(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 112583860 B
(45)授权公告日 2021.05.18

(21)申请号 202110225129.2

(22)申请日 2021.03.02

(65)同一申请的已公布的文献号
申请公布号 CN 112583860 A

(43)申请公布日 2021.03.30

(73)专利权人 北京智慧易科技有限公司
地址 100085 北京市海淀区信息路甲28号11层D座11A-063

(72)发明人 赵利国 向永清 关涛

(51)Int.Cl.
H04L 29/06(2006.01)
G06N 3/08(2006.01)
G06N 3/04(2006.01)

(56)对比文件
CN 110213227 A,2019.09.06
CN 110225001 A,2019.09.10
WO 2018132178 A1,2018.07.19
US 10778705 B1,2020.09.15
赵智阳等.基于卷积神经网络的电网工控系统入侵检测算法.《计算机系统应用》.2020,第29卷(第8期),

审查员 张华晶

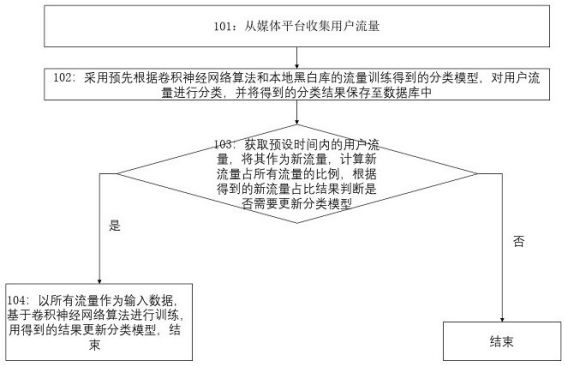
权利要求书2页 说明书9页 附图3页

(54)发明名称

一种互联网流量异常检测的方法、装置及设备

(57)摘要

本发明公开一种互联网流量异常检测的方法、装置及设备，属于互联网的流量技术领域，特别涉及一种互联网流量异常检测的方法，包括，从媒体平台收集用户流量；采用预先根据卷积神经网络算法和本地黑、白库的流量训练得到的分类模型，对所述用户流量进行分类，并将得到的分类结果保存至数据库中；获取预设时间内的用户流量，将其作为新流量，计算所述新流量占所有流量的比例，根据得到的新流量占比结果判断是否需要更新所述分类模型；在需要更新所述分类模型的情形下，以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型。



CN 112583860 B

1.一种互联网流量异常检测的方法,其特征在于,包括:

从媒体平台收集用户流量;

采用预先根据卷积神经网络算法和本地黑、白库的流量训练得到的分类模型,对所述用户流量进行分类,并将得到的分类结果保存至数据库中;

获取预设时间内的用户流量,将其作为新流量,计算所述新流量占所有流量的比例,根据得到的新流量占比结果判断是否需要更新所述分类模型;

在需要更新所述分类模型的情形下,以所有流量作为输入数据,基于卷积神经网络算法进行训练,用得到的结果更新分类模型。

2.如权利要求1所述的方法,其特征在于,

所述从媒体平台收集用户流量后还包括,根据用户的设备ID判断是否为新用户。

3.如权利要求2所述的方法,其特征在于,

根据用户的设备ID判断不是新用户的情形下,若用户的设备ID只出现在黑库中,则根据用户的设备ID最近出现在黑库中的时间范围对用户流量进行分类,并将得到的分类结果保存至数据库中。

4.如权利要求2所述的方法,其特征在于,

根据用户的设备ID判断不是新用户的情形下,若用户的设备ID出现在白库,则采用预先根据卷积神经网络算法和本地黑白库的流量训练得到的分类模型,对用户流量进行分类,并将得到的分类结果保存至数据库中。

5.如权利要求2所述的方法,其特征在于,

所述根据用户的设备ID判断是否为新用户,包括,根据用户的设备ID查询数据库,判断用户的设备ID是否在已经存在数据库中,若存在于数据库中,则判定不是新用户;若未存在于数据库中,则判定为新用户。

6.如权利要求1所述的方法,其特征在于,

所述采用预先根据卷积神经网络算法和本地黑、白库的流量训练得到的分类模型,对所述用户流量进行分类,并将得到的分类结果保存至数据库中,包括,

采用词向量工具对用户流量进行向量化,将得到的向量化结果输入到分类模型中,对用户流量进行分类,保存得到的分类结果至数据库中;

所述分类模型为采用预先以卷积神经网络算法为基础,以本地黑白库的流量作为数据,训练得到的模型。

7.如权利要求6所述的方法,其特征在于,

所述保存得到的分类结果至数据库中,包括,将异常流量保存至黑库中,将正常流量保存至白库中。

8.一种互联网流量异常检测的装置,其特征在于,包括:

流量采集模块,用于从媒体平台收集用户流量;

分类模块,与所述流量采集模块连接,用于采用预先根据卷积神经网络算法和本地黑、白库的流量训练得到的分类模型,对所述用户流量进行分类,并将得到的分类结果保存至数据库中;

判断模块,与所述流量采集模块连接,用于获取预设时间内的用户流量,将其作为新流量,计算所述新流量占所有流量的比例,根据得到的新流量占比结果判断是否需要更新所

述分类模型；

更新模块，分别与所述流量采集模块和所述判断模块连接，用于在需要更新所述分类模型的情形下，以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型。

9.一种计算机设备，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权利要求1-7任一项所述方法的步骤。

10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有程序，所述程序被执行时，能够实现如权利要求1-7任一项所述的方法。

一种互联网流量异常检测的方法、装置及设备

技术领域

[0001] 本发明属于互联网的流量技术领域，特别涉及一种互联网流量异常检测的方法、装置及设备。

背景技术

[0002] 随着网络技术的飞速发展，在当今互联网领域，每天都在源源不断的产生庞大的流量。在这庞大的流量中，隐藏着相当大一部分的异常流量，异常流量通常是由团体或个人伪造用户或用户行为产生的，异常流量的存在能够扰乱正常的业务进行。

[0003] 在当今，互联网流量对广告投放有着至关重要的作用，在广告的精准投放过程中，很容易被异常流量误导而做出错误的推送，影响广告业务的回报。对此，识别流量中的异常流量和作弊行为有着十分重要的意义。

[0004] 本发明经研究发现，现有技术中，检测和识别异常流量的方法准确率低。

发明内容

[0005] 为了至少解决上述技术问题，本发明提供了一种互联网流量异常检测的方法、装置及设备。

[0006] 根据本发明第一方面，提供了一种互联网流量异常检测的方法，包括：

[0007] 从媒体平台收集用户流量；

[0008] 采用预先根据卷积神经网络算法和本地黑、白库的流量训练得到的分类模型，对所述用户流量进行分类，并将得到的分类结果保存至数据库中；

[0009] 获取预设时间内的用户流量，将其作为新流量，计算所述新流量占所有流量的比例，根据得到的新流量占比结果判断是否需要更新所述分类模型；

[0010] 在需要更新所述分类模型的情形下，以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型。

[0011] 进一步地，所述从媒体平台收集用户流量后还包括，根据用户的设备ID判断是否为新用户。

[0012] 进一步地，所述根据用户的设备ID判断不是新用户的情形下，若用户的设备ID只出现在黑库中，则根据用户的设备ID最近出现在黑库中的时间范围对用户流量进行分类，并将得到的分类结果保存至数据库中。

[0013] 进一步地，所述根据用户的设备ID判断不是新用户的情形下，若用户的设备ID出现在白库，则采用预先根据卷积神经网络算法和本地黑白库的流量训练得到的分类模型，对用户流量进行分类，并将得到的分类结果保存至数据库中。

[0014] 进一步地，所述根据用户的设备ID判断是否为新用户，包括，根据用户的设备ID查询数据库，判断用户的设备ID是否在已经存在数据库中，若存在于数据库中，则判定不是新用户；若未存在于数据库中，则判定为新用户。

[0015] 进一步地，所述采用预先根据卷积神经网络算法和本地黑、白库的流量训练得到

的分类模型，对所述用户流量进行分类，并将得到的分类结果保存至数据库中，包括，

[0016] 采用词向量工具对用户流量进行向量化，将得到的向量化结果输入到分类模型中，对用户流量进行分类，保存得到的分类结果至数据库中；

[0017] 所述分类模型为采用预先以卷积神经网络算法为基础，以本地黑白库的流量作为数据，训练得到的模型。

[0018] 进一步地，所述保存得到的分类结果至数据库中，包括，将异常流量保存至黑库中，将正常流量保存至白库中。

[0019] 在本发明第二方面，一种互联网流量异常检测的装置，包括：

[0020] 流量采集模块，用于从媒体平台收集用户流量；

[0021] 分类模块，与所述流量采集模块连接，用于采用预先根据卷积神经网络算法和本地黑、白库的流量训练得到的分类模型，对所述用户流量进行分类，并将得到的分类结果保存至数据库中；

[0022] 判断模块，与所述流量采集模块连接，用于获取预设时间内的用户流量，将其作为新流量，计算所述新流量占所有流量的比例，根据得到的新流量占比结果判断是否需要更新所述分类模型；

[0023] 更新模块，分别与所述流量采集模块和所述判断模块连接，用于在需要更新所述分类模型的情形下，以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型。

[0024] 在本发明第三方面，一种计算机设备，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，

[0025] 所述处理器执行所述程序时实现如上任一项所述方法的步骤。

[0026] 在本发明第四方面，一种计算机可读存储介质，所述计算机可读存储介质存储有程序，所述程序被执行时，能够实现如上任一项所述的方法。

[0027] 本发明的有益效果：本发明基于卷积神经网络算法进行训练，用得到的结果更新分类模型，以保证分类模型在对流量进行分类时的精度和准确度本发明能够实现对异常流量准确高效的检测，并对互联网的流量进行实时的监测，为精准的广告投放提供可靠的流量，保障业务的正常开展。

附图说明

[0028] 本发明上述的和 / 或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中，

[0029] 图1为本发明提供的一种互联网流量异常检测的方法流程图；

[0030] 图2为本发明提供的另一种互联网流量异常检测的方法流程图；

[0031] 图3为本发明提供的一种分类模型工作流程图。

具体实施方式

[0032] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能解释为对本发明的限制。

[0033] 为了更清楚地说明本发明，下面结合优选实施例和附图对本发明做进一步的说明。附图中相似的部件以相同的附图标记进行表示。本领域技术人员应当理解，下面所具体描述的内容是说明性的而非限制性的，不应以此限制本发明的保护范围。

[0034] 在本发明的第一方面，提供了一种互联网流量异常检测的方法，如图1所示，包括：

[0035] 在本发明的第一方面，提供了一种互联网流量异常检测的方法，如图1所示，包括：

[0036] 步骤101：从媒体平台收集用户流量；

[0037] 在本发明实施例中，从媒体平台上收集流量，其中，流量包括，包括用户的基本信息和用户在媒体上的行为信息。本发明中，媒体平台包括手机端的各种应用程序，如：抖音，快手，今日头条，腾讯新闻等，用户流量包括用户基本信息和用户在应用程序端的行为。

[0038] 步骤102：采用预先根据卷积神经网络算法和本地黑白库的流量训练得到的分类模型，对用户流量进行分类，并将得到的分类结果保存至数据库中；

[0039] 在本发明实施例中，系统可以预先以卷积神经网络算法为基础，以本地黑白库的流量作为数据，训练出分类模型。其中，卷积神经网络算法可以为循环神经网络(RNN，Recurrent Neuron Network)。在对用户流量进行分类时，采用词向量工具对用户流量进行向量化，将得到的向量化结果输入到分类模型中，对用户流量进行分类。

[0040] 进一步地，可以采用词向量工具中的Word2Vec对用户流量进行向量化，包括，将用户流量映射为实数向量，得到文本的向量化。本实施例中，采用Word2vec能够高效准确的将自然语言映射为实数向量，也就是将文本数据转换为向量形式，即实现了文本的向量化。文本被映射为向量后，即可应用各种AI学习算法对其进行处理。

[0041] 本发明方法，通过使用Word2vec开源工具，使用简单高效，能够实现词嵌入的同时，比普通的词向量工具维度更低，训练速度更快，而且通用性强，几乎适用于所有的语言场景。

[0042] 在本发明的另一实施例中，根据分类模型输出的结果对用户流量进行分类，具体包括：在分类模型输出的结果为第一预设值的情形下，则判定用户流量为正常流量。相对应的，在分类模型输出的结果为第二预设值的情形下，则判定用户流量为异常流量。其中，第一预设值可以为1，第二预设值可以为0。

[0043] 在本发明的又一个实施例中，将异常流量保存至黑库中，将正常流量保存至白库中，进一步地，将保存异常流量的时间保存至黑库中，将保存正常流量的时间保存至白库中。

[0044] 在本发明中，数据库包括，黑库和白库，其中，黑库是用于存放异常流量的数据库，白库是用于存放正常流量的数据库。

[0045] 本发明中，通过采用分类模型，能够在训练过程使用分布式的方式，运行速度快，能够挖掘数据的时间序列和前后信息，而且可以处理任意长度的输入序列，没有序列长度的制约。

[0046] 步骤103：获取预设时间内的用户流量，将其作为新流量，计算新流量占所有流量的比例，根据得到的新流量占比结果判断是否需要更新分类模型，是则执行步骤104；否则结束。

[0047] 在本发明中，获取预设时间内的用户流量，具体可以获取最近一周收集的用户流量，作为新流量，一周以前的流量作为旧流量。

[0048] 计算新流量占所有流量的比例，判断得到的新流量占比是否大于等于30%，在大于等于30%的情形下，则判定新流量占比足够多，分类模型的时效性容易受到影响，此时需更新分类模型。在小于30%的情形下，则分类模型仍具有较好的时效性，不需要更新分类模型。

[0049] 步骤104：以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型，结束。

[0050] 本发明中，在得到的新流量占比大于等于30%的情形下，则判定新流量占比足够多，此时的分类模型的时效性容易受到影响，此时以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型，以保证分类模型在对流量进行分类时的精度和准确度。

[0051] 在本发明的另一实施例中，还包括判断是否有新的流量产生，是则返回步骤101；否则结束。

[0052] 本发明能够实现对异常流量准确高效的检测，并对互联网的流量进行实时的监测，为精准的广告投放提供可靠的流量，保障业务的正常开展。

[0053] 本发明的另一实施例提供一种互联网流量异常检测的方法，包括：

[0054] 步骤201：从媒体平台收集用户流量；

[0055] 在本发明实施例中，从媒体平台上收集流量，其中，流量包括，包括用户的基本信息和用户在媒体上的行为信息。本发明中，媒体平台包括手机端的各种应用程序，如：抖音，快手，今日头条，腾讯新闻等，用户流量包括用户基本信息和用户在应用程序端的行为。

[0056] 步骤202：根据用户的设备ID判断是否为新用户，是则执行步骤203；否则执行步骤204；

[0057] 在本发明中，根据用户的设备ID查询数据库，其中，设备ID是可以唯一识别用户身份的描述信息。

[0058] 判断用户的设备ID是否在已经存在数据库中，若存在于数据库中，则判定不是新用户，即为老用户；若未存在于数据库中，则判定为新用户。

[0059] 步骤203：采用预先根据卷积神经网络算法和本地黑白库的流量训练得到的分类模型，对用户流量进行分类，并将得到的分类结果保存至数据库中，执行步骤206；

[0060] 在本发明实施例中，系统可以预先以卷积神经网络算法为基础，以本地黑白库的流量作为数据，训练出分类模型。其中，卷积神经网络算法可以为循环神经网络(RNN，Recurrent Neuron Network)。在判定为新用户的情形下，采用词向量工具对用户流量进行向量化，将得到的向量化结果输入到分类模型中，对用户流量进行分类。

[0061] 进一步地，可以采用词向量工具中的Word2Vec对用户流量进行向量化，包括，将用户流量映射为实数向量，得到文本的向量化。本实施例中，采用Word2vec能够高效准确的将自然语言映射为实数向量，即实现了文本的向量化。文本被映射为向量后，即可应用各种AI学习算法对其进行处理。

[0062] 本发明方法，通过使用Word2vec开源工具，使用简单高效，能够实现词嵌入的同时，比普通的词向量工具维度更低，训练速度更快，而且通用性强，几乎适用于所有的语言场景。

[0063] 在本发明的另一实施例中，根据分类模型输出的结果对用户流量进行分类，具体包括：在分类模型输出的结果为第一预设值的情形下，则判定用户流量为正常流量，也就是

说，正常流量是真实的用户产生的真实的行为。相对应的，在分类模型输出的结果为第二预设值的情形下，则判定用户流量为异常流量，异常流量是指虚假的用户、真实用户的虚假流量。其中，第一预设值可以为1，第二预设值可以为0。

[0064] 在本发明的又一个实施例中，将异常流量保存至黑库中，将正常流量保存至白库中，进一步地，将保存异常流量的时间保存至黑库中，将保存正常流量的时间保存至白库中。

[0065] 在本发明中，数据库包括，黑库和白库，其中，黑库是用于存放异常流量的数据库，白库是用于存放正常流量的数据库。

[0066] 本发明中，通过采用分类模型，能够在训练过程使用分布式的方式，运行速度快，能够挖掘数据的时间序列和前后信息，而且可以处理任意长度的输入序列，没有序列长度的制约。

[0067] 步骤204：用户的设备ID只出现在黑库中的情形下，根据用户的设备ID最近出现在黑库中的时间范围对用户流量进行分类，并将得到的分类结果保存至数据库中，执行步骤206；

[0068] 在本发明实施例中，在用户的设备ID最近出现在黑库中的时间符合预设时间的情形下，判定用户流量为异常流量；在用户的设备ID最近出现在黑库中的时间不符合预设时间的情形下，采用预先根据卷积神经网络算法和本地黑白库的流量训练得到的分类模型，对用户流量进行分类，并将得到的分类结果保存至数据库中。

[0069] 步骤205：用户的设备ID出现在白库的情形下，采用预先根据卷积神经网络算法和本地黑白库的流量训练得到的分类模型，对用户流量进行分类，并将得到的分类结果保存至数据库中，执行步骤206；

[0070] 在本发明实施例中，用户的设备ID只出现在白库中的情形下，并不代表着用户本次的流量不是异常的，因此采取将用户输入训练的模型进行分类的方法，能够有效避免检测不准确的问题，提供检测精准度。

[0071] 在用户的设备ID在黑库和白库中都出现过的情形下，采取将用户输入训练的分类模型进行分类，能够有效克服用户的设备ID在黑库和白库中都出现过无法判定用户本次的流量是否正常的情况。

[0072] 步骤206：获取预设时间内的用户流量，将其作为新流量，计算新流量占所有流量的比例，根据得到的新流量占比结果判断是否需要更新分类模型，是则执行步骤207；否则执行步骤208。

[0073] 在本发明中，获取预设时间内的用户流量，具体可以获取最近一周收集的用户流量，作为新流量，一周以前的流量作为旧流量。

[0074] 计算新流量占所有流量的比例，判断得到的新流量占比是否大于等于30%，在大于等于30%的情形下，则判定新流量占比足够多，分类模型的时效性容易受到影响，此时需要重新训练分类模型，并进行更新。在小于30%的情形下，则分类模型仍具有较好的时效性，不需要更新分类模型。

[0075] 步骤207：以所有流量作为输入数据，更新分类模型，执行步骤208；

[0076] 本发明中，在得到的新流量占比大于等于30%的情形下，则判定新流量占比足够多，此时分类模型的时效性容易受到影响，以所有流量作为输入数据，重新训练分类模型，

更新分类模型。以保证分类模型的精度和准确度。

[0077] 步骤208:判断是否有新的流量产生,是则返回步骤201;否则结束。

[0078] 本发明能够实现对异常流量准确高效的检测,并对互联网的流量进行实时的监测,为精准的广告投放提供可靠的流量,保障业务的正常开展。

[0079] 在本发明的又一实施例中,提供一种互联网流量异常检测的方法,包括:

[0080] 从所在媒体平台收集用户流量,包括用户的基本信息和用户在媒体上的行为信息;用户的基本信息和用户的行为信息在下面有具体的描述。

[0081] 由于设备ID可以唯一的指定用户,故通过用户的设备ID查询黑白数据库,判断用户的设备ID是否在已经存在数据库中;

[0082] 若用户的ID存在数据库中,则标记为用户为已经存在的用户;

[0083] 若用户的ID不存在数据库中,则标记为用户为新的用户;

[0084] 对于新的用户而言,没有自己的历史流量作为参考,故直接使用已经训练的RNN模型对流量进行分类;流量需要先通过word2Vec进行向量化,然后输入模型中得出结果,模型输出为1,则属于正常流量,将该流量存储到白库中。模型输出为0,则属于异常流量,将该流量存储到黑库中。

[0085] 对于已经存在的用户而言,用户以往的流量可能存在如下情况:

[0086] 1、只出现在黑库中,即属于异常流量;

[0087] (1)用户最近出现在黑库中在一周内;

[0088] (2)用户最近出现在黑库中在一周以前;

[0089] 2、只出现在白库中,即属于正常流量;

[0090] 3、在黑库和白库中都出现过,即该用户曾经出现过异常行为;

[0091] 为了更准确的识别异常流量,在最大程度上减少业务的损失,保障正常工作的进行。针对上述几种情况,采取以下的处理方法,包括:

[0092] 对于1而言,用户只出现在异常流量中,说明一直属于异常流量。但用户可能由会由异常转为正常,故我们进一步细化用户出现在黑库中的最近时间,同样以周为单位,判断用户最近一次出现在黑库中是否属于一周以内;

[0093] 若是,则直接判断为异常流量;

[0094] 若不是,则用户有本次流量可能会发生变化,由异常转换为正常,故使用RNN算法模型,也就是分类模型进行分类。

[0095] 对于2而言,用户曾经出现在白库中,并不代表着用户本次的流量不是异常的,因此采取将用户输入训练的模型进行分类;

[0096] 对于3而言,用户曾经出现正常和异常流量,不能直接判断,因此使用RNN算法模型进行分类;

[0097] 对流量进行分类之后,将正常流量存储到白库中,将异常流量存储到黑库中;

[0098] 对用户流量进行分类后将最近一周收集的流量,作为新流量,一周以前的流量作为旧流量。计算新流量占所有流量的比例,判断是否比例超过30%;

[0099] 若计算比例超过30%,则说明新流量占比足够多,模型的时效性容易受到影响,此时重新训练RNN模型,更新RNN算法模型。

[0100] 若比例小于30%,则模型仍具有较好的时效性,不需要更新模型。

[0101] 本发明可以在对流量分类和存储数据库完成后，判断是否继续收集流量并处理；

[0102] 若继续，则转入收集步骤继续进行即可；

[0103] 若不继续，则结束程序。

[0104] 在本发明另一实施例中，用户的主要流量构成如下，包括：

[0105] 性别：男/女

[0106] 年龄：13岁以下，13-18周岁，19-26周岁，26-35周岁，36-45周岁，45-60周岁；60-75周岁；75周岁以上；

[0107] 职业：教育/培训、IT/互联网、医辽/制药、制造业、餐饮/服务业、事业单位、学生、法律行业、娱乐/传媒、金融/保险

[0108] 地理信息：农村、县城、北上广深、新一线/省会、普通地级市

[0109] 使用设备：手机端：Android /IOS、PC端

[0110] 收入水平：5k以下、5k-10k、10k-20k、20-30k、30k以上

[0111] 浏览偏好：教育/培训、IT/互联网、医疗/制药、制造业、餐饮/服务业、事业单位、学生、法律行业、娱乐/传媒、金融/保险

[0112] 娱乐爱好：音乐、舞蹈、运动、电竞游戏、动漫、电影、电视剧、综艺节目

[0113] 上线频率：以12h为最小单位，统计平均上线频率；

[0114] 上线行为：浏览媒体内容：如文章、视频、音乐；浏览时间、点击广告次数、是否没有任何行为；

[0115] 在线时长：以小时为单位，统计在线总时长。

[0116] 本发明通过以本地数据库为基础，用户流量被各种维度的数据描述，使用word2vec工具将用户流量进行向量化，后续作为RNN的数据输入，数据库分为白库和黑库，分别存储正常流量和异常流量。白库中的流量标签设置为1，黑库中的流量标签设置为0。

[0117] 训练RNN模型，以用户流量向量化后的数据作为输入，0、1标签作为输出，对RNN进行有监督的学习，训练得到RNN模型。

[0118] 对于待检测的用户流量，使用word2vec将流量嵌入向量空间。输入训练好的RNN模型，得到输出结果。

[0119] 根据RNN模型的输出结果，判断是否属于异常流量。0表示异常流量，1表示正常流量。

[0120] RNN模型训练需要在检测之前完成，以便快速完成输入流量的检测。

[0121] 在新流量占所有流量的比例超过30%时，需要重新训练RNN模型，以保证模型的时效性。

[0122] 在本发明的第二方面，提供一种互联网流量异常检测的装置，包括：

[0123] 流量采集模块，用于从媒体平台收集用户流量；

[0124] 在本发明实施例中，流量采集模块从媒体平台上收集流量，其中，流量包括，包括用户的基本信息和用户在媒体上的行为信息。本发明中，媒体平台包括手机端的各种应用程序，如：抖音，快手，今日头条，腾讯新闻等，用户流量包括用户基本信息和用户在应用程序端的行为。

[0125] 进一步地，本发明实施例中，还包括第一判断模块，与流量采集模块连接，用于根据用户的设备ID判断是否为新用户。

[0126] 在本发明中，第一判断模块根据用户的设备ID查询数据库，判断用户的设备ID是否在已经存在数据库中，若存在于数据库中，则判定不是新用户，即为老用户；若未存在于数据库中，则判定为新用户。

[0127] 分类模块，与所述流量采集模块连接，用于采用预先根据卷积神经网络算法和本地黑、白库的流量训练得到的分类模型，对所述用户流量进行分类，并将得到的分类结果保存至数据库中；

[0128] 在本发明实施例中，分类模块具体用于采用词向量工具对用户流量进行向量化，将得到的向量化结果输入到分类模型中，对用户流量进行分类，保存得到的分类结果至数据库中；

[0129] 所述分类模型为采用预先以卷积神经网络算法为基础，以本地黑白库的流量作为数据，训练得到的模型。

[0130] 进一步地，装置可以预先以卷积神经网络算法为基础，以本地黑白库的流量作为数据，训练出分类模型。其中，卷积神经网络算法可以为循环神经网络(RNN，Recurrent Neuron Network)。分类模块可以采用词向量工具对用户流量进行向量化，将得到的向量化结果输入到分类模型中，对用户流量进行分类。

[0131] 进一步地，分类模块可以采用词向量工具中的Word2Vec对用户流量进行向量化，包括，将用户流量映射为实数向量，得到文本的向量化。本实施例中，采用Word2vec能够高效准确的将自然语言映射为实数向量，即实现了文本的向量化。文本被映射为向量后，即可应用各种AI学习算法对其进行处理。

[0132] 本发明通过使用Word2vec开源工具，使用简单高效，能够实现词嵌入的同时，比普通的词向量工具维度更低，训练速度更快，而且通用性强，几乎适用于所有的语言场景。

[0133] 在本发明的另一实施例中，分类模块在分类模型输出的结果为第一预设值的情形下，则判定用户流量为正常流量。相对应的，在分类模型输出的结果为第二预设值的情形下，则判定用户流量为异常流量。其中，第一预设值可以为1，第二预设值可以为0。

[0134] 在本发明的又一个实施例中，将异常流量保存至黑库中，将正常流量保存至白库中，进一步地，将保存异常流量的时间保存至黑库中，将保存正常流量的时间保存至白库中。

[0135] 在本发明中，数据库包括，黑库和白库，其中，黑库是用于存放异常流量的数据库，白库是用于存放正常流量的数据库。

[0136] 本发明中，通过采用分类模型，能够在训练过程使用分布式的方式，运行速度快，能够挖掘数据的时间序列和前后信息，而且可以处理任意长度的输入序列，没有序列长度的制约。

[0137] 判断模块，与流量采集模块连接，用于获取预设时间内的用户流量，将其作为新流量，计算所述新流量占所有流量的比例，根据得到的新流量占比结果判断是否需要更新所述分类模型；

[0138] 在本发明中，判断模块，用于获取预设时间内的用户流量，具体可以获取最近一周收集的用户流量，作为新流量，一周以前的流量作为旧流量。

[0139] 判断模块还用于计算新流量占所有流量的比例，判断得到的新流量占比是否大于等于30%，在大于等于30%的情形下，则判定新流量占比足够多，分类模型的时效性容易受到

影响，此时需更新分类模型。在小于30%的情形下，则分类模型仍具有较好的时效性，不需要更新分类模型。

[0140] 更新模块，分别与所述流量采集模块和所述判断模块连接，用于在需要更新所述分类模型的情形下，以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型。

[0141] 本发明中，更新模块在得到的新流量占比大于等于30%的情形下，则判定新流量占比足够多，此时的分类模型的时效性容易受到影响，此时以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型，以保证分类模型在对流量进行分类时的精度和准确度。

[0142] 本发明的另一实施例中，还包括第一分类模块，与第一判断模块连接，用于所述根据用户的设备ID判断不是新用户的情形下，若用户的设备ID只出现在黑库中，则根据用户的设备ID最近出现在黑库中的时间范围对用户流量进行分类，并将得到的分类结果保存至数据库中。

[0143] 在本发明的另一实施例中，还包括第二分类模块，与第一判断模块连接，用于根据用户的设备ID判断不是新用户的情形下，若用户的设备ID出现在白库，则采用预先根据卷积神经网络算法和本地黑白库的流量训练得到的分类模型，对用户流量进行分类，并将得到的分类结果保存至数据库中。

[0144] 在本发明的第三方面，提供一种计算机设备，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如上任一项所述方法的步骤。

[0145] 在本发明的第四方面，提供一种计算机可读存储介质，所述计算机可读存储介质存储有程序，所述程序被执行时，能够实现如上任一项所述的方法。

[0146] 本技术领域技术人员可以理解，除非特意声明，这里使用的单数形式“一”、“一个”、“所述”和“该”也可包括复数形式。应该进一步理解的是，本发明的说明书中使用的措辞“包括”是指存在所述特征、整数、步骤、操作、元件和 / 或组件，但是并不排除存在或添加 一个或多个其他特征、整数、步骤、操作、元件、组件和 / 或它们的组。应该理解，当我们称元件被“连接”或“耦接”到另一元件时， 它可以直接连接或耦接到其他元件，或者也可以存在 中间元件。此外，这里使用的“连接”或“耦接”可以包括无线连接或无线耦接。这里使用 的措辞“和 / 或”包括一个或更多个相关联的列出项的全部或任一单元和全部组合。

[0147] 本技术领域技术人员可以理解，除非另外定义，这里使用的所有术语 (包括技术 术语和科学术语)，具有与本发明所属领域中的普通技术人员的一般理解相同的意义。还应 该理解的是，诸如通用字典中定义的那些术语，应该被理解为具有与现有技术的上下文中 的意义一致的意义，并且除非像这里一样被特定定义，否则不会用理想化或过于正式的含义来解释。

[0148] 应当理解，以上借助优选实施例对本发明的技术方案进行的详细说明是示意性的而非限制性的。本领域的普通技术人员在阅读本发明说明书的基础上可以对各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

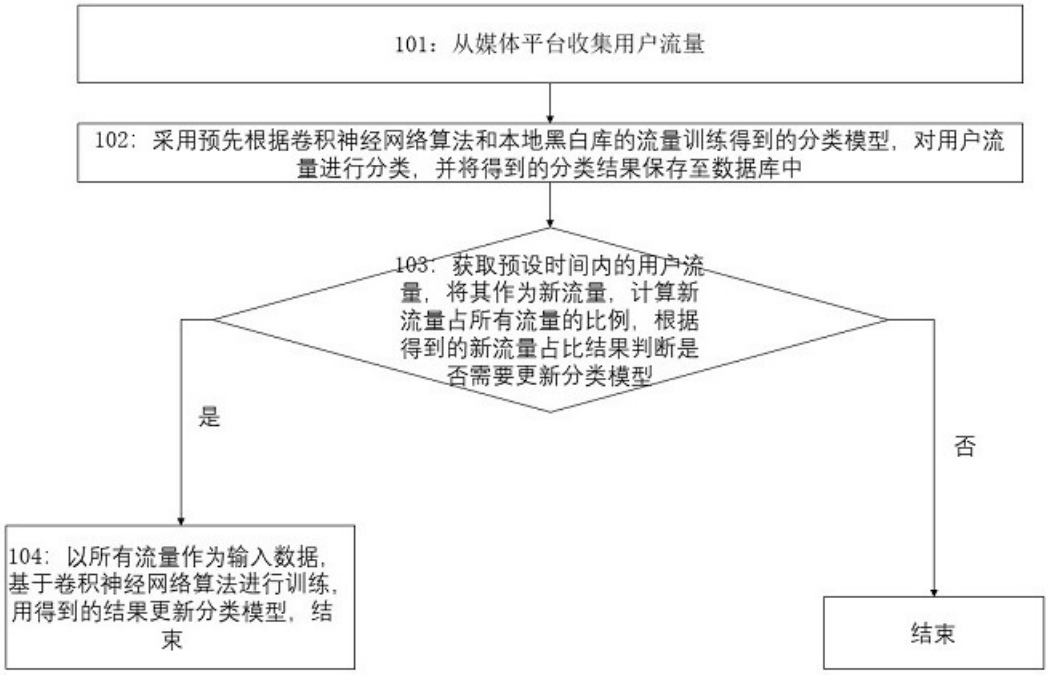

101：从媒体平台收集用户流量
102：采用预先根据卷积神经网络算法和本地黑白库的流量训练得到的分类模型，对用户流量进行分类，并将得到的分类结果保存至数据库中
103：获取预设时间内的用户流量，将其作为新流量，计算新流量占所有流量的比例，根据得到的新流量占比结果判断是否需要更新分类模型
是
否
104：以所有流量作为输入数据，基于卷积神经网络算法进行训练，用得到的结果更新分类模型，结束
结束

图1

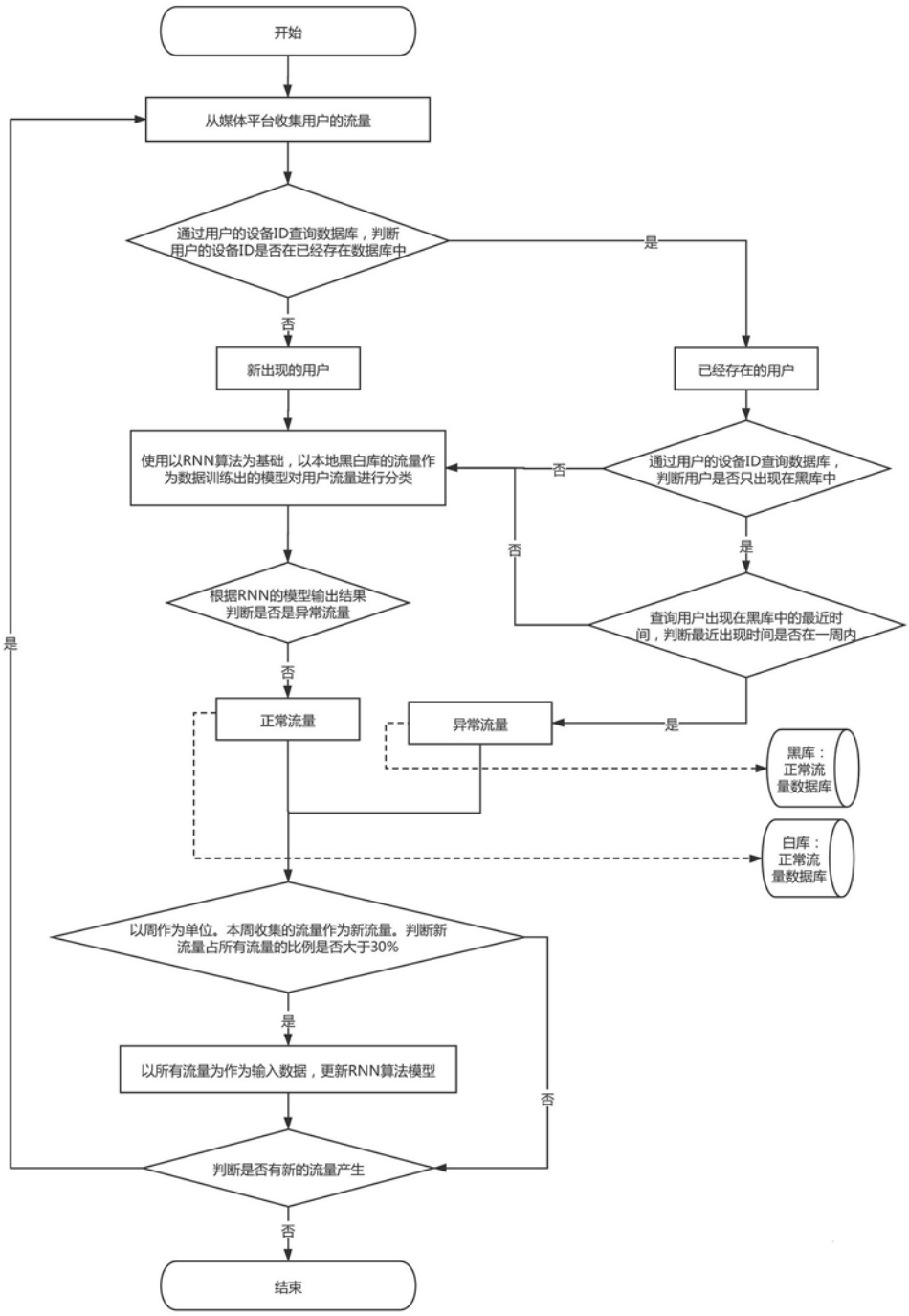

开始
从媒体平台收集用户的流量
通过用户的设备ID查询数据库，判断用户的设备ID是否在已经存在数据库中
是
否
新出现的用户
已经存在的用户
使用以RNN算法为基础，以本地黑白库的流量作为数据训练出的模型对用户流量进行分类
否
通过用户的设备ID查询数据库，判断用户是否只出现在黑库中
否
是
根据RNN的模型输出结果判断是否是异常流量
查询用户出现在黑库中的最近时间，判断最近出现时间是否在一周内
否
是
正常流量
异常流量
是
黑库：正常流量数据库
白库：正常流量数据库
以周作为单位，本周收集的流量作为新流量。判断新流量占所有流量的比例是否大于30%
是
否
以所有流量为作为输入数据，更新RNN算法模型
判断是否有新的流量产生
是
否
结束

图2

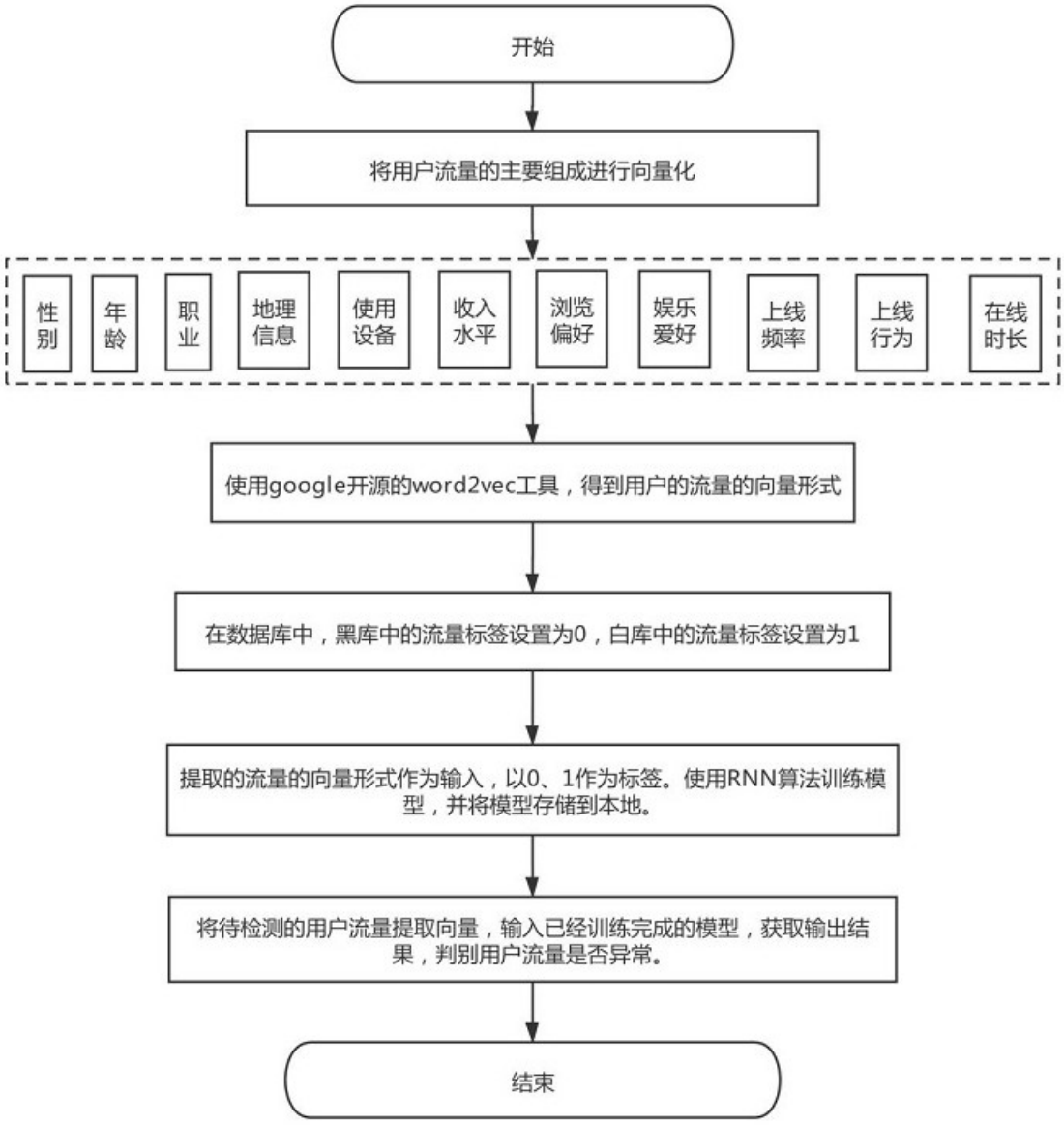

开始
将用户流量的主要组成进行向量化
性别
年龄
职业
地理信息
使用设备
收入水平
浏览偏好
娱乐爱好
上线频率
上线行为
在线时长
使用google开源的word2vec工具，得到用户的流量的向量形式
在数据库中，黑库中的流量标签设置为0，白库中的流量标签设置为1
提取的流量的向量形式作为输入，以0、1作为标签。使用RNN算法训练模型，并将模型存储到本地。
将待检测的用户流量提取向量，输入已经训练完成的模型，获取输出结果，判别用户流量是否异常。
结束

图3