(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108470071 B
(45)授权公告日 2022.02.18

(21)申请号 201810271496.4

(22)申请日 2018.03.29

(65)同一申请的已公布的文献号
申请公布号 CN 108470071 A

(43)申请公布日 2018.08.31

(73)专利权人 联想(北京)有限公司
地址 100085 北京市海淀区上地信息产业基地创业路6号

(72)发明人 杨帆 金宝宝 张成松

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 王宝筠

(51)Int.Cl.
G06F 16/21(2019.01)

(56)对比文件
CN 107527070 A,2017.12.29
CN 105590157 A,2016.05.18
CN 102156983 A,2011.08.17
CN 105915555 A,2016.08.31
US 2006123389 A1,2006.06.08
刘晓然.《基于文件的数据分级存储的研究与实现》.《中国优秀硕士学位论文全文数据库信息科技辑》.2014,

审查员 陈学元

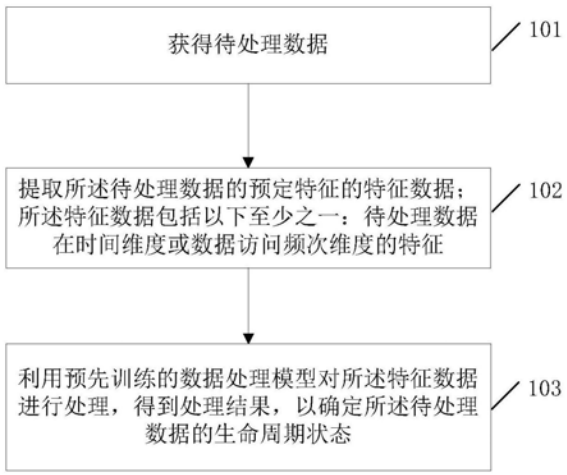
权利要求书2页 说明书15页 附图5页

(54)发明名称

一种数据处理方法及装置

(57)摘要

本申请提供的数据处理方法及装置,在获得待处理数据后,提取待处理数据的特征数据,该特征数据包括待处理数据在时间维度或数据访问频次维度特征中的至少之一;在此基础上,利用预先训练的数据处理模型对特征数据进行处理以确定待处理数据的生命周期状态。本申请方案由于利用了预先训练的数据处理模型基于待处理数据在时间维度或数据访问频次维度的特征中的至少之一,来确定待处理数据的生命周期状态,从而,在确定待处理数据的生命周期状态时,具体是利用了大数据在时间维度或数据访问频次维度的特征中的至少之一与大数据生命周期状态的对应关系的规律,这与现有技术相比,既降低了人力成本,又可使得所确定的数据生命周期状态具有较高的准确度。



CN 108470071 B

1.一种数据处理方法,其特征在于,包括:

获得待处理数据;

提取所述待处理数据的预定特征的特征数据;所述特征数据包括以下至少之一:待处理数据在时间维度或数据访问频次维度的特征;

利用预先训练的数据处理模型描述的大数据在时间维度或数据访问频次维度的特征中的至少之一与大数据生命周期状态的对应关系的规律,对所述特征数据进行生命周期状态识别处理,得到所述待处理数据的处理结果,基于所述处理结果确定所述待处理数据的生命周期状态;

所述基于所述处理结果确定所述待处理数据的生命周期状态,包括:在将所述待处理数据的生命周期状态划分为生命周期已结束、生命周期未结束两种状态的情况下,获得所述处理结果中包括的待处理数据生命周期已结束的第一置信度信息,及待处理数据生命周期未结束的第二置信度信息;基于所述第一置信度信息及所述第二置信度信息,确定所述待处理数据的生命周期状态;

其中,所述数据处理模型的训练样本的数据类型至少包括所述待处理数据的数据类型。

2.根据权利要求1所述的方法,其特征在于,所述提取所述待处理数据的预定特征的特征数据,包括:

提取所述待处理数据自创建时间点起至当前的存活时长、自最近一次访问的时间点起至当前的时长或在最近的至少一个预定时间段内的数据访问频次中的至少一个。

3.根据权利要求1所述的方法,其特征在于,所述对所述特征数据进行生命周期状态识别处理,得到所述待处理数据的处理结果,包括:

获得所述待处理数据的数据类型;

从所述数据处理模型中确定出与所述数据类型相对应的子处理模型;其中,所述数据处理模型包括多于一个的子处理模型,不同的子处理模型与不同的数据类型相对应,且不同的子处理模型所对应的数据特征类型和/或特征权重不同;

将所述待处理数据的特征数据输入与所述数据类型相对应的所述子处理模型中,得到所述待处理数据的处理结果,以基于所述处理结果得到所述待处理数据的生命周期状态。

4.根据权利要求1所述的方法,其特征在于,所述数据处理模型的训练过程包括:

获得多条训练样本;

提取每条训练样本的所述预定特征的特征数据;所述训练样本的特征数据包括以下至少之一:训练样本在所述时间维度或所述数据访问频次维度的特征;

标注每条训练样本的数据生命周期状态,得到每条训练样本的数据生命周期状态标注结果;

建立每条训练样本的特征数据与数据生命周期状态标注结果间的对应关系,得到每条训练样本的特征数据与数据生命周期状态标注结果的对应关系数据;

基于预定的机器学习算法,利用各条训练样本的所述对应关系数据训练一数据处理模型,使得所述数据处理模型能够基于输入的特征数据输出相应的数据生命周期状态预测结果。

5.根据权利要求4所述的方法,其特征在于,所述获得多条训练样本,包括:

从生产环境中随机选择一批数据，并将随机选择的数据作为训练样本。

6.一种数据处理装置，其特征在于，包括：

获取单元，用于获得待处理数据；

提取单元，用于提取所述待处理数据的预定特征的特征数据；所述特征数据包括以下至少之一：待处理数据在时间维度或数据访问频次维度的特征；

处理单元，用于利用预先训练的数据处理模型描述的大数据在时间维度或数据访问频次维度的特征中的至少之一与大数据生命周期状态的对应关系的规律，对所述特征数据进行生命周期状态识别处理，得到所述待处理数据的处理结果，基于所述处理结果确定所述待处理数据的生命周期状态；

所述处理单元在基于所述处理结果确定所述待处理数据的生命周期状态时，具体用于：在将所述待处理数据的生命周期状态划分为生命周期已结束、生命周期未结束两种状态的情况下，获得所述处理结果中包括的待处理数据生命周期已结束的第一置信度信息，及待处理数据生命周期未结束的第二置信度信息；基于所述第一置信度信息及所述第二置信度信息，确定所述待处理数据的生命周期状态；

其中，所述数据处理模型的训练样本的数据类型至少包括所述待处理数据的数据类型。

7.根据权利要求6所述的装置，其特征在于，所述提取单元，具体用于：

提取所述待处理数据自创建时间点起至当前的存活时长、自最近一次访问的时间点起至当前的时长或在最近的至少一个预定时间段内的数据访问频次中的至少一个。

8.根据权利要求6所述的装置，其特征在于，所述处理单元，具体用于：

获得所述待处理数据的数据类型；

从所述数据处理模型中确定出与所述数据类型相对应的子处理模型；其中，所述数据处理模型包括多于一个的子处理模型，不同的子处理模型与不同的数据类型相对应，且不同的子处理模型所对应的数据特征类型和/或特征权重不同；

将所述待处理数据的特征数据输入与所述数据类型相对应的所述子处理模型中，得到所述待处理数据的处理结果，以基于所述处理结果得到所述待处理数据的生命周期状态。

9.根据权利要求6所述的装置，其特征在于，还包括：

预处理单元，用于：

获得多条训练样本；

提取每条训练样本的所述预定特征的特征数据；所述训练样本的特征数据包括以下至少之一：训练样本在所述时间维度或所述数据访问频次维度的特征；

标注每条训练样本的数据生命周期状态，得到每条训练样本的数据生命周期状态标注结果；

建立每条训练样本的特征数据与数据生命周期状态标注结果间的对应关系，得到每条训练样本的特征数据与数据生命周期状态标注结果的对应关系数据；

基于预定的机器学习算法，利用各条训练样本的所述对应关系数据训练一数据处理模型，使得所述数据处理模型能够基于输入的特征数据输出相应的数据生命周期状态预测结果。

一种数据处理方法及装置

技术领域

[0001] 本发明属于机器学习领域，尤其涉及一种数据处理方法及装置。

背景技术

[0002] 在生产环境中，经常需要删除生命周期已结束的数据，如表、日志文件等，以释放存储空间、保证系统正常运行，这就必然需要确定数据的生命周期是否已经结束。

[0003] 当前主流的确定数据生命周期是否结束的方法有两种，一种是预先设定一个阈值T，将数据自产生时刻起的总存活时长t与设定的阈值T进行比较，并将t>T的数据认为是生命周期已结束的数据将其删除，反之则认为数据的生命周期未结束；另一种是人工审查数据是否有用，将人工审查无用的数据认为是生命周期已结束的数据并删除。

[0004] 上述两种方法均存在一些缺陷，第一种方法中的阈值T难以确定，相应地会导致数据生命周期结束与否的判断准确度低，进而会为数据删除操作带来不利影响，其中，若T设置过大，可能会导致很多无用的数据无法删除，若T设置过小，则在删除无用数据的同时往往会删除有用的数据；第二种方法由于需要人工逐数据审查、判断，从而存在人力成本过高的问题。

发明内容

[0005] 有鉴于此，本发明的目的在于提供一种数据处理方法及装置，旨在克服现有技术存在的上述问题，以使得在低人力成本前提下，所确定的数据生命周期状态具有较高的准确度。

[0006] 为此，本发明公开如下技术方案：

[0007] 一种数据处理方法，包括：

[0008] 获得待处理数据；

[0009] 提取所述待处理数据的预定特征的特征数据；所述特征数据包括以下至少之一：待处理数据在时间维度或数据访问频次维度的特征；

[0010] 利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果，以确定所述待处理数据的生命周期状态。

[0011] 上述方法，优选的，所述提取所述待处理数据的预定特征的特征数据，包括：

[0012] 提取所述待处理数据自创建时间点起至当前的存活时长、自最近一次访问的时间点起至当前的时长或在最近的至少一个预定时间段内的数据访问频次中的至少一个。

[0013] 上述方法，优选的，所述利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果，包括：

[0014] 获得所述待处理数据的数据类型；

[0015] 从所述数据处理模型中确定出与所述数据类型相对应的子处理模型；其中，所述数据处理模型包括多于一个的子处理模型，不同的子处理模型与不同的数据类型相对应，且不同的子处理模型所对应的数据特征类型和/或特征权重不同；

[0016] 将所述待处理数据的特征数据输入与所述数据类型相对应的所述子处理模型中，得到所述待处理数据的处理结果。

[0017] 上述方法，优选的，所述确定所述待处理数据的生命周期状态，包括：

[0018] 获得所述处理结果中包括的待处理数据生命周期已结束的第一置信度信息，及待处理数据生命周期未结束的第二置信度信息；

[0019] 基于所述第一置信度信息及所述第二置信度信息，确定所述待处理数据的生命周期状态。

[0020] 上述方法，优选的，所述数据处理模型的训练过程包括：

[0021] 获得多条训练样本；

[0022] 提取每条训练样本的所述预定特征的特征数据；所述训练样本的特征数据包括以下至少之一：训练样本在所述时间维度或所述数据访问频次维度的特征；

[0023] 标注每条训练样本的数据生命周期状态，得到每条训练样本的数据生命周期状态标注结果；

[0024] 建立每条训练样本的特征数据与数据生命周期状态标注结果间的对应关系，得到每条训练样本的特征数据与数据生命周期状态标注结果的对应关系数据；

[0025] 基于预定的机器学习算法，利用各条训练样本的所述对应关系数据训练一数据处理模型，使得所述数据处理模型能够基于输入的特征数据输出相应的数据生命周期状态预测结果。

[0026] 上述方法，优选的，所述获得多条训练样本，包括：

[0027] 从生产环境中随机选择一批数据，并将随机选择的数据作为训练样本。

[0028] 一种数据处理装置，包括：

[0029] 获取单元，用于获得待处理数据；

[0030] 提取单元，用于提取所述待处理数据的预定特征的特征数据；所述特征数据包括以下至少之一：待处理数据在时间维度或数据访问频次维度的特征；

[0031] 处理单元，用于利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果，以确定所述待处理数据的生命周期状态。

[0032] 上述装置，优选的，所述提取单元，具体用于：

[0033] 提取所述待处理数据自创建时间点起至当前的存活时长、自最近一次访问的时间点起至当前的时长或在最近的至少一个预定时间段内的数据访问频次中的至少一个。

[0034] 上述装置，优选的，所述处理单元，具体用于：

[0035] 获得所述待处理数据的数据类型；

[0036] 从所述数据处理模型中确定出与所述数据类型相对应的子处理模型；其中，所述数据处理模型包括多于一个的子处理模型，不同的子处理模型与不同的数据类型相对应，且不同的子处理模型所对应的数据特征类型和/或特征权重不同；

[0037] 将所述待处理数据的特征数据输入与所述数据类型相对应的所述子处理模型中，得到所述待处理数据的处理结果。

[0038] 上述装置，优选的，还包括：

[0039] 预处理单元，用于：

[0040] 获得多条训练样本；

[0041] 提取每条训练样本的所述预定特征的特征数据;所述训练样本的特征数据包括以下至少之一:训练样本在所述时间维度或所述数据访问频次维度的特征;

[0042] 标注每条训练样本的数据生命周期状态,得到每条训练样本的数据生命周期状态标注结果;

[0043] 建立每条训练样本的特征数据与数据生命周期状态标注结果间的对应关系,得到每条训练样本的特征数据与数据生命周期状态标注结果的对应关系数据;

[0044] 基于预定的机器学习算法,利用各条训练样本的所述对应关系数据训练一数据处理模型,使得所述数据处理模型能够基于输入的特征数据输出相应的数据生命周期状态预测结果。

[0045] 根据以上方案可知,本申请提供的数据处理方法及装置,在获得待处理数据后,提取待处理数据的特征数据,该特征数据包括待处理数据在时间维度或数据访问频次维度的特征中的至少之一;在此基础上,利用预先训练的数据处理模型对特征数据进行处理,以确定待处理数据的生命周期状态。本申请方案由于利用了预先训练的数据处理模型基于待处理数据在时间维度或数据访问频次维度的特征中的至少之一,来确定待处理数据的生命周期状态,从而,在确定待处理数据的生命周期状态时,具体是利用了大数据在时间维度或数据访问频次维度的特征中的至少之一与大数据生命周期状态的对应关系的规律,这与现有技术相比,既降低了人力成本,又可使得所确定的数据生命周期状态具有较高的准确度。

附图说明

[0046] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述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,还可以根据提供的附图获得其他的附图。

[0047] 图1是本申请实施例一提供的数据处理方法流程图;

[0048] 图2是本申请实施例二提供的训练数据处理模型并基于训练的模型进行数据生命周期预测的整体框架图;

[0049] 图3是本申请实施例二提供的训练数据处理模型的实现过程示意图;

[0050] 图4是本申请实施例三提供的数据处理方法流程图;

[0051] 图5是本申请实施例四提供的数据处理方法流程图;

[0052] 图6是本申请实施例五提供的数据处理装置的结构示意图;

[0053] 图7是本申请实施例六提供的数据处理装置的结构示意图。

具体实施方式

[0054] 下面将结合本发明实施例中的附图,对本发明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发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,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,都属于本发明保护的范围。

[0055] 本申请提供了一种数据处理方法及装置,用于解决现有技术中通过设置阈值T的方式所判断出的数据生命周期状态准确率低以及人工审查方式所存在的人力成本过高的

问题，以实现在低人力成本前提下，所确定的数据生命周期状态具有较高的准确度，以下将通过多个实施例对本申请方案进行说明。

[0056] 参考图1，为本申请提供的一种数据处理方法实施例一的流程图，该方法可应用于智能手机、平板电脑(PAD，Portable Android Device)、个人数字助理(PDA(Personal Digital Assistant)、笔记本、台式机或一体机等各种终端设备中，或者还可以应用于各种通用或专用服务器中。如图1所示，该数据处理方法包括以下处理步骤：

[0057] 步骤101、获得待处理数据。

[0058] 所述待处理数据可以是实际生产环境中所产生或创建的各类型数据，例如可以是但不限于数据表数据或日志文件的日志数据等。

[0059] 对于这些数据，为了达到特定的目的，往往存在获知其生命周期状态的需求，例如，为了删除生命周期已结束的数据(或称失效数据)，以释放存储空间，则需要获知数据的生命周期是否已结束；为了根据数据访问的频繁程度进行数据的分类存储，则需要获知这些数据是处于频繁访问状态还是很少访问状态还是已失效。针对该情况，本申请的目的就在于能够低人力成本、高准确率地确定出数据的生命周期状态。

[0060] 数据生命周期状态的确定需以数据生命周期状态的划分为基础，划分方式不同，在对数据进行生命周期状态预测时，所对应的候选状态不同。

[0061] 例如，按数据失效与否进行划分，可将数据的生命周期状态划分为生命周期未结束及生命周期已结束这两种状态；按数据访问的频繁程度进行划分，则可将数据的生命周期状态划分为频繁访问、很少访问以及失效等状态，具体应用中，可根据实际的状态预测需求选择合适的划分方式对数据的生命周期状态进行划分。

[0062] 步骤102、提取所述待处理数据的预定特征的特征数据；所述特征数据包括以下至少之一：待处理数据在时间维度或数据访问频次维度的特征。

[0063] 待处理数据在时间维度的特征可以包括但不限于：待处理数据自创建时间点起至当前的存活时长、自最近一次访问的时间点起至当前的时长。

[0064] 待处理数据在数据访问频次维度的特征可以包括但不限于：待处理数据在至少一个预定时间段内的数据访问频次，例如待处理数据在最近1天、最近1周、最近15天、最近1月、最近1年或最近5年等时间粒度内的访问频次等。

[0065] 针对不同类型的待处理数据，当对其进行数据特征提取时，可以对其进行相同的数据特征提取，或者还可以对其进行不同的数据特征提取，本申请并不对此进行局限。

[0066] 具体地，例如针对数据表、日志文件这些不同类型的数据，可以设定两者需提取的特征均为上述时间维度和/或数据访问频次维度的特征。

[0067] 或者，针对数据表、日志文件这些不同类型的数据，还可以分别为其设定不同的需提取的特征，如对于数据表，可设定其需提取的特征为时间维度的特征、数据访问频次维度的特征以及价值维度的特征，对于日志文件，则可以设定其需提取的特征为时间维度的特征及数据访问频次维度的特征等。

[0068] 所述价值维度的特征，具体地，可以是但不限于数据的价值等级，例如预先设定的高、中、低三个价值等级等，示例性地，在进行数据特征提取时，针对中间数据的数据表，由于中间数据在实际生产环境中重要程度往往较低，企业人员或用户一般不会对中间数据给予过多关注，从而可将其价值特征的特征值提取为“低”，针对结果数据的数据表，由于结果

数据在实际生产环境中重要程度往往较高，其一般是企业人员或用户的重点关注对象，从而可将其价值特征的特征值提取为“高”。

[0069] 步骤103、利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果，以确定所述待处理数据的生命周期状态。

[0070] 所述数据处理模型，为预先利用数据表或日志文件等类型的大批量数据所训练的模型，该数据处理模型能够描述数据的数据特征(如时间维度和/或访问频次维度的特征等)与其生命周期状态间的对应关系规律。

[0071] 鉴于此，本步骤中，可将待处理数据的特征数据作为所述数据处理模型的输入提供给所述数据处理模型，由所述数据处理模型基于学习所得的数据特征与数据生命周期状态间的对应关系规律，输出能够表明所述待处理数据的生命周期状态的处理结果数据，进而可基于所述数据处理模型的处理结果数据，确定所述待处理数据的生命周期状态，如确定所述待处理数据的生命周期是否已结束，或者确定所述待处理数据是处于频繁访问、很少访问还是已失效的状态等等。

[0072] 根据以上方案可知，本申请提供的数据处理方法，在获得待处理数据后，提取待处理数据的特征数据，该特征数据包括待处理数据在时间维度或数据访问频次维度的特征中的至少之一；在此基础上，利用预先训练的数据处理模型对特征数据进行处理，以确定待处理数据的生命周期状态。本申请方案由于利用了预先训练的数据处理模型基于待处理数据在时间维度或数据访问频次维度的特征中的至少之一，来确定待处理数据的生命周期状态，从而，在确定待处理数据的生命周期状态时，具体是利用了大数据在时间维度或数据访问频次维度的特征中的至少之一与大数据生命周期状态的对应关系的规律，这与现有技术相比，既降低了人力成本，又可使得所确定的数据生命周期状态具有较高的准确度。

[0073] 本申请的数据处理方法需要在预先训练出所述数据处理模型的基础上展开，即，如图2所示，需要利用训练样本预先训练出一数据处理模型，在此基础上，方可基于所述数据处理模型进行数据的生命周期状态预测。

[0074] 本实施例二提供训练所述数据处理模型的实现过程，参考图3，为本申请提供的数据处理模型的训练过程实施例二的流程图，该训练过程包括：

[0075] 步骤301、获得多条训练样本。

[0076] 所述训练样本可以是但不限于从实际生产环境中随机选择的批量数据，如批量的日志文件数据，或者批量的数据表数据等。

[0077] 需要说明的是，所述训练样本的数据类型应视拟训练的数据处理模型所需具备的功能而定，若拟训练的数据处理模型需具备数据表数据的生命周期状态预测功能，则所述训练样本应是数据表类型的数据，若拟训练的数据处理模型需具备日志数据的生命周期状态预测功能，则所述训练样本应是日志类型的数据。

[0078] 步骤302、提取每条训练样本的所述预定特征的特征数据；所述训练样本的特征数据包括以下至少之一：训练样本在所述时间维度或所述数据访问频次维度的特征。

[0079] 训练样本在时间维度的特征可以包括但不限于：训练样本自创建时间点起至当前的存活时长、自最近一次访问的时间点起至当前的时长。

[0080] 训练样本在数据访问频次维度的特征可以包括但不限于：训练样本在至少一个预定时间段内的数据访问频次，如训练样本在最近1天、最近1周、最近15天、最近1月、最近1

年、最近5年等时间粒度内的访问频次等。

[0081] 针对不同类型的训练样本，当对其进行数据特征提取时，可以对其进行相同的数据特征提取，或者还可以对其进行不同的数据特征提取，本申请并不对此进行局限。

[0082] 具体地，例如针对数据表、日志文件这些不同类型的训练样本，可以设定两者需提取的特征均为上述时间维度和/或数据访问频次维度的特征。

[0083] 或者，针对数据表、日志文件这些不同类型的训练样本，还可以分别为其设定不同的需提取的特征，如对于数据表，可设定其需提取的特征为时间维度的特征、数据访问频次维度的特征以及价值维度的特征，对于日志文件，则可以设定其需提取的特征为时间维度的特征及数据访问频次维度的特征等。

[0084] 所述价值维度的特征，具体地，可以是但不限于数据的价值等级，例如预先设定的高、中、低三个价值等级等，示例性地，在对训练样本进行数据特征提取时，针对中间数据的数据表，由于中间数据在实际生产环境中重要程度往往较低，企业人员或用户一般不会对中间数据给予过多关注，从而可将其价值特征的特征值提取为“低”，针对结果数据的数据表，由于结果数据在实际生产环境中重要程度往往较高，其一般是企业人员或用户的重点关注对象，从而可将其价值特征的特征值提取为“高”。

[0085] 步骤303、标注每条训练样本的数据生命周期状态，得到每条训练样本的数据生命周期状态标注结果。

[0086] 具体可采用人工标注等方式来标注每条训练样本的生命周期状态。

[0087] 训练样本的生命周期状态标注结果应视其生命周期状态的划分方式而定，示例性地，若按数据失效与否将其划分为生命周期已结束及生命周期未结束两种状态，则训练样本的生命周期状态标注结果应是该两种状态的其中之一，若其生命周期按访问的频繁程度划分为频繁访问、很少访问及已失效这三种状态，则训练样本的生命周期状态标注结果应是该三种状态的其中之一

[0088] 需要说明的是，本步骤提供的标注训练样本生命周期状态的处理过程与上一步骤提供的提取训练样本数据特征的处理过程，不必局限于本实施例提供的先后执行次序，实际应用中，既可以先执行特征提取过程，后执行状态标注过程，还可以先执行状态标注过程，后执行特征提取过程，或者还可以同时执行上述两个处理过程，本实施例对此不作限定。

[0089] 步骤304、建立每条训练样本的特征数据与数据生命周期状态标注结果间的对应关系，得到每条训练样本的特征数据与数据生命周期状态标注结果的对应关系数据。

[0090] 在提取出训练样本的特征数据并标注出训练样本的生命周期状态的基础上，可建立每条训练样本的特征数据与其生命周期状态标注结果间的对应关系，以使得为数据处理模型的训练提供支持。

[0091] 步骤305、基于预定的机器学习算法，利用各条训练样本的所述对应关系数据训练一数据处理模型，使得所述数据处理模型能够基于输入的特征数据输出相应的数据生命周期状态预测结果。

[0092] 所述机器学习算法可以是但不限于随机森林、KNN(k-NearestNeighbor，k最邻近分类算法)、逻辑回归、SVM(Support Vector Machine，支持向量机)等中的任意一种，具体应用中，可由技术人员基于实际的模型特点需求，选择合适的机器学习算法。

[0093] 在完成建立每条训练样本的特征数据与其生命周期状态标注结果间的对应关系后，可对所选择的机器学习算法进行模型参数的初始化，如初始化模型的各特征权重，以及初始化模型的学习速率这些超参数等，在此基础上，可利用各条训练样本的特征数据与其生命周期状态标注结果间的对应关系，对初始化后的模型进行训练，以使得数据处理模型能够不断学习各训练样本的数据特征与其生命周期状态间的对应关系的规律，进而不断调整/优化各特征的特征权重，最终得到能够基于输入的特征数据输出相应的生命周期状态预测结果的数据处理模型。

[0094] 训练所得的所述数据处理模型能够处理的数据的数据类型，一般来说需与训练样本的数据类型相一致，例如，若训练样本的数据类型为数据表类型，则训练所得的数据处理模型能够处理数据表数据，即能够依据输入的数据表数据的数据特征，输出数据表数据的生命周期状态预测结果；若训练样本的数据类型为日志类型，则训练所得的数据处理模型能够处理日志类型的数据，即能够依据输入的日志数据的数据特征，输出日志数据的生命周期状态预测结果。

[0095] 也就是说，基于不同数据类型的训练样本训练所得的数据处理模型，一般来说仅适用于与训练样本数据类型相同的数据的生命周期状态预测。

[0096] 为了向用户提供同时适用于不同类型数据的生命周期状态预测功能，还可以将训练所得的分别适用于不同数据类型的数据处理模型作为子模型进行集成，在对各个子模型完成集成后得到一总模型，该总模型则能够同时适用于对不同类型的数据进行生命周期状态预测，实际应用中可通过为所述总模型提供一个位于各子模型上层的总接口(如设计一个面向各类型数据的数据输入及状态预测界面等)，使得为用户提供同时适用于不同类型数据的生命周期状态预测功能，该总接口至少应具备识别/判断待预测数据的数据类型，并根据待预测数据的数据类型调起相应的子模型进行数据生命周期状态预测的功能。

[0097] 本实施例通过基于预定的机器学习算法，利用相应数据类型的训练样本进行数据模型训练，为所述相应数据类型的数据提供了生命周期状态预测功能；通过将分别适用于不同数据类型的数据处理模型作为子模型进行集成，并为集成所得的总模型提供一个位于各子模型上层的总接口，为多种类型的数据提供了通用的生命周期状态预测功能。

[0098] 参考图4，为本申请提供的一种数据处理方法实施例三的流程图，在本实施例中，所述数据处理模型优选地为能够同时适用于对不同数据类型(如数据表数据、日志数据)的数据进行生命周期状态预测的模型，具体地，所述数据处理模型包括多于一个的子处理模型，不同的子处理模型与不同的数据类型相对应，用于对不同数据类型的数据进行生命周期状态预测，且不同的子处理模型所对应的数据特征类型和/或特征权重不同。

[0099] 如图4所示，本实施例中，所述步骤103(利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果)，可以通过以下的处理过程实现：

[0100] 步骤1031、获得所述待处理数据的数据类型。

[0101] 所述待处理数据可以是但不限于数据表数据或者日志数据。

[0102] 其中，可以但不限于基于待处理数据的数据格式、待处理数据所在文档的文档类型(如数据表数据及日志数据在数据格式上以及文档类型上一般有较大差别)或者预先标注的类型信息等，识别待处理数据的数据类型。

[0103] 步骤1032、从所述数据处理模型中确定出与所述数据类型相对应的子处理模型；

其中，所述数据处理模型包括多于一个的子处理模型，不同的子处理模型与不同的数据类型相对应，且不同的子处理模型所对应的数据特征类型和/或特征权重不同。

[0104] 在获得所述待处理数据的数据类型后，可根据该数据类型，从所述数据处理模型包括的多个子处理模型中确定出与其相匹配的子处理模型。

[0105] 例如，如果所述待处理数据的数据类型为数据表数据，则可从所述数据处理模型包括的多个子处理模型中确定出能够对数据表数据进行生命周期状态预测的第一子处理模型；如果所述待处理数据的数据类型为日志数据，则可从所述数据处理模型包括的多个子处理模型中确定出能够对日志数据进行生命周期状态预测的第二子处理模型。

[0106] 步骤1033、将所述待处理数据的特征数据输入与所述数据类型相对应的所述子处理模型中，得到所述待处理数据的处理结果。

[0107] 在确定出与待处理数据的数据类型相对应的子处理模型后，可将待处理数据的特征数据输入该子处理模型中，最终由该子处理模型基于数据特征与数据生命周期状态间的对应关系的规律，输出与待处理数据的特征数据(如时间维度的特征、数据访问频次维度的特征和/或价值维度的特征等)相对应的生命周期状态预测结果。

[0108] 其中，所输出的生命周期状态预测结果，会因数据生命周期状态划分方式的不同而不同。

[0109] 示例性地，若数据生命周期状态按失效与否划分为生命周期已结束、生命周期未结束两种状态，则所述生命周期状态预测结果包括分别对应于数据生命周期已结束的置信度信息及生命周期未结束的置信度信息；若数据生命周期状态按访问频繁程度划分为频繁访问、很少访问及失效三种状态，则所述生命周期状态预测结果包括分别对应于频繁访问、很少访问及失效三种状态的三种置信度信息。

[0110] 本实施例通过使用能够适用于不同数据类型的数据处理模型，为不同类型数据提供了通用的生命周期状态预测功能。且由于本实施例是利用了大数据在时间维度、数据访问频次维度和/或价值维度的特征与大数据生命周期状态的对应关系规律，对待处理数据进行生命周期状态预测，从而与现有技术相比，既降低了人力成本，又可使得所确定的数据生命周期状态具有较高的准确度。

[0111] 参考图5，为本申请提供的一种数据处理方法实施例四的流程图，本实施例以数据的生命周期状态按失效与否被划分为生命周期已结束、生命周期未结束两种状态为例，介绍根据所述数据处理模型的处理结果，确定待处理数据的生命周期状态的实现过程，如图5所示，可以通过以下处理过程来确定待处理数据的生命周期状态：

[0112] 步骤501、获得所述处理结果中包括的待处理数据生命周期已结束的第一置信度信息，及待处理数据生命周期未结束的第二置信度信息。

[0113] 在数据的生命周期状态按失效与否被划分为生命周期已结束、生命周期未结束两种状态的情况下，一般来说，数据处理模型在对待处理数据进行生命周期状态预测后可输出以下结果数据：待处理数据生命周期已结束的第一置信度信息、待处理数据生命周期未结束的第二置信度信息。

[0114] 所述第一置信度信息、第二置信度信息，可以是但不限于概率值或百分比数值等能够表明待处理数据属于相应状态的可能性程度的信息。

[0115] 步骤502、基于所述第一置信度信息及所述第二置信度信息，确定所述待处理数据

的生命周期状态。

[0116] 在获得待处理数据生命周期已结束的第一置信度信息，以及待处理数据生命周期未结束的第二置信度信息基础上，可基于所述第一置信度信息及所述第二置信度信息的取值大小并结合预定的状态确定策略，确定所述待处理数据的生命周期状态。

[0117] 示例性地，所述状态确定策略可以是：待处理数据的生命周期状态为数值较大的置信度所对应的状态。例如，假设待处理数据对应于生命周期未结束这一状态的置信度为70％，对应于生命周期已结束这一状态的置信度为30％，则可确定出待处理数据的生命周期状态为生命周期未结束。

[0118] 所述状态确定策略还可以是：待处理数据的生命周期状态为数值较大的置信度所对应的状态，且为了避免识别不够准确，限定所述数值较大的置信度需不低于预定阈值，否则若数值较大的置信度低于所述预定阈值，则认为未识别。

[0119] 例如，假设所述预定阈值为70％，若待处理数据对应于生命周期未结束这一状态的置信度为45％，对应于生命周期已结束这一状态的置信度为55％，则由于数值较大的置信度(55％)低于所述预定阈值为(70％)，从而所述待处理数据的状态未能识别，若待处理数据对应于生命周期未结束这一状态的置信度为10％，对应于生命周期已结束这一状态的置信度为90％，由于数值较大的置信度(90％)高于所述预定阈值为(70％)，从而待处理数据的数据状态为数值较大的置信度(90％)所对应的状态，即生命周期已结束的状态。

[0120] 具体实施时，可基于实际需求(是否允许有未能识别的情况存在等)，来确定所述状态确定策略，并不以本实施例提供的上述两种策略为限。

[0121] 由于本实施例在确定待处理数据的生命周期状态时，数据处理模型具体是利用了大数据在各维度的特征与大数据生命周期状态的对应关系规律，进行生命周期状态预测，从而与现有技术相比，既降低了人力成本，又可使得所确定的数据生命周期状态具有较高的准确度。

[0122] 参考图6，为本申请提供的一种数据处理装置实施例五的结构示意图，该装置可应用于智能手机、平板电脑、个人数字助理、笔记本、台式机或一体机等各种终端设备中，或者还可以应用于各种通用或专用服务器中。如图1所示，该数据处理装置包括：

[0123] 获取单元601，用于获得待处理数据。

[0124] 所述待处理数据可以是实际生产环境中所产生或创建的各类型数据，例如可以是但不限于数据表数据或日志文件的日志数据等。

[0125] 对于这些数据，为了达到特定的目的，往往存在获知其生命周期状态的需求，例如，为了删除生命周期已结束的数据(或称失效数据)，以释放存储空间，则需要获知数据的生命周期是否已结束；为了根据数据访问的频繁程度进行数据的分类存储，则需要获知这些数据是处于频繁访问状态还是很少访问状态还是已失效。针对该情况，本申请的目的就在于能够低人力成本、高准确率地确定出数据的生命周期状态。

[0126] 数据生命周期状态的确定需以数据生命周期状态的划分为基础，划分方式不同，在对数据进行生命周期状态预测时，所对应的候选状态不同。

[0127] 例如，按数据失效与否进行划分，可将数据的生命周期状态划分为生命周期未结束及生命周期已结束这两种状态；按数据访问的频繁程度进行划分，则可将数据的生命周期状态划分为频繁访问、很少访问以及失效等状态，具体应用中，可根据实际的状态预测需

求选择合适的划分方式对数据的生命周期状态进行划分。

[0128] 提取单元602，用于提取所述待处理数据的预定特征的特征数据；所述特征数据包括以下至少之一：待处理数据在时间维度或数据访问频次维度的特征。

[0129] 待处理数据在时间维度的特征可以包括但不限于：待处理数据自创建时间点起至当前的存活时长、自最近一次访问的时间点起至当前的时长。

[0130] 待处理数据在数据访问频次维度的特征可以包括但不限于：待处理数据在至少一个预定时间段内的数据访问频次，例如待处理数据在最近1天、最近1周、最近15天、最近1月、最近1年或最近5年等时间粒度内的访问频次等。

[0131] 针对不同类型的待处理数据，当对其进行数据特征提取时，可以对其进行相同的数据特征提取，或者还可以对其进行不同的数据特征提取，本申请并不对此进行局限。

[0132] 具体地，例如针对数据表、日志文件这些不同类型的数据，可以设定两者需提取的特征均为上述时间维度和/或数据访问频次维度的特征。

[0133] 或者，针对数据表、日志文件这些不同类型的数据，还可以分别为其设定不同的需提取的特征，如对于数据表，可设定其需提取的特征为时间维度的特征、数据访问频次维度的特征以及价值维度的特征，对于日志文件，则可以设定其需提取的特征为时间维度的特征及数据访问频次维度的特征等。

[0134] 所述价值维度的特征，具体地，可以是但不限于数据的价值等级，例如预先设定的高、中、低三个价值等级等，示例性地，在进行数据特征提取时，针对中间数据的数据表，由于中间数据在实际生产环境中重要程度往往较低，企业人员或用户一般不会对中间数据给予过多关注，从而可将其价值特征的特征值提取为“低”，针对结果数据的数据表，由于结果数据在实际生产环境中重要程度往往较高，其一般是企业人员或用户的重点关注对象，从而可将其价值特征的特征值提取为“高”。

[0135] 处理单元603，用于利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果，以确定所述待处理数据的生命周期状态。

[0136] 所述数据处理模型，为预先利用数据表或日志文件等类型的大批量数据所训练的模型，该数据处理模型能够描述数据的数据特征(如时间维度和/或访问频次维度的特征等)与其生命周期状态间的对应关系规律。

[0137] 鉴于此，可将待处理数据的特征数据作为所述数据处理模型的输入提供给所述数据处理模型，由所述数据处理模型基于学习所得的数据特征与数据生命周期状态间的对应关系规律，输出能够表明所述待处理数据的生命周期状态的处理结果数据，进而可基于所述数据处理模型的处理结果数据，确定所述待处理数据的生命周期状态，如确定所述待处理数据的生命周期是否已结束，或者确定所述待处理数据是处于频繁访问、很少访问还是已失效的状态等等。

[0138] 根据以上方案可知，本申请提供的数据处理装置，在获得待处理数据后，提取待处理数据的特征数据，该特征数据包括待处理数据在时间维度或数据访问频次维度的特征中的至少之一；在此基础上，利用预先训练的数据处理模型对特征数据进行处理，以确定待处理数据的生命周期状态。本申请方案由于利用了预先训练的数据处理模型基于待处理数据在时间维度或数据访问频次维度的特征中的至少之一，来确定待处理数据的生命周期状态，从而，在确定待处理数据的生命周期状态时，具体是利用了大数据在时间维度或数据访

问频次维度的特征中的至少之一与大数据生命周期状态的对应关系的规律，这与现有技术相比，既降低了人力成本，又可使得所确定的数据生命周期状态具有较高的准确度。

[0139] 本申请的数据处理装置需要在预先训练出所述数据处理模型的基础上使用，即，如图2所示，需要利用训练样本预先训练出一数据处理模型，在此基础上，方可基于所述数据处理模型进行数据的生命周期状态预测。

[0140] 鉴于此，参考图7，为本申请提供的一种数据处理装置实施例六的结构示意图，本实施例中，所述数据处理装置还包括：

[0141] 预处理单元604，用于：

[0142] 获得多条训练样本；提取每条训练样本的所述预定特征的特征数据；所述训练样本的特征数据包括以下至少之一：训练样本在所述时间维度或所述数据访问频次维度的特征；标注每条训练样本的数据生命周期状态，得到每条训练样本的数据生命周期状态标注结果；建立每条训练样本的特征数据与数据生命周期状态标注结果间的对应关系，得到每条训练样本的特征数据与数据生命周期状态标注结果的对应关系数据；基于预定的机器学习算法，利用各条训练样本的所述对应关系数据训练一数据处理模型，使得所述数据处理模型能够基于输入的特征数据输出相应的数据生命周期状态预测结果。

[0143] 所述训练样本可以是但不限于从实际生产环境中随机选择的批量数据，如批量的日志文件数据，或者批量的数据表数据等。

[0144] 需要说明的是，所述训练样本的数据类型应视拟训练的数据处理模型所需具备的功能而定，若拟训练的数据处理模型需具备数据表数据的生命周期状态预测功能，则所述训练样本应是数据表类型的数据，若拟训练的数据处理模型需具备日志数据的生命周期状态预测功能，则所述训练样本应是日志类型的数据。

[0145] 训练样本在时间维度的特征可以包括但不限于：训练样本自创建时间点起至当前的存活时长、自最近一次访问的时间点起至当前的时长。

[0146] 训练样本在数据访问频次维度的特征可以包括但不限于：训练样本在至少一个预定时间段内的数据访问频次，如训练样本在最近1天、最近1周、最近15天、最近1月、最近1年、最近5年等时间粒度内的访问频次等。

[0147] 针对不同类型的训练样本，当对其进行数据特征提取时，可以对其进行相同的数据特征提取，或者还可以对其进行不同的数据特征提取，本申请并不对此进行局限。

[0148] 具体地，例如针对数据表、日志文件这些不同类型的训练样本，可以设定两者需提取的特征均为上述时间维度和/或数据访问频次维度的特征。

[0149] 或者，针对数据表、日志文件这些不同类型的训练样本，还可以分别为其设定不同的需提取的特征，如对于数据表，可设定其需提取的特征为时间维度的特征、数据访问频次维度的特征以及价值维度的特征，对于日志文件，则可以设定其需提取的特征为时间维度的特征及数据访问频次维度的特征等。

[0150] 所述价值维度的特征，具体地，可以是但不限于数据的价值等级，例如预先设定的高、中、低三个价值等级等，示例性地，在对训练样本进行数据特征提取时，针对中间数据的数据表，由于中间数据在实际生产环境中重要程度往往较低，企业人员或用户一般不会对中间数据给予过多关注，从而可将其价值特征的特征值提取为“低”，针对结果数据的数据表，由于结果数据在实际生产环境中重要程度往往较高，其一般是企业人员或用户的重点

关注对象，从而可将其价值特征的特征值提取为“高”。

[0151] 具体可采用人工标注等方式来标注每条训练样本的生命周期状态。

[0152] 训练样本的生命周期状态标注结果应视其生命周期状态的划分方式而定，示例性地，若按数据失效与否将其划分为生命周期已结束及生命周期未结束两种状态，则训练样本的生命周期状态标注结果应是该两种状态的其中之一，若其生命周期按访问的频繁程度划分为频繁访问、很少访问及已失效这三种状态，则训练样本的生命周期状态标注结果应是该三种状态的其中之一。

[0153] 需要说明的是，本步骤提供的标注训练样本生命周期状态的处理过程与上一步骤提供的提取训练样本数据特征的处理过程，不必局限于本实施例提供的先后执行次序，实际应用中，既可以先执行特征提取过程，后执行状态标注过程，还可以先执行状态标注过程，后执行特征提取过程，或者还可以同时执行上述两个处理过程，本实施例对此不作限定。

[0154] 在提取出训练样本的特征数据并标注出训练样本的生命周期状态的基础上，可建立每条训练样本的特征数据与其生命周期状态标注结果间的对应关系，以使得为数据处理模型的训练提供支持。

[0155] 所述机器学习算法可以是但不限于随机森林、KNN(k-NearestNeighbor，k最邻近分类算法)、逻辑回归、SVM(Support Vector Machine，支持向量机)等中的任意一种，具体应用中，可由技术人员基于实际的模型特点需求，选择合适的机器学习算法。

[0156] 在完成建立每条训练样本的特征数据与其生命周期状态标注结果间的对应关系后，可对所选择的机器学习算法进行模型参数的初始化，如初始化模型的各特征权重，以及初始化模型的学习速率这些超参数等，在此基础上，可利用各条训练样本的特征数据与其生命周期状态标注结果间的对应关系，对初始化后的模型进行训练，以使得数据处理模型能够不断学习各训练样本的数据特征与其生命周期状态间的对应关系的规律，进而不断调整/优化各特征的特征权重，最终得到能够基于输入的特征数据输出相应的生命周期状态预测结果的数据处理模型。

[0157] 训练所得的所述数据处理模型能够处理的数据的数据类型，一般来说需与训练样本的数据类型相一致，例如，若训练样本的数据类型为数据表类型，则训练所得的数据处理模型能够处理数据表数据，即能够依据输入的数据表数据的数据特征，输出数据表数据的生命周期状态预测结果；若训练样本的数据类型为日志类型，则训练所得的数据处理模型能够处理日志类型的数据，即能够依据输入的日志数据的数据特征，输出日志数据的生命周期状态预测结果。

[0158] 也就是说，基于不同数据类型的训练样本训练所得的数据处理模型，一般来说仅适用于与训练样本数据类型相同的数据的生命周期状态预测。

[0159] 为了向用户提供同时适用于不同类型数据的生命周期状态预测功能，还可以将训练所得的分别适用于不同数据类型的数据处理模型作为子模型进行集成，在对各个子模型完成集成后得到一总模型，该总模型则能够同时适用于对不同类型的数据进行生命周期状态预测，实际应用中可通过为所述总模型提供一个位于各子模型上层的总接口(如设计一个面向各类型数据的数据输入及状态预测界面等)，使得为用户提供同时适用于不同类型数据的生命周期状态预测功能，该总接口至少应具备识别/判断待预测数据的数据类型，并

根据待预测数据的数据类型调起相应的子模型进行数据生命周期状态预测的功能。

[0160] 本实施例通过基于预定的机器学习算法，利用相应数据类型的训练样本进行数据模型训练，为所述相应数据类型的数据提供了生命周期状态预测功能；通过将分别适用于不同数据类型的数据处理模型作为子模型进行集成，并为集成所得的总模型提供一个位于各子模型上层的总接口，为多种类型的数据提供了通用的生命周期状态预测功能。

[0161] 在本申请接下来的实施例七中，所述数据处理模型优选地为能够同时适用于对不同数据类型(如数据表数据、日志数据)的数据进行生命周期状态预测的模型，具体地，所述数据处理模型包括多于一个的子处理模型，不同的子处理模型与不同的数据类型相对应，用于对不同数据类型的数据进行生命周期状态预测，且不同的子处理模型所对应的数据特征类型和/或特征权重不同。

[0162] 在此基础上，所述处理单元603利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果，具体包括：

[0163] 获得所述待处理数据的数据类型；从所述数据处理模型中确定出与所述数据类型相对应的子处理模型；其中，所述数据处理模型包括多于一个的子处理模型，不同的子处理模型与不同的数据类型相对应，且不同的子处理模型所对应的数据特征类型和/或特征权重不同；将所述待处理数据的特征数据输入与所述数据类型相对应的所述子处理模型中，得到所述待处理数据的处理结果。

[0164] 所述待处理数据可以是但不限于数据表数据或者日志数据。

[0165] 其中，可以但不限于基于待处理数据的数据格式、待处理数据所在文档的文档类型(如数据表数据及日志数据在数据格式上以及文档类型上一般有较大差别)或者预先标注的类型信息等，识别待处理数据的数据类型。

[0166] 在获得所述待处理数据的数据类型后，可根据该数据类型，从所述数据处理模型包括的多个子处理模型中确定出与其相匹配的子处理模型。

[0167] 例如，如果所述待处理数据的数据类型为数据表数据，则可从所述数据处理模型包括的多个子处理模型中确定出能够对数据表数据进行生命周期状态预测的第一子处理模型；如果所述待处理数据的数据类型为日志数据，则可从所述数据处理模型包括的多个子处理模型中确定出能够对日志数据进行生命周期状态预测的第二子处理模型。

[0168] 在确定出与待处理数据的数据类型相对应的子处理模型后，可将待处理数据的特征数据输入该子处理模型中，最终由该子处理模型基于数据特征与数据生命周期状态间的对应关系的规律，输出与待处理数据的特征数据(如时间维度的特征、数据访问频次维度的特征和/或价值维度的特征等)相对应的生命周期状态预测结果。

[0169] 其中，所输出的生命周期状态预测结果，会因数据生命周期状态划分方式的不同而不同。

[0170] 示例性地，若数据生命周期状态按失效与否划分为生命周期已结束、生命周期未结束两种状态，则所述生命周期状态预测结果包括分别对应于数据生命周期已结束的置信度信息及生命周期未结束的置信度信息；若数据生命周期状态按访问频繁程度划分为频繁访问、很少访问及失效三种状态，则所述生命周期状态预测结果包括分别对应于频繁访问、很少访问及失效三种状态的三种置信度信息。

[0171] 本实施例通过使用能够适用于不同数据类型的数据处理模型，为不同类型数据提

供了通用的生命周期状态预测功能。且由于本实施例是利用了大数据在时间维度、数据访问频次维度和/或价值维度的特征与大数据生命周期状态的对应关系规律，对待处理数据进行生命周期状态预测，从而与现有技术相比，既降低了人力成本，又可使得所确定的数据生命周期状态具有较高的准确度。

[0172] 在本申请接下来的实施例八中，以数据的生命周期状态按失效与否被划分为生命周期已结束、生命周期未结束两种状态为例，介绍处理单元603根据所述数据处理模型的处理结果，确定待处理数据的生命周期状态的实现过程。其中，所述处理单元603可以通过以下处理过程来确定待处理数据的生命周期状态：

[0173] 获得所述处理结果中包括的待处理数据生命周期已结束的第一置信度信息，及待处理数据生命周期未结束的第二置信度信息；基于所述第一置信度信息及所述第二置信度信息，确定所述待处理数据的生命周期状态。

[0174] 在数据的生命周期状态按失效与否被划分为生命周期已结束、生命周期未结束两种状态的情况下，一般来说，数据处理模型在对待处理数据进行生命周期状态预测后可输出以下结果数据：待处理数据生命周期已结束的第一置信度信息、待处理数据生命周期未结束的第二置信度信息。

[0175] 所述第一置信度信息、第二置信度信息，可以是但不限于概率值或百分比数值等能够表明待处理数据属于相应状态的可能性程度的信息。

[0176] 在获得待处理数据生命周期已结束的第一置信度信息，以及待处理数据生命周期未结束的第二置信度信息基础上，可基于所述第一置信度信息及所述第二置信度信息的取值大小并结合预定的状态确定策略，确定所述待处理数据的生命周期状态。

[0177] 示例性地，所述状态确定策略可以是：待处理数据的生命周期状态为数值较大的置信度所对应的状态。例如，假设待处理数据对应于生命周期未结束这一状态的置信度为70％，对应于生命周期已结束这一状态的置信度为30％，则可确定出待处理数据的生命周期状态为生命周期未结束。

[0178] 所述状态确定策略还可以是：待处理数据的生命周期状态为数值较大的置信度所对应的状态，且为了避免识别不够准确，限定所述数值较大的置信度需不低于预定阈值，否则若数值较大的置信度低于所述预定阈值，则认为未识别。

[0179] 例如，假设所述预定阈值为70％，若待处理数据对应于生命周期未结束这一状态的置信度为45％，对应于生命周期已结束这一状态的置信度为55％，则由于数值较大的置信度(55％)低于所述预定阈值为(70％)，从而所述待处理数据的状态未能识别，若待处理数据对应于生命周期未结束这一状态的置信度为10％，对应于生命周期已结束这一状态的置信度为90％，由于数值较大的置信度(90％)高于所述预定阈值为(70％)，从而待处理数据的数据状态为数值较大的置信度(90％)所对应的状态，即生命周期已结束的状态。

[0180] 具体实施时，可基于实际需求(是否允许有未能识别的情况存在等)，来确定所述状态确定策略，并不以本实施例提供的上述两种策略为限。

[0181] 由于本实施例在确定待处理数据的生命周期状态时，数据处理模型具体是利用了大数据在各维度的特征与大数据生命周期状态的对应关系规律，进行生命周期状态预测，从而与现有技术相比，既降低了人力成本，又可使得所确定的数据生命周期状态具有较高的准确度。

[0182] 需要说明的是，本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0183] 为了描述的方便，描述以上系统或装置时以功能分为各种模块或单元分别描述。当然，在实施本申请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。

[0184] 通过以上的实施方式的描述可知，本领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件平台的方式来实现。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，该计算机软件产品可以存储在存储介质中，如ROM/RAM、磁碟、光盘等，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或者实施例的某些部分所述的方法。

[0185] 最后，还需要说明的是，在本文中，诸如第一、第二、第三和第四等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，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、物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、物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、物品或者设备中还存在另外的相同要素。

[0186] 以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

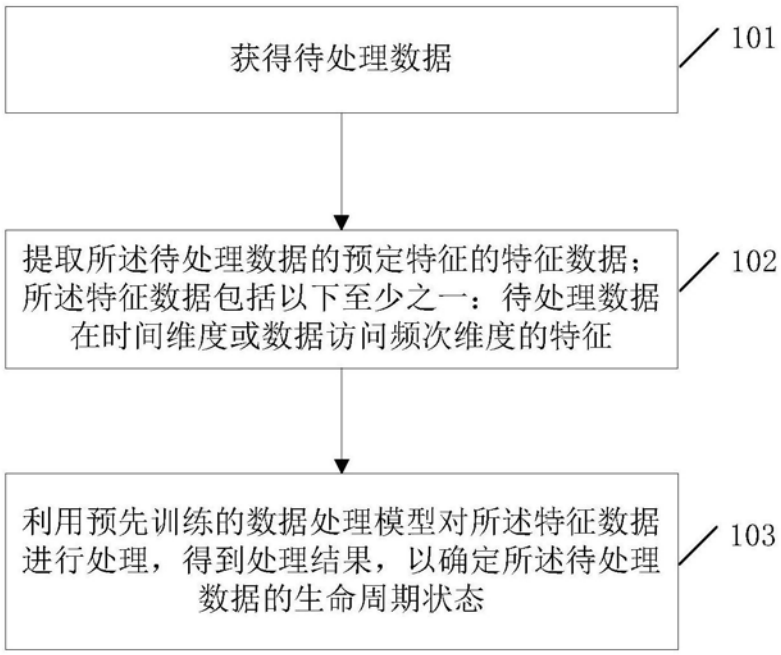

获得待处理数据
101
提取所述待处理数据的预定特征的特征数据；所述特征数据包括以下至少之一：待处理数据在时间维度或数据访问频次维度的特征
102
利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果，以确定所述待处理数据的生命周期状态
103

图1

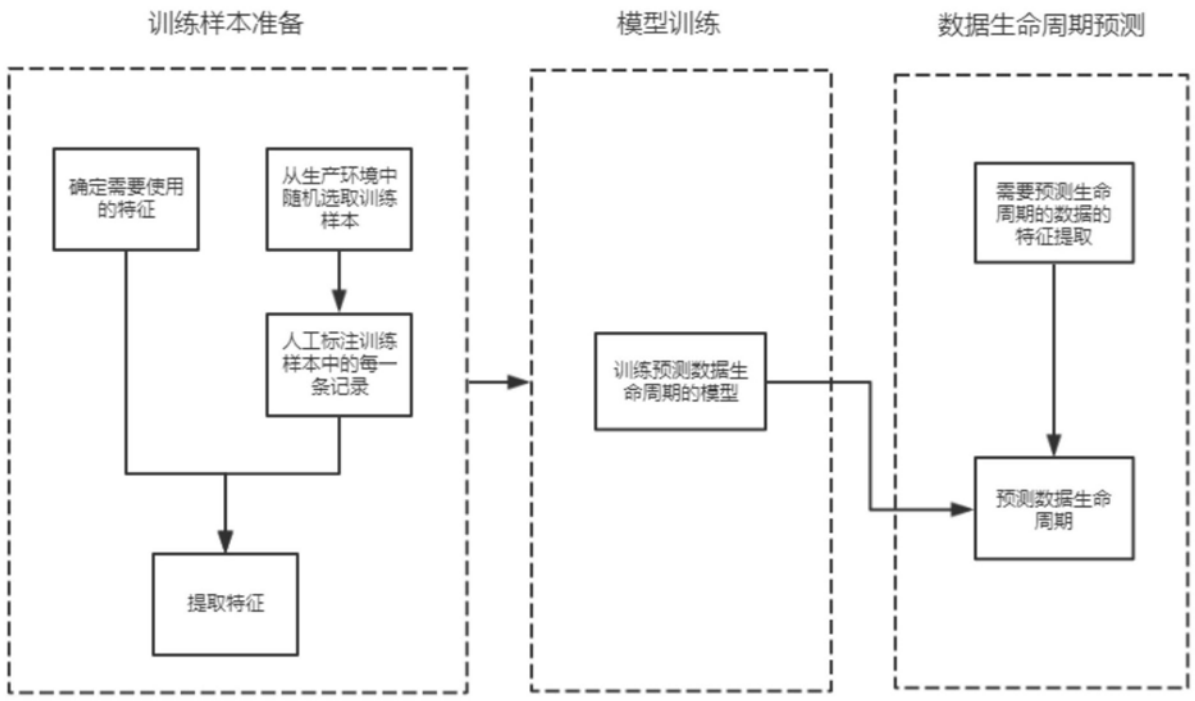

训练样本准备
模型训练
数据生命周期预测
确定需要使用的特征
从生产环境中随机选取训练样本
人工标注训练样本中的每一条记录
提取特征
训练预测数据生命周期的模型
需要预测生命周期的数据的特征提取
预测数据生命周期

图2

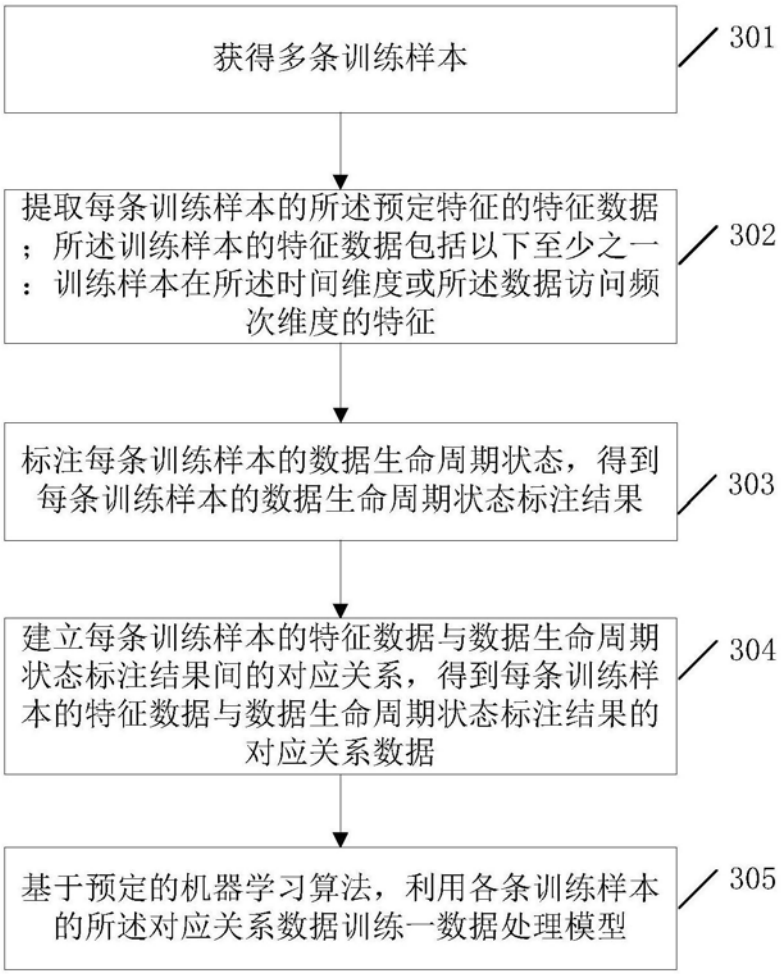

获得多条训练样本
301
提取每条训练样本的所述预定特征的特征数据；所述训练样本的特征数据包括以下至少之一：训练样本在所述时间维度或所述数据访问频次维度的特征
302
标注每条训练样本的数据生命周期状态，得到每条训练样本的数据生命周期状态标注结果
303
建立每条训练样本的特征数据与数据生命周期状态标注结果间的对应关系，得到每条训练样本的特征数据与数据生命周期状态标注结果的对应关系数据
304
基于预定的机器学习算法，利用各条训练样本的所述对应关系数据训练一数据处理模型
305

图3

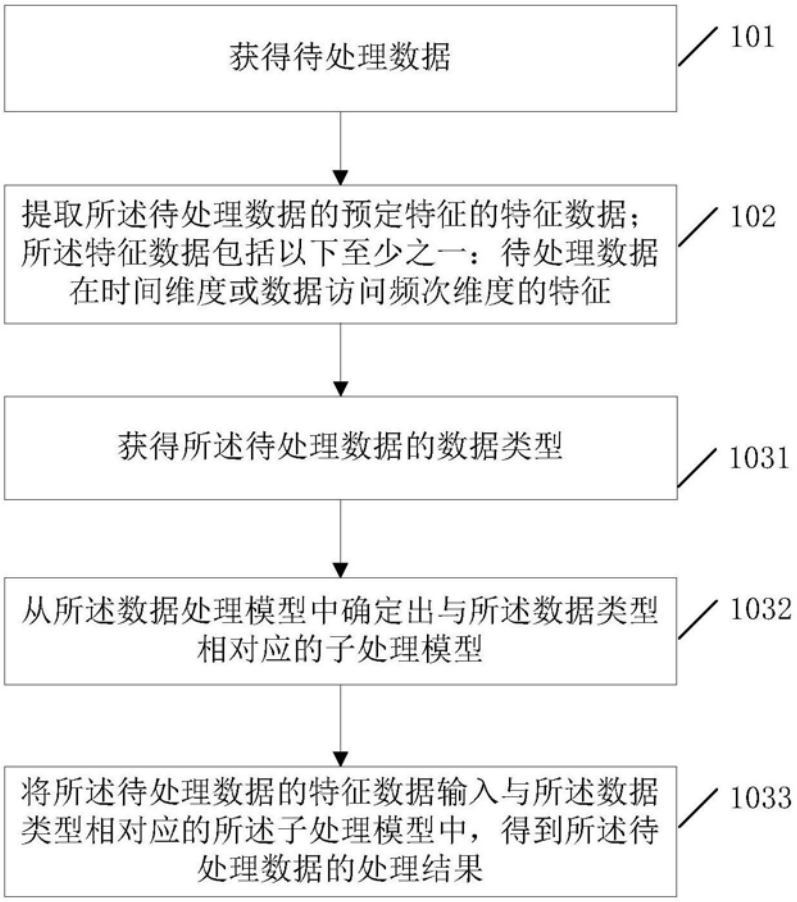

获得待处理数据
101
提取所述待处理数据的预定特征的特征数据；所述特征数据包括以下至少之一：待处理数据在时间维度或数据访问频次维度的特征
102
获得所述待处理数据的数据类型
1031
从所述数据处理模型中确定出与所述数据类型相对应的子处理模型
1032
将所述待处理数据的特征数据输入与所述数据类型相对应的所述子处理模型中，得到所述待处理数据的处理结果
1033

图4

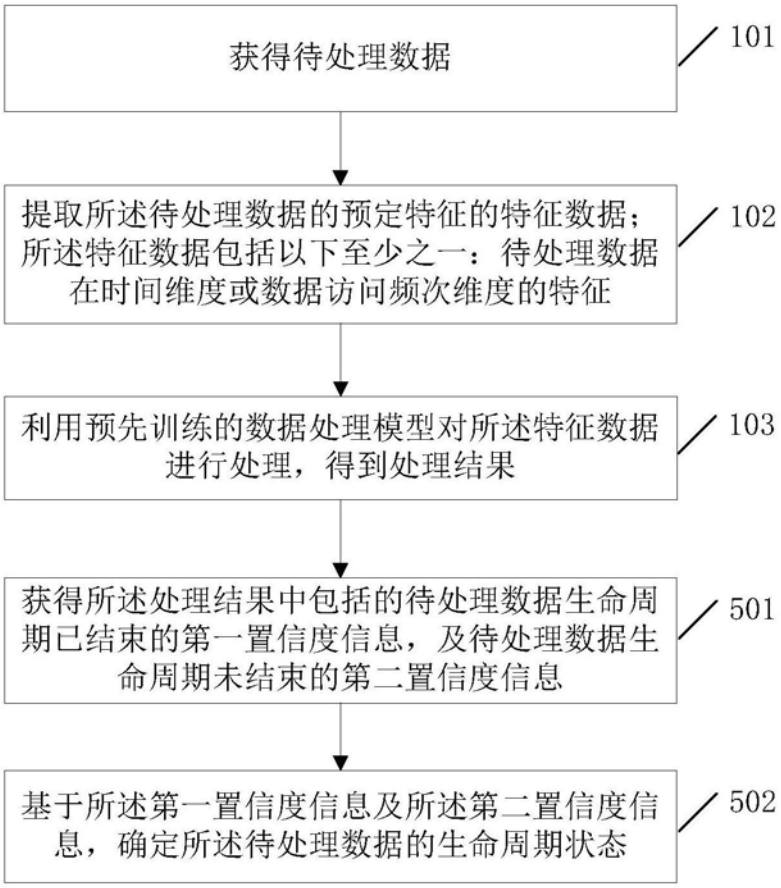

获得待处理数据
101
提取所述待处理数据的预定特征的特征数据；所述特征数据包括以下至少之一：待处理数据在时间维度或数据访问频次维度的特征
102
利用预先训练的数据处理模型对所述特征数据进行处理，得到处理结果
103
获得所述处理结果中包括的待处理数据生命周期已结束的第一置信度信息，及待处理数据生命周期未结束的第二置信度信息
501
基于所述第一置信度信息及所述第二置信度信息，确定所述待处理数据的生命周期状态
502

图5

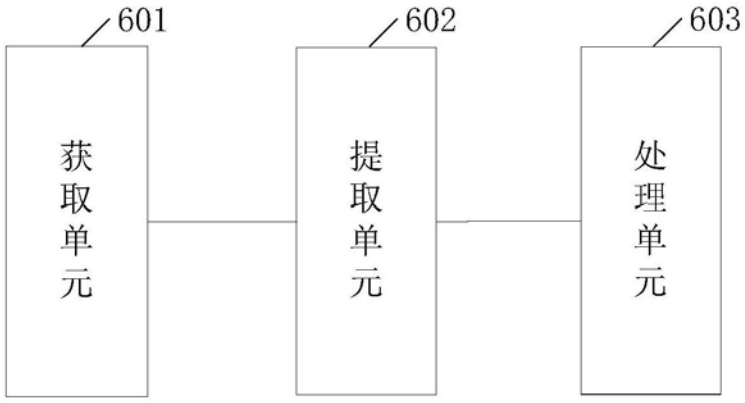

601
602
603
获取单元
提取单元
处理单元

图6

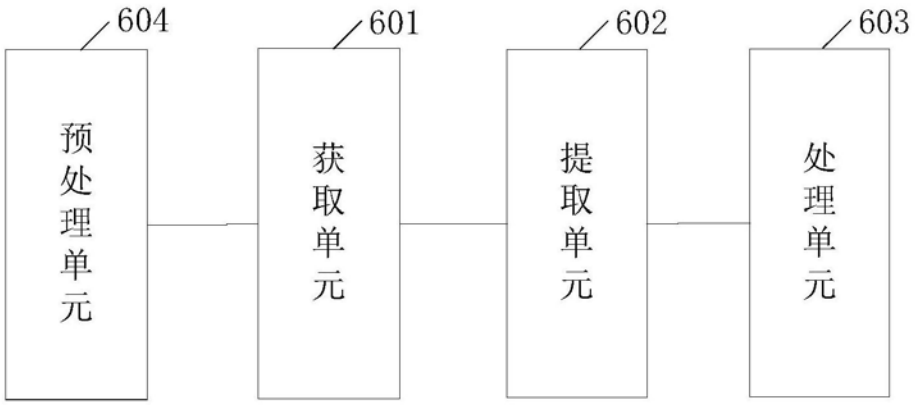


图7