

发明名称：网络异常数据检测方法、装置、计算机设备和存储介质

相关申请的交叉引用

本申请要求于 2018 年 01 月 04 日提交中国专利局、申请号为
5 201810008923X、发明名称为“网络异常数据检测方法、装置、计算机设备
及存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请
中。

技术领域

10 本申请涉及一种网络异常数据检测方法、装置、计算机设备及存储介质。

背景技术

互联网的快速发展给人们的生活带来了极大便利的同时，也带来了新的
网络安全挑战，网站不可避免地会遭受恶意入侵或攻击，人们的财产、信息
15 安全受到威胁，当今的网络安全问题已引起了人们的高度重视。目前的网站
攻击威胁分析检测方法及系统大多基于网络日志分析，采取将日志数据归转
化为结构化数据进行存储，并基于特定的规则进行字符匹配。但是，上述检
测方法处理效率低、容易让攻击者发现匹配规律，造成网络攻击的漏报和误
报。

20

发明内容

根据本申请公开的各种实施例，提供一种网络异常数据检测方法、装置、
计算机设备及存储介质。

一种网络异常数据检测方法，包括：

25 接收客户端发送的访问请求数据；
查找访问请求数据中用户会话标识对应的历史访问请求数据；
获取所述访问请求数据的报文头字符串；
根据预设步长对所述报文头字符串进行分词处理得到分词集合；
根据所述历史访问请求数据和所述分词集合得到分词权重矩阵；
30 将所述分词权重矩阵输入异常数据检测模型得到数据异常概率；及
根据所述数据异常概率判断所述报文头字符串中是否存在异常数据。
一种网络异常数据检测装置包括：
请求接收模块，用于接收客户端发送的访问请求数据；
历史数据查找模块，用于查找访问请求数据中用户会话标识对应的历史

访问请求数据;

字符串获取模块, 用于获取所述访问请求数据的报文头字符串;

分词模块, 用于根据预设步长对所述报文头字符串进行分词处理得到分词集合;

5 权重获得模块, 用于根据所述历史访问请求数据和所述分词集合得到分词权重矩阵;

概率获得模块, 用于将所述分词权重矩阵输入预训练异常数据检测模型得到数据异常概率; 及

10 异常判断模块, 用于根据所述数据异常概率判断所述报文头字符串中是否存在异常数据。

一种计算机设备, 包括存储器和一个或多个处理器, 存储器中存储有计算机可读指令, 计算机可读指令被处理器执行时实现本申请任意一个实施例中提供的网络异常数据检测方法的步骤。

15 一个或多个存储有计算机可读指令的非易失性计算机可读指令存储介质, 计算机可读指令被一个或多个处理器执行时, 使得一个或多个处理器实现本申请任意一个实施例中提供的网络异常数据检测方法的步骤。

本申请的一个或多个实施例的细节在下面的附图和描述中提出。本申请的其它特征和优点将从说明书、附图以及权利要求书变得明显。

20 附图说明

为了更清楚地说明本申请实施例中的技术方案, 下面将对实施例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申请的一些实施例, 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, 还可以根据这些附图获得其它的附图。

25 图 1 为根据一个或多个实施例中网络异常数据检测方法的应用环境图。

图 2 为根据一个或多个实施例中网络异常数据检测方法的方法流程图。

图 3 为根据一个或多个实施例中网络异常数据检测装置的框图。

图 4 为根据一个或多个实施例中计算机设备的框图。

30 具体实施方式

为了使本申请的技术方案及优点更加清楚明白, 以下结合附图及实施例, 对本申请进行进一步详细说明。应当理解, 此处描述的具体实施例仅仅用以解释本申请, 并不用于限定本申请。

35 本申请提供的网络异常数据检测方法, 可以应用于如图 1 所示的应用环境中。服务器通过网络与客户端通过网络进行通信。客户端向服务器发起网

络访问请求，服务器接收客户端发送的访问请求数据，服务器根据接收的访问请求数据中的用户会话标识查找对应的用户的历史访问请求数据。服务器根据预设步长对访问请求数据中的报文头字符串进行分词处理，并计算出分词权重矩阵，将分词权重矩阵输入异常数据检测模型得到数据异常概率，根据数据异常概率判断客户端发送的访问请求数据中是否存在网络异常数据。从而能够提高网络异常数据检测的准确性，避免网络攻击的漏检、漏报。终端 102 可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

10 在一个实施例中，如图 2 所示，提供了一种网络异常数据检测方法，以该方法应用于图 1 中的服务器为例进行说明，包括以下步骤：

步骤 210，接收客户端发送的访问请求数据。

用户在客户端浏览网页时，客户端根据用户的网页访问地址向服务器进行网络访问请求，向服务器发送访问请求数据。访问请求数据中可以包括请求方法协议数据，请求报文头字符串和请求正文数据等。服务器接收客户端发送的访问请求数据。

步骤 220，查找访问请求数据中用户会话标识对应的历史访问请求数据。

客户端在向服务器发送访问请求数据时携带用户会话标识，用户会话标识用于对客户端发起的该用户的网络会话进行唯一标识，网络会话可以包括多次网络请求。客户端向服务器发起网络会话的第一次网络访问请求时并不携带用户会话标识，服务器接收到第一次网络访问请求生成请求会话标识，向客户端返回响应数据时携带生成的请求会话标识，请求会话标识用于唯一标识该网络会话。客户端接收请求会话标识后，根据请求会话标识、客户端信息及用户信息生成用户会话标识，在下次进行网络访问请求时携带用户会话标识。用户标识除了包含请求会话标识之外，还包含用户及其所使用客户端的特征标识。服务器在接收到客户端发送的访问请求数据之后，将访问请求数据与请求中携带的用户会话标识进行关联记录。

服务器在接收到客户端发送的访问请求数据之后，读取访问请求数据中的用户会话标识，查找用户会话标识所关联的此次网络会话的历史网路访问请求的历史访问请求数据。

步骤 230，获取访问请求数据的报文头字符串。

服务器从接收的访问请求数据中提取出报文头字符串。在一些实施例中，报文头字符串可以为所请求的 URL。在其他实施例中，报文头字符串也可以包括更多的请求头信息。

35 步骤 240，根据预设步长对报文头字符串进行分词处理得到分词集合。

服务器预先设定分词步长，例如可以将预设步长设定为 3 个字符、5 个字符等等。具体的预设步长可以根据实际经验进行设置。服务器根据预设步长按照预设切分顺序将提取出的报文头字符串依次切分成多个分词，预设切分顺序可以为由右至左的顺序，也可以为由左至右的顺序，也可以为其他顺序。

例如，报文头字符串为“/chapter17/user.html”，预设步长为 4 个四个字符，按照从左至右的顺序将报文头字符串切分成“/cha”、“pter”、“17/u”、“ser.”、“html” 5 个分词。若最后一个分词的字符数小于预设步长时，可以将最后一个分词的末尾填充空格字符使其字符数等于预设步长。

在一些实施例中，在对报文头字符串进行分词之前，对报文头字符串中的常见字符进行预处理，剔除其中的常见字符，对预处理后的报文头字符串再进行分词处理。例如，常见字符可以为“html”、“htm”等字符。服务器获取常见字符表，依次从常见字符表中取出常见字符，判断报文头字符串中是否存在常见字符，若存在，将相应的常见字符从报文头字符串中剔除。在一些实施例中，在进行常见字符的查找匹配时，需按照常见字符表中常见字符的排列顺序进行匹配，如在常见字符表中，“html”字符的排序高于“htm”字符，则先查找“html”，若在报文头字符串中查找不到，再查找“htm”字符。由此，可以提高常见字符剔除的合理性和有效性。

服务器得到切分后的分词后生成分词集合，分词集合可以为分词矩阵的形式，也可以为其他形式。

步骤 250，根据历史访问请求数据和分词集合得到分词权重矩阵。

服务器得到分词集合后，计算出各分词在报文头字符串中的出现的词频，计算出各分词在所有历史访问请求数据中出现的频率，根据上述计算得到的两个频率计算得到各分词的词权重并生成分词权重矩阵。

在一些实施例中，在计算各分词在所有历史访问请求数据中出现的频率之前，先从历史访问请求数据中提取出各自的报文头字符串，只计算各分词在各历史访问请求数据的报文头字符串中出现的频率。进一步地，若对本次访问请求数据的报文头字符串进行切词之前进行了预处理工作，则对历史访问请求数据的报文头字符串也进行预处理工作，从中剔除常见字符后再进行各分词出现频率的计算。

步骤 260，将分词权重矩阵输入异常数据检测模型得到数据异常概率。

服务器将分词权重矩阵输入异常数据检测模型得到访问请求数据的数据异常概率。异常数据检测模型是通过将已知样本数据输入初始异常数据检测模型进行模型训练、模型参数优化而得到的。

步骤 270，根据数据异常概率判断报文头字符串中是否存在异常数据。

服务器获取预设概率临界值，将计算得到的数据异常概率与预设概率临界值进行比较，并根据预设判断规则对比较结果进行判断。如，预设概率临界值设定为 0.5，当数据异常概率大于等于 0.5 时，则判断报文头字符串中存在异常数据，当数据异常概率小于 0.5 时，则判断报文头字符串中不存在异常数据。需要说明的是，预设概率临界值和预设判断规则均是根据具体的异常数据检测模型进行设定的，不限于上述示例。在一些实施例中，当服务器判断出访问请求数据中存在异常数据时，生成网络访问请求异常警报并进行提示。当服务器判断出访问请求数据中不存在异常数据时，将此次接收的访问请求数据与用户会话标识进行关联存储，便于后续进行该会话的访问请求数据的查找。

在本实施例中，服务器对预先训练的异常数据检测模型进行存储，当接收到访问请求后，查找此次访问请求所属会话的历史访问请求数据，对访问请求中的报文头字符串进行分词处理并得到分词权重矩阵，将矩阵输入异常数据检测模型中得到数据异常概率，根据数据异常概率判断请求中是否存在网络攻击，从而能够根据访问请求所属会话的访问请求数据并利用异常数据检测模型对访问请求进行用户行为分析，有效地识别出网络攻击行为和正常访问行为，提高异常数据检测的准确性，减少网络攻击漏报误报的概率。

在一些实施例中，接收客户端发送的访问请求数据的步骤 210 之后，还包括：读取访问请求数据中的用户会话标识；当用户会话标识中包括请求会话标识时，将用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比较；当用户指纹标识与上一访问请求数据的用户指纹标识不一致时，则生成新的请求会话标识，并在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端。

在本实施例中，用户会话标识包括请求会话标识和用户指纹标识两部分。请求会话标识是服务器在建立网络会话时随机生成的，请求会话标识可以为 cookie 字段，采用 setcookie 的方式下发给客户端。而用户指纹标识由客户端根据客户端特征信息生成的指纹标识。客户端特征信息可以包括用户账号信息、客户端浏览器的型号、版本、使用语言、字体、安装插件等信息，网络环境信息如公网和内网的 ip 地址等，以及客户端屏幕分辨率等其他相关信息。用户指纹信息可以反映当前客户端所处的环境信息。当使用的浏览器型号、所处的网络或者其他客户端特征发生变化时，客户端生成的用户指纹信息也会随之发生改变。

在一些实施例中，客户端生成用户指纹标识的步骤包括：客户端提取出客户端特征，查找各客户端特征对应的特征值，获取预先设置的各客户端特征对应的特征权重，根据各客户端特征对应的特征值和特征权重进行计算得

到用户指纹标识，例如，可以采用哈希算法进行计算生成指纹值作为用户指纹标识。由此可以对使用该客户端的用户进行唯一标识。

服务器读取用户会话标识时，首先判断用户会话标识中是否包括请求会话标识，当判断出没有请求会话标识时，表明是客户端发出的新的会话的访问请求。服务器随机生成请求会话标识并建立新的会话，在向客户端返回响应数据时将新生成的请求会话标识一同返回。

当服务器判断出用户会话标识包括请求会话标识时，服务器获取与请求会话标识对应的客户端的上一访问请求的访问请求数据，并从中读取用户指纹标识。服务器将上一访问请求数据中的用户指纹标识与本次接收的用户指纹标识进行比较，当两个用户指纹标识不一致时，服务器生成新的会话标识并建立新的会话，在向客户端返回响应数据时将新的请求会话标识发送给客户端，以使客户端在下次发送访问请求时携带新的会话标识及客户端的用户指纹标识。

当服务器判断出用户指纹标识与上一访问请求数据的用户指纹标识一致时，则继续执行查找访问请求数据中用户会话标识对应的历史访问请求数据的步骤。

在本实施例中，服务器将客户端发送的访问请求数据按照用户会话标识进行分类存储。根据用户会话标识可以查找到与其对应的会话的访问请求数据。通过请求会话标识可以查找到本次会话上一次访问请求的访问请求数据，通过将上一次访问请求数据中的用户指纹标识与本次请求的用户指纹标识进行比较，可以判断客户端所处环境是否发生变化，当两次指纹不一致时，判断客户端环境发生变化，可能会存在网络攻击的风险，则终止连续的网络会话，生成新的会话标识并新建会话。从而可以通过用户指纹标识对客户端环境进行验证，降低网络攻击的可能。此外，通过将客户端的访问请求数据与用户会话标识对应存储，可以在检测到网络异常数据时迅速追踪到相应的历史访问请求数据，对网络攻击行为进行快速定位，提高工作效率。

在一些实施例中，接收客户端发送的访问请求的步骤 210 之后，还包括：读取访问请求数据中的用户会话标识；当用户会话标识中包括请求会话标识时，将用户会话标识中的用户指纹标识与接收的上一访问请求数据中的用户指纹标识进行比较；当用户指纹标识与上一访问请求的用户指纹标识一致时，获取当前会话的会话时长；当会话时长未超过预设会话时长阈值时，执行查找访问请求数据中用户会话标识对应的历史访问请求数据的步骤。

当服务器判断出用户指纹标识与上一次访问请求的用户指纹标识一致时，判断当前客户端所处环境没有发生变化，可以继续当前会话。服务器获取当前会话的会话时长。具体地，服务器查找与用户会话标识对应的第一次

访问请求的时间及当前时间，计算两个时间的时间间隔得到当前会话的会话时长。服务器获取预设会话时长阈值，预设会话时长阈值表示一个有效会话的最大存续时间，例如，预设会话时长阈值可以为 10 分钟、20 分钟、1 个小时、1 天等。预设会话时长阈值可以根据具体的访问请求的类型进行设定。

5 服务器将当前会话的会话时长与预设会话时长阈值进行比较，当会话时长未超过预设会话时长阈值时，表示当前会话为有效会话，客户端可以继续访问。服务器继续执行查找访问请求数据中用户会话标识对应的历史访问请求数据的步骤，进行网络异常数据的检测过程。

10 在一些实施例中，所述方法还包括：当会话时长超过预设会话时长阈值时，则生成新的请求会话标识，在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端。

当会话时长超过预设会话时长阈值时，服务器判断当前会话已经超过有效时间，为无效会话，服务器终止当前会话，生成新的请求会话标识并建立新的会话，在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端，以使客户端下一次发送访问请求数据时携带新的请求会话标识及客户端的用户指纹标识。

在本实施例中，通过设定预设会话时长阈值，可以对一次会话的最大时长进行限定，以保证会话的有效性，防止由于会话时间过长所造成的对用户网络信息的窃取、攻击行为，提高用户进行网络访问的安全性。

20 在一些实施例中，根据所述历史访问请求数据和所述分词集合得到分词权重矩阵的步骤，包括：计算分词集合中各分词在报文头字符串中的词频；计算分词集合中各分词在历史访问请求数据中的逆向文件频率；根据计算得到的各分词的词频和逆向文件频率计算得到各分词的词权重；根据各分词的词权重生成分词集合的分词权重矩阵。

25 在本实施例中，采用词频-逆向文件频率的信息检索方法来作为文本的特征提取。各分词在报文头字符串中的词频是指该分词在报文头字符串中出现的次数除以该报文头字符串的分词总数，如一个分词“/ch”在报文头字符串中出现了 2 次，而该报文头字符串的分词总数为 10 个，则分词“/ch”在报文头字符串中的词频为 $2/10=0.2$ 。逆向文件频率是指历史访问请求数据中报文头字符串的总份数除以所有报文头字符串中出现过该词频的个数再求对数。例如，“/ch”分词在 20 个报文头字符串中出现了 2 次，则逆向文件频率为 $\log(20/2)=1$ 。各分词的词权重为各分词的词频与逆向文件频率的乘积。参照上述示例，分词“/ch”的词权重为 $0.2*1=0.2$ 。服务器按照上述方法计算出分词集合中所有分词的词权重，并得到分词结合的分词权重矩阵。在一些
35 实施例中，分词权重矩阵的行和列的排布按照异常数据检测模型的输入要求

进行设定，在此不做特别限定。

本实施例中采用的词频-逆向文件频率的信息检索方法能够很好地度量一个词语的普遍重要度，因此，采用上述方法能够对异常分词在分词权重上进行区别，为后续的异常数据检测模型提供良好的数据。需要说明的是，在其他实施例中，也可以采用其他能够度量分词普遍重要度的特征计算方法计算分词集合的分词权重，并不限于本实施例中描述的方法。

在一些实施例中，所述方法还包括：获取访问请求数据样本集以及访问请求数据样本集中各访问请求数据样本的期望值；计算各访问请求数据样本的分词权重矩阵；将各访问请求数据样本的期望值和分词权重矩阵输入初始异常数据检测模型进行训练并得到最优回归系数矩阵；根据最优回归系数矩阵修正初始异常数据检测模型得到异常数据检测模型。

服务器中预先存储了异常数据检测模型，异常数据检测模型是根据样本数据对初始异常数据检测模型进行训练得到的。在本实施例中，初始异常数据检测模型为逻辑回归算法模型： $h_{\theta_0}(x) = \frac{1}{1 + e^{-\theta_0^T x}}$ ，其中， $h_{\theta_0}(x)$ 为数据异常概率， $h_{\theta_0}(x)$ 的取值范围为(0,1)， x 为分词权重矩阵， θ_0 为初始回归系数矩阵。在其他实施例中，异常数据检测模型也可以为其他算法模型，并不限于逻辑回归算法模型。

服务器收集正常网络访问请求数据样本和异常网络访问请求数据样本形成访问请求数据样本，并将正常网络访问数据和异常网络访问数据的期望值进行设置。例如，设定正常网络访问数据的期望值为0，异常网络访问数据的期望值为1，也可以设定为其他数值，并不限于上述示例。

服务器计算收集的各访问请求数据样本的分词权重矩阵，根据预设步长对各访问请求数据样本的报文头字符串进行分词处理得到分词集合，计算出分词集合中各分词在报文头字符串的词频，再计算出各分词在所有访问数据请求样本中的出现的逆向文件频率，再根据计算出的词频和逆向文件频率计算出各分词的词权重，由此得到该访问请求数据样本的分词权重矩阵。根据上述方法计算出各访问请求数据样本的分词权重矩阵。

服务器将各访问请求数据样本设定的期望值与对应的分词权重矩阵输入上述初始异常数据检测模型进行训练，根据训练结果不断对模型中的回归系数矩阵进行优化，最终得到该模型的最优回归系数矩阵 θ_F 。根据最优回归系数矩阵 θ_F 修正初始异常数据检测模型得到异常数据检测模型 $h_{\theta_F}(x) = \frac{1}{1 + e^{-\theta_F^T x}}$ ，并将异常数据检测模型进行存储，以进行后续的网络异常数据检测。

在本实施例中，通过实现采集大量的访问请求数据样本对初始的异常数据检测模型进行训练，能够对回归系数矩阵进行不断优化，得到最终的检测模型，从而提高异常数据检测模型的准确率。

应该理解的是，虽然图 2 的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图 2 中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

在一个实施例中，如图 3 所示，提供了一种网络异常数据检测装置，包括：请求接收模块 310、历史数据查找模块 320、字符串获取模块 330、分词模块 340、权重获得模块 350、概率获得模块 360 和异常判断模块 370，其中：

请求接收模块 310，用于接收客户端发送的访问请求数据。

历史数据查找模块 320，用于查找访问请求数据中用户会话标识对应的历史访问请求数据。

字符串获取模块 330，用于获取访问请求数据的报文头字符串。

分词模块 340，用于根据预设步长对报文头字符串进行分词处理得到分词集合。

权重获得模块 350，用于根据历史访问请求数据和分词集合得到分词权重矩阵。

概率获得模块 360，用于将分词权重矩阵输入预训练异常数据检测模型得到数据异常概率。

异常判断模块 370，用于根据数据异常概率判断报文头字符串中是否存在异常数据。

在一些实施例中，所述装置还包括：

标识读取模块，用于读取访问请求数据中的用户会话标识。

标识比较模块，用于当用户会话标识中包括请求会话标识时，将用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比较。

标识返回模块，用于当用户指纹标识与上一访问请求数据的用户指纹标识不一致时，则生成新的请求会话标识，并在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端。

在一些实施例中，所述装置还包括：

标识读取模块，用于读取访问请求数据中的用户会话标识。

标识比较模块，用于当用户会话标识中包括请求会话标识时，将用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比

较。

时长获取模块，用于当用户指纹标识与上一访问请求的用户指纹标识一致时，获取当前会话的会话时长。

5 时长比较模块，用于当会话时长未超过预设会话时长阈值时，执行查找访问请求数据中用户会话标识对应的历史访问请求数据的步骤。

在一些实施例中，所述装置还包括：

标识生成模块，用于当会话时长超过预设会话时长阈值时，则生成新的请求会话标识，在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端。

10 在一些实施例中，所述权重获得模块 350 还包括：

词频计算模块，用于计算所述分词集合中各分词在所述报文头字符串中的词频。

文件频率模块，用于计算所述分词集合中各分词在所述历史访问请求数据中的逆向文件频率。

15 词权重计算模块，用于根据计算得到的各分词的词频和逆向文件频率计算得到各分词的词权重。

矩阵生成模块，用于根据各分词的词权重生成所述分词集合的分词权重矩阵。

在一些实施例中，所述装置还包括：

20 期望值获取模块，用于获取访问请求数据样本集以及所述访问请求数据样本集中各访问请求数据样本的期望值。

矩阵计算模块，用于计算各访问请求数据样本的分词权重矩阵。

系数矩阵获得模块，用于将各访问请求数据样本的期望值和分词权重矩阵输入所述初始异常数据检测模型进行训练并得到最优回归系数矩阵。

25 模型修正模块，用于根据所述最优回归系数矩阵修正所述初始异常数据检测模型得到异常数据检测模型。

关于网络异常数据检测装置的具体限定可以参见上文中对于网络异常数据检测方法的限定，在此不再赘述。上述网络异常数据检测装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

30 在一些实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图 4 所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。

该非易失性存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为非易失性存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储访问请求数据、异常数据检测模型等数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机可读指令被处理器执行时以实现一种网络异常数据检测方法。

本领域技术人员可以理解，图4中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

一种计算机设备，包括存储器和一个或多个处理器，存储器中储存有计算机可读指令，计算机可读指令被处理器执行时，使得一个或多个处理器执行以下步骤：接收客户端发送的访问请求数据；查找访问请求数据中用户会话标识对应的历史访问请求数据；获取访问请求数据的报文头字符串；根据预设步长对报文头字符串进行分词处理得到分词集合；根据历史访问请求数据和分词集合得到分词权重矩阵；将分词权重矩阵输入异常数据检测模型得到数据异常概率；及根据数据异常概率判断报文头字符串中是否存在异常数据。

在一些实施例中，处理器执行计算机可读指令时还实现以下步骤：读取访问请求数据中的用户会话标识；当用户会话标识中包括请求会话标识时，将用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比较；及当用户指纹标识与上一访问请求数据的用户指纹标识不一致时，则生成新的请求会话标识，并在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端。

在一些实施例中，处理器执行计算机可读指令时还实现以下步骤：读取访问请求数据中的用户会话标识；当用户会话标识中包括请求会话标识时，将用户会话标识中的用户指纹标识与接收的上一访问请求数据中的用户指纹标识进行比较；当用户指纹标识与上一访问请求的用户指纹标识一致时，获取当前会话的会话时长；及当会话时长未超过预设会话时长阈值时，执行查找访问请求数据中用户会话标识对应的历史访问请求数据的步骤。

在一些实施例中，处理器执行计算机可读指令时还实现以下步骤：当会话时长超过预设会话时长阈值时，则生成新的请求会话标识，在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端。

在一些实施例中，处理器执行根据历史访问请求数据和分词集合得到分词权重矩阵的步骤时还用以实现：计算分词集合中各分词在报文头字符串中的词频；计算分词集合中各分词在历史访问请求数据中的逆向文件频率；根

据计算得到的各分词的词频和逆向文件频率计算得到各分词的词权重；及根据各分词的词权重生成分词集合的分词权重矩阵。

5 在一些实施例中，处理器执行计算机可读指令时还实现以下步骤：获取访问请求数据样本集以及访问请求数据样本集中各访问请求数据样本的期望值；计算各访问请求数据样本的分词权重矩阵；将各访问请求数据样本的期望值和分词权重矩阵输入初始异常数据检测模型进行训练并得到最优回归系数矩阵；及根据最优回归系数矩阵修正初始异常数据检测模型得到异常数据检测模型。

10 一个或多个存储有计算机可读指令的非易失性计算机可读指令存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行以下步骤：接收客户端发送的访问请求数据；查找访问请求数据中用户会话标识对应的历史访问请求数据；获取访问请求数据的报文头字符串；根据预设步长对报文头字符串进行分词处理得到分词集合；根据历史访问请求数据和分词集合得到分词权重矩阵；将分词权重矩阵输入异常数据检测模型得到数据异常概率；及根据数据异常概率判断报文头字符串中是否存在异常数据。
15

20 在一些实施例中，计算机可读指令被处理器执行时还实现以下步骤：读取访问请求数据中的用户会话标识；当用户会话标识中包括请求会话标识时，将用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比较；及当用户指纹标识与上一访问请求数据的用户指纹标识不一致时，则生成新的请求会话标识，并在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端。

25 在一些实施例中，计算机可读指令被处理器执行时还实现以下步骤：读取访问请求数据中的用户会话标识；当用户会话标识中包括请求会话标识时，将用户会话标识中的用户指纹标识与接收的上一访问请求数据中的用户指纹标识进行比较；当用户指纹标识与上一访问请求的用户指纹标识一致时，获取当前会话的会话时长；及当会话时长未超过预设会话时长阈值时，执行查找访问请求数据中用户会话标识对应的历史访问请求数据的步骤。

30 在一些实施例中，计算机可读指令被处理器执行时还实现以下步骤：当会话时长超过预设会话时长阈值时，则生成新的请求会话标识，在向客户端返回与访问请求数据对应的响应数据时，将新的请求会话标识发送至客户端。

35 在一些实施例中，计算机可读指令被处理器执行根据历史访问请求数据和分词集合得到分词权重矩阵的步骤时还用以实现：计算分词集合中各分词在报文头字符串中的词频；计算分词集合中各分词在历史访问请求数据中的逆向文件频率；根据计算得到的各分词的词频和逆向文件频率计算得到各分

词的词权重；及根据各分词的词权重生成分词集合的分词权重矩阵。

5 在一些实施例中，计算机可读指令被处理器执行时还实现以下步骤：获取访问请求数据样本集以及访问请求数据样本集中各访问请求数据样本的期望值；计算各访问请求数据样本的分词权重矩阵；将各访问请求数据样本的期望值和分词权重矩阵输入初始异常数据检测模型进行训练并得到最优回归系数矩阵；及根据最优回归系数矩阵修正初始异常数据检测模型得到异常数据检测模型。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性计算机可读取存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器（ROM）、可编程 ROM（PROM）、电可编程 ROM（EPROM）、电可擦除可编程 ROM（EEPROM）或闪存。易失性存储器可包括随机存取存储器（RAM）或者外部高速缓冲存储器。作为说明而非局限，RAM 以多种形式可得，诸如静态 RAM（SRAM）、动态 RAM（DRAM）、同步 DRAM（SDRAM）、双数据率 SDRAM（DDRSDRAM）、增强型 SDRAM（ESDRAM）、同步链路（Synchlink）DRAM（SLDRAM）、存储器总线（Rambus）直接 RAM（RDRAM）、直接存储器总线动态 RAM（DRDRAM）、以及存储器总线动态 RAM（RDRAM）等。

以上所述实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

25 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

权利要求书

- 1、一种网络异常数据检测方法，包括：
接收客户端发送的访问请求数据；
查找访问请求数据中用户会话标识对应的历史访问请求数据；
获取所述访问请求数据的报文头字符串；
5 根据预设步长对所述报文头字符串进行分词处理得到分词集合；
根据所述历史访问请求数据和所述分词集合得到分词权重矩阵；
将所述分词权重矩阵输入异常数据检测模型得到数据异常概率；及
根据所述数据异常概率判断所述报文头字符串中是否存在异常数据。
- 2、根据权利要求1所述的方法，其特征在于，所述接收客户端发送的访问请求数据之后，还包括：
10 读取所述访问请求数据中的用户会话标识；
当所述用户会话标识中包括请求会话标识时，将所述用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比较；及
当所述用户指纹标识与上一访问请求数据的用户指纹标识不一致时，则
15 生成新的请求会话标识，并在向所述客户端返回与所述访问请求数据对应的响应数据时，将所述新的请求会话标识发送至所述客户端。
- 3、根据权利要求1所述的方法，其特征在于，所述接收客户端发送的访问请求之后，还包括：
读取所述访问请求数据中的用户会话标识；
20 当所述用户会话标识中包括请求会话标识时，将所述用户会话标识中的用户指纹标识与接收的上一访问请求数据中的用户指纹标识进行比较；
当所述用户指纹标识与上一访问请求的用户指纹标识一致时，获取当前会话的会话时长；及
当所述会话时长未超过预设会话时长阈值时，执行查找访问请求数据中
25 用户会话标识对应的历史访问请求数据。
- 4、根据权利要求3所述的方法，其特征在于，所述方法还包括：
当所述会话时长超过预设会话时长阈值时，则生成新的请求会话标识，在向所述客户端返回与所述访问请求数据对应的响应数据时，将所述新的请求会话标识发送至所述客户端。
- 30 5、根据权利要求1所述的方法，其特征在于，所述根据所述历史访问请求数据和所述分词集合得到分词权重矩阵，包括：
计算所述分词集合中各分词在所述报文头字符串中的词频；
计算所述分词集合中各分词在所述历史访问请求数据中的逆向文件频率；

根据计算得到的各分词的词频和逆向文件频率计算得到各分词的词权重；及

根据各分词的词权重生成所述分词集合的分词权重矩阵。

6、根据权利要求1所述的方法，其特征在于，所述方法还包括：

5 获取访问请求数据样本集以及所述访问请求数据样本集中各访问请求数据样本的期望值；

计算各访问请求数据样本的分词权重矩阵；

将各访问请求数据样本的期望值和分词权重矩阵输入所述初始异常数据检测模型进行训练并得到最优回归系数矩阵；及

10 根据所述最优回归系数矩阵修正所述初始异常数据检测模型得到异常数据检测模型。

7、一种网络异常数据检测装置，包括：

请求接收模块，用于接收客户端发送的访问请求数据；

15 历史数据查找模块，用于查找访问请求数据中用户会话标识对应的历史访问请求数据；

字符串获取模块，用于获取所述访问请求数据的报文头字符串；

分词模块，用于根据预设步长对所述报文头字符串进行分词处理得到分词集合；

20 权重获得模块，用于根据所述历史访问请求数据和所述分词集合得到分词权重矩阵；

概率获得模块，用于将所述分词权重矩阵输入预训练异常数据检测模型得到数据异常概率；及

25 异常判断模块，用于根据所述数据异常概率判断所述报文头字符串中是否存在异常数据。

8、根据权利要求7所述的装置，其特征在于，所述装置还包括：

标识读取模块，用于读取所述访问请求数据中的用户会话标识；

30 标识比较模块，用于当所述用户会话标识中包括请求会话标识时，将所述用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比较；及

标识返回模块，用于当所述用户指纹标识与上一访问请求数据的用户指纹标识不一致时，则生成新的请求会话标识，并在向所述客户端返回与所述访问请求数据对应的响应数据时，将所述新的请求会话标识发送至所述客户端。

35 9、一种计算机设备，包括存储器及一个或多个处理器，所述存储器中储

存有计算机可读指令，所述计算机可读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执行以下步骤：

接收客户端发送的访问请求数据；

查找访问请求数据中用户会话标识对应的历史访问请求数据；

5 获取所述访问请求数据的报文头字符串；

根据预设步长对所述报文头字符串进行分词处理得到分词集合；

根据所述历史访问请求数据和所述分词集合得到分词权重矩阵；

将所述分词权重矩阵输入异常数据检测模型得到数据异常概率；及

根据所述数据异常概率判断所述报文头字符串中是否存在异常数据。

10 10、根据权利要求 9 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

读取所述访问请求数据中的用户会话标识；

当所述用户会话标识中包括请求会话标识时，将所述用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比较；及

15 当所述用户指纹标识与上一访问请求数据的用户指纹标识不一致时，则生成新的请求会话标识，并在向所述客户端返回与所述访问请求数据对应的响应数据时，将所述新的请求会话标识发送至所述客户端。

11、根据权利要求 9 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

20 读取所述访问请求数据中的用户会话标识；

当所述用户会话标识中包括请求会话标识时，将所述用户会话标识中的用户指纹标识与接收的上一访问请求数据中的用户指纹标识进行比较；

当所述用户指纹标识与上一访问请求的用户指纹标识一致时，获取当前会话的会话时长；及

25 当所述会话时长未超过预设会话时长阈值时，执行查找访问请求数据中用户会话标识对应的历史访问请求数据。

12、根据权利要求 11 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

30 当所述会话时长超过预设会话时长阈值时，则生成新的请求会话标识，在向所述客户端返回与所述访问请求数据对应的响应数据时，将所述新的请求会话标识发送至所述客户端。

13、根据权利要求 9 所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时实现所述根据所述历史访问请求数据和所述分词集合得到分词权重矩阵，还包括：

35 计算所述分词集合中各分词在所述报文头字符串中的词频；

计算所述分词集合中各分词在所述历史访问请求数据中的逆向文件频率；

根据计算得到的各分词的词频和逆向文件频率计算得到各分词的词权重；及

5 根据各分词的词权重生成所述分词集合的分词权重矩阵。

14、根据权利要求9所述的计算机设备，其特征在于，所述处理器执行所述计算机可读指令时还执行以下步骤：

获取访问请求数据样本集以及所述访问请求数据样本集中各访问请求数据样本的期望值；

10 计算各访问请求数据样本的分词权重矩阵；

将各访问请求数据样本的期望值和分词权重矩阵输入所述初始异常数据检测模型进行训练并得到最优回归系数矩阵；及

根据所述最优回归系数矩阵修正所述初始异常数据检测模型得到异常数据检测模型。

15、一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行以下步骤：

15 接收客户端发送的访问请求数据；

查找访问请求数据中用户会话标识对应的历史访问请求数据；

获取所述访问请求数据的报文头字符串；

根据预设步长对所述报文头字符串进行分词处理得到分词集合；

根据所述历史访问请求数据和所述分词集合得到分词权重矩阵；

20 将所述分词权重矩阵输入异常数据检测模型得到数据异常概率；及

根据所述数据异常概率判断所述报文头字符串中是否存在异常数据。

16、根据权利要求15所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

读取所述访问请求数据中的用户会话标识；

25 当所述用户会话标识中包括请求会话标识时，将所述用户会话标识中的用户指纹标识与接收的上一访问请求数据的用户指纹标识进行比较；及

当所述用户指纹标识与上一访问请求数据的用户指纹标识不一致时，则生成新的请求会话标识，并在向所述客户端返回与所述访问请求数据对应的响应数据时，将所述新的请求会话标识发送至所述客户端。

30 17、根据权利要求15所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

读取所述访问请求数据中的用户会话标识；

当所述用户会话标识中包括请求会话标识时，将所述用户会话标识中的用户指纹标识与接收的上一访问请求数据中的用户指纹标识进行比较；

当所述用户指纹标识与上一访问请求的用户指纹标识一致时，获取当前会话的会话时长；及

- 5 当所述会话时长未超过预设会话时长阈值时，执行查找访问请求数据中用户会话标识对应的历史访问请求数据。

18、根据权利要求 16 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

- 10 当所述会话时长超过预设会话时长阈值时，则生成新的请求会话标识，在向所述客户端返回与所述访问请求数据对应的响应数据时，将所述新的请求会话标识发送至所述客户端。

19、根据权利要求 15 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器实现所述根据所述历史访问请求数据和所述分词集合得到分词权重矩阵，还包括：

- 15 计算所述分词集合中各分词在所述报文头字符串中的词频；
 计算所述分词集合中各分词在所述历史访问请求数据中的逆向文件频率；
 根据计算得到的各分词的词频和逆向文件频率计算得到各分词的词权重；及
20 根据各分词的词权重生成所述分词集合的分词权重矩阵。

20、根据权利要求 16 所述的存储介质，其特征在于，所述计算机可读指令被所述处理器执行时还执行以下步骤：

获取访问请求数据样本集以及所述访问请求数据样本集中各访问请求数据样本的期望值；

- 25 计算各访问请求数据样本的分词权重矩阵；
 将各访问请求数据样本的期望值和分词权重矩阵输入所述初始异常数据检测模型进行训练并得到最优回归系数矩阵；及

根据所述最优回归系数矩阵修正所述初始异常数据检测模型得到异常数据检测模型。