

URL 异常定位方法、装置、服务器及存储介质

相关申请的交叉引用

[01]本专利申请要求于 2018 年 03 月 06 日提交的、申请号为 201810182571.X、发明名称为“URL 异常定位方法、装置、服务器及存储介质”的中国专利申请的优先权，该申请

5 请的全文以引用的方式并入本文中。

技术领域

[02]本说明书实施例涉及互联网技术领域，尤其涉及一种 URL 异常定位方法、装置、服务器及存储介质。

背景技术

10 [03]在互联网的应用场景中，每天会有大量的对于网址 URL (Uniform Resource Locator, 统一资源定位符) 的访问；与此同时，不乏不法分子试图通过不合法的 URL 访问进行攻击。

发明内容

[04]本说明书实施例提供及一种 URL 异常定位方法、装置、服务器及存储介质。

15 [05]第一方面，本说明书实施例提供一种 URL 异常定位方法，包括：对 URL 进行字段切分，得到由对应各个字段的各个示例组成的多示例包；将所述多示例包输入基于多示例学习的 URL 异常定位模型进行异常示例预测；根据异常示例定位出对应的异常字段。

[06]第二方面，本说明书实施例提供一种 URL 异常定位训练方法，包括：收集由多个 URL 样本组成的 URL 样本集；对 URL 样本集中各个 URL 样本进行字段切分，针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包；集合各个 URL 样本的多

20 示例包得到多示例包集；基于多示例学习算法，对多示例包集进行异常示例及非异常示例分类训练；基于所述分类训练，得到所述 URL 异常定位模型。

[07]第三方面，本说明书实施例提供一种 URL 异常定位装置，包括：切分单元，用于对 URL 进行字段切分，得到由对应各个字段的各个示例组成的多示例包；预测单元，用于

25 将所述多示例包输入基于多示例学习的 URL 异常定位模型进行异常示例预测；定位单

元，用于根据异常示例定位出对应的异常字段。

[08]第四方面，本说明书实施例提供一种 URL 异常定位训练装置，包括：样本获取单元，用于收集由多个 URL 样本组成的 URL 样本集；样本切分单元，用于对 URL 样本集中各个 URL 样本进行字段切分，针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包；示例包集合单元，用于集合各个 URL 样本的多示例包得到多示例包集；训练单元，用于基于多示例学习算法，对多示例包集进行异常示例及非异常示例分类训练，得到所述 URL 异常定位模型。

[09]第五方面，本说明书实施例提供一种服务器，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任一项所述方法的步骤。

[10]第六方面，本说明书实施例提供一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现上述任一项所述方法的步骤。

[11]本说明书实施例有益效果如下：

[12]在本说明书实施例中，通过对 URL 表示为对应多个字段的多个示例组成的包，利用 URL 异常定位模型对异常示例进行预测，从而定位出 URL 中的异常字段。基于多示例学习的 URL 异常定位，能够较好的预测出每天的访问数据中未发现的潜在威胁，由于可以针对异常 URL 确定异常所在，从而可为发现潜在威胁、建立新的安全规则、构建安全系统提供巨大的支持。

附图说明

[13]图 1 为本说明书实施例的 URL 异常定位场景示意图；

[14]图 2 为本说明书实施例第一方面的 URL 异常定位方法流程图；

[15]图 3 为本说明书实施例第二方面的 URL 异常定位训练方法流程图；

[16]图 4 为本说明书实施例第三方面的 URL 异常定位装置结构示意图；

[17]图 5 为本说明书实施例第四方面的 URL 异常定位训练装置结构示意图；

[18]图 6 为本说明书实施例第五方面提供的 URL 异常定位服务器结构示意图。

具体实施方式

[19]为了更好的理解上述技术方案，下面通过附图以及具体实施例对本说明书实施例的技术方案做详细的说明，应当理解本说明书实施例以及实施例中的具体特征是对本说明书实施例技术方案的详细的说明，而不是对本说明书技术方案的限定，在不冲突的情况下，

5 下，本说明书实施例以及实施例中的技术特征可以相互组合。

[20]本说明书实施例 URL 异常定位场景示意图请参见图 1。客户端（用户侧）向服务端（网络侧）提出 URL 访问请求，服务端解析 URL 访问请求得到 URL，并对 URL 进行基于多示例学习的异常定位。具体的，服务端预先基于多示例学习算法训练出 URL 异常定位模型，通过 URL 异常定位模型预测出异常示例，再由异常示例确定出对应的 URL
10 所在字段，达到定位 URL 异常的目的。本发明实施例中所谓“URL 异常定位”，区别于仅仅给出整个 URL 是否异常的方式，而是确定出 URL 中具体的异常字段位置，从而可便于对异常进行更为准确的分析和预防。

[21]第一方面，本说明书实施例提供一种 URL 异常定位方法。

[22]请参考图 2，本说明书实施例提供的 URL 异常定位方法包括如下步骤 S201-S203。

15 [23]S201：对 URL 进行字段切分，得到由对应各个字段的各个示例组成的多示例包。

[24]基于多示例学习（Multiple Instance Learning，MIL），对 URL 进行切分得到多示例构成的包。区别于传统的监督学习，在多示例学习中，数据是以包（bag）的形式给出，一个包（bag）中通常具有多个示例（instance）。可以理解，本说明书实施例中每一个 URL 对应一个包（bag）、而包中又包括多个示例，因此以“多示例包”表示一个 URL。

20 [25]基本 URL 包含模式（或称协议）、服务器名称（或 IP 地址）、路径和文件名，如“协议://授权/路径?查询”。完整的、带有授权部分的普通统一资源标志符语法可如下：协议://用户名:密码@子域名.域名.顶级域名:端口号/目录/文件名.文件后缀?参数=值#标志。本说明书实施例中可以对 URL 整个构成进行字段切分，也可以仅对高风险字段进行切分。

25 [26]例如，仅对服务器名称字段进行进一步切分为多个示例；或者仅对#（井号）后面的部分进行字段切分得到多个示例。

[27]其中，示例可由对应字段的特征向量表示。例如，将某字段的 pattern、字符数、字母数等表示为特征向量，得到该字段对应的示例。

[28]S202：将多示例包输入预设的 URL 异常定位模型进行异常示例预测。

[29]本说明书实施例中,首先,可预先根据多示例学习算法基于多个 URL 样本进行训练,得到 URL 异常定位模型(训练过程请参考图 3 及相关描述);然后,将待预测 URL 对应的多示例包输入 URL 异常定位模型,根据 URL 异常定位模型,预测多示例包中的各个示例的异常标记的值,从而预测多示例包中的各个示例是否为异常示例。

- 5 [30]在传统多示例学习算法中,仅在包的层面给出标记(label)。例如,对于标准的多示例学习而言,以二分类为例,对于一个包,只要其中有一个示例是正示例,该包就是正包;而负包中的所有示例都是负示例。要说明的是,包的标记是已知的,样本的标记是未知的。因此,相比传统监督学习,多示例学习的监督信息更少,难度更大。

- 10 [31]本说明书实施例中,通过给出示例的标记的方式,区别出异常示例和非异常示例(而不仅限于包层面的标记)。

- [32]对于一个正包(positive bag),其中至少有一个正示例(positive instance),而负包(negative bag)中的所有示例都是负示例(negative instance)。需要说明的是,本说明书实施例中,所谓正示例是指异常示例(例如异常标记的值为 1 或者具有正标记),正包也即异常包;所谓负示例是指非异常示例(例如异常标记的值为 0 或者具有负标记),
15 负包也即非异常包。

[33]S203: 根据异常示例定位出对应的异常字段。

[34]由于每一个示例都是确定对应某个字段的,因此在预测出异常示例之后,即可确定出异常示例对应的异常字段,也即确定出 URL 中异常位置。

- 20 [35]本发明实施例中: 对于一个包,只要其中有一个示例是正的,该包就是正包;而负包中的所有示例都是负的。因此,如果某个 URL 对应的多示例包中,只要包括一个异常示例,则确定该 URL 为异常 URL;如果该 URL 对应的多示例包中不包括任何异常示例,则确定该 URL 为非异常 URL。因此,如果确定出 URL 对应异常示例,则可确定 URL 为异常 URL。

[36]参考图 3,为本说明书实施例第二方面的 URL 异常定位训练方法流程图,包括:

- 25 [37]S301: 收集由多个 URL 样本组成的 URL 样本集。

[38]S302: 对 URL 样本集中各个 URL 样本进行字段切分,针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包。

[39]对于一条 URL 样本原始数据,将其切分,对于每一个切分好的字段,可以提取不同

的特征来对字段以示例进行表示，最终集合一个 URL 的不同示例作为一个包（bag）。

[40]S303: 集合各个 URL 样本的多示例包得到多示例包集。

[41]S304: 基于多示例学习算法，对多示例包集进行异常示例及非异常示例分类训练。

5 [42]对示例分类训练的过程可以理解为，是从每个包里面试图挖掘其示例的标记，从而来训练对于示例进行分类的分类器。其中：异常示例和非异常示例通过示例的异常标记的值进行区分。

[43]在一种可选方式中，对多示例包集进行异常示例及非异常示例分类训练的具体过程包括：对多示例包集中每个示例的异常标记的值进行初始化，并对异常标记的值进行迭代学习，更新调整出每个示例最终的异常标记的值。

10 [44]由于示例的标记不可见，因此先初始化一个可能的标记，进而在训练的过程中迭代的更新示例标记，以达到优化结果的目的。随着训练的进行，示例标记会越来越趋于接近事实，从而使模型能够渐渐发现可能的异常示例（有害示例），使得最终的模型能对未来的 URL 预测其可能的异常字段。

15 [45]具体而言，如前所述的，对于一个正包（positive bag），其中至少有一个正示例（positive instance），而负包（negative bag）中的所有示例都是负示例（negative instance）。需要说明的是，本说明书实施例中，所谓正示例是指异常示例（例如异常标记的值为 1 或者具有正标记），正包也即异常包；所谓负示例是指非异常示例（例如异常标记的值为 0 或者具有负标记），负包也即非异常包。

20 [46]因而，在初始化时，随机选择正包（positive bag）中的一个或几个示例（instance）赋予正标记（positive label），正包（positive bag）中的其余示例赋予负标记（negative label），对于负包（negative bag），其中的所有示例（instance）均赋予负标记（negative label）。

25 [47]基于以上的初始化，每个示例（instance）都会获得一个示例标记（label），依据这些示例训练一个分类器（Classifier）。当得到分类器（Classifier）之后，又可以对所有的示例（instance）来进行预测，从而更新示例（instance）的标记（label）。例如，具体的更新过程为，对于负包（negative bag），其中示例（instance）的标记（label）保持不变（均为负标记），对于正包（positive bag）中的示例（instance），首先根据分类器（Classifier）的结果来修改示例的标记；之后进行一轮检查，即如果某一个正包（positive bag）的所有示例（instance）都被赋予了负标记（negative label），需要选择其中得分

最低的（即判断为负示例的置信度最低的），赋予其正标记（positive label）。如此，完成对示例标记的更新，再给予更新的结果训练新的分类器（Classifier）。

[48]循环以上过程，直至前后两轮之间示例的标记不发生改变为止。

[49]需要注意的是，一开始的示例标记（instance label）的赋予是比较随机的，而在分类器训练的过程中，会渐渐的对示例标记（instance label）进行修正，修正之后的示例标记（instance label）又会使得训练的分类器更加准确。从而达到优化的效果。

[50]S305：基于分类训练，得到 URL 异常定位模型。

[51]下面针对一个具体 url 进行异常定位的过程，对本说明书实施例做示例性说明。

[52]对于一条正常的 URL，例如：

10 `http://render.alipay.com/p/s/alipay_site/wait?mintime=3&maxtime=5&fromspanner=goldetfprod_502`

[53]对于攻击者而言，往往通过修改其中的任何可能的部分，来达到攻击的目的，在实际业务中，由于域名字段（例如上述：`http://render.alipay.com`）基本都是固定的，只有参数请求部分会存在被修改的可能性，因而攻击往往由此产生。如下例中，攻击者通过
15 手动的修改一个正常的 URL，来达到执行 script 的目的。

[54]例如，异常 URL 示例为：

`http://render.alipay.com/p/s/alipay_site/wait?mintime=3>SCRiPT={Sleep}&maxtime=5
&fromspanner=goldetfprod_502`

[55]这里的异常字段就是“`mintime=3>SCRiPT={Sleep}`”。

20 [56]传统的技术手段，都是试图将 URL 进行统一的表示，即将每个 URL 表示成一个一定维度的特征向量的形式，这样的方法，可能发现异常的 URL，但没法准确定位到上面的示例中的异常字段。而准确的发现异常存在的字段，是希望能够看到的，因而，这一方法的目的就在于对异常字段进行检测。

[57]具体而言，将这一问题形式化成一个多示例学习的问题，在获取数据时，往往可得
25 知哪些 URL 是正常的，哪些是异常的，这也就意味着，对于一个 URL 的标记，往往是在 URL 整个层面的，对于每个字段是否有问题，往往是不可见的。这刚好符合多示例学习的假设：在多示例学习中，数据是以包（bag）的形式给出，一个包具有一个标记（label），包中具有多个示例（instance），示例（instance）的标记（label）是不知道

的，如果一个包中有一个示例（instance）是正示例（positive instance），那么这个包就是正包（positive bag），若所有示例（instance）都是负示例（negative instance），那么这个包就是负包（negative bag）。

- [58]还是以上述正常 URL 举例来说明：通过对 URL 的切分，得到不同的子字段，即
- 5 “mintime=3”；“maxtime=5”；“fromspanner=goldetfprod_502”这三部分（instance）（其中这三部分的 value 分别为“3”、“5”、“goldetfprod_502”）；这里一个 URL 就是一个包，它被表示成包含三个 instance 的形式。进一步的，对于每一字段的信息提取和表示，可以有各种不同的模式，例如，提取 value 不同的 pattern：字符总数、字母总数、数字总数、符号总数、不同字符数、不同字母数、不同数字数、不同符号数，由
- 10 此，以上示例中的包中的三个向量就可以分别表示为 [1,0,1,0,1,0,1,0], [1,0,1,0,1,0,1,0], [15,11,3,1,13,9,3,1]，该 URL 是一个正常的 URL，因而该包的标记是负的（即非异常 URL）。

- [59]同理，对于上述被修改的异常 URL，可通过模型识别出具体哪个示例异常，由该异常示例确定出对应的异常字段（异常子字段），即“SCRiPT={Sleep}”。由此便于异常
- 15 分析和异常预防。

- [60]在本说明书实施例中，通过对 URL 表示为对应多个字段的多个示例组成的包，利用 URL 异常定位模型对异常示例进行预测，从而定位出 URL 中的异常字段。基于多示例学习的 URL 异常定位，能够较好的预测出每天的访问数据中未发现的潜在威胁，由于
- 20 可以针对异常 URL 确定异常所在，从而可为发现潜在威胁、建立新的安全规则、构建安全系统提供巨大的支持。

[61]第三方面，基于同一发明构思，本说明书实施例提供一种 URL 异常定位装置，请参考图 4，包括：

- [62]切分单元 401，用于对所述 URL 进行字段切分，得到由对应各个字段的各个示例组成的多示例包；
- 25 [63]预测单元 402，用于将所述多示例包输入预设的 URL 异常定位模型进行异常示例预测；

[64]定位单元 403，用于根据异常示例定位出对应的异常字段。

[65]在一种可选方式中，还包括：

[66]模型训练单元 404，用于基于多示例学习算法对多个 URL 样本进行训练，得到所述

URL 异常定位模型。

[67]在一种可选方式中，所述模型训练单元 404 包括：

[68]样本切分在单元 4041，用于对 URL 样本集中各个 URL 样本进行字段切分，针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包；

5 [69]示例包集合子单元 4042，用于集合各个 URL 样本的多示例包得到多示例包集；

[70]训练子单元 4043，用于基于多示例学习算法，对多示例包集进行异常示例及非异常示例分类训练，得到所述 URL 异常定位模型。

[71]在一种可选方式中，所述异常示例和非异常示例通过示例的异常标记的值进行区分；

10 [72]所述训练子单元 4043 具体用于：对多示例包集中每个示例的异常标记的值进行初始化，并对异常标记的值进行迭代学习，更新调整出每个示例最终的异常标记的值。

[73]在一种可选方式中，所述预测单元 402 具体用于：根据所述 URL 异常定位模型，预测所述多示例包中的各个示例的异常标记的值，从而确定各个示例是否为异常示例。

[74]在一种可选方式中，还包括：

15 [75]异常 URL 确定单元 405，用于确定所述 URL 是否为异常 URL：如果所述 URL 对应的示例包中包括异常示例，则确定所述 URL 为异常 URL；如果所述 URL 对应的示例包中不包括异常示例，则确定所述 URL 为非异常 URL。

[76]在一种可选方式中，所述示例是由对应字段的特征向量表示的。

[77]在一种可选方式中，所述字段为 URL 中参数请求字段。

20 [78]第四方面，基于同一发明构思，本说明书实施例提供一种 URL 异常定位训练装置，请参考图 5，包括：

[79]样本获取单元 501，用于收集由多个 URL 样本组成的 URL 样本集；

[80]样本切分单元 502，用于对 URL 样本集中各个 URL 样本进行字段切分，针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包；

[81]示例包集合单元 503，用于集合各个 URL 样本的多示例包得到多示例包集；

25 [82]训练单元 504，用于基于多示例学习算法，对多示例包集进行异常示例及非异常示例分类训练，得到所述 URL 异常定位模型。

[83]在一种可选方式中，所述异常示例和非异常示例通过示例的异常标记的值进行区分；

[84]所述训练单元 504 具体用于:对多示例包集中每个示例的异常标记的值进行初始化,并对异常标记的值进行迭代学习,更新调整出每个示例最终的异常标记的值。

[85]第四方面,基于与前述实施例中基于流式计算的数据处理控制方法同样的发明构思,本发明还提供一种服务器,如图 6 所示,包括存储器 604、处理器 602 及存储在存储器 5 604 上并可在处理器 602 上运行的计算机程序,所述处理器 602 执行所述程序时实现前文所述 URL 异常定位方法的步骤。

[86]其中,在图 6 中,总线架构(用总线 600 来代表),总线 600 可以包括任意数量的互联的总线和桥,总线 600 将包括由处理器 602 代表的一个或多个处理器和存储器 604 代表的存储器的各种电路链接在一起。总线 600 还可以将诸如外围设备、稳压器和功率 10 管理电路等之类的各种其他电路链接在一起,这些都是本领域所公知的,因此,本文不再对其进行进一步描述。总线接口 606 在总线 600 和接收器 601 和发送器 603 之间提供接口。接收器 601 和发送器 603 可以是同一个元件,即收发机,提供用于在传输介质上与各种其他装置通信的单元。处理器 602 负责管理总线 600 和通常的处理,而存储器 604 可以被用于存储处理器 602 在执行操作时所使用的数据。

15 [87]第六方面,基于与前述实施例中 URL 异常定位方法的发明构思,本发明还提供一种计算机可读存储介质,其上存储有计算机程序,该程序被处理器执行时实现前文所述 URL 异常定位的任一方法的步骤。

[88]本说明书是参照根据本说明书实施例的方法、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 20 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设备。

25 [89]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设备的制造品,该指令设备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。

[90]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。

5 [91] 尽管已描述了本说明书的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例作出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本说明书范围的所有变更和修改。

[92] 显然，本领域的技术人员可以对本说明书进行各种改动和变型而不脱离本说明书的精神和范围。这样，倘若本说明书的这些修改和变型属于本说明书权利要求及其等同技术的范围之内，则本说明书也意图包含这些改动和变型在内。

权利要求书

- 1、一种 URL 异常定位方法，包括：
对 URL 进行字段切分，得到由对应各个字段的各个示例组成的多示例包；
将所述多示例包输入基于多示例学习的 URL 异常定位模型进行异常示例预测；
5 根据异常示例定位出对应的异常字段。
- 2、根据权利要求 1 所述的方法，还包括：
基于多示例学习算法对多个 URL 样本进行训练，得到所述 URL 异常定位模型。
- 3、根据权利要求 2 所述的方法，所述基于多示例学习算法对多个 URL 样本进行训练，得到所述 URL 异常定位模型，包括：
10 对 URL 样本集中各个 URL 样本进行字段切分，针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包；
集合各个 URL 样本的多示例包得到多示例包集；
基于多示例学习算法，对多示例包集进行异常示例及非异常示例分类训练；
基于所述分类训练，得到所述 URL 异常定位模型。
- 4、根据权利要求 3 所述的方法，所述异常示例和非异常示例通过示例的异常标记
15 的值进行区分；
所述对多示例包集进行异常示例及非异常示例分类训练包括：
对多示例包集中每个示例的异常标记的值进行初始化，并对异常标记的值进行迭代学习，更新调整出每个示例最终的异常标记的值。
- 5、根据权利要求 4 所述的方法，所述将所述多示例包输入预设的 URL 异常定位模型进行异常示例预测包括：
20 根据所述 URL 异常定位模型，预测所述多示例包中的各个示例的异常标记的值，从而预测所述多示例包中的各个示例是否为异常示例。
- 6、根据权利要求 1-5 任一项所述的方法，还包括：
25 确定所述 URL 是否为异常 URL：
如果所述 URL 对应的示例包中包括异常示例，则确定所述 URL 为异常 URL；
如果所述 URL 对应的示例包中不包括异常示例，则确定所述 URL 为非异常 URL。
- 7、根据权利要求 1-5 任一项所述的方法，所述示例是由对应字段的特征向量表示的。
- 8、根据权利要求 1-5 任一项所述的方法，所述字段为 URL 中参数请求字段。
- 9、一种 URL 异常定位训练方法，包括：

收集由多个 URL 样本组成的 URL 样本集;

对 URL 样本集中各个 URL 样本进行字段切分, 针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包;

集合各个 URL 样本的多示例包得到多示例包集;

5 基于多示例学习算法, 对多示例包集进行异常示例及非异常示例分类训练;

基于所述分类训练, 得到所述 URL 异常定位模型。

10、根据权利要求 9 所述的方法, 所述异常示例和非异常示例通过示例的异常标记的值进行区分;

所述对多示例包集进行异常示例及非异常示例分类训练包括:

10 对多示例包集中每个示例的异常标记的值进行初始化, 并对异常标记的值进行迭代学习, 更新调整出每个示例最终的异常标记的值。

11、一种 URL 异常定位装置, 包括:

切分单元, 用于对 URL 进行字段切分, 得到由对应各个字段的各个示例组成的多示例包;

15 预测单元, 用于将所述多示例包输入基于多示例学习的 URL 异常定位模型进行异常示例预测;

定位单元, 用于根据异常示例定位出对应的异常字段。

12、根据权利要求 11 所述的装置, 还包括:

模型训练单元, 用于基于多示例学习算法对多个 URL 样本进行训练, 得到所述 URL

20 异常定位模型。

13、根据权利要求 12 所述的装置, 所述模型训练单元, 包括:

样本切分单元, 用于对 URL 样本集中各个 URL 样本进行字段切分, 针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包;

示例包集合子单元, 用于集合各个 URL 样本的多示例包得到多示例包集;

25 训练子单元, 用于基于多示例学习算法, 对多示例包集进行异常示例及非异常示例分类训练, 得到所述 URL 异常定位模型。

14、根据权利要求 13 所述的装置, 所述异常示例和非异常示例通过示例的异常标记的值进行区分;

所述训练子单元具体用于: 对多示例包集中每个示例的异常标记的值进行初始化,

30 并对异常标记的值进行迭代学习, 更新调整出每个示例最终的异常标记的值。

15、根据权利要求 14 所述的装置, 所述预测单元具体用于: 根据所述 URL 异常定

位模型，预测所述多示例包中的各个示例的异常标记的值，从而确定各个示例是否为异常示例。

16、根据权利要求 11-15 任一项所述的装置，还包括：

- 5 异常 URL 确定单元，用于确定所述 URL 是否为异常 URL：如果所述 URL 对应的示例包中包括异常示例，则确定所述 URL 为异常 URL；如果所述 URL 对应的示例包中不包括异常示例，则确定所述 URL 为非异常 URL。

17、根据权利要求 11-15 任一项所述的装置，所述示例是由对应字段的特征向量表示的。

18、根据权利要求 11-15 任一项所述的装置，所述字段为 URL 中参数请求字段。

- 10 19、一种 URL 异常定位训练装置，包括：

样本获取单元，用于收集由多个 URL 样本组成的 URL 样本集；

样本切分单元，用于对 URL 样本集中各个 URL 样本进行字段切分，针对每个 URL 样本得到由对应各个字段的各个示例组成的多示例包；

示例包集合单元，用于集合各个 URL 样本的多示例包得到多示例包集；

- 15 训练单元，用于基于多示例学习算法，对多示例包集进行异常示例及非异常示例分类训练，得到所述 URL 异常定位模型。

20、根据权利要求 19 所述的装置，所述异常示例和非异常示例通过示例的异常标记的值进行区分；

- 20 所述训练单元具体用于：对多示例包集中每个示例的异常标记的值进行初始化，并对异常标记的值进行迭代学习，更新调整出每个示例最终的异常标记的值。

21、一种服务器，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权利要求 1-10 任一项所述方法的步骤。

22、一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现权利要求 1-10 任一项所述方法的步骤。