(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 105281973 A
(43)申请公布日 2016.01.27

(21)申请号 201510481183.8

(22)申请日 2015.08.07

(71)申请人 南京邮电大学
地址 210023 江苏省南京市栖霞区文苑路9号

(72)发明人 陈伟 李晨阳 沈婧 张伟 杨庚

(74)专利代理机构 南京知识律师事务所 32207
代理人 汪旭东

(51)Int.Cl.
H04L 12/26(2006.01)
H04L 12/24(2006.01)

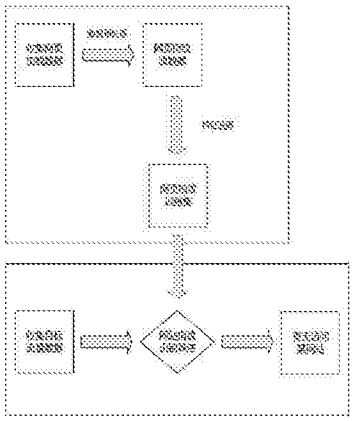
权利要求书3页 说明书6页 附图3页

(54)发明名称

一种针对特定网站类别的网页指纹识别方法

(57)摘要

本发明公开了一种针对特定网站类别的网页指纹识别方法，该方法为基于分类效果的特征选择方法以及基于训练集划分和结果集成相结合的分类方法，解决了特定网站类别网页指纹识别时出现的不平衡分类问题，并且改进了网页指纹收集方法，使其能够应对缓存机制下的网页指纹识别，该方法简单易行，在数据收集时充分考虑了不同的浏览器操作所生成的不同指纹数据，大大增强了指纹识别系统应对实际应用环境的能力，对网络行为监控有着很重要作用。



CN 105281973 A

1. 一种针对特定网站的网页指纹识别方法，其特征在于，所述方法包括如下步骤：

步骤 1：训练数据收集；在通信链路上对目标可能访问的所有网站的不同浏览器操作方式下的网页指纹数据进行采集；

步骤 2：数据预处理；消除噪音数据与冗余数据，包括：重传数据包、坏数据包，冗余数据包括协议控制数据；

步骤 3：构造训练集；首先进行特征提取操作，根据指纹特征从预处理后的网页加载数据流中提取出相应的特征值，然后将各特征或特征值组合成特征值向量，并将该网页加载实例所属的网站类别作为该特征值向量的分类类别添加在特征向量末尾构成训练实例，最终所有的训练实例构成了指纹原始训练集；

步骤 4：特征选择；将指纹数据集分为正类和负类，其中需要识别的网站类别为正类，其它网站类别为负类；

步骤 5：训练集划分；首先将整个训练集按正类和负类划分为正类训练集和负类训练集，用 C 和 $\overline{C}$ 分别表示正类训练集和负类训练集：

$$C=\{(c_i,+)\}_{i=1}^{n},\quad \overline{C}=\{(\overline{c}_i,-)\}_{i=1}^{m}$$

其中，c_i 表示第 i 个正类样本，n 表示正类样本数；$\overline{c}_i$ 表示第 i 个负类样本，m 表示负类样本数；之后，对负类训练集使用随机划分法进行划分：

$$\overline{C}i=\{(\overline{c}_k^i,-)\}_{k=1}^{l_i},i=1,2,...,N$$

其中，$\overline{C}i$ 代表划分后的第 i 块负类子训练集，l_i 表示第 i 块负类子训练集的样本数目，其中划分的块数 N 由以下公式决定：$N=\frac{m}{n}$，m 是负类训练集样本数，n 是正类训练集样本数，最后，将正类训练集和各个负类子训练集合并，得到 N 个子训练集：

$$T^i=C\bigcup\overline{C}_i,i=1,2,...,N$$

其中，T_i 是最终划分完成后的子训练集，在子训练集中，正类样本数目等于负类样本数目，在这些训练集上使用传统分类器进行分类；

步骤 6：分类；训练集划分完毕后，使用传统分类器在各个训练子集上对目标产生的待分类指纹数据进行分类；

步骤 7：结果集成；经过训练集划分并用分类器对每个子训练集进行分类后，产生 N 个分类结果，该分类结果数与训练集划分块数相同，最后基于最大化的思想对这些分类结果进行整合，得到最终的分类结果，该步骤如下：

$W^i = F(T^i)$，$i = 1, 2, \ldots, N$

$W = MAX(W^1, W^2, \ldots, W^N)$

经过对各个划分后的训练子集进行分类得到各子集分类结果为 W^i，该结果由两部分组成：待分类指纹所属网站类别 c 和待分类指纹属于该类别的类别权值 p，选取所有分类结果中 p 值最大的 W^k 作为最终分类结果。

2. 根据权利要求 1 所述的一种针对特定网站的网页指纹识别方法，其特征在于，所述步骤 4 的特征选择如下表所示，包括：

特征T	正类(C_i)	负类$(\overline{C}_i)$
t	A	C
$\overline{t}$	B	D

其中，C_i表示需要识别的正类，$\overline{C}_i$表示样本集中除了需要识别的网站以外的所有类别，A 表示正类中含有特征 t 的样本频率；B 表示正类中不含有特征 t 的样本频率；C 表示负类中含有特征 t 的样本频率，D 表示负类中不含有特征 t 的样本频率。

3. 根据权利要求 2 所述的一种针对特定网站的网页指纹识别方法，其特征在于，所述步骤 4 的特征选择是基于以下几种假设，包括：a) 如果特征在正类的指纹样本中广泛出现，即：在正类中该特征属性分布越均匀，那么该特征越能代表该网站类别；b) 如果特征在正类指纹样本中广泛出现，而在负类指纹样本中很少出现，则该特征区分类别的效果好；c) 如果正类指纹样本中广泛含有该特征，而负类样本中很少含有该特征，则该特征为该网站类别的识别特性；d) 从特征在样本中出现频数的角度考虑；如果该特征在正类样本中的出现频数远大于在负类样本中的出现频数，则该特征区分类别的效果好；使用衡量流属性和网站类别相关性的 DCR 算法为：

$$DCR(t,C_i)=\frac{A}{A+B}\times\frac{A}{A+C}\times(\frac{A}{A+B}-\frac{C}{C+D})\times\frac{FC_i}{F\overline{C}_i}$$

其中，$\frac{A}{A+B}$表示正类中含有流属性 t 的样本数占正类指纹样本集的比值，该值越大，流属性 t 表示正类的能力就越强；$\frac{A}{A+C}$表示正类中含有流属性 t 的样本数占整个指纹样本集的比例，该值越大，表示该流属性区分类别的效果好；$(\frac{A}{A+B}-\frac{C}{C+D})$表示正类中含有流属性 t 的样本数多，而负类中含有流属性 t 的样本数少，该值越大，表明该流属性类别区分的效果好；$\frac{FC_i}{F\overline{C}_i}$表示该流属性 t 在正类中平均出现的频数和在负类中平均出现的频数的比值，该比值越大表示该流属性和正类有着很强的相关性；将计算出每个指纹数据集中每个特征正类的分类效果值按照分类效果值大小进行降序排序，选择排名前 N 个特征作为最终分类用特征。

4. 根据权利要求 1 所述的一种针对特定网站的网页指纹识别方法，其特征在于，所述步骤 6 包括：使用的是基于余弦相似度匹配算法的 KNN 分类器，KNN 即 K-Nearest Neighbor，通过计算待分类实例与训练集中已知类别信息的训练样本间的相似度，筛选出与待分类样本最接近的 K 个训练样本，如果这 K 个样本均属于同一类别，则带分类样本也属于该类别，否则对于每个类别进行评估，最终按照某种规则确定待分类样本数据所属类别；当目标网页加载流量到达后，返回步骤 2 和步骤的操作，将其转化为待分类网页指纹实例，之后使用特征选择出的特征参与相似度计算，计算待分类指纹实例与训练集中各实例间的余弦相似度值，筛选出与待分类指纹最相似的 K 个指纹，其相似度计算公式：

$$Sim(d_i,d_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}}$$

其中，d_i为待分类指纹，d_j为训练集中某指纹，W_{ik}为待分类指纹中第 k 个特征值，W_{jk}为训练集指纹中第 k 个特征值，分别计算待分类指纹 K 个邻居的类别权重，计算公式：$p(d_i,C_j)=\sum_{\overline{d_i}\in KNN}Sim(d_i,\overline{d_i})y(\overline{d_i},C_j)$，其中 d_i为待分类指纹，$\overline{d_i}$ 为 K 个相邻指纹中的一个，$Sim(d_i,\overline{d_i})$为相似度匹配算法，$y(\overline{d_i},C_j)$为类别函数，当$\overline{d_i}$ 属于类别 C_j时，该函数值为 1，否则为 0，最后通过比较各类的类别权重，确定该待分类指纹的网站类别。

5. 根据权利要求 1 所述的一种针对特定网站的网页指纹识别方法，其特征在于，所述应用于应对缓存机制下的网页指纹识别。

一种针对特定网站类别的网页指纹识别方法

技术领域

[0001] 本发明涉及一种针对特定网站类别的网页指纹识别方法，属于数据挖掘和信息安全技术领域。

背景技术

[0002] 随着信息化建设的深入和互联网的深层发展，以互联网为载体的各种行为与交流愈发活跃，但是，事物都有它的两面性，人们在享受互联网带来的信息获取和活动交流便利的同时，各种形式的互联网犯罪也呈现出愈演愈烈之势，例如：网络间谍、网络诈骗、网络色情、网络赌博等，这些网络犯罪活动严重危害了国家安全和社会稳定。为了应对这些潜在的安全威胁，对目标人群的网络行为进行判别和监控就变得异常重要。但是，由于目前互联网行为是跨境互联网行为，即使访问操作的 web 服务器的物理地点并不在境内，由于我国长城防火墙的限制以及用户本身对于保密性的考虑，用户在进行跨境互联网操作时往往采用基于匿名通信技术的通信工具。由于传统的网络行为监控大多基于流量分析技术，这些基于数据包载荷特征的流量分类技术在数据包载荷为明文时是有效的，但由于匿名通信技术的广泛使用，该技术使用加密算法对数据包载荷部分进行加密，数据探针无法获得数据包明文信息，使得基于数据包载荷的流量分类方法失效，目前，传统的加密流量分析主要采用基于统计学的流量分析技术，网页指纹识别技术是该技术在实际场景下的具体应用。

[0003] 在许多网络信息交互活动中，网站是重要的信息载体，而网页作为网站的基本元素扮演着非常重要的角色，人们进入网站浏览网页，以获取信息或是在网页上留下自己的信息，因此，对于网络行为的监控而言，判定目标人群浏览的网站类别十分重要，而网站类别的判别离不开网页类别的判别。当目标访问某网页，浏览器开始加载网页时，虽然匿名通信工具将加载流中数据包的载荷部分用加密算法进行了加密，但是数据包其它特征信息并没有被掩盖，如：数据包大小值、数据包传输方向以及各个数据包传输次序和传输时间间隔，网页指纹识别技术正是通过分析通信链路加密流量中的这些信道特征来识别目标浏览的网页类别，所谓的网页“指纹”指的是加载网页时，通信链路中加载数据流产生的信道特征实例，由于加载不同内容的网页时，通信链路中加载数据流信道特征和网页内容呈现一一对应关系，在预先收集浏览相关网页的加密会话数据流，抽取统计规律信息构造训练集，使用有监督分类算法对目标产生的加密网页加载数据流进行分类，从而识别出目标访问的网页类别。

[0004] 网页指纹识别技术还在不断成熟的过程中，当前最核心的问题是如何提高在实际环境下的网页指纹识别性能和适用性，这里的环境包括系统使用环境和系统应用环境，在实际应用 中往往需要判断网页指纹是否属于某个特定网站，而适用性主要指的是网页指纹识别方法应对新型 web 技术的能力，现在大多数浏览器都默认开启缓存机制，而缓存机制的存在会形成不稳定的网页指纹，造成判断指纹准确性的降低。

[0005] 因此，在满足实际缓存条件下，特定网站类别网页指纹识别需求，对提高网页指纹识别技术实际应用环境的适用性至关重要。而本发明能够很好地解决上面的问题。

发明内容

[0006] 本发明目的在于提出了一种针对特定网站类别的网页指纹识别方法，该方法针对特定网站类别网页指纹在缓存环境下的识别问题，主要采用基于分类效果的特征选择方法以及基于训练集划分和结果集成相结合的分类方法，解决了特定网站类别网页指纹识别时出现的不平衡分类问题，改进了网页指纹收集方法，该方法能够应用于应对缓存机制下的网页指纹识别。

[0007] 本发明解决其技术问题所采取的技术方案是：一种针对特定网站类别的网页指纹识别方法，包括基于分类效果值的特征选择方法与训练集划分和结果集成相结合的分类方法，以及基于用户操作的网页指纹收集方法，

[0008] 方法流程：

[0009] 步骤 1：训练数据收集。在通信链路上对目标可能访问的所有网站的不同浏览器操作方式下的网页指纹数据进行采集。

[0010] 步骤 2：数据预处理。消除噪音数据与冗余数据，包括重传数据包、坏数据包，冗余数据包括协议控制数据。

[0011] 步骤 3：构造训练集。首先进行特征提取操作，根据指纹特征从预处理后的网页加载数据流中提取出相应的特征值，然后将各特征、特征值组合成特征值向量，并将该网页加载实例所属的网站类别作为该特征值向量的分类类别添加在特征向量末尾构成训练实例，最终所有的训练实例构成了指纹原始训练集。

[0012] 步骤 4：特征选择。本发明提出了一种基于分类效果值的特征选择方法，该方法将指纹数据集分为正类和负类，其中需要识别的网站类别为正类，其它网站类别为负类，如下表所示：

[0013]

特征T	正类(C_i)	负类$(\overline{C}_i)$
t	A	C
$\overline{t}$	B	D

[0014] 其中，C_i表示需要识别的正类，$\overline{C}_i$表示样本集中除了需要识别的网站以外的所有类别。A 表示正类中含有特征 t 的样本频率；B 表示正类中不含有特征 t 的样本频率；C 表示负类中含有特征 t 的样本频率，D 表示负类中不含有特征 t 的样本频率。本发明使用衡量流属性和网 站类别相关性的 DCR(Distinguish Classification Result) 算法为：

[0015] $$DCR(t,C_i)=\frac{A}{A+B}\times\frac{A}{A+C}\times(\frac{A}{A+B}-\frac{C}{C+D})\times\frac{FC_i}{F\overline{C}_i}$$

[0016] 其中，$\frac{A}{A+B}$表示正类指纹实例中含有特征属性 t 的比值，该值越大，特征属性 t 表示正类的能力就越强；$\frac{A}{A+C}$表示含有特征属性 t 的指纹实例中正类实例比例，该值越大，表示该流属性区分类别的效果较好。$(\frac{A}{A+B}-\frac{C}{C+D})$表示正类中含有特征属性 t 的样本数

较多，而负类中含有特征属性 t 的样本数较少，该值越大，表明该特征属性类别区分的效果较好。$\frac{FC_i}{\overline{FC_i}}$表示该特征属性 t 在正类中平均出现的频数和在负类中平均出现的频数的比值，该比值越大表示该特征属性和正类有着较强的相关性。将计算出每个指纹数据集中每个特征正类的分类效果值按照分类效果值大小进行降序排序，选择排名前 N 个特征作为最终分类用特征。

[0017] 步骤 5：训练集划分。首先将整个训练集按正类和负类划分为正类训练集和负类训练集，用 C 和$\overline{C}$分别表示正类训练集和负类训练集：

[0018] $C=\{(c_i,+)\}_{i=1}^{n}$，$\overline{C}=\{(\overline{c_i},-)\}_{i=1}^{m}$

[0019] 其中，c_i表示第 i 个正类样本，n 表示正类样本数；$\overline{c_i}$表示第 i 个负类样本，m 表示负类样本数。之后，对负类训练集使用随机划分法进行划分：

[0020] $\overline{C}i=\{(\overline{c}_k^i,-)\}_{k=1}^{l_i},i=1,2,...,N$

[0021] 其中，$\overline{C}i$代表划分后的第 i 块负类子训练集，l_i表示第 i 块负类子训练集的样本数目。其中划分的块数 N 由以下公式决定：$N=\frac{m}{n}$，m 是负类训练集样本数，n 是正类训练集样本数。最后，将正类训练集和各个负类子训练集合并，得到 N 个子训练集：

[0022] $T^i=C\bigcup\overline{C}_i,i=1,2,...,N$

[0023] 其中，T_i是最终划分完成后的子训练集。由于在子训练集中，正类样本数目等于负类样本数目，可以在这些训练集上使用传统分类器进行分类。

[0024] 步骤 6：分类。训练集划分完毕后，使用传统分类器在各个训练子集上对目标产生的待分类指纹数据进行分类。本发明使用的是基于余弦相似度匹配算法的 KNN 分类器，当目标网页加载流量到达后，返回步骤 2 和步骤 3 的操作，将其转化为待分类网页指纹实例，之后使用特征选择出的特征参与相似度计算，计算待分类指纹实例与训练集中各实例间的余弦相 似度值，筛选出与待分类指纹最相似的 K 个指纹，其相似度计算公式为：

[0025] $$\mathrm{Sim}(d_i,d_j)=\frac{\sum_{k=1}^{M}W_{ik}\times W_{jk}}{\sqrt{(\sum_{k=1}^{M}W_{ik}^2)(\sum_{k=1}^{M}W_{jk}^2)}}$$

[0026] 其中，d_i为待分类指纹，d_j为训练集中某指纹，W_{ik}为待分类指纹中第 k 个特征值，W_{jk}为训练集指纹中第 k 个特征值。之后分别计算待分类指纹 K 个邻居的类别权重，计算公式：$p(d_i,C_j)=\sum_{\overline{d_i}\in KNN}Sim(d_i,\overline{d_i})y(\overline{d_i},C_j)$，其中 d_i为待分类指纹，$\overline{d_i}$为 K 个相邻指纹中的一个，$Sim(d_i,\overline{d_i})$为相似度匹配算法，$y(\overline{d_i},C_j)$为类别函数，当$\overline{d_i}$属于类别 C_j时，该函数值为 1，否则为 0。最后通过比较各类的类别权重，确定该待分类指纹的网站类别。

[0027] 步骤 7：结果集成。分类器对每个子训练集进行分类后，产生 N 个分类结果，该分类结果数与训练集划分块数相同。最后基于最大化的思想对这些分类结果进行整合，得到最终的分类结果，具体步骤如下：

[0028] $W^i = F(T^i), i = 1,2,\ldots,N$

[0029] $W = MAX(W^1, W^2, \ldots, W^N)$

[0030] 经过对各个划分后的训练子集进行分类得到各子集分类结果为 W^i，该结果由两部分组成：待分类指纹所属网站类别 c 和待分类指纹属于该类别的类别权值 p。选取所有分类结果中 p 值最大的 W^k 作为最终分类结果。

[0031] 有益效果：

[0032] 1、本发明简单易行，在数据收集时充分考虑了不同的浏览器操作所生成的不同指纹数据，大大增强了指纹识别系统应对实际应用环境的能力。

[0033] 2、本发明使用特征选择方法对指纹特征进行特征选择，降低了指纹相似度计算时的计算复杂度，并且有效地提高了识别性能。

[0034] 3、本发明针对特定网站识别出现的不平衡分类情况，提出了一种训练集划分—分类—集成方法，有效地提高了特定网站识别性能和网站类别网页指纹的识别率。

附图说明

[0035] 图 1 为本发明网页指纹识别的方法流程图。

[0036] 图 2 为训练集划分示意图。

[0037] 图 3 为使用环境示意图。

[0038] 图 4 为某训练集中 DCR 值前 20 的特征情况。

[0039] 图 5 为子训练集中指纹实例相似度情况。

具体实施方式

[0040] 下面结合说明书附图对本发明创造作进一步的详细说明。

[0041] 本发明包含的技术术语包括如下：

[0042] 有监督分类：是根据已标记类别的训练数据来学习或建立一个决策模式，并依据此模式推测为标记实例的所属类别。在有监督学习中，每个训练实例都由特征值向量和所属类别所构成，通过在训练数据上使用监督学习算法产生可以判断新的未标记实例类别的模式分类方法，以下是其相关技术定义。

[0043] 训练集：在验证有监督分类的有效性时中一般将样本数据分为两部分：训练集和测试集，其中训练集用于建立分类模型，测试集则用于检验分类性能。在大多数情况下，原始数据不能直接作为训练集，而要经过数据预处理、特征选择等步骤后构建而成。

[0044] 分类器：是指作用于训练集之上的分类算法所形成的模式分类规则，分类器可以将待分类实例映射到给定类别中的某一个。本发明使用 KNN 分类器作为网页指纹识别分类器。

[0045] 不平衡分类：非平衡分类是指在各类别间分布不平衡的训练样本上进行分类，其中某些类别的样本数量远少于其他类，但是这些类别却是需要重点关注的类别，由于其样本数稀少，传统分类器对该类别的敏感度很低，因此需要特定的方法来解决稀有类样本识别问题。

[0046] 特征选择：是指从已有的 M 个特征中选择 N 个特征使得分类系统的某些性能指标最优化，其从原始特征集中选择出一些最有效特征以降低分类用训练集维度，起到降低分

类复杂度、提高分类性能的作用。

[0047] 在实际环境下应用该方法时，由于存在大量商业匿名通信工具和浏览器，本发明选取 shadowocks 翻墙软件以及 chrome 浏览器进行说明。如图 3 所示，首先目标通过使用 shadowsocks 翻墙工具访问浏览境外网站，shadowsocks 工具连接到远端 SOCKS 代理服务器，并使用 chrome 浏览器，此时 shadowsocks 工具在目标用户和远端代理服务器之间建立了一个匿名加密通信信道，该信道通过某个监控者可控的交换设备，该交换设备被配置有镜像端口，可由监控者抓取目标用户的流量数据，监控者通过从流量数据中提取出相应的网页加载数据，并对其进行分析。其中，监控者与目标用户处于相同的通信链路环境，目标所产生的流量数据可被监控方获取，流量数据载荷部分被加密。本发明基于该环境依据如图 1 所示的方法流程进行工作，具体的分析步骤包括 ：

[0048] 步骤 1 ：监控者通过利用可控交换设备使用数据采集工具进行数据收集，包括目标数据收集和训练数据收集。目标数据收集主要是通过数据采集工具对目标浏览网站的流量数据进 行收集，并从流量中提取出网页加载流量。训练数据收集主要是监控者使用浏览器访问目标可能会浏览的网站并在通信链路上收集并提取出相应网站的网页加载流量，其中每个网站分别使用 4 种不同的浏览器操作方式对其进行访问，每种浏览器操作方式分别收集 10 次指纹数据。指纹数据由数据采集工具导入至 CSV 文件中，在每个指纹数据记录了一次网页加载过程中浏览器与远端 Web 服务器间所有的流量活动，这些指纹数据由一些系列 TCP 数据包构成，由于加密技术的影响无法得到 TCP 数据包载荷部分的信息，其余 TCP 数据包信息的数据结构包括 ：数据包序号、数据包传输时间、源 IP 地址、目的 IP 地址、数据包大小以及数据包描述。

[0049] 步骤 2 ：收集完指纹数据后，需要对其进行数据预处理操作。指纹数据中的 TCP 数据中含有大量的协议控制数据，这些协议控制数据主要用于控制 TCP 数据的建立和断开，除此之外原始的指纹数据中还包含有其它的冗余和噪音数据，包括 ：TCP 重传数据包以及 TCP 坏包。本案例将 shadowsocks 指纹数据中数据包大小小于 70 的数据包视为协议控制数据包予以清除，并将数据包描述中含有“Retransmission”，“Dup”，“Out-of-order”描述字段的数据包视为坏包和重传数据包予以清除。

[0050] 步骤 3 ：在清除冗余和噪音数据后，需要将原始数据构造为数据集。首先提取出所有网页加载流量中数据包中不同的 < 数据包大小，数据包方向 > 向量，将所有不同的 < 数据包大小，数据包方向 > 向量作为原始特征。之后将每个网页加载流量中符合某 < 数据包大小，数据包方向 > 特征的数据包个数作为特征值，并加上该网页指纹流量所属的网站类别，最终构成网页指纹训练集 ：

[0051]

(1048,-)	(116,-)	.....	(148,-)	Website Class
449	3	.....	0	Zhaori News
438	4	.....	0	Zhaori News
.....	.....	.....	.....	.....
678	0	.....	2	youtube
578	0	.....	0	New York Time

[0052] 表中每行数据表示一个完整的网页加载数据流实例，特征即数据包大小与数据包传输方向组成的特征对，例如 (1048,-) 表示数据包大小为 1048，数据包传输方向为由服务器传输至浏览器，而每个特征值表示的是该指纹实例中满足该特征的数据包数目，例如 449

表示，在某个网站类别为朝日新闻的网页加载流量中数据包大小为1048，数据包传输方向为从服务器端到浏览器端的数据包共有449个。

[0053] 步骤4：特征选择。假设本案例中需要判别的特定网站类别为facebook，某训练集包含特征：(962，+)，其中网站类别为facebook的指纹实例中含有该特征的实例数目为19；网站类别非facebook的指纹实例中含有该特征的实例数目为0；网站类别为facebook的指纹实例 中不含该特征的实例数目为21；网站类别非facebook的指纹实例中不含该特征的实例数目为80；网站类别为facebook的指纹实例中符合该特征数据包个数为38；网站类别非facebook的指纹实例中符合该特征的数据包个数为0；因此最终特征(962,+)的DCR值为：

[0054] $$DCR\ (962+)=\frac{19}{19+21}\times\frac{19}{19+0}\times\left(\frac{19}{19+21}-\frac{0}{0+80}\right)\times\frac{38}{(0+1)}=8.57375$$

[0055] 对训练集中每个特征均计算出DCR值，选出DCR值前20的特征作为分类特征，如图4所示。

[0056] 步骤5：训练集划分。如图2所示，假设本案例供收集了40个目标潜在访问网站的网页指纹数据，那么网站类别为facebook的网页指纹数据占训练集中所有数据的1/40，正类指纹数据和负类指纹数据在训练集中的比例为1：39。对负类指纹数据进行随机划分，划分为39个子负类数据集，之后再合并每个子负类数据集和正类数据集构成子训练集，在这些子训练集中，正类指纹数目与负类相同。

[0057] 步骤6：分类。计算待分类目标指纹与各子训练集中指纹实例分内特征的余弦相似度，如图5所示，以('facebook.59',0.995499913195721)为例，其表示训练集中序号为59的类别为facebook的指纹实例与待测指纹的相似度为0.995499913195721。将子训练集各指纹实例按照与待测指纹相似度值从高到低排列，取排名前5的相似度计算结果参与分类决策，当前排名前5的结果为：(facebook,0.995)，(facebook,0.692)，(zhaori,0.626)，(facebook,0.614)，(facebook,0.610)，之后计算该目标指纹的类别权重，即相应类别的相似度之和：facebook的类别权值为0.995+0.692+0.614+0.610 = 2.911，非facecook的类别权值为0.626，子分类选择类别权值最大的分类结果作为子分类分类结果，得出最终分类器输出的分类结果为facebook。

[0058] 步骤7：结果集成。当待测指纹在每个子训练集上使用KNN分类器得出分类结果后，我们基于最大化思想汇总各个子分类器分类结果，即选择各个子分类器中最大类别权值的子分类结果作为最终的分类结果，如图5中两个子训练集，第一个训练集的分类结果是(facebook,2.911)，(非facebook,0.626)，第二个训练集的分类结果是(facebook,2.157)，(非facebook,1.261)，其中最大类别权值为2.911，因而将其对应的分类结果facebook作为这两个子分类器集成后的结果。

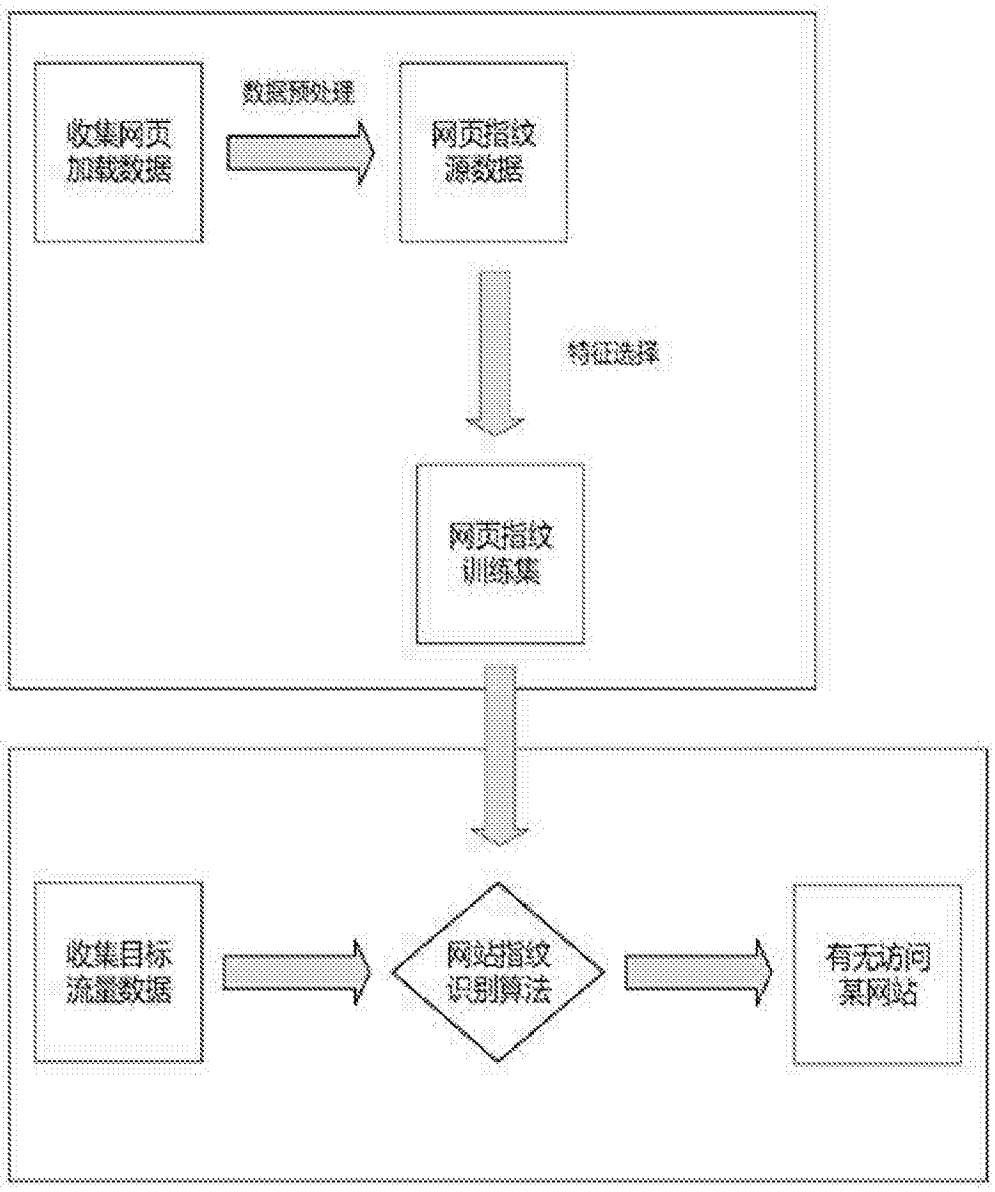

收集网页
加载数据
数据预处理
网页指纹
源数据
特征选择
网页指纹
训练集
收集目标
流量数据
网站指纹
识别算法
有无访问
某网站

图 1

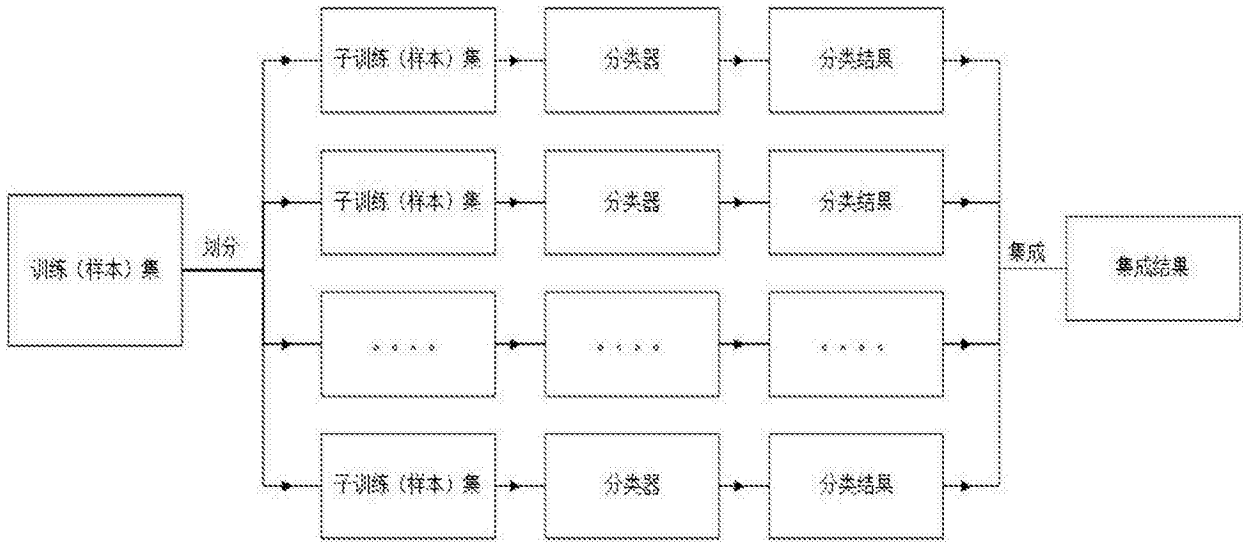

训练（样本）集
划分
子训练（样本）集
分类器
分类结果
子训练（样本）集
分类器
分类结果
。。。。
。。。。
。。。。
子训练（样本）集
分类器
分类结果
集成
集成结果

图 2

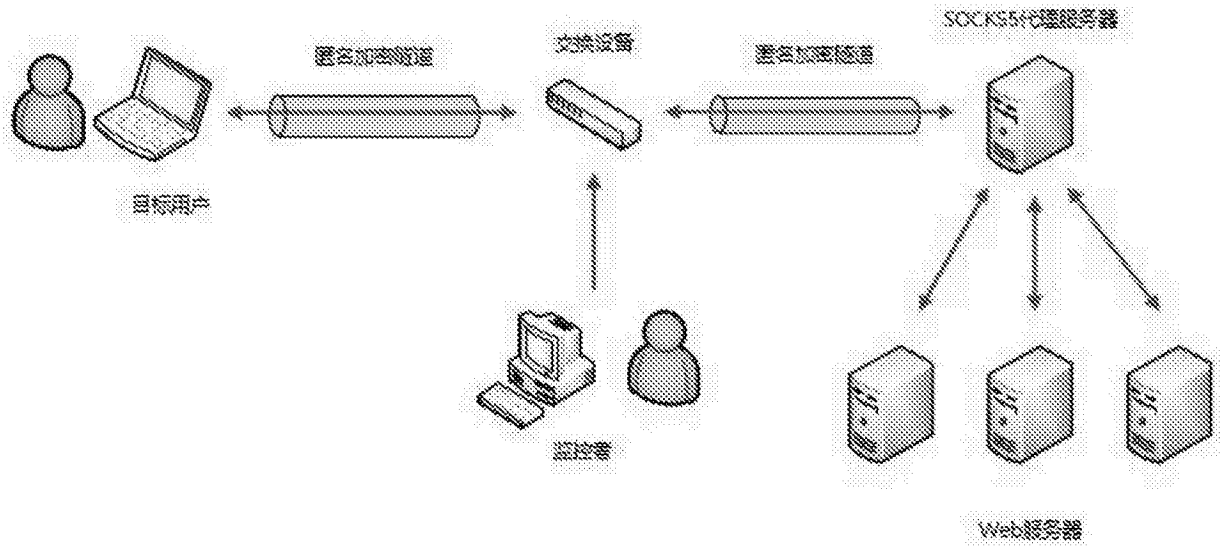

匿名加密隧道
交换设备
匿名加密隧道
SOCKS5代理服务器
目标用户
监控者
Web服务器

图 3

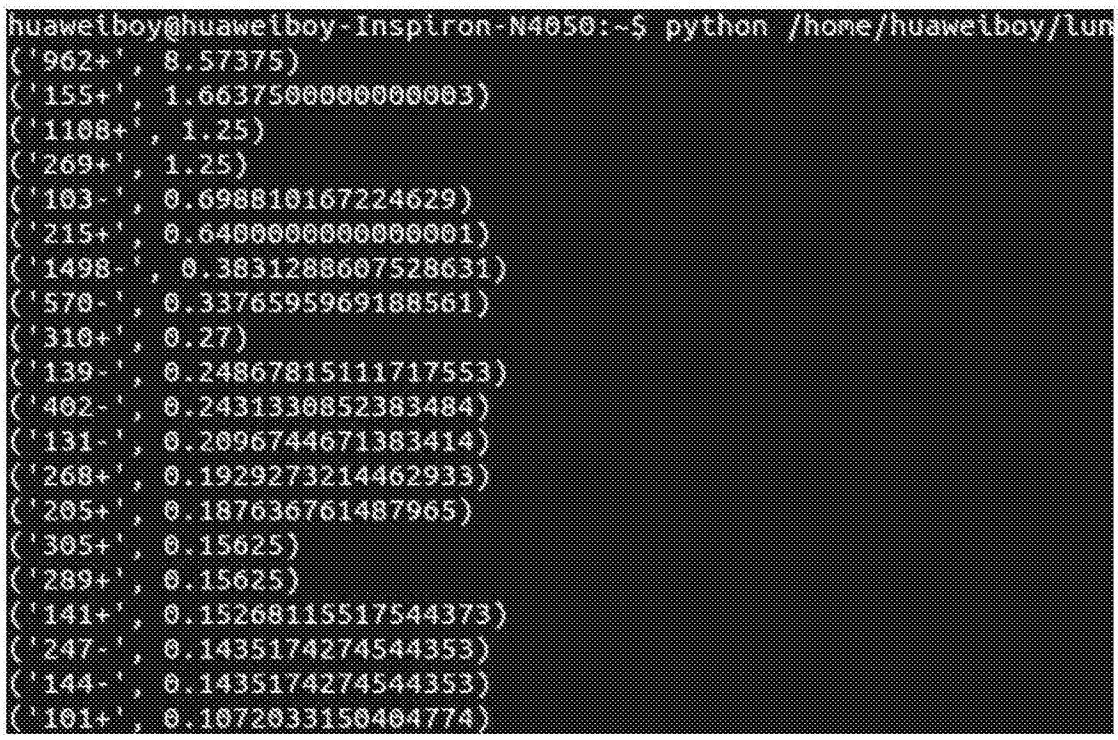

huaweiboy@huaweiboy-Inspiron-N4050:~$ python /home/huaweiboy/lun
('962+', 8.57375)
('155+', 1.6637500000000003)
('1108+', 1.25)
('269+', 1.25)
('103-', 0.698810167224629)
('215+', 0.6400000000000001)
('1498-', 0.38312886075286 31)
('570-', 0.33765959691885 61)
('310+', 0.27)
('139-', 0.24867815111717553)
('402-', 0.24313308523834 84)
('131-', 0.20967446713834 14)
('268+', 0.19292732144629 33)
('205+', 0.18763676148796 5)
('305+', 0.15625)
('289+', 0.15625)
('141+', 0.15268115517544373)
('247-', 0.14351742745443 53)
('144-', 0.14351742745443 53)
('101+', 0.10720331504047 74)

图 4

图 5