(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2019 年 6 月 20 日 (20.06.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/114700 A1

(51) 国际专利分类号:
H04L 12/851 (2013.01)

(21) 国际申请号: PCT/CN2018/120321

(22) 国际申请日: 2018 年 12 月 11 日 (11.12.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201711354592.7 2017年12月15日 (15.12.2017) CN

(71) 申请人: 华 为 技 术 有 限 公 司 (**HUAWEI TECHNOLOGIES CO., LTD.**) [CN/CN]; 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(72) 发明人: 欧 阳 黜 霏 (**OUYANG, Chufei**); 中国广东省深圳市龙岗区坂田华为总部办公楼, Guangdong 518129 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

WO 2019/114700 A1

(54) **Title:** TRAFFIC ANALYSIS METHOD, PUBLIC SERVICE TRAFFIC ATTRIBUTION METHOD AND CORRESPONDING COMPUTER SYSTEM

(54) **发明名称**: 流量分析方法、公共服务流量归属方法及相应的计算机系统

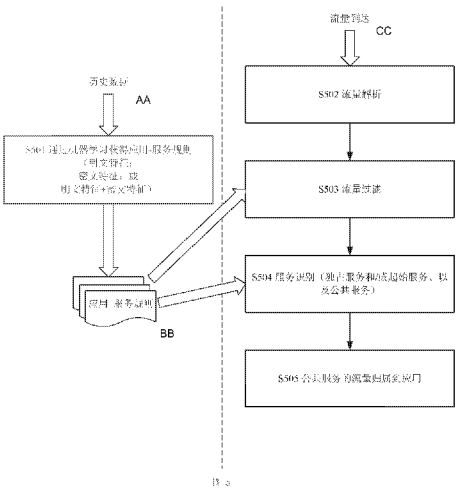

图 5

S501 BY MEANS OF MACHINE LEARNING, OBTAIN AN APPLICATION-SERVICE RULE (PLAINTEXT CHARACTERISTICS, CIPHERTEXT CHARACTERISTICS, OR PLAINTEXT CHARACTERISTICS + CIPHERTEXT CHARACTERISTICS)
S502 TRAFFIC ANALYSIS
S503 TRAFFIC FILTERING
S504 SERVICE IDENTIFICATION (EXCLUSIVE SERVICE AND/OR STARTING SERVICE, AND PUBLIC SERVICE)
S505 THE TRAFFIC OF A PUBLIC SERVICE IS ATTRIBUTED TO AN APPLICATION
AA HISTORICAL DATA
BB APPLICATION-SERVICE RULE
CC TRAFFIC ARRIVES

(57) **Abstract:** Provided in the present application are a traffic analysis method and device and a computer system. The method comprises: acquiring plaintext characteristics and ciphertext characteristics of a message in traffic, the ciphertext characteristics comprising length characteristics of encrypted fields in the message; and analyzing the traffic according to the plaintext characteristics and the ciphertext characteristics so as to identify the service or application to which the traffic is attributed. The method may be used for service identification as well as application identification. The accuracy of traffic identification is increased in the message encryption scenario by means of introducing ciphertext characteristics into traffic analysis. On the other hand, also provided in the present application are a method and device for attributing traffic of a public service and a computer system. The traffic of the public service is attributed to a corresponding application by identifying initial service and/or exclusive service of the application, thus providing application traffic attribution accuracy and providing a basis for subsequent services such as application traffic billing.

(57) **摘要**: 本申请提供一种流量分析方法、装置和计算机系统。该方法包括获取流量中的报文的明文特征和密文特征，所述密文特征包括所述报文中加密后的字段的长度特征；根据所述明文特征和所述密文特征对所述流量执行分析，以识别该流量所归属的服务或应用。该方法可以用于服务识别，也可以应用于应用识别。通过在流量分析中引入密文特征，在报文加密的场景下增加了流量识别的准确度。另一方面，本申请还提供一种公共服务的流量归属方法、装置和计算机系统，通过识别应用的起始服务和/或独占服务，将公共服务的流量归属到相应的应用，提供了应用流量归属的准确度，为后续应用流量计费等服务提供了基础。

[见续页]

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

流量分析方法、公共服务流量归属方法及相应的计算机系统

技术领域

本申请涉及网络流量分析领域，尤其涉及流量分析方法、公共服务流量归属方法以及相应的计算机系统等。

背景技术

近年来随着移动互联网的发展迅速，网络流量占比逐年上升。为了保障网络用户享受到可靠的服务，方便网络管理方实时管理和监控各个网元间的活动，流量分析技术广泛应用于各种网络设备，包括网关、路由等网络中转或分组设备。当前运营商利用该技术为其开展计费、阻断、策略控制等网络业务提供信息保障。

流量分析的其中一个目的是将流量中包含的报文（或数据流）归属到不同的应用，本申请中将该过程称之为应用识别。例如：运营商可能需要对某个用户的某个手机应用的流量进行计费，这就需要计算出该用户在某个时段内由于使用该应用而产生的流量的大小，亦即属于该应用的流量的大小。要计算这个流量大小，就需要先分析该时段内的流量，从中识别出属于该应用的报文。基于应用识别，运营商就能够进行差异化计费，为消费者提供更丰富的服务，又能够监控网络实时状况，动态调整自己的网络资源分配。

然而，随着网络协议技术和应用程序技术的演进，流量分析技术面临到一些挑战。

一方面，现有技术通常使用报文的明文特征做应用识别，报文的明文特征即报文中能够被直接解析得到的字符或数字组成的特征。但是随着网络协议加密技术的广泛使用，原有非加密协议下报文的部分明文特征被隐藏，而应用识别仅使用未被隐藏的特征，使得应用识别的准确度下降。

另一方面，一个应用可能调用多个服务，现有的流量分析技术能够一定程度上区分属于不同应用的流量，但很少涉及更细粒度的区分，比如区分属于不同服务的流量（本申请称之为服务识别，作为流量分析的另一个目的），特别是网络协议加密技术引入之后，报文的很多原有的明文特征被隐藏，使得服务识别难上加难。

进一步的，当多个应用调用同一服务时，会产生相似度较高的流量，即公共服务流量，如何将这些相似的公共服务流量识别出来并归属到各自的应用是目前业界没有解决的一个难题。

发明内容

下面介绍本申请提供的流量分析方法、服务识别方法、以及相应的装置等。应理解的是，以下方面未必涵盖本申请提出的所有实现方式，并且不同方面的实现方式和有益效果可互相参考。

第一方面，本申请提供一种流量分析方法，该方法用于在网络协议加密的场景下提高流量分析的精确度，具体的，可以提高应用识别或服务识别的准确度。该方法可以应用在网关或其它类型的网络设备中。该方法包括：获取流量中的报文的特征，所述特征包括密文特征，所述密文特征包括所述报文中加密后的报文的顺序、长度、传输方向中的任意一个或多个。根据所述特征对所述流量执行分析，以识别该流量所归属的服务或

应用。

“流量”是泛指，可以是一条或多条数据流，被提取特征的报文的数量可以是一个或多个。本申请中提到的“报文”包括“数据包”和其它类型的报文，比如不带数据，只有报头的报文。

“加密后的报文”根据加密方式的不同可以指不同的报文，本实施例不做限定。“服务”指的是比应用小或等同于应用的功能组件，服务提供的功能通常通过某种接口被应用调用。一个应用可以调用一个或多个服务。

报文的顺序指的是单个报文出现在一条数据流中的位置，或是多个报文的先后顺序。

报文的长度指的是包长，包长通常可以从报文中的某个字段中获取到；若没有表示包长的字段，也可以实时计算包长。

报文的传输方向包括上行或下行。

举例来说，密文特征可以为一条流量中第一个 A 报文（A 报文为加密后的一种报文）的包长为 m 字节且第二个 A 报文的包长为 n 字节；或者，第一个方向为上行的 A 报文的长度为 m 且第二个方向为下行的 A 报文（或其他类型的报文）的包长为 n 字节；或者第一个方向为上行的 A 报文为该流量中的第 n 个报文，等等，可能的特征组合方式有很多，这里不再一一列举。

可见，该方法提供了一种从流量中高效识别服务或应用的方法，解决了在网络协议加密的场景下服务或应用无法识别的问题，并且通过在服务识别或应用识别的过程中考虑报文的密文特征，提高了识别的准确度。

在一些实施例中，密文特征还可以和明文特征组合使用。所述明文特征包括所述报文中能够被直接解析得到的字符和/或数字组成的特征。这里提到的“字符和/或数字组成的特征”包括单个字符或单个数字，也包括字符串以及其它可能的组合。在具体工程实现中，字符类型例如为 char， 字符串类型例如为 string。

在一些实施例中，该方法应用于深度报文解析（deep packet inspection, DPI）设备，该 DPI 设备可以是独立的网络设备，也可以内置在网关 GRPS 支持节点 (gateway GPRS support node, GGSN) 中。DPI 设备应用该方法之后，可以配合策略和计费网关（policy and charging rules function, PCRF）进行计费。在其它一些实施例中，该方法应用于其它网络流量解析设备。

在一些实施例中，被提取明文特征的报文包括安全传输层协议(transport layer security，TLS)握手报文中（部分或全部信息）。

在一些实施例中，被提取密文特征的报文包括 application data 这种数据包。

在一些实施例中，所述根据所述特征对所述流量执行分析，以识别该流量所归属的服务或应用，包括：将所述特征与服务的识别规则或应用的识别规则匹配，以识别所述流量所归属的服务或应用。这里可能用到的这两种识别规则是基于所述特征，通过机器学习算法获得的。通过机器学习的方法获得服务或引用的识别规则，使得整个流量分析过程更加智能，提高了流量分析的准确度。

应理解的是，通过机器学习算法学习识别规则时用到的报文并非当前待分析的报文，是从历史流量中获得的报文，或者通过其他方法获得的与现实流量具有相同或相似特征的仿真流量，从中获取的报文。

在一些实施例中，在对流量分析之前可以对流量执行过滤，过滤方法可参考下面几个方面所提供的过滤方法。

在一些实施例中，流量分析之后可继续对各种服务的流量进行归属，可参考下面几个方面所提供的流量归属方法。

第二方面，本申请提供一种公共服务流量归属方法。公共服务是被两个或多个应用调用的服务，所以在流量分析中识别出来的属于某一个公共服务的这部分流量需要确定是被哪个应用调用而产生的，也就是说该部分流量属于哪个应用，这样才能支持后续的应用流量计费等操作。

首先，获取流量中的报文的特征，所述特征包括密文特征，所述密文特征包括所述报文中加密后的报文的顺序、长度、传输方向中的任意一个或多个。然后，根据所述特征对所述流量执行分析，识别该流量中的起始服务、独占服务和公共服务。根据识别时间在起始服务 A 和起始服务 B 的识别时间之间的独占服务确定一个应用。确定识别时间在所述起始服务 A 和所述起始服务 B 的识别时间之间的公共服务的流量归属到所述应用。这的起始服务 A 和起始服务 B 是在流量中被识别到的时间（即识别时间）具有先后顺序的任意两个相邻服务，换句话说，所述起始服务 B 为识别时间在所述第一起始服务之后的第一个起始服务，两者可以是同样的服务，也可以是不同服务。这里的 A 和 B 仅是为了区分两个服务，便于理解，也可以用“第一”和“第二”来区分。这里的“在...之间”可以包括端点。

在一些实施例中，所述特征还包括明文特征。

在一些实施例中，应用调用的服务被分为三种类型：起始服务，即运行在应用启动阶段的服务（并非仅指应用启动服务），例如启动服务、登陆服务、注册服务等；公共服务，即可以被多个应用调用的服务。和公共服务不同，独占服务是仅被一个应用调用的服务，所以通过这样独占服务可以唯一确定一个应用。

应理解的是，公共服务和独占服务是互相排斥的，但是起始服务可能是独占服务，也可能是公共服务。

在一些实施例中，“识别时间”在具体实现时可以不用真的时间值来表示，只要记录各个服务的识别顺序即可。

在一些实施例中，服务的识别时间并非是真实识别到服务的时间，可以用其他一些能够标识服务被识别到的先后顺序的数字或其它类型的信息表示。

由于起始服务是应用启动阶段的服务，所以通过两个起始服务的识别时间可以确定一个时间段，该时间段内产生的流量属于前一个起始服务所归属的应用。但是这个应用是哪个应用单纯通过起始服务无法确定，因为起始服务也可能是公共服务，所以需要通过该时间段内的一个独占服务来确定，由于独占服务的特点，所以该时间段内独占服务对应的应用就是前一个起始服务所归属的应用，也是该时间段内产生的所有流量所归属的应用，所以该时间段内的公共服务产生的流量理所当然也被归属到该应用。公共服务虽然被多个应用调用，但是通过本申请提供的方法就可以确定识别到的每个公共服务到底是被哪个应用调用，换句话说，其产生的流量到底应该归属给哪个应用。

在一些实施例中，在执行流量分析之前还可以对流量进行过滤。具体的，在一些实现方式中，过滤条件为最大进包数量，该最大进包数量为根据识别规则确定的一个报文

的数量值。识别规则由于是预先通过某些方法获得的，所以该识别规则要运用的话需要的最大报文的数量可以直接或通过某种计算方式获得。在另一些实现方式下，根据所述流量的网络协议（internet protocol, IP）信息对所述流量执行过滤。具体的，通过IP信息计算获得需要进行流量分析的某类应用的自治系统号(autonomous system number, ASN)域，通过该ASN域对流量执行过滤。通过在分析之前执行过滤可以减少待分析的报文数量，提高流量分析的效率；而且最大进包数量可根据识别规则适应性调整，提高了过滤的灵活性。

在一些实施例中，所述特征对所述流量执行分析，识别该流量中的起始服务、独占服务和公共服务，包括：将所述特征与第一识别规则、第二识别规则和第三识别规则分别匹配以识别所述流量中的所述起始服务、所述独占服务和所述公共服务，其中所述第一识别规则、所述第二识别规则和所述第三识别规则是基于所述特征，通过机器学习算法获得的。

在一些实施例中，第一识别规则、第二识别规则和第三识别规则为分别识别起始服务、独占服务和公共服务三种服务的识别规则，这三种识别规则是在流量分析之前基于报文的明文特征和密文特征组合，通过机器学习算法获得的。这里的报文如前述所述，来源于历史流量数据或通过某种数学方法仿真的数据。

在一些实施例中，以上三种识别规则中每种识别规则都可以包括多个识别规则，分别用于识别同类型的多个服务。换句话说，本申请并没有限定第一、第二或第三识别规则仅是一个识别规则。

在一些实施例中，提取应用的流量的特征（作为样本），并将特征输入到机器学习算法（通常为有监督的机器学习算法），通过机器学习过程输出三种服务识别规则：用于识别起始服务的第一识别规则，用于识别应用独占服务的第二识别规则，以及用于识别公共服务的第三识别规则。机器学习的过程中输入的特征包含报文的明文特征和/或密文特征。

在一些实施例中，机器学习的过程可以是线下的，也可以是线上的。多个应用所涉及的服务识别规则的学习可以同时进行，也可以分别单独进行。可选的，根据同一种应用内学习到的服务识别规则建立应用-服务规则的关联关系。

在一些实施例中，在实时系统中，当待分析的流量到达之后，根据之前学习到的三种服务识别规则区分三种服务，识别结果中包含各个服务的位置信息（相当于服务的识别时间）。然后根据其中起始服务的位置将单个用户下的流量分段，根据每个分段中的独占服务确定该分段所属的应用，然后将位置位于各个分段内的公共服务归属于各个分段所属的应用。其中，服务的位置信息用于指示服务的先后关系，具体可以用识别到服务的时间表示。

在一些实施例中，独占服务和调用该服务的应用的对应关系可以预先存储在存储器中，根据该对应信息确定所述应用。

第三方面，本申请提供另一种公共服务流量归属方法，与前一种方法相比，该方法起始服务和独占服务可以仅使用其中一种。

在一些实施例中，获取流量中的报文的特征，所述特征包括密文特征，所述密文特征包括所述报文中加密后的报文的顺序、长度、传输方向中的任意一个或多个。根据所

述特征对所述流量执行分析，识别该流量中的独占服务和公共服务，其中所述独占服务是仅被一个应用调用的服务，所述公共服务是被两个或多个应用调用的服务。根据识别到的独占服务 A 确定一个应用，所述应用为调用所述独占服务 A 的应用。将识别时间在所述独占服务 A 和下一个识别到的独占服务 B 之间的公共服务的流量归属到所述应用。应理解的是，调用该独占服务 B 的应用可能是跟之前一样的同一个应用。这里的 A 和 B 仅是为了区分两个服务，便于理解，也可以用“第一”和“第二”来区分。这里的“在...之间”可以包括端点。

在其它一些实施例中，使用起始服务替代上述过程中的独占服务。

第三方面的任意实施例可以应用于同一个应用中起始服务和独占服务（指的是起始服务之后的第一个独占服务）之间不存在任何其它流量的情况，因为此时独占服务相当于起始服务。

第三方面的其它实现方式可参考第二方面提供的实施例，在此不再赘述。第三方面和第二方面在产品中可以单独实现，也可以同时实现，同时实现时可以先判断起始服务和独占服务之间存不存在其它流量，若不存在，则使用第三方面或第三方面任意实施例的方法；若存在，则使用第二方面或第二方面任意实施例提供的方法。

第四方面，本申请还提供一种流量过滤方法，该方法通常在流量分析之前执行。在流量分析中用到服务或应用的识别规则（或其它类型的分析规则），该过滤方法则是在流量分析之前就先通过后面会用到的这些识别规则中的一个或多个确定流量分析时的最大进包数量，多于该数量的数据包被过滤掉。例如，识别规则均涉及 application data 数据包（此处仅是一种识别规则的举例），按照流量中报文的顺序来看，涉及到的最靠后的 application data 数据包是第 n 个，那么第 n 个 application data 数据包之后的 application data 数据包就可以被过滤掉了，这里 n 就可以作为最大进包数量的值。该最大进包数量可以是一个数值，也可以是不同类型的数据包分别对应的多个数值。

第五方面，本申请还提供一种流量过滤方法，该方法中根据流量中的 IP 信息计算 ASN 域，然后通过该 ASN 域进行流量过滤。

第六方面，本申请还提供一种流量分析装置，包括一个或多个单元，用于实现第一方面任意一种方法。另外，本申请还提供一种公共服务流量归属装置，包括一个或多个单元，用于实现第二方面任意一种方法或第三方面任意一种方法。再者，本申请还提供一种流量过滤装置，包括一个或多个单元，用于实现第四方面任意一种方法或第五方面任意一种方法。

第七方面，本申请提供一种计算机系统，包括存储器和处理器，所述存储器用于存储计算机可读指令，所述处理器用于读取所述计算机可读指令并实现前述第一方面到第五方面任意一种或多种方法。

第八方面，本申请还提供一种计算机可读存储介质，该介质通常是非易失性的，该介质用于存储计算机可读指令，该计算机可读指令在被一个或多个处理器读取之后实现前述第一方面到第五方面任意一种或多种方法。

本申请中出现的“多个”或“多次”若无特殊说明则意指“两个或两个以上”，或“两次或两次以上”。本申请中出现的“第一”和“第二”并无限定顺序的意思，仅是为了在某些描述上下文中区分两个主体，以方便理解，但是其所指示的主体并非在所有

实施例中都必须是不同的主体。本申请中出现的“A/B”、“A 和/或 B”包括 A、B 以及 A 和 B 三种情况。本申请中“A®”意指 A 为一个商标名称，但没有带®的词语也有可能是商标名称。

附图说明

为了更清楚地说明本申请提供的技术方案，下面将对附图作简单地介绍。显而易见地，下面描述的附图仅仅是本申请的一些实施例。

图 1 为流量的层次示意图；

图 2 为 HTTP 请求报文和 HTTP 响应报文的示例；

图 3 为 TLS 握手过程示意图；

图 4 为本申请一实施例提供的流量分析装置的逻辑结构示意图；

图 5 为本申请一实施例提供的流量分析方法的流程示意图；

图 6 为本申请一实施例提供的流量归属方法的原理示意图；

图 7 为本申请一实施例提供的流量分析装置的逻辑结构示意图；

图 8 为本申请一实施例提供的流量特征构造方法的流程示意图；

图 9 为本申请一实施例提供的服务或应用的识别规则学习的流程示意图；

图 10 为本申请一实施例提供的流量过滤方法的流程示意图；

图 11 为本申请一实施例提供的三种服务类型识别的流程示意图；

图 12 为本申请一实施例提供的公共服务流量归属方法的流程示意图；

图 13 为本申请一实施例提供的计算机系统的逻辑结构示意图。

具体实施方式

为了方便理解本申请提出的技术方案，首先在此介绍本申请描述中会引入的几个要素。应理解的是，以下介绍仅方便理解这些要素，以期理解实施例的内容，并非一定涵盖所有可能的情况。

流量：通过网络连通的设备之间发生交互，就会产生网络通信报文，这些报文被称为流量。流量是一个泛指。

数据流：在服务器与客户端的一次完整通信过程（从连接建立到连接结束）中产生的数据包，称为该次连接的数据流，应用使用过程中通常会执行多次交互，因此会产生多条数据流，组成应用流量。

例如，以建立 TLS 握手开始，以传输控制协议（transmission control protocol，TCP） FIN（finish）报文为终止的一次会话期间产生的流量。数据流表示两个主体之间的一次交互过程，例如应用进程与服务器间的一次交互。

公共服务：部署在服务器上供多个应用程序调用的 API，公开提供完成某些功能的服务，如地图导航、云存储、视频传输等等。

流量分析：通过监听、抓取、拷贝等手段获取网络通信报文，并对其进行解析、重组、切分等还原其原本通信内容的操作，以了解网络通信双方的即时状态。

明文特征：报文中能够被直接解析得到的字符和/或数字组成的特征，区别于密文特征。

请参考图 1，为流量的层次结构示意图。图 1 以移动应用脸书（Facebook®）为例，该应用的流量从层次结构上可以分为三层，第一层是数据流层，即以建立 TLS 握手开始，以 TCP FIN 包为终止的一次会话期间产生的流量，表示应用进程与服务器间的一次交互。第二层是服务层，即该应用中与服务器进行交互的子模块，其对应的进程与服务器交互产生的所有流量即为该服务模块的流量，比如 Facebook®云存储服务、消息服务等。第三层是应用层，即该应用程序 Facebook®。Facebook®中还包括公共服务，例如登陆服务、云服务、消息推送服务等。Facebook®中的公共服务可以被其它应用程序调用，这意味着属于公共服务的流量不一定全部属于 Facebook®。当新流量到达之后，流量分析模块在识别出公共服务之后，还需要通过一定方法将公共服务的流量做归属，划分到应该归属的应用程序中，这样才能准确计算应用程序的流量。

当前流量识别技术主要集中在应用层面的流量识别，基本没有涉及服务层面的流量识别，但是目前应用市场公共服务流量已经占据了总体流量的 60%以上，使用公共服务模块的应用数量占总体的 95%以上。其中最突出的服务识别问题为 Google®类服务识别，例如：所有使用 Google®地图服务的应用程序，都会出现公共服务流量，例如 Google®地图流量，的识别冲突，严重影响运营商的业务。然而在实际应用中，应用层流量识别技术不能精确识别服务，产生较高的误识别率。

现有普遍使用的流量分析方案为明文特征识别方法，利用超文本传输协议（hypertext transfer protocol，HTTP）报文的明文特征和 TLS 握手报文的明文特征识别流量。HTTP 报文包括请求报文和响应报文。图 2 为 HTTP 请求报文（a）和 HTTP 响应报文（b）的示例。HTTP 报文由三部分组成，分别是：起始行、消息首部和主体。表 1 示出了起始行可能的动作。

表 1

动作	含义
GET	请求获取 URI 所标识的资源
POST	在 URI 所标识的资源后附加新的数据
HEAD	请求获取由 URI 所标识的资源的响应消息报头
PUT	请求服务器存储一个资源，并用 URI 作为其标识
DELETE	请求服务器删除 URI 所标识的资源
TRACE	请求服务器回送收到的请求信息，主要用于测试或诊断
CONNECT	保留将来使用
OPTIONS	请求查询服务器的性能，或者查询与资源相关的选项和需求

在流量分析中，通过上述动作就可以判断客户端和服务器端正在进行的交互行为。例如，利用统一资源标识符（uniform resource identifier, URI）标识的资源可以确定交互的内容，首部字段中的 Host 字段可以用来判断该报文是否属于某个应用，等等。因此明文特征分析技术通常直接利用这些可以被解析的字符或数字特征去推测网络通信双

方的状态，后续当加密技术引入网络通信协议后，只有少部分未加密的流量能够继续使用该项技术。

由于协议加密技术的应用，原有 HTTP 报文的明文特征字段，全部被加密变成了基于超文本传输安全协议（hypertext transfer protocol secure, HTTPS）的字段。当前网络流量的 90%以上全部为 HTTPS 协议，其结构是在原来的 HTTP 报文之上封装了一层 TLS 协议，其握手过程如图 3 所示，类此 TCP 协议的三次握手过程。如图 3 所示，TLS 协议客户端首先发送 ClientHello 给服务端，服务端返回 ServerHello 和证书，客户端接受证书后生成加密用的公钥，发送公钥和加密算法给服务端，服务端确认后结束握手过程。之后双方开始发送加密的应用数据（application data）报文。在协议加密的前提下明文特征包括 TLS 握手报文的特征，而密文特征包括加密的应用数据的特征。现有技术仅使用了流量中明文特征来执行应用识别。

TLS 握手报文的基本类型主要包括 10 种（还有其它扩展类型）。下文中关于 TLS 握手报文的特征构造主要基于这 10 种报文中的一个或多个。这 10 种报文包括图 3 中示出的（1）-（5）以及（7）（相当于（9）），另外还包括图 3 中没有示出的 HelloRequest、ServerKeyExchange、CertificateRequest 以及 CertificateVerify。下面通过表 2 对这 10 种报文进行简单的介绍。表 2 中的报文有些是根据服务器或客户端的要求需要的，并非所有场景下都是必须的。

表 2

报文类型	含义或功能
HelloRequest	由服务器主动发起的握手，这种情况不常发生，主要是用在这种情况下：一个 session 已经持续了很长时间，服务器为了降低安全隐患，重新与客户端建立新的连接。
ClientHello	客户端向服务器发送的打招呼消息，包含一个 session ID。
ServerHello	服务器向客户端发送的打招呼消息，包含服务器选定使用的加密算法和压缩算法。
Certificate	服务器向客户端发送的证书链。
ServerKeyExchange	由客户端从服务器段接收，里面携带一个参数，用于建立对称加密。这是一个可选的参数，并不是所有的秘钥交换算法都需要这个参数。
CertificateRequest	服务器要求客户端提供证书，这种在 web 服务器中并不常见。
ServerHelloDone	打招呼结束消息
ClientKeyExchange	负责向服务器发送下面三项信息： 一个随机数，该随机数用服务器公钥加密，防止被窃听； 编码改变通知，表示随后的信息都将用双方商定

	的加密方法和密钥发送； 客户端握手结束通知，表示客户端的握手阶段已经结束。这一项同时也是前面发送的所有内容的hash值，用来供服务器校验。
CertificateVerify	客户端需要对服务端的证书进行检查，是否可信机构颁布、证书中的域名是否与实际域名一致、或者证书是否已经过期。如果证书检查通过，客户端就会从服务器证书中取出服务器的公钥。
Finished	这个消息发送时，已经是被加密的了。 因为 协商已经结束，ChangeCipherSpec 消息已经发送并激活双方的加密通信。

需要说明的是，ChangeCipherSpec 协议并不属于握手协议的一部分，发送它表明双方的加密状态已经准备好了，接下来的通信使用双方协商好的密文加密通信，在本申请中不再详细介绍。另外，这里的 Finished 包表示握手过程结束，并非前文所述 TCP FIN 报文。客户端与服务器的通信过程实际是在 TCP 层先建立 TCP 握手，然后以 TCP 协议传送图 3 所示的 TLS 握手报文，然后传送业务报文，最后以 TCP FIN 报文结束本次交互。

现有方案可利用上述 TLS 握手报文中的一种或多种来构造特征，将特征转化为机器可读的规则，如 XML，并存储这些规则。当网络流量被解析完成后，读入这些规则按对应的协议格式过滤流量，过滤方式可以为顺序过滤，从 ClientHello 报文开始，到 Finish 报文结束，建立全量的匹配规则(即报文中所有明文字段全部输入)。当过滤完成后，将过滤流量送入业务逻辑匹配模块，根据规则对应的应用 ID 识别该流量所归属的应用，将匹配结果输出。

但是，对于某一些同类型的应用，仅靠上述TLS握手报文的特征来建立规则可能存在不能区分应用的问题，因为同类型应用在某些特征（例如证书）上相似度比较高。同时，仅靠上述TLS握手报文的特征也不能识别同一应用下不同服务的流量。尤其是不同应用使用同一个服务产生的公共流量会被识别为单个应用流量，特别是服务内部也存在嵌套服务的情况下，会产生大量的误识别。当前的这些明文特征无法细分服务流量，当出现公共服务流量时也无法完成识别，所以在公共服务出现后，统计流量时往往将下一个或者上一个应用的公共流量统计到当前应用中来，导致误识别率较高。

这里同类型的应用指的是调用的服务相同或相似的应用，因为服务器会给同一种服务颁发同一类证书，此时仅靠上述TLS握手报文不能识别。同类型的应用可能是相同类型的应用，例如两个同公司或不同公司的地图应用，也可能是同一公司的不同类型应用但调用了相同的服务。

请参考图 4，为本实施例提供的一种流量分析装置 400 的逻辑结构示意图。该装置包括特征学习模块 410、服务识别模块 420、流量归属模块 430。

进一步的，该流量分析装置还可以和一个流量解析装置 300 相连，该流量解析装置 300 用于对接收到的流量执行解析，然后将解析后的结果输入到流量分析装置 400。流

量解析过程就是按照协议格式，一步一步提取字段的值域信息（具体由图 4 中解析模块执行），具体可采用现有技术，在本实施例中不做详细介绍。

进一步的，该流量分析装置 400 中还可以包括一个流量过滤模块 440，用于对流量解析装置 300 的输出结果按照特征学习模块 410 获得的全部或部分规则执行过滤，将过滤后的流量再输入到服务识别模块 420，以减少服务识别模块 420 的处理量，提高处理效率。解析过程还可以结合硬件实现，例如结合硬件加速装置加速解析过程。

图 4 中的多个模块可以部署在同一台物理机上，也可以部署在不同的物理机上。

以流量分析装置 400 装置为例，下面介绍本申请提供的流量分析方法，该流量分析方法属于流量分析装置 400 提供的部分或全部功能。

请参考图 5，为本实施例提供的流量分析方法的方法流程示意图。

S501、特征学习模块 410 根据采集的历史流量数据或通过其他方式获得的流量数据，执行机器学习，通过机器学习获得每个应用的应用-服务规则。

机器学习的过程中需要提取报文特征，这里的报文特征包括报文的明文特征以及密文特征中的其中一个或两个。所述明文特征包括所述报文中能够被直接解析得到的字符和/或数字组成的特征。所述密文特征包括所述报文中加密后的报文的顺序、长度、传输方向中的任意一个或多个。

一个应用的应用-服务规则包含该应用所调用的三种服务的识别规则。三种服务包括起始服务、应用独占服务、以及公共服务。该应用-服务规则用于执行服务识别，同时因为这三种规则和特定的应用关联，通过该规则也能获知识别的流量属于哪一个应用。两个或更多不同的应用，它们的起始服务和公共服务可能是部分或完全相同的，所以学习获得的识别规则可能是存在部分重复的。

该机器学习过程可以在线下执行，即并非实时的；也可以实时执行。当实时执行时可以周期性地获取一些流量数据，通过机器学习的方法生成或更新应用-服务规则。

在其他一些实施例中，管理者可以通过一个管理配置模块（图中未示出）来管理特征学习模块 410 获得的规则，例如管理者可以自己增加、删除、修改或查看这些规则。

S502、当流量到达之后，流量解析装置 300 从存储器（例如内存）中读取流量中的报文，按照报文的协议格式解析报文，将解析后的报文（或称流量）传输给流量过滤模块 440。

该解析过程主要涉及传输层以上的协议，即 TCP/IP 层，例如 TLS 协议。基于 TLS 协议的报文按照格式可以分为 TLS 握手部分和 TLS 记录（record）部分。在本实施例中，握手部分主要包含有 7 种类型的数据包，包括 Client Hello，ServerHello，Certificate 等数据包。前述已经提到过，10 种类型的数据包不一定会全部都用到。

S503、流量过滤模块 440 从流量解析装置 300 接收流量，并从特征学习模块 410 获得应用-服务规则，将接收到的报文根据该应用-服务规则执行过滤，并将过滤后的报文发送给服务识别模块 420。

具体的，特征学习模块 410 将应用-服务规则通过文件或其他形式存储在存储器中，流量过滤模块 440 从该存储器中读取该应用-服务规则后根据该应用-服务规则执行流量过滤。

流量过滤模块 440 主要是为了在进行服务识别之前对流量进行预处理，例如过滤或

分流等，以减少系统开销，提高服务识别模块420的处理效率。流量过滤模块440能够支持按HTTP、TLS等不同报文中的不同字段进行解析，也能够支持自定义正则过滤模式。

在其他一些实施例中，流量过滤模块440可以不需要。

S504、服务识别模块420从流量过滤模块440接收过滤后的流量，并从特征学习模块410获得应用-服务规则，根据该应用-服务规则对过滤后的流量执行服务识别，获得识别结果。该识别结果中包含流量所归属的服务的类型：起始服务、应用独占服务、或公共服务，以及每种服务的“位置”。最后，将该识别结果发送给流量归属模块430。

这里服务的位置“位置”并非地理位置的意思。服务的位置信息可以理解为一个标记或指示，用于表示该服务相对于其它服务被识别到的时间的先后顺序，例如服务的位置信息可以为识别到该服务的时间点，或是可以反映先后顺序的数字等。

例如，确定某条数据流S1的特征匹配某个应用的起始服务的特征，则该数据流S1的流量均归属到起始服务中，然后将数据流 S1、启示类服务、服务位置这种对应关系记录到存储器中。

S505、流量归属模块430接收到服务识别模块420发送的识别结果，根据其中起始服务和独占服务（或仅靠独占服务）来确定公共服务的流量归属到哪个应用。

具体的，服务识别模块420将识别结果记录到存储器中，可以是缓存，也可以是其它类型的存储器，之后流量归属模块430在从该存储器中读取识别结果。

在一种实现方式下，不用考虑应用切分时间（即起始服务的位置），当识别到独占服务时在存储器中记录该独占服务对应的应用（例如应用ID），从该时间点往后出现的公共服务，其流量都归属到该应用。后续再识别到下一个独占服务时记录下新的应用（也可能和之前的应用一样，因为同一个应用可能有两个或多个独占服务）。这种方法适宜于起始服务和独占服务之间没有流量，独占服务相当于起始服务的场景。

在另一种实现方式下，先识别起始服务，确定应用切分时间，将切分时间存入存储器。需要说明的是，这里的“时间”并非一定是一个时间的值。当识别到独占服务时在存储器中记录该独占服务对应的应用，从该时间点往后出现的公共服务，其流量都归属到该应用。后续再识别到下一个起始服务之后才考虑更新存储器中记录的应用。

以上两种实现方式中，为了节省存储器的存储空间，可以在实现方法的前提下设置存储的内容的老化时间，或设置存储的内容的条目的数量等。

下面以第二种实现方式为例介绍，第一种实现方式仅存在微小区别，本领域技术人员参考第二种实现方式即可得知第一种如何实现。

首先，根据所识别到的所有起始服务的位置信息将当前接收到的流量分段。举例来说，起始服务 SS_a 和起始服务 SS_b 之间属于第一分段；起始服务 SS_b 和起始服务 SS_c 之间属于第二分段。

然后，根据独占服务的位置信息确定一个分段所对应的应用。例如，独占服务 OS_b 位于第二分段内，而独占服务 OS_b 为应用B所独占，那么确定第二分段对应应用B。应理解的是，分段和应用并非一一对应关系，第二分段对应应用B，但不代表应用B的流量仅存在于第二分段中，应用B可能会启动多次。

最后，根据公共服务的位置信息和分段对应的应用确定公共服务的归属应用。例如，

公共服务 PS_a 位于第二分段，而已知第二分段对应应用 B，那么公共服务 PS_a 的流量被归属到应用 B。

S502-S505 通常为实时处理过程。

为便于理解，图 6 用示意图的方式示例出公共服务的流量归属过程。图中用一个箭头代表一条数据流，也代表一种服务，服务的位置指的是箭头的起始位置。箭头上的方框代表上行报文和下行报文，多个方框组合形成不同的报文特征。如图 6 所示，假设通过步骤 S504 已经识别出三个起始服务 SS_a、SS_b 和 SS_c ；两个独占服务 OS_a 和 OS_b；以及两个公共服务 PS_a 和 PS_b。

在起始服务 SS_b 之后，下一个起始服务 SS_c 之前，存在一个独占服务 OS_b，而 OS_b 已知为应用 B 所独占，所以可确定起始服务 SS_b 为应用 B 的起始服务，进而可确定应用 B 的启动时间大约为起始服务 SS_b 的位置所指示的时间。同理，独占服务 OS_a 为应用 A 所独占，所以可确定启示类服务 SS_a 为应用 A 的起始服务。

公共服务 PS_a 位于第二分段，是应用 B 启动之后，所以其流量应该被归属到应用 B。而另一个公共服务 PS_b 虽然有大部分数据流的到达时间与第二分段重合，但是从图中看到其初始位置（识别到该公共服务的位置）是位于第一分段的，而此时应用 B 还未启动，所以 PS_b 的流量不应该被归属到应用 B，而应该被归属到应用 A。

需要说明的是，识别到服务的时间（即服务的位置所表示的时间）并非是应用启动运行或服务启动运行的确切时间，但是识别到服务的先后顺序与服务运行的先后顺序通常是一致的。

以上对方案进行了整体性地说明，下面将以Google®应用(例如Google Map)为例，详细介绍服务识别方法以及服务流量归属的方法，涉及以上各个步骤的具体实现。 当前的技术对 Google®类应用的流量识别准确度较低，无法正确确定公共服务流量的归属，影响正常的运营商流量识别业务的开展，因此，本申请以 Google®类应用为例介绍流量分析方法。

下面将要介绍的方法的目标是确定Google公共服务流量的归属，以提高Google®应用的流量识别准确度。

方法大致过程与图5类似，包括：首先通过加密流量特征构造和特征学习技术，得到应用-服务规则，具体包括三类规则，即用于识别起始服务的第一识别规则，用于识别独占服务的第二识别规则和用于识别公共服务的第三识别规则（具体规则学习过程见后续）；然后利用应用-服务规则过滤技术减少待匹配的流量、动态设定进包数量等，以减少系统性能开销；之后利用应用-服务规则识别三种类型的服务，并根据不同类型服务的位置确定公共服务归属到哪个应用。

请参考图7，为本实施例提供的流量分析装置700的逻辑结构示意图。该流量分析装置700从流量解析装置800中接收被解析后的流量，执行流量分析。具体的，该流量分析装置700包括特征学习模块710、服务识别模块720、流量归属模块730以及流量过滤模块740。下面结合详细的方法介绍该装置。

图8示出了特征向量的确定方法。该方法由特征学习模块710的构造器711执行。首先由构造器711构造特征矩阵（S801），其中每一列为一种特征。

构造特征矩阵的方法可以使用如下三种中的一种或多种。第一种：根据报文的明文

构造特征矩阵，比如ClientHello报文中的SNI(server name indication)字段作为一列特征。第二种：根据协议的密文特征构造特征矩阵，比如上行应用数据（application data）的第一包数据长度和/或下行数据包长度等，无需获取密文内容。第三种：将明文和密文组合起来构造的特征矩阵。第一次构造特征矩阵，可以为人工构造，后续步骤中可以根据学习到的特征值范围再进行调整。

得到特征矩阵后，开始生成特征向量（S802）。具体的，检查应用流量中每条数据流的特征，如果该数据流在对应特征列出现了该特征，则标记1，没有出现则标记为0。这样最终能够得到全部数据流的特征矩阵，矩阵的每一行表示一条数据流的特征向量。比如，Google Map的应用流量包含20条数据流，构造特征列为30列，则输出20×30的由0和1构成的特征矩阵。

图9示出了基于特征向量，通过机器学习算法获得应用-服务规则的方法。该方法由特征学习模块710的学习器712执行。学习器712从构造器711获取特征向量，并根据机器学习算法寻找能够区分服务的特征向量，搜索该服务的特征向量对应的特征列和特征值，将搜索结果转换为用于识别该服务的规则（或称为服务的识别规则）(S901)。具体的，搜索三种类型的识别规则：第一识别规则，第二识别规则和第三识别规则，分别对应前述实施例提到的启示类服务识别规则、独占服务识别规则和公共服务识别规则。

当学习器712寻找到用于区分服务的特征向量时（S902），则输出该特征向量对应的识别规则，并将针对同一种应用学到的服务识别规则合并为该应用的应用-服务规则（S903）。当学习器712没有寻找到用于区分服务的特征向量时（S902），向构造器发送重新构造特征矩阵的请求(S904)，请求重新构造特征矩阵。参考图8，当构造器711确定接收到该请求之后(S803)，采用一些预定的方法重新构造特征矩阵(S804)，例如，将密文的特征（例如数字特征）进行等长的分割，然后根据分割结果再次构造特征矩阵，并重新输出特征向量。迭代图8和图9所示的步骤，直到输出应用-服务规则。

本实施例可以使用的机器学习算法例如决策树、人工神经网络、支持向量机、聚类、贝叶斯分类、马尔科夫链和概率图模型等。

规则包括三种：第一识别规则、第二识别规则和第三识别规则。一种规则包括一个或多个字段，如下表3-表5所示。

需要说明的是，表3-表5中的“字段”指的是规则中的字段，是自定义的。“位置”是实际的数据包中的字段，该字段通常由互联网协议小组约定，在对应协议的RFC(Request For Comments)文档中可见，是领域内共识；通过该字段可以获取到值来匹配规则中的字段的预设值。

表3

第一识别规则	字段	位置	说明	举例
	SNI	TLS Handshake	该字段为服务器名	“clients4.google.com”
	TLS record	TLS record Length	包长特征	比如第一包 record 长度 254，可以判定为 Google Map 起始

第一识别规则示例：

SNI=www.googleapis.com && TLS record =512

使用该规则的时候，从接收到的数据包的TLS handshake字段获取值，以及从TLS record length字段获取值，与该识别规则匹配，确定获取的两个值是否分别就是www.googleapis.com和512，若是，则匹配成功；若否，则匹配不成功。以下其它规则的使用方法与此类似，不再赘述。

表4

第二识别规则	字段	位置	说明	举例
	SNI	TLS Handshake	该字段为服务器名	“clients4.google.com”
	CertCommonName	Certificate	证书别名	“blackberry.com”
	UserAgent	HTTP head	浏览器、系统名称（单包识别）	“com.google.android.youtube”
	UDP-UserAgent	HTTP head	浏览器、系统名称（单包识别）	“com.google.android.youtube”
	Client application data (cAppD)	TLS record length	客户端发送给服务器端的数据（考虑分包和性能可用TCP.length替代）	0-1300（同方向有序匹配，支持TCP、TLS报文）
	Server application data（sAppD）	TLS record length	服务器端发送给客户端的数据（考虑分包和性能可用TCP.length替代）	0-1300（该方向最大匹配4包，同方向有序匹配，TCP、TLS报文）
	Other	TLS handshake	其它可能的握手特征	已有TLS识别（指纹）规则

第二识别规则示例：

iOS®系统：SNI= clients4.google.com && sAppD[1]==62 && sAppD[2]==42 && sAppD[3]==38 && sAppD[4]>=242 && sAppD[4]<=243 && cAppD[1]==53 && cAppD[2]==50 && cAppD[3]>=301 && cAppD[3]<=308

Android®系统: SNI= clients4.google.com && sAppD[1]== 376 && nCAppD>=1 && cAppD[1]>=848 && cAppD[1]<=849

其中，sAppD[x]表示server端发送给client端第x个application data数据包的长度；cAppD[x] 表示client端发送给server端第x个的application data数据包长度。

表5

第三识别规则	字段	位置	说明	举例
	SNI	TLS Handshake	该字段为服务器名（单包识别）	“clients4.google.com”
	CertCommonName	Certificate	证书别名（单包识别）	“blackberry.com”
	Other	TLS handshake	其它可能的握手特征	已有TLS识别（指纹）规则

第三识别规则示例：

SNI_googleadservices.com

SNI_www.googleapis.com

CertCommonName_google-analytics.com

以上为服务识别规则的获取过程，该过程是线下执行的。下面介绍实时的流量分析过程。在该实时流量分析过程中，以下获取流量、过滤流量、识别服务、以及公共服务流量归属等过程是实时的、顺序的执行过程。

图10示出了流量过滤的方法，该方法不是必须的，但可以减少待匹配的流量，提高处理效率。该方法由流量过滤模块740中的域过滤模块741执行。该模块741的输入有两部分，一部分来源于流量解析装置800对网络流量的解析得到的报文（即待过滤的流量），另一部分来源于学习期712输出的应用-服务规则。该模块741的输出为过滤后的流量。

具体的，接收到应用-服务规则和待过滤的流量之后，根据应用-服务规则确定该规则识别服务时需要的最大进包数量（S1001），另外根据待过滤的流量的IP信息计算Google的ASN域（S1001），根据该判断结果和前述最大进包数量执行流量的过滤(S1002)，过滤后的流量属于 Google 的 ASN 域且满足最大进包数量的要求。

这里的最大进包数量指的是流量分析装置700从一条数据流中读取的报文的最大数量，例如最大进包数量为5，那么读取的报文数量小于或等于5，超过则不会读取了，也就是说，流量过滤的时候多余5个的其他数据包被过滤掉。

图11示出了对过滤后的流量执行服务识别的方法。本方法由服务识别模块720执行，输入为域过滤模块741对当前网络流量的过滤结果以及特征学习模块710输出的应用-服务规则；输出为服务分类识别结果。首先由单用户识别模块721根据过滤后的流量中的IP、Session ID、设备ID、用户ID或者其它身份识别信息，将单个用户的应用流量区分出来并输入服务分类模块722(S1101)；服务分类模块722根据应用-服务规则将单个用户流量中每个应用的起始服务、独占服务和公共服务识别出来(S1102)，将识别结果送入流量归属模块730。识别过程可以是将当个用户流量中的报文特征提取出来，和应用-服务规则一一匹配，匹配成功则结束并输出匹配成功的那个规则所对应的服务类型和应用。

需要说明的是，在其他一些实施例中，单用户识别模块721及其执行过程不是必须的，比如流量本来就来源于一个用户，或者流量虽然来源于多个用户，但是对方案提出的要求中不包括区分不同用户的流量。

图12示出了将流量归属到应用的方法。该方法由流量归属模块730执行。该模块的输入为单用户下的服务识别结果，输出为公共服务流量的应用归属。

具体的，获取起始服务的位置（S1201），利用该位置将单个用户下的流量分段(S1202)；获取位置位于该分段（即当前分段）内的独占服务，得到该段流量所归属的应用(S1203)，该应用就是调用该独占服务的应用。然后建立缓存表，在缓存表中记录的信息包括该分段对应的应用ID、用户ID和起始服务的位置(S1204)。

为节省存储空间，缓存表中仅存储上一个分段和当前分段对应的应用ID、用户ID和起始服务的位置信息。

应理解的是，缓存表指的是以表格的形式存储在缓存中的一张表，在其它一些实施例中，这些信息也可以通过其它形式存储在其它存储空间中。

若前一个模块识别到公共服务，则获取识别到的公共服务的位置（S1205）。根据公共服务的位置判断是否属于当前分段（S1206），属于则输出该公共服务所归属的应用(S1207)；如果不属于当前分段，则用自己的位置信息在缓存表中查询对应位置的应用信息(S1208)，并输出该公共服务所归属的应用。或者，根据公共服务的位置直接在缓存表中查询对应位置的应用信息，并输出该公共服务所归属的应用。

需要说明的是，本实施例中某一项的ID指的是用于标识该项的信息，可以是数字、文本、代码或其它类型的信息。本实施例中某一种服务的位置指的是识别到该服务的时间，参考图6中表示服务的箭头的起始位置。

前述实施例提供的任意方法可以实现在一台或多台物理计算机上，前述实施例中提出的装置可以部署在一台或多台物理计算机上，装置内部的单元模块划分仅作为一种示例性的示出，各个单元模块可以部署在同一台物理计算机上，也可以部署在不同的物理计算机上。

请参考图13，为本实施例提供的一种计算机系统的逻辑结构示意图。该计算机系统可以是网络设备（例如DPI设备）、服务器、移动终端、个人计算机、车载计算机等任意一种类型的计算机系统。该计算机系统1300包括处理器1310、存储器1320以及网络接口1330（也被称为网卡或网络适配器等）等组件。该计算机系统可以与其他设备互联实现更多的功能，例如流量计费等。

处理器1310可以为单核处理器，也可以为多核处理器。当处理器1310为多核处理器时，本申请提供的方法可以运行在一个核上，也可以分布运行在不同的核上。处理器1310可以为一个，也可以为多个，多个处理器的类型可以相同或不相同。处理器的类型有中央处理器（central processing unit, CPU）、图形处理器、微处理器或协处理器等。

网络接口1330用于连接其他网络设备，包括无线连接和有线连接。在本实施例中，网络接口1330可以用于从网络上获取流量，以执行流量解析或分析。

存储器1320包括易失性和非易失性存储器，通常非易失性存储器上存储有本申请提供的流量分析装置1322和/或流量解析装置1321的计算机可读指令，还可以存储其他程序模块123（例如操作系统）的计算机可读指令。当这些计算机可读指令被处理器1310读取和运行之后可实现本申请前述实施例提供的任意一种或多种方法。流量分析装置1322和流量解析装置1321的具体实现可参考前述实施例。在其他实施例中，流量分析装置1322和流量解析装置1321可分别部署在不同的物理计算机上。

以上组件通过总线140连接。总线140可以是一条，也可以是多条。总线140包括高级微控制器总线（advance microcontroller bus architecture, AMBA）工业标准结构(industry standard architecture, ISA)总线，微通道结构(micro channel architecture, MCA)总线，扩展ISA(extended-ISA)总线，视频电子标准协会(video electronics standards association, VESA)局域总线，以及外围器件互联(peripheral component interconnect, PCI)总线等。

本申请提供的流量分析方法，不同于现有技术中仅针对应用识别的TLS握手方案，本申请提供了更细粒度的服务识别。在服务识别的过程中应用了报文的密文特征，提高了服务识别的准确度。相应的，在规则学习的过程中加入了密文特征的学习，利用密文

特征（例如application data数据包的长度、顺序、传输方向等）对服务识别的影响，构造特征矩阵，以及学习特征向量，最终生成应用-服务规则，增加了识别粒度，因此，解决了 TLS握手部分特征不足以区分和识别流量的问题。进一步的，本申请提供的流量分析方法将加密的HTTP会话部分的特征与TLS握手明文特征结合，通过将数值化特征和符号特征结合的自适应分箱方法，学习特征向量，用于识别应用或服务流量，提高识别准确率和精度。

本申请提供的公共流量归属方法，通过三种服务协同合作的方式解决归属问题，利用起始服务定位流量分段，利用独占服务获取应用标签，利用分段信息归属公共服务的流量，解决了公共服务的流量无法归属到应用的问题。

本申请还提供了基于最大进包数和流量的ASN域的过滤方法，减少需要分析的流量的大小，同时在规则生成过程中就考虑到效率问题，合并冗余规则，减少判断次数，因此，解决了规则复杂度过高，导致性能下降严重的问题。TLS握手规则需要对证书的全流程字段进行解析，消耗大量内存，不能做到针对单独字段进行精确匹配因此增加识别开销，需要优化解析的字段，并降低规则的复杂度。本申请提供的过滤方法效果在于根据识别规则提供的参数自适应的调整过滤策略，减少冗余规则对性能造成的影响，设计了过滤模块，减少读取次数和性能开销，改进了当前技术方案全字段特征建立规则的缺点，适配了高速实时的流量识别环境。

在主干核心网的高速环境中，对流量识别需要的报文数量有较大限制，因此本申请在阐述的过程中，并没有应用全流量特征，但是如果硬件技术进步或者任何特殊构造的环境能够支持这种特征学习方式，本申请能够自然的扩展到这种流量识别环境中，其识别的核心步骤依然与本申请前述实施例类似，即便存在不同之处本领域技术人员也容易想到。另外，由于TLS协议的任意包装，或者层级更低的TCP协议或人为构造的私有协议，可能会部分改变识别时的特征值，此方案仍然属于本申请的保护范围。

本申请提供的技术方案除可应用于运营商的策略和计费控制场景，还可以用于视频关键质量指标（key quality indicators, KQI）场景，例如内容分发网络（content delivery network，CDN）流量的区分，这种场景下公共流量产生的原因与前述实施例类似，基本可以按照前述实施例提供的方法去识别和区分CDN中不同应用使用的公共流量的归属，从而准确的完成视频KQI统计需求。更广泛的，任何由一个公共服务产生的公共流量需要区分的场景都可以适用本申请提供的方案。

需要说明的是，前述实施例中提出模块或单元的划分仅作为一种示例性的示出，所描述的各个模块的功能仅是举例说明，本申请并不以此为限。本领域普通技术人员可以根据需求合并其中两个或更多模块的功能，或者将一个模块的功能拆分从而获得更多更细粒度的模块，以及其他变形方式。

以上描述的各个实施例之间相同或相似的部分可相互参考。

以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理模块，即可以位于一个地方，或者也可以分布到多个网络模块上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。另外，本申请提供的装置实施例附图中，模块之间的连接关系表示它们之间具有通信连接，具体可以实现为一条或多条通信总线或信号线。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

以上所述，仅为本申请的一些具体实施方式，但本申请的保护范围并不局限于此。

权利要求

1.一种流量分析方法，其特征在于，包括：

获取流量中的报文的特征，所述特征包括密文特征，所述密文特征包括加密后的报文的顺序、长度、传输方向中的任意一个或多个；

根据所述特征对所述流量执行分析，以识别该流量所归属的服务或应用。

2.根据所述权利要求 1 所述的方法，其特征在于，所述加密后的报文包括以下类型的报文：应用数据 application data。

3.根据所述权利要求 1 或 2 所述的方法，其特征在于，所述根据所述特征对所述流量执行分析，以识别该流量所归属的服务或应用，包括：

将所述特征与服务的识别规则或应用的识别规则匹配，以识别所述流量所归属的服务或应用，所述服务的识别规则或应用的识别规则是基于所述特征，通过机器学习算法获得的。

4. 根据所述权利要求 1-3 任意一项所述的方法，其特征在于，所述特征还包括明文特征，所述明文特征包括所述报文中能够被直接解析得到的字符和/或数字组成的特征。

5. 根据所述权利要求 1-4 任意一项所述的方法，其特征在于，根据所述特征对所述流量执行分析，以识别该流量所归属的服务，包括：

将所述特征与第一识别规则、第二识别规则和第三识别规则分别匹配以识别所述流量中的起始服务、独占服务和公共服务，其中所述第一识别规则、所述第二识别规则和所述第三识别规则是基于所述特征，通过机器学习算法获得的，所述起始服务是一个应用启动阶段所调用的服务，所述独占服务是仅被一个应用调用的服务，所述公共服务是被两个或多个应用调用的服务。

6. 根据所述权利要求 5 所述的方法，其特征在于，还包括：

确定识别时间在起始服务 A 和起始服务 B 之间的公共服务的流量归属到一个应用，该应用为调用识别时间在所述起始服务 A 和所述起始服务 B 之间的独占服务的应用，所述起始服务 A 为识别到的任意一个起始服务，所述起始服务 B 为识别时间在所述起始服务 A 之后的第一个起始服务。

7.一种公共服务流量归属方法，其特征在于，包括：

获取流量中的报文的特征，所述特征包括密文特征，所述密文特征包括加密后的报文的顺序、长度、传输方向中的任意一个或多个；

根据所述特征对所述流量执行分析，识别该流量中的起始服务、独占服务和公共服务，其中所述起始服务是一个应用启动阶段所调用的服务，所述独占服务是仅被一个应用调用的服务，所述公共服务是被两个或多个应用调用的服务；

将识别时间在起始服务 A 和起始服务 B 之间的公共服务的流量归属到一个应用，该应用为调用识别时间在所述起始服务 A 和所述起始服务 B 之间的独占服务的应用，所述起始服务 A 为识别到的任意一个起始服务，所述起始服务 B 为识别时间在所述起始服务 A 之后的第一个起始服务。

8.根据所述权利要求 7 所述的方法，其特征在于，在获取所述特征之前，还包括：

根据识别规则确定所述流量分析所需的最大进包数量，所述识别规则为基于所述特

征，通过机器学习算法获得的，所述识别规则用于从所述流量中识别不同的服务；

基于所述最大进包数量对所述流量执行过滤。

9.根据所述权利要求 7 所述的方法，其特征在于，在获取所述特征之前，还包括：

根据所述流量的网络协议 IP 信息对所述流量执行过滤。

10.根据所述权利要求 7-9 任意一项所述的方法，其特征在于，根据所述特征对所述流量执行分析，识别该流量中的起始服务、独占服务和公共服务，包括：

将所述特征与第一识别规则、第二识别规则和第三识别规则分别匹配以识别所述流量中的所述起始服务、所述独占服务和所述公共服务，其中所述第一识别规则、所述第二识别规则和所述第三识别规则是基于所述特征，通过机器学习算法获得的。

11.根据所述权利要求 7-10 任意一项所述的方法，其特征在于，根据识别时间在第一起始服务和第二起始服务的识别时间之间的独占服务确定一个应用，包括：

根据所述独占服务和对应信息确定所述应用，所述对应信息包括所述独占服务和调用所述独占服务的应用的对应关系。

12. 根据所述权利要求 7-11 任意一项所述的方法，其特征在于，所述特征还包括明文特征，所述明文特征包括所述报文中能够被直接解析得到的字符和/或数字组成的特征。

13.一种公共服务流量归属方法，其特征在于，包括：

获取流量中的报文的特征，所述特征包括密文特征，所述密文特征包括加密后的报文的顺序、长度、传输方向中的任意一个或多个；

根据所述特征对所述流量执行分析，识别该流量中的独占服务和公共服务，其中所述独占服务是仅被一个应用调用的服务，所述公共服务是被两个或多个应用调用的服务；

将识别时间在独占服务 A 和独占服务 B 之间的公共服务的流量归属到一个应用，该应用为调用所述独占服务 A 的应用，所述独占服务 A 为识别到的任意一个独占服务，所述独占服务 B 包括识别时间在所述独占服务 A 之后的第一个独占服务。

14.根据所述权利要求 13 所述的方法，其特征在于，在获取所述特征之前，还包括：

根据识别规则确定执行所述流量分析所需的最大进包数量，所述识别规则为基于所述特征，通过机器学习算法获得的，所述识别规则用于从所述流量中识别不同的服务；

基于所述最大进包数量对所述流量执行过滤。

15.根据所述权利要求 13 所述的方法，其特征在于，在获取所述特征之前，还包括：

根据所述流量的网络协议 IP 信息对所述流量执行过滤。

16. 根据所述权利要求 13-15 任意一项所述的方法，其特征在于，根据所述特征对所述流量执行分析，识别该流量中的独占服务和公共服务，包括：

将所述特征与第二识别规则和第三识别规则分别匹配以识别所述流量中的所述独占服务和所述公共服务，其中所述第二识别规则和所述第三识别规则是基于所述特征组合，通过机器学习算法获得的。

17. 根据所述权利要求 13-16 任意一项所述的方法，其特征在于，所述特征还包括明文特征，所述明文特征包括所述报文中能够被直接解析得到的字符和/或数字组成的特征。

18. 一种计算机系统，其特征在于，包括存储器和处理器，所述存储器用于存储计算机可读指令，所述处理器用于读取所述计算机可读指令并实现如权利要求 1-6 任意一项所述的方法。

19. 一种计算机系统，其特征在于，包括存储器和处理器，所述存储器用于存储计算机可读指令，所述处理器用于读取所述计算机可读指令并实现如权利要求 7-12 任意一项所述的方法。

20. 一种计算机系统，其特征在于，包括存储器和处理器，所述存储器用于存储计算机可读指令，所述处理器用于读取所述计算机可读指令并实现如权利要求 13-17 任意一项所述的方法。

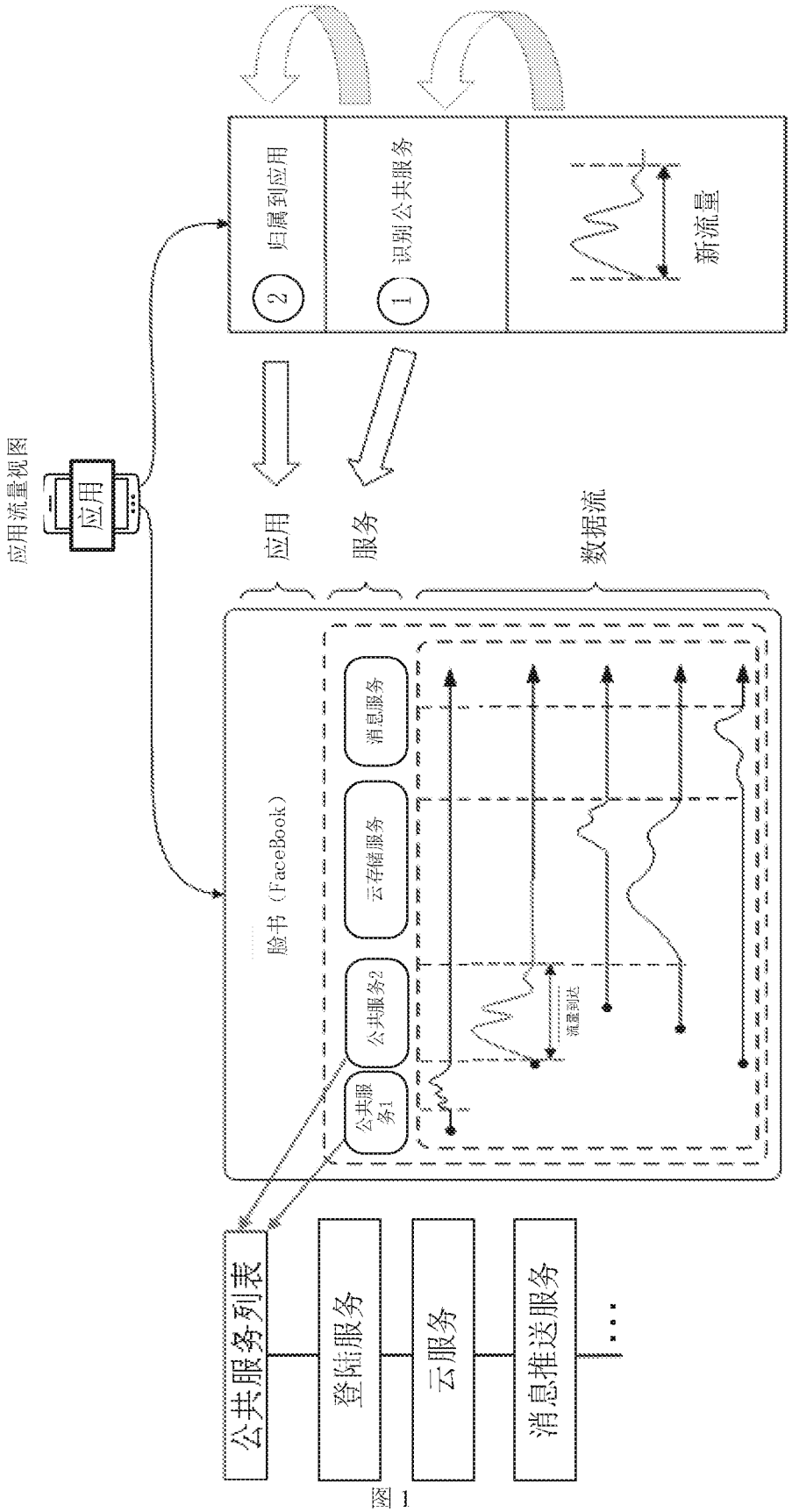
应用流量视图
应用
应用
服务
数据流
归属到应用
识别公共服务
新流量
脸书（FaceBook）
公共服务1
公共服务2
云存储服务
消息服务
流量到达
公共服务列表
登陆服务
云服务
消息推送服务
...

图 1

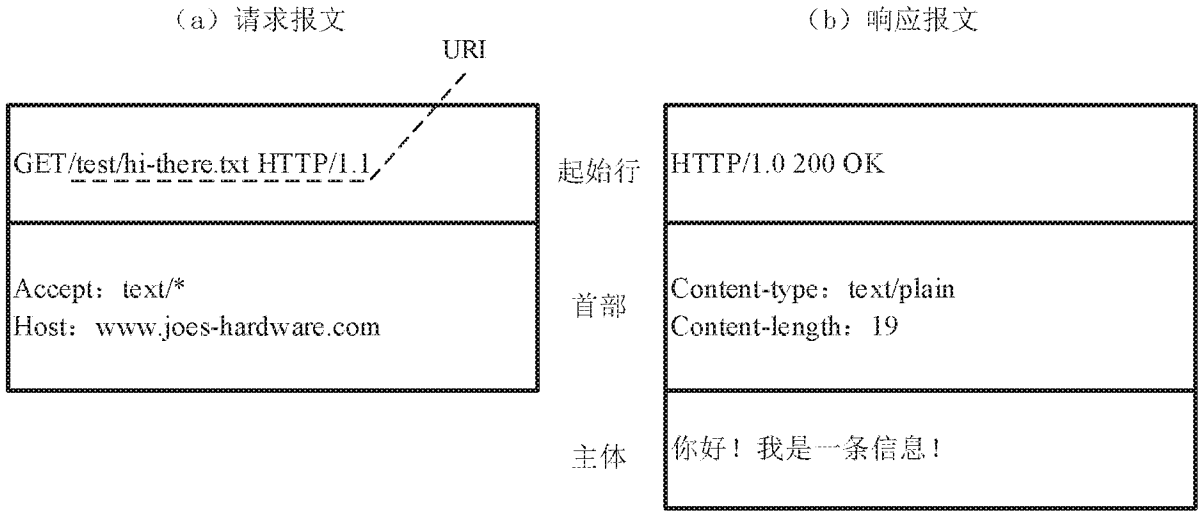
（a）请求报文
（b）响应报文
URI
GET/test/hi-there.txt HTTP/1.1
起始行
HTTP/1.0 200 OK
Accept：text/*
Host：www.joes-hardware.com
首部
Content-type：text/plain
Content-length：19
主体
你好！我是一条信息！

图 2


客户端
服务器
(1)客户端打招呼消息（ClientHello）
(2)服务器打招呼消息（ServerHello）
(3)证书链（Certificate）
(4)服务器打招呼结束（ServerHelloDone）
(5)客户端秘钥交换（ClientKeyExchange）
(6)交换密文（ChangeCipherSpec）
(7)结束（Finished）
(8)交换密文（ChangeCipherSpec）
(9)结束（Finished）

图 3

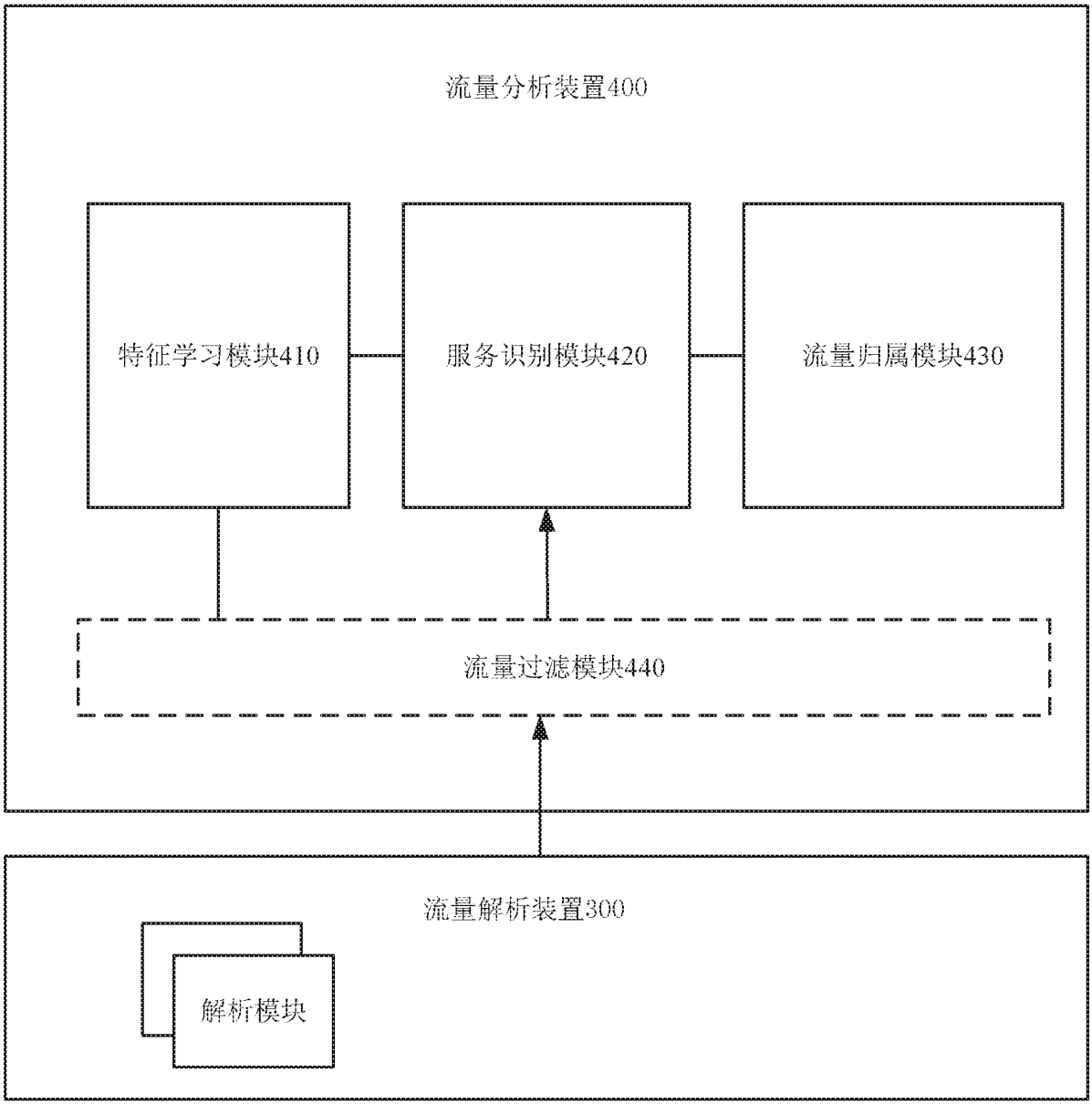
流量分析装置400
特征学习模块410
服务识别模块420
流量归属模块430
流量过滤模块440
流量解析装置300
解析模块

图 4

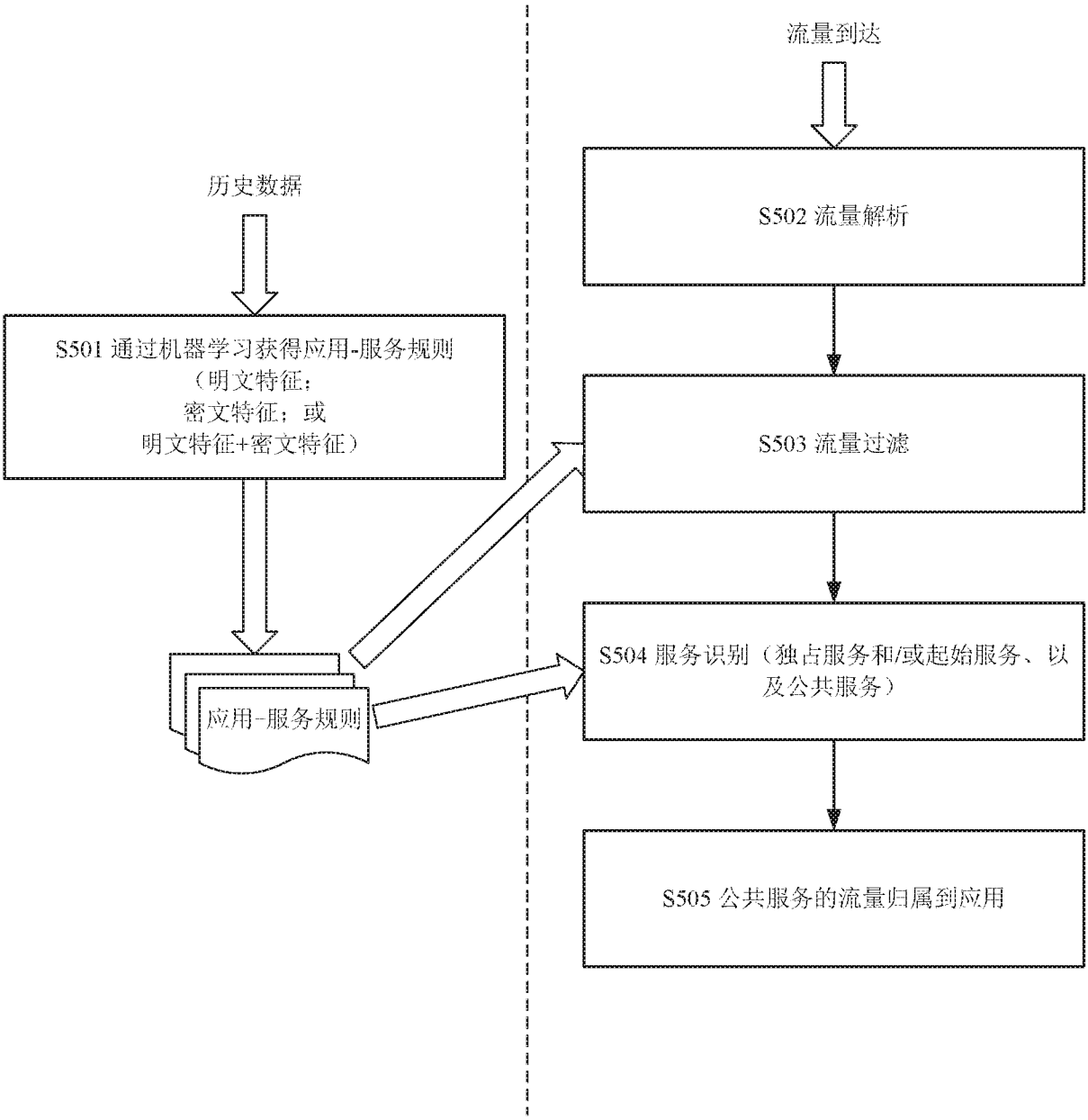
流量到达
历史数据
S502 流量解析
S501 通过机器学习获得应用-服务规则
（明文特征；
密文特征；或
明文特征+密文特征）
S503 流量过滤
S504 服务识别（独占服务和/或起始服务、以及公共服务）
应用-服务规则
S505 公共服务的流量归属到应用

图 5

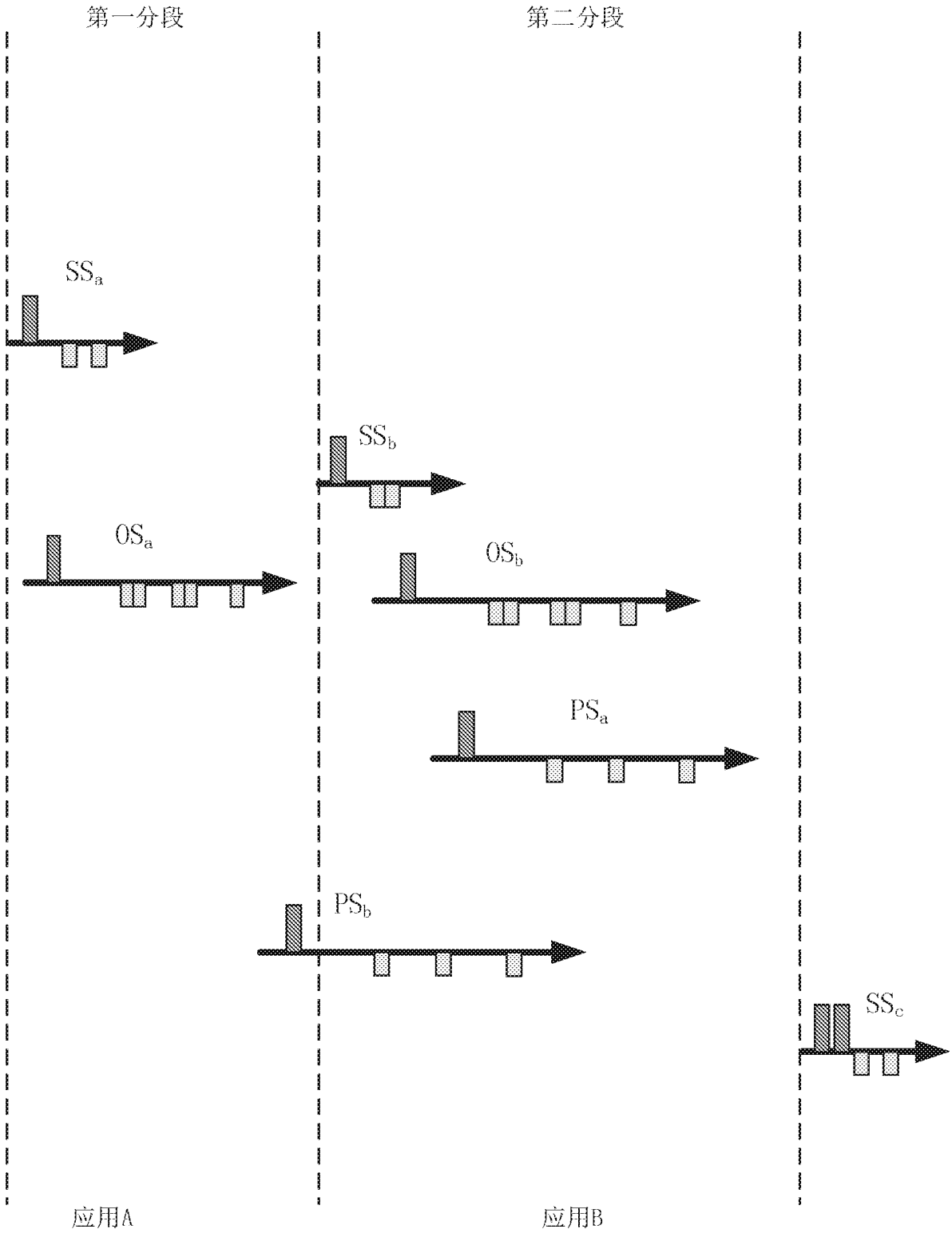

图 6

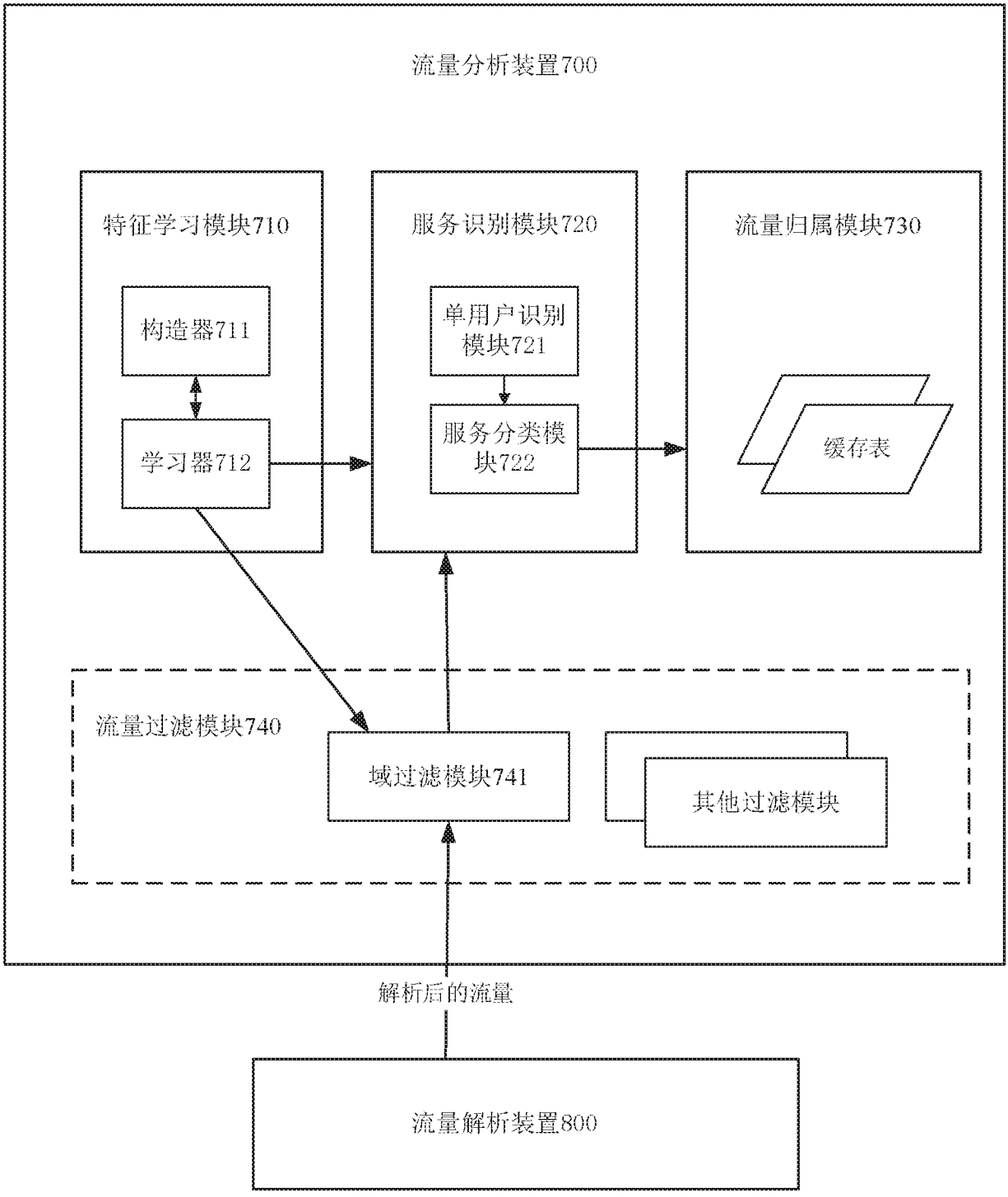
流量分析装置700
特征学习模块710
构造器711
学习器712
服务识别模块720
单用户识别模块721
服务分类模块722
流量归属模块730
缓存表
流量过滤模块740
域过滤模块741
其他过滤模块
解析后的流量
流量解析装置800

图 7

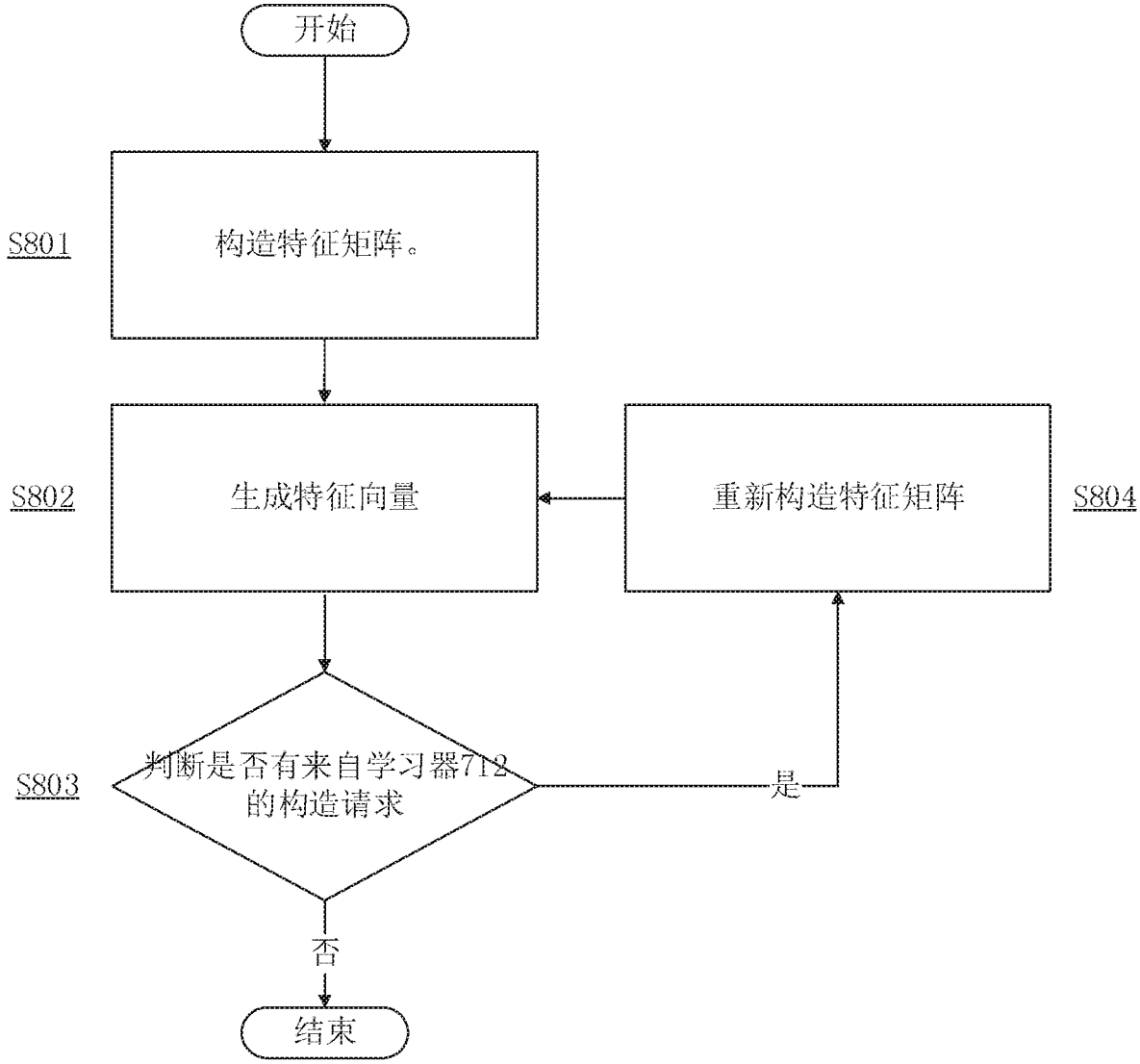
开始
S801
构造特征矩阵。
S802
生成特征向量
重新构造特征矩阵
S804
S803
判断是否有来自学习器712的构造请求
是
否
结束

图 8

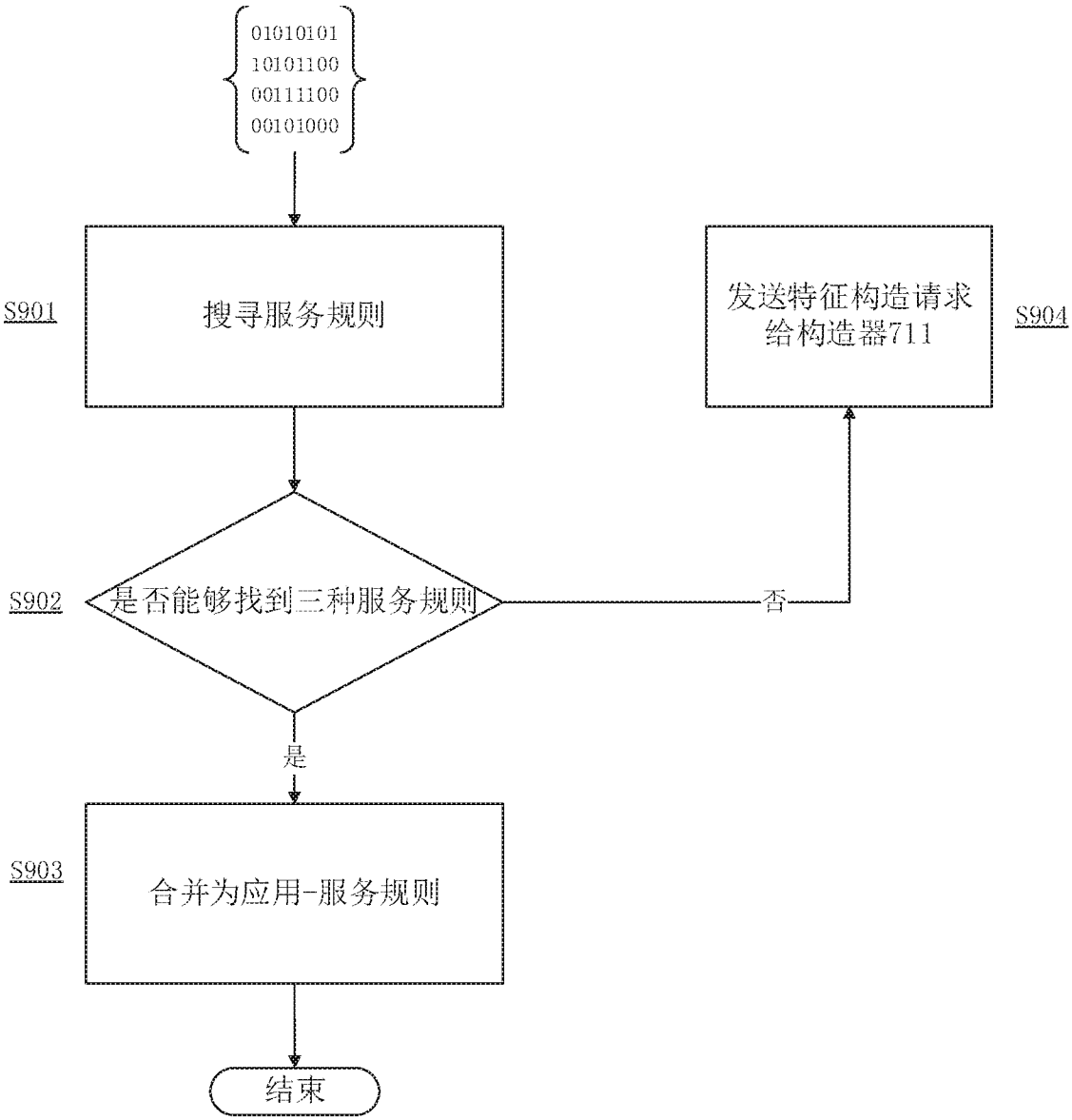
01010101
10101100
00111100
00101000
S901
搜寻服务规则
S902
是否能够找到三种服务规则
否
发送特征构造请求
给构造器711
S904
是
S903
合并为应用-服务规则
结束

图 9

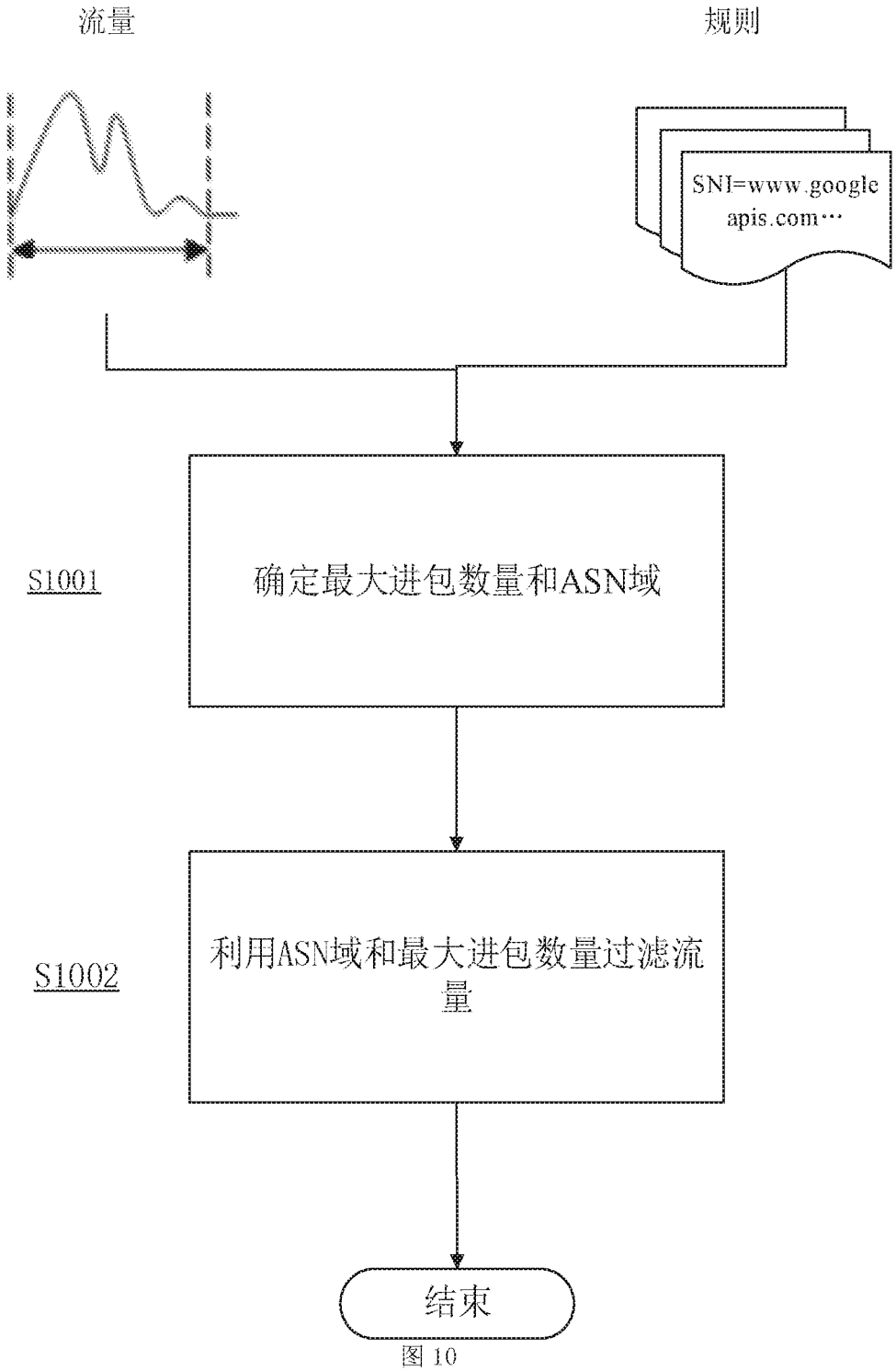
流量
规则
SNI=www.google
apis.com…
S1001
确定最大进包数量和ASN域
S1002
利用ASN域和最大进包数量过滤流量
结束

图 10

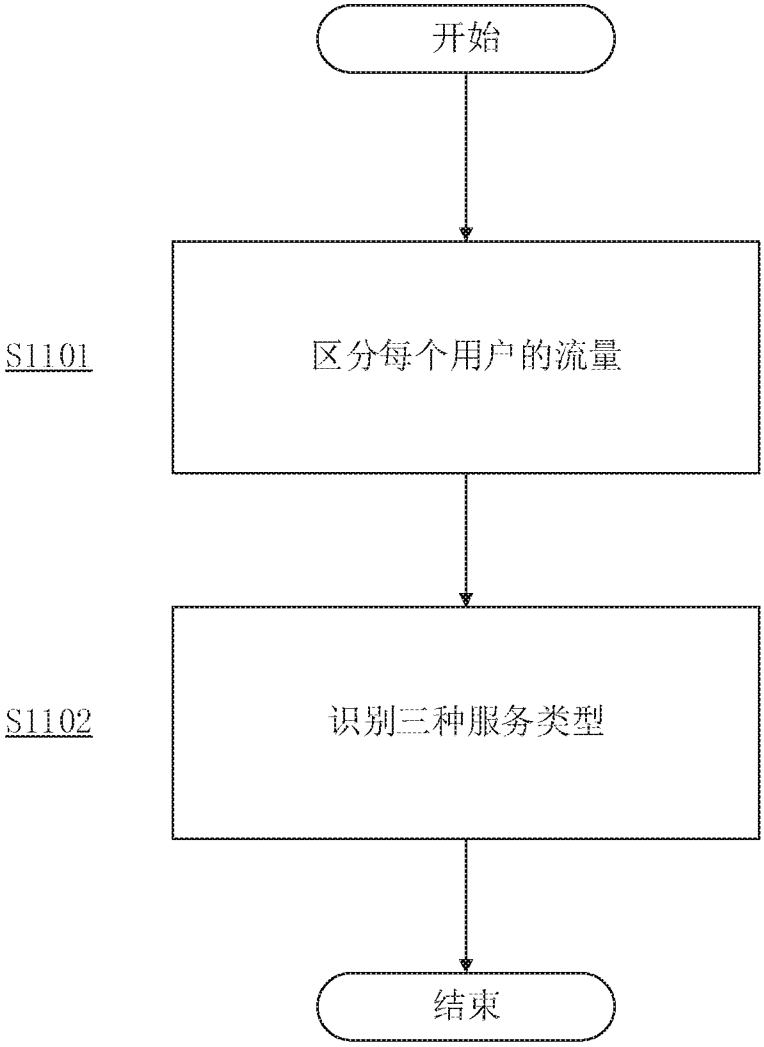

图 11

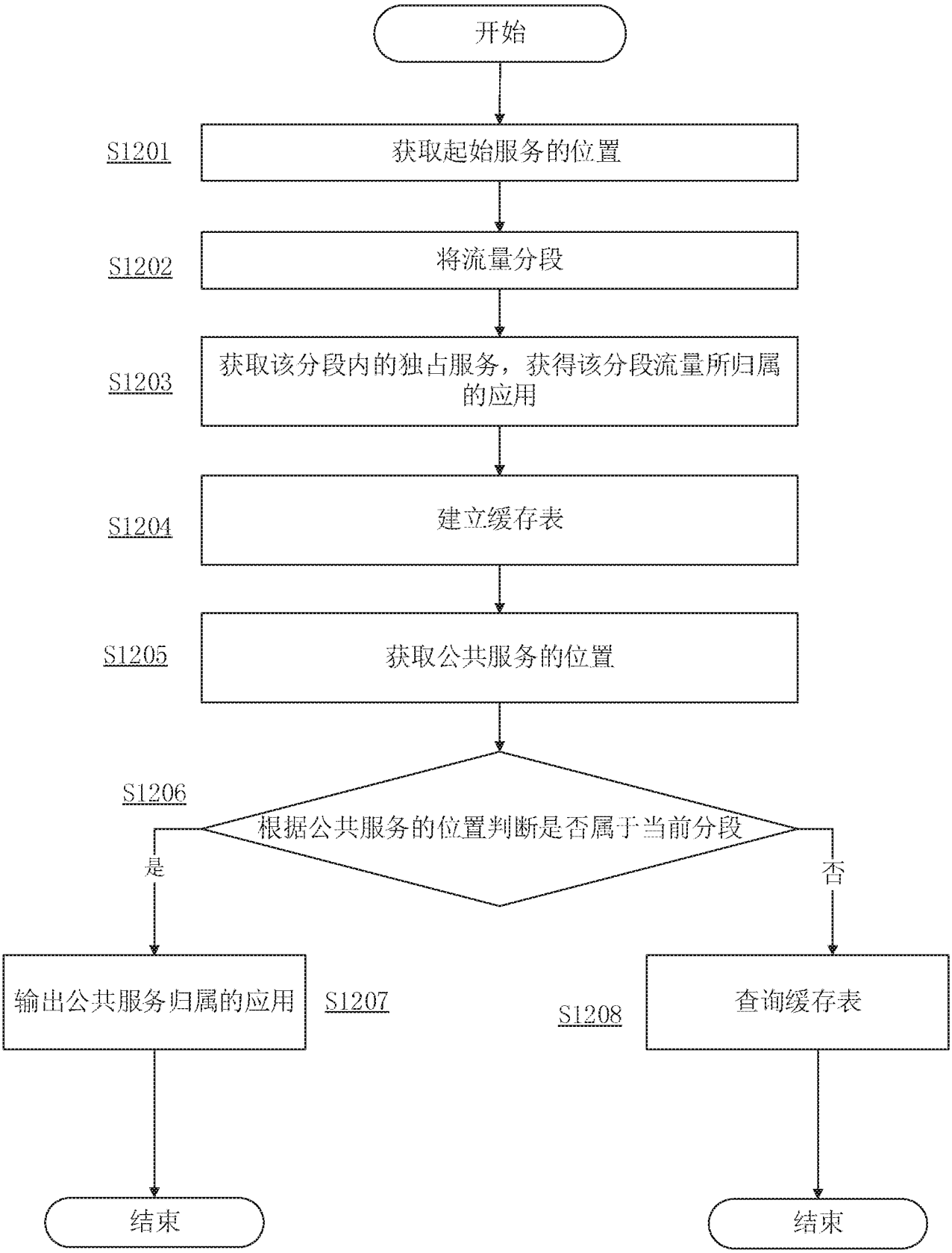
开始
S1201
获取起始服务的位置
S1202
将流量分段
S1203
获取该分段内的独占服务，获得该分段流量所归属的应用
S1204
建立缓存表
S1205
获取公共服务的位置
S1206
根据公共服务的位置判断是否属于当前分段
是
否
输出公共服务归属的应用
S1207
S1208
查询缓存表
结束
结束

图 12

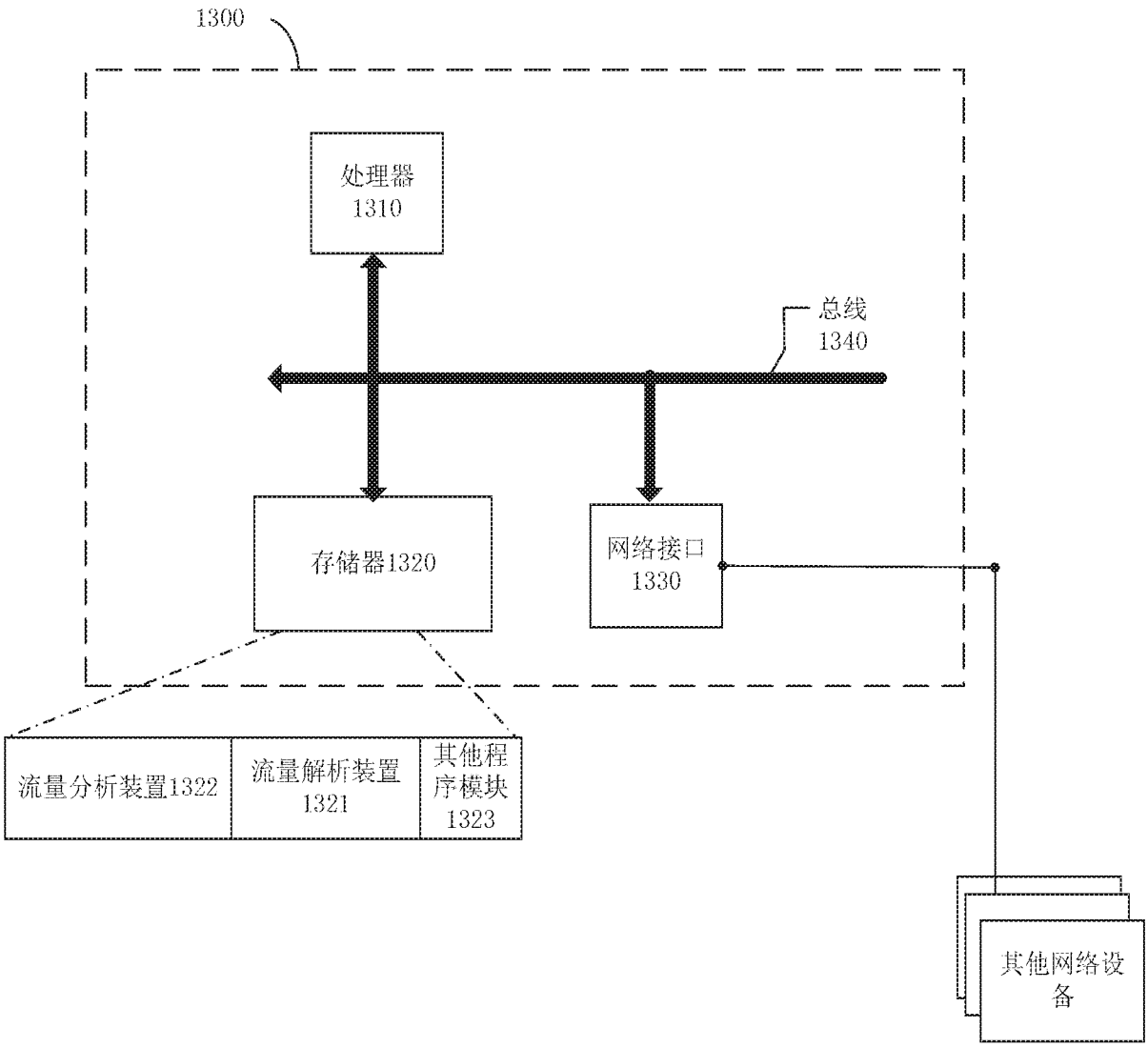
1300
处理器
1310
总线
1340
存储器1320
网络接口
1330
流量分析装置1322
流量解析装置
1321
其他程
序模块
1323
其他网络设
备

图 13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/120321

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/851(2013.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: 归属, 识别, 属于, 匹配, 数据包, 流量, 报文, 应用, 服务, 业务, 密文, 加密, 分析, 解析, attribut+, identi+, match, packet, message, traffic, application, APP, service

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 103873320 A (BEIJING TOPSEC SCIENCE & TECHNOLOGY CO., LTD.) 18 June 2014 (2014-06-18) description, paragraphs 9-24, 91-109, and 137	1-5
A	CN 105871832 A (BEIJING INSTITUTE OF TECHNOLOGY) 17 August 2016 (2016-08-17) entire document	1-20
A	CN 101505276 A (HANGZHOU H3C TECHNOLOGIES CO., LTD.) 12 August 2009 (2009-08-12) entire document	1-20
A	CN 102201982 A (BEIJING NETENTSEC, INC.) 28 September 2011 (2011-09-28) entire document	1-20
A	WO 2016201876 A1 (ZTE CORP.) 22 December 2016 (2016-12-22) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
07 January 2019	**31 January 2019**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/120321

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	103873320	A	18 June 2014	CN	103873320	B	13 June 2017
CN	105871832	A	17 August 2016	CN	105871832	B	02 November 2018
CN	101505276	A	12 August 2009	CN	101505276	B	01 June 2011
				CN	101505276	K1	12 August 2009
CN	102201982	A	28 September 2011		None		
WO	2016201876	A1	22 December 2016	CN	106257867	A	28 December 2016

国际检索报告

国际申请号 PCT/CN2018/120321

A. 主题的分类

H04L 12/851(2013.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

H04L

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT,CNKI,WPI,EPODOC:归属，识别，属于，匹配，数据包，流量，报文，应用，服务，业务，密文，加密，分析，解析，attribut+，identi+，match，packet，message，traffic，application，APP，service

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
X	CN 103873320 A (北京天融信科技有限公司) 2014年 6月 18日 (2014-06-18) 说明书第9-24，91-109，137段	1-5
A	CN 105871832 A (北京理工大学) 2016年 8月 17日 (2016-08-17) 全文	1-20
A	CN 101505276 A (杭州华三通信技术有限公司) 2009年 8月 12日 (2009-08-12) 全文	1-20
A	CN 102201982 A (北京网康科技有限公司) 2011年 9月 28日 (2011-09-28) 全文	1-20
A	WO 2016201876 A1 (ZTE CORP.) 2016年 12月 22日 (2016-12-22) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2019年 1月 7日	2019年 1月 31日

ISA/CN的名称和邮寄地址	受权官员
中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	蒋莉
传真号 (86-10)62019451	电话号码 (86-10) 53961751

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/120321

检索报告引用的专利文件			公布日（年/月/日）	同族专利			公布日（年/月/日）
CN	103873320	A	2014年 6月 18日	CN	103873320	B	2017年 6月 13日
CN	105871832	A	2016年 8月 17日	CN	105871832	B	2018年 11月 2日
CN	101505276	A	2009年 8月 12日	CN	101505276	B	2011年 6月 1日
				CN	101505276	K1	2009年 8月 12日
CN	102201982	A	2011年 9月 28日		无		
WO	2016201876	A1	2016年 12月 22日	CN	106257867	A	2016年 12月 28日