(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 105429950 A
(43)申请公布日 2016.03.23

(21)申请号 201510725503.X

(22)申请日 2015.10.29

(71)申请人 国家计算机网络与信息安全管理中心
地址 100029 北京市朝阳区裕民路甲3号

(72)发明人 王啸 王大伟 贺龙涛 曹首峰 刘培朋 赵咏 苟高鹏

(74)专利代理机构 北京安博达知识产权代理有限公司 11271
代理人 徐国文

(51)Int.Cl.
H04L 29/06(2006.01)

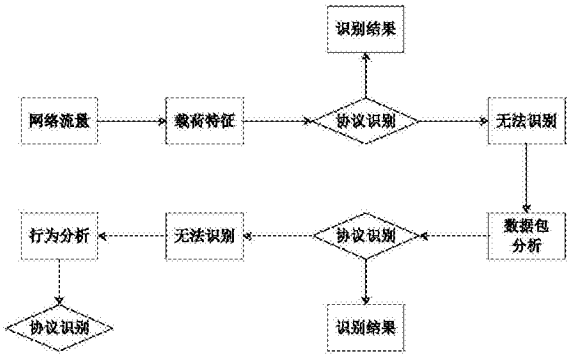
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种基于动态数据包采样的网络流量识别系统和方法

(57)摘要

本发明提供一种基于动态数据包采样的网络流量识别系统和方法，系统包括网络流量识别服务器、数据包分析模块和行为分析模块；网络流量识别服务器、数据包分析模块和行为分析模块依次单向连接。本发明解决了传统的网络流量识别方法面对不断变化的流量环境无法及时调整识别策略的矛盾，使得在网络流量识别的过程中，可以通过感知数据包的变化，来调整当前网络流量识别的策略，是根据首包信息结合协议识别方法来进行识别，还是根据数据包分析结合协议识别方法来进行识别，还是根据网络行为分析结合协议识别方法来进行识别，根据运行环境变化并自动选择适合当前流量特征的网络流量协议识别策略，从而保证在任意流量环境下网络流量识别的准确率与处理效率。



CN 105429950 A

1. 一种基于动态数据包采样的网络流量识别系统，其特征在于：所述系统包括网络流量识别服务器、数据包分析模块和行为分析模块；所述网络流量识别服务器、数据包分析模块和行为分析模块依次单向连接。

2. 根据权利要求 1 所述的基于动态数据包采样的网络流量识别系统，其特征在于：所述网络流量识别服务器获取网络流量，并从获取的网络流量中提取数据包首包的载荷特征，网络流量识别服务器根据提取的载荷特征识别网络流量；如果网络流量识别服务器能够识别网络流量，则不再进行网络流量的获取，否则采用数据包分析模块进行网络流量的识别。

3. 根据权利要求 2 所述的基于动态数据包采样的网络流量识别系统，其特征在于：所述网络流量识别服务器根据提取的载荷特征识别网络流量，包括：

所述网络流量识别服务器比较数据包首包载荷特征和网络流量识别服务器中网络流量行为特征之间的相似度，最相似的网络流量类型即为网络流量识别服务器识别出的网络流量类型。

4. 根据权利要求 2 所述的基于动态数据包采样的网络流量识别系统，其特征在于：所述数据包分析模块对首包之后的数据包进行均匀采样，根据采样的数据包的载荷特征识别网络流量，如果数据包采样数超过采样阈值还未识别出网络流量，则采用行为分析模块进行网络流量的识别。

5. 根据权利要求 4 所述的基于动态数据包采样的网络流量识别系统，其特征在于：所述数据包分析模块根据采样的数据包的载荷特征识别网络流量，包括：

所述数据包分析模块首先比较采样的第一个数据包与分类器中网络流量，确定第一个数据包载荷特征与网络流量行为特征之间的相似度，然后对确定的相似度进行归一化处理；所述数据包分析模块比较采样的第二个数据包与分类器中网络流量，确定第二个数据包载荷特征与网络流量行为特征之间的相似度，对确定的相似度进行归一化处理；经过归一化处理后的第一个数据包载荷特征与网络流量行为特征之间的相似度和第二个数据包载荷特征与网络流量行为特征之间的相似度相乘，之后再次归一化处理，依次进行同样的操作，直到所有数据包采样运行结束；最后，如果当前均匀采样的数据包载荷特征与网络流量行为特征之间的相似度大于 90%，则认为当前网络流量的类型为数据包分析模块识别出的网络流量类型。

6. 根据权利要求 4 所述的基于动态数据包采样的网络流量识别系统，其特征在于：所述行为分析模块对随后的数据包采用随机递增的间隔抽样策略进行随机抽样，提取该网络流量的行为特征，并且将网络流量的行为特征与数据包的载荷特征相融合，进行网络流量的识别。

7. 根据权利要求 6 所述的基于动态数据包采样的网络流量识别系统，其特征在于：所述行为分析模块识别网络流量包括：

将数据包首包载荷特征与网络流量行为特征之间的相似度和行为分析模块采样得到的数据包载荷特征与网络流量行为特征之间的相似度进行累加，累加之后的相似度大于 90%，则认为当前网络流量的类型为行为分析模块识别出的网络流量类型。

8. 一种基于动态数据包采样的网络流量识别方法，其特征在于：所述方法包括以下步骤：

步骤 1 :通过网络流量识别服务器获取网络流量,从获取的网络流量中提取数据包首包的载荷特征,如果网络流量识别服务器能够识别网络流量,则不再进行网络流量的获取,否则执行步骤 2 ;

步骤 2 :通过数据包分析模块对首包之后的数据包进行均匀采样,均匀采样的参数包括采样数据包间隔、总采样数据包数 m 和采样数据包范围,且在第 2 个数据包和第 k 个数据包之间进行均匀采样, $k \leqslant m$;根据采样的数据包的载荷特征识别网络流量,当其中某个数据包已经识别出网络流量,则终止首包之后数据包的均匀采样,如果数据包采样数超过采样阈值还未识别出网络流量,则执行步骤 3 ;

步骤 3 :通过行为分析模块对第 k 个数据包之后的数据包进行随机抽样,采用随机递增的间隔抽样策略完成随机抽样,提取该网络流量的行为特征,并且将网络流量的行为特征与数据包的载荷特征相融合,进行网络流量的识别。

一种基于动态数据包采样的网络流量识别系统和方法

技术领域

[0001] 本发明属于信息安全技术领域，具体涉及一种基于动态数据包采样的网络流量识别系统和方法。

背景技术

[0002] 随着信息技术特别是互联网技术的快速发展，网络应用的数量也在快速的增长。网络应用的发展给人们的生活带来了极大的方便，但是网络应用的复杂性和多样性也给网络应用管理、流量控制等带来巨大的挑战。为了有效的应对网络应用快速发展所带来的挑战，实时、准确的网络应用识别研究成为当前网络管理研究领域的重要研究问题之一。

[0003] 目前存在的协议识别技术主要存在如下几种：(1) 深度报文检测技术；(2) 多模式匹配方法；(3) 正则表达式匹配方法。

[0004] 深度报文检测 (Deep Packet Inspection 简称 DPI) 技术主要相对传统的基于五元组信息浅层报文检测技术而言，基于 DPI 的协议识别技术将检测深入到应用层负载内容，通过匹配数据包负载内容是否包含协议的特征对流量进行识别，深度包检测技术能够识别 http 伪装、端口协商和随机端口下载的 P2P 流量，具有较好的健壮性。

[0005] 多模式匹配算法是经典的多模匹配算法。该算法的主要思路是对特征串集合进行预处理，通过算法寻找特征串之间的内部关联关系，当匹配失效时通过对后缀包含进行处理，直接对下一个待匹配字符进行匹配而不需要在特征串中进行回溯。该算法的核心包括三张表：goto 表、failure 表和 output 表。

[0006] 正则表达式是正则语言的一种描述模型，在用正则表达式进行匹配的算法当中，普遍采用将正则表达式转换为有穷自动机 (FA) 的方式。有穷自动机是指一种进行文法识别的逻辑结构，其结构可以采用编程方式实现，与正则表达式作为主要的正则文法描述方式不同，有穷状态机主要用于的正则文法识别和匹配领域，有穷状态自动机又分为确定有穷状态，因此基于正则表达式进行协议识别的方法可以分为基于 NFA 正则表达式匹配算法和基于 DFA 正则表达式匹配算法。

[0007] 在目前现有的协议识别方法中，深度报文检测虽然有较好的准确性和健壮性，但是其识别速度太慢，无法满足流量协议进行实时识别的需求，标准的多模式匹配算法虽然具有较高的匹配速度和效率，但是只能对字符串形式的协议特征进行匹配，无法应用于正则表达式协议识别领域，当前普遍使用正则表达式进行协议特征描述，主要采用基于正则表达式匹配的协议识别方法，而采用 NFA 方式对正则表达式进行识别时间开销较高，无法满足需求；使用 DFA 识别方式会面临状态图爆炸问题，因此需要对算法进行改进；而目前基于状态图进行优化方式很难满足需求，因此需要结合协议识别中协议特征的特性，对正则表达式匹配算法进行更深入的研究。

[0008] 结合协议特征的识别方法有如下方法：(1) 基于端口的协议识别技术；(2) 基于应用层负载签名特征的协议识别技术；(3) 基于流特征的协议识别技术几个阶段。

[0009] 基于端口的协议识别技术根据常见的网络应用或者网络流量使用的固定端口号

来识别网络应用或者协议，但是这种技术无法应对越来越多的采用动态端口的网络应用。

[0010] 基于应用层负载签名的协议识别技术通过识别网络应用的应用层负载签名来识别网络应用或者协议，这种方法克服了动态端口技术给协议识别带来的困难，但是对部分数据流加密的网络应用或者协议仍然无法有效识别。

[0011] 基于流特征的协议识别技术根据网络数据流中的数据包长短、连接比等特点识别网络应用或者协议，但是这种方法准确度没有基于应用层负载签名的协议识别方法高并且开销较大。

发明内容

[0012] 为解决现有协议识别无法适应实际网络环境中不断变化且不可预测的网络流量的问题，本发明提供一种基于动态数据包采样的网络流量识别系统和方法，可以感知网络流量特征变化并自动选择适应数据包采样，实现网络流量的识别。

[0013] 为了实现上述发明目的，本发明采取如下技术方案：

[0014] 本发明提供一种基于动态数据包采样的网络流量识别系统，所述系统包括网络流量识别服务器、数据包分析模块和行为分析模块；所述网络流量识别服务器、数据包分析模块和行为分析模块依次单向连接。

[0015] 所述网络流量识别服务器获取网络流量，并从获取的网络流量中提取数据包首包的载荷特征，网络流量识别服务器根据提取的载荷特征识别网络流量；如果网络流量识别服务器能够识别网络流量，则不再进行网络流量的获取，否则采用数据包分析模块进行网络流量的识别。

[0016] 所述网络流量识别服务器根据提取的载荷特征识别网络流量，包括：

[0017] 所述网络流量识别服务器比较数据包首包载荷特征和网络流量识别服务器中网络流量行为特征之间的相似度，最相似的网络流量类型即为网络流量识别服务器识别出的网络流量类型。

[0018] 所述数据包分析模块对首包之后的数据包进行均匀采样，根据采样的数据包的载荷特征识别网络流量，如果数据包采样数超过采样阈值还未识别出网络流量，则采用行为分析模块进行网络流量的识别。

[0019] 所述数据包分析模块根据采样的数据包的载荷特征识别网络流量，包括：

[0020] 所述数据包分析模块首先比较采样的第一个数据包与分类器中网络流量，确定第一个数据包载荷特征与网络流量行为特征之间的相似度，然后对确定的相似度进行归一化处理；所述数据包分析模块比较采样的第二个数据包与分类器中网络流量，确定第二个数据包载荷特征与网络流量行为特征之间的相似度，对确定的相似度进行归一化处理；经过归一化处理后的第一个数据包载荷特征与网络流量行为特征之间的相似度和第二个数据包载荷特征与网络流量行为特征之间的相似度相乘，之后再次归一化处理，依次进行同样的操作，直到所有数据包采样运行结束；最后，如果当前均匀采样的数据包载荷特征与网络流量行为特征之间的相似度大于90％，则认为当前网络流量的类型为数据包分析模块识别出的网络流量类型。

[0021] 所述行为分析模块对随后的数据包采用随机递增的间隔抽样策略进行随机抽样，提取该网络流量的行为特征，并且将网络流量的行为特征与数据包的载荷特征相融合，进

行网络流量的识别。

[0022] 所述行为分析模块识别网络流量包括：

[0023] 将数据包首包载荷特征与网络流量行为特征之间的相似度和行为分析模块采样得到的数据包载荷特征与网络流量行为特征之间的相似度进行累加，累加之后的相似度大于90％，则认为当前网络流量的类型为行为分析模块识别出的网络流量类型。

[0024] 本发明提供一种基于动态数据包采样的网络流量识别方法，所述方法包括以下步骤：

[0025] 步骤1：通过网络流量识别服务器获取网络流量，从获取的网络流量中提取数据包首包的载荷特征，如果网络流量识别服务器能够识别网络流量，则不再进行网络流量的获取，否则执行步骤2；

[0026] 步骤2：通过数据包分析模块对首包之后的数据包进行均匀采样，均匀采样的参数包括采样数据包间隔、总采样数据包数m和采样数据包范围，且在第2个数据包和第k个数据包之间进行均匀采样，$k \leqslant m$；根据采样的数据包的载荷特征识别网络流量，当其中某个数据包已经识别出网络流量，则终止首包之后数据包的均匀采样，如果数据包采样数超过采样阈值还未识别出网络流量，则执行步骤3；

[0027] 步骤3：通过行为分析模块对第k个数据包之后的数据包进行随机抽样，采用随机递增的间隔抽样策略完成随机抽样，提取该网络流量的行为特征，并且将网络流量的行为特征与数据包的载荷特征相融合，进行网络流量的识别。

[0028] 与现有技术相比，本发明的有益效果在于：

[0029] 本发明利用对数据包的不同采样策略，解决传统的网络流量识别方法面对不断变化的流量环境无法及时调整识别策略的矛盾，使得在网络流量识别的过程中，可以通过感知数据包的变化，来调整当前网络流量识别的策略，是根据首包信息结合协议识别方法来进行识别，还是根据数据包分析结合协议识别方法来进行识别，还是根据网络行为分析结合协议识别方法来进行识别，根据运行环境变化并自动选择适合当前流量特征的网络流量协议识别策略，从而保证在任意流量环境下网络流量识别的准确率与处理效率。

附图说明

[0030] 图1是本发明实施例中基于动态数据包采样的网络流量识别方法流程图；

[0031] 图2是本发明实施例中网络流量识别服务器工作流程图；

[0032] 图3是本发明实施例中数据包分析模块工作流程图；

[0033] 图4是本发明实施例中行为分析模块工作流程图。

具体实施方式

[0034] 下面结合附图对本发明作进一步详细说明。

[0035] 本发明提供一种基于动态数据包采样的网络流量识别系统，所述系统包括网络流量识别服务器、数据包分析模块和行为分析模块；所述网络流量识别服务器、数据包分析模块和行为分析模块依次单向连接。

[0036] 所述网络流量识别服务器获取网络流量，并从获取的网络流量中提取数据包首包的载荷特征，网络流量识别服务器根据提取的载荷特征识别网络流量；如果网络流量识别

服务器能够识别网络流量，则不再进行网络流量的获取，否则采用数据包分析模块进行网络流量的识别。

[0037] 所述网络流量识别服务器根据提取的载荷特征识别网络流量，包括：

[0038] 所述网络流量识别服务器比较数据包首包载荷特征和网络流量识别服务器中网络流量行为特征之间的相似度，最相似的网络流量类型即为网络流量识别服务器识别出的网络流量类型。

[0039] 所述数据包分析模块对首包之后的数据包进行均匀采样，根据采样的数据包的载荷特征识别网络流量，如果数据包采样数超过采样阈值还未识别出网络流量，则采用行为分析模块进行网络流量的识别。

[0040] 所述数据包分析模块根据采样的数据包的载荷特征识别网络流量，包括：

[0041] 所述数据包分析模块首先比较采样的第一个数据包与分类器中网络流量，确定第一个数据包载荷特征与网络流量行为特征之间的相似度，然后对确定的相似度进行归一化处理；所述数据包分析模块比较采样的第二个数据包与分类器中网络流量，确定第二个数据包载荷特征与网络流量行为特征之间的相似度，对确定的相似度进行归一化处理；经过归一化处理后的第一个数据包载荷特征与网络流量行为特征之间的相似度和第二个数据包载荷特征与网络流量行为特征之间的相似度相乘，之后再次归一化处理，依次进行同样的操作，直到所有数据包采样运行结束；最后，如果当前均匀采样的数据包载荷特征与网络流量行为特征之间的相似度大于90％，则认为当前网络流量的类型为数据包分析模块识别出的网络流量类型。

[0042] 所述行为分析模块对随后的数据包采用随机递增的间隔抽样策略进行随机抽样，提取该网络流量的行为特征，并且将网络流量的行为特征与数据包的载荷特征相融合，进行网络流量的识别。

[0043] 所述行为分析模块识别网络流量包括：

[0044] 将数据包首包载荷特征与网络流量行为特征之间的相似度和行为分析模块采样得到的数据包载荷特征与网络流量行为特征之间的相似度进行累加，累加之后的相似度大于90％，则认为当前网络流量的类型为行为分析模块识别出的网络流量类型。

[0045] 本发明提供一种基于动态数据包采样的网络流量识别方法，所述方法包括以下步骤：

[0046] 步骤1：通过网络流量识别服务器获取网络流量，从获取的网络流量中提取数据包首包的载荷特征，如果网络流量识别服务器能够识别网络流量，则不再进行网络流量的获取，否则执行步骤2；

[0047] 步骤2：通过数据包分析模块对首包之后的数据包进行均匀采样，均匀采样的参数包括采样数据包间隔、总采样数据包数m和采样数据包范围，且在第2个数据包和第k个数据包之间进行均匀采样，$k \leqslant m$；根据采样的数据包的载荷特征识别网络流量，当其中某个数据包已经识别出网络流量，则终止首包之后数据包的均匀采样，如果数据包采样数超过采样阈值还未识别出网络流量，则执行步骤3；

[0048] 步骤3：通过行为分析模块对第k个数据包之后的数据包进行随机抽样，采用随机递增的间隔抽样策略完成随机抽样，提取该网络流量的行为特征，并且将网络流量的行为特征与数据包的载荷特征相融合，进行网络流量的识别。

[0049] 步骤 1 中，所述网络流量识别服务器根据提取的载荷特征识别网络流量，包括：

[0050] 所述网络流量识别服务器比较数据包首包载荷特征和网络流量识别服务器中网络流量行为特征之间的相似度，最相似的网络流量类型即为网络流量识别服务器识别出的网络流量类型。

[0051] 步骤 2 中，所述数据包分析模块根据采样的数据包的载荷特征识别网络流量，包括：

[0052] 所述数据包分析模块首先比较采样的第一个数据包与分类器中网络流量，确定第一个数据包载荷特征与网络流量行为特征之间的相似度，然后对确定的相似度进行归一化处理；所述数据包分析模块比较采样的第二个数据包与分类器中网络流量，确定第二个数据包载荷特征与网络流量行为特征之间的相似度，对确定的相似度进行归一化处理；经过归一化处理后的第一个数据包载荷特征与网络流量行为特征之间的相似度和第二个数据包载荷特征与网络流量行为特征之间的相似度相乘，之后再次归一化处理，依次进行同样的操作，直到所有数据包采样运行结束；最后，如果当前均匀采样的数据包载荷特征与网络流量行为特征之间的相似度大于 90％，则认为当前网络流量的类型为数据包分析模块识别出的网络流量类型。

[0053] 步骤 3 中，所述行为分析模块识别网络流量包括：

[0054] 将数据包首包载荷特征与网络流量行为特征之间的相似度和行为分析模块采样得到的数据包载荷特征与网络流量行为特征之间的相似度进行累加，累加之后的相似度大于 90％，则认为当前网络流量的类型为行为分析模块识别出的网络流量类型。

[0055] 最后应当说明的是：以上实施例仅用以说明本发明的技术方案而非对其限制，所属领域的普通技术人员参照上述实施例依然可以对本发明的具体实施方式进行修改或者等同替换，这些未脱离本发明精神和范围的任何修改或者等同替换，均在申请待批的本发明的权利要求保护范围之内。

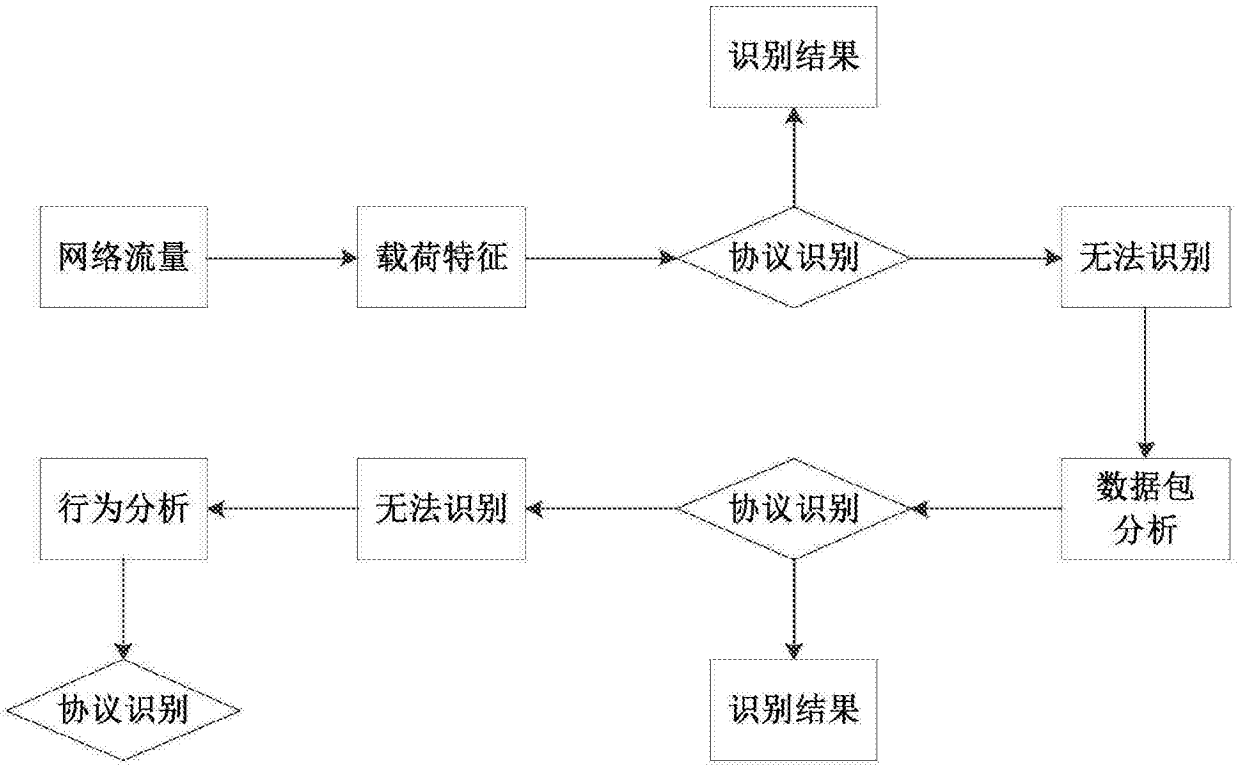

识别结果
网络流量
载荷特征
协议识别
无法识别
行为分析
无法识别
协议识别
数据包分析
协议识别
识别结果

图 1

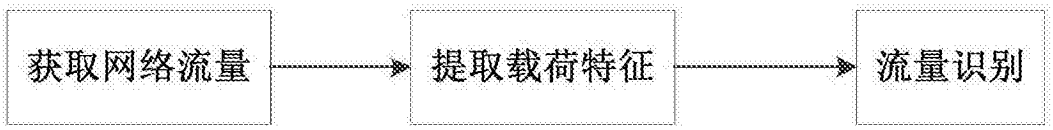

获取网络流量
提取载荷特征
流量识别

图 2

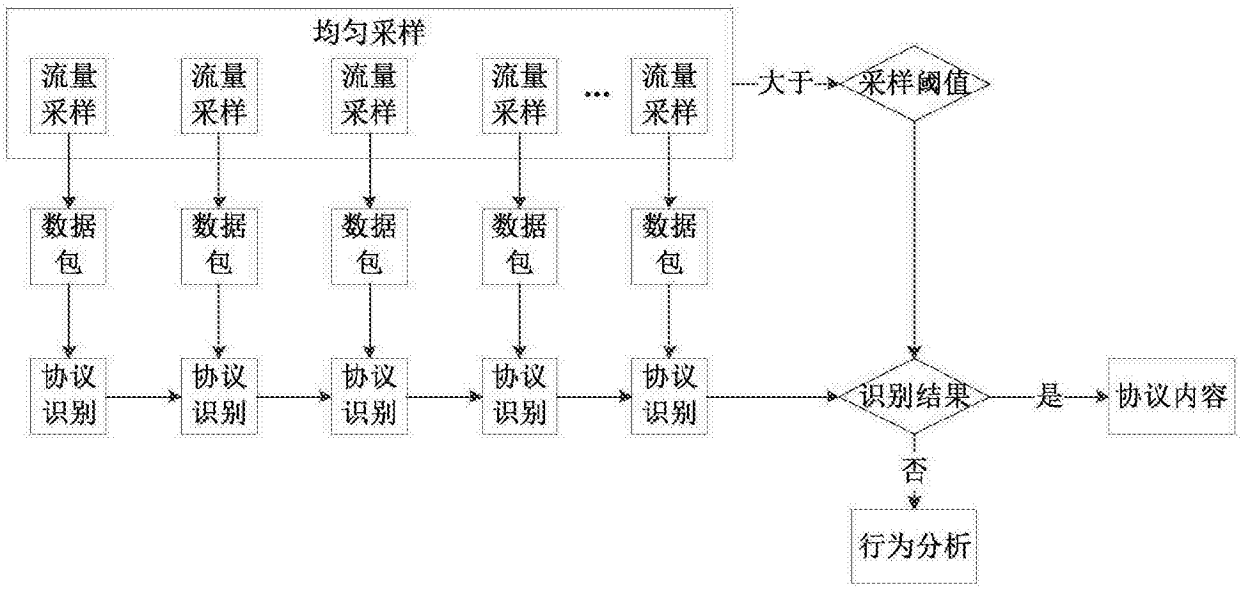

均匀采样
流量采样
流量采样
流量采样
流量采样
...
流量采样
大于
采样阈值
数据包
数据包
数据包
数据包
数据包
协议识别
协议识别
协议识别
协议识别
协议识别
识别结果
是
协议内容
否
行为分析

图 3

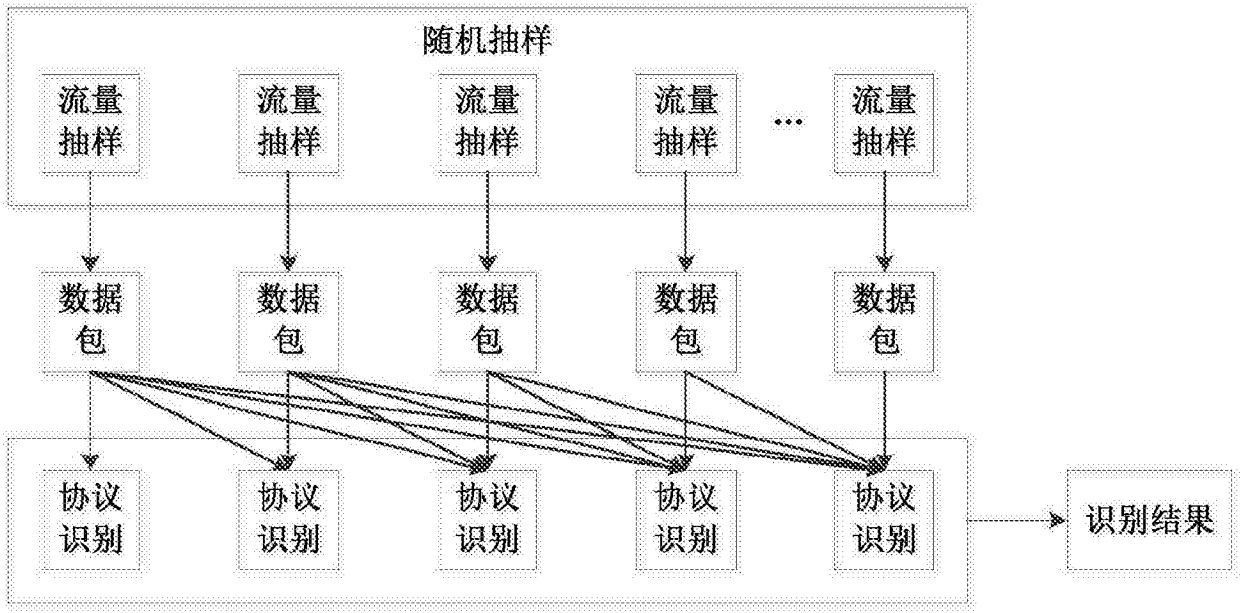

随机抽样
流量抽样
流量抽样
流量抽样
流量抽样
...
流量抽样
数据包
数据包
数据包
数据包
数据包
协议识别
协议识别
协议识别
协议识别
协议识别
识别结果

图 4