(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 101714952 A
(43)申请公布日 2010.05.26

(21)申请号 200910259394.1

(22)申请日 2009.12.22

(71)申请人 北京邮电大学
地址 100876 北京市海淀区西土城路 10 号

(72)发明人 寿国础 张剑 胡怡红 郭志刚
钱宗珏 宁帆

(74)专利代理机构 北京市隆安律师事务所
11323
代理人 权鲜枝

(51)Int.Cl.
H04L 12/56(2006.01)
H04L 12/26(2006.01)
H04L 29/08(2006.01)

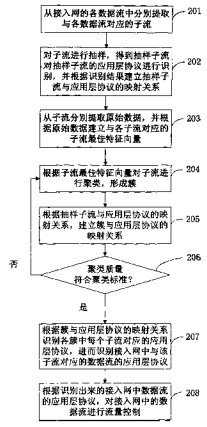
权利要求书 5 页 说明书 14 页 附图 3 页

(54)发明名称

一种接入网的流量识别方法和装置

(57)摘要

本发明公开了一种接入网的流量识别方法和装置。该方法包括：从接入网的每个数据流提取与其对应的子流；对所有子流进行抽样，得到抽样子流，对每个抽样子流的应用层协议进行识别，根据识别结果建立抽样子流与应用层协议的映射关系；从每个子流提取原始数据，并根据原始数据建立与该子流对应的子流特征向量；根据所有子流特征向量将子流进行聚类，形成多个簇；根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系；根据簇与应用层协议的映射关系，识别各簇中的每个子流对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议。本发明能够对接入网进行高速在线的流量识别，并能识别接入网中加密数据流的应用层协议。



CN 101714952 A

1. 一种接入网的流量识别方法，其特征在于，该方法包括：

从接入网的每个数据流中提取与该数据流对应的子流；

对所有子流进行抽样，得到抽样子流，对每个抽样子流的应用层协议进行识别，并根据识别结果建立抽样子流与应用层协议的映射关系；

从每个子流提取原始数据，并根据原始数据建立与该子流对应的子流特征向量；

根据所有子流特征向量将对应的子流进行聚类，形成多个簇；

根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系；其中，如果一个簇不能与任何已知应用层协议建立映射关系，则该簇对应的应用层协议类型为未知；

根据簇与应用层协议的映射关系，识别各簇中的每个子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议；其中，对于与未知应用层协议类型相对应的簇，该簇中每个子流所对应的应用层协议类型为未知，进而接入网中与该子流对应的数据流的应用层协议类型也为未知。

2. 根据权利要求 1 所述的方法，其特征在于，在所述进而识别接入网中与该子流对应的数据流的应用层协议之后，该方法进一步包括：

根据识别出来的接入网中数据流的应用层协议，对接入网中的数据流进行流量控制；其中，对于接入网中与未知应用层协议类型对应的数据流，首先检测其对应的子流所在的簇是否为异常数据流，如果是异常数据流，则对接入网中与该未知应用层协议类型对应的数据流进行接入控制，如果不是异常数据流，则将该未知应用层协议类型定义为新的应用层协议，将接入网中与该未知应用层协议类型对应的数据流与所述新的应用层协议建立映射关系，然后对接入网中与所述新的应用层协议建立映射关系的数据流进行流量控制。

3. 根据权利要求 1 所述的方法，其特征在于，所述从接入网的每个数据流中提取与该数据流对应的子流的步骤包括：

对于接入网中的每一个数据流，提取该数据流起始位置的预设数量的数据包作为该数据流的子流。

4. 根据权利要求 1 所述的方法，其特征在于，所述对每个抽样子流的应用层协议进行识别的步骤包括：

用深度包检测技术对每个抽样子流进行检测，从而识别出抽样子流所对应的应用层协议，并将不能与已知应用层协议相对应的抽样子流与未知应用层协议相对应。

5. 根据权利要求 1 所述的方法，其特征在于，从每个子流提取原始数据，并根据原始数据建立与该子流对应的子流特征向量的步骤包括：

从子流中提取数据包的特征数据作为原始数据，所述特征数据包括：协议字段、子流的持续时间、数据包的长度、数据包中有效载荷的大小、相邻数据包到达接入点的时间间隔中的任意一个或多个；

根据所述原始数据，用统计学的方法计算子流的统计特征，所述统计特征包括：最大前向数据包的长度、数据包包头中协议的类型、相邻前向数据包到达接入点的最小时间间隔、相邻前向数据包到达接入点的时间间隔的标准差、相邻后向数据包到达接入点的时间间隔的标准差、最大后向数据包的长度以及后向数据包长度的均值中的任意一个或多个；

将所述统计特征排列成一个向量，得到子流特征向量。

6. 根据权利要求 1 或 4 所述的方法，其特征在于，所述根据所有子流特征向量将对应的

子流进行聚类，形成多个簇的步骤包括以下步骤：

第一步，将所有子流特征向量的状态都设置为未归类，并为每一个子流设置邻近特征向量距离 eps 和最小特征向量数目 minpts 这两个参数；

第二步，选定任一未归类的子流特征向量 p，则 p 具有邻近特征向量距离 p.eps 和最小特征向量数目 p.minpts 两个参数，计算 p 与其他所有未归类的子流特征向量的距离；

将参数 p.eps 的大小与 p 和其他子流特征向量之间的距离作比较，并统计与 p 的距离小于参数 p.eps 的子流特征向量的数目，然后将该数目与参数 p.minpts 作比较，如果该数目不小于参数 p.minpts，那么称 p 为核对象，与 p 的距离小于 p.eps 的所有子流特征向量称为 p 的直接密度可达对象，将 p 与所有 p 的直接密度可达对象组成以 p 为核对象的簇 a，然后将 p 和所有 p 的直接密度可达对象以及其所对应的各子流的状态都设置为属于簇 a；如果该数目小于 p.minpts，那么称 p 为噪声对象，并将 p 以及其对应的子流的状态设置为未知；

第三步，判断是否存在未归类的子流特征向量，如果存在，则重复执行第二步，如果不存在，则结束流程；

其中，所述 p 与其他任一子流特征向量的距离，为 p 与其他任一子流特征向量的欧氏距离。

7. 根据权利要求 6 所述的方法，其特征在于，

所述根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系的步骤包括：根据抽样子流与应用层协议的映射关系，统计簇内各抽样子流所对应的应用层协议；如果一个簇内各抽样子流所对应的应用层协议至少包括一个已知应用层协议，则计算簇内各抽样子流所对应的已知应用层协议的概率，然后将概率最大的已知应用层协议与该簇建立映射；

所述如果一个簇不能与任何应用层协议建立映射关系，则该簇对应的应用层协议类型为未知的步骤包括：根据抽样子流与应用层协议的映射关系，统计簇内各抽样子流所对应的应用层协议；如果一个簇内各抽样子流所对应的应用层协议均为未知应用层协议，那么该簇不能与任何已知应用层协议建立映射关系，则将该簇与未知应用层协议类型相对应。

8. 根据权利要求 7 所述的方法，其特征在于，

在根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系之后，并且在根据簇与应用层协议的映射关系，识别各簇中的每个子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议之前，该方法进一步包括：

评判聚类质量是否符合聚类标准，如果聚类质量符合聚类标准，则执行所述根据簇与应用层协议的映射关系，识别各簇中子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议；否则，重新为每一个子流设置 eps 和 minpts 这两个参数，然后重新根据子流特征向量对子流进行聚类，形成簇，并重新根据抽样子流与应用层协议的映射关系建立簇与应用层协议的映射关系，直至评判结果为聚类质量符合聚类标准；

所述聚类质量符合聚类标准，是指聚类质量同时达到三个标准，如果不能满足其中任一标准，都为聚类质量不符合聚类标准，所述三个标准为：

第一：状态为未知的子流特征向量的数目占所有子流特征向量数目的比例小于 10%；

第二：簇的数目为抽样子流与应用层协议的映射关系中的已知应用层协议的数目的两

倍以上；

第三：根据抽样子流与应用层协议的映射关系建立簇与应用层协议或未知应用层协议类型的映射关系之后，与应用层协议映射的任一簇中，概率最大的应用层协议的概率不低于 60％。

9. 一种接入网的流量识别装置，其特征在于，该装置包括：

应用层协议模块：用于存储应用层协议；向抽样子流生成和应用识别模块提供应用层协议查询服务；

子流特征生成模块：用于从接入网的各数据流中分别提取与所述各数据流对应的子流；从所述子流分别提取原始数据，并根据原始数据建立与各子流对应的子流特征向量；

子流聚类模块：用于从子流特征生成模块接收子流及其子流特征向量；根据子流特征向量对子流进行聚类，形成簇；向簇的应用识别模块发送簇；

抽样子流生成和应用识别模块：用于从子流特征生成模块接收子流；对子流进行抽样，得到抽样子流；向应用层协议模块查询应用层协议；对抽样子流的应用层协议进行识别，并根据识别结果建立抽样子流与应用层协议的映射关系；

簇的应用识别模块：用于从抽样子流生成和应用识别模块获取抽样子流与应用层协议的映射关系；从子流聚类模块接收簇；根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系；

业务识别模块：用于从簇的应用识别模块获取簇与应用层协议的映射关系；根据簇与应用层协议的映射关系，识别各簇中子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议。

10. 根据权利要求 9 所述的装置，其特征在于，该装置进一步包括：

异常流量检测模块：用于从业务识别模块接收与未知应用层协议类型映射的簇及其映射的未知应用层协议类型；判断簇是否为异常数据流；将异常数据流的簇发送到异常流量控制模块；将不是异常数据流的簇所映射的未知应用层协议类型定义为新的应用层协议，发送到应用层协议模块；将不是异常数据流的簇与所述新的应用层协议建立映射；将不是异常数据流的簇及其映射的新的应用层协议发送到转发策略模块；

异常流量控制模块：用于从异常流量检测模块接收异常数据流的簇；向接入网发送子流接入控制信息；

转发策略模块：用于从业务识别模块接收与已知应用层协议映射的簇及其映射的已知应用层协议；从异常流量检测模块接收不是异常数据流的簇及其映射的新的应用层协议；将簇内所有子流都与该簇映射的应用层协议建立映射；根据簇映射的应用层协议的类型，向接入网发送子流流量控制信息。

11. 根据权利要求 9 所述的装置，其特征在于，

子流特征生成模块，用于从接入网传输的每一个数据流中提取该数据流起始位置的预设数量的数据包作为该数据流的子流。

12. 根据权利要求 9 所述的装置，其特征在于，

抽样子流生成和应用识别模块，用深度包检测技术对抽样子流进行检测，从而识别出抽样子流所对应的应用层协议，并将不能与已知应用层协议相对应的抽样子流与未知应用层协议相对应，从而建立抽样子流与应用层协议的映射关系。

13. 根据权利要求 9 所述的装置，其特征在于，

所述子流特征生成模块，首先从子流中提取数据包的特征数据作为原始数据，所述特征数据包括：协议字段、子流的持续时间、数据包的长度、数据包中有效载荷的大小、相邻数据包到达接入点的时间间隔中的任意一个或多个；然后，根据所述原始数据，用统计学的方法计算子流的统计特征，所述统计特征包括：最大前向数据包的长度、数据包包头中协议的类型、相邻前向数据包到达接入点的最小时间间隔、相邻前向数据包到达接入点的时间间隔的标准差、相邻后向数据包到达接入点的时间间隔的标准差、最大后向数据包的长度以及后向数据包长度的均值中的任意一个或多个；最后，将所述统计特征排列成一个向量，得到子流特征向量。

14. 根据权利要求 9 或 12 所述的装置，其特征在于，所述子流聚类模块根据子流特征向量对子流进行聚类、形成簇包括：

第一步，所述子流聚类模块将所有子流特征向量的状态都设置为未归类，并为每一个子流设置邻近特征向量距离 eps 和最小特征向量数目 minpts 这两个参数；

第二步，所述子流聚类模块选定任一未归类的子流特征向量 p，则 p 具有邻近特征向量距离 p.ep s 和最小特征向量数目 p.minpts 两个参数，计算 p 与其他所有未归类的子流特征向量的距离；

所述子流聚类模块将参数 p.eps 的大小与 p 和其他子流特征向量之间的距离作比较，并统计与 p 的距离小于参数 p.eps 的子流特征向量的数目，然后将该数目与参数 p.minpts 作比较，如果该数目不小于参数 p.minpts，那么称 p 为核对象，与 p 的距离小于 p.eps 的所有任一子流特征向量称为 p 的直接密度可达对象，所述子流聚类模块将 p 与所有 p 的直接密度可达对象组成以 p 为核对象的簇 a，然后将 p 和所有 p 的直接密度可达对象以及其所对应的各子流的状态都设置为属于簇 a；如果与 p 的距离小于 p.eps 的子流特征向量的数目小于 p.minpts，那么称 p 为噪声对象，并将 p 以及其对应的子流的状态设置为未知；

第三步，所述子流聚类模块判断是否存在未归类的子流特征向量，如果存在，则重复执行第二步，如果不存在，结束流程；

其中，所述 p 与其他任一子流特征向量的距离，为 p 与其他任一子流特征向量的欧氏距离。

15. 根据权利要求 14 所述的装置，其特征在于，

所述簇的应用识别模块，用于根据抽样子流与应用层协议的映射关系，统计簇内各抽样子流所对应的应用层协议，如果一个簇内各抽样子流所对应的应用层协议至少包括一个已知应用层协议，则所述簇的应用识别模块计算簇内各抽样子流所对应的已知应用层协议的概率，然后将概率最大的已知应用层协议与该簇建立映射；其中，如果一个簇不能与任何已知应用层协议建立映射关系，则所述簇的应用识别模块将该簇与未知应用层协议类型相对应。

16. 根据权利要求 15 所述的装置，其特征在于，该装置进一步包括：

聚类质量评判模块：用于从抽样子流生成和应用识别模块接收抽样子流与应用层协议的映射关系；从子流聚类模块接收簇、子流及其子流特征向量；从簇的应用识别模块接收簇与应用层协议的映射关系；

所述聚类质量评判模块，用于评判聚类质量是否符合聚类标准，如果评判结果为聚类

质量符合聚类标准，则向子流聚类模块发送聚类质量评判结果；

所述子流聚类模块，只有在聚类质量评判结果为聚类质量符合聚类标准时，向簇的应用识别模块发送簇；

所述子流聚类模块，在聚类质量评判结果为聚类质量不符合聚类标准时，重新为每一个子流设置 eps 和 minpts 这两个参数，然后重新根据子流特征向量对子流进行聚类，形成簇，并将子流及其子流特征向量以及新的簇发送给聚类质量评判模块重新进行聚类质量的评判，直至所述聚类质量评判模块评判聚类质量符合聚类标准为止；

所述聚类质量符合聚类标准，是指聚类质量同时达到三个标准，如果不能满足其中任一标准，都为聚类质量不符合聚类标准，所述三个标准为：

第一：状态为未知的子流特征向量的数目占所有子流特征向量数目的比例小于 10％；

第二：簇的数目为抽样子流与应用层协议的映射关系中的已知应用层协议的数目的两倍以上；

第三：所述簇的应用识别模块根据抽样子流与应用层协议的映射关系建立簇与应用层协议的映射关系之后，在与已知应用层协议映射的任一簇中，概率最大的已知应用层协议的概率不低于 60％。

一种接入网的流量识别方法和装置

技术领域

[0001] 本发明涉及网络通信技术领域，特别是涉及一种接入网的流量识别方法和装置。

背景技术

[0002] 随着网络业务类型的多样化，网络流量的识别技术成为业界关注的热点。接入网是连接核心网和用户终端，或连接核心网和用户驻地网的网络，及时准确地对接入网中不同类型的网络流量进行识别，对于流量工程、服务质量 (QoS) 以及网络安全管理都有重要的意义。

[0003] 在网络通信过程中，原始数据信息需要被分割成为许多小的数据包，以便能高速地通过网络，因此，接入网中传输的网络流量表现为数据包的形式。数据包分为包头和有效载荷两部分，包头是记录数据包附加信息的部分，如协议字段等；网络要传输的具体信息存在于数据包的有效载荷部分。

[0004] 现有技术采用深度包检测 (DPI，Deep Packet Inspection) 技术对接入网中传输的网络流量进行检测，其具体过程如下：用专门的通信管理设备将接入网中正在传输的数据包全部加以复制、存储，然后将存储起来的数据包重新组装成为原始数据，再将这些原始数据进行缓存；通信管理设备通过扮演特定的应用程序数据流代理，寻找已经定义的变量，再根据这些变量决定采取的动作，从而找到原始数据所对应的应用程序的类型或信源和信宿。当通信管理设备找到了数据包中有效载荷所携带的信息，它就会向能以最佳效果处理该数据包的应用程序发送数据包。另外，通信管理设备还用于检测已定义的变量的正确性，从而确定数据包是否为病毒或网络入侵等异常数据流，对于异常数据流，通信管理设备将向管理员发送警报。这种 DPI 方法能够对接入网中传输的全部数据包加以深入检测，因而具有较高的准确性。

[0005] 但是，上述的 DPI 实现过程中需要对接入网中传输的全部数据包进行复制、存储，再进行计算处理，其存储开销和计算量都很大，因而对网络流量的检测速度很慢，不能用于对接入网进行高速在线的流量识别。

[0006] 另外，在上述的 DPI 实现过程中，通信管理设备是通过扮演特定应用程序数据流代理的方式进行检测，而加密的数据包是无法用这种方式找到已定义的变量的，因此，无法对接入网中传输的加密数据包进行检测。

发明内容

[0007] 本发明提出了一种接入网的流量识别方法，能够对接入网进行高速在线的流量识别，并能识别接入网中加密数据流的应用层协议。

[0008] 本发明还提供了一种接入网的流量识别装置，能够对接入网进行高速在线的流量识别，并能识别接入网中加密数据流的应用层协议。

[0009] 为了达到上述目的，本发明的技术方案是这样实现的：

[0010] 一种接入网的流量识别方法，该方法包括：

[0011] 从接入网的每个数据流中提取与该数据流对应的子流；

[0012] 对所有子流进行抽样，得到抽样子流，对每个抽样子流的应用层协议进行识别，并根据识别结果建立抽样子流与应用层协议的映射关系；

[0013] 从每个子流提取原始数据，并根据原始数据建立与该子流对应的子流特征向量；

[0014] 根据所有子流特征向量将对应的子流进行聚类，形成多个簇；

[0015] 根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系；其中，如果一个簇不能与任何已知应用层协议建立映射关系，则该簇对应的应用层协议类型为未知；

[0016] 根据簇与应用层协议的映射关系，识别各簇中的每个子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议；其中，对于与未知应用层协议类型相对应的簇，该簇中每个子流所对应的应用层协议类型为未知，进而接入网中与该子流对应的数据流的应用层协议类型也为未知。

[0017] 在所述进而识别接入网中与该子流对应的数据流的应用层协议之后，该方法进一步包括：

[0018] 根据识别出来的接入网中数据流的应用层协议，对接入网中的数据流进行流量控制；其中，对于接入网中与未知应用层协议类型对应的数据流，首先检测其对应的子流所在的簇是否为异常数据流，如果是异常数据流，则对接入网中与该未知应用层协议类型对应的数据流进行接入控制，如果不是异常数据流，则将该未知应用层协议类型定义为新的应用层协议，将接入网中与该未知应用层协议类型对应的数据流与所述新的应用层协议建立映射关系，然后对接入网中与所述新的应用层协议建立映射关系的数据流进行流量控制。

[0019] 所述从接入网的每个数据流中提取与该数据流对应的子流的步骤包括：

[0020] 对于接入网中的每一个数据流，提取该数据流起始位置的预设数量的数据包作为该数据流的子流。

[0021] 所述对每个抽样子流的应用层协议进行识别的步骤包括：

[0022] 用深度包检测技术对每个抽样子流进行检测，从而识别出抽样子流所对应的应用层协议，并将不能与已知应用层协议相对应的抽样子流与未知应用层协议相对应。

[0023] 从每个子流提取原始数据，并根据原始数据建立与该子流对应的子流特征向量的步骤包括：

[0024] 从子流中提取数据包的特征数据作为原始数据，所述特征数据包括：协议字段、子流的持续时间、数据包的长度、数据包中有效载荷的大小、相邻数据包到达接入点的时间间隔中的任意一个或多个；

[0025] 根据所述原始数据，用统计学的方法计算子流的统计特征，所述统计特征包括：最大前向数据包的长度、数据包包头中协议的类型、相邻前向数据包到达接入点的最小时间间隔、相邻前向数据包到达接入点的时间间隔的标准差、相邻后向数据包到达接入点的时间间隔的标准差、最大后向数据包的长度以及后向数据包长度的均值中的任意一个或多个；

[0026] 将所述统计特征排列成一个向量，得到子流特征向量。

[0027] 所述根据所有子流特征向量将对应的子流进行聚类，形成多个簇的步骤包括以下步骤：

[0028] 第一步，将所有子流特征向量的状态都设置为未归类，并为每一个子流设置邻近特征向量距离 eps 和最小特征向量数目 minpts 这两个参数；

[0029] 第二步，选定任一未归类的子流特征向量 p，则 p 具有邻近特征向量距离 p.eps 和最小特征向量数目 p.minpts 两个参数，计算 p 与其他所有未归类的子流特征向量的距离；

[0030] 将参数 p.eps 的大小与 p 和其他子流特征向量之间的距离作比较，并统计与 p 的距离小于参数 p.eps 的子流特征向量的数目，然后将该数目与参数 p.minpts 作比较，如果该数目不小于参数 p.minpts，那么称 p 为核对象，与 p 的距离小于 p.eps 的所有子流特征向量称为 p 的直接密度可达对象，将 p 与所有 p 的直接密度可达对象组成以 p 为核对象的簇 a，然后将 p 和所有 p 的直接密度可达对象以及其所对应的各子流的状态都设置为属于簇 a；如果该数目小于 p.minpts，那么称 p 为噪声对象，并将 p 以及其对应的子流的状态设置为未知；

[0031] 第三步，判断是否存在未归类的子流特征向量，如果存在，则重复执行第二步，如果不存在，则结束流程；

[0032] 其中，所述 p 与其他任一子流特征向量的距离，为 p 与其他任一子流特征向量的欧氏距离。

[0033] 所述根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系的步骤包括：根据抽样子流与应用层协议的映射关系，统计簇内各抽样子流所对应的应用层协议；如果一个簇内各抽样子流所对应的应用层协议至少包括一个已知应用层协议，则计算簇内各抽样子流所对应的已知应用层协议的概率，然后将概率最大的已知应用层协议与该簇建立映射；

[0034] 所述如果一个簇不能与任何应用层协议建立映射关系，则该簇对应的应用层协议类型为未知的步骤包括：根据抽样子流与应用层协议的映射关系，统计簇内各抽样子流所对应的应用层协议；如果一个簇内各抽样子流所对应的应用层协议均为未知应用层协议，那么该簇不能与任何已知应用层协议建立映射关系，则将该簇与未知应用层协议类型相对应。

[0035] 在根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系之后，并且在根据簇与应用层协议的映射关系，识别各簇中的每个子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议之前，该方法进一步包括：

[0036] 评判聚类质量是否符合聚类标准，如果聚类质量符合聚类标准，则执行所述根据簇与应用层协议的映射关系，识别各簇中子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议；否则，重新为每一个子流设置 eps 和 minpts 这两个参数，然后重新根据子流特征向量对子流进行聚类，形成簇，并重新根据抽样子流与应用层协议的映射关系建立簇与应用层协议的映射关系，直至评判结果为聚类质量符合聚类标准；

[0037] 所述聚类质量符合聚类标准，是指聚类质量同时达到三个标准，如果不能满足其中任一标准，都为聚类质量不符合聚类标准，所述三个标准为：

[0038] 第一：状态为未知的子流特征向量的数目占所有子流特征向量数目的比例小于 10％；

[0039] 第二：簇的数目为抽样子流与应用层协议的映射关系中的已知应用层协议的数目的两倍以上；

[0040] 第三：根据抽样子流与应用层协议的映射关系建立簇与应用层协议或未知应用层协议类型的映射关系之后，与应用层协议映射的任一簇中，概率最大的应用层协议的概率不低于 60％。

[0041] 一种接入网的流量识别装置，该装置包括：

[0042] 应用层协议模块：用于存储应用层协议；向抽样子流生成和应用识别模块提供应用层协议查询服务；

[0043] 子流特征生成模块：用于从接入网的各数据流中分别提取与所述各数据流对应的子流；从所述子流分别提取原始数据，并根据原始数据建立与各子流对应的子流特征向量；

[0044] 子流聚类模块：用于从子流特征生成模块接收子流及其子流特征向量；根据子流特征向量对子流进行聚类，形成簇；向簇的应用识别模块发送簇；

[0045] 抽样子流生成和应用识别模块：用于从子流特征生成模块接收子流；对子流进行抽样，得到抽样子流；向应用层协议模块查询应用层协议；对抽样子流的应用层协议进行识别，并根据识别结果建立抽样子流与应用层协议的映射关系；

[0046] 簇的应用识别模块：用于从抽样子流生成和应用识别模块获取抽样子流与应用层协议的映射关系；从子流聚类模块接收簇；根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系；

[0047] 业务识别模块：用于从簇的应用识别模块获取簇与应用层协议的映射关系；根据簇与应用层协议的映射关系，识别各簇中子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议。

[0048] 该装置进一步包括：

[0049] 异常流量检测模块：用于从业务识别模块接收与未知应用层协议类型映射的簇及其映射的未知应用层协议类型；判断簇是否为异常数据流；将异常数据流的簇发送到异常流量控制模块；将不是异常数据流的簇所映射的未知应用层协议类型定义为新的应用层协议，发送到应用层协议模块；将不是异常数据流的簇与所述新的应用层协议建立映射；将不是异常数据流的簇及其映射的新的应用层协议发送到转发策略模块；

[0050] 异常流量控制模块：用于从异常流量检测模块接收异常数据流的簇；向接入网发送子流接入控制信息；

[0051] 转发策略模块：用于从业务识别模块接收与已知应用层协议映射的簇及其映射的已知应用层协议；从异常流量检测模块接收不是异常数据流的簇及其映射的新的应用层协议；将簇内所有子流都与该簇映射的应用层协议建立映射；根据簇映射的应用层协议的类型，向接入网发送子流流量控制信息。

[0052] 子流特征生成模块，用于从接入网传输的每一个数据流中提取该数据流起始位置的预设数量的数据包作为该数据流的子流。

[0053] 抽样子流生成和应用识别模块，用深度包检测技术对抽样子流进行检测，从而识别出抽样子流所对应的应用层协议，并将不能与已知应用层协议相对应的抽样子流与未知应用层协议相对应，从而建立抽样子流与应用层协议的映射关系。

[0054] 所述子流特征生成模块，首先从子流中提取数据包的特征数据作为原始数据，所述特征数据包括：协议字段、子流的持续时间、数据包的长度、数据包中有效载荷的大小、相

邻数据包到达接入点的时间间隔中的任意一个或多个；然后，根据所述原始数据，用统计学的方法计算子流的统计特征，所述统计特征包括：最大前向数据包的长度、数据包包头中协议的类型、相邻前向数据包到达接入点的最小时间间隔、相邻前向数据包到达接入点的时间间隔的标准差、相邻后向数据包到达接入点的时间间隔的标准差、最大后向数据包的长度以及后向数据包长度的均值中的任意一个或多个；最后，将所述统计特征排列成一个向量，得到子流特征向量。

[0055] 所述子流聚类模块根据子流特征向量对子流进行聚类、形成簇包括：

[0056] 第一步，所述子流聚类模块将所有子流特征向量的状态都设置为未归类，并为每一个子流设置邻近特征向量距离 eps 和最小特征向量数目 minpts 这两个参数；

[0057] 第二步，所述子流聚类模块选定任一未归类的子流特征向量 p，则 p 具有邻近特征向量距离 p.eps 和最小特征向量数目 p.minpts 两个参数，计算 p 与其他所有未归类的子流特征向量的距离；

[0058] 所述子流聚类模块将参数 p.eps 的大小与 p 和其他子流特征向量之间的距离作比较，并统计与 p 的距离小于参数 p.eps 的子流特征向量的数目，然后将该数目与参数 p.minpts 作比较，如果该数目不小于参数 p.minpts，那么称 p 为核对象，与 p 的距离小于 p.eps 的所有任一子流特征向量称为 p 的直接密度可达对象，所述子流聚类模块将 p 与所有 p 的直接密度可达对象组成以 p 为核对象的簇 a，然后将 p 和所有 p 的直接密度可达对象以及其所对应的各子流的状态都设置为属于簇 a；如果与 p 的距离小于 p.eps 的子流特征向量的数目小于 p.minpts，那么称 p 为噪声对象，并将 p 以及其对应的子流的状态设置为未知；

[0059] 第三步，所述子流聚类模块判断是否存在未归类的子流特征向量，如果存在，则重复执行第二步，如果不存在，结束流程；

[0060] 其中，所述 p 与其他任一子流特征向量的距离，为 p 与其他任一子流特征向量的欧氏距离。

[0061] 所述簇的应用识别模块，用于根据抽样子流与应用层协议的映射关系，统计簇内各抽样子流所对应的应用层协议，如果一个簇内各抽样子流所对应的应用层协议至少包括一个已知应用层协议，则所述簇的应用识别模块计算簇内各抽样子流所对应的已知应用层协议的概率，然后将概率最大的已知应用层协议与该簇建立映射；其中，如果一个簇不能与任何已知应用层协议建立映射关系，则所述簇的应用识别模块将该簇与未知应用层协议类型相对应。

[0062] 该装置进一步包括：

[0063] 聚类质量评判模块：用于从抽样子流生成和应用识别模块接收抽样子流与应用层协议的映射关系；从子流聚类模块接收簇、子流及其子流特征向量；从簇的应用识别模块接收簇与应用层协议的映射关系；

[0064] 所述聚类质量评判模块，用于评判聚类质量是否符合聚类标准，如果评判结果为聚类质量符合聚类标准，则向子流聚类模块发送聚类质量评判结果；

[0065] 所述子流聚类模块，只有在聚类质量评判结果为聚类质量符合聚类标准时，向簇的应用识别模块发送簇；

[0066] 所述子流聚类模块，在聚类质量评判结果为聚类质量不符合聚类标准时，重新为

每一个子流设置 eps 和 minpts 这两个参数，然后重新根据子流特征向量对子流进行聚类，形成簇，并将子流及其子流特征向量以及新的簇发送给聚类质量评判模块重新进行聚类质量的评判，直至所述聚类质量评判模块评判聚类质量符合聚类标准为止；

[0067] 所述聚类质量符合聚类标准，是指聚类质量同时达到三个标准，如果不能满足其中任一标准，都为聚类质量不符合聚类标准，所述三个标准为：

[0068] 第一：状态为未知的子流特征向量的数目占所有子流特征向量数目的比例小于 10％；

[0069] 第二：簇的数目为抽样子流与应用层协议的映射关系中的已知应用层协议的数目的两倍以上；

[0070] 第三：所述簇的应用识别模块根据抽样子流与应用层协议的映射关系建立簇与应用层协议的映射关系之后，在与已知应用层协议映射的任一簇中，概率最大的已知应用层协议的概率不低于 60％。

[0071] 由于本发明从数据流中提取出子流，又从子流中抽样出抽样子流，然后将抽样子流所对应的应用层协议与聚类形成的簇建立映射，因而相对于现有技术，本发明的检测工作量很小，存储开销和计算量也都很小，从而可以通过低速在线检测少量抽样子流对应的应用层协议的方式，来获知所有子流对应的应用层协议，进而据此对接入网中的数据流进行高速在线的流量识别，因此，本发明可以有效地对接入网进行高速在线的流量识别。

[0072] 另外，由于本发明利用子流的统计特征对子流进行聚类，然后将聚类形成的簇与簇内概率最大的应用层协议建立映射，并以此为基础进行网络流量控制，因此，本发明可以根据加密数据流的统计特征将其聚类到相应的簇，然后得到该加密数据所映射的应用层协议，从而实现对接入网中传输的加密数据包的流量识别和控制。

附图说明

[0073] 图 1 为本发明实施例提供的接入网的流量识别流程的示意图；

[0074] 图 2 为本发明实施例提供的接入网的流量识别方法的流程图；

[0075] 图 3 为本发明实施例提供的接入网的流量识别装置的结构图。

具体实施方式

[0076] 为了使本发明的目的、技术方案和优点更加清楚，下面结合附图和具体实施例对本发明进行详细描述。

[0077] 图 1 为本发明实施例中的接入网的流量识别流程的示意图。如图 1 所示，本发明实施例的主要思想为：步骤 101，从接入网中的数据流中提取子流，并生成子流特征；步骤 102，通过对子流进行抽样，得到抽样子流，然后识别抽样子流所对应的应用层协议；步骤 103，对子流进行聚类，形成簇；步骤 104，根据各簇中的抽样子流所对应的应用层协议，识别各簇所对应的应用层协议；步骤 105，判断簇所对应的应用层协议是否为已知的应用层协议，是则执行步骤 106，否则执行步骤 107；步骤 106，根据该簇所对应的已知应用层协议，对接入网中传输的与簇中的子流相对应的数据流的应用层协议进行识别，并制定和执行相应的转发策略；步骤 107，对于应用层协议为未知的簇，对该簇进行异常流量检测，对确认是异常数据流的簇进行步骤 108，对于确认不是异常数据流的簇，则将其对应的应用层协议

定义为新的应用层协议，并根据该新的应用层协议，对接入网中传输的与该簇中的子流相对应的数据流，制定并执行与新的应用层协议对应的转发策略；步骤 108，对确认是异常数据流的簇进行异常流量控制，限制其对应的数据流在接入网中的传输。

[0078] 如图 1 所示，由于本发明实施例的方案中对接入网数据流的子流进行了抽样，然后对抽样子流进行应用层协议识别，因此大大减少了需要进行应用层识别的数据量。这样，只需采用在线低速进行抽样子流的应用层协议识别，就能够与生成子流特征、子流聚类、应用映射、业务识别、转发策略、异常流量检测以及异常流量控制等需要在线高速处理的工作相匹配。因此，本发明能够对接入网进行高速在线的流量识别。

[0079] 此外，本发明实施例的方案采用了聚类这种统计方法，将具有一定相似性的子流聚类为一个簇，并使这些子流与相同的应用层协议相对应，这样就可以识别出接入网中与簇内所有子流对应的数据流的应用层协议，因此，本发明能够对接入网中传输的加密数据流进行应用层协议的识别。

[0080] 本发明实施例中的接入网的流量识别方法中涉及的应用层协议是网络和用户终端之间的接口，用于向用户终端提供各种实际的网络应用服务。常见的应用层协议包括超文本传输协议 (http)、文件传输协议 (ftp)、电子邮件协议 (smtp 和 pop3) 等。在网络通信过程中，信源和信宿都使用应用层协议，并且所使用的应用层协议必须相同。

[0081] 图 2 为本发明实施例中的接入网的流量识别方法的流程图。如图 2 所示，本发明实施例中的接入网的流量识别方法包括以下步骤：

[0082] 步骤 201：从接入网的各数据流中分别提取与各数据流对应的子流。

[0083] 这里，对于接入网中的每一个数据流，提取该数据流起始位置的预设数量的数据包，作为与各数据流对应的子流，例如，预设数量可以为 5 或 6。

[0084] 步骤 202：对子流进行抽样，得到抽样子流，对抽样子流的应用层协议进行识别，并根据识别结果建立抽样子流与应用层协议的映射关系。

[0085] 这里，对子流进行抽样的方法包括随机抽样和重尾抽样。

[0086] 随机抽样，是指按照随机性原则，从总体中抽取部分对象作为样本进行调查，以样本的调查结果推断总体有关指标的一种抽样方法。随机性原则是指从总体中抽取样本时，每个样本被抽取的概率是相同的。

[0087] 重尾抽样是指按照重尾分布的规律对总体进行抽样，重尾抽样的特点如下：大量的小抽样取值和少量的大抽样取值并存，在这些抽样数据集中，虽然大部分抽样取值是小的，但是对抽样的均值和方差起决定作用的是那些少量的大抽样取值。

[0088] 随机抽样和重尾抽样技术均为现有技术，本发明中不再详细描述。

[0089] 对子流进行随机抽样或重尾抽样后，即可得到抽样子流。

[0090] 对抽样子流的应用层协议进行识别是采用深度包检测技术进行的，深度包检测技术属于现有技术。用深度包检测技术对抽样子流进行检测，从而识别出抽样子流所对应的应用层协议，并将不能与已知应用层协议相对应的抽样子流，使其与未知应用层协议相对应，从而建立抽样子流与应用层协议的映射关系。

[0091] 步骤 203：从子流分别提取原始数据，并根据原始数据建立与各子流对应的子流特征向量。

[0092] 首先，从步骤 201 提取到的各子流中提取数据包的特征数据作为原始数据，用于

计算各子流的统计特征。这些作为原始数据的特征数据包括：协议字段、子流的持续时间、数据包的长度、数据包中有效载荷的大小、相邻数据包到达接入点的时间间隔。将这些特征数据全部提取出来可以计算得到子流最佳特征向量，如果不提取全部特征数据，仅提取其中的一个或多个，得到的子流特征向量虽然不是子流最佳特征向量，但也能实现本发明的功能；

[0093] 然后，根据提取到的原始数据，用统计学的方法计算子流的统计特征，这些统计特征包括：最大前向数据包的长度、数据包包头中协议的类型、相邻前向数据包到达接入点的最小时间间隔、相邻前向数据包到达接入点的时间间隔的标准差、相邻后向数据包到达接入点的时间间隔的标准差、最大后向数据包的长度以及后向数据包长度的均值。其中，前向是指数据流信源向信宿的传输方向，后向是指数据流信宿向信源的传输方向。这七项统计特征可以组成子流最佳特征向量，如果仅采用其中的一项或多项，得到的子流特征向量虽然不是子流最佳特征向量，但也能实现本发明的功能。

[0094] 最后，将计算得到的统计特征排列成一个向量，得到子流特征向量。

[0095] 步骤 204：根据子流特征向量对子流进行聚类，形成簇。

[0096] 这里，聚类是一种统计学的方法，用于将一个集合中的多个对象按照相似性分为若干类，每一个类称为一个簇，同一个簇内的对象具有某种相似性，并与其他簇内的对象相异。本发明根据子流特征向量对子流进行聚类、形成簇的步骤包括：

[0097] 第一步：将步骤 203 中建立的所有子流特征向量的状态都设置为未归类，并为每一个子流设置邻近特征向量距离 eps 和最小特征向量数目 minpts 这两个参数。

[0098] 第二步：选择任一未归类的子流特征向量 p，则由第一步可知，p 具有邻近特征向量距离 p.eps 和最小特征向量数目 p.minpts 这两个参数，然后，计算 p 与其他所有未归类的子流特征向量之间的距离，该距离为欧氏距离，欧氏距离的计算方法属于现有技术，本发明不做赘述。

[0099] 将参数 p.eps 的大小与 p 和其他未归类的子流特征向量之间的距离作比较，并统计与 p 的距离小于参数 p.eps 的子流特征向量的数目，然后将该数目与参数 p.minpts 作比较，如果该数目不小于 p.minpts，那么 p 称为核对象，与 p 的距离小于 p.eps 的所有子流特征向量称为 p 的直接密度可达对象，p 与所有 p 的直接密度可达对象组成以 p 为核对象的簇 a；如果与 p 的距离小于 p.eps 的子流特征向量的数目小于 p.minpts，那么 p 称为噪声对象。

[0100] 如果 p 为簇 a 的核对象，那么将 p 和所有 p 的直接密度可达对象以及其所对应的各子流的状态都设置为属于簇 a；如果 p 为噪音对象，那么将 p 以及其对应的子流的状态设置为未知 (unknown)。

[0101] 第三步，判断是否存在未归类的子流特征向量，如果存在，则重复执行第二步，直至不存在未归类的子流特征向量，如果不存在，表示聚类完成，则结束聚类流程，执行步骤 205。

[0102] 聚类完成后，所有子流特征向量的状态只能为属于某一个簇或噪音对象中的一种，不存在未归类的子流特征向量。

[0103] 步骤 205：根据抽样子流与应用层协议的映射关系，按照应用层协议概率优势原则，建立簇与应用层协议的映射关系。

[0104] 这里，应用层协议概率优势原则是指，将簇内各抽样子流所对应的已知应用层协议中概率最大的应用层协议，作为该簇所对应的应用层协议，从而建立簇与应用层协议的映射关系。例如，某簇由 100 个子流聚类形成，其中有 10 个子流为抽样后得到的抽样子流，在这 10 个抽样子流中，有 7 个抽样子流与应用层协议 A 映射，有 2 个抽样子流与应用层协议 B 映射，另外 1 个抽样子流与未知应用层协议映射，那么该簇中各抽样子流所对应的应用层协议中概率最大的为应用层协议 A，其概率计算如下 :7/10 = 70%，因此，利用应用层协议概率优势原则即可决定该簇与应用层协议 A 建立映射。再例如，某簇由 100 个子流聚类形成，其中有 10 个抽样子流，在这 10 个抽样子流中，有 2 个抽样子流与应用层协议 A 映射，有 1 个抽样子流与应用层协议 B 映射，另外 7 个抽样子流与未知应用层协议映射，那么该簇中各抽样子流所对应的已知应用层协议中概率最大的仍为应用层协议 A。

[0105] 根据抽样子流与应用层协议的映射关系，统计簇内各抽样子流所对应的应用层协议。

[0106] 如果簇内各抽样子流所对应的应用层协议至少包括一个已知应用层协议，则计算簇内各抽样子流所对应的已知应用层协议的概率，然后根据应用层协议概率优势原则，将概率最大的已知应用层协议与该簇建立映射。

[0107] 如果簇内各抽样子流所对应的应用层协议均为未知应用层协议，即该簇不能与任何已知应用层协议建立映射关系，那么将该簇对应的应用层协议类型为未知，即将该簇与未知应用层协议类型相对应。

[0108] 这样，就建立起簇与应用层协议的映射关系。

[0109] 另外，如果簇内各抽样子流所对应的应用层协议至少包括一个已知应用层协议，那么该簇内概率最大的已知应用层协议的概率有最优值，该最优值可以最好地保证按照应用层协议概率优势原则所建立的簇与应用层协议的映射关系的覆盖全面性，该最优值为 70% -80%中的任一值。

[0110] 步骤 206 :评判聚类质量是否符合聚类标准。

[0111] 这里，聚类质量符合聚类标准意味着聚类质量同时达到以下三个标准，如果不能满足其中任一标准，都判定为聚类质量不符合聚类标准，这三个标准为 :

[0112] 第一 :步骤 204 设置的状态为未知 (unknown) 的子流特征向量的数目占所有子流特征向量数目的比例小于 10%。

[0113] 控制状态为 unknown 的子流特征向量的比例，可以提高子流聚类成的簇的数量，从而使本发明根据簇与应用层协议或未知应用层协议类型的映射来对接入网进行流量识别的方法更有代表性，也更有效。

[0114] 第二 :步骤 204 形成的簇的数目为步骤 202 建立的抽样子流与应用层协议的映射关系中的已知应用层协议的数目的两倍以上。

[0115] 如果步骤 204 形成的簇的数目过少，达不到步骤 202 建立的抽样子流与应用层协议的映射关系中映射的数目的两倍及以上，那么每个簇内的抽样子流所映射的应用层协议的数目就会比较多，比如超过 5 个，这样会直接导致簇中概率最大的应用层协议的概率比较低，比如低于 60%，从而使步骤 205 所建立的簇与应用层协议的映射关系不具有代表性，进而影响对接入网流量识别的质量。

[0116] 第三 :步骤 205 中根据抽样子流与应用层协议的映射关系，按照应用层协议概率

优势原则，建立簇与已知应用层协议或未知应用层协议类型的映射关系之后，与已知应用层协议映射的任一簇中，概率最大的应用层协议的概率不低于 60％。

[0117] 簇中概率最大的应用层协议的概率如果低于 60％，那么所建立的该簇与应用层协议或未知应用层协议类型的映射的代表性太差，不足以满足对接入网进行流量识别的质量要求。

[0118] 评判聚类质量是否符合聚类标准，如果聚类质量符合聚类标准，那么执行步骤 207 ；如果评判结果为聚类质量不符合聚类标准，那么重新设置步骤 204 中每一个子流的 eps 和 minpts 这两个参数，然后重新根据子流特征向量对子流进行聚类，形成簇，并重新根据抽样子流与应用层协议的映射关系，按照应用层协议概率优势原则，建立簇与应用层协议的映射关系，直至评判结果为聚类质量符合聚类标准。

[0119] 通过评判聚类质量是否符合聚类标准，可以提高步骤 204 形成的簇的质量，进而提高步骤 207 对接入网中数据流的应用层协议进行识别的质量。

[0120] 步骤 207 ：根据簇与应用层协议的映射关系，识别各簇中子流所对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议。

[0121] 根据步骤 205 建立的簇与应用层协议的映射关系，可以得到各簇所对应的应用层协议，将各簇中所有子流均与相应的簇所对应的应用层协议建立映射，然后就可以识别出接入网中与各子流对应的数据流的应用层协议。

[0122] 对于与未知应用层协议类型相对应的簇，该簇中所有子流所对应的应用层协议类型均为未知，进而在接入网中与这些子流对应的数据流的应用层协议类型也为未知。

[0123] 步骤 208 ：根据识别出来的接入网中数据流的应用层协议，对接入网中的数据流进行流量控制。

[0124] 这里，数据流的应用层协议不同，对数据流的流量控制方法也有所不同。

[0125] 如果数据流与已知的应用层协议相对应，则根据识别出来的接入网中数据流的应用层协议，在接入网的接入点处对数据流进行相适应的流量控制。

[0126] 如果数据流与未知应用层协议类型相对应，则首先采用深度包检测技术检测该数据流对应的子流所在的簇是否为异常数据流。如果该簇是异常数据流，则在接入网的接入点处对接入网中传输的与该簇内子流相对应的数据流进行接入控制，比如，对于蠕虫、木马等计算机病毒以及端口扫描等网络入侵类型的异常数据流，通过本发明的高速在线识别，可以及时地在接入网的接入点处进行接入限制，从而避免用户或网络受到病毒或网络入侵的危害。如果该簇不是异常数据流，则将该未知应用层协议类型定义为新的应用层协议，将该簇与新的应用层协议相对应，从而建立该簇与所述新的应用层协议的映射，然后将该簇中所有的子流以及接入网中与这些子流相应的数据流也与新的应用层协议建立映射，根据这种新的应用层协议，在接入网的接入点处对接入网中传输的与该簇内子流相应的数据流进行流量控制。

[0127] 基于图 2 所述的接入网的流量识别方法，本发明还提出了一种接入网的流量识别装置。图 3 为本发明实施例提供的接入网的流量识别装置的结构图。如图 3 所示，该装置包括 ：

[0128] 应用层协议模块 301 ：用于存储应用层协议 ；向抽样子流生成和应用识别模块 304 提供应用层协议查询服务 ；

[0129] 子流特征生成模块 302 :用于从接入网的各数据流中分别提取与所述各数据流对应的子流 ;从所述子流分别提取原始数据,并根据原始数据建立与各子流对应的子流特征向量 ;

[0130] 子流聚类模块 303 :用于从子流特征生成模块 302 接收子流及其子流特征向量 ;根据子流特征向量对子流进行聚类,形成簇 ;向簇的应用识别模块发送簇 ;

[0131] 抽样子流生成和应用识别模块 304 :用于从子流特征生成模块 302 接收子流 ;对子流进行抽样,得到抽样子流 ;向应用层协议模块 301 查询应用层协议 ;对抽样子流的应用层协议进行识别,并根据识别结果建立抽样子流与应用层协议的映射关系 ;

[0132] 簇的应用识别模块 305 :用于从抽样子流生成和应用识别模块 304 获取抽样子流与应用层协议的映射关系 ;从子流聚类模块 303 接收簇 ;根据抽样子流与应用层协议的映射关系,按照应用层协议概率优势原则,建立簇与应用层协议的映射关系 ;

[0133] 业务识别模块 306 :用于从簇的应用识别模块 305 获取簇与应用层协议的映射关系 ;根据簇与应用层协议的映射关系,识别各簇中子流所对应的应用层协议,进而识别接入网中与该子流对应的数据流的应用层协议。

[0134] 其中,子流特征生成模块 302 从接入网的各数据流中分别提取与所述各数据流对应的子流的方法包括 :子流特征生成模块 302 从接入网传输的每一个数据流中提取该数据流起始位置的预设数量的数据包作为该数据流的子流。

[0135] 抽样子流生成和应用识别模块 304 对抽样子流的应用层协议进行识别,并根据识别结果建立抽样子流与应用层协议的映射关系的方法包括 :

[0136] 抽样子流生成和应用识别模块 304 用深度包检测技术对抽样子流进行检测,从而识别出抽样子流所对应的应用层协议,并将不能与已知应用层协议相对应的抽样子流与未知应用层协议相对应,从而建立抽样子流与应用层协议的映射关系。

[0137] 子流特征生成模块 302 从所述子流分别提取原始数据,并根据原始数据建立与各子流对应的子流特征向量的方法包括 :

[0138] 所述子流特征生成模块 302,首先从子流中提取数据包的特征数据作为原始数据,所述特征数据包括 :协议字段、子流的持续时间、数据包的长度、数据包中有效载荷的大小、相邻数据包到达接入点的时间间隔,将这些特征数据全部提取出来可以计算得到子流最佳特征向量,如果不提取全部特征数据,仅提取其中的一个或多个,得到的子流特征向量虽然不是子流最佳特征向量,但也能实现本发明的功能 ;然后,根据所述原始数据,用统计学的方法计算子流的统计特征,所述统计特征包括 :最大前向数据包的长度、数据包包头中协议的类型、相邻前向数据包到达接入点的最小时间间隔、相邻前向数据包到达接入点的时间间隔的标准差、相邻后向数据包到达接入点的时间间隔的标准差、最大后向数据包的长度以及后向数据包长度的均值,这七项统计特征可以组成子流最佳特征向量,如果仅采用其中的一项或多项,得到的子流特征向量虽然不是子流最佳特征向量,但也能实现本发明的功能 ;最后,将所述统计特征排列成一个向量,得到子流特征向量。

[0139] 子流聚类模块 303 根据子流特征向量对子流进行聚类、形成簇的方法包括 :

[0140] 第一步,所述子流聚类模块 303 将所有子流特征向量的状态都设置为未归类,并为每一个子流设置邻近特征向量距离 eps 和最小特征向量数目 minpts 这两个参数 ;

[0141] 第二步,所述子流聚类模块 303 选定任一未归类的子流特征向量 p,则 p 具有邻近

特征向量距离 p.eps 和最小特征向量数目 p.minpts 两个参数，计算 p 与其他所有未归类的子流特征向量的距离；

[0142] 所述子流聚类模块 303 将参数 p.eps 的大小与 p 和其他子流特征向量之间的距离作比较，并统计与 p 的距离小于参数 p.eps 的子流特征向量的数目，然后将该数目与参数 p.minpts 作比较，如果该数目不小于参数 p.minpts，那么称 p 为核对象，与 p 的距离小于 p.eps 的所有子流特征向量称为 p 的直接密度可达对象，所述子流聚类模块 303 将 p 与所有 p 的直接密度可达对象组成以 p 为核对象的簇 a，然后将 p 和所有 p 的直接密度可达对象以及其所对应的各子流的状态都设置为属于簇 a；如果与 p 的距离小于 p.eps 的子流特征向量的数目小于 p.minpts，那么称 p 为噪声对象，并将 p 以及其对应的子流的状态设置为未知；

[0143] 第三步，所述子流聚类模块 303 判断是否存在未归类的子流特征向量，如果存在，则重复执行第二步，直至不存在未归类的子流特征向量，如果不存在，那么所述子流聚类模块 303 根据子流特征向量对子流进行聚类、形成簇的流程结束；

[0144] 其中，所述 p 与其他任一子流特征向量的距离，为 p 与其他任一子流特征向量的欧氏距离。

[0145] 簇的应用识别模块 305 根据抽样子流与应用层协议的映射关系，按照应用层协议概率优势原则，建立簇与应用层协议的映射关系的方法包括：所述簇的应用识别模块 305 根据抽样子流与应用层协议的映射关系，统计簇内各抽样子流所对应的应用层协议，如果一个簇内各抽样子流所对应的应用层协议至少包括一个已知应用层协议，则所述簇的应用识别模块 305 计算簇内各抽样子流所对应的已知应用层协议的概率，然后将概率最大的已知应用层协议与该簇建立映射；其中，如果一个簇不能与任何已知应用层协议建立映射关系，则所述簇的应用识别模块 305 将该簇与未知应用层协议类型相对应。

[0146] 本发明实施例中接入网的流量识别装置将接入网中数据流的应用层协议识别出来之后，还可以进一步对接入网中的数据流进行流量控制。由于数据流的应用层协议不同，对数据流的流量控制方法也有所不同，因此，如图 3 所示，该装置进一步包括：

[0147] 异常流量检测模块 307：用于从业务识别模块 306 接收与未知应用层协议类型映射的簇及其映射的未知应用层协议类型；判断簇是否为异常数据流；将异常数据流的簇发送到异常流量控制模块 308；将不是异常数据流的簇所映射的未知应用层协议类型定义为新的应用层协议，发送到应用层协议模块 301；将不是异常数据流的簇与所述新的应用层协议建立映射；将不是异常数据流的簇及其映射的新的应用层协议发送到转发策略模块 309；

[0148] 异常流量控制模块 308：用于从异常流量检测模块 307 接收异常数据流的簇；向接入网发送子流接入控制信息。

[0149] 转发策略模块 309：用于从业务识别模块 306 接收与已知应用层协议映射的簇及其映射的已知应用层协议；从异常流量检测模块 307 接收不是异常数据流的簇及其映射的新的应用层协议；将簇内所有子流都与该簇映射的应用层协议建立映射；根据簇映射的应用层协议的类型，向接入网发送子流流量控制信息。

[0150] 在子流聚类模块 303 根据子流特征向量对子流进行聚类、形成簇之后，为了提高聚类质量，进而提高本发明对接入网的流量识别质量，如图 3 所示，本发明实施例中的接入

网的流量识别装置进一步包括：

[0151] 聚类质量评判模块 310：用于从抽样子流生成和应用识别模块 304 接收抽样子流与应用层协议的映射关系；从子流聚类模块 303 接收簇、子流及其子流特征向量；从簇的应用识别模块 305 接收簇与应用层协议的映射关系；评判聚类质量是否符合聚类标准；向子流聚类模块 303 发送聚类质量评判结果。

[0152] 聚类质量评判模块 310 用于评判聚类质量是否符合聚类标准，如果评判结果为聚类质量符合聚类标准，则向子流聚类模块 303 发送聚类质量评判结果；所述子流聚类模块，只有在聚类质量评判结果为聚类质量符合聚类标准时，才向簇的应用识别模块发送簇；

[0153] 所述子流聚类模块，在聚类质量不符合聚类标准时，那么所述聚类质量评判模块 310 向子流聚类模块 303 发送聚类质量评判结果，所述子流聚类模块 303 重新为每一个子流设置 eps 和 minpts 这两个参数，然后所述子流聚类模块 303 重新根据子流特征向量对子流进行聚类，形成簇，并将子流及其子流特征向量以及新的簇发送给聚类质量评判模块重新进行聚类质量的评判，直至所述聚类质量评判模块 310 评判聚类质量符合聚类标准为止；

[0154] 所述聚类质量符合聚类标准，是指聚类质量同时达到三个标准，如果不能满足其中任一标准，都为聚类质量不符合聚类标准，所述三个标准为：

[0155] 第一：状态为未知的子流特征向量的数目占所有子流特征向量数目的比例小于 10％；

[0156] 第二：簇的数目为抽样子流与应用层协议的映射关系中的已知应用层协议的数目的两倍以上；

[0157] 第三：所述簇的应用识别模块 305 根据抽样子流与应用层协议的映射关系、按照应用层协议概率优势原则建立簇与应用层协议的映射关系之后，在与已知应用层协议映射的任一簇中，概率最大的已知应用层协议的概率不低于 60％。

[0158] 由此可见，本发明具有以下优点：

[0159] (1) 本发明从数据流中提取出子流，又从子流中抽样出抽样子流，然后将抽样子流所对应的应用层协议与聚类形成的簇建立映射，因而相对于现有技术，本发明的检测工作量很小，存储开销和计算量也都很小，从而可以通过低速在线检测少量抽样子流对应的应用层协议的方式，来获知所有子流对应的应用层协议，进而据此对接入网中的数据流进行高速在线的流量识别，因此，本发明可以有效地对接入网进行高速在线的流量识别。

[0160] (2) 本发明利用子流和抽样子流来实现实时的网络流量检测，因而不需要采用现有技术那种先完整接收整个数据流再生成统计特征的方法来进行检测，因此，本发明可以实时完成数据流的识别和控制操作。

[0161] (3) 本发明利用子流的统计特征对子流进行聚类，然后将聚类形成的簇与簇内概率最大的应用层协议建立映射，并以此为基础进行网络流量控制，因此，本发明可以根据加密数据流的统计特征将其聚类到相应的簇，然后得到该加密数据所映射的应用层协议，从而实现对接入网中传输的加密数据包的流量识别和控制。

[0162] (4) 本发明对于与未知应用层协议类型映射的簇，采取深度包检测技术检测其是否为异常数据流，因而本发明能够实时检测出蠕虫、木马等计算机病毒以及端口扫描等网络入侵类型的异常数据流，并对这些网络流量进行实时的流量控制，进而及时产生告警通知网络管理员，因此，本发明能够实时地避免计算机病毒及网络入侵等异常网络数据流对

网络和用户的危害。

[0163] (5) 本发明可以评判聚类质量是否符合聚类标准，从而自动调整聚类参数，改善子流的聚类质量，因此，本发明能自动适应数据流的变化，保证了对接入网中网络流量识别效果的可靠性。

[0164] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明保护的范围之内。

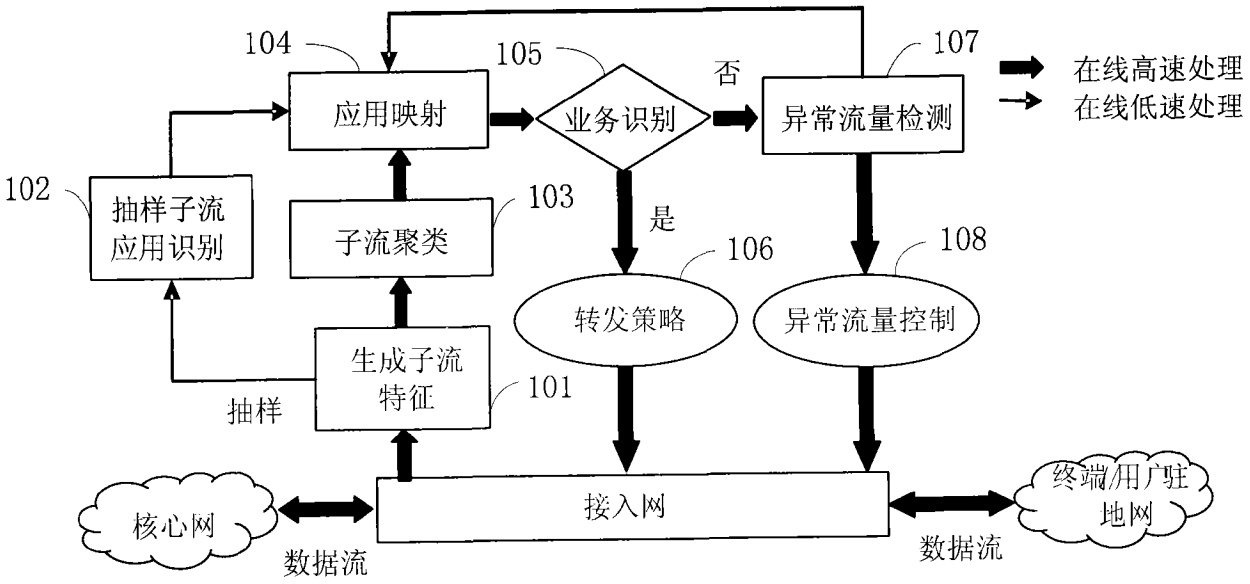

104
105
107
在线高速处理
在线低速处理
应用映射
业务识别
否
异常流量检测
102
抽样子流
应用识别
子流聚类
103
是
106
108
转发策略
异常流量控制
生成子流
特征
101
抽样
核心网
接入网
终端/用户驻
地网
数据流
数据流

图 1

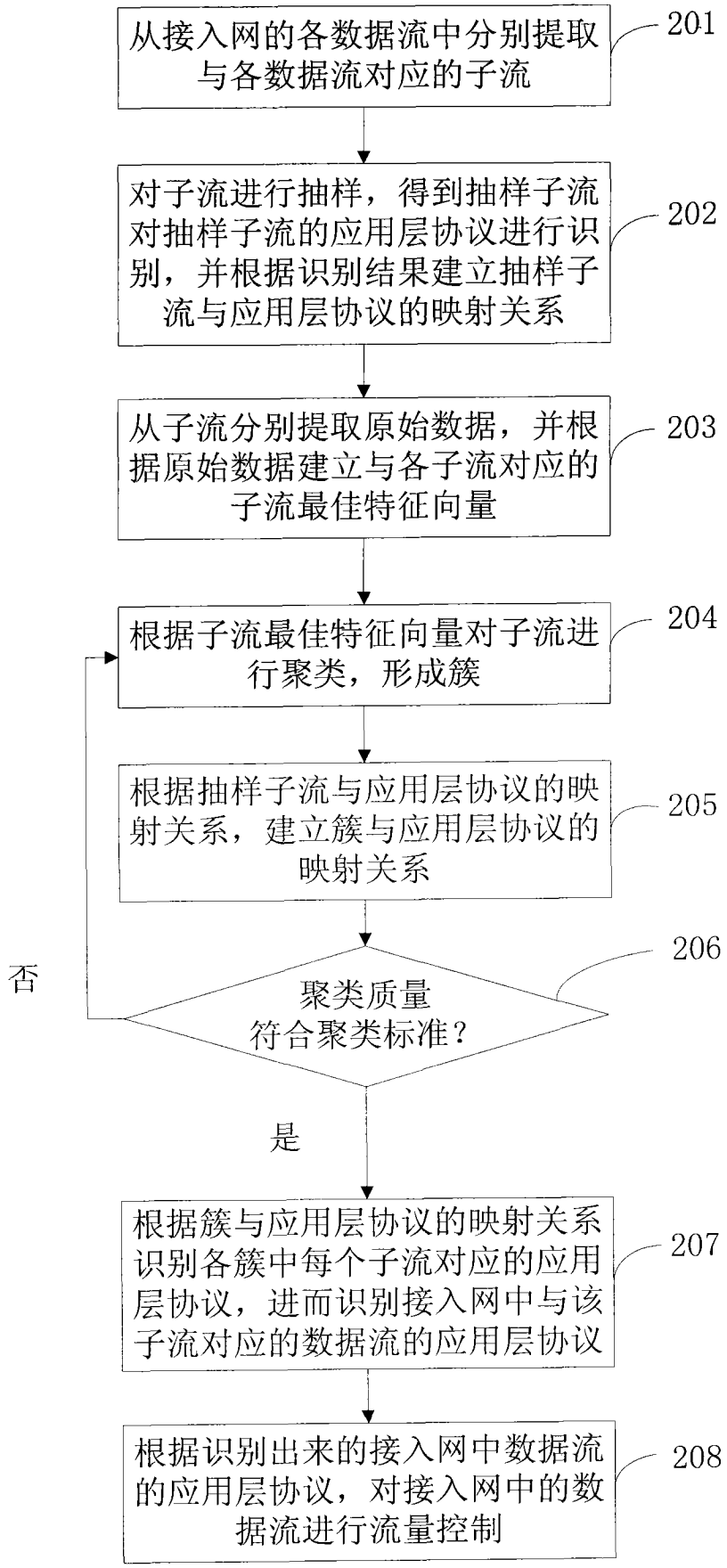

从接入网的各数据流中分别提取与各数据流对应的子流
201
对子流进行抽样，得到抽样子流对抽样子流的应用层协议进行识别，并根据识别结果建立抽样子流与应用层协议的映射关系
202
从子流分别提取原始数据，并根据原始数据建立与各子流对应的子流最佳特征向量
203
根据子流最佳特征向量对子流进行聚类，形成簇
204
根据抽样子流与应用层协议的映射关系，建立簇与应用层协议的映射关系
205
聚类质量符合聚类标准？
206
否
是
根据簇与应用层协议的映射关系识别各簇中每个子流对应的应用层协议，进而识别接入网中与该子流对应的数据流的应用层协议
207
根据识别出来的接入网中数据流的应用层协议，对接入网中的数据流进行流量控制
208

图 2

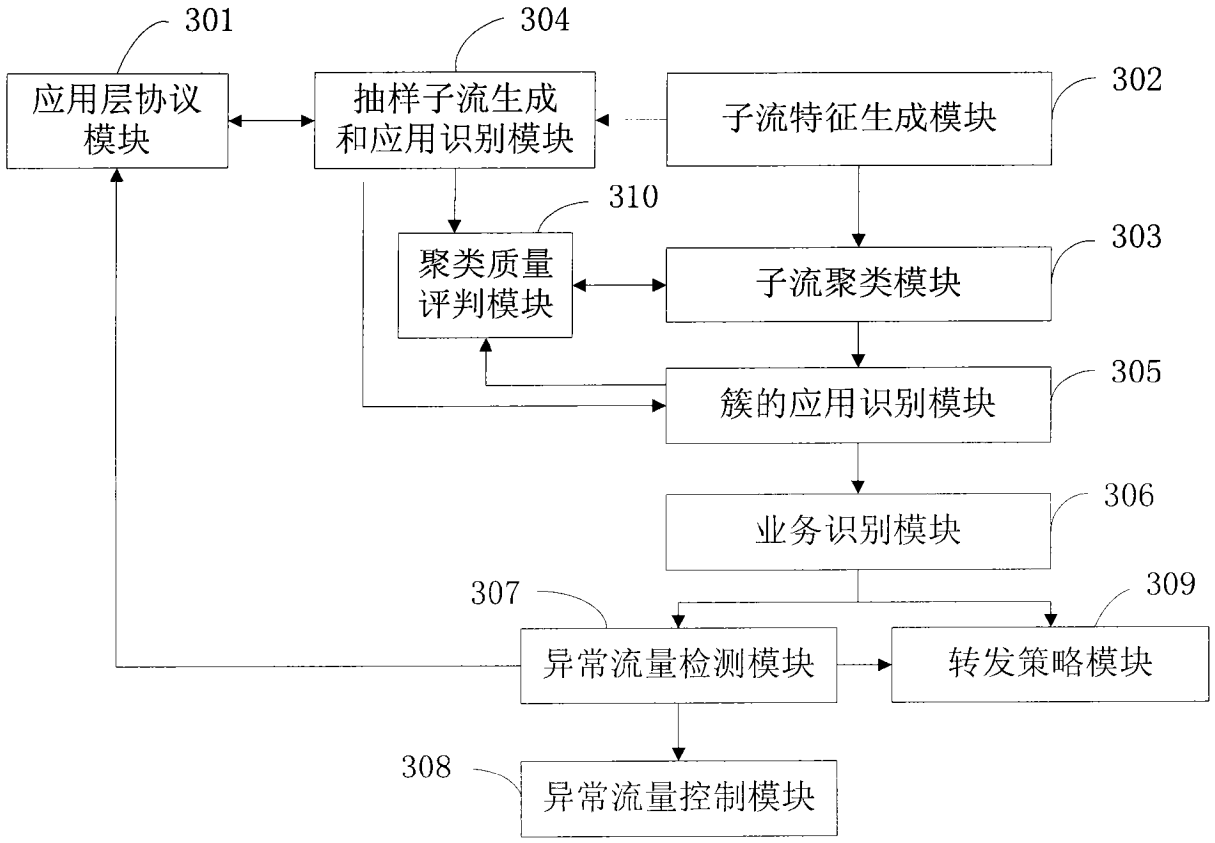

301
304
应用层协议模块
抽样子流生成和应用识别模块
子流特征生成模块
302
310
聚类质量评判模块
子流聚类模块
303
簇的应用识别模块
305
业务识别模块
306
307
309
异常流量检测模块
转发策略模块
308
异常流量控制模块

图 3