基于会话的应用特征自适应提取

王变琴[1,2] 余顺争[1]
(中山大学信息科学与技术学院 广州 510006)[1] (中山大学东校区教学实验中心 广州 510006)[2]

摘 要 提取网络应用特征,对于准确地识别应用层流量,进一步提供差异性服务、QoS保障、入侵检测、流量监控以及计费管理等应用具有很重要的意义。然而目前还没有有效的应用特征自动提取方法。提出了一种自动提取应用特征的新方法,该方法能够从应用层的会话中提取频繁项集,经过冗余项过滤及基于识别率的自适应特征选择获取识别应用协议的最小特征集。采用识别率和正确率对所提取的特征进行评估。实验结果表明,该方法是有效的,所提取的特征具有准确性,能用于应用层流量的精确识别。

关键词 网络应用,特征自适应提取,频繁项挖掘,会话
中图法分类号 TP393 **文献标识码** A

Adaptively Extracting Application Signatures from Session

WANG Bian-qin[1,2] YU Shun-zheng[1]
(School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China)[1]
(Education & Experiment Center, Sun Yat-sen University, Guangzhou 510006, China)[2]

Abstract Accurate identification of application layer traffic is important for many Internet applications, such as provision of differentiated services, quality of service(QoS) guarantee, intrusion detection, traffic monitoring, accounting, and so on. However, there is no effective method which can automatically extract application signatures. In this paper, a novel method based on frequent item mining was presented, which can automatically extract frequent set from sessions of any application protocol, reduce redundancy of the frequent set based on adaptive filtering rules, and obtain the application signatures. Identification rate and precision rate were applied to verity the extracted signatures. Experiment results show that this method is effective and the extracted signatures are subtle. Therefore it can be used to accurately identify the corresponding application.

Keywords Network application, Adaptable extraction of signatures, Frequent item mining, Session

1 引言

准确识别网络流量是提供差异性服务(differentiated services)、QoS保障、入侵检测、流量监控及计费管理等应用的前提基础。然而,面对当今网络应用的快速发展,传统的端口(Port)识别法已逐渐失效。因此,近年来许多的研究工作致力于探索识别应用层流量的新方法。基于流(flows)行为特征统计的识别方法[1-3]不能进行应用的精确判定,而且判定结果滞后,难以用其对流进行实时跟踪控制。基于载荷(payload)的识别方法[4,5]准确度高,但是如何高效地、准确地获取应用特征,是这种方法面临的最大问题。

网络应用日新月异,种类繁多,不同应用有不同的格式规范,使其特征提取变得非常困难。通过查阅应用层协议的有效文档(例如,HTTP协议的标准文档RFC2616)寻找已知应用特征的方法,对于不公开文档的应用(例如,MSN应用)及不断出现的新应用无能为力。通过wireshark,tcpdump等工具对网络上采集的应用层数据进行对比统计来获取应用特征,虽然具有一定的自动化,但其寻找应用特征的效率与可信度较低,同时应用协议版本不断更新,新应用不断涌现的现实给这种半人工分析方法带来巨大挑战。

ACAS方法[5]将TCP数据流(flow)的前64Bytes作为特征向量,这种方法的缺点是特征向量的数量大、训练时间长。同时基于Naive Bayes, AdaBoost, Maximum Entropy等机器学习的识别方法本身就存在一定的误差和不确定性,并且随着应用数量增多误差将迅速增大。利用反向工程(reverse engineering)的识别方法[6-9]为应用识别提出了新的途径,但该方法需要了解应用协议报文格式的语法规范。基于正则表达式(regular expression)的应用识别(例如Linux中的L7-filter[10])相对于其它方法,其识别的正确性、效率有很大提高,但目前识别的应用数量有限,并且需要人工归纳应用特征来构造每种应用的正则表达式。刘兴彬等人[11]提出基于Apriori算法的流量识别特征自动提取方法,使特征提取可以基于

到稿日期:2010-03-24 返修日期:2010-06-19 本文受国家自然科学基金—广东联合基金重点项目(U0735002),国家高技术研究发展计划(863)(2007AA01Z449),国家自然科学基金面上项目(60970146)资助。

王变琴(1963—),博士生,高级工程师,主要研究方向为网络应用协议特征提取及其应用协议流量识别与跟踪;**余顺争**(1958—),教授,博士生导师,主要研究方向为Internet流量异常检测、网络应用个体行为规范与防御。

应用层流量自动进行。本文提出一种应用特征自适应提取方法，它以会话为研究对象，提取识别应用的最小特征集，取得了较好的效果。

2 问题定义

2.1 应用特征

定义 1(应用特征，application signature，AS) 即在应用层数据中频繁出现的并且具有位置特性的字节或字节组合。主要有两类：①应用层协议报头中的特征串，包括协议名称、版本号等。例如 HTTP 协议报文中的“HTTP/1.”代表 HTTP 协议的名称及版本号，BitTorrent 协议报文中第 2－20 字节的值“BitTorrent Protocol”代表 BitTorrent 的协议名称。②应用层协议控制信息中的特征串，包括命令码、状态码及定界符等，例如 FTP 协议报文中的命令码“PASS”、响应码“220”以及回车换行符号“\0x0d0a”等。

在通信过程中，应用特征具有如下特性：它们不一定出现在传递的每个报文中，但会频繁地出现在相应会话(定义 2)中；此外，应用特征还具有位置特性。一般情况下，特征出现在会话建立后的开始若干报文中，也有少数特征出现在会话结束的某个或几个报文中。特征在报文中的位置也有其特点：对于多数应用，特征出现在报文开始几个字节处(例如 SMTP 协议中的“HELO”，“250”等)，但在某些应用中报文结尾的几个字节是结尾标志码(例如 QQ 协议报文的最后一个字节为“0x03”)。

2.2 会话及会话子集

在同种网络应用的流量中，相应的应用特征会频繁地出现在各个会话中，同时其在会话中也有其位置特点，这就需要应用特征提取应以会话报文序列为单位进行。然而实际流量数据是多用户、多种应用会话的无序混杂数据，因此需要从捕获的训练 Trace 中提取完整的待测应用会话集，并且按时间顺序对其会话报文序列进行重组(reassembly)。

定义 2(会话，session，S) 即一次通信建立和结束之间所有报文构成的序列。待测应用协议的所有会话组成其样本会话集。

定义 3(会话子集，session subset，SS) 即具有相同二元组(source IP，destination IP)的会话构成的会话子集。

应用特征提取就是从应用层数据中提取能够代表某种应用的全部特征的集合。

3 应用特征的自适应提取

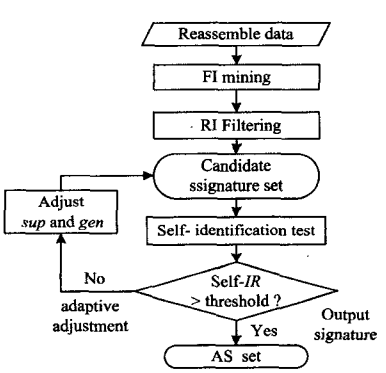

图 1 特征提取算法结构

本文提出的基于会话的应用特征提取算法由频繁项挖掘算法、冗余项过滤算法及自适应反馈调整机制的特征选择算法组成，如图 1 所示。频繁项挖掘算法能够从会话集中挖掘出满足最小支持度和通用度的频繁集；冗余项过滤算法根据设置的规则对频繁集中的冗余项进行过滤，产生候选特征集；自适应反馈调整机制基于自识别度(定义 8)反馈信息自动选择满足需要的特征。

3.1 频繁项挖掘算法

结合网络流量的特点，提出基于会话的频繁项挖掘(frequent item mining based on session，FI-mining-on-S)算法，此方法是对经典的 Apriori 算法[12]的改进，使其适用于应用会话的特征挖掘，其主要思想为：将待测应用的会话集视为交易数据库(transaction database)，其中的会话视为交易，在给定的支持度(定义 4)和通用度(定义 5)阈值下，挖掘出待测应用会话中的频繁项(frequent item，FI)。算法涉及的相关定义如下：

设 $I=\{I_1, I_2, \cdots, I_k\}$ 为项的集合，其中 I_i 为单字节项。对任意 X 满足 $X\subseteq I$，称 X 为一个项集。如果 X 包含 k 个项(或字节)，则称 X 为 k 项集，记为 *k-itemset*。设 $D=\{d_{ij}\}$ 为 I 上的单一应用会话数据库，其中 d_{ij} ($i=1,2,\cdots,m$; $j=1,2,\cdots,n$)为第 i 个会话子集中的第 j 个会话，记为 $d_{ij}=\{I_{ij1}, I_{ij2}, \cdots, I_{ijh}\}$，并且 $I_{ijl}\in I$，D 中共有 m 个会话子集，第 i 个会话子集有 n_i 个会话(令 $n=\Sigma n_i$ 代表会话集中会话总数)。

定义 4(支持度，support，记为 sup) 给定 D 和 X，称 $\sup(X)=n(X)/n$ 为 X 在 D 上的支持度，简记为 $\sup(X)$，其中 n 为 D 中会话总数，$n(X)$ 为 D 中包含 X 的会话数。

定义 5(通用度，generality，记为 gen) 给定 D 和 X，且 $\sup(X)\geqslant \min\_sup$，称 $gen(X)=m(X)/m$ 为 X 在数据集 D 上的通用度，简记为 $gen(X)$，其中 m 为 D 中会话子集的总数，$m(X)$ 为 D 中包含 X 的会话子集数。

定义 6(频繁项集，frequent itemset，FIS) 给定 D、X、最小支持度 $\min\_sup\in(0,1)$ 和最小通用度 $\min\_gen\in(0,1)$，当 $(\sup(X)\geqslant \min\_sup)\wedge(gen(X)\geqslant \min\_gen)$ 时，称 X 为 D 上的频繁项子集，D 上所有的频繁项子集组成的整个集合称为频繁项集，记为：$FIS=\{X\mid X\subseteq I\wedge(\sup(X)\geqslant \min\_sup)\wedge(gen(X)\geqslant \min\_gen)\}$。

性质 1 频繁项集的任何子集也是频繁的。例如，字符串 $(abc)\in 3\text{-}itemset$ 意味着其子串 $(ab)\in 2\text{-}itemset$，$(bc)\in 2\text{-}itemset$。

算法在原有支持度的基础上增加通用度，可以在一定程度上消除仅仅在少数会话子集中出现的频繁特征项(由特定用户行为引起)及其非特征的冗余 *FI*，以增强特征项的普适性，提高算法的挖掘效率(由于逐层搜索时候选项的减少)。

假设待测协议的会话集中有 m 个会话子集、n 个会话，最小支持度为 min_sup，最小通用度为 min_gen，利用 FI-mining-on-S 算法提取 FIS 的原理步骤如下。

(1)挖掘 1-*itemset*：计算 $\sup(1\text{-}item)$ 和 $gen(1\text{-}item)$，当 $(\sup(1\text{-}item)\geqslant \min\_sup)\wedge(gen(1\text{-}item)\geqslant \min\_gen)$ 时，将其列入 1-*itemset*；(2)利用 1-*itemset* 获取 2-*itemset*：计算 $\sup(2\text{-}item)$ 和 $gen(2\text{-}item)$，当 $(\sup(2\text{-}item)\geqslant \min\_sup)\wedge(gen(2\text{-}item)\geqslant \min\_gen)$ 时，将其列入 2-*itemset*；(3)由 *k-itemset*

获取($k+1$)-*itemset*($k\geqslant 2$)：对于 k-*itemset* 中的任意两项 l_1 和 l_2，如果满足连接条件，则将其合成一个($k+1$)-*item*，然后计算 sup(($k+1$)-*item*)和 gen(($k+1$)-*item*)，当[sup(($k+1$)-*item*))$\geqslant$min_sup]$\wedge$[(gen(($k+1$)-*item*))$\geqslant$min_gen]时，将其列入($k+1$)-*itemset*；如此进行，直到没有更长频繁字串为止，设最长频繁字串为 l；(4)获得 FIS：将 1-*itemset*，2-*itemset*，3-*itemset*，…，l-*itemset* 合成为 FIS，即 FIS=(1-*itemset*)$\cup$(2-*itemset*)$\cup$(3-*itemset*)…$\cup$(l-*itemset*)。

要挖掘出最完备的可能特征，在初次挖掘 FI 时，sup 和 gen 阈值要尽可能小，可以取最小值，即 min_sup=$1/n$，min_gen=$1/m$。但是在实际中，为了提高算法的效率，在大多数情况下，二者的初始值均可以设为 0.5(通过实验获得的经验值)，然后根据结果进行上下调整。

3.2 冗余项过滤

根据性质 1，FI-mining-on-S 算法会产生许多包含关系的 FI。设 FIS(x)(FIS(x)$\subset$FIS)是 FIS(y)(FIS(y)$\subset$FIS 的子串，则有[sup(FIS(x))$\geqslant$sup(FIS(y))]$\wedge$[gen(FIS(x))$\geqslant$gen(FIS(y))]。在一组包含关系的 FI 中，往往只有一项可能是特征项，其余项都可以视为冗余项(redundant item，RI)。例如 SMTP 流量中，如果特征码“250”被选中，则其子串“2”，“5”，“0”，“25”及“50”也一定会被选中，对此设置规则 1 对其进行过滤。

规则 1 若 FIS(x)(FIS(x)$\subset$FIS)是 FIS(y)(FIS(y)$\subset$FIS)的子串，并且[sup(FIS(x))=sup(FIS(y))]$\wedge$[gen(FIS(x))=gen(FIS(y))]，即 FIS(y)出现的次数与 FIS(x)相等时，则认为所有的 FIS(y)出现时都包含了 FIS(x)，此时过滤 FIS(x)。

此外，一些应用协议的请求/响应报文本身携带实体数据(例如 HTTP 协议)，或流量中的其它报文数据会引起高频数据项，这类冗余与应用特征的根本区别在于频繁特征在会话中出现的位置是较为固定的，而冗余的分布则一般是随机的。因此，在原算法的基础上增加对 FI 的位置标记，通过记录每个 FI 在报文中的偏移量，确定其分布的随机程度，据此区分特征和冗余。

规则 2 扫描频繁项 FIS(i)在每个报文中的偏移量，计算其位置自由度(定义 7)，通过设置其阈值(严格地讲，pos_fre(FIS(i))=$1/n\_meg$(FIS(i))，即固定偏移)对其进行过滤。

定义 7(位置自由度，position freedom，记为 pos_fre) 描述 FIS(i)在报文中出现的位置的随机性。假设包含 FIS(i)的报文数量为 n_meg(FIS(i))，FIS(i)在报文中出现的不同偏移位置数为 n_Os(FIS(i))，则 FIS(i)的位置自由度 pos_fre(FIS(i))计算公式为

$$\text{pos\_fre}(\text{FIS}(i))=\frac{n\_Os(\text{FIS}(i))}{n\_meg(\text{FIS}(i))}$$
$$\frac{1}{n\_meg(\text{FIS}(i))}\leqslant\text{pos\_fre}(\text{FIS}(i))\leqslant 1 \quad (1)$$

根据特征项在报文中具有相对固定的偏移这一特性，pos_fre(FIS(i))越大，说明 FIS(i)出现的位置随机性越大，其不是特征项的可能性越大。

特征项作为应用的标识，一般要求其具有应用唯一性，即仅在待测应用中出现，这就需要过滤具有负支持度(即在负例子集中的支持度)的 FI，因此设置了过滤规则 3。

规则 3 检查频繁项 FIS(i)(⊂FIS)在负例子集中是否出现。一旦出现，则将其从 FIS 中删除。

负例子的选取对规则 3 的过滤结果影响较大，通常选择与待测应用相近的或者容易混淆的应用。在实验中，通常将训练数据集中待测应用之外的其他应用流量作为负例子集，可能不能 100%地保证应用特征的唯一性。对于漏网的非唯一特征项，在混合应用流量测试阶段可以根据识别率和准确率的反馈进行排查消除。

经过以上规则的过滤得到不含冗余或冗余最少的应用特征集(signature set)，用于待测应用的识别。

3.3 自适应机制的特征选择

基于 FIS 的特征选择，其特征项的数量受 FI 的数量控制。sup 和 gen 的阈值设置直接决定 FIS 的规模。然而，实际中对于不同应用无法预知其最优值。虽然可以通过多次反复试验得出最优值，但这个过程需要人工参与，因此设计了自适应反馈调整机制使特征选择过程尽量自动化，以减少人工参与。

自适应机制首先由 FI-mining-on-S 算法产生 FIS，然后根据 FI 的 sup 和 gen 高低及其对待测应用流量的识别效果进行自适应特征选择。这样可以兼顾识别率和识别效率，选择出识别应用流量的最小特征集。图 1 显示了自适应特征选择的大致流程，其关键在于只需要进行一次 FI 挖掘，并在该次挖掘中提取尽可能多的 FI。具体的自适应过程如下：

(1) 输入 min_sup 和 min_gen，利用主算法 FI-mining-on-S 挖出 FIS；

(2) 利用过滤规则对 FIS 进行过滤，获得候选特征集；

(3) 从高到低对 sup 和 gen 进行自适应调整，其初值均设为最高值(100%)；

(4) 从(2)的输出结果中选择满足 sup 和 gen 的准特征；

(5) 利用准特征对待测应用流量进行自识别，若自识别率(定义 8)低于阈值，则 gen 不变，将 sup 调低一个阶度，重复(4)；若 sup 调到最低值($1/n$)后仍未达到要求，则 gen 调低一个阶度，重复(4)。

对 sup 和 gen 的调整过程实际为两重循环：外循环从高到低历遍调整支持度，内循环从高到低历遍调整 gen。每一个 gen 阶度为 $1/m$(m 为会话子集数)，每一个 sup 阶度为 $1/n$(n 为会话数)。

定义 8(自识别率，self-identification rate，记为 self-IR)是指利用获得的准特征对待测应用会话(测试数据集)的识别程度。如果 self-IR 达到比较高的标准(经验值为 95%)，则认为得到的特征集是具有应用代表性的，能够准确地识别该种应用流量。

通过自适应调整机制获得的特征仅为准特征，最后还需要经过多应用综合识别测试来对其进行验证确认。

4 实验评估

选择了目前较为流行的 7 种应用(HTTP，FTP，SMTP，POP，MSN，BT and SSL，它们分别代表网页浏览类、数据传输类、邮件收发类、即时通信类、P2P 下载类和加密类)对算法的性能进行评估。实验环境为一台个人 PC 机(CPU：Intel Core 2 Duo E6550 2.33 GHz，内存：0.99GB)，其测试工具为 matlab7.1。

4.1 数据及其预处理

测试数据(见表1)主要源于中国教育与科研网(CERNET)某小区的网络出入口处的真实流量,有少量训练数据(FTP应用)是在网络实验室中仿造的模拟流量。数据均用Wireshark采集,数据内容包括头部信息(header)和载荷(payload)。

表1 训练数据集和测试数据集

Application	Train dataset			Test dataset	
	Size(Mb)	SS	S	size(Mb)	S
HTTP	19.30	166	220	17.76	250
FTP	20.53	13	166	25.56	230
SMTP	15.08	89	163	10.10	154
POP3	18.20	6	128	16.80	110
MSN	04.63	82	139	03.06	211
BT	14.90	89	395	13.60	382
SSL	13.08	51	200	10.72	197

在特征提取时,每次输入纯净的单一应用流量数据,挖掘出该应用特征。提取的应用特征集的完备性、可靠性要根据多种应用流量混合时的性能指标判定,同时可以根据各指标值的优劣对特征集进行优化。

4.2 应用识别效果

为了确保测试结果的可信性,测试数据全部来源于上述真实网络流量(见表1),采集时间为2008年3月。

识别测试是基于会话进行的,为此定义了2个相关度量指标:识别率(Identification Rate,IR)即应该正确识别的会话中有多少被正确识别,可以反映特征集的完备程度,其计算见式(2);正确率(Precision Rate,PR)即识别结果中有多少是正确的,可以反映特征集的区分度好坏,其计算见式(3)。

$$IR=\frac{TP}{TP+FP} \tag{2}$$

$$PR=\frac{TP}{TP+FN} \tag{3}$$

式中,TP(True positive)表示将 c_i 应用的会话识别为 c_i 应用的数量,FN(False negative)表示将 c_i 应用的会话识别为非 c_i 应用的数量,FP(False Positive)表示将非 c_i 应用的会话识别为 c_i 应用的数量。

采用特征匹配的识别方法。由于网络流量识别是在线实时环境,缓冲区读进的报文数量会随着时间的增长而增加,因此测试过程模拟了读取不同报文数量的识别效果,据此可以确定精确识别会话所需要的报文数。实验结果表明,当会话中的报文偏移N=4时(不含建立会话的协商报文),各种测试应用的度量指标几乎都达到最佳(见图2、图3)。此时,测试应用的识别率均在94.57%~100%之间(即漏报率<5.43%),并且多数大于97%(即漏报率<3%),表明获得的特征集是较为完备的。测试应用的准确率均在97.17%~100%之间(即误报率<3%),表明所提取的特征的大部分区分度(独特性)很好。产生高误报率的特征集,可能存在区分度较差的特征或特征冗余项,可以根据实际情况对其进行取舍。对于载荷加密的应用协议,其加密内容只针对数据部分,协议自身的控制命令信息部分一般不加密。*SSL* 应用流量的高准确率说明这种挖掘算法对类似 *SSL* 的加密应用同样有效。

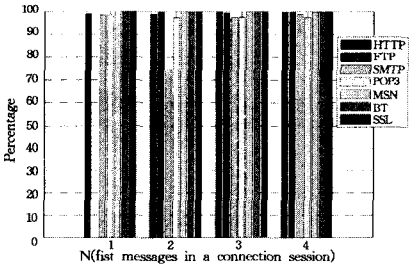

图2 识别率

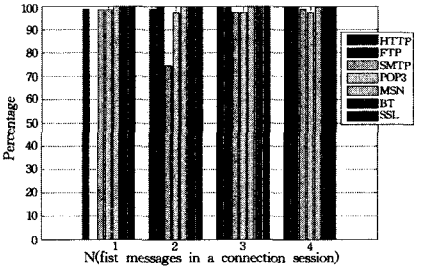

图3 正确率

实验进一步将本文方法(简称为 *AS-Mining_gen+sup*)与不含通用度的方法(简称为 *AS-Mining_Aprori*)进行了比较。准确率的比较结果见表2。多数情况下,二者有相近的识别率,但某些情况下(例如 *HTTP* 应用、*SMTP* 应用),*AS-Mining_gen+sup* 的正确率略高于 *AS-Mining_Aprori*,这是由于通用度的引入,在不降低识别率的情况下,能进一步减少冗余,精简特征集,保证较高的正确率。

表2 两种方法的准确率比较

Application	*AS-Mining_gen+sup*		*AS-Mining_Aprori*	
	IR%	*PR%*	*IR%*	*PR%*
HTTP	100	100	100	99.20
FTP	99.57	100	99.57	99.45
SMTP	95.10	98.55	94.59	97.80
POP3	98.55	97.14	98.55	97.14
MSN	100	100	100	100
BT	100	100	100	100
SSL	94.57	100	94.57	100

4.3 算法效率评估

特征提取算法除了提取的特征具有高准确度外,本身也需要具有较高的效率。图4对两种特征提取算法的时间性能进行了比较。从中可以看出,AS-Mining_gen+sup 算法与 AS-Mining_Aprori 算法相比,提取特征的时间并没有明显增加。相反,在某些情况下,有所缩短(例如HTTP应用、SMTP应用、FTP应用),这是由于算法在早期阶段排除了某些候选项,使连接过程减少,由此加速了算法的收敛。

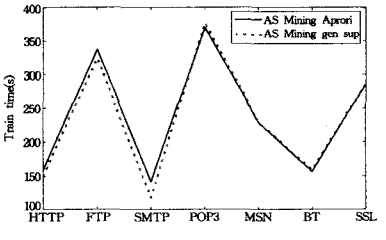

图4 两种方法的时间性能比较

结束语 本文研究了网络应用特征的提取,提出了基于

(下转第118页)

up 和 Primary with Bypass 预置 LSP，在“ Failure/Recovery”的属性中配置好链路的故障和恢复动作：链路 Cincinnati<->Washionton 设定在 400s 时失效，460s 时恢复（不要求业务恢复到原 LSP）。当网络运行时，起初数据流都通过主 LSP 进行传输，链路在 400s 故障时网络便通过快速路由改变数据流的通道路径，启用备用的 LSP，从而可以在不预先占用资源的情况下，通过快速重路由实现 LSP 路径的快速恢复，保护倒换时间只有新建 LSP 时间的 1/3（小于 50ms）；而通过局部路由器实现 Bypass Tunnel LSP 的旁路保护，恢复时间约为全局重路由恢复时间的 1/10，业务无明显丢失。图 10 为链路在 400s 故障时全局/局部快速重路由耗费时间的对比；图 11 为网络在 400s 故障前后主用 LSP/预置 LSP 中的数据流状况。

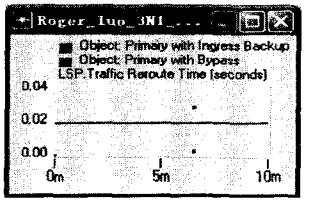

图 10 全局/局部快速重路由的耗时

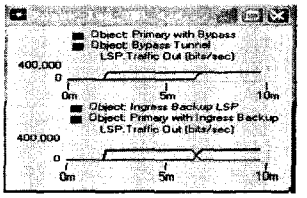

图 11 LSP 中数据流状况

仿真结果表明，RSVP-TE 信令协议能够快速处理网络元素的变化情况，可以在不预先占用资源的情况下，通过快速重路由到备用的 LSP 上工作，从而避免在 LSP 故障时数据资源的大量丢失。

结束语 本文讨论了基于 OPNET 的 G-LOBS 通用仿真平台开发的可行性及总体设计思想，最后运用 OPNET 软件对 RSVP-TE 运行效果进行了进一步的仿真分析，为深入开展 RSVP-TE 在 G-LOBS 网络中的研究提供了思路和蓝本。

参 考 文 献

[1] Ben Yoo S J. Optical Packet and Burst Switching Technologies for the Future Photonic Internet[J]. Journal of Lightwave Technology, 2006, 24(12): 4468-4492

[2] 李志先，张灵玲. OPNET 在网络仿真中的应用研究[J]. 洛阳理工学院学报：自然科学版，2009(3)：44-47

[3] 陈岩，董淑福，蒋磊. OPNET 网络仿真技术及其应用研究[J]. 计算机技术与发展，2009(02)：199-201，204

[4] Jue J P, Yang W-H, et al. Optical Packet and Burst Switched Networks: a review[J]. IET Communication, 2009, 3(3): 334-352

[5] 罗洪斌. OBS 网络边缘节点实现方案及相关技术研究[D]. 成都：电子科技大学，2004

[6] 邢子杰，孙卫强，金耀辉，等. 基于 RSVP-TE 信令的 GMPLS 动态性能研究[J]. 光通信技术，2009(08)：36-38

[7] 伍杰明. 智能光网络信令技术[J]. 电脑与电信，2009(05)：35-36

[8] 杨春勇. GMPLS 智能光网络中波长路由器的研究[D]. 武汉：华中科技大学，2005

（上接第 85 页）

应用会话的特征的自动提取方法。实验结果表明，提取的应用特征序准确、可靠，可用于应用层流量的实时精确识别。本方法的最大优势是不需要应用协议先验知识，可用于未知应用和新出现的应用特征的提取；同时仅依据应用流前期报文就可以识别应用类型，可以满足实时性要求，有利于流量的控制与跟踪；此外，算法也比较简单，可操作性强。然而，方法的研究还不够精细，算法的复杂性及优化问题将值得进一步探讨研究。

参 考 文 献

[1] Karagiannis T, Papagiannaki K, Faloutsos M. BLINK: Multilevel Traffic Classification in the Dark[J]. Proceedings of the 2005 Conference on Applications, Ttechnologies, Architectures, and Protocols for Computer Communications, New York, NY: ACM, 2005, 35(4): 229-240

[2] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly[J]. ACM SIGCOMM Computer Communication Review, 2006, 36(2): 23-26

[3] Crotti M, Dusi M, Gringoli F, et al. Traffic Classification Through Simple Statistical Fingerprinting[J]. ACM SIGCOMM Computer Communication Review, 2007, 37(1): 1-16

[4] Sen S, Spatscheck O, Wang Dongmei. Accurate, Scalable In-network Identification of P2P Traffic Using Application Signatures [C]//Proc. of the 13th International Conference on World Wide Web. New York, NY: ACM, 2004: 512-521

[5] Haffner P, Sen S, Spatscheck O, et al. ACAS: Automated Construction of Application Signatures[C]//Proc of the 2005 ACM SIGCOMM Workshop on Mining Network Data. New York, NY: ACM, 2005: 167-202

[6] Caballero J, Yin Heng, Liang Zhenhai, et al. Polyglot: Automatic Extraction of Protocol Message Format Using Dynamic Binary Analysis[C]//Proc of the 14th ACM Conference on Computer and Communications Security. New York, NY: ACM, 2007: 317-329

[7] Cui Weidong, Kannan J, Wang H J. Discoverer: Automatic Protocol Description Generation from Network Traces[C]//Proc of 16th USENIX Security Symposium on USENIX Security Symposium. Berkeley, CA: USENIX Association, 2007

[8] Ma J, Levchenko K, Kreibich C, et al. Unexpected Means of Protocol Inference[C]//Proc of the 6th ACM SIGCOMM Conference on Internet Measurement. New York, NY: ACM, 2006: 313-326

[9] 赵咏，姚秋林，张志斌，等. TPCAD：一种文本类多协议特征自动发现方法[J]. 通信学报，2009，30(10)：29-35

[10] Application Layer Packet Classifier for Linux [OL]. http://17-filter. Sourceforge. net

[11] 刘兴彬，杨建华，谢高岗，等. 基于 Apriori 算法的流量识别特征自动提取方法[J]. 通信学报，2008，29(12)：51-59

[12] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases [C]//Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publishers Inc, 1994: 487-499