互联网中基于用户连接图的流量分类机制

张 震* 汪斌强 陈鸿昶 马海龙
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要：针对机器学习分类算法的“概念漂移”现象，该文提出了一种基于用户连接图的(Host Connection Graph, HCG)流量分类机制。算法将{IP Address, Port}作为用户唯一标识，构建了用户连接图，提出了“用户相似度”的概念；应用“图挖掘”理论将用户连接图划分为互不相交的行为子簇，使得用户之间的相互通信抽象为一种“社会团体”；通过定义基于信息熵的“用户行为模式”(UBM)，分析了各个行为子簇背后表现出的业务特征，并使用“UBM+Port”对用户行为子簇进行了业务标签映射，实现了流量分类的目的。仿真实验表明：在不牺牲识别准确率的前提下，算法不仅能克服“概念漂移”问题，还能有效降低算法的计算复杂度。

关键词：流量分类；用户连接图；用户相似度；图挖掘
中图分类号：TP393 **文献标识码**： A **文章编号**：1009-5896(2013)04-0958-07
DOI: 10.3724/SP.J.1146.2012.01040

Internet Traffic Classification Based on Host Connection Graph

Zhang Zhen Wang Bin-qiang Chen Hong-chang Ma Hai-long
(*National Digital Switching System Engineering & Technological R&D Center*, *Zhengzhou* 450002, *China*)

Abstract: Considering at the concept drift issue of machine learning identification, a novel algorithm called traffic classification based on Host Connection Graph (HCG) is proposed. Considering {IP Address, Port} as the unique user identifier, HCG constructs a host connection graph and innovates the concept of user similarity. Based on the theory of graph mining, social community is abstracted from communications among hosts by partitioning the graph into mutually intersectant behavior clusters. In order to reach traffic classification, HCG not only conceives a definition called User Behavior Mode (UBM) to analyse the implicit traffic characteristics, but also maps application labels to every host behavior by employing UBM and Port. Finally, simulations are conducted based on the real network trace. Results demonstrate that HCG can circumvent the concept shift problem and ameliorate gracefully computational complication without sacrificing accuracy.

Key words: Traffic classification; Host Connection Graph (HCG); User similarity; Graph mining

1 引言

随着 P2P 业务和新兴业务的出现，大量随机端口和加密技术被用于数据传输，导致基于端口和深度报文检测(Deep Packet Inspect, DPI)的流量分类方法逐步失效。为了不依赖于报文负载进行检测，基于机器学习的流量分类技术受到了关注。它通过提取网络流的统计特征(如平均报文长度、流的持续时间等)，将网络流抽象为由一组统计特征值构成的属性向量，实现了由流量分类向机器学习的转化，代表性方法有 K-means[1]、朴素贝叶斯分类器(Naïve Bayesian Classifier, NBC)[2]。但是基于机器学习的流量分类方法最大的问题在于：高速网络中，提取流量特征的时间复杂度一般随所统计的报文数量线性增加；概念飘移问题[3]，即在时刻 t 得到的最佳分类模型 y_t，与前一时刻 $t-1$ 得到的最佳分类模型 y_{t-1} 不一致，导致这种现象的原因是网络时空环境和网络应用分布发生变化。

针对机器学习方法过分依赖网络环境和计算复杂度过高的问题，用户级的流量分类方法不再关注应用层负载、流量特征的提取和统计，而是从用户行为学的角度进行研究，为流量分类开辟了新的研究思路。Karagiannis 等人[4]分别从社会、功能、应用 3 个层面逐步深入分析了各种应用的行为模式，提出了一种基于“用户交互行为”的流量分类方法—BLINC(BLINd Classification)。BLINC 利用不同业务在传输层连接模式的差异，构建了主机交互关系图，每一个图代表一个主机与其他主机间的交互

2012-08-14 收到，2012-12-31 改回
国家 973 规划项目(2012CB312901，2012CB312905)和国家 863 计划项目(2011AA01A103)资助课题
*通信作者：张震 zhangzhen2096@163.com

行为，并以“图匹配”的方式来划分网络流量。BLINC 方法利用了网络应用的行为属性，不依赖于报文载荷内容和流统计特性，具有良好的可扩展性。但是，BLINC 方法具有以下缺陷：需要事先确定交互关系图，对于新出现的业务模式无能为力；创建和匹配模式图的计算复杂度较高，需要消耗大量的时间，必须进行离线分析处理，不适合对骨干网中的流量进行分类。

基于用户行为的分析方法还包括：文献[5]将 P2P 节点信息的复制和分发来对用户行为进行建模，并基于此计算了 P2P 业务的动态变化对网络的影响；文献[6,7]根据用户行为特征对 P2P 僵尸网络等异常流量进行了检测，取得了很好的效果。不同于 BLINC 方法，本文提出了一种基于用户连接图(Host Connection Graph, HCG)的流量分类方法。

2　预备知识

2.1 信息熵

在信息论领域，信息熵从平均意义上表征了信源的总体信息测度和不确定性[8]。为了描述用户的连接行为特征，特引入“信息熵”的概念。考虑随机变量 X 取值于离散集 $A=\{a_1,a_2,\cdots,a_n\}$，假设 X 按照某种概率分布共产生 m 个观测值，m_i 表示变量 X 取值 a_i 的次数，可得 X 取值 a_i 的经验概率值为 $p(a_i)=m_i/m$。则变量 X 的信息熵可定义为

$$H(X)=-\sum_{i=1}^{n}p(a_i)\lg p(a_i) \tag{1}$$

假设 $n\geq 2$ 和 $m\geq 2$ 成立(即存在取值的不确定性)，为了对信息熵进行归一化处理，特定义相对熵 $R(X)$：

$$R(X)=\frac{H(X)}{H_{\max}(X)}=\frac{H(X)}{\lg\min\{n,m\}} \tag{2}$$

$R(X)$ 表示了变量 X 的统计不确定性，且存在以下特点：当 $R(X)$ 趋于比较小的值时，说明 X 取值确定，即在某一个或几个值上发生频率较高；若 $R(X)$ 趋于比较大的值时，说明 X 取值比较均匀，不确定性较大；若 $R(X)=0$，则必然存在 $a_i\in A$，使得 $p(a_i)=1$ 成立，即变量 X 不存在任何变化。

2.2 谱聚类

谱聚类将聚类中的每个数据 x_i 看作无向图 $G(V,E)$ 的一个顶点，数据 x_i, x_j 之间的关联看成是这个无向图 $G(V,E)$的边 e_{ij}，并按照数据间的相似性对相应的边赋予不同的权重 w_{ij}；利用关联矩阵的谱分解所传达的信息，来揭示无向图的结构属性。通过最优化一个有效的图划分判据——规范切判据[9]，使得同一类的点具有较高的相似性，不同类的点具有较低的相似性，进而得到若干互不连通的子图。

谱聚类利用了拉普拉斯矩阵的特征向量，是一种配对聚类方法，主要优势表现为：算法仅与数据点的数目有关，与维数无关，可以避免由特征向量的过高维数所造成的奇异性问题；通过谱分解，可以获得聚类判据在放松了的连续域中的全局最优解；谱聚类不用对数据的全局结构作假设，就能识别出任意形状数据集的聚类，非常适合于许多实际问题。基于以上优势，该方法已成功应用于语音识别、图像分割等领域[9,10]。

3　基于谱聚类的用户行为子簇划分

3.1 构造用户连接图

HCG方法将用户之间的交互行为通过连接图 $G(V,E)$ 的形式进行了抽象，将互联网中每个用户抽象为一个点 $v_i\in V$。若用户 v_i，v_j 之间相互通信，则将对应的点连成一条边 $e_{ij}\in E$，于是不同的互联网应用将形成各种模式的拓扑结构。下面给出了HCG创建用户连接图 $G(V,E)$ 的步骤。

步骤 1　**构造用户连接图 $G(V,E)$的点**　用户主机一般是由 IP 地址表示的，端口号则与某种服务相关联。基于端口进行分析具有以下 4 方面优势：区别于仅用 IP 地址标识用户的方法，加入端口能够识别一个用户可能开启多个业务的情况；从业务识别的层面，用端口来细化业务流，能更好地掌握业务绑定端口的规律；从网络功能层面，若大量主机和用户 A 的某一固定端口建立了连接，说明用户 A 很有可能为服务器；从连接行为层面，若用户 B 和用户 C 的大量端口建立了连接，说明用户 B 很可能是端口扫描病毒。因此，HCG 方法将{IP Address, Port}作为用户的唯一标识，并抽象为连接图中的一个结点。

步骤 2　**构造用户连接图 $G(V,E)$的边**　连接图的边表征了用户之间的交互，可根据不同的研究目标对边进行定义。考虑以下原则建立 1 条边：(1) 对于 UDP 流，若 A 向 B 传送了第 1 个报文；(2) 对于 TCP 流，若第 1 个 SYN 报文发送的时候；(3) 若 A, B 之间通信的字节总数或者报文总数大于某一阈值；(4) 若 A 和 B 三次握手成功，TCP 连接建立。从行为学的角度，只要用户 A 向用户 B 发送了报文，则 A 必然存在向 B 索取某种信息或资源的意图，因此，HCG 方法选择(1)和(2)作为构建边的准则。

3.2 基于谱聚类的行为子簇划分

基于谱聚类的行为子簇划分的核心问题就是如何定义用户之间的相似性。为了便于描述，下面首先给出以下定义：

定义 1 相邻用户(neighbor user) 在用户连接图 $G(V,E)$ 中，若用户 v_i 与用户 v_j 直接相连，则称 v_i 和 v_j 为相邻用户；若 v_i 与用户 v_1, v_2, v_3 相连，则定义用户 v_i 的相邻用户集合为 $U_i = \{v_1, v_2, v_3\}$。

定义 2 用户相似度(user similarity) 若用户 v_i 与 v_j 为相邻用户，则定义 v_i 与 v_j 的用户相似度为无穷大 ∞ (即两个用户具有共同的业务应用)；若 v_i 与 v_j 不相邻，且 v_i， v_j 对应的相邻用户集合为 U_i 和 U_j，则 v_i 与 v_j 的用户相似度定义为两个相邻集合共享用户的个数 $|U_i \cap U_j|$。

用户相似度从用户的周围连接环境出发，不仅仅孤立地计算单一用户对之间的关联，而是引入了以“用户之间共享最近邻”为指标的相似性度量。如图1所示，用户 A 和 B 之间共享4个用户连接，根据用户相似度的定义，A 和 B 之间的相似度为4。用户相似度基于如下原理：若用户 v_1 与用户 v_i 直接相连，用户 v_n 和用户 v_i 直接相连，则认定 v_1 与 v_n 具有较高的业务相似度。谱聚类的过程是基于数据点的相似度矩阵进行的，HCG算法则是以“用户相似度”来度量用户之间的业务交互行为，并基于谱聚类将用户连接图 $G(V,E)$ 进行行为子簇的划分。表1给出了基于谱聚类的用户行为子簇划分的详细流程。

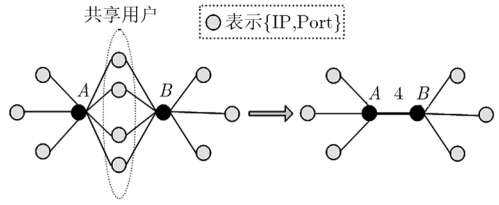

图1 用户相似度示意图

表1 用户行为子簇划分详细流程图

用户行为子簇划分详细流程图

(1) 输入：用户集合 $U = \{u_1, \cdots, u_N\}$，聚簇个数 K，尺度参数 σ
输出：行为聚簇集合 $C = \{C_1, \cdots, C_K\}$

(2) 若用户 u_i, u_j 为相邻用户，则用户相似度 $u_{ij} = \infty$，即 u_i, u_j 具有相同的业务交互行为；

(3) 若用户 u_i, u_j 不相邻，则用户相似度 $u_{ij} = |U_i \cap U_j|$，其中 U_i 和 U_j 分别为 u_i, u_j 的相邻用户集合；

(4) 计算相似度矩阵 $\boldsymbol{S} \equiv [s_{ij}]_{N\times N}$，其中 $s_{ij} = \exp\{-u_{ij}^2/\sigma^2\}$，令用户 u_i 的自相似值为最大值 $u_{ii} = \infty$, $s_{ii} = 0$；

(5) 构建拉普拉斯矩阵 $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$，其中 $\boldsymbol{D}$ 为对角矩阵，定义为 $D_{ii} = \sum_{j=1}^{N} s_{ij}$；

(6) 找出 $\boldsymbol{L}$ 最大的 K 个本征值 $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_K$ 和对应的本征向量 $[\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_K]$，构成矩阵 $\boldsymbol{F}_{N\times K} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_K]$；

(7) 对矩阵 $\boldsymbol{F}_{N\times K}$ 进行规范化处理，得到矩阵 $\boldsymbol{Y}_{N\times K}$，其中

$$Y_{ij} = F_{ij} \Big/ \left(\sum_{j=1}^{K} F_{ij}^2\right)^{1/2}$$

(8) 将 $\boldsymbol{Y}_{N\times K}$ 的每一行看成 $\mathbb{R}^K$ 空间中的一点，使用 K-means 将其聚为 K 类 $C_1, \cdots, C_K$；

(9) 如果 $\boldsymbol{Y}_{N\times K}$ 的第 i 行属于 C_j，则将用户 u_i 也划分到子簇集合 C_j 中。

4 基于信息熵的用户行为分析

4.1 基于信息熵的行为定义

在用户连接图 $G(V,E)$ 中，用{SrcIP, SrcPort, DstIP, DstPort}4维元素来标识一条用户连接。若固定SrcIP，则{SrcIP, SrcPort, *, *}、{SrcIP, *, DstIP, *}、{ SrcIP, *, *, DstPort}的连接数会呈现不同值，其中“*”代表任意值。分别计算相对熵 R(SrcIP, SrcPort, *, *)、R(SrcIP, *, DstIP, *)、R(SrcIP, *, *, DstPort)。为方便描述，特将 R(SrcIP, SrcPort, *, *) 简写为 R(SrcPort)，其它依次类比。由于 R(SrcPort) 表示在某一SrcIP下，以SrcPort为变量的相对熵值，所以 R(SrcPort) 实质上体现了SrcPort平均不确定性。基于此，下面给出用户行为模式的定义。

定义3 用户行为模式(User Behavior Mode, UBM) 给定某一用户地址SrcIP，分别计算相对熵 R(SrcPort)、R(DstIP)、R(DstPort)，并给出式(3)的准则进行量化：

$$\widehat{R}(X) = \begin{cases} 0, & 0 \le R(X) \le \delta \\ 1, & \delta < R(X) \le 1-\delta \\ 2, & 1-\delta < R(X) \le 1 \end{cases} \tag{3}$$

其中 $\delta \in [0, 0.5)$，定义 M(SrcIP) = {$\widehat{R}$(SrcPort), $\widehat{R}$(DstIP), $\widehat{R}$(DstPort)} 为SrcIP的用户行为模式。

基于定义3，下面给出了常见的3种用户模式：

(1)Server行为 若某用户行为模式满足 M(SrcIP)={0,1,2},{0,2,2}，则认定用户SrcIP为服务提供者(即为Server角色)。这主要由于 M(SrcIP) 体现了SrcIP在使用固定的端口与某个或某几个客户端进行交互。

(2)P2P模式 若行为子簇中多数用户的行为模式满足 M(SrcIP)={0,0,1}, $\cdots$,{0,1,0}或{1,0,0}，则可认定该子簇表现为P2P行为。多数用户的 M(SrcIP) 中每一维相对熵都小于2，表明行为子簇中连接图任意结点的连接度比较均匀。这正体现了每个P2P用户既可以为Server，也可以为Client的特点。

(3)Scan病毒 若某用户行为模式满足 M(SrcIP) ={0,2,0},{2,2,0}，则可认定该用户为IP Scan病毒；若用户模式等于{0,0,2},{2,0,2}，则该用户为Port

Scan病毒。IP Scan病毒表现的行为特征是用户使用设定的目的端口与受害者通信；Port Scan病毒则对固定的用户地址进行端口扫面。

4.2 行为子簇的业务标签映射

基于谱聚类的方法将样本集合划分为不同的子簇 $C=\{C_1,\cdots,C_q\}$，流量分类需要根据用户行为模式进一步确定任意子簇对应的业务标签 L_i，$L_i\in L=\{L_1,\cdots,L_m\}$。其中，业务标签为样本集合的所属类别。如图 2 所示，HCG 算法对行为子簇进行标签映射的核心步骤如下：

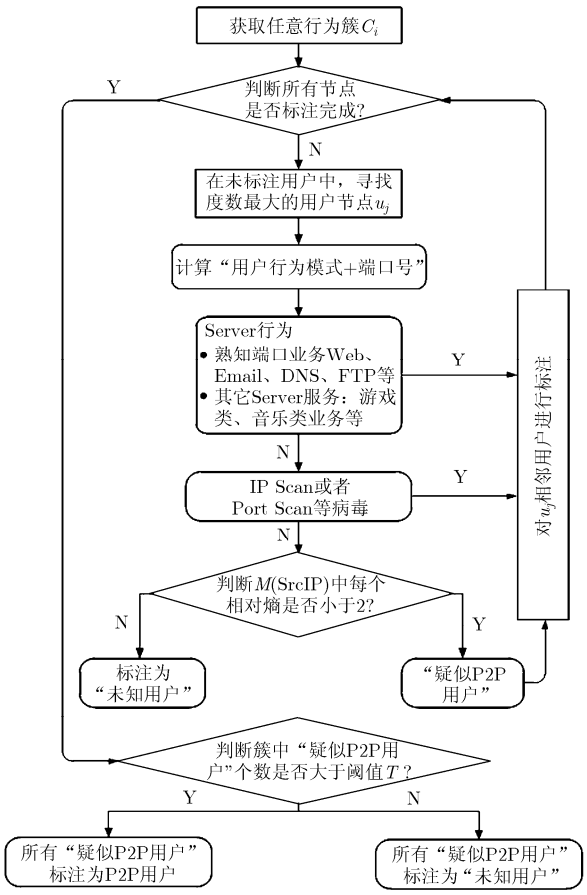

图 2 对用户结点进行标注

步骤 1　**寻求行为子簇中度数最大的用户结点 u_i**　HCG 算法按照行为子簇中用户的度数大小进行标注：一方面，在用户连接图中，用户结点的度数越大表明其相邻用户越多，作为服务器的可能性也就越大；另一方面，若连接度数最大的用户结点被标注，则与其直接相连的用户也相应地得到了识别，从而能有效降低算法的计算复杂度。

步骤 2　**应用“用户行为模式+端口号”标注结点 u_i**　HCG 算法使用“用户行为模式+端口号”的方式对用户结点 u_i 进行类型标注。根据定义 3，首先计算 u_i 的行为模式，然后结合熟知端口号进行识别：

(1)对于 Server 行为，结合熟知端口号进行业务识别，如：FTP, DNS, HTTP, Email, Telnet 等；非熟知端口业务结合固定端口号识别，如：游戏(如“魔兽世界”对应端口 3424)、数据库服务(如 Mysql 服务对应端口 3306)等。

(2)对于 P2P 业务，若某用户的行为模式满足 $M(\text{SrcIP})=\{A,B,C\}$，其中 $A<2,B<2,C<2$，则该用户被定义为“疑似 P2P 用户”。如果行为子簇中存在大于 T 个疑似 P2P 用户，则该子簇表现为 P2P 行为，并将所有“疑似 P2P 用户”标注为 P2P 用户。

(3)只需分析用户行为模式 $M(\text{SrcIP})=\{0,2,0\}$, $\{2,2,0\}$,$\{0,0,2\}$,$\{2,0,2\}$，即可判定病毒型业务。

(4)不属于上述 3 种行为的结点被标注为“未知用户”，该用户很有可能属于新兴业务。

(5)最后，由于相邻用户结点具有相同的业务，所以将与 u_i 相邻的所有用户结点进行标注。

步骤 3　**对子簇进行业务标签映射**　通过对行为子簇中的用户结点进行标注，可以得到子簇中每个用户的业务类型。流量识别算法还需要将行为子簇集合 C 和业务标签集合 L 进行映射。HCG 采用条件概率 $P(l=L_j|C_i)$ 来表示子簇 C_i 中任意用户结点属于类别标签 L_j 的概率，并且利用已标记用户集对其进行估计，其中 $i\in[1,q],j\in[1,m]$。

令 N_i^j 表示在子簇 C_i 中属于类别 L_j 的用户总数，N_i 表示子簇 C_i 中所有的用户总数，则根据最大似然估计，可得条件概率的估计表达式为：$\tilde{P}(l=L_j|C_i)=N_i^j/N_i$。基于此，可得到子簇的标签映射函数为

$$l(C_i)=\arg\max_{j=1,\cdots,m}\tilde{P}\left(l=L_j\middle|C_i\right) \tag{4}$$

映射函数就是将子簇中包含最多用户的类别标签赋给该簇。若该子簇中未知用户数目最大，则认定其为“未知流量类型”。

4.3 流量分类详细流程

图 3 描述了 HCG 算法的整体流程，算法的核心思想体现为：将{IP Address, Port}作为用户标识，使得流量识别不依赖与应用层负载和流的统计信息；基于谱聚类的用户行为挖掘，将用户划分为各个子簇，抽象了用户的“社会团体”行为；引入了“信息熵”的概念对用户行为模式进行了量化，分析了行为子簇背后表现出的业务行为特征；基于“用户行为模式+端口”对用户结点进行了标注，并对子簇进行了业务标签映射，从而实现了流量识别的目标。

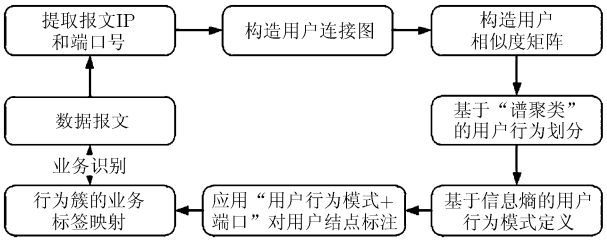

图3 HCG算法流量识别流程图

5 实验结果及分析

流量分类方法有多种多样，其中基于机器学习和基于用户行为的分类方法受到越来越多人的关注。本文采用此两种方法的典型应用算法——NBC[2]和 BLINC[4]与 HCG 流量分类方法进行仿真对比。

5.1 性能评价指标

实验中采用以下两种评价指标：

定义 4 整体准确率 对于任意业务类型 $l_i \in L = \{L_1, \cdots, L_m\}$，假设被正确分类为 l_i 的样本流总数为 N_i'，分类前属于 l_i 的样本流总数为 N_i，则业务类型 l_i 的分类准确率为：$P_i = N_i'/N_i$，整体准确率为：$P_{\text{all}} = \sum_{i=1}^{q} N_i' \Big/ \sum_{i=1}^{q} N_i$。

定义 5 计算复杂度 针对某一数据集，分类算法计算出所有数据流的类型标签所消耗的时间。

5.2 实验数据说明

实验中的流量数据采集于校园网的 Internet 接入点，该接入点的接入带宽为 100 Mbps。我们对接入链路进行了旁路，然后利用一台安装了 Ethereal 的笔记本电脑。为了仿真不同时间段的算法性能，观测了从 2012 年 6 月 20 日 12 时至 2012 年 6 月 21 日 12 时的 24 个小时流量数据，每个小时只采集了连续 10 min 的数据。所采集的流量文件的大小为 24 byte, 562 byte, 469 byte, 311 byte，并且流量数据包含完整的数据包负载。

文献[4]给出了 Web, FTP, Email 等多种主流业务的特征字段。因为 Campus Trace 与文献[4]的数据集采集环境不同，所以本文在文献[4]所给出的特征字段基础上进行了扩充，其中病毒特征采用 Snort 库[11]，表 2 给出了扩充业务字段相关说明。并使用 DPI 方法对采集的数据报文进行识别和标注，称为 Campus Trace 样本集。然后，以 Campus Trace 样本集作为基准，计算各种仿真算法的准确性。

图 4 给出了在采集的流量数据中，各种业务所占的比例：能够识别出业务类型的数据包占总包数的 84.4%，病毒包占 1.4%，不能识别出业务类型的数据包占总包数的 7.2%。某些数据包不能识别的原因是数据包无负载或者某些业务对数据包进行了加密。另外，为了更好地对比仿真，Campus Trace 样本集剔除了不被识别的数据包，只使用已标注的业务报文。

表 2 扩充业务特征说明

业务应用	特征字段
eDonkey(P2P 行为)	‘0x319010000’
Thunder(P2P 行为)	‘0x32000000’
Kazaa(P2P 行为)	‘X-Kazaa’, ‘Kazaa’,
BT(P2P 行为)	‘0x13BitTorrent protocol’
DirectConnect (P2P行为)	‘$Sen/$Get/$Fit’
Fasttrack(P2P 行为)	‘Get ./hash’, ‘GIVE’
PPLive(P2P 行为)	‘0xe903’,‘0x98ab0102’
PPStream(P2P 行为)	‘0x2104430000a20001’
QVOD(P2P 行为)	‘0x0000040d5b’
UUSee(P2P 行为)	‘0x14070b04’,‘0x13500709’
Games(Server 行为)	Warcraft(port3724), CounterStrike (port27000-27020)等
Chat(Server行为)	QQ(port4000或8000), MSN(port1863)等
Attack(病毒行为)	Snort特征集[11]

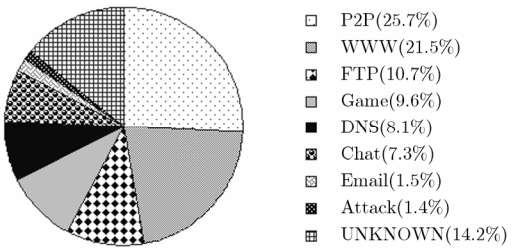

图4 Campus Trace流量数据业务构成

5.3 分类准确性仿真

(1)准确性的比较 针对整个 Campus Trace 样本集，图 5 仿真了 NBC, BLINC 和 HCG 3 种算法的识别准确性。其中，不仅包括每个业务的分类准确性，还计算了算法的整体准确率(图中“All”代表整体准确率)。可以看出：相比 NBC 和 BLINC, HCG 算法的整体准确率最高，特别在 P2P 业务上表现出了良好的分类性能；在 Server 业务上(WWW, FTP, Game, DNS, Chat, Email 等业务)，HCG 和 BLINC 要好于 NBC 分类算法；在 Attack 业务上，NBC 要优于 HCG 和 BLINC。

这主要是因为：HCG 算法使用“用户行为+端口”的方式进行识别，这对于表现显著行为(P2P 行为、Server 行为、Attack 行为)的业务具有良好的分类效果；BLINC 方法使用匹配模式图的方法进行业务识别，其在 Server 业务上的模式图和 HCG 基于

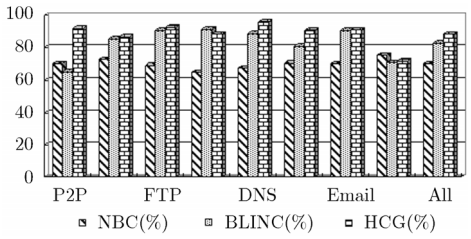

图 5 各业务分类准确性比较

信息熵定义的行为模式比较类似，所以两者在 Server 业务上表现出了相似的分类效果；在 Attack 业务上，HCG 只描述了 Scan 病毒行为，还不能识别 Worm 和木马等病毒。

(2)时间变化的影响　为了仿真算法性能随时间的变化情况，特将 Campus Trace 样本集按照时间分割成 24 份数据，每一份对应不同的一个小时内采集到的 10 分钟数据，并将其编号为 1 至 24。针对不同编号的数据流量，图 6 对 3 种算法进行整体准确率比较。其中 NBC 方法采用编号为 1 的数据为训练集(由于所有数据流均已事先正确标注，所以准确率为 100%)，编号 2–24 的数据作为测试集。可以看出：随着时间的推移，NBC 性能会大幅下降，HCG 和 BLINC 会有些许震荡，但不会有下降趋势。这主要是由于 NBC 分类算法完全依赖于流的统计信息，会产生概念漂移现象。而 HCG 和 BLINC 均基于用户的行为进行业务分类，不依赖于流的统计信息，能够克服概念漂移问题。

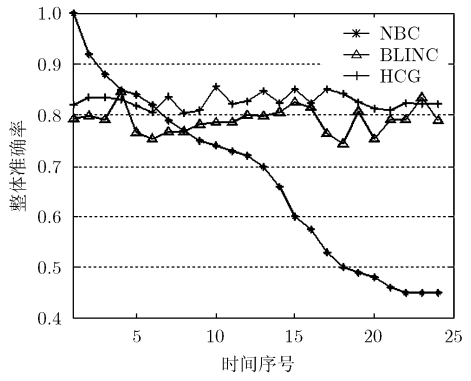

图6　整体准确率随时间的变化曲线

5.4 计算复杂度比较

(1)NBC 的计算复杂度　NBC 的处理过程主要包括提取流的统计特征和创建贝叶斯分类器：若令 Q 为统计的数据包个数，则提取流的统计特征平均需要 $O(Q)$[12]；假设 L 为类别标签的个数，N 为样本流的个数，文献[2]推导了创建贝叶斯分类器的复杂度为 $O(L\times N)$。因此，NBC 的计算复杂度为 $O(Q+L\times N)$，其中 $Q>>N$。

(2)BLINC 的计算复杂度　BLINC 算法的计算复杂度主要由两部分组成：构建用户交互模式图和对模式图进行逐一匹配。构建用户交互图涉及到图挖掘的相关知识，业界一般采用频繁子图挖掘技术来搜寻样本图集中模式图[13]。而目前最好的频繁子图挖掘算法的时间复杂度为 $O(N^4\cdot M\cdot 2^M)$[13]，其中 N 为数据流的个数(即图中点的个数)，M 为图中边的个数。另外，假设 BLINC 算法图库中包括 α 个模式图，则逐一匹配所需操作至少为 $O(\alpha N)$。因此，BLINC 算法的计算复杂度为 $O(N^4\cdot M\cdot 2^M+\alpha N)\approx O(N^4\cdot M\cdot 2^M)$。

(3)HCG 算法的计算复杂度　HCG 构建用户连接图的处理复杂度为 $O(M)$；谱聚类的计算主要消耗于用户相似度矩阵的本征分解上，若采用高效的 Lanczos 算法，计算量为 $O(N^{3/2})$[9]；基于“用户行为模式+端口”的用户结点标注和子簇业务标签映射的处理操作均在 $O(N)$级别上。因此，HCG 算法的计算复杂度为 $O(N^{3/2}+M+N)$，假设具有 N 个结点的完全图包含 $M=N(N-1)/2$ 条边，则 HCG 算法的计算复杂度可以表示为 $O(N^2)$。

基于以上分析可知：HCG 算法的计算复杂度要小于 BLINC 的计算复杂度，而 NBC 则依赖于报文总数 Q。在高速发网络中，报文总数 Q 是一个非常大的值。并且网络流量一般服从重尾分布[14]，即少数流占总流量的大部分比率，因此，Q 正比于 N^β，且 $\beta>2$，即 $O(Q+L\times N)>O(N^2)$。针对以上 3 种算法，图 7 依托编号为 1–10 的数据集进行了计算复杂度的比较。可以看出：与 NBC 和 BLINC 不同，HCG 算法不需要提取流的统计特征、创建和逐一匹配连接模式图，而是通过谱聚类和信息熵的方法分别对用户进行子簇划分和业务标签映射，有效地降低了算法的计算复杂度。

6　结论

针对“概念漂移”问题，本文提出了基于用户连接图的 HCG 方法。HCG 方法将研究对象从传统的 IP 流转移到用户，将用户之间的相互通信抽象为

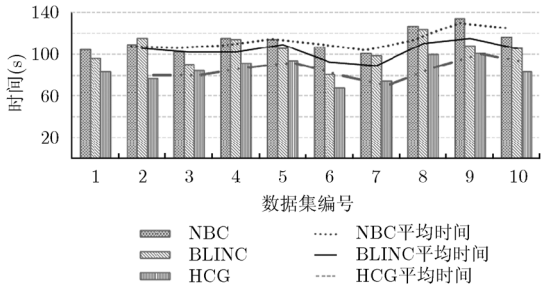

图 7 算法计算复杂度仿真

一种“社会团体行为”，构建了用户连接图模型；基于“图挖掘”的谱聚类将用户连接图划分为互不相交的行为子簇；引入“信息熵”对用户行为进行了刻画；设计了一种行为子簇到业务标签的自动映射机制，降低了算法的计算复杂度。论文还需进一步研究：由于 HCG 算法只能分析 Scan 病毒行为，对 Worm 和木马等病毒的检测，还需要对用户行为模式从不同的角度进行研究；另外，HCG 还可以尝试使用其它业务特征进行联合识别，比如 P2P 的 TCP/UDP 并发连接特征、Attack 业务报文的固定长度特征等。

参 考 文 献

[1] Zander S, Nguyen T, and Armitage G. Automated traffic classification and application identification using machine learning[C]. Proceedings of the 30th IEEE Conference on Local Computer Networks, Sydney, Australia, 2005: 250–257.

[2] Roughan M, Sen S, Spatscheck O, *et al.*. Class-of-service mapping for QoS: a statistical signature-based approach to IP traffic classification[C]. Proceedings of ACM SIGCOMM Internet Measurement Conference, Taormina, Sicily, Italy, 2004: 135–148.

[3] Williams N, Zander S, and Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification[J]. *ACM SIGCOMM Computer Communication Review*, 2006, 36(5): 5–15.

[4] Karagiannis T, Papagiannaki K, and Faloutsos M. BLINC: multilevel traffic classification in the dark [C]. Proceedings of SIGCOMM, Philadelphia, PA, USA, 2005: 229–240.

[5] Altman E, Nain P, and Shwart A. Predicting the impact of measures against P2P networks on the transient behaviors [C]. Proceedings of INFOCOM, Shanghai, 2011: 1440–1448.

[6] Jin Zhi-gang, Wang Ying, and Wei Bo. P2P Botnets detection based on user behavior sociality and traffic entropy function[C]. Proceedings of Consumer Electronics, Communications and Networks (CECNet), Yichang, 2012: 1953–1955.

[7] Saad S, Traore I, Ghorbani A, *et al.*. Detecting P2P botnets through network behavior analysis and machine learning [C]. Proceedings of Privacy, Security and Trust (PST), Montreal, QC, 2011: 174–180.

[8] 格雷. 熵与信息论[M]. 北京: 科学出版社, 2012: 10–100.

[9] Thedoridis S and Koutroumbas K. 模式识别 [M]. 北京: 电子工业出版社, 2010: 389–407.

[10] 朱云峰, 章毓晋. 直推式多视图协同分割[J]. 电子与信息学报, 2011, 33(4): 763–768.
Zhu Yun-feng and Zhang Yun-jin. Transductive co-segmentation of multi-view images [J]. *Journal of Electronics & Information Technology*, 2011, 33(4): 763–768.

[11] Snort. Network intrusion prevention and detection system [EB/OL]. http://www.snort.org, 2012.

[12] 鲁刚, 张宏莉, 叶麟. P2P 流量识别[J]. 软件学报, 2011, 22(6): 1281–1298.
Lu Gang, Zhang Hong-li, and Ye Lin. P2P traffic identification[J]. *Journal of Software*, 2011, 22(6): 1281–1298.

[13] 李先通, 李建中, 高宏. 一种高效频繁子图挖掘算法[J]. 软件学报, 2007, 18(10): 2469–2480.
Li Xian-tong, Li Jian-zhong, and Gao Hong. An efficient frequent subgraph mining algorithm[J]. *Journal of Software*, 2007, 18(10): 2469–2480.

[14] Zang Y, Breslau L, Paxson V, *et al.*. On the characteristics and origins of internet flow rates[C]. Proceedings of the Conference on Applications, Technologies, Architectures and Protocols for Computer Communications, New York, 2002: 309–322.

张 震： 男，1985 年生，博士生，研究方向为网络测量.

汪斌强： 男，1963 年生，教授，博士生导师，研究方向为宽带信息网络.

陈鸿昶： 男，1964 年生，教授，博士生导师，研究方向为计算机应用、程控交换技术.

马海龙： 男，1980 年生，讲师，研究方向为网络管理.