(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
2011 年 11 月 24 日 (24.11.2011)

PCT

(10) 国际公布号
WO 2011/143817 A1

(51) 国际专利分类号:
H04L 29/06 (2006.01) *H04L 12/24* (2006.01)

(21) 国际申请号: PCT/CN2010/072923

(22) 国际申请日: 2010 年 5 月 19 日 (19.05.2010)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人 (对除美国外的所有指定国): **阿尔卡特朗讯 (ALCATEL LUCENT)** [FR/FR]; 法国巴黎市波艾蒂耶大街 54 号, Paris F-75008 (FR)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **刘芳 (LIU, Fang)** [CN/CN]; 中国北京市博雅西园农大南路, Beijing 100012 (CN)。

(74) 代理人: **北京市金杜律师事务所 (KING & WOOD PRC LAWYERS)**; 中国北京市东三环中路 39 号建外 SOHOA 座 31 层, Beijing 100022 (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, 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。

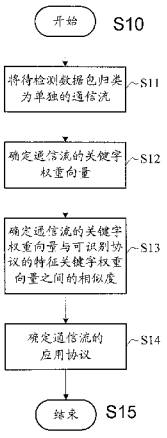
(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF,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

(54) Title: METHOD AND APPARATUS FOR IDENTIFYING APPLICATION PROTOCOL

(54) 发明名称 : 应用协议识别方法及装置



S10 START
S11 CLASSIFYING THE PACKETS TO BE DETECTED AS SEPARATE COMMUNICATION FLOWS
S12 DETERMINING THE KEYWORD WEIGHT VECTORS OF THE COMMUNICATION FLOWS
S13 DETERMINING THE SIMILARITY DEGREE BETWEEN THE KEYWORD WEIGHT VECTORS OF SAID COMMUNICATION FLOWS AND THE CHARACTERISTIC KEYWORD WEIGHT VECTORS OF THE IDENTIFIABLE PROTOCOLS
S14 DETERMINING THE APPLICATION PROTOCOLS OF SAID COMMUNICATION FLOWS
S15 END

图 1 / FIG. 1

(57) **Abstract**: The present invention involves a method and apparatus for identifying an application protocol. In one embodiment, the method for identifying the application protocol includes the following steps: S11. classifying the packets to be detected as separate communication flows; S12. searching the keywords in the payload of said communication flows based on the keyword database of the identifiable application protocols and determining the keyword weight vectors of said communication flows, wherein the weight of one keyword relates to the location of said keyword in the payload of one communication flow; S13. determining the similarity degree between said keyword weight vectors of said communication flows and the characteristic keyword weight vectors of the identifiable application protocols; S14. if the predefined condition is met, determining the application protocols corresponding to the characteristic keyword weight vectors which have the maximum similarity degree with the keyword weight vectors of said communication flows as the application protocols of said communication flows. Using said method and apparatus in the present invention, the accuracy of protocol identification can be improved without significant performance cost compared with the prior art.

[见续页]

WO 2011/143817 A1

(57) **摘要**:

本发明涉及应用协议识别方法及装置。在一个实施例中，识别应用协议的方法包括以下步骤：S11. 将待检测数据包归类到单独的通信流；S12. 基于可识别的应用协议的关键字库检索所述通信流的有效载荷中的关键字，并确定所述通信流的关键字权重向量；其中，一个关键字的权重与其在一个通信流的有效载荷中的位置有关；S13. 确定所述通信流的关键字权重向量与可识别的应用协议的特征关键字权重向量之间的相似度；S14. 如果满足预定条件，将与所述通信流的关键字权重向量之间具有最高相似度的特征关键字权重向量所对应的应用协议确定为所述通信流的应用协议。采用本发明中的方法及其装置，可以改善协议识别的精确性，同时相比于现有机制不引入显著的性能开销。

应用协议识别方法及装置

技术领域

本发明涉及网络通信技术，尤其涉及协议识别技术。

背景技术

本发明提出了一种在诸如 TCP/IP 网络的数据网络中，对由接入网关或其他设备接收到的数据中的应用层协议类型进行识别和分类的方法。

应用协议识别用于确定网络中所承载的数据的协议类型。这是一种提供网络流量的信息特征的非常重要的技术，其在多种不同场合都是必不可少的，例如有效的网络规划和设计、诸如合法监控和网络屏蔽的安全策略、诸如流量整型和服务区分的服务质量保障、以及收费策略设计、等。

当今的通信网络通常遵循分层模型，如 OSI 参考模型或 TCP/IP 参考模型。大多数数据网络采用 TCP/IP 参考模型，其包括五层：物理层、数据链路层、网络层、传输层和应用层。诸如接入网关的中继节点通常仅涉及 IP 层的传输和中继，而对于上层（传输层和应用层）承载的内容并不知晓。然而，在某些场合，例如屏蔽某种特定类型的应用，中继节点必须找到一个有效的途径来识别和确定应用层所承载的协议类型。

发明内容

通常的应用协议识别方案主要包括以下三类：

基于端口的协议识别

通过端口号进行协议分类是最简单、最传统的方法。该方法通过传输层报头所包含的端口号来识别协议或应用类型。对于标准协议，端口号与协议之间的关系由互联网号码分配机构（IANA）定义，例如，HTTP 协议通常使用端口 80，SMTP 协议使用端口 25。而对于私有协议，端口号通常由协议或应用自身定义。因为协议和端口

号之间的这种对应关系，通过应用层协议报头，例如 TCP 报头，中的源端口域和目的端口域的端口号，就可以确定协议类型。

尽管基于端口号的协议识别是高效而易于实施的，但其存在几个明显的缺点。1）在某些协议中，实际使用的端口号是在运行过程中动态分配的。2）防火墙的广泛使用使得某些端口比另外一些端口更容易穿越防火墙；因此，为了保证连接，无论是直接使用还是封装于原始协议之中，为私有协议使用流行的端口呈上升趋势。3）一些协议明确地使用非标准端口以避免被识别，甚至一些 P2P 应用允许用户改变默认端口，还有一些采用隧道技术和随机端口选择的组合以避免被检测。因此，仅仅基于端口号的协议识别已经不再可靠。

基于有效载荷的协议识别

基于有效载荷的分类是通过深度包检测（DPI）技术来检测承载于通信数据包中的协议的有效载荷。该方法包括：在应用层数据包中寻找确定的特征字符串，例如字符串‘http/1’对应于 HTTP 应用，字符串‘0xe319010000’对应于电驴应用。然而，单个特征码对于确定协议类型不够可靠，例如字符串‘http/1’也可以出现在 kazza 协议中。

为了改善匹配精确度，有人提出了采用正则表达式匹配来识别协议。正则表达式提供了灵活、强大的表达形式，能够提供高精确度的协议识别。在实际应用中通常使用确定有限自动机（DFA）来实现正则表达式。然而，取决于具体的模式，将一个正则表达式完全编译到一个确定有限自动机可能导致指数数量的 DFA 状态，这会降低性能。

基于行为的协议识别

基于行为的协议识别并不检测通信的内容，而是根据观察到的通信的行为或特征，例如数据包的大小、连接数量、等，来识别协议。一种通用的基于行为的协议识别就应用而言是采用统计特性来对通信进行识别和分类。例如一篇文献中提出了采用监督的机器学习来识别互联网通信流，分类的基本目标是通信流，其被表达为给定主机对之间的一个或多个数据包。每一个通信流具有若干描述其

行为的特征（参数）。这些参数构成了应用识别的输入鉴别器。例如，流周期、包到达时间、有效载荷尺寸、包到达时间的傅里叶变换、等，均可作为输入鉴别器。该文献中声称采用贝叶斯机器学习得到的识别率为 65%；当引入两个改进，即核密度估计（KDE）以及基于快速相关的滤波器（FCBF），识别率达到 95%。还有人提出了一些其他的基于统计特性的识别机制。

尽管基于行为的协议识别相比于基于内容的协议识别引起较少的性能开销，但其存在多方面的限制。1）基于行为的协议识别的精确度通常低于基于内容的协议识别。由于它主要依赖统计描述， 相对于基于有效载荷的确定方法，其识别精确度并不稳定。2）另外，观察到的行为总是依赖于外部环境，诸如网络类型、主机处理能力。例如，在一个无线局域网（WLAN）中的包到达间隔时间可能不同于以太网。3）恶意用户相对地比较容易通过修改上述的通信流参数，例如通过补零来改变包的长度、通过延迟来改变包到达时间，以此来规避识别。

为了克服现有技术中的上述缺陷，本发明提出了基于关键字向量匹配来实现应用协议识别的方法及装置。

在本发明的一个实施例中，提供了一种识别应用协议的方法。该方法包括以下步骤：A. 将待检测数据包归类到单独的通信流；B. 基于可识别的应用协议的关键字库检索所述通信流的有效载荷中的关键字，并确定所述通信流的关键字权重向量；其中，一个关键字的权重与其在一个通信流的有效载荷中的位置有关；C. 确定所述通信流的关键字权重向量与可识别的应用协议的特征关键字权重向量之间的相似度；D. 如果满足预定条件，将与所述通信流的关键字权重向量之间具有最高相似度的特征关键字权重向量所对应的应用协议确定为所述通信流的应用协议。

在本发明的另一个实施例中，识别应用协议的方法在所述步骤 A 之前还包括步骤：a. 基于可识别的应用协议的关键字库检索多个训练通信流的有效载荷中的关键字，并确定所述多个训练通信流的关键字权重向量；b. 根据所述多个训练通信流的关键字权重向量来确

定每一个可识别的应用协议所对应的特征关键字权重向量。

在本发明的一个实施例中，提供了一种识别应用协议的协议识别装置。该协议识别装置包括：第一装置，用于将待检测数据包归类到单独的通信流；第二装置，用于基于可识别的应用协议的关键字库检索所述通信流的有效载荷中的关键字，并确定所述通信流的关键字权重向量；其中，一个关键字的权重与其在一个通信流的有效载荷中的位置有关；第三装置，用于确定所述通信流的关键字权重向量与可识别的应用协议的特征关键字权重向量之间的相似度；第四装置，用于：如果满足预定条件，将与所述通信流的关键字权重向量之间具有最高相似度的特征关键字权重向量所对应的应用协议确定为所述通信流的应用协议。

在本发明的一个实施例中，提供了一种网络设备，其包括本发明上述的协议识别装置。

采用本发明中的方法及其装置，可以改善协议识别的精确性，同时相比于现有机制不引入显著的额外性能开销。

附图说明

参考下面的图和说明，将更好地理解该系统。图中的元件不一定按比例绘制，而是重点用于说明典型模型的原理。在图中，贯穿不同的示图，类似的参考标号表示对应的特征。

图 1 示出了根据本发明的一个实施例的识别应用协议的方法流程图；

图 2 示出了根据本发明的一个实施例的识别应用协议的方法中确定特征关键字权重向量的流程图；

图 3 示出了根据本发明的一个实施例的协议识别装置的结构图。

具体实施方式

不失一般性地，本发明中的以下实施例均是应用于数据通信网络，例如 Internet 网络，之中。

在本发明中，对于每一个需要识别的协议，将选择一组可能出

现在该协议中的关键字。例如但不限于，超文本传输协议的关键字包括“http/”，文件传输协议的关键字包括“ftp/”，等。接收到的数据包被根据五元组归类为单独的通信流，五元组包括：源地址、目的地址、源端口号、目的端口号、传输协议类型。对于每一个单独的通信流，检测其有效载荷中的关键字可以得到一个关键字权重向量，将该关键字权重向量和可识别协议的特征关键字权重向量进行匹配就可以实现对这一单独的通信流的协议的识别。

图 1 示出了根据本发明的一个实施例的识别应用协议的方法流程图。该方法通常应用于数据通信网络的网络设备之中。如图 1 所示，在该实施例中，识别应用协议的方法包括四个步骤 S11、S12、S13、S14。

在步骤 S11 中，将待检测数据包归类到单独的通信流。

在步骤 S12 中，基于可识别的应用协议的关键字库检索所述通信流的有效载荷中的关键字，并确定所述通信流的关键字权重向量；其中，一个关键字的权重与其在一个通信流的有效载荷中的位置有关。

具体地，例如但不限于，在可识别的应用协议的关键字库中共有 M 个不同的关键字，对于每一可识别协议，均有若干可能出现在该协议中的关键字 k_m ，$m \in \{1,\cdots,M\}$。例如但不限于，超文本传输协议的关键字包括“http/”，文件传输协议的关键字包括“ftp/”，等。例如，一个通信流 f_d，其关键字权重向量可以表示为 v_d。

在步骤 S13 中，确定所述通信流的关键字权重向量与可识别的应用协议的特征关键字权重向量之间的相似度。

具体地，例如但不限于，可识别的协议共有 N 个，每一可识别协议的特征关键字权重向量可以表示为 V_n ，$n \in \{1,\cdots,N\}$。

在步骤 S14 中，如果满足预定条件，将与所述通信流的关键字权重向量之间具有最高相似度的特征关键字权重向量所对应的应用协议确定为所述通信流的应用协议。

本领域技术人员应能理解，不同的可识别协议可以具有相同的关键字，也可以具有互不重叠的关键字；N 和 M 的取值没有必然的

大小关系。优选地，每一可识别应用协议具有至少一个独有的关键字，则 M≥N，这样可以提高不同协议之间的区分度。

在本发明的一个实施例中，具体地，在步骤 S11 中，待检测数据包将被根据五元组来归类，所述五元组包括：源地址、目的地址、源端口号、目的端口号、传输协议类型。需要理解的是，用于承载同一个应用程序在源（节点）和目的（节点）之间的数据的数据包，其源地址、目的地址、源端口号、目的端口号、传输协议类型应是相同的。通常，不同的应用程序的数据包之间源端口号和目的端口号不同；不同的源（节点）和目的（节点）之间的数据包之间的源地址和目的地址不同。因此，具有相同五元组内容的不同数据包应该属于一对源和目的之间的同一应用程序，将被归类到同一个单独的通信流。

在本发明的一个实施例中，具体地，步骤 S13 中的相似度包括余弦相似度。例如，通信流 f_d 的关键字权重向量 v_d 与任一可识别协议的特征关键字权重向量之间的余弦相似度将被分别计算，可以用公式表示为：

$$\cos(v_d, V_n) = \frac{v_d \cdot V_n}{\|v_d\| \times \|V_n\|} \quad 。$$

在本发明的一个实施例中，具体地，在步骤 S14 中，所述预定条件包括：所述最高相似度高于一个预定值。例如，通信流 f_d 的关键字权重向量 v_d 与特征关键字权重向量 V_2 的之间的余弦相似度最高，则当该最高相似度高于一个预定值时，V_2 所代表的应用协议将被确定为通信流 f_d 的应用协议；当最高相似度低于所述预定值时，则通信流 f_d 将被识别为未知协议，也就是通信流 f_d 被确定为不属于任一可识别协议。

在本发明的一个实施例中，具体地，步骤 S12 中，一个关键字在一个通信流的有效载荷中首次出现的位置越靠前，其权重越大。

通常，识别过程包括上述步骤 S11、S12、S13、S14，每一可识别协议的特征关键字权重向量应在识别过程之前的训练过程中确定。图 2 示出了根据本发明的一个实施例的识别应用协议的方法中确定特征关键字权重向量，也就是训练过程，的流程图。如图所示，

该训练过程包括两个步骤 S21、S22。

在步骤 S21 中，将基于可识别的应用协议的关键字库检索多个训练通信流的有效载荷中的关键字，并确定所述多个训练通信流的关键字权重向量。

在步骤 S22 中，将根据所述多个训练通信流的关键字权重向量来确定每一个可识别的应用协议所对应的特征关键字权重向量。具体地，因为每一训练通信流的应用协议都是预知的，可以对同属于一个应用协议的所有训练通信流的关键字权重向量取平均值，以确定该应用协议所对应的特征关键字权重向量。

在本发明的一个实施例中，具体地，步骤 S21 中，一个关键字在一个通信流的有效载荷中首次出现的位置越靠前，其权重越大。

更具体地，在本发明的一个实施例的步骤 S21 中，关键字的权重采用如下的 itp-idf（反向文本位置-反向文件频率）算法来检测，第 i 个关键字对于第 j 个训练通信流的权重表示如下：

$$\omega_{ij} = itp_{ij} \times idf_i \tag{1}$$

其中，itp_{ij}是反向文本位置的度量，idf_i是反向文件频率的度量。

$$itp_{ij} = 1/o_{ij} \tag{2}$$

其中，o_{ij}表示第 i 个关键字 k_i在第 j 个训练通信流中首次出现时的位置，该位置可以是比特位置或字节位置，itp_{ij}表示 o_{ij}的倒数；如果在第 j 个训练通信流中没有出现第 i 个关键字 k_i，则 itp_{ij}取值为零；

$$idf_i = \log \frac{|F|}{|\{f : k_i \in f\}|} \tag{3}$$

其中，$|F|$表示训练通信流的总数目，$|\{f : k_i \in f\}|$表示包含第 i 个关键字 k_i的训练通信流的数目；则第 j 个训练通信流的关键字权重向量可以表示为：

$$\mathrm{v}_j = (\omega_{1j}, \omega_{2j}, \cdots, \omega_{Mj}) \tag{4}$$

其中，M 表示关键字库中关键字的总数量。

在步骤 S22 中，将根据所述多个训练通信流的关键字权重向量来确定每一个可识别的应用协议所对应的特征关键字权重向量。对于某一个可识别的应用协议 p，从给定的训练通信流中可以计算出一个特征关键字权重向量，或者可以称为质心向量，表示如下：

$$V_p = \frac{1}{|F_p|}\sum_{v_j \in F_p} v_j \qquad (5)$$

其中，F_p表示属于协议 p 的训练通信流的集合，$|F_p|$表示属于协议 p 的训练通信流的数量，v_j表示集合F_p中的一个通信流的关键字权重向量。通过平均集合F_p中的各训练通信流的关键字权重向量中出现的各关键字的权重，就可以得到V_p。

在该实施例的步骤 S21 和 S22 中，基于上述的公式（1）至（5），可以很容易地基于训练通信流计算出每一可识别的应用协议的特征关键字权重向量。

本领域技术人员应能理解，训练通信流也是将大量训练数据包根据五元组归类而成的；在训练通信流中，应包括每一可识别应用程序的若干通信流。

本领域技术人员应能理解，上述训练过程，也就是步骤 S21 和步骤 S22，可以定期地或不定期地重复执行，每次执行时针对更新的训练通信流，以更新可识别应用协议的特征关键字权重向量。在某一次更新的训练过程中，可以引入新的可识别应用协议；则相应地，关键字库中的关键字数量可能会增加，更新的训练通信流还应包括新引入的可识别应用协议的若干通信流，更新结果还包括新引入的可识别应用协议的特征关键字权重向量。

本领域技术人员应能理解，在识别过程中的步骤 S12 中，将采用与训练过程中的步骤 S21 中同样的算法来确定通信流的关键字权重向量。

在本发明的一个实施例中，具体地，在步骤 S12 中，通信流f_d的关键字权重向量v_d将采用如下的 itp-idf 算法来检测，第 i 个关键字对于通信流f_d的权重表示如下：$\omega_{id} = itp_{id} \times idf_i$。其中，$itp_{id} = 1/o_{id}$；其中，$o_{id}$表示第 i 个关键字$k_i$在通信流$f_d$中首次出现时的位置，该位置可以是比特位置或字节位置，itp_{id}表示o_{id}的倒数；如果在通信流f_d中没有出现第 i 个关键字k_i，则itp_{id}取值为零。其中，

$$idf_i = \log\frac{|F|}{|\{f : k_i \in f\}|},$$

其中，$|F|$表示训练通信流的总数目，$|\{f : k_i \in f\}|$表示包含第 i 个关键字

k_i的训练通信流的数目。则通信流f_d的关键字权重向量可以表示为：$v_d=(\omega_{1d},\omega_{2d},\cdots,\omega_{Md})$，其中，M表示关键字库中关键字的总数量。

对于一个通信流而言，通常，对于识别该通信流的协议具有重要作用的关键字位于其有效载荷的头部。在本发明的一个实施例中，具体地，在步骤S12或步骤S21中，一个通信流或训练通信流的关键字权重向量将基于其有效载荷头部预定长度的内容来确定，关键字的检索和权重计算仅在该有效载荷头部预定长度的内容中进行即可。例如但不限于，该预定长度为128字节、或256字节。本领域技术人员应能理解，该有效载荷头部预定长度的内容可以容纳在一个或多个数据包中。

图3示出了根据本发明的一个实施例的协议识别装置的结构图。如图所示，该实施例中的协议识别装置10包括：第一装置11、第二装置12、第三装置13、第四装置14。该协议识别装置10通常设置于数据通信网络中的网络设备之中。

第一装置11，用于将待检测数据包归类到单独的通信流。

第二装置12，用于基于可识别的应用协议的关键字库检索所述通信流的有效载荷中的关键字，并确定所述通信流的关键字权重向量；其中，一个关键字的权重与其在一个通信流的有效载荷中的位置有关。

具体地，例如但不限于，在可识别的应用协议的关键字库中共有M个不同的关键字，对于每一可识别协议，均有若干可能出现在该协议中的关键字k_m ，$m\in\{1,\cdots,M\}$。例如但不限于，超文本传输协议的关键字包括“http/”，文件传输协议的关键字包括“ftp/”，等。例如，一个通信流f_d，其关键字权重向量可以表示为v_d。

第三装置13，用于确定所述通信流的关键字权重向量与可识别的应用协议的特征关键字权重向量之间的相似度。

具体地，例如但不限于，可识别的协议共有N个，每一可识别协议的特征关键字权重向量可以表示为V_n ，$n\in\{1,\cdots,N\}$。

第四装置14，用于：如果满足预定条件，将与所述通信流的关键字权重向量之间具有最高相似度的特征关键字权重向量所对应的应用协议确定为所述通信流的应用协议。

本领域技术人员应能理解，不同的可识别协议可以具有相同的关键字，也可以具有互不重叠的关键字；N 和 M 的取值没有必然的大小关系。优选地，每一可识别应用协议具有至少一个独有的关键字，则 M≥N，这样可以提高不同协议之间的区分度。

在本发明的一个实施例中，具体地，第一装置 11 根据五元组将待检测数据包归类为单独的通信流，所述五元组包括：源地址、目的地址、源端口号、目的端口号、传输协议类型。需要理解的是，用于承载同一个应用协议在源（节点）和目的（节点）之间的数据的数据包，其源地址、目的地址、源端口号、目的端口号、传输协议类型应是相同的。通常，不同的应用协议的数据包之间源端口号和目的端口号不同；不同的源（节点）和目的（节点）之间的数据包的源地址和目的地址不同。因此，具有相同五元组内容的不同数据包应该属于一对源和目的之间的同一应用协议，将被归类到同一个单独的通信流。

在本发明的一个实施例中，具体地，第三装置 13 确定的相似度包括余弦相似度。例如，通信流 f_d 的关键字权重向量 v_d 与任一可识别协议的特征关键字权重向量之间的余弦相似度将被分别计算，可以用公式表示为：

$$\cos(v_d, V_n) = \frac{v_d \cdot V_n}{\|v_d\| \times \|V_n\|} \quad 。$$

在本发明的一个实施例中，具体地，第四装置 14 中的所述预定条件包括：所述最高相似度高于一个预定值。例如，通信流 f_d 的关键字权重向量 v_d 与特征关键字权重向量 V_2 的之间的余弦相似度最高，则当该最高相似度高于一个预定值时，V_2 所代表的应用协议将被确定为通信流 f_d 的应用协议；当最高相似度也低于所述预定值时，则通信流 f_d 将被识别为未知协议，也就是通信流 f_d 被第四装置 14 确定为不属于任一可识别协议。

在本发明的一个实施例中，具体地，第二装置 12 中，一个关键字在一个通信流的有效载荷中首次出现的位置越靠前，其权重越大。

通常，识别过程包括上述第一装置 11、第二装置 12、第三装置 13、第四装置 14 所分别执行的操作，每一可识别协议的特征关键字

权重向量应在识别过程之前的训练过程中预先确定。

该训练过程由第二装置 12 执行，包括两个子操作 21、22（图中未示出）。

在子操作 21 中，第二装置 12 将基于可识别的应用协议的关键字库检索多个训练通信流的有效载荷中的关键字，并确定所述多个训练通信流的关键字权重向量。

在子操作 22 中，第二装置 12 将根据所述多个训练通信流的关键字权重向量来确定每一个可识别的应用协议所对应的特征关键字权重向量。具体地，因为每一训练通信流的应用协议都是预知的，可以对同属于一个应用协议的所有训练通信流的关键字权重向量取平均值，以确定该应用协议所对应的特征关键字权重向量。

在本发明的一个实施例中，具体地，第二装置 12 所执行的子操作 21 中，一个关键字在一个通信流的有效载荷中首次出现的位置越靠前，其权重越大。

更具体地，在本发明的一个实施例中，第二装置 12 所执行的子操作 21 中，关键字的权重采用如下的 itp-idf（反向文本位置-反向文件频率）算法来检测，第 i 个关键字对于第 j 个训练通信流的权重表示如下：

$$\omega_{ij} = itp_{ij} \times idf_i \qquad (1)$$

其中，itp_{ij} 是反向文本位置的度量，idf_i 是反向文件频率的度量。

$$itp_{ij} = 1/o_{ij} \qquad (2)$$

其中，o_{ij} 表示第 i 个关键字 k_i 在第 j 个训练通信流中首次出现时的位置，该位置可以是比特位置或字节位置，itp_{ij} 表示 o_{ij} 的倒数；如果在第 j 个训练通信流中没有出现第 i 个关键字 k_i，则 itp_{ij} 取值为零；

$$idf_i = \log \frac{|F|}{|\{f : k_i \in f\}|} \qquad (3)$$

其中，$|F|$ 表示训练通信流的总数目，$|\{f : k_i \in f\}|$ 表示包含第 i 个关键字 k_i 的训练通信流的数目；则第 j 个训练通信流的关键字权重向量可以表示为：

$$\mathrm{v}_j = (\omega_{1j}, \omega_{2j}, \cdots, \omega_{Mj}) \qquad (4)$$

其中，M 表示关键字库中关键字的总数量。

在子操作 22 中，第二装置 12 将根据所述多个训练通信流的关键字权重向量来确定每一个可识别的应用协议所对应的特征关键字权重向量。对于某一个可识别的应用协议 p，从给定的训练通信流中可以计算出一个特征关键字权重向量，或者可以称为质心向量，表示如下：

$$V_p = \frac{1}{|F_p|} \sum_{v_j \in F_p} v_j \qquad (5)$$

其中，F_p表示属于协议 p 的训练通信流的集合，$|F_p|$表示属于协议 p 的训练通信流的数量，v_j表示集合F_p中的一个通信流的关键字权重向量。通过平均集合F_p中的各训练通信流的关键字权重向量中出现的各关键字的权重，就可以得到V_p。

在该实施例中，第二装置 12 所执行的子操作 21 和 22 中，基于上述的公式（1）至（5），可以很容易地基于训练通信流计算出每一可识别的应用协议的特征关键字权重向量。

本领域技术人员应能理解，训练通信流也是将大量训练数据包根据五元组归类而成的；在训练通信流中，应包括每一可识别应用程序的若干通信流。

本领域技术人员应能理解，上述训练过程，也就是第二装置 12 所执行的子操作 21 和子操作 22，可以定期地或不定期地重复执行，每次执行时针对更新的训练通信流，以更新可识别应用协议的特征关键字权重向量。在某一次更新的训练过程中，可以引入新的可识别应用协议；则相应地，关键字库中的关键字数量可能会增加，更新的训练通信流还应包括新引入的可识别应用协议的若干通信流，更新结果还包括新引入的可识别应用协议的特征关键字权重向量。

本领域技术人员应能理解，在识别过程中，第二装置 12 将采用与训练过程中的子操作 21 中同样的算法来确定通信流的关键字权重向量。

在本发明的一个实施例中，具体地，识别过程中，第二装置 12 将采用如下的 itp-idf 算法来通信流f_d的关键字权重向量v_d，第 i 个关键字对于通信流f_d的权重表示如下：$\omega_{id} = itp_{id} \times idf_i$。其中，$itp_{id} = 1/o_{id}$；其中，$o_{id}$表示第 i 个关键字$k_i$在通信流$f_d$中首次出现时的位置，该位

置可以是比特位置或字节位置，itp_{id}表示o_{id}的倒数；如果在通信流f_d中没有出现第i个关键字k_i，则itp_{id}取值为零。其中，

$$idf_i = \log \frac{|F|}{|\{f : k_i \in f\}|},$$

其中，$|F|$表示训练通信流的总数目，$|\{f : k_i \in f\}|$表示包含第i个关键字k_i的训练通信流的数目。则通信流f_d的关键字权重向量可以表示为：$v_d = (\omega_{1d}, \omega_{2d}, \cdots, \omega_{Md})$，其中，M表示关键字库中关键字的总数量。

对于一个通信流而言，通常，对于识别该通信流的协议具有重要作用的关键字位于其有效载荷的头部。在本发明的一个实施例中，具体地，第二装置11在识别过程中的关键字权重向量的确认操作或训练过程中的子操作21中，一个通信流或训练通信流的关键字权重向量将基于其有效载荷头部预定长度的内容来确定。第二装置12仅在该有效载荷头部预定长度的内容中进行关键字的检索和权重计算即可。例如但不限于，该预定长度为128字节、或256字节。本领域技术人员应能理解，该有效载荷头部预定长度的内容可以容纳在一个或多个数据包中。

在本发明的一个实施例中，提供了设置有上述协议识别装置10的网络设备。该网络设备例如但不限于：交换机、路由器、网关、终端机。

本领域技术人员应能理解，本发明中所称的装置，可以包括：软件功能模块、硬件模块、或两者的结合。

本领域技术人员应能理解，上述实施例均是示例性而非限制性的。在不同实施例中出现的不同技术特征可以进行组合以取得有益效果。本领域技术人员在研究附图、说明书及权利要求书的基础上，应能理解并实现所揭示的实施例的其他变化的实施例。在权利要求书中，术语“包括”并不排除其他装置或步骤；不定冠词“一个”不排除多个；术语“第一”、“第二”用于标示名称而非用于表示任何特定的顺序。权利要求中的任何附图标记均不应被理解为对保护范围的限制。权利要求中出现的多个部分的功能可以由一个单独的硬件或软件模块来实现。某些技术特征出现在不同的从属权利要求中并不意味着不能将这些技术特征进行组合以取得有益效果。

权 利 要 求 书

1. 一种识别应用协议的方法，包括以下步骤：

A. 将待检测数据包归类到单独的通信流；

B. 基于可识别的应用协议的关键字库检索所述通信流的有效载荷中的关键字，并确定所述通信流的关键字权重向量；其中，一个关键字的权重与其在一个通信流的有效载荷中的位置有关；

C. 确定所述通信流的关键字权重向量与可识别的应用协议的特征关键字权重向量之间的相似度；

D. 如果满足预定条件，将与所述通信流的关键字权重向量之间具有最高相似度的特征关键字权重向量所对应的应用协议确定为所述通信流的应用协议。

2. 根据权利要求 1 所述的方法，其中，所述步骤 B 中，一个关键字在一个通信流的有效载荷中首次出现的位置越靠前，其权重越大。

3. 根据权利要求 2 所述的方法，其中，通信流 f_d 的关键字权重向量中的第 i 个关键字的权重表示为 $\omega_{id} = itp_{id} \times idf_i$；

其中，$itp_{id} = 1/o_{id}$；o_{id} 表示第 i 个关键字在通信流 f_d 中首次出现时的位置；

其中，$idf_i = \log \frac{|F|}{|\{f : k_i \in f\}|}$；$|F|$ 表示通信流的总数目，$|\{f : k_i \in f\}|$ 表示包含第 i 个关键字 k_i 的通信流的数目。

4. 根据权利要求 1 至 3 中任一项所述的方法，其中，在所述步骤 A 之前还包括步骤：

a. 基于可识别的应用协议的关键字库检索多个训练通信流的有效载荷中的关键字，并确定所述多个训练通信流的关键字权重向量；

b. 根据所述多个训练通信流的关键字权重向量来确定每一个可识别的应用协议所对应的特征关键字权重向量。

5. 根据权利要求 4 所述的方法，其中，所述步骤 a 中采用与所述步骤 B 中同样的算法来确定训练通信流的关键字权重向量；

其中，第 j 个通信流的关键字权重向量中的第 i 个关键字的权重

表示为 $\omega_{ij} = itp_{ij} \times idf_i$；

其中，$itp_{ij} = 1/o_{ij}$；o_{ij}表示第 i 个关键字在第 j 个通信流中首次出现时的位置；

其中，$idf_i = \log \frac{|F|}{|\{f : k_i \in f\}|}$；$|F|$表示通信流的总数目，$|\{f : k_i \in f\}|$表示包含第 i 个关键字 k_i 的通信流的数目。

6. 根据权利要求 1 至 3 中任一项所述的方法，其中，所述步骤 A 中根据所述待检测数据包的五元组来归类，所述五元组包括：源地址、目的地址、源端口号、目的端口号、传输协议类型；所述步骤 C 中的相似度包括余弦相似度。

7. 根据权利要求 1 至 3 中任一项所述的方法，其中，所述预定条件包括：所述最高相似度高于一个预定值。

8. 一种识别应用协议的协议识别装置，包括：

第一装置，用于将待检测数据包归类到单独的通信流；

第二装置，用于基于可识别的应用协议的关键字库检索所述通信流的有效载荷中的关键字，并确定所述通信流的关键字权重向量；其中，一个关键字的权重与其在一个通信流的有效载荷中的位置有关；

第三装置，用于确定所述通信流的关键字权重向量与可识别的应用协议的特征关键字权重向量之间的相似度；

第四装置，用于：如果满足预定条件，将与所述通信流的关键字权重向量之间具有最高相似度的特征关键字权重向量所对应的应用协议确定为所述通信流的应用协议。

9. 根据权利要求 8 所述的协议识别装置，其中，所述第二装置中，一个关键字在一个通信流的有效载荷中首次出现的位置越靠前，其权重越大。

10. 根据权利要求 9 所述的协议识别装置，其中，通信流 f_d 的关键字权重向量中的第 i 个关键字的权重表示为 $\omega_{id} = itp_{id} \times idf_i$；

其中，$itp_{id} = 1/o_{id}$；o_{id}表示第 i 个关键字在通信流 f_d 中首次出现时的位置；

其中，$idf_i = \log \frac{|F|}{|\{f : k_i \in f\}|}$；$|F|$表示通信流的总数目，$|\{f : k_i \in f\}|$表示包含第 i 个关键字 k_i 的通信流的数目。

11. 根据权利要求 8 至 10 中任一项所述的协议识别装置，其中，所述第二装置还用于：

- 基于可识别的应用协议的关键字库检索多个训练通信流的有效载荷中的关键字，并确定所述多个训练通信流的关键字权重向量；

- 根据所述多个训练通信流的关键字权重向量来确定每一个可识别的应用协议所对应的特征关键字权重向量。

12. 根据权利要求 11 所述的协议识别装置，其中，所述第二装置采用同样的算法来确定训练通信流以及待检测数据包的通信流的关键字权重向量；

其中，第 j 个通信流的关键字权重向量中的第 i 个关键字的权重表示为 $\omega_{ij} = itp_{ij} \times idf_i$；

其中，$itp_{ij} = 1/o_{ij}$；o_{ij}表示第 i 个关键字在第 j 个通信流中首次出现时的位置；

其中，$idf_i = \log \frac{|F|}{|\{f : k_i \in f\}|}$；$|F|$表示通信流的总数目，$|\{f : k_i \in f\}|$表示包含第 i 个关键字 k_i 的通信流的数目。

13. 根据权利要求 8 至 10 中任一项所述的协议识别装置，其中，所述第一装置根据所述待检测数据包的五元组来归类，所述五元组包括：源地址、目的地址、源端口号、目的端口号、传输协议类型；所述第三装置中的相似度包括余弦相似度。

14. 根据权利要求 8 至 10 中任一项所述的协议识别装置，其中，所述预定条件包括：所述最高相似度高于一个预定值。

15. 一种网络设备，其包括权利要求 8 至 14 中任一项所述的协议识别装置。

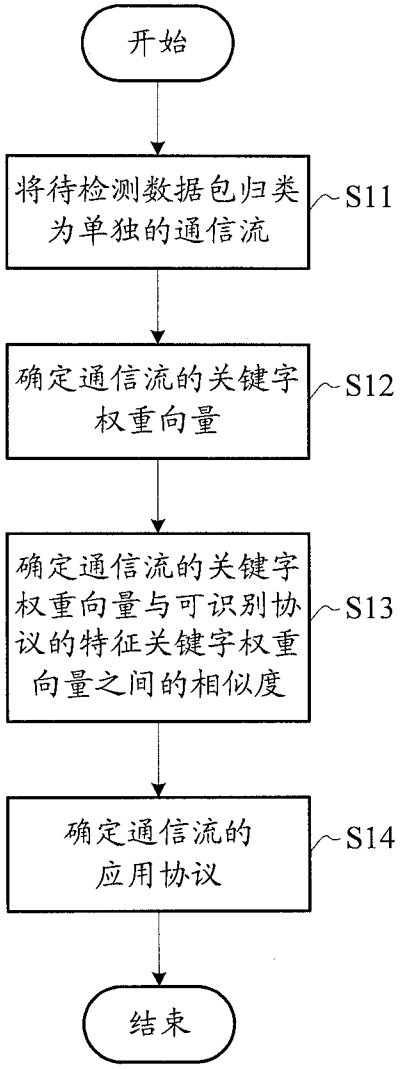

开始
将待检测数据包归类为单独的通信流
S11
确定通信流的关键字权重向量
S12
确定通信流的关键字权重向量与可识别协议的特征关键字权重向量之间的相似度
S13
确定通信流的应用协议
S14
结束

图 1

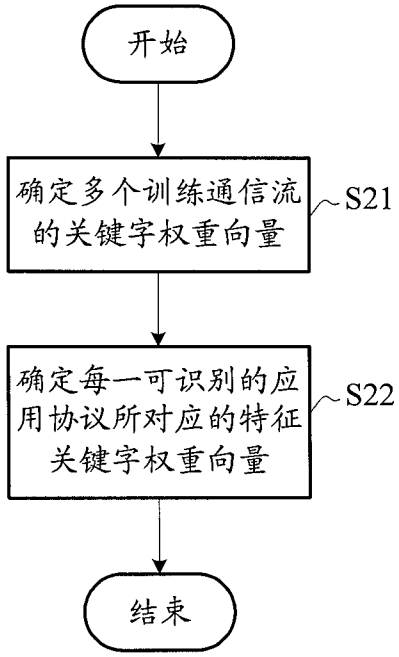

开始
确定多个训练通信流的关键字权重向量
S21
确定每一可识别的应用协议所对应的特征关键字权重向量
S22
结束

图 2

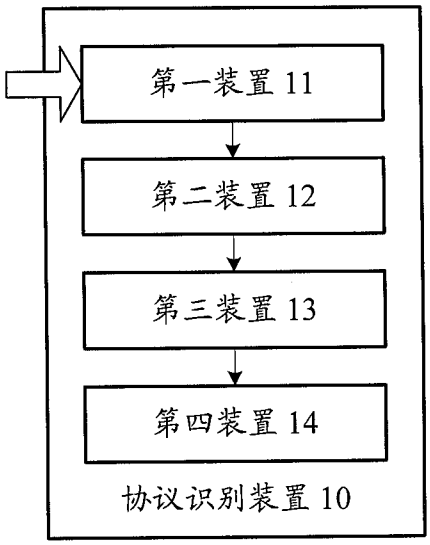

第一装置 11
第二装置 12
第三装置 13
第四装置 14
协议识别装置 10

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2010/072923

A. CLASSIFICATION OF SUBJECT MATTER

See extra sheet

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)

IPC：H04L,G06F,H04Q,H04W

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)

WPI,EPODOC,CNPAT,CNKI,IEEE: protocol, application, type, class, identify+, determin+, distinguish+, keyword?, vector?, position, location, match???, weight

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN101184089A（UNIV. ZHEJIANG POLYTECHNIC）21 May 2008（21.05.2008）the whole document	1-15
A	CN101488861A（UNIV. SUN YET-SEN）22 Jul. 2009（22.07.2009）the whole document	1-15
A	SEN, Subhabrata et.al, Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures, WWW2004, 22 May 2004 (22.05.2004), pages 512-517	1-15
A	CHEN, Liang et.al, A survey of Application-level Protocol Identification Algorithm, Computer Science, 2007,Vol.34,No.7, pages73-75	1-15

☐ 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
28 Jan. 2011(28.01.2011)	**24 Feb. 2011 (24.02.2011)**
Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer CHEN,Chen Telephone No. (86-10)62413487

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2010/072923

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN101184089A	21.05.2008	none	
CN101488861A	22.07.2009	none	

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2010/072923

A. CLASSIFICATION OF SUBJECT MATTER

H04L29/06(2006.01)i

H04L12/24(2006.01)i

国际检索报告

国际申请号
PCT/CN2010/072923

A. 主题的分类

见附加页

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)

IPC：H04L,G06F,H04Q,H04W

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）)

WPI,EPODOC,CNPAT,CNKI,IEEE:协议,应用,类型,识别,确认,辨别,辨识,关键词,关键字,矢量,向量,位置,匹配, 权重, protocol, application, type, class, identify+, determin+, distinguish+, keyword?, vector?, position, location, match???, weight

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
A	CN101184089A（浙江工业大学）21.5 月 2008（21.05.2008）全文	1-15
A	CN101488861A（中山大学）22.7 月 2009（22.07.2009）全文	1-15
A	SEN, Subhabrata 等, Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures, WWW2004, 22. 5 月 2004 (22.05.2004), 512-517 页	1-15
A	陈亮等,应用层协议识别算法综述,计算机科学, 2007 年,34 卷,第 7 期,73-75 页	1-15

☐ 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
28.1 月 2011（28.01.2011）	**24.2 月 2011 (24.02.2011)**
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	受权官员 **陈晨** 电话号码：(86-10) **62413487**

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2010/072923

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN101184089A	21.05.2008	无	
CN101488861A	22.07.2009	无	

国际检索报告

国际申请号
PCT/CN2010/072923

A. 主题的分类

H04L29/06(2006.01)i

H04L12/24(2006.01)i