0078] 从上述方法中可以看出,由于针对每个业务场景,可以根据推荐信息以及用户在该业务场景下针对该推荐信息的操作结果,确定出该推荐信息在该业务场景下对应的待填充样本并保存,后续需要进行模型训练时,可以对保存的各待填充样本进行填充,从而生成用于对该业务场景下模型训练的训练样本。这样就实现了按照统一的方式维护不同业务场景的训练样本,给训练样本的维护带来了极大的便利,并且,极大的提高了训练样本的生成效率。

[0079] 针对每个业务场景,服务器在生成用于训练该业务场景的预测模型的训练样本后,可以将该业务场景的预测模型作为待训练模型,并从该训练样本所包含的目标数据中提取出该待训练模型所需的特征数据。而后,服务器可以根据提取出的特征数据以及该训练样本包含的样本标签,对该待训练模型进行训练。

[0080] 需要说明的是,在本说明书中,一个业务场景对应线上服务系统和线下测试系统,其中,线上服务系统用于对用户在该业务场景下进行信息推荐,而线下测试系统用于对线上服务系统进行维护进行更新。具体的,线上服务系统中设有预测模型,线上服务系统中的服务器需要通过该预测模型,向用户发送推荐信息。而在线下测试系统中设有与线上服务系统中预测模型相对应的镜像模型,该镜像模型的作用在于,可以通过对该镜像模型进行训练,实现对线上服务系统中预测模型各项参数的更新。

[0081] 基于此,在本说明书中,线上服务系统向用户进行信息推荐后,可以将该用户的用户标识以及所发送的推荐信息的数据标识发送给线下测试系统,即,不必向线下测试系统发送用户信息和推荐信息的完整数据,从而极大的节省了数据带宽。线下测试系统在获取到用户标识和数据标识后,可以根据获取到的该用户在该业务场景下针对该推荐信息的操作结果,确定出后续生成的待填充样本的样本标签,进而根据该样本标签、用户标识以及数据标号,生成相应的待填充样本。测试系统在生成该待填充样本对应的训练样本时,可以通过该待填充样本中记录的用户标识以及数据标识,查询出相应的用户信息和推荐信息,进而基于查询到的信息,对该待填充样本进行填充,从而得到用于对线下测试系统中设置的

镜像模型进行训练的训练样本。即,上述提到的在该业务场景下进行模型训练,可以是指对线下测试系统中的镜像模型进行训练。而训练后的镜像模型可以用于对该业务场景下线上服务系统中设置的预测模型进行更新。

[0082] 在本说明书中,一个业务场景在线上服务系统中可能设有多个预测模型,同理,在线下测试系统中也将设有该业务场景下的多个镜像模型。而需要指出的是,由于服务器在从查询出的用户信息和推荐信息所提取出的目标数据,可以是该业务场景下所有预测模型所需特征维度的数据,所以,对于同一业务场景来说,无论具体采用该业务场景的哪个预测模型对用户进行信息推荐,服务器生成的该业务场景的所有训练样本,均可以用于对该业务场景所有的预测模型在线下测试系统中对应的镜像模型进行训练。

[0083] 以上为本说明书的一个或多个实施例提供的生成训练样本的方法,基于同样的思路,本说明书还提供了相应的生成训练样本的装置,如图4所示。

[0084] 图4为本说明书提供的一种生成训练样本的装置示意图,具体包括:

[0085] 确定模块401,用于针对每个业务场景,确定在该业务场景下推荐给用户的至少一个推荐信息;

[0086] 保存模块402,用于针对每个推荐信息,根据该推荐信息对应的数据标识、所述用户的用户标识以及所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的待填充样本并保存;

[0087] 查询模块403,用于在该业务场景下进行模型训练时,根据所述待填充样本中的数据标识以及用户标识,查询所述数据标识对应的推荐信息以及所述用户标识对应的用户信息;

[0088] 提取模块404,用于从查询出的推荐信息以及所述用户信息中分别提取出目标数据;

[0089] 生成模块405,用于通过提取出的目标数据对所述待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

[0090] 可选地,所述保存模块402具体用于,根据所述用户在该业务场景下针对该推荐信息的操作结果,确定该推荐信息在该业务场景下对应的样本标签;确定所述用户的用户信息对应的数据类别作为第一数据类别,以及确定该推荐信息对应的数据类别作为第二数据类别;根据所述样本标签、所述用户标识与所述第一数据类别的对应关系、所述数据标识与所述第二数据类别的对应关系,确定该推荐信息在该业务场景下对应的待填充样本并保存。

[0091] 可选地,所述提取模块404具体用于,根据所述待填充样本中包含的第一数据类别,从所述用户信息中提取出所述第一数据类别所需的数据作为目标数据,根据所述待填充样本中包含的第二数据类别,从查询出的推荐信息中提取出所述第二数据类别所需的数据作为目标数据。

[0092] 可选地,所述生成模块405具体用于,将从所述用户信息中提取出的目标数据作为所述用户标识对应的第一键值,将从查询出的推荐信息中提取出的目标数据作为所述数据标识对应的第二键值;通过所述第一键值和所述第二键值,对该业务场景对应的待填充样本进行填充,生成用于对该业务场景下进行模型训练的训练样本。

[0093] 可选地,所述装置还包括:

[0094] 训练模块406,用于针对每个业务场景,确定该业务场景下线上服务系统中设置的预测模型在该业务场景下线下测试系统中对应的镜像模型,作为待训练模型;从所述训练样本所包含的目标数据中提取出所述待训练模型所需的特征数据;根据所述特征数据以及所述训练样本中包含的样本标签,对所述待训练模型进行训练;

[0095] 通过训练后的所述待训练模型,对所述预测模型进行更新。

[0096] 本说明书还提供了一种计算机可读存储介质,该存储介质存储有计算机程序,计算机程序可用于执行上述生成训练样本的方法。

[0097] 本说明书还提供了图5所示的电子设备的示意结构图。如图5所述,在硬件层面,该电子设备包括处理器、内部总线、网络接口、内存以及非易失性存储器,当然还可能包括其他业务所需要的硬件。处理器从非易失性存储器中读取对应的计算机程序到内存中然后运行,以实现上述生成训练样本的方法。当然,除了软件实现方式之外,本说明书并不排除其他实现方式,比如逻辑器件抑或软硬件结合的方式等等,也就是说以下处理流程的执行主体并不限定于各个逻辑单元,也可以是硬件或逻辑器件。

[0098] 在20世纪90年代,对于一个技术的改进可以很明显地区分是硬件上的改进(例如,对二极管、晶体管、开关等电路结构的改进)还是软件上的改进(对于方法流程的改进)。然而,随着技术的发展,当今的很多方法流程的改进已经可以视为硬件电路结构的直接改进。设计人员几乎都通过将改进的方法流程编程到硬件电路中来得到相应的硬件电路结构。因此,不能说一个方法流程的改进就不能用硬件实体模块来实现。例如,可编程逻辑器件(Programmable Logic Device, PLD)(例如现场可编程门阵列(Field Programmable Gate Array, FPGA))就是这样一种集成电路,其逻辑功能由用户对器件编程来确定。由设计人员自行编程来把一个数字系统“集成”在一片PLD上,而不需要请芯片制造厂商来设计和制作专用的集成电路芯片。而且,如今,取代手工地制作集成电路芯片,这种编程也多半改用“逻辑编译器(logic compiler)”软件来实现,它与程序开发撰写时所用的软件编译器相类似,而要编译之前的原始代码也得用特定的编程语言来撰写,此称之为硬件描述语言(Hardware Description Language, HDL),而HDL也并非仅有一种,而是有许多种,如ABEL(Advanced Boolean Expression Language)、AHDL(Altera Hardware Description Language)、Confluence、CUPL(Cornell University Programming Language)、HDCal、JHDL(Java Hardware Description Language)、Lava、Lola、MyHDL、PALASM、RHDH(Ruby Hardware Description Language)等,目前最普遍使用的是VHDL(Verilog-High-Speed Integrated Circuit Hardware Description Language)与Verilog。本领域技术人员也应该清楚,只需要将方法流程用上述几种硬件描述语言稍作逻辑编程并编程到集成电路中,就可以很容易得到实现该逻辑方法流程的硬件电路。

[0099] 控制器可以按任何适当的方式实现,例如,控制器可以采取例如微处理器或处理器以及存储可由该(微)处理器执行的计算机可读程序代码(例如软件或固件)的计算机可读介质、逻辑门、开关、专用集成电路(Application Specific Integrated Circuit, ASIC)、可编程逻辑控制器和嵌入微控制器的形式,控制器的例子包括但不限于以下微控制器:ARC 625D、Atmel AT91SAM、Microchip PIC18F26K20以及Silicone Labs C8051F320,存储器控制器还可以被实现为存储器的控制逻辑的一部分。本领域技术人员也知道,除了以纯计算机可读程序代码方式实现控制器以外,完全可以通过将方法步骤进行逻辑编程来使得控制器

以逻辑门、开关、专用集成电路、可编程逻辑控制器和嵌入微控制器等的形式来实现相同功能。因此这种控制器可以被认为是一种硬件部件,而对其内包括的用于实现各种功能的装置也可以视为硬件部件内的结构。或者甚至,可以将用于实现各种功能的装置视为既可以是实现方法的软件模块又可以是硬件部件内的结构。

[0100] 上述实施例阐明的系统、装置、模块或单元,具体可以由计算机芯片或实体实现,或者由具有某种功能的产品来实现。一种典型的实现设备为计算机。具体的,计算机例如可以为个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任何设备的组合。

[0101] 为了描述的方便,描述以上装置时以功能分为各种单元分别描述。当然,在实施本说明书时可以把各单元的功能在同一个或多个软件和/或硬件中实现。

[0102] 本领域内的技术人员应明白,本发明的实施例可提供为方法、系统、或计算机程序产品。因此,本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0103] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0104] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0105] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0106] 在一个典型的配置中,计算设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0107] 内存可能包括计算机可读介质中的非永久性存储器,随机存取存储器(RAM)和/或非易失性内存等形式,如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。

[0108] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括,但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除

可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带, 磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质, 可用于存储可以被计算设备访问的信息。按照本文中的界定, 计算机可读介质不包括暂存电脑可读媒体 (transitory media), 如调制的数据信号和载波。

[0109] 还需要说明的是, 术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含, 从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素, 而且还包括没有明确列出的其他要素, 或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下, 由语句“包括一个……”限定的要素, 并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0110] 本领域技术人员应明白, 本说明书的实施例可提供为方法、系统或计算机程序产品。因此, 本说明书可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且, 本说明书可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质 (包括但不限于磁盘存储器、CD-ROM、光学存储器等) 上实施的计算机程序产品的形式。

[0111] 本说明书可以在由计算机执行的计算机可执行指令的一般上下文中描述, 例如程序模块。一般地, 程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本说明书, 在这些分布式计算环境中, 由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中, 程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[0112] 本说明书中的各个实施例均采用递进的方式描述, 各个实施例之间相同相似的部分互相参见即可, 每个实施例重点说明的都是与其他实施例的不同之处。尤其, 对于系统实施例而言, 由于其基本相似于方法实施例, 所以描述的比较简单, 相关之处参见方法实施例的部分说明即可。

[0113] 以上所述仅为本说明书的实施例而已, 并不用于限制本说明书。对于本领域技术人员来说, 本说明书可以有各种更改和变化。凡在本说明书的精神和原理之内所作的任何修改、等同替换、改进等, 均应包含在本说明书的权利要求范围之内。

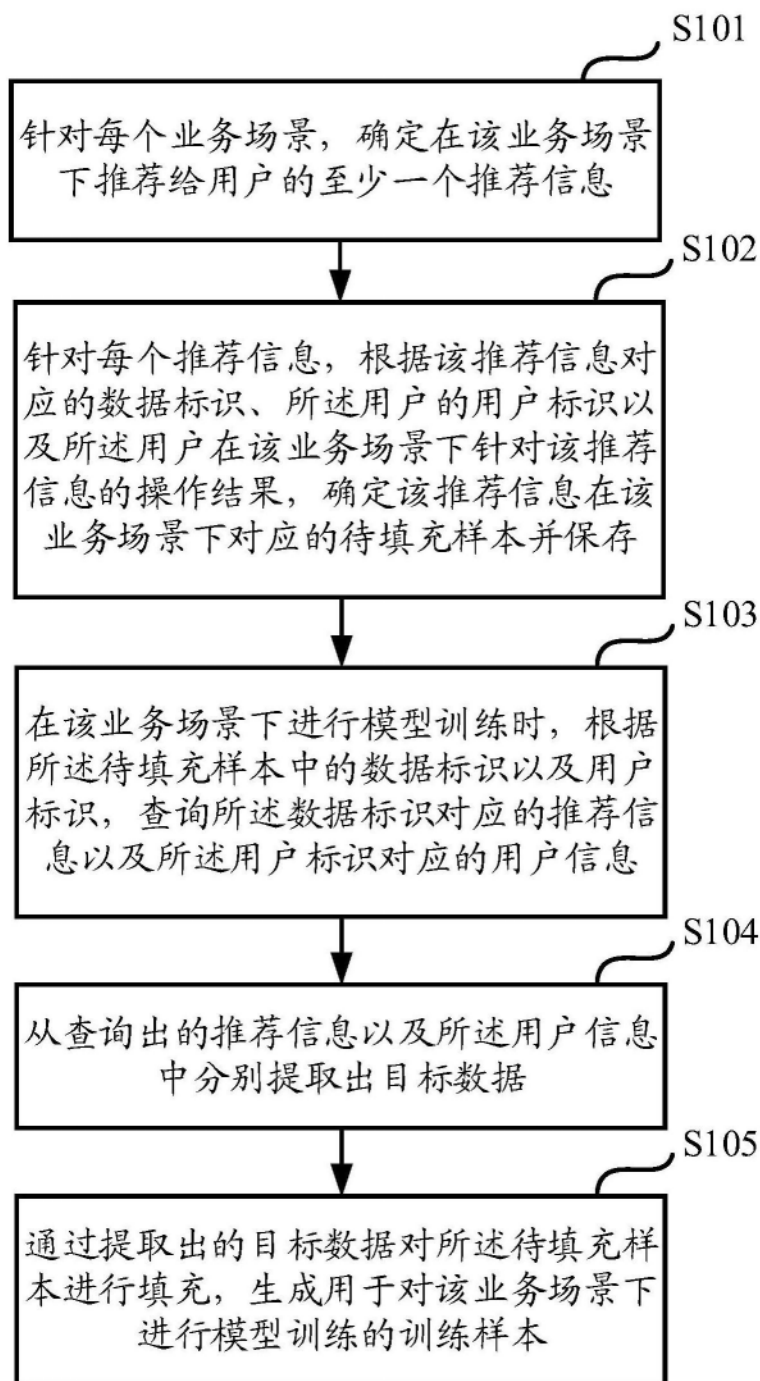


图1

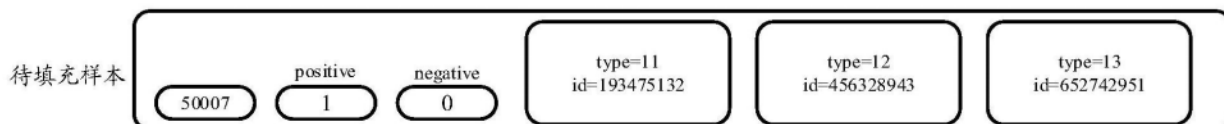


图2

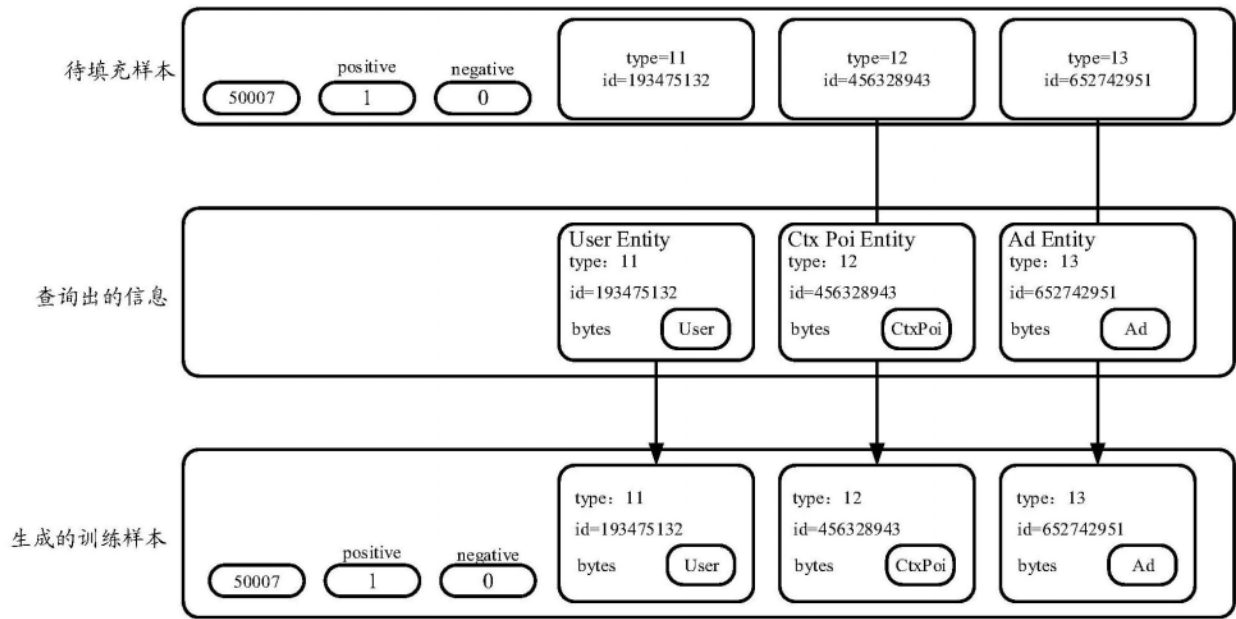


图3

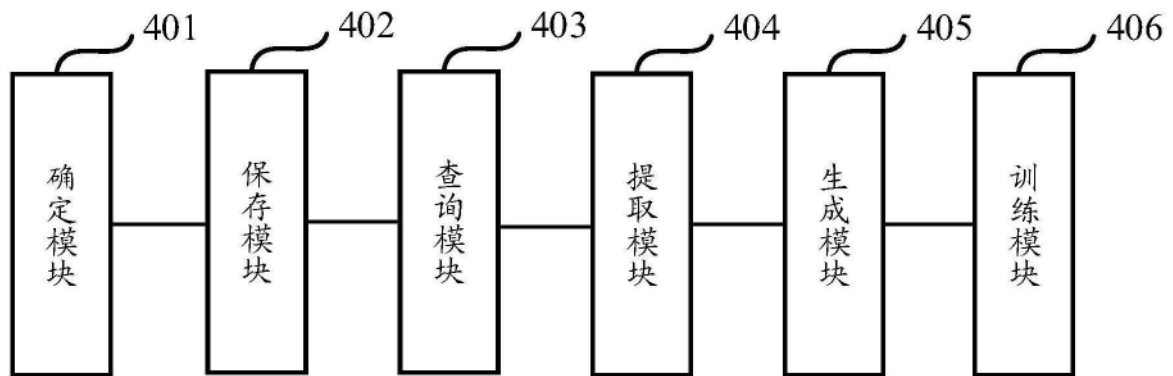


图4

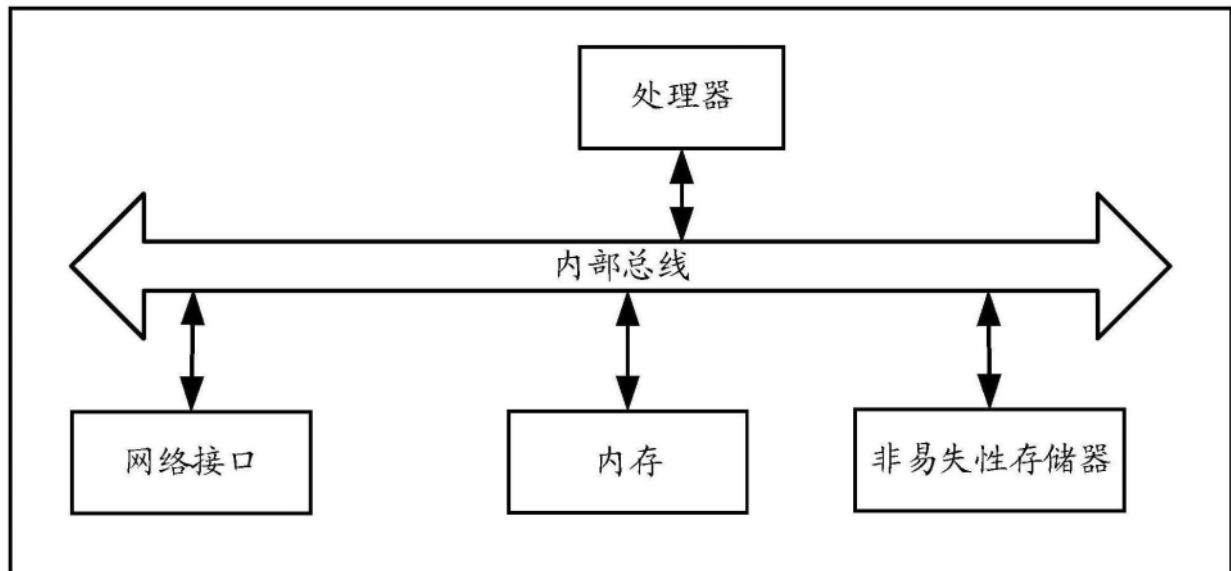


图5