(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 111666171 A
(43)申请公布日 2020.09.15

(21)申请号 202010502727.5

(22)申请日 2020.06.04

(71)申请人 中国工商银行股份有限公司
地址 100140 北京市西城区复兴门内大街55号

(72)发明人 徐晨灿 夏刚 袁宁 宫晨

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 鄢功军

(51)Int.Cl.
G06F 11/07(2006.01)
G06F 16/35(2019.01)

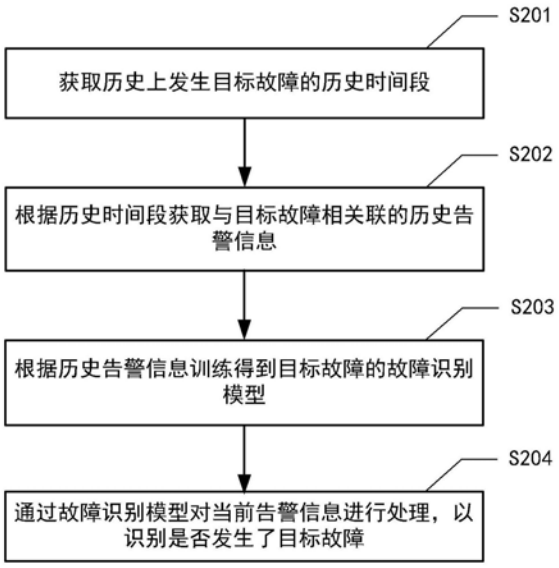
权利要求书3页 说明书18页 附图8页

(54)发明名称

故障识别方法及装置、电子设备和可读存储介质

(57)摘要

本公开提供了一种故障识别方法，包括：获取历史上发生目标故障的历史时间段；根据历史时间段获取与目标故障相关联的历史告警信息；根据历史告警信息训练得到目标故障的故障识别模型；以及通过故障识别模型对当前告警信息进行处理，以识别是否发生了目标故障。本公开还提供了一种故障识别装置、一种电子设备和一种计算机可读存储介质。



CN 111666171 A

1.一种故障识别方法,包括:

获取历史上发生目标故障的历史时间段;

根据所述历史时间段获取与所述目标故障相关联的历史告警信息;

根据所述历史告警信息训练得到所述目标故障的故障识别模型;以及

通过所述故障识别模型对当前告警信息进行处理,以识别是否发生了所述目标故障。

2.根据权利要求1所述的方法,其中,

所述根据所述历史时间段获取与所述目标故障相关联的历史告警信息包括:

确定包含所述历史时间段的告警时间段;以及

获取所述告警时间段内发生的历史告警信息;

所述根据所述历史告警信息训练得到所述目标故障的故障识别模型包括:

按照预设分割时长将所述告警时间段分割为多个子时间段;

将与所述历史时间段存在交叉,且在一个子时间段内的告警条数大于预设条数的子时间段确定为故障时间段;

将与所述历史时间段不存在交叉,或者在一个子时间段内的告警条数小于或等于所述预设条数的子时间段确定为非故障时间段;以及

根据所述故障时间段内的告警信息和所述非故障时间段内的告警信息训练得到所述目标故障的故障识别模型。

3.根据权利要求2所述的方法,其中,所述根据所述故障时间段内的告警信息和所述非故障时间段内的告警信息训练得到所述目标故障的故障识别模型包括:

从所述故障时间段内的告警信息中选取特征词的多个候选词;

确定每个所述候选词的相似词;

计算每个所述候选词的软词频和软逆文档频率;

根据每个所述候选词的软词频和软逆文档频率确定所述目标故障的特征词;以及

根据所述故障时间段内关于所述目标故障的特征词和所述非故障时间段内关于所述目标故障的特征词,训练得到所述目标故障的故障识别模型。

4.根据权利要求3所述的方法,其中,所述故障时间段包括多个,所述从所述故障时间段内的告警信息中选取特征词的候选词包括:

获取多个所述故障时间段中每个所述故障时间段内的告警信息的文本;

对每个所述故障时间段内的告警信息的文本进行分词;以及

根据多个所述故障时间段内的每个词汇的总出现次数选取候选词。

5.根据权利要求3所述的方法,其中,所述根据每个所述候选词的软词频和软逆文档频率确定所述目标故障的特征词包括:

将所述候选词在所述故障时间段的软词频与所述候选词在所述非故障时间段的软逆文档频率的乘积作为所述候选词在所述故障时间段内的软TF-IDF值;以及

根据每个所述候选词在不同所述故障时间段内的软TF-IDF值确定所述目标故障的特征词。

6.根据权利要求3所述的方法,其中,所述计算每个所述候选词的软词频包括:

将所述故障时间段内的多条所述告警信息进行排序;

按照排序顺序遍历每条所述告警信息,采用衰减策略累加每个所述候选词和所述候选

词的相似词的频次，其中，所述衰减策略包括根据每个所述候选词和所述候选词的相似词在每条所述告警信息中出现的次序和每条所述告警信息指向的同一个网络地址出现的次数，计算用于统计每个所述候选词和所述候选词的相似词在每条所述告警信息中的频次的贡献值；以及

在遍历完多条所述告警信息后，将累加得到的所述候选词和所述候选词的相似词的频次作为所述候选词在所述故障时间段的软词频。

7.根据权利要求3所述的方法，其中，所述非故障时间段包括多个，所述计算每个所述候选词的软逆文档频率包括：

从多个所述非故障时间段选取预设数量个非故障时间段；

按照如下方式计算每个所述候选词在一个所述非故障时间段内的告警信息中的软词频：

对所述非故障时间段内的多条告警信息进行排序；

按照排序顺序，对于所述非故障时间段内的每条告警信息，采用衰减策略累加每个所述候选词和所述候选词的相似词的频次，其中，所述衰减策略包括根据每个所述候选词和所述候选词的相似词在每条所述告警信息中出现的次序、每条所述告警信息指向的同一个网络地址出现的次数和不同网络地址在所述非故障时间段内的出现的次序，计算用于统计每个所述候选词和所述候选词的相似词在每条所述告警信息中的频次的贡献值；

在遍历完多条所述告警信息后，将累加得到的所述候选词和所述候选词的相似词的频次作为所述候选词在所述非故障时间段的软词频；

将所述候选词和所述候选词的相似词在每个所述非故障时间段内的软词频进行求和，得到软词频总和；

计算所述候选词和所述候选词的相似词在所述非故障时间段内的软词频不为0的非故障时间段的个数；以及

根据所述预设数量、所述软词频总和、所述软词频不为0的非故障时间段的个数计算所述候选词的软逆文档频率。

8.根据权利要求2所述的方法，其中，所述根据所述故障时间段内的告警信息和所述非故障时间段内的告警信息训练得到所述目标故障的故障识别模型包括：

获取所述故障时间段内的告警信息中的特征词对应的第一特征向量，其中，所述第一特征向量的标签为故障；

获取所述非故障时间段内的告警信息中的特征词对应的第二特征向量，其中，所述第二特征向量的标签为非故障；以及

将所述第一特征向量和所述第一特征向量对应的标签，所述第二特征向量和所述第二特征向量对应的标签输入到支持向量机中，以训练得到所述目标故障的故障识别模型。

9.根据权利要求8所述的方法，其中，所述获取所述非故障时间段内的告警信息中的特征词对应的第二特征向量包括：

计算所述非故障时间段内的告警信息中的每个特征词对应的第二特征向量的模；以及

根据每个所述特征词对应的第二特征向量的模的大小选择指定数量的第二特征向量。

10.一种故障识别装置，包括：

第一获取模块，用于获取历史上发生目标故障的历史时间段；

第二获取模块，用于根据所述历史时间段获取与所述目标故障相关联的历史告警信息；

训练模块，用于根据所述历史告警信息训练得到所述目标故障的故障识别模型；以及

处理模块，用于通过所述故障识别模型对当前告警信息进行处理，以识别是否发生了所述目标故障。

11.一种电子设备，包括：

一个或多个处理器；

存储器，用于存储一个或多个指令，

其中，当所述一个或多个指令被所述一个或多个处理器执行时，使得所述一个或多个处理器实现权利要求1至9中任一项所述的方法。

12.一种计算机可读存储介质，其上存储有可执行指令，该指令被处理器执行时使处理器实现权利要求1至9中任一项所述的方法。

故障识别方法及装置、电子设备和可读存储介质

技术领域

[0001] 本公开涉及计算机技术领域，更具体地，涉及一种故障识别方法、一种故障识别装置、一种电子设备和一种计算机可读存储介质。

背景技术

[0002] 信息系统安全稳定地运行对于业务来说至关重要，但信息系统的故障总会不可避免地发生。较为快速准确地识别故障有助于故障快速处理并恢复业务，降低业务影响。

[0003] 监控人员可以通过告警信息来发现信息系统运行中的异常，通过对告警信息进行分析来判断信息系统发生了什么故障。然而，当突然发生较多地或者大量告警时，监控人员难以从较多地或者大量告警中较为快速准确地分析出故障情况，导致故障恢复的时间延长，可能对业务产生不利影响。

[0004] 因此，突发较多地或者大量告警时的故障判断是一个亟待解决的技术问题。

发明内容

[0005] 有鉴于此，本公开提供了一种故障识别方法、一种故障识别装置、一种电子设备和一种计算机可读存储介质。

[0006] 本公开的一个方面提供了一种故障识别方法，包括：获取历史上发生目标故障的历史时间段；根据上述历史时间段获取与上述目标故障相关联的历史告警信息；根据上述历史告警信息训练得到上述目标故障的故障识别模型；以及通过上述故障识别模型对当前告警信息进行处理，以识别是否发生了上述目标故障。

[0007] 根据本公开的实施例，上述根据上述历史时间段获取与上述目标故障相关联的历史告警信息包括：确定包含上述历史时间段的告警时间段；以及获取上述告警时间段内发生的历史告警信息；上述根据上述历史告警信息训练得到上述目标故障的故障识别模型包括：按照预设分割时长将上述告警时间段分割为多个子时间段；将与上述历史时间段存在交叉，且在一个子时间段内的告警条数大于预设条数的子时间段确定为故障时间段；将与上述历史时间段不存在交叉，或者在一个子时间段内的告警条数小于或等于上述预设条数的子时间段确定为非故障时间段；以及根据上述故障时间段内的告警信息和上述非故障时间段内的告警信息训练得到上述目标故障的故障识别模型。

[0008] 根据本公开的实施例，上述根据上述故障时间段内的告警信息和上述非故障时间段内的告警信息训练得到上述目标故障的故障识别模型包括：从上述故障时间段内的告警信息中选取特征词的多个候选词；确定每个上述候选词的相似词；计算每个上述候选词的软词频和软逆文档频率；根据每个上述候选词的软词频和软逆文档频率确定上述目标故障的特征词；以及根据上述故障时间段内关于上述目标故障的特征词和上述非故障时间段内关于上述目标故障的特征词，训练得到上述目标故障的故障识别模型。

[0009] 根据本公开的实施例，上述故障时间段包括多个，上述从上述故障时间段内的告警信息中选取特征词的候选词包括：获取多个上述故障时间段中每个上述故障时间段内的

告警信息的文本;对每个上述故障时间段内的告警信息的文本进行分词;以及根据多个上述故障时间段内的每个词汇的总出现次数选取候选词。

[0010] 根据本公开的实施例,上述根据每个上述候选词的软词频和软逆文档频率确定上述目标故障的特征词包括:将上述候选词在上述故障时间段的软词频与上述候选词在上述非故障时间段的软逆文档频率的乘积作为上述候选词在上述故障时间段内的软TF-IDF值;以及根据每个上述候选词在不同上述故障时间段内的软TF-IDF值确定上述目标故障的特征词。

[0011] 根据本公开的实施例,上述计算每个上述候选词的软词频包括:将上述故障时间段内的多条上述告警信息进行排序;按照排序顺序遍历每条上述告警信息,采用衰减策略累加每个上述候选词和上述候选词的相似词的频次,其中,上述衰减策略包括根据每个上述候选词和上述候选词的相似词在每条上述告警信息中出现的次序和每条上述告警信息指向的同一个网络地址出现的次数,计算用于统计每个上述候选词和上述候选词的相似词在每条上述告警信息中的频次的贡献值;以及在遍历完多条上述告警信息后,将累加得到的上述候选词和上述候选词的相似词的频次作为上述候选词在上述故障时间段的软词频。

[0012] 根据本公开的实施例,上述非故障时间段包括多个,上述计算每个上述候选词的软逆文档频率包括:从多个上述非故障时间段选取预设数量个非故障时间段;

[0013] 按照如下方式计算每个上述候选词在一个上述非故障时间段内的告警信息中的软词频:对上述非故障时间段内的告警信息进行排序;按照排序顺序,对于上述非故障时间段内的每条告警信息,采用衰减策略累加每个上述候选词和上述候选词的相似词的频次,其中,上述衰减策略包括根据每个上述候选词和上述候选词的相似词在每条上述告警信息中出现的次序、每条上述告警信息指向的同一个网络地址出现的次数和不同网络地址在上述非故障时间段内的出现的次序,计算用于统计每个上述候选词和上述候选词的相似词在每条上述告警信息中的频次的贡献值;在遍历完多条上述告警信息后,将累加得到的上述候选词和上述候选词的相似词的频次作为上述候选词在上述非故障时间段的软词频;将上述候选词和上述候选词的相似词在每个上述非故障时间段内的软词频进行求和,得到软词频总和;计算上述候选词和上述候选词的相似词在上述非故障时间段内的软词频不为0的非故障时间段的个数;以及根据上述预设数量、上述软词频总和、上述软词频不为0的非故障时间段的个数计算上述候选词的软逆文档频率。

[0014] 根据本公开的实施例,上述根据上述故障时间段内的告警信息和上述非故障时间段内的告警信息训练得到上述目标故障的故障识别模型包括:获取上述故障时间段内的告警信息中的特征词对应的第一特征向量,其中,上述第一特征向量的标签为故障;获取上述非故障时间段内的告警信息中的特征词对应的第二特征向量,其中,上述第二特征向量的标签为非故障;以及将上述第一特征向量和上述第一特征向量对应的标签,上述第二特征向量和上述第二特征向量对应的标签输入到支持向量机中,以训练得到上述目标故障的故障识别模型。

[0015] 根据本公开的实施例,上述获取上述非故障时间段内的告警信息中的特征词对应的第二特征向量包括:计算上述非故障时间段内的告警信息中的每个特征词对应的第二特征向量的模;以及根据每个上述特征词对应的第二特征向量的模的大小选择指定数量的第二特征向量。

[0016] 本公开的另一个方面提供了一种故障识别装置，包括：第一获取模块，用于获取历史上发生目标故障的历史时间段；第二获取模块，用于根据上述历史时间段获取与上述目标故障相关联的历史告警信息；训练模块，用于根据上述历史告警信息训练得到上述目标故障的故障识别模型；以及处理模块，用于通过上述故障识别模型对当前告警信息进行处理，以识别是否发生了上述目标故障。

[0017] 本公开的另一个方面提供了一种电子设备，包括：一个或多个处理器；存储器，用于存储一个或多个指令，其中，当上述一个或多个指令被上述一个或多个处理器执行时，使得上述一个或多个处理器实现如上所述的方法。

[0018] 本公开的另一方面提供了一种计算机可读存储介质，存储有计算机可执行指令，上述指令在被执行时用于实现如上所述的方法。

[0019] 本公开的另一方面提供了一种计算机程序，上述计算机程序包括计算机可执行指令，上述指令在被执行时用于实现如上上述的方法。

[0020] 根据本公开的实施例，采用了根据历史上发生目标故障的历史时间段获取与目标故障相关联的历史告警信息；根据历史告警信息训练得到目标故障的故障识别模型；通过故障识别模型对当前告警信息进行处理，以识别是否发生了目标故障的技术手段，所以至少部分地克服了突发较多地或者大量告警时的故障判断的技术问题，进而达到了当目标故障发生时，可以及时地通知监控人员，进而缩短故障恢复时间，保障业务平稳运行的技术效果。

附图说明

[0021] 通过以下参照附图对本公开实施例的描述，本公开的上述以及其他目的、特征和优点将更为清楚，在附图中：

[0022] 图1示意性示出了根据本公开实施例的可以应用故障识别方法及装置的示例性系统架构；

[0023] 图2示意性示出了根据本公开实施例的故障识别方法的流程图；

[0024] 图3示意性示出了根据本公开实施例的训练得到目标故障的故障识别模型的流程图；

[0025] 图4示意性示出了根据本公开实施例的根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型的流程图；

[0026] 图5示意性示出了根据本公开另一实施例的根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型的流程图；

[0027] 图6示意性示出了根据本公开实施例的从故障时间段内的告警信息中选取特征词的候选词的流程图；

[0028] 图7示意性示出了根据本公开实施例的计算每个候选词的软词频的流程图；

[0029] 图8示意性示出了根据本公开实施例的计算每个候选词的软逆文档频率的流程图；

[0030] 图9示意性示出了根据本公开实施例的根据软TF-IDF值确定目标故障的特征词的流程图；

[0031] 图10示意性示出了根据本公开实施例的故障识别装置的框图；以及

[0032] 图11示意性示出了根据本公开实施例的适于实现上文描述的方法的计算机系统的框图。

具体实施方式

[0033] 以下，将参照附图来描述本公开的实施例。但是应该理解，这些描述只是示例性的，而并非要限制本公开的范围。在下面的详细描述中，为便于解释，阐述了许多具体的细节以提供对本公开实施例的全面理解。然而，明显地，一个或多个实施例在没有这些具体细节的情况下也可以被实施。此外，在以下说明中，省略了对公知结构和技术的描述，以避免不必要地混淆本公开的概念。

[0034] 在此使用的术语仅仅是为了描述具体实施例，而并非意在限制本公开。在此使用的术语“包括”、“包含”等表明了所述特征、步骤、操作和/或部件的存在，但是并不排除存在或添加一个或多个其他特征、步骤、操作或部件。

[0035] 在此使用的所有术语(包括技术和科学术语)具有本领域技术人员通常所理解的含义，除非另外定义。应注意，这里使用的术语应解释为具有与本说明书的上下文相一致的含义，而不应以理想化或过于刻板的方式来解释。

[0036] 在使用类似于“A、B和C等中至少一个”这样的表述的情况下，一般来说应该按照本领域技术人员通常理解该表述的含义来予以解释(例如，“具有A、B和C中至少一个的系统”应包括但不限于单独具有A、单独具有B、单独具有C、具有A和B、具有A和C、具有B和C、和/或具有A、B、C的系统等)。在使用类似于“A、B或C等中至少一个”这样的表述的情况下，一般来说应该按照本领域技术人员通常理解该表述的含义来予以解释(例如，“具有A、B或C中至少一个的系统”应包括但不限于单独具有A、单独具有B、单独具有C、具有A和B、具有A和C、具有B和C、和/或具有A、B、C的系统等)。

[0037] 本公开的实施例提供了一种故障识别方法，包括：获取历史上发生目标故障的历史时间段；根据历史时间段获取与目标故障相关联的历史告警信息；根据历史告警信息训练得到目标故障的故障识别模型；以及通过故障识别模型对当前告警信息进行处理，以识别是否发生了目标故障。

[0038] 图1示意性示出了根据本公开实施例的可以应用故障识别方法及装置的示例性系统架构100。需要注意的是，图1所示仅为可以应用本公开实施例的系统架构的示例，以帮助本领域技术人员理解本公开的技术内容，但并不意味着本公开实施例不可以用于其他设备、系统、环境或场景。

[0039] 如图1所示，根据该实施例的系统架构100可以包括终端设备101、102、103，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线和/或无线通信链路等等。

[0040] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。终端设备101、102、103上可以安装有各种信息系统，例如交易系统，数据库系统等其他业务系统。

[0041] 终端设备101、102、103可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0042] 服务器105可以是提供各种服务的服务器，例如对用户利用终端设备101、102、103

所浏览的网站提供支持的后台管理服务器(仅为示例)。后台管理服务器可以对接收到的用户请求等数据进行分析等处理,并将处理结果(例如根据用户请求获取或生成的网页、信息、或数据等)反馈给终端设备。

[0043] 需要说明的是,本公开实施例所提供的故障识别方法一般可以由服务器105执行。相应地,本公开实施例所提供的故障识别装置一般可以设置于服务器105中。本公开实施例所提供的故障识别方法也可以由不同于服务器105且能够与终端设备101、102、103和/或服务器105通信的服务器或服务器集群执行。相应地,本公开实施例所提供的故障识别装置也可以设置于不同于服务器105且能够与终端设备101、102、103和/或服务器105通信的服务器或服务器集群中。或者,本公开实施例所提供的故障识别方法也可以由终端设备101、102、或103执行,或者也可以由不同于终端设备101、102、或103的其他终端设备执行。相应地,本公开实施例所提供的故障识别装置也可以设置于终端设备101、102、或103中,或设置于不同于终端设备101、102、或103的其他终端设备中。

[0044] 例如,与目标故障相关联的历史告警信息可以原本存储在终端设备101、102、或103中的任意一个(例如,终端设备101,但不限于此)之中,或者存储在外部存储设备上并可以导入到终端设备101中。然后,终端设备101可以在本地执行本公开实施例所提供的故障识别方法,或者将与目标故障相关联的历史告警信息发送到其他终端设备、服务器、或服务器集群,并由接收该与目标故障相关联的历史告警信息的其他终端设备、服务器、或服务器集群来执行本公开实施例所提供的故障识别方法。

[0045] 应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器。

[0046] 信息系统安全稳定运行对于业务来说至关重要,但故障总会不可避免地发生,快速准确地识别故障有助于故障快速处理恢复,降低业务影响。对于那些重要的且会引发大量告警的故障,需要自动识别是否发生了这样的故障,从而缩短故障恢复时间,保障业务平稳运行。

[0047] 本公开的实施例提供了一种故障识别方法及装置,能够通过对告警信息的学习,建立目标故障的故障识别模型,从而在大量告警发生时能够快速准确地判断出是否发生了目标故障,进而缩短故障恢复时间,保障业务平稳运行。

[0048] 图2示意性示出了根据本公开实施例的故障识别方法的流程图。

[0049] 如图2所示,该方法包括操作S201～S204。

[0050] 在操作S201,获取历史上发生目标故障的历史时间段。

[0051] 根据本公开的实施例,对于某个目标故障,获取历史上发生目标故障的时间段。目标故障可以包括重要的且会引发大量告警的故障。历史上发生目标故障的时间段可以是由运维人员在故障发生后记录下来的,也可以通过关键词搜索历史告警选出备选时间段后再由运维人员进行确认得到的。

[0052] 在操作S202,根据历史时间段获取与目标故障相关联的历史告警信息。

[0053] 根据本公开的实施例,根据历史时间段获取与目标故障相关联的历史告警信息,而不是直接获取所有与目标故障相关联的信息,可以使得获取的告警信息与目标故障的关联度更高,从而可以提高目标故障的故障识别模型的识别准确度。

[0054] 在操作S203,根据历史告警信息训练得到目标故障的故障识别模型。

[0055] 根据本公开的实施例，可以利用支持向量机训练得到目标故障的故障识别模型。当然，本公开不限于支持向量机，也可以采用相关技术中的其他分类模型进行训练。

[0056] 在操作S204，通过故障识别模型对当前告警信息进行处理，以识别是否发生了目标故障。

[0057] 根据本公开的实施例，当前告警信息可以是实时扫描得到的告警信息，将告警信息输入到故障识别模型，得出识别结果后，可以通知监控人员。根据本公开的实施例，也可以仅在识别结果为目标故障的情况下，通知监控人员。

[0058] 根据本公开的实施例，在将告警信息输入到故障识别模型之前，可以预先对告警信息进行处理，例如，将告警信息进行向量化处理，然后将向量化处理后的告警信息输入故障识别模型。

[0059] 根据本公开的实施例，将实时扫描得到的告警输入到故障识别模型输出识别结果的流程如下：获取最近一段时间发生的告警信息，如3分钟，扫描时长可以与之前子时间段的长度保持一致。计算告警信息中的特征词在这一时间段的告警中软词频并组成特征向量。将该特征向量输入到故障识别模型，故障识别模型可以输出故障或非故障的结果。

[0060] 根据本公开的实施例，采用了根据历史上发生目标故障的历史时间段获取与目标故障相关联的历史告警信息；根据历史告警信息训练得到目标故障的故障识别模型；通过故障识别模型对当前告警信息进行处理，以识别是否发生了目标故障的技术手段，不需要大量人力的介入，自动从历史数据中学习故障识别模型，所以至少部分地克服了突发较多地或者大量告警时的故障判断的技术问题，进而达到了当目标故障发生时，可以及时地通知监控人员，进而缩短故障恢复时间，保障业务平稳运行的技术效果。进一步的，可以减轻运维人员工作量、提高监控及时性和有效性、减少故障恢复时间，进而提升安全生产水平，在监控告警行业具有重要的应用价值。

[0061] 下面参考图3～图9，结合具体实施例对图2所示的方法做进一步说明。

[0062] 图3示意性示出了根据本公开实施例的训练得到目标故障的故障识别模型的流程图。

[0063] 如图3所示，该方法包括操作S301～S306。

[0064] 在操作S301，确定包含历史时间段的告警时间段。

[0065] 在操作S302，获取告警时间段内发生的历史告警信息。

[0066] 根据本公开的实施例，操作S301～S302可以是对操作S202的进一步说明。

[0067] 根据本公开的实施例，例如，可以选定包含历史上发生目标故障的历史时间段的较长一段时间，然后获取选定的较长一段时间内的告警信息。

[0068] 在操作S303，按照预设分割时长将告警时间段分割为多个子时间段。

[0069] 根据本公开的实施例，可以按照设定的分割时长将选定的较长一段时间平均分割为多个小时间段(即子时间段)。设定的分割时长可根据具体情况调整，例如可以是3分钟。

[0070] 在操作S304，将与历史时间段存在交叉，且在一个子时间段内的告警条数大于预设条数的子时间段确定为故障时间段。

[0071] 在操作S305，将与历史时间段不存在交叉，或者在一个子时间段内的告警条数小于或等于预设条数的子时间段确定为非故障时间段。

[0072] 根据本公开的实施例，与历史上发生目标故障的历史时间段有交叉且告警条数大

于预设条数的小时间段可以记为故障时间段。

[0073] 根据本公开的实施例，与历史时间段没有交叉的，或者在一个子时间段内的告警条数小于或等于预设条数的子时间段可以记为非故障时间段。预设条数可根据具体情况调整，例如可以为20条。

[0074] 根据本公开的实施例，在一个子时间段内的告警条数小于或等于预设条数的子时间段也可以不作使用，即不将这些子时间段记为非故障时间段，而仅将与历史时间段没有交叉的子时间段确定为非故障时间段。根据本公开的另一实施例，也可以将与历史时间段有交叉但是告警条数小于或等于预设条数的子时间段不作使用。

[0075] 在操作S306，根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型。

[0076] 根据本公开的实施例，其中，操作S303～S306可以是对操作S203的进一步说明。

[0077] 根据本公开的实施例，通过历史时间段确定了划分故障与非故障的最佳超平面，能够在最大程度上区分两者，使得目标故障的故障识别模型具有更强的泛化能力。

[0078] 根据本公开的实施例，根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型包括：获取故障时间段内的告警信息中的特征词对应的第一特征向量，其中，第一特征向量的标签为故障；获取非故障时间段内的告警信息中的特征词对应的第二特征向量，其中，第二特征向量的标签为非故障；以及将第一特征向量和第一特征向量对应的标签，第二特征向量和第二特征向量对应的标签输入到支持向量机中，以训练得到目标故障的故障识别模型。

[0079] 根据本公开的实施例，获取非故障时间段内的告警信息中的特征词对应的第二特征向量包括：计算非故障时间段内的告警信息中的每个特征词对应的第二特征向量的模，根据每个特征词对应的第二特征向量的模的大小选择指定数量的第二特征向量。

[0080] 图4示意性示出了根据本公开实施例的根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型的流程图。

[0081] 如图4所示，该方法包括操作S401～S407。

[0082] 在操作S401，获取所有故障时间段的特征向量，标签为故障。

[0083] 在操作S402，计算所有非故障时间段的特征向量并计算特征向量的模的大小。

[0084] 在操作S403，选取非故障时间段的特征向量的模的大小排名靠前的指定数量的非故障时间段作为候选训练集。

[0085] 在操作S404，从候选训练集中随机获取指定数量个非故障时间段的特征向量，标签为非故障。非故障时间段的指定数量与故障时间段的数量在数量级上相当，当然，数量上可以相同。

[0086] 在操作S405，将故障时间段的特征向量和标签，非故障时间段的特征向量和标签输入到支持向量机训练。

[0087] 在操作S406，训练指定数量个支持向量机，每次随机选取非故障时间段。

[0088] 在操作S407，如果输出故障的支持向量机数量比输出非故障的多，那么判断为目标故障，否则判断为非目标故障。

[0089] 根据本公开的实施例，通过对获得的历史告警信息进行处理，分为故障时间段内的告警信息和非故障时间段内的告警信息，利用故障时间段内的告警信息和非故障时间段

内的告警信息训练得到目标故障的故障识别模型，而不是直接将获得的告警信息去训练故障识别模型而言，提出了一种较为有效的故障识别模型的模型训练方法。

[0090] 根据本公开的实施例，在得到一个或多个故障时间段和一个或多个非故障时间段之后，可以确定每个故障时间段内的告警信息的特征词，该特征词用于训练得到目标故障的故障识别模型。

[0091] 在得到每个故障时间段内的告警信息的特征词之前，可以先确定特征词的一个或多个候选词，然后从一个或多个候选词中选择较好的候选词作为特征词。

[0092] 图5示意性示出了根据本公开另一实施例的根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型的流程图。

[0093] 如图5所示，该方法包括操作S501～S505。

[0094] 在操作S501，从故障时间段内的告警信息中选取特征词的多个候选词。

[0095] 图6示意性示出了根据本公开实施例的从故障时间段内的告警信息中选取特征词的候选词的流程图。

[0096] 根据本公开的实施例，如图6所示，从故障时间段内的告警信息中选取特征词的候选词包括操作S601～操作S603。

[0097] 在操作S601，获取多个故障时间段中每个故障时间段内的告警信息的文本。

[0098] 在操作S602，对每个故障时间段内的告警信息的文本进行分词。

[0099] 根据本公开的实施例，在对告警文本进行分词时，可以去掉停用词。

[0100] 在操作S603，根据多个故障时间段内的每个词汇的总出现次数选取候选词。

[0101] 根据本公开的实施例，例如，包括10个故障时间段，可以将每个故障时间段内的告警信息的文本分别进行分词，然后统计分词得到的每个词汇在该10个故障时间段内的所有文本中出现的总出现次数，将每个词汇的总出现次数进行排序，选取排名靠前的一定数量的词作为候选词。候选词的个数可根据具体情况调整，可以为100个。

[0102] 在确定多个候选词之后，在操作S502，确定每个候选词的相似词。

[0103] 根据本公开的实施例，可以采用如下方式确定每个候选词的相似词。例如，先获取所有故障时间段和非故障时间段的告警文本，然后对告警文本进行分词，并去掉停用词，再然后分别从故障时间段和非故障时间段中选取总出现次数靠前的指定数量的词。选取所有故障时间段中总出现次数靠前的指定数量的词例如可以是选取故障时间段中总出现次数靠前的1000个词。选取非故障时间段中总出现次数靠前的指定数量的词例如可以是选取非故障时间段中总出现次数靠前的1000个词，当然，这些数值可以根据实际情况进行调整。然后将上述选取的词形成词汇表。获取词汇表中每个词的词向量，可以使用word2vec获取词向量。接下来，对于每个候选词，计算该词的词向量与词汇表中其他词的词向量的余弦相似度。选取相似度大于预设阈值的词作为该候选词的相似词。预设阈值可以根据实际情况设定。每个候选词可以具有相应的一个或多个相似词。

[0104] 根据本公开的实施例，由于分别从故障时间段和非故障时间段中选取总出现次数靠前的指定数量的词，而不是将故障时间段和非故障时间段内的所有词汇混合在一起之后，再按照总出现次数进行排序进行选词，可以尽量避免由于将故障时间段和非故障时间段内的所有词汇混合在一起后选词，而导致一些与目标故障强关联的词汇由于总出现次数较少，被一些与目标故障弱关联，但总出现次数较多的词汇所替代的问题，使得候选词的相

似词较为准确有效。

[0105] 在操作S503,计算每个候选词的软词频和软逆文档频率。

[0106] 根据本公开的实施例,可以把一个故障时间段内的所有告警信息当作一个文档,软词频是指在计算候选词的词频时,将候选词的相似词与候选词视为相同的词,把候选词和候选词的相似词都计入词频,同时,还可以考虑到文档内部的层次结构,采用衰减策略累加每个候选词和候选词的相似词的频次。

[0107] 根据本公开的实施例,可以把一个非故障时间段内的所有告警信息当作一个文档,软逆文档频率是指在计算候选词的逆文档频率时,将候选词的相似词与候选词视为相同的词,把候选词和候选词的相似词都计入,同时,还可以考虑到文档内部的层次结构。

[0108] 根据本公开的实施例,相比于相关技术中仅对特定词本身计算词频而言,得到的词频可以更好的反映出告警信息的特征,充分挖掘隐藏在告警文本中的故障特征,使特征向量更为充分地代表了目标故障的信息,减少了噪音。

[0109] 在操作S504,根据每个候选词的软词频和软逆文档频率确定目标故障的特征词。

[0110] 根据本公开的实施例,根据每个候选词的软词频和软逆文档频率确定目标故障的特征词包括:将候选词在故障时间段的软词频与候选词在非故障时间段的软逆文档频率的乘积作为候选词在故障时间段内的软TF-IDF值;根据每个候选词在不同故障时间段内的软TF-IDF值确定目标故障的特征词。

[0111] 根据本公开的实施例,TF-IDF(Term Frequency-Inverse Document Frequency)是指词频-逆文本频率指数,软TF-IDF值可以是指候选词的软词频-逆文本频率,即软词频与软逆文档频率的乘积。

[0112] 通过本公开的实施例,创造性地结合TF-IDF与词向量形成软TF-IDF的计算方法,为提取文档中重要特征词提供了更好的方式。

[0113] 根据本公开的实施例,可以根据每个候选词的软TF-IDF值的大小,选出目标故障的特征词。

[0114] 在操作S505,根据故障时间段内关于目标故障的特征词和非故障时间段内关于目标故障的特征词,训练得到目标故障的故障识别模型。

[0115] 通过构建故障识别模型能够快速准确地判断出是否发生了目标故障,当目标故障发生时,及时地通知监控人员,进而缩短故障恢复时间,保障业务平稳运行。

[0116] 图7示意性示出了根据本公开实施例的计算每个候选词的软词频的流程图。

[0117] 根据本公开的实施例,可以计算每个候选词在一个故障时间段的软词频。需要说明的是,故障时间段可以包括多个,在包括多个故障时间段的情况下,对于每个故障时间段,都要计算该候选词在该故障时间段的软词频。

[0118] 根据本公开的实施例,如图7所示,计算每个候选词的软词频包括操作S701～操作S703。

[0119] 在操作S701,将一个故障时间段内的多条告警信息进行排序。

[0120] 根据本公开的实施例,在将一个故障时间段内的多条告警信息进行排序之前,可以将一个故障时间段内出现过该候选词或其相似词的告警信息先筛选出来,然后对包括该候选词或其相似词的告警信息进行排序。

[0121] 根据本公开的实施例,将一个故障时间段内的多条告警信息进行排序的方式不做

限定,例如可以包括以下方式。

[0122] 例如,在一个故障时间段内可以包括一条或多条告警信息,可以计算候选词和候选词的相似词在一个故障时间段内的每条告警信息中出现的总次数,按照候选词和候选词的相似词在一个故障时间段内的每条告警信息中出现的总次数,将该故障时间段内的多条告警信息进行排序。

[0123] 根据本公开的实施例,也可以计算候选词和候选词的相似词分别在故障时间段内的每条告警信息中出现的次数。

[0124] 在计算候选词和候选词的相似词分别在故障时间段内的每条告警信息中出现的次数的情况下,可以先根据候选词出现次数优先将告警信息进行排序,在候选词出现次数相同时,再按相似词出现次数将候选词出现次数相同的告警信息进行排序。

[0125] 在操作S702,按照排序顺序遍历每条告警信息,采用衰减策略累加每个候选词和候选词的相似词的频次。其中,衰减策略包括根据每个候选词和候选词的相似词在每条告警信息中出现的次序和每条告警信息指向的网络地址出现的次数,计算用于统计每个候选词和候选词的相似词在每条告警信息中的频次的贡献值。

[0126] 在操作S703,在遍历完多条告警信息后,将累加得到的候选词和候选词的相似词的频次作为候选词在该故障时间段的软词频。

[0127] 根据本公开的实施例,在按照排序顺序遍历每条告警信息的过程中,在当前一条告警信息中遇到该候选词及其相似词时,如果该告警信息指向的IP地址是第n次出现,且该候选词及其相似词在该告警信息中第一次出现,那么在累加该候选词及其相似词的软词频贡献值时,需要加上相似度/n;如果该候选词及其相似词在同一条该告警信息中第2次出现,那么在累加该候选词及其相似词的软词频贡献值时,需要加上相似度/2n;如果该候选词及其相似词在同一条该告警信息中第k次出现,那么在累加该候选词及其相似词的软词频贡献值时,需要加上相似度/kn。

[0128] 需要说明的是,衰减策略具体可以包括指向不同IP地址的不同告警信息中的候选词及其相似词具有更大的权重,指向相同IP地址的相同告警信息中的候选词及其相似词随着告警信息的出现次序具有更小的权重的策略。

[0129] 上述衰减策略也可以称之为双层同质衰减策略,双层的第一层在IP地址,如果前述词(即候选词及其相似词)出现的两条告警信息有相同的IP地址,那么后一条告警信息中的候选词及其相似词的权重就会降低。双层的第二层在告警信息,如果前述词在同一条告警信息中出现多次,那么后面出现的权重就会降低。

[0130] 在上述示例中,两层的衰减函数是1/n,但是也可以使用其他衰减函数,不同的层也可以选择不同的衰减函数。

[0131] 根据本公开的实施例,采用衰减策略累加每个候选词和候选词的相似词的频次可以包括如下具体示例。

[0132] 在一故障时间段内,将告警信息进行排序后存在告警信息1、告警信息2和告警信息3。其中,告警信息1和告警信息2都是指向IP地址1的告警,告警信息3是指向IP地址2的告警。告警信息1包括第一候选词及其相似词1,告警信息2包括第一候选词及其相似词2,告警信息3包括第一候选词。

[0133] 针对告警信息1中的第一候选词,告警信息1指向的IP地址1是第一次出现,且该第

一候选词在告警信息1中第一次出现，那么在计算该第一候选词的软词频贡献值时，由于相似度为1，n也等于1，k也等于1，因此该第一候选词的软词频贡献值为1。

[0134] 针对相似词1，告警信息1指向的IP地址1是第一次出现，且该相似词1在告警信息1中第二次出现(本公开将相似词和对应的候选词作为同一个词，由于第一候选词已经出现过一次，因此相似词再出现时应该记为第二次出现)，那么在计算该相似词1的软词频贡献值时，相似词1和第一候选词的相似度为x，n等于1，k等于2，因此该相似词1的软词频贡献值为x/2。

[0135] 针对告警信息2中的第一候选词，告警信息2指向的IP地址1是第二次出现，且该第一候选词在告警信息2中第一次出现，那么在计算该第一候选词的软词频贡献值时，由于相似度为1，n等于2，k等于1，因此该第一候选词的软词频贡献值为1/2。

[0136] 针对相似词2，告警信息2指向的IP地址1是第二次出现，且该相似词2在告警信息2中第二次出现，那么计算该相似词2的软词频贡献值时，第一候选词和相似词2的相似y，n等于2，k等于2，因此该相似词2的软词频贡献值为y/4。

[0137] 针对告警信息3中的第一候选词，告警信息3指向的IP地址2是第一次出现，且该第一候选词在告警信息3中第一次出现，那么在计算该第一候选词的软词频贡献值时，由于相似度为1，n等于1，k等于1，因此该第一候选词的软词频贡献值为1。

[0138] 假设上述故障时间段内只有上述3条告警信息，那么该第一候选词在上述故障时间段内的软词频等于上述所有第一候选词及其相似词的软词频贡献值的总和，即1+x/2+1/2+y/4+1。

[0139] 图8示意性示出了根据本公开实施例的计算每个候选词的软逆文档频率的流程图。

[0140] 根据本公开的实施例，每个候选词的软逆文档频率可以只计算一次，计算得到的候选词的软逆文档频率可以适用于每个故障时间段，用于计算一故障时间段内的候选词的软TF-IDF值。当然，根据本公开的实施例，每个候选词的软逆文档频率也可以计算多次，例如，针对每个故障时间段都计算一次候选词的软逆文档频率，但在计算软逆文档频率时每次随机选取的非故障时间段不同。需要说明的是，对于一个故障时间段的不同候选词，选取的非故障时间段需要相同。

[0141] 根据本公开的实施例，如图8所示，计算每个候选词的软逆文档频率包括操作S801～操作S807。

[0142] 在操作S801，从多个非故障时间段选取预设数量个非故障时间段。

[0143] 根据本公开的实施例，例如，可以随机选取D个非故障时间段。

[0144] 然后，按照如下操作方式计算每个所述候选词在一个所述非故障时间段内的告警信息中的软词频。

[0145] 在操作S802，对非故障时间段内的告警信息进行排序。

[0146] 在操作S803，按照排序顺序，对于非故障时间段内的每条告警信息，采用衰减策略累加每个候选词和候选词的相似词的频次，其中，该衰减策略包括根据每个候选词和候选词的相似词在每条告警信息中出现的次序、每条告警信息指向的同一个网络地址出现的次数和不同网络地址在非故障时间段内的出现的次序，计算用于统计每个候选词和候选词的相似词在每条告警信息中的频次的贡献值。

[0147] 在操作S804，在遍历完多条非故障时间段内的告警信息后，将累加得到的候选词和候选词的相似词的频次作为候选词在非故障时间段的软词频。

[0148] 在操作S805，将候选词和候选词的相似词在每个非故障时间段内的软词频进行求和，得到软词频总和，结果记为f。

[0149] 在操作S806，计算候选词和候选词的相似词在非故障时间段内的软词频不为0的非故障时间段的个数。可以将结果记为d。

[0150] 在操作S807，根据预设数量、软词频总和、软词频不为0的非故障时间段的个数计算候选词的软逆文档频率。

[0151] 根据本公开的实施例，例如，可以按照公式log((D+f-d)/(f+1))计算该候选词的软逆文档频率。这里实际上是把软词频作为每个文档的权重，软词频为0的非故障时间段的权重为1，软词频不为0的非故障时间段的权重为软词频。

[0152] 根据本公开的实施例，计算候选词和候选词的相似词在每个非故障时间段内的软词频的计算方法与计算候选词和候选词的相似词在每个故障时间段内的软词频的方式相似，但是多了一层时间段层，即考虑了不同网络地址在非故障时间段内的出现的次序。这是为了区别例如在10个时间段中各出现1次候选词和相似词与在一个时间段中出现10次候选词和相似词，在10个时间段中各出现1次候选词和相似词的分布更广泛更能说明该词的普遍存在。

[0153] 在同一个非故障时间段内，该候选词及其相似词在第n1个IP的第n2条告警信息中第n3次出现时，软词频的贡献值为相似度/(n1*n2*n3)。这里每层使用的衰减函数都是1/n，当然，也可以选择不同的衰减函数。

[0154] 具体地，计算候选词和候选词的相似词在每个非故障时间段内的软词频的计算方法可以包括如下具体示例。

[0155] 首先，先计算一个非故障时间段内的每条告警信息中出现该候选词及其相似词的总次数。

[0156] 然后，再选取出现过该候选词或其相似词的告警信息并按该出现总次数进行排序。

[0157] 当然，也可以分别计算该候选词或其相似词在告警信息中出现的次数，然后将多条告警信息以候选词出现次数优先排序，候选词次数相同时按相似词出现次数排序。

[0158] 再然后，按次序遍历每条告警信息，采用衰减策略累加每个候选词和候选词的相似词的频次。

[0159] 最后，在遍历完多条告警信息后，将累加得到的候选词和候选词的相似词的频次作为候选词在非故障时间段的软词频。

[0160] 具体地，例如，在一个非故障时间段内，将多条告警信息进行排序后存在如下顺序的告警信息：告警信息1，告警信息2，告警信息3。其中，告警信息1和告警信息2都是指向IP地址1的告警，告警信息3是指向IP地址2的告警。告警信息1包括第一候选词及其相似词1，告警信息2包括第一候选词及其相似词2，告警信息3包括第一候选词。

[0161] 针对告警信息1中的第一候选词，告警信息1指向的IP地址1是上述非故障时间段内的第1个IP地址(即IP地址1在非故障时间段内的出现的次序为第1个，n1等于1)，且告警信息1是IP地址1的第1条告警信息(即告警信息1指向的IP地址1是第一次出现，且告警信息

1是IP地址1的第一条告警信息,n2等于1),且该第一候选词在告警信息1中第一次出现(即n3等于1),那么在计算该第一候选词的软词频贡献值时,由于相似度为1,n1等于1,n2也等于1,n3也等于1,因此该第一候选词的软词频贡献值为1。

[0162] 针对相似词1,告警信息1指向的IP地址1是上述非故障时间段内的第1个IP地址(即n1等于1),且告警信息1是IP地址1的第1条告警信息(即n2等于1),且该相似词1在告警信息1中第二次出现(本公开将相似词和对应的候选词作为同一个词,由于第一候选词已经出现过一次,因此相似词再出现时应该记为第二次出现,即n3等于2),那么在计算该相似词1的软词频贡献值时,第一候选词和相似词1的相似度为x,n1等于1,n2等于1,n3等于2,因此该相似词1的软词频贡献值为x/2。

[0163] 针对告警信息2中的第一候选词,告警信息2指向的IP地址1是上述非故障时间段内的第1个IP地址(即n1等于1),且告警信息2指向的IP地址1是第二次出现(即n2等于2),且该第一候选词在告警信息2中第一次出现(即n3等于1),那么在计算该第一候选词的软词频贡献值时,由于相似度为1,n1等于1,n2等于2,n3等于1,因此该第一候选词的软词频贡献值为1/2。

[0164] 针对相似词2,告警信息2指向的IP地址1是上述非故障时间段内的第1个IP地址(即n1等于1),且告警信息2指向的IP地址1是第二次出现(即n2等于2),且该相似词2在告警信息2中第二次出现(即n3等于2),那么计算该相似词2的软词频贡献值时,第一候选词和相似词2的相似y,n1等于1,n2等于2,n3等于2,因此该相似词2的软词频贡献值为y/4。

[0165] 针对告警信息3中的第一候选词,告警信息3指向的IP地址2是上述非故障时间段内的第2个IP地址(即n1等于2),且告警信息3指向的IP地址2是第一次出现(即n2等于1),且该第一候选词在告警信息3中第一次出现(即n3等于1),那么在计算该第一候选词的软词频贡献值时,由于相似度为1,n1等于2,n2等于1,n3等于1,因此该第一候选词的软词频贡献值为1/2。

[0166] 假设上述非故障时间段内只有上述3条告警信息,那么该第一候选词在上述故障时间段内的软词频等于上述所有第一候选词及其相似词的软词频贡献值的总和,即1+x/2+1/2+y/4+1/2。

[0167] 图9示意性示出了根据本公开实施例的根据软TF-IDF值确定目标故障的特征词的流程图。

[0168] 根据本公开的实施例,如图9所示,该方法包括操作S901～操作S903。

[0169] 在操作S901,对每个故障时间段,分别选出软TF-IDF值属于异常值(例如软TF-IDF值异常大)的候选词作为提名候选词。可以先用孤立森林选出异常值,再判断是否大于平均值来选出异常大的值。

[0170] 在操作S902,给每个提名候选词计算投票值。如果被一个故障时间段提名,那么就加一票,如果若干个故障时间段属于同一次故障,那么投票值可以除以同一次故障的故障时间段数。

[0171] 例如,存在3个故障时段:故障时间段1,故障时间段2,故障时间段3,其中故障时间段2和故障时间段3属于同一次故障。故障时间段1的提名候选词为提名候选词1,提名候选词2,提名候选词3;故障时间段2的提名候选词为提名候选词1,提名候选词2,提名候选词4;故障时间段3的提名候选词为提名候选词1,提名候选词2,提名候选词5。

[0172] 由于故障时间段2和故障时间段3属于同一次故障，所以它们的投票值要除以2。因此，提名候选词1的投票值为1+1/2+1/2＝2，提名候选词2的投票值为1+1/2+1/2＝2，提名候选词3的投票值为1+0+0＝1，提名候选词4的投票值为0+1/2+0＝1/2，提名候选词5的投票值为0+0+1/2＝1/2。

[0173] 在操作S903，选取投票值最高的t个提名候选词作为目标故障的特征词。t不宜过大，例如可以选取2到8之间。

[0174] 图10示意性示出了根据本公开实施例的故障识别装置的框图。

[0175] 如图10所示，故障识别装置包括：第一获取模块1010、第二获取模块1020、训练模块1030和处理模块1040。

[0176] 第一获取模块1010用于获取历史上发生目标故障的历史时间段。

[0177] 第二获取模块1020用于根据历史时间段获取与目标故障相关联的历史告警信息。

[0178] 训练模块1030用于根据历史告警信息训练得到目标故障的故障识别模型。

[0179] 处理模块1040用于通过故障识别模型对当前告警信息进行处理，以识别是否发生了目标故障。

[0180] 根据本公开的实施例，采用了根据历史上发生目标故障的历史时间段获取与目标故障相关联的历史告警信息；根据历史告警信息训练得到目标故障的故障识别模型；通过故障识别模型对当前告警信息进行处理，以识别是否发生了目标故障的技术手段，不需要大量人力的介入，自动从历史数据中学习故障识别模型，所以至少部分地克服了突发较多地或者大量告警时的故障判断的技术问题，进而达到了当目标故障发生时，可以及时地通知监控人员，进而缩短故障恢复时间，保障业务平稳运行的技术效果。进一步的，可以减轻运维人员工作量、提高监控及时性和有效性、减少故障恢复时间，进而提升安全生产水平，在监控告警行业具有重要的应用价值。

[0181] 根据本公开的实施例，根据历史时间段获取与目标故障相关联的历史告警信息包括：确定包含历史时间段的告警时间段；以及获取告警时间段内发生的历史告警信息。

[0182] 根据历史告警信息训练得到目标故障的故障识别模型包括：按照预设分割时长将告警时间段分割为多个子时间段；将与历史时间段存在交叉，且在一个子时间段内的告警条数大于预设条数的子时间段确定为故障时间段；将与历史时间段不存在交叉，或者在一个子时间段内的告警条数小于或等于预设条数的子时间段确定为非故障时间段；以及根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型。

[0183] 根据本公开的实施例，根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型包括：从故障时间段内的告警信息中选取特征词的多个候选词；确定每个候选词的相似词；计算每个候选词的软词频和软逆文档频率；根据每个候选词的软词频和软逆文档频率确定目标故障的特征词；以及根据故障时间段内关于目标故障的特征词和非故障时间段内关于目标故障的特征词，训练得到目标故障的故障识别模型。

[0184] 根据本公开的实施例，故障时间段可以包括多个，从故障时间段内的告警信息中选取特征词的候选词包括：获取多个故障时间段中每个故障时间段内的告警信息的文本；对每个故障时间段内的告警信息的文本进行分词；以及根据多个故障时间段内的每个词汇

的总出现次数选取候选词。

[0185] 根据本公开的实施例，根据每个候选词的软词频和软逆文档频率确定目标故障的特征词包括：将候选词在故障时间段的软词频与候选词在非故障时间段的软逆文档频率的乘积作为候选词在故障时间段内的软TF-IDF值；以及根据每个候选词在不同故障时间段内的软TF-IDF值确定目标故障的特征词。

[0186] 根据本公开的实施例，计算每个候选词的软词频包括：将故障时间段内的多条告警信息进行排序；按照排序顺序遍历每条告警信息，采用衰减策略累加每个候选词和候选词的相似词的频次，其中，衰减策略包括根据每个候选词和候选词的相似词在每条告警信息中出现的次序和每条告警信息指向的同一个网络地址出现的次数，计算用于统计每个候选词和候选词的相似词在每条告警信息中的频次的贡献值；以及在遍历完多条告警信息后，将累加得到的候选词和候选词的相似词的频次作为候选词在故障时间段的软词频。

[0187] 根据本公开的实施例，非故障时间段包括多个，计算每个候选词的软逆文档频率包括：从多个非故障时间段选取预设数量个非故障时间段；按照如下方式计算每个候选词在一个非故障时间段内的告警信息中的软词频：对非故障时间段内的告警信息进行排序；按照排序顺序，对于非故障时间段内的每条告警信息，采用衰减策略累加每个候选词和候选词的相似词的频次，其中，衰减策略包括根据每个候选词和候选词的相似词在每条告警信息中出现的次序、每条告警信息指向的同一个网络地址出现的次数和不同网络地址在非故障时间段内的出现的次序，计算用于统计每个候选词和候选词的相似词在每条告警信息中的频次的贡献值；在遍历完多条告警信息后，将累加得到的候选词和候选词的相似词的频次作为候选词在非故障时间段的软词频；将候选词和候选词的相似词在每个非故障时间段内的软词频进行求和，得到软词频总和；计算候选词和候选词的相似词在非故障时间段内的软词频不为0的非故障时间段的个数；以及根据预设数量、软词频总和、软词频不为0的非故障时间段的个数计算候选词的软逆文档频率。

[0188] 根据本公开的实施例，根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型包括：获取故障时间段内的告警信息中的特征词对应的第一特征向量，其中，第一特征向量的标签为故障；获取非故障时间段内的告警信息中的特征词对应的第二特征向量，其中，第二特征向量的标签为非故障；以及将第一特征向量和第一特征向量对应的标签，第二特征向量和第二特征向量对应的标签输入到支持向量机中，以训练得到目标故障的故障识别模型。

[0189] 根据本公开的实施例，获取非故障时间段内的告警信息中的特征词对应的第二特征向量包括：计算非故障时间段内的告警信息中的每个特征词对应的第二特征向量的模；以及根据每个特征词对应的第二特征向量的模的大小选择指定数量的第二特征向量。

[0190] 根据本公开的实施例，还提供了一种基于文本挖掘的故障识别系统。该系统可以包括：历史告警装置、故障特征词及特征向量确定装置、故障识别模型训练装置、故障识别装置和故障结果通知装置。

[0191] 其中，历史告警装置负责获取和存储历史告警。

[0192] 故障特征词及特征向量确定装置负责确定目标故障的特征词及特征向量。

[0193] 故障识别模型训练装置负责训练故障识别模型。

[0194] 故障识别装置负责根据当前告警信息识别是否发生了目标故障。

[0195] 故障结果通知装置负责当识别结果为故障时将结果通知给监控人员。

[0196] 根据本公开的实施例的模块、子模块、单元、子单元中的任意多个、或其中任意多个的至少部分功能可以在一个模块中实现。根据本公开实施例的模块、子模块、单元、子单元中的任意一个或多个可以被拆分成多个模块来实现。根据本公开实施例的模块、子模块、单元、子单元中的任意一个或多个可以至少被部分地实现为硬件电路，例如现场可编程门阵列(FPGA)、可编程逻辑阵列(PLA)、片上系统、基板上的系统、封装上的系统、专用集成电路(ASIC)，或可以通过对电路进行集成或封装的任何其他的合理方式的硬件或固件来实现，或以软件、硬件以及固件三种实现方式中任意一种或以其中任意几种的适当组合来实现。或者，根据本公开实施例的模块、子模块、单元、子单元中的一个或多个可以至少被部分地实现为计算机程序模块，当该计算机程序模块被运行时，可以执行相应的功能。

[0197] 例如，第一获取模块1010、第二获取模块1020、训练模块1030和处理模块1040中的任意多个可以合并在一个模块/单元/子单元中实现，或者其中的任意一个模块/单元/子单元可以被拆分成多个模块/单元/子单元。或者，这些模块/单元/子单元中的一个或多个模块/单元/子单元的至少部分功能可以与其他模块/单元/子单元的至少部分功能相结合，并在一个模块/单元/子单元中实现。根据本公开的实施例，第一获取模块1010、第二获取模块1020、训练模块1030和处理模块1040中的至少一个可以至少被部分地实现为硬件电路，例如现场可编程门阵列(FPGA)、可编程逻辑阵列(PLA)、片上系统、基板上的系统、封装上的系统、专用集成电路(ASIC)，或可以通过对电路进行集成或封装的任何其他的合理方式等硬件或固件来实现，或以软件、硬件以及固件三种实现方式中任意一种或以其中任意几种的适当组合来实现。或者，第一获取模块1010、第二获取模块1020、训练模块1030和处理模块1040中的至少一个可以至少被部分地实现为计算机程序模块，当该计算机程序模块被运行时，可以执行相应的功能。

[0198] 需要说明的是，本公开的实施例中故障识别装置部分与本公开的实施例中故障识别方法部分是相对应的，故障识别装置部分的描述具体参考故障识别方法部分，在此不再赘述。

[0199] 图11示意性示出了根据本公开实施例的适于实现上文描述的方法的计算机系统的框图。图11示出的计算机系统仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

[0200] 如图11所示，根据本公开实施例的计算机系统1100包括处理器1101，其可以根据存储在只读存储器(ROM)1102中的程序或者从存储部分1108加载到随机访问存储器(RAM)1103中的程序而执行各种适当的动作和处理。处理器1101例如可以包括通用微处理器(例如CPU)、指令集处理器和/或相关芯片组和/或专用微处理器(例如，专用集成电路(ASIC))，等等。处理器1101还可以包括用于缓存用途的板载存储器。处理器1101可以包括用于执行根据本公开实施例的方法流程的不同动作的单一处理单元或者是多个处理单元。

[0201] 在RAM 1103中，存储有系统1100操作所需的各种程序和数据。处理器1101、ROM 1102以及RAM 1103通过总线1104彼此相连。处理器1101通过执行ROM 1102和/或RAM 1103中的程序来执行根据本公开实施例的方法流程的各种操作。需要注意，所述程序也可以存储在除ROM 1102和RAM 1103以外的一个或多个存储器中。处理器1101也可以通过执行存储在所述一个或多个存储器中的程序来执行根据本公开实施例的方法流程的各种操作。

[0202] 根据本公开的实施例，系统1100还可以包括输入/输出(I/O)接口1105，输入/输出(I/O)接口1105也连接至总线1104。系统1100还可以包括连接至I/O接口1105的以下部件中的一项或多项：包括键盘、鼠标等的输入部分1106；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1107；包括硬盘等的存储部分1108；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1109。通信部分1109经由诸如因特网的网络执行通信处理。驱动器1110也根据需要连接至I/O接口1105。可拆卸介质1111，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1110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1108。

[0203] 根据本公开的实施例，根据本公开实施例的方法流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，其包括承载在计算机可读存储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1109从网络上被下载和安装，和/或从可拆卸介质1111被安装。在该计算机程序被处理器1101执行时，执行本公开实施例的系统中限定的上述功能。根据本公开的实施例，上文描述的系统、设备、装置、模块、单元等可以通过计算机程序模块来实现。

[0204] 本公开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设备/装置/系统中所包含的；也可以是单独存在，而未装配入该设备/装置/系统中。上述计算机可读存储介质承载有一个或者多个程序，当上述一个或者多个程序被执行时，实现根据本公开实施例的方法。

[0205] 根据本公开的实施例，计算机可读存储介质可以是非易失性的计算机可读存储介质。例如可以包括但不限于：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)、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。

[0206] 例如，根据本公开的实施例，计算机可读存储介质可以包括上文描述的ROM 1102和/或RAM 1103和/或ROM 1102和RAM 1103以外的一个或多个存储器。

[0207] 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，上述模块、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，框图或流程图中的每个方框、以及框图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。本领域技术人员可以理解，本公开的各个实施例和/或权利要求中记载的特征可以进行多种组合和/或结合，即使这样的组合或结合没有明确记载于本公开中。特别地，在不脱离本公开精神和教导的情况下，本公开的各个实施例和/或权利要求中记载的特征可以进行多种组合和/或结合。所有这些组合和/或结合均落入本公开的范围。

[0208] 以上对本公开的实施例进行了描述。但是，这些实施例仅仅是为了说明的目的，而并非为了限制本公开的范围。尽管在以上分别描述了各实施例，但是这并不意味着各个实施例中的措施不能有利地结合使用。本公开的范围由所附权利要求及其等同物限定。不脱离本公开的范围，本领域技术人员可以做出多种替代和修改，这些替代和修改都应落在本公开的范围之内。

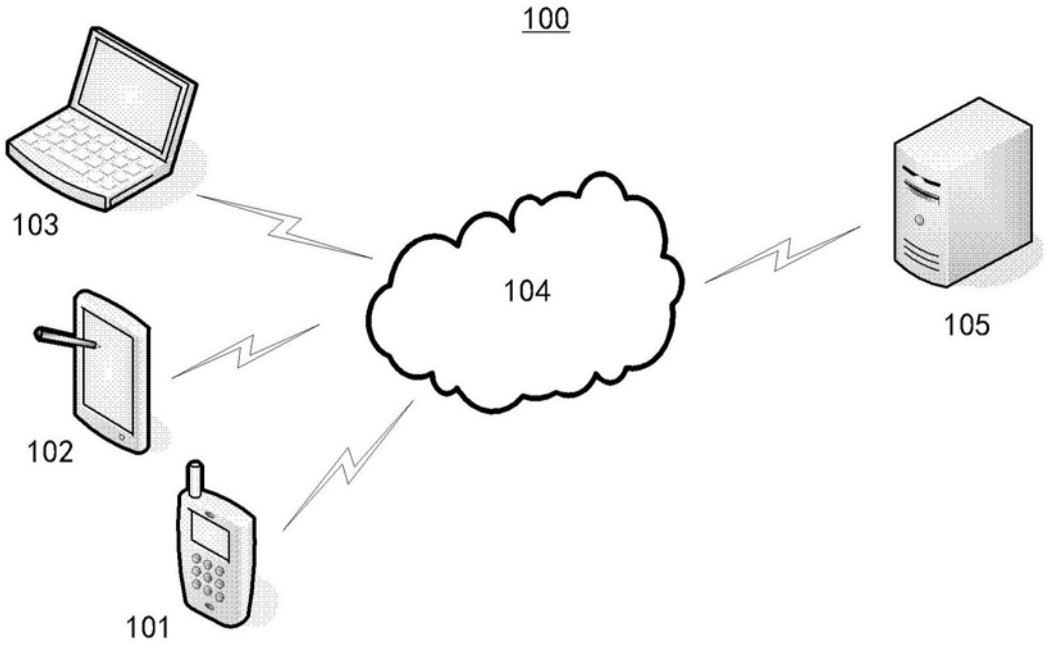

100
103
104
105
102
101

图1

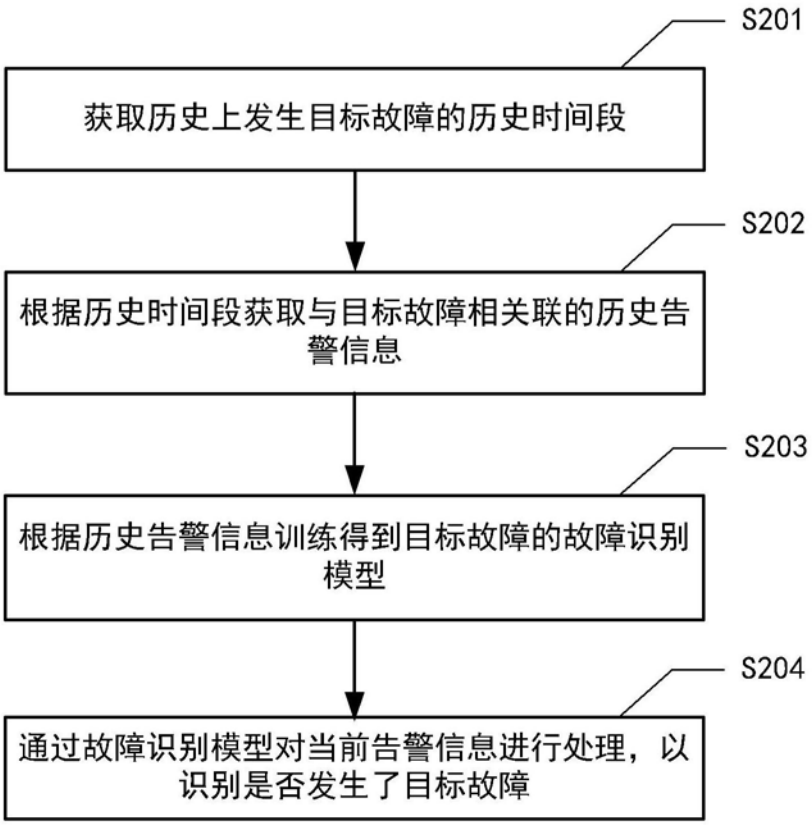

S201
获取历史上发生目标故障的历史时间段
S202
根据历史时间段获取与目标故障相关联的历史告警信息
S203
根据历史告警信息训练得到目标故障的故障识别模型
S204
通过故障识别模型对当前告警信息进行处理，以识别是否发生了目标故障

图2

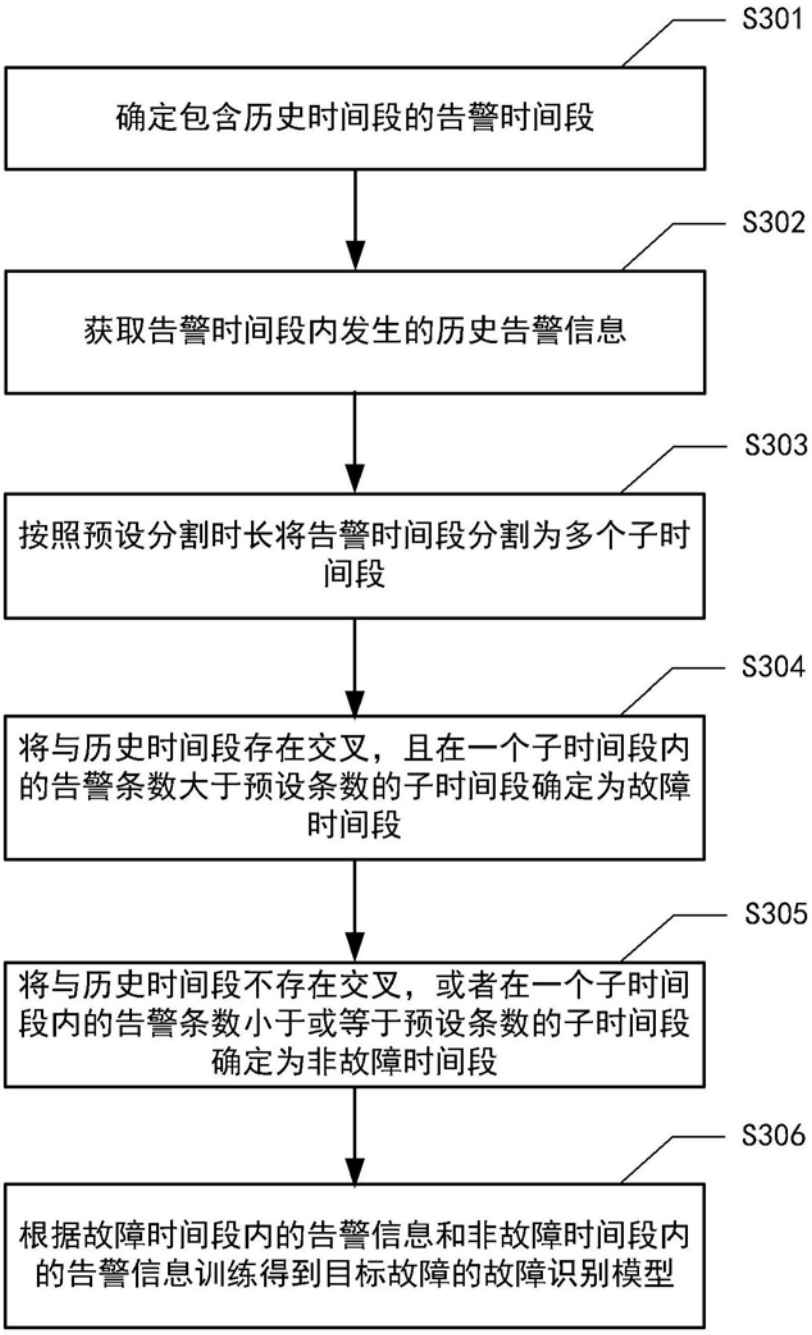

S301
确定包含历史时间段的告警时间段
S302
获取告警时间段内发生的历史告警信息
S303
按照预设分割时长将告警时间段分割为多个子时间段
S304
将与历史时间段存在交叉，且在一个子时间段内的告警条数大于预设条数的子时间段确定为故障时间段
S305
将与历史时间段不存在交叉，或者在一个子时间段内的告警条数小于或等于预设条数的子时间段确定为非故障时间段
S306
根据故障时间段内的告警信息和非故障时间段内的告警信息训练得到目标故障的故障识别模型

图3

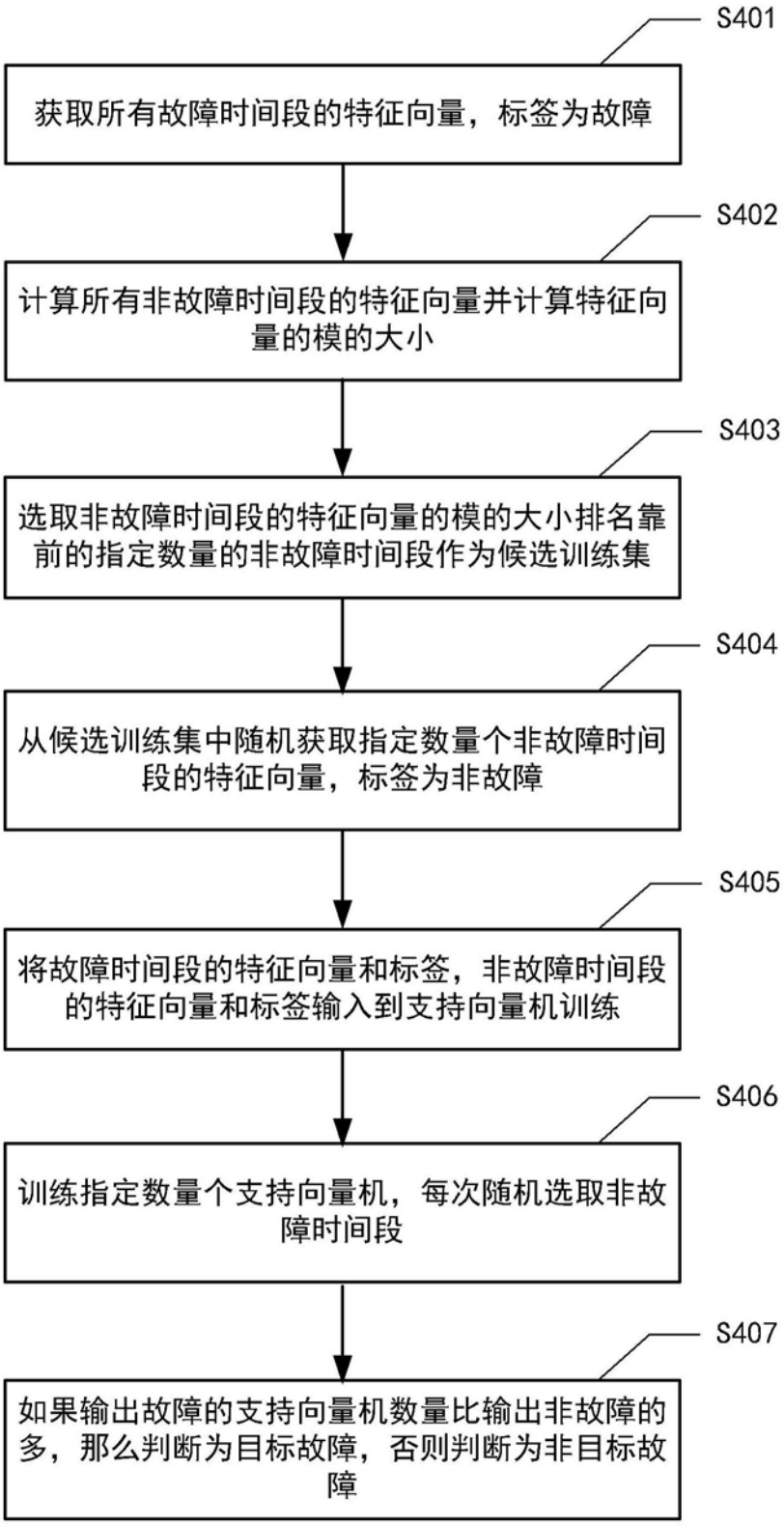

S401
获取所有故障时间段的特征向量，标签为故障
S402
计算所有非故障时间段的特征向量并计算特征向量的模的大小
S403
选取非故障时间段的特征向量的模的大小排名靠前的指定数量的非故障时间段作为候选训练集
S404
从候选训练集中随机获取指定数量个非故障时间段的特征向量，标签为非故障
S405
将故障时间段的特征向量和标签，非故障时间段的特征向量和标签输入到支持向量机训练
S406
训练指定数量个支持向量机，每次随机选取非故障时间段
S407
如果输出故障的支持向量机数量比输出非故障的多，那么判断为目标故障，否则判断为非目标故障

图4

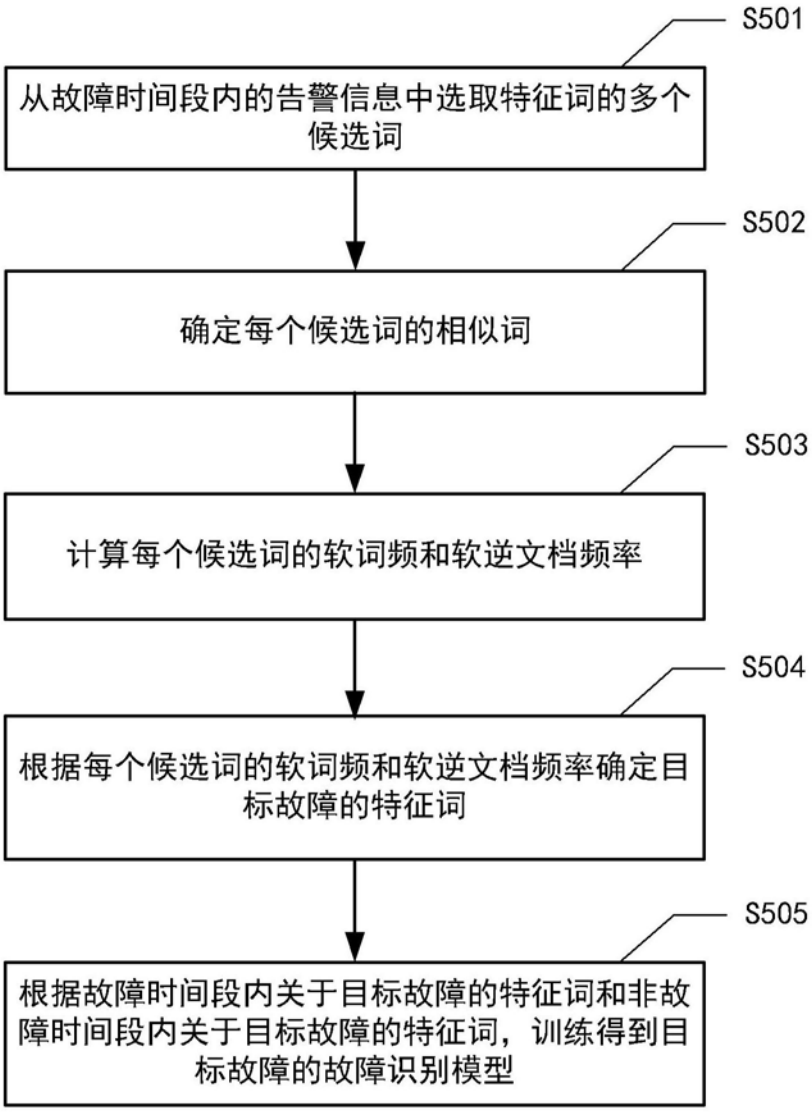

S501
从故障时间段内的告警信息中选取特征词的多个候选词
S502
确定每个候选词的相似词
S503
计算每个候选词的软词频和软逆文档频率
S504
根据每个候选词的软词频和软逆文档频率确定目标故障的特征词
S505
根据故障时间段内关于目标故障的特征词和非故障时间段内关于目标故障的特征词，训练得到目标故障的故障识别模型

图5

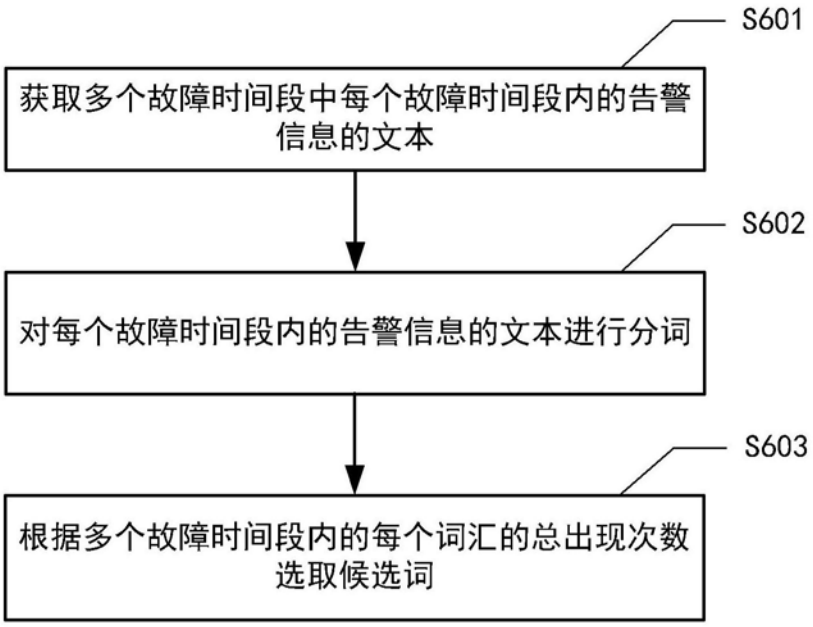

S601
获取多个故障时间段中每个故障时间段内的告警信息的文本
S602
对每个故障时间段内的告警信息的文本进行分词
S603
根据多个故障时间段内的每个词汇的总出现次数选取候选词

图6

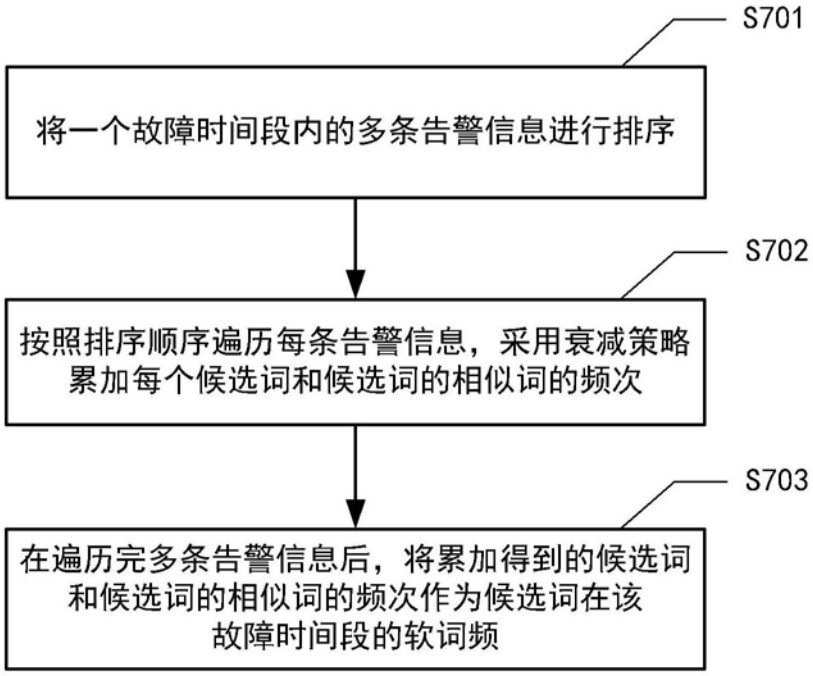

S701
将一个故障时间段内的多条告警信息进行排序
S702
按照排序顺序遍历每条告警信息，采用衰减策略累加每个候选词和候选词的相似词的频次
S703
在遍历完多条告警信息后，将累加得到的候选词和候选词的相似词的频次作为候选词在该故障时间段的软词频

图7

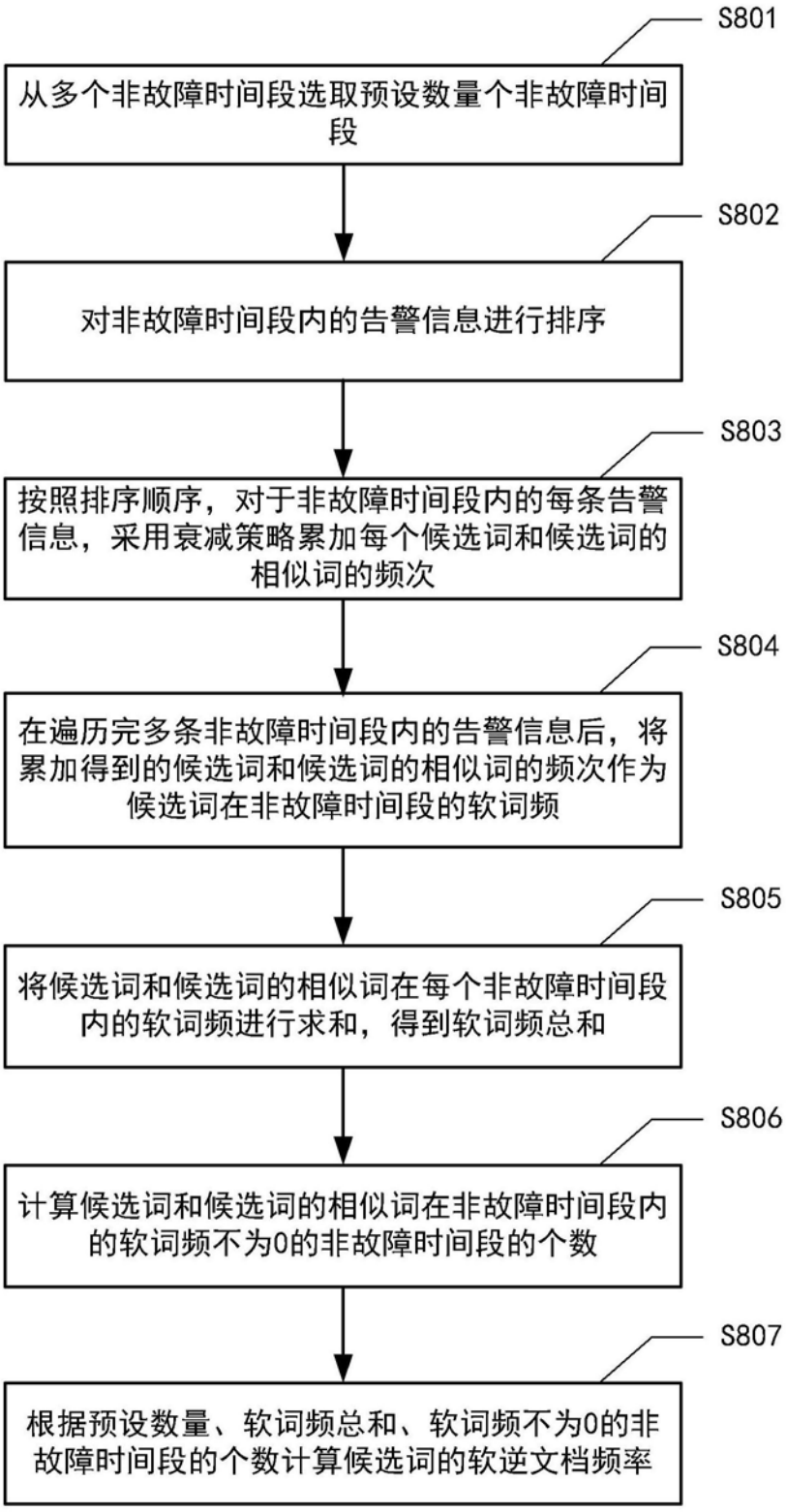

S801
从多个非故障时间段选取预设数量个非故障时间段
S802
对非故障时间段内的告警信息进行排序
S803
按照排序顺序，对于非故障时间段内的每条告警信息，采用衰减策略累加每个候选词和候选词的相似词的频次
S804
在遍历完多条非故障时间段内的告警信息后，将累加得到的候选词和候选词的相似词的频次作为候选词在非故障时间段的软词频
S805
将候选词和候选词的相似词在每个非故障时间段内的软词频进行求和，得到软词频总和
S806
计算候选词和候选词的相似词在非故障时间段内的软词频不为0的非故障时间段的个数
S807
根据预设数量、软词频总和、软词频不为0的非故障时间段的个数计算候选词的软逆文档频率

图8

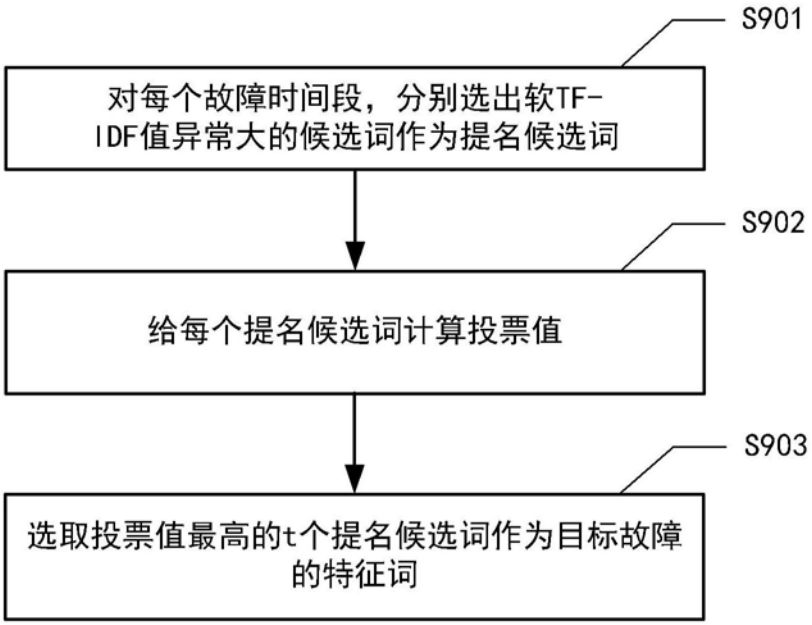

S901
对每个故障时间段，分别选出软TF-IDF值异常大的候选词作为提名候选词
S902
给每个提名候选词计算投票值
S903
选取投票值最高的t个提名候选词作为目标故障的特征词

图9

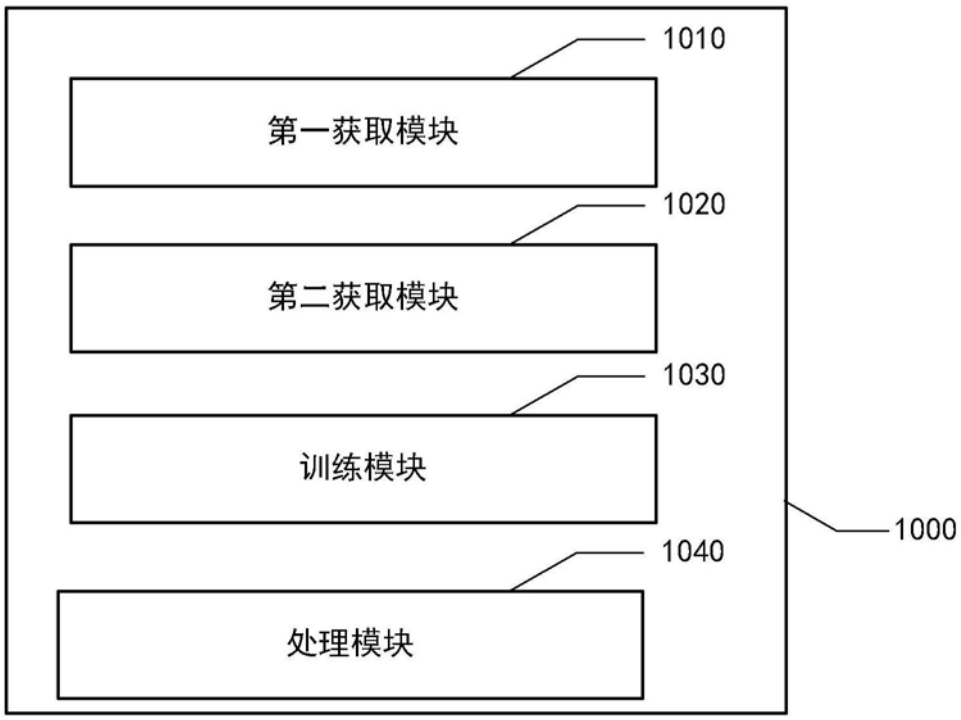

1010
第一获取模块
1020
第二获取模块
1030
训练模块
1040
处理模块
1000

图10

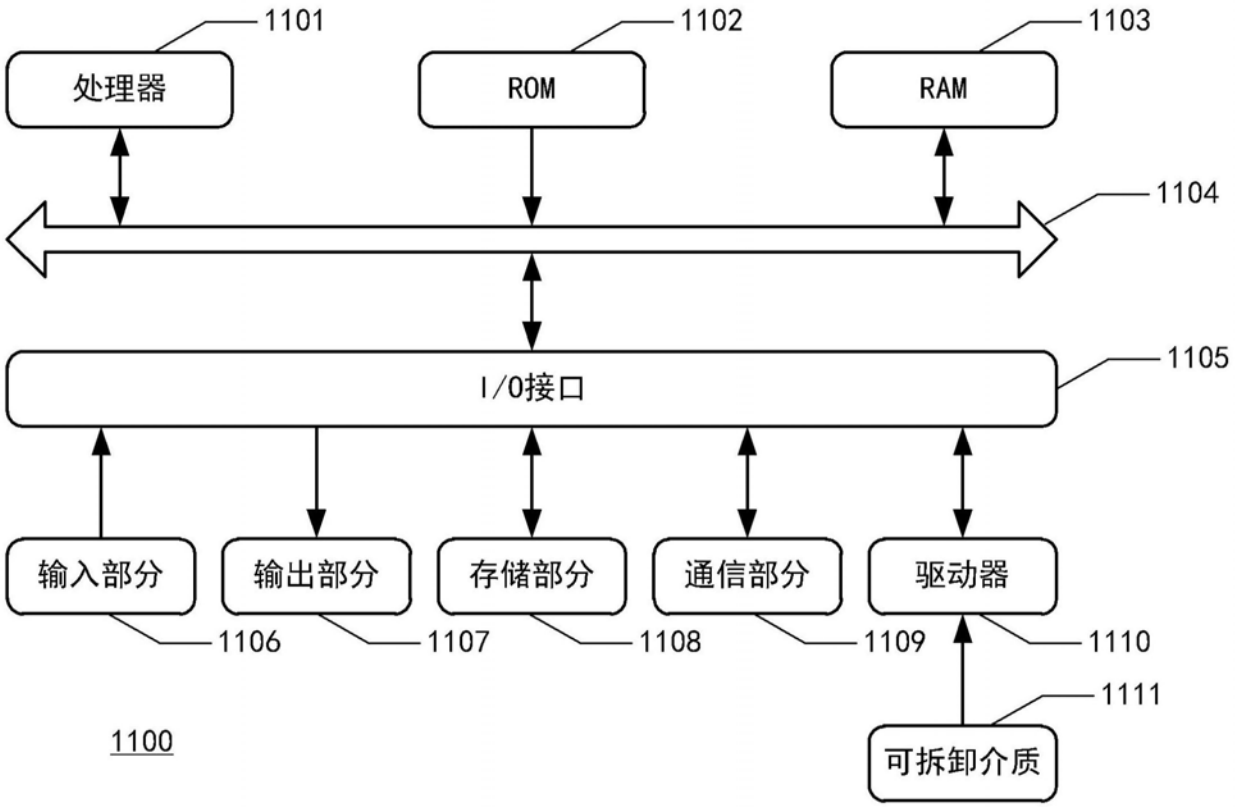

1101
1102
1103
处理器
ROM
RAM
1104
1105
I/O接口
输入部分
输出部分
存储部分
通信部分
驱动器
1106
1107
1108
1109
1110
1111
1100
可拆卸介质

图11