论 文

基于复杂网络挖掘的用户行为感知机制

张震*, 李玉峰, 王晶, 兰巨龙

国家数字交换系统工程技术研究中心, 郑州 450002

* 通信作者. E-mail: zhangzhenhigh@gmail.com

收稿日期: 2013-12-02; 接受日期: 2014-01-03

国家自然科学基金项目 (批准号: 61372121) 和国家高技术研究发展计划 (863 计划) (批准号: 2011AA01A103) 资助项目

摘要 用户行为感知是进行网络管理、安全检测以及应用趋势分析的基础. 针对基于流量统计特征检测方法具有计算复杂度高和“概念漂移”的缺陷, 提出了一种基于用户复杂网络图的用户行为感知机制算法 (UBP-CN). 算法将用户标识 {IP, Port} 和用户交互分别抽象为一个点和一条边, 构建了用户复杂网络图; 应用社团挖掘算法将复杂网络图划分为互不相交的行为子簇, 使得用户之间的通信抽象为一种“社会团体”; 通过定义基于相对熵的“用户行为模式”(UBM), 表征了各个子簇背后表现出的行为特性, 并使用“UBM + Port”对各个子簇进行标签映射, 实现了用户行为的有效感知. 仿真结果表明: 在不牺牲用户行为分类准确率的前提下, 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 用户行为感知 复杂网络 社团发现 用户行为模式 概念漂移 相对熵

1 引言

Everything over IP 思想推动了各种异构网络向互联网的融合, 互联网已经发展成为信息资源、存储资源以及计算资源共享的新型综合平台, 并朝着宽带化和高速化的方向发展. 但另一方面, 互联网在惠及人们工作生活的同时也带来了一系列亟待解决的问题, 具体表现为以下几点:

1) 互联网用户数量的规模化增长给网络管理带来了前所未有的挑战.

根据中国互联网络信息中心 (China Internet network information center, CNNIC) 第 32 次调查报告显示¹⁾: 截止到 2013 年 6 月底, 中国网民数量已达到 5.91 亿, 较 2012 年底新增网民 2656 万, 互联网普及率提升了 2 个百分点, 达到 44.1%. 由于采用开放的 TCP/IP 网络层协议, 用户规模将会继续增加. 在用户规模膨胀的背景之下, 由于硬件处理速度和后端存储能力的限制, 对用户行为特征的准确测量带来了巨大的挑战.

2) 网络应用类型日趋复杂, 给用户行为的精细化感知提出了新课题.

随着下一代网络 (next generation network, NGN) 和第三代移动通信 (3rd generation, 3G) 等电信技术的出现, 各种新的网络应用层出不穷, 互联网的运营环境发生了很大变化. 网络视频、网络购物以及网上支付等电子商务类应用的用户规模增幅明显; “微博”和网络视频用户增长保持强劲势头. 随着

1) 中国互联网络信息中心. 第 32 次中国互联网络发展状况统计报告. <http://www.cnnic.net.cn>, 2013.

引用格式: 张震, 李玉峰, 王晶, 等. 基于复杂网络挖掘的用户行为感知机制. 中国科学: 信息科学, 2014, 44: 1069–1083, doi: 10.1360/N112013-00081

“三网融合”进程的不断推进²⁾, 用户可以通过电信固定宽带或拨号、电信 3G 无线网络、广电网络等多种方式接入, 势必衍生出很多新的业务应用, 导致用户上网行为呈现多样化特点. 而这些用户业务多为新型应用, 其特征尚不明确, 难以有效感知.

3) 用户行为的多样化发展趋势, 对网络运营维护产生了较大影响.

从网络用户所使用的业务来看, 不同年龄、学历、职业的用户对业务的兴趣均有所不同. 网络运营商需要关注用户上网行为分析和用户业务趋向分析, 为制定相应的市场营销策略提供依据. 这对传统粗放式的网络运营方式提出了严峻的挑战. 与此同时, 当前网络基础设施也时常面临着拒绝式服务攻击、蠕虫传播、僵尸网络等安全威胁³⁾, 常常导致网络瘫痪, 造成了极大的损失. 由于缺乏有效的用户行为感知手段, 致使恶意信息能够通过网络广泛传播, 对网络的安全和可信带来了严重威胁.

面对用户规模快速膨胀和业务环境日趋复杂的网络形势, 迫切需要深入探寻网络行为的内在关系、规律和动态特性, 通过行为表征与建模、行为上下文感知与推理、行为特征挖掘、行为情感分析与推理、行为关联分析与演变分析等手段, 发现隐藏的可疑行为或者异常事件, 对于理解和认识网络行为规律、优化网络结构和改善网络 QoS(quality of service) 具有极其重要的意义.

基于流量统计特征的用户行为检测方法在业界得到广泛应用^[1]. 这种方法首先根据业务流特征(如每流字节数、每流报文数、流的开始时间、流的结束时间等)对流量进行统计测量, 并采用机器学习方法提炼流量特征规律, 应用自动分类和无监督学习技术识别用户行为特征. 基于朴素 Bayes 网络^[2]、支持向量机^[3]、神经网络^[4]等方法进行用户行为检测都属于此范畴. 但是, 该类方法的最大问题在于:

(i) 统计流量特征需要较高计算复杂度. 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, 特别在骨干链路百万级并发流的条件下, 实时提取每流统计值将面临巨大的存储和计算开销; 另一方面, 机器学习方法需要避免冗余属性和无关属性对分类结果的影响, 这也增加了算法的额外复杂度.

(ii) 概念漂移 (concept drift) 问题^[5]. 即 t 时刻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的动态变化特性: 网络时空环境发生变化、网络流量模式发生变化等.

针对流量特征的检测方法过分依赖网络环境和计算复杂度过高的问题, 基于用户行为的检测越来越受重视. 用户级检测从用户行为学的角度研究, 其最终目标是通过分析用户交互的行为的语义描述与理解. 因此, 本文的研究内容主要是行为抽象与定义、行为深度挖掘以及行为有效识别.

Williams 等^[5]提出了一种基于“交互行为”的用户行为感知方法——BLINC (blind classification), 该方法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, 并以“图匹配”的方式来划分网络用户. 图 1 给出了用户的各种连接模式图, 每一个图代表一个主机与其他主机间的交互行为^[6]. BLINC 方法利用了网络应用的行为属性, 具有良好的可扩展性. 但是, BLINC 方法需要构建和匹配连接模式图, 算法的计算复杂度相对较高; 需要事先确定连接模式图, 对于新出现的行为模式无能为力; 仅适合部署到单宿主边缘网络, 不适合对骨干网中的用户进行行为感知. Iliofotou 等^[7]认为同一时刻相互通信的结点间以很高的概率运行同一种互联网应用, 于是, 以网络中某些已知结点作为种子即可推测“相邻”结点的应用, 最终获得整个网络中所有用户结点的业务行为. 文献 [8,9] 根据用户行为特征对 P2P 僵尸网络等异常行为进行了检测, 取得了很好的效果.

2) 国家高技术研究发展规划. 面向三网融合的统一安全管控网络. <http://www.863.gov.cn/news/1011/05/1011053773.htm>, 2010.

3) CNCERT/CC. CNCERT 互联网安全威胁报告. <http://www.cert.org.cn/publish/main/>, 2013.

图 1 业务通信模式图, 修改自文献 [6]

Figure 1 Communication modes of services, modified from [6]

不同于 BLINC 方法, 本文提出了一种基于复杂网络挖掘的用户行为感知机制 (user behavior perception based on mining complex network, UBP-CN), 该方法将复杂网络社团发现算法和信息熵的概念应用于用户行为表征和划分, 其具有的优势如下:

(i) 将用户之间的通信抽象为一种“社会团体行为”, 构建了用户复杂网络图, 提出了“用户行为模式”的概念, 不再依赖于流量统计特征的任何信息.

(ii) 基于复杂网络 CNM 社团发现算法将用户复杂网络图划分为互不相交的用户行为子簇, 各个行为子簇背后都具有的某种共同社会行为特性. 因此, 不需要任何先验信息, 既能发现出新兴行为模式, 又使得行为感知算法不受网络环境的影响.

(iii) 引入“信息熵”概念对用户行为进行了刻画, 设计了一种子簇到行为标签的映射机制, 降低了算法的实现复杂度. 立足于实际的网络链路数据进行了仿真实验, 验证了 UBP-CN 算法的有效性.

2 相对熵

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^[10]. 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 $A = \{a_1, a_2, \dots, a_n\}$, 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, 可计算 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

$$H(X) = - \sum_{i=1}^n p(a_i) \log p(a_i). \quad (1)$$

上式中若 $p(a_i) = 0$, 则令 $0 \log 0 = 0$. 信息熵描述了事件发生的不确定性与事件发生的概率有关, 事件发生的概率越小, 猜测其发生的困难程度就越大, 含有的信息量也就越多. 相反, 事件发生的概率

越大, 不确定性就越小. 对于发生概率为 1 的必然事件, 就不存在不确定性, 因此所含的信息量也就等于 0. 当变量 X 的取值等概率分布时, 信息熵具有最大值^[10]: $H_{\max}(X) = \log \min\{n, m\}$. 一般情况下观测值数量 $m > n$, 则变量 X 取不同值的个数 n 满足 $2^{H_{\max}(X)}$. 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 $R(X)$:

$$R(X) = \frac{H(X)}{H_{\max}(X)} = \frac{H(X)}{\log \min\{n, m\}}. \quad (2)$$

$R(X)$ 表示了变量 X 的统计不确定性: 当 $R(X)$ 趋于比较小的值时, 说明 X 取值比较确定, 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; 当 $m > n$ 时, 当且仅当 $m_i = m/n$, 即 $p(a_i) = 1/n$ 时, 等式 $R(X) = 1$ 成立; 当 $m < n$ 时, 当且仅当 $p(a_i) = 1/m$ 时, 等式 $R(X) = 1$ 成立.

3 基于 CNM 的用户行为子簇划分

3.1 构造用户复杂网络

复杂网络是一种非线性结构图, 是由一些结点以及连接结点之间的边或线构成的. 在基于图的数据挖掘中, 它可以反映对象之间的相互关系. 复杂网络已成为当前最重要的多学科交叉研究领域之一, 已广泛应用于社会网络分析^[11]、生物网络分析^[12]、Web 社区挖掘^[13] 等领域. 基于此, UBP-CN 方法将用户之间的交互行为通过复杂网络 $G(V, E)$ 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⁴⁾. 下面给出了用户复杂网络图 $G(V, E)$ 的创建步骤.

步骤 1 构造 $G(V, E)$ 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. 区别于仅用 IP 地址标识用户的方法, 基于 $\{\text{IP}, \text{Port}\}$ 进行分析具有以下 4 方面优势: 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, 对网络安全分析具有重要意义; 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, UBP-CN 将 $\{\text{IP}, \text{Port}\}$ 作为用户的唯一标识, 并抽象为复杂网络图中的一个结点.

步骤 2 构造 $G(V, E)$ 的边. 复杂网络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. 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, 考虑以下原则对用户 A 和 B 构建一条边: 对于 UDP 流, 若 A 向 B 传送了第一个报文的时候; 对于 TCP 流, 若第一个 SYN 报文发送的时候. 另外, 由于基于无向图 (由没有方向的边组成) 的理论较为丰富, 且可以利用图聚类对用户行为进行分析, 所以将 UBP-CN 复杂网络图中的边看作是无方向的.

步骤 3 用户复杂网络的直观分析. 如图 2 所示, P2P 用户形成的网络拓扑在结点度分布、网络拓扑直径等几何性质方面明显不同于 Client/Server(C/S 模式) 模式下的 Web 用户, P2P 网络拓扑中

4) 某高速链路上 Web 和 P2P 用户的连接拓扑图请参考: Faloutsos M. Traffic monitoring and application classification: a novel approach. <http://www.ece.uci.edu/athina/netsys/08/slides/faloutsos-slides.pdf>, 2011.

图 2 UBP-CN 算法整体流程图

Figure 2 Flow chart of UBP-CN scheme

表 1 各种社团发现算法性能比较

Table 1 Comparing performance of classical community detection algorithms

Algorithm	KL	GN	Newman	CNM
Accuracy	Lower	Lower	Lower	Better
Apriority	Required	Not required	Not required	Not required

结点度 (关联于某一点的边的个数) 的分布更加均匀, 网络的最大直径也更大. 这是因为在 P2P 网络中, 每个结点同时扮演客户机与服务器的角色, 结点从服务器获得信息后亦可作为分布式服务器继续向外提供服务, 这种“病毒式”传播形成的网络拓扑直径必然大于 C/S 类应用单点传播形成的网络; C/S 类应用中, 服务器是唯一的信息源, 所有客户机均向其请求服务, 因此其网络拓扑中, 服务器结点的度异乎寻常的高, 客户机节点的度较低.

3.2 基于 CNM 的行为子簇划分

复杂网络通常会存在不同的社团, 社团内部结点连接相对紧密, 而社团间的结点只有少部分连接. 复杂网络挖掘旨在揭示出真实存在的社团结构、理解复杂网络的功能、发现复杂网络中的隐藏规律以及预测复杂网络的行为. 按照所采用的求解策略, 复杂网络社团发现算法可以分为两大类: 基于优化的方法和基于启发式的方法. 前者将复杂网络聚类问题转化为优化问题, 通过最优化预定义的目标函数来计算复杂网络的簇结构, 代表性算法是 KL 算法^[14]、Newman 算法^[15]、CNM 算法^[16]; 后者将复杂网络的聚类问题转化为启发式规则的设计问题, 代表性算法是 GN 算法^[17].

表 1 给出了各个典型算法的性能比较, 其中 m 和 n 分别表示网络的连接数和结点数. 由于 CNM 算法不需要任何先验知识, 且准确性较高; 并且, CNM 算法计算复杂度只有 $O(n\log^2 n)$, 已接近线性, 适合于大规模网络的社团发现. 因此, 在用户复杂网络 $G(V, E)$ 中, UBP-CN 将每个社团抽象为一个“用户行为子簇”, 并采用 CNM 算法对用户行为子簇进行划分.

CNM 采用了模块度来评估用户行为子簇的紧密程度, 模块度的含义是指网络中子簇内部结点的边所占总边数比例与某一随机网络中子簇内部结点的边所占比例的差值, 公式定义如下^[18]:

$$Q = \sum_i (e_{ii} - a_i^2), \quad (3)$$

其中, e_{ii} 表示子簇 C_i 中内部边占总边数的比例, $a_i = \sum_j e_{ij}$ 表示所有连接了子簇 C_i 结点的边占总边数的比例. CNM 基于贪婪算法思想对子簇进行划分, 从每个结点占据一个社团开始, 沿着使模块度增加最大或者减少最小的方向不断合并社团, 期望得到最大的网络模块度. 为了提高执行效率, 算法采用了 3 种数据结构^[18]:

1) 模块度增量矩阵 ΔQ .

ΔQ 的每一行都以平衡二叉树的形式存储, 每个元素 Δq_{ij} 都可以在 $O(\log n)$ 时间内定位, Δq_{ij} 表示子簇 C_i 和子簇 C_j 的模块度增量.

2) 最大堆 H .

H 中包含了 ΔQ 中每一行的最大元素, 以及该元素对应的用户行为子簇编号 i 和 j .

3) 向量 $\mathbf{a} = (a_1, a_2, \dots, a_n)$.

CNM 算法步骤如下^[18]:

步骤 1 初始化. 初始化网络每个用户结点为一个行为子簇, 令初始模块度 $Q = 0$, $a_i = k_i/2m$, k_i 为结点 i 的连接度数. 模块度增量矩阵 ΔQ 中元素 Δq_{ij} 初始化为

$$\Delta q_{ij} = \begin{cases} 1/2M - k_i k_j / (2M)^2, & \text{if node } i \text{ and } j \text{ connect,} \\ 0, & \text{else.} \end{cases} \quad (4)$$

步骤 2 合并. 从最大堆中合并选择最大的 Δq_{ij} , 合并相应的用户行为子簇 C_i 和 C_j , 并标记合并后的子簇为 C_j , 同时按照以下方式更新 Δq_{ij} , 最大堆 H 和向量 $\mathbf{a}$.

① ΔQ 的更新: 删除第 i 行和第 i 列的元素, 以下列方式更新第 j 行和第 j 列的元素:

$$\Delta q_{jk} = \begin{cases} \Delta q_{ik} + \Delta q_{jk} & \text{if } C_k \text{ connects to } C_i \text{ and } C_j, \\ \Delta q_{ik} - 2a_j a_k, & \text{if } C_k \text{ connects to } C_i, \text{ not to } C_j, \\ \Delta q_{jk} - 2a_i a_k, & \text{if } C_k \text{ connects to } C_j, \text{ not to } C_i. \end{cases} \quad (5)$$

②更新最大堆 H : ΔQ 更新后, H 中相应的每行元素也要更新.

③更新向量 $\mathbf{a}$: $a_i \leftarrow a_i + a_j$, $a_i \leftarrow 0$.

记录合并后的模块度值为: $Q \leftarrow Q + \max\{\Delta q_{ij}\}$.

步骤 3 重复步骤 2, 直至 $\max\{\Delta q_{ij}\} < 0$ 算法停止, 最终得到 l 个用户行为子簇.

4 基于信息熵的用户行为分析

基于 CNM 划分得到的用户行为子簇需要进行量化, 以便映射到相应的业务标签上, 达到用户行为感知的目的. 论文此节将首先利用“信息熵”的概念对行为进行定义, 并分析了几种典型的行为模式; 然后, 对行为子簇进行业务标签映射.

表 2 用户行为模式分析

Table 2 Analysis of user behavior mode

User behavior mode $M(\text{SrcIP})$	Behavior classification	Examples
$\{0,1,2\}, \{0,2,2\}$	Server behavior	Web, DNS, Email
$\{0,0,1\}, \dots, \{0,1,0\}$	P2P behavior	BitTorrent, eDonkey, Gnutella
$\{0,2,0\}, \{2,2,0\}$	Virus behavior	IP Scan
$\{2,0,2\}, \{0,0,2\}$	Virus behavior	Port Scan

4.1 基于信息熵的行为表征

在用户复杂网络图 $G(V, E)$ 中, 任何一条连接是由一对 $\{\text{IP}, \text{Port}\}$ 标识的. 分析某用户 $\{\text{IP}, \text{Port}\}$ 的连接行为时, 可将其假定为源端 $\{\text{SrcIP}, \text{SrcPort}\}$, 与其相连的用户认定为目的端 $\{\text{DstIP}, \text{DstPort}\}$, 即用 $\{\text{SrcIP}, \text{SrcPort}, \text{DstIP}, \text{DstPort}\}$ 四维元素来标识一条用户连接. 固定用户连接中的 SrcIP, 则 $\{\text{SrcIP}, \text{SrcPort}, *, *\}, \{\text{SrcIP}, *, \text{DstIP}, *\}, \{\text{SrcIP}, *, *, \text{DstPort}\}$ 的连接数将会呈现出不同值, 其中 “*” 代表任意值. 根据论文第 2 部分的有关定义, 分别计算相对熵 $R(\text{SrcIP}, \text{SrcPort}, *, *)$, $R(\text{SrcIP}, *, \text{DstIP}, *)$, $R(\text{SrcIP}, *, *, \text{DstPort})$. 为了方便描述, 特将 $R(\text{SrcIP}, \text{SrcPort}, *, *)$ 简写为 $R(\text{SrcPort})$, 其他依次类推. 由于 $R(\text{SrcPort})$ 表示在某一 SrcIP 下, 以 SrcPort 为变量的相对熵值, 所以 $R(\text{SrcPort})$ 实质上体现了 SrcPort 平均不确定性. 若 $R(\text{SrcPort})$ 趋于比较小的值时, 说明 SrcIP 只在少数几个源端口上的连接数较多; 若 $R(\text{SrcPort})$ 趋于比较大的值时, 说明 SrcIP 在每个源端口的连接数几乎均等; 同理可知, $R(\text{DstIP})$ 和 $R(\text{DstPort})$ 具有同样的性质. 基于此, 下面给出用户连接的行为模式定义.

定义 1 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 $R(\text{SrcPort})$, $R(\text{DstIP})$, $R(\text{DstPort})$, 并给出以下准则进行量化:

$$\hat{R} = \begin{cases} 0, & 0 \leq R(X) \leq \delta, \\ 1, & \delta < R(X) \leq 1 - \delta, \\ 2, & 1 - \delta < R(X) \leq 1. \end{cases} \quad (6)$$

则定义 $M(\text{SrcIP}) = \{\hat{R}(\text{SrcPort}), \hat{R}(\text{DstIP}), \hat{R}(\text{DstPort})\}$ 为用户 SrcIP 的行为模式.

基于定义 1, 表 2 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 $M(\text{SrcIP}) = \{0,0,2\}, \{0,1,2\}, \{0,2,2\}$, 则可认定此用户 SrcIP 为服务提供者 (即为 Server 角色). 这主要由于 $M(\text{SrcIP})$ 体现了 SrcIP 在使用固定的端口与某个、某几个与多数客户端进行交互, 交互的业务可以是 Web、DNS 或 Email 等.

2) P2P 行为. 若行为子簇中多数用户的行为模式满足 $M(\text{SrcIP}) = \{0,0,1\}, \dots, \{0,1,0\}$ 或 $\{1,0,0\}$, 则可认定该子簇表现为 P2P 行为. 其中, 多数用户的 $M(\text{SrcIP})$ 中每一维相对熵都小于 2, 表明行为子簇的连接图中结点度的分布比较均匀. 这正体现了 P2P 业务中每个用户既可以为 Server, 也可以为 Client 的特点.

3) Scan 病毒. 若某用户行为模式满足 $M(\text{SrcIP}) = \{0,2,0\}, \{2,2,0\}$, 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描.

表 3 用户结点标注流程

Table 3 Labelling steps of every user node

UBP-CN Flow chart of labeling user nodes with service tags	
1	Take out user nodes belong to behavior cluster C_i .
2	If all user nodes of C_i have been labeled completely.
3	Directly go to the 19th step.
4	Else search the user node u_j with the maximal degree among the unlabeled ones.
5	Computer the user behavior mode $M(\text{SrcIP})$ of u_j .
6	Switch the user behavior mode $M(\text{SrcIP})$.
7	Case server behavior, namely $M(\text{SrcIP})=\{0,1,2\}$ or $\{0,2,2\}$.
8	Based on normative ports, label u_j with service tags such as web, email, DNS, ftp and so on.
9	Based on familiar ports, label u_j with service tags such as game, music and so on.
10	Case virus behavior, namely $M(\text{SrcIP})=\{0,2,0\}$ or $\{2,2,0\}$.
11	Label u_j with IP scan.
12	Case virus behavior, namely $M(\text{SrcIP})=\{2,0,2\}$ or $\{0,0,2\}$.
13	Label u_j with port scan.
14	Case every quantitative entropy of $M(\text{SrcIP})$ is less than 2.
15	Label u_j with the suspicious P2P user.
16	Default label u_j with the unknown user.
17	Repeat executing the 3rd step.
18	End If
19	If the number of suspicious P2P users is more than T .
20	Label all suspicious P2P users with the P2P service tag.
21	Else Label all suspicious P2P users with the unknown service tag.
22	End If

4.2 行为子簇的标签映射

已知用户行为标签集合 $L = \{L_1, \dots, L_h\}$ 和行为子簇集合 $C = \{C_1, \dots, C_l\}$, 流量分类需要确定行为子簇集合 C 和行为标签集合 L 的映射关系, 以便对用户进行行为感知. 如表 3 所示, UBP-CN 算法对行为子簇进行标签映射的核心步骤如下:

步骤 1 寻求行为子簇中度数最大的用户结点 u_i .

UBP-CN 算法按照行为子簇中用户的度数大小进行标注: 一方面, 在用户复杂网络图中, 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 .

UBP-CN 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 1, 首先计算 u_i 的行为模式, 然后结合熟知端口号和常用端口号进行识别:

图 3 整体准确率随时间的变化曲线

Figure 3 Simulating the overall accuracy curve with diverse time points

1) 对于 Server 行为, 需结合熟知端口号进行业务识别, 如: FTP、DNS、HTTP、Email、Telnet 等; 非熟知端口业务需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, 则该聚簇表现为 P2P 行为, 并将所有“疑似 P2P 用户”标注为 P2P 用户。

3) 如表 2 所示, 只需分析用户行为模式, 即判定病毒型业务。

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行为模式。

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业务标注。

步骤 3 对聚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, 可以得到聚簇中每个用户的业务类型。行为感知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。UBP-CN 采用条件概率 $P(l = L_j | C_i)$ 来表示聚簇 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 $i \in [1, l], j \in [1, h]$ 。令 N_i^j 表示在聚簇 C_i 中属于类别 L_j 的用户总数, N_i 表示聚簇 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 $\hat{P}(l = L_j | C_i) = N_i^j / N_i$ 。基于此, 可得到聚簇的标签映射函数为

$$l = \arg \max_{j=1, \dots, h} \hat{P}(l = L_j | C_i). \quad (7)$$

映射函数就是将聚簇中包含最多用户的类别标签赋给该簇。若该聚簇中未知用户数目最大, 则认定其为“未知行为模式”。

4.3 用户行为感知详细流程

图 3 描述了 UBP-CN 算法的整体流程, 其主要步骤包括: 提取用户业务报文的 IP 地址和端口号、构造用户复杂网络图、基于 CNM 的用户行为划分、基于信息熵的行为模式定义、用户结点业务类型标注以及行为子簇的业务标签映射。其中, UBP-CN 算法的核心思想体现为: 将 $\{\text{IP}, \text{Port}\}$ 作为用户标识, 使得行为感知不依赖与应用层负载和流的统计信息; 基于 CNM 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

表 4 数据集详细信息

Table 4 Detailed information of data sets

Data sets	Link type	Date	Duration (h)	Traffic size (MB)
Campus Trace	100 M Internet access	2012-06-20 12:00	24	24
IDC Trace	GE China Unicom BGP line	2013-11-15 15:00	24	150

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, 进而实现了用户业务感知的目的.

5 实验结果及分析

本文采用基于流量统计特征的典型检测算法——朴素贝叶斯网络 (naive Bayesian network, NBN)^[2]、支持向量机 (support vector machine, SVM)^[3], 与 BLINC^[6] 和 UBP-CN 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 2 整体准确率

对于任意行为类型 $l_i \in L = \{L_1, \dots, L_h\}$, 类型 l_i 的分类准确率 P_i 等于被正确分类为 l_i 的样本用户总数 N'_i 除以分类前属于 l_i 的样本用户总数 N_i , 即 $P_i = \frac{N'_i}{N_i}$. 则整体准确率 P_{all} 等于所有被正确分类的样本用户总数除以样本总量, 即 $P_{\text{all}} = \sum_{i=1}^q N'_i / \sum_{i=1}^q N_i$.

定义 3 计算复杂度

针对某一数据集, 算法计算出所有用户标签所消耗的时间.

5.2 实验数据说明

实验中的流量数据分别采集于校园网和某互联网数据中心 (Internet data center, IDC) 机房, 我们对接入链路进行了旁路, 然后利用安装了数据捕获软件的电脑进行数据采集. 为了仿真不同时间段的算法性能, 观测了一天中 24 个小时的流量数据, 每个小时只采集了连续 10 分钟和 5 分钟的数据, 并将其按时间顺序编号为 1 至 24. 为方便描述, 校园网和互联网数据中心链路分别称为 Campus Trace 和 IDC Trace, 所采集的流量文件的大小分别 24 Mbytes 和 50 Mbytes 左右, 并且流量数据包含完整的数据包负载, 详细信息参照表 4.

文献 [6] 给出了 Web、FTP、Email 等多种主流业务的特征字段. 因为国内 Campus Trace 和 IDC Trace 与文献 [9] 的数据集采集环境不同, 所以本文在文献 [6] 所给出的特征字段基础上进行了扩充. 表 5 给出了各种扩充的业务特征字段, 其中病毒特征采用 Snort 库⁵⁾. 基于以上特征, 使用 DPI 方法对采集的数据报文进行业务识别和标注. 然后, 以 Campus Trace 和 IDC Trace 样本集作为基准, 计算各种仿真算法的准确性. 另外, 由于数据包无负载或者某些业务对数据包进行了加密, 某些数据部报文不能被识别标注. 为了更好地对比仿真, Campus Trace 和 IDC 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.

5) Snort. Network intrusion prevention and detection system. <http://www.snort.org>, 2012.

表 5 扩充业务行为特征说明

Table 5 Complementary interpretation for service characteristics

Service behavior	Characters
Gnutella(P2P behavior)	'Gnutella'
eDonkey(P2P behavior)	'0x319010000'
Thunder(P2P behavior)	'0x32000000'
BT(P2P behavior)	'0x13BitTorrent protocol'
Kazaa(P2P behavior)	'X-Kazaa', 'Kazaa'
DirectConnect(P2P behavior)	'\$Sen/\$Get/\$Fit'
Fasttrack(P2P behavior)	'Get ./hash', 'GIVE'
PPLive(P2P behavior)	'0xe903', '0x98ab0102'
PPStream(P2P behavior)	'0x2104430000a20001'
QVOD(P2P behavior)	'0x0000040d5b'
UUSee(P2P behavior)	'0x14070b04', '0x13500709'
Games(Server behavior)	Warcraft(port3724), CounterStrike(port27000-27020) and etc
Chat(Server behavior)	QQ(port4000 or 8000), MSN(port1863) and etc
Attack(Virus behavior)	Snort rule set

表 6 Campus Trace 业务分类准确性比较 (%)

Table 6 Overall accuracy of service classification for Campus Trace (%)

Algorithm	P2P	HTTP	FTP	Game	DNS	Chat	Email	Attack	All
NBN	65.05	83.78	75.02	75.02	58.44	70.28	70.21	60.33	70.01
SVM	60.09	85.33	80.65	89.36	73.66	75.90	82.04	62.35	76.17
BLINC	85.26	82.17	80.45	90.08	80.37	90.67	88.58	88.76	85.79
UBP-CN	91.59	86.45	85.31	92.04	85.66	90.37	90.70	91.54	89.21

表 7 IDC Trace 业务分类准确性比较 (%)

Table 7 Overall accuracy of service classification for IDC Trace (%)

Algorithm	P2P	HTTP	FTP	Game	DNS	Chat	Email	Attack	All
NBN	64.34	76.05	75.47	70.38	57.95	71.34	76.29	60.00	68.97
SVM	70.23	84.39	83.95	88.20	75.66	76.87	81.05	62.75	77.88
BLINC	83.47	80.83	81.80	90.17	79.85	90.40	86.73	85.30	84.81
UBP-CN	88.75	90.26	85.00	94.59	80.99	89.44	92.10	89.80	88.86

5.3 分类准确性仿真

1) 准确性的比较

针对整个 Campus Trace 和 IDC Trace 样本集, 表 6 和表 7 仿真了 NBN、SVM、BLINC 和 UBP-CN 3 种算法的识别准确性. 其中, 不仅包括每个业务行为的分类准确性, 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 NBN、SVM 和 BLINC, 算法 UBP-CN 的整体准确率

图 4 算法计算复杂度仿真

Figure 4 Simulating computational complexity of algorithms

最高, 特别在 P2P 和 Attack 行为上表现出了良好的分类性能; 在 Server 行为上 (WWW, FTP, Game, Chat, DNS, Email), UBP-CN 和 BLINC 具有相似的分类效果, 但两者都要好于 NBN 和 SVM 分类算法。

这主要是因为: UBP-CN 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、Scan 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行为识别, 其在 Server 行为上的模式图和 UBP-CN 中基于信息熵定义的行为模式比较类似, 所以两者在 Server 行为上表现出了相似的分类效果。

2) 时间变化的影响

针对 IDC Trace 中 1 至 24 编号的数据流量, 图 4 对整体准确率和时间的变化关系进行了仿真。其中 NBN、SVM 方法采用编号为 1 的数据作为训练集 (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, NBN、SVM 性能会大幅下降, UBP-CN 和 BLINC 会有些许震荡, 但不会有下降趋势。这主要是由于 NBN、SVM 算法完全依赖于流量统计信息, 会产生概念漂移现象。而 UBP-CN 和 BLINC 均基于用户的行为进行分类, 不依赖于流的统计信息, 克服了概念漂移问题。

5.4 计算复杂度比较

1) NBN、SVM 的计算复杂度

NBN 和 SVM 的处理过程主要包括提取流的统计特征, 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 $O(Q)$ 。另外 NBN 和 SVM 分别需要创建 Bayes 分类器和计算支持向量, 因此, 两者的计算复杂度均大于 $O(Q)$ 。

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逐一匹配模式图。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^[19]。而目前最好的频繁子图挖掘算法的时间复杂度为 $O(n^4 \cdot m \cdot 2^m)$ ^[19], 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) UBP-CN 算法的计算复杂度

根据图 4, UBP-CN 算法复杂度的计算过程如下: UBP-CN 构建用户复杂网络图的处理复杂度为 $O(m)$; CNM 社团发现算法的计算复杂度为 $O(n \cdot \log^2 n)$; 基于“用户行为模式 + 端口”的用户结点标注和子簇行为标签映射的处理操作均在 $O(n)$ 级别上. 因此, UBP-CN 算法的计算复杂度为 $O(n \cdot \log^2 n + m + n)$, 假设具有 n 个结点的完全图包含 $\frac{n(n-1)}{2}$ 条边, 则 UBP-CN 算法的计算复杂度可以表示为 $O(n^2)$.

基于以上分析可知: UBP-CN 算法的计算复杂度要小于 BLINC 的计算复杂度, 而 NBN 和 SVM 则依赖于报文总数 Q . 在高速 IP 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^[20], 即少数流占总流量的大部分比率, 因此, Q 正比于 $n\beta$, 且 $\beta > 2$, 即 $O(Q) > O(n^2)$.

针对以上 3 种算法, 图 4 依托 Campus Trace 数据集中编号为 1~10 的数据集进行了计算复杂度的比较, 实验的计算机环境为: 处理器为 Core4 3.3 GHz, 内存 4 GB, 硬盘 1 TB, 操作系统为 Windows Server 2008 企业版. 可以看出: 与 NBN、SVM 和 BLINC 不同, UBP-CN 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 CNM 社团发现和熵的方法分别对用户进行子簇划分和行为标签映射, 有效地降低了算法的计算复杂度.

6 结论

UBP-CN 方法将研究对象从传统的流量统计特征转移到用户, 将用户之间的相互通信抽象为一种“社会团体行为”. 算法首先将 {IP, Port} 作为用户标识, 构建了用户复杂网络图模型; 其次, 应用 CNM 社团发现算法将用户复杂网络图划分为各个子簇, 反映了用户的“团体行为”; 然后, 引入“信息熵”的概念对用户行为进行了刻画, 提出了“用户行为模式”的定义, 并基于“用户行为模式 + 端口”对用户结点进行了标注; 最后, 设计了一种行为子簇到行为标签的映射机制. UBP-CN 方法不依赖于流的统计信息, 克服了“概念漂移”现象, 提高了业务行为分类准确性能; 同时, 算法也不需要提取流的统计特征、事先构造和逐一匹配用户模式图, 有效地降低算法的计算复杂度. 论文还需要进一步研究: 对用户行为模式从不同的角度进行研究, 使用其他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 Dainotti A, Pescapé A, Claffy K C. Issues and future directions in traffic classification. *IEEE Network*, 2012, 26: 35–40
- 2 Xiang C, Yong P C, Meng L S. Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees. *Pattern Recogn Lett*, 2008, 29: 918–924
- 3 Xu Q Z, Yang L X. A supervised local decision hierarchical support vector machine based anomaly intrusion detection method. *J Electron Inform Tech*, 2010, 32: 2383–2387 [徐琴珍, 杨绿溪. 一种基于有监督局部决策分层支持向量机的异常检测方法. *电子与信息学报*, 2010, 32: 2383–2387]
- 4 Ramadas M, Ostermann S, Tjaden B. Detecting anomalous network traffic with self-organizing maps. In: *Proceedings of 6th International Symposium on Recent Advances in Intrusion Detection*, Pittsburgh, 2003. 36–54
- 5 Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Comput Commun Rev*, 2006, 36: 5–15
- 6 Karagiannis T, Papagiannaki K, Faloutsos M. BLINC: multilevel traffic classification in the dark. In: *Proceedings of SIGCOMM*, Philadelphia, 2005. 229–240

- 7 Iliofotou M, Kim H, Faloutsos M, et al. Graph-based P2P traffic classification at the Internet backbone. In: Proceedings of IEEE INFOCOM, Rio de Janeiro, 2009. 1–6
- 8 Jin Z G, Wang Y, Wei B. P2P botnets detection based on user behavior sociality and traffic entropy function. In: Proceedings of Consumer Electronics, Communications and Networks (CECNet), Yichang, 2012. 1953–1955
- 9 Saad S, Traore I, Ghorbani A, et al. Detecting P2P botnets through network behavior analysis and machine learning. In: Proceedings of Privacy, Security and Trust (PST), Montreal, 2011. 174–180
- 10 Gray R M. Entropy and Information Theory. Beijing: Science Press, 2012 [Gray R M. 熵与信息论. 北京: 科学出版社, 2012]
- 11 Cao H H, Zhu J M, Pan Y, et al. Context-aware P2P mobile social network structure and discovery algorithm. Chinese J Comput, 2012, 35: 1223–1234 [曹怀虎, 朱建明, 潘耘, 等. 情景感知的 P2P 移动社交网络构造及发现算法. 计算机学报, 2012, 35: 1223–1234]
- 12 Farutin V, Robison K, Lightcap E, et al. Edge-count probabilities for the identification of local protein communities and their organization. Proteins, 2006, 62: 800–818
- 13 Ma W D, Li Y P, Ma J G, et al. Empirical study of region user behaviors for Web pages. Chinese J Comput, 2008, 31: 960–967 [马卫东, 李幼平, 马建国, 等. 面向 Web 网页的区域用户行为实证研究. 计算机学报, 2008, 31: 960–967]
- 14 Newman M E J. Detecting community structure in networks. Eur Phys J B, 2004, 38: 321–330
- 15 Newman M E J. Fast algorithm for detecting community structure in networks. Phys Rev E, 2004, 69: 066133
- 16 Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. Phys Rev E, 2004, 70: 066111
- 17 Girvan M, Newman M E J. Community structure in social and biological networks. P Natl Acad Sci USA, 2002, 99: 7821–7826
- 18 Guo S Z, Lu Z M. Complex Network Theory. Beijing: Science Press, 2012 [郭世泽, 陆哲明. 复杂网络基础理论. 北京: 科学出版社, 2012]
- 19 Li X T, Li J Z, Gao H. An efficient frequent subgraph mining algorithm. J Softw, 2007, 18: 2469–2480 [李先通, 李建中, 高宏. 一种高效频繁子图挖掘算法. 软件学报, 2007, 18: 2469–2480]
- 20 Zhang Y, Breslau L, Paxson V, et al. On the characteristics and origins of Internet flow rates. In: Proceedings of ACM SIGCOMM, New York, 2002. 309–322

User behavior perception based on mining complex network

ZHANG Zhen*, LI YuFeng, WANG Jing & LAN JuLong

National Digital Switching System Engineering and Technological Center, Zhengzhou 450002, China

*E-mail: zhangzhenhigh@gmail.com

Abstract User behavior perception is fundamental to network management, security diagnosis and application prediction analysis. Considering that detecting algorithms using flow statistics have deficiencies of high complexity and concept drift, a novel scheme called user behavior perception based on mining complex network (UBP-CN) is proposed. UBP-CN constructs a host complex graph with abstracting every node and every edge from the user identifier IP, Port and communications between users. Based on the theory of community detection, the host complex graph is partitioned into mutually exclusive behavior clusters which represent social communities. To nicely depict implicit traffic characteristics for every intersected community, user behavior mode (UBM) is defined by introducing the idea of relative entropy. Both UBM and protocol port are utilized to map the behavior label to every host. Experiments demonstrate that UBP-CN can circumvent the concept shift problem and gracefully ameliorate computational complication without sacrificing accuracy.

Keywords user behavior perception, complex network, community detection, user behavior mode, concept shift, relative entropy

ZHANG Zhen was born in 1985. He received Ph.D. degree in 2012 from PLA Information Engineering University, Zhengzhou city. Currently, he is a senior researcher at National Digital Switching System Engineering & Technological R & D Center. His research interests include network management, traffic measurement and network architectural.