基于流量特征的用户互联网访问类型识别①

陈 康[1], 黄晓宇[2,3], 陶彩霞[1], 关迎晖[1], 李 磊[3], 王爱宝[4]

[1](中国电信股份有限公司广东研究院, 广州 510630)
[2](华南理工大学 经济与贸易学院, 广州 510006)
[3](中山大学 软件研究所, 广州 510275)
[4](中国电信集团公司, 北京 100032)

摘 要: 近年来, 对互联网用户在网络上的行为分析研究吸引了广泛的兴趣, 分析的结果对网络运营商和普通用户都有重要的意义. 研究用户在网络上的访问行为的类型识别问题,分析了一个由 22 万个网络数据包组成的数据集, 从中提取统计特征, 设计用户网络访问的类型识别算法, 实验结果显示本文算法具有相当高的识别准确率.

关键词: 网络流量; 类型识别; 特征选择; 决策树

Traffic Features Based Categories Identification on Users' Network Behavior

CHEN Kang[1], HUANG Xiao-Yu[2,3], TAO Cai-Xia[1], GUAN Ying-Hui[1], LI Lei[3], WANG Ai-Bao[4]

[1](Academy of Guangdong Telecom Co. Ltd, Guangzhou 510630, China)
[2](School of Economics and Commerce, South China University of Technology, Guangzhou 510006, China)
[3](Software Institute, SUN Yat-Sen University, Guangzhou 510275, China)
[4](China Telecom, Beijing 100032, China)

Abstract: In recent years, the research on analyzing the users' network behavior has attracted much attention. In this paper, we study the problem of identifying users' network behavior categories. The research is based on a dataset that consists of 220 thousand network packets, with which we extract the statistical features needed for the identifications. We propose the identifying algorithm, and we also apply the algorithm to make categories identification on the network dataset. The results show the presented work can achieve very high accuracy.

Key words: network traffic; category identification; feature selection; decision tree

1 前言

近年来, 随着互联网上的内容服务的日益丰富, 对用户在网络上的访问行为的分析也引起了广泛的兴趣[1-4], 一方面, 对于网络运营商, 这些分析作为增强网络可控性的基础技术之一, 不仅可以帮助他们提供更好的服务质量, 而且能够对网络进行有效的监督管理, 确保网络安全; 另一方面, 即使对于一般的网络用户, 这些分析的结果也具有十分重要的意义. 如对于企业的管理人员, 他们可以由此得知员工在上班时间是否有访问与工作内容无关的网站; 对未成年人的父母, 他们也能够了解自己的孩子日常使用网络的主要用途. 然而, 注意到互联网上有数以亿计的内容服务应用, 而且新的应用还在不断的涌现, 因此, 要具体的掌握用户所使用的每一个网络服务的细节, 事实上是不可行的. 所以, 在当前的研究中, 主要关心的是如何识别用户在网络上的访问行为的类型, 如浏览网页、观看网络视频、网络游戏等.

对用户在网络上的访问行为类型的识别, 目前常用的有基于内容的识别和基于流量特征的识别两种类型. 对于前者, 其主要实现是对明文传输的用户通信数据作内容分析, 如以支持向量机[5,6]、贝叶斯方法[7,8]等手段作内容分类. 然而, 在实际的应用中, 这一策略

① 基金项目:国家高技术研究发展计划(863)(2012AA12A203)
收稿时间:2013-12-24;收到修改稿时间:2014-02-20

有非常大的局限性, 体现在: (1)它需要访问用户通信的明文数据, 因而严重侵犯了用户个人的隐私; (2)不同的网络应用通常采用自定义的数据格式, 并对敏感信息作了加密处理, 因而用户通信的明文内容也是难以获取的.

基于网络流量的识别是另一种被广泛使用的用户行为类型识别方法, 与基于内容的策略不同, 此类算法仅以用户产生的网络流量特征为依据进行识别. 根据算法使用的特征在网络协议中所处的层次, 这一类模型又可进一步细分为如基 Packet-level 特征的识别[9]、基于 Flow-level 特征的识别[10]以及基于 Stream-level 特征的识别[11]. 由于不需要分析用户数据的内容, 所以基于网络流量的用户行为识别在最大程度上减少了对用户隐私的侵犯, 而且也不受加密数据的影响[12-14]. 然而, 在当前的研究中, 这一类工作主要关心的是网络流量在整个互联网络上的分布而非特定的用户对象, 因而其结果也不能直接应用于对个体用户的网络访问行为分析中[4].

在本文的研究中, 我们将提出一种新的用户网络行为识别算法. 与现有的工作相比, 我们的工作以用户访问产生的网络流量为基础, 因而不受数据内容的限制; 此外, 本文算法关心的是对个体用户的行为识别, 因而也有别于已有的其它基于网络流量特征的识别研究. 此外, 我们还把本文的算法应用于一个由 22 万个网络数据包组成的数据集中进行用户的行为识别, 实验的结果显示本文算法具有非常高的准确率.

本文余下部分的组织为: 在论文的第 2 部分, 我们将以一个真实的网络数据包集合为对象, 研究用户网络行为的特征; 在论文的第 3 部分, 我们将提出本文的算法; 论文的第 4 部分是实验结果和分析; 最后一部分是全文工作的总结.

2 数据分析

本文研究所用的用户上网数据使用 Wireshark[15]抓取, 数据总量约为 22 万个数据包, 每个数据包都由报头与正文两部分内容组成, 其中, 报头部分主要包括时戳、源 IP 地址、目的 IP 地址、协议与数据包长度等信息. 根据报文的内容, 这些数据包可以分为五种类型: 网页浏览数据包(W)、网络游戏数据包(G)、在线视频(音频)数据包(V)、即时聊天数据包(T)以及由用户收发电子邮件或后台软件更新而产生的其它数据包, 具体的数量统计如表 1 所示.

表 1 网络访问数据集的类型组成

类型	W	G	V	T	Others
数量	51668	88114	70420	7231	1132

从表 1 可以看出, W、G、、V、T 这四类数据共占了所有数据包总数的 99.5%, 对余下的数据, 由于其数量不大且以软件的后台服务启动为主, 一般可以从数据包的报头中直接获知其服务类型, 所以我们工作中, 我们只关心对 W、G、V、T 这四种类型包的识别.

容易知道, 在不依赖于数据包中的数据内容的情况下, 仅以单个数据包的报头信息为依据, 是难以确定该数据包所属的类型的. 这一基础事实启发我们考察每一对“用户端 IP-服务器端 IP”(注: 在一般的描述中, 我们通常称之为“客户-服务器”, 但由于客户机与服务器两者间的地位是对称的, 所以在这里我们采用此表述方式)的连续交互行为.

我们首先对比以上 4 种访问行为在数据包长度上的特征, 为方便阐述, 在下文中, 对 W、G、V、T 每一种类型, 我们都分别从原始数据集中选取一对“用户—服务器”地址对, 对每个地址对间的通信数据包, 我们都分别截取了以用户端为始发 IP 和以服务器端为始发 IP 的 200 个连续数据包进行说明.

在图 1 中, 横坐标对应的是数据包的编号, 纵坐标对应的是数据包的长度; 红色折线对应的是由服务器端发出的数据包, 蓝色折线对应的数据包则由用户端发出. 其中, 1-1 是用户使用即时聊天工具(腾讯 QQ、淘宝旺旺等)产生的数据包, 1-2 是用户浏览网页产生的数据包, 1-3 是用户玩网络游戏产生的数据包, 1-4 是用户在观看网络视频(或收听在线音乐)产生的数据包.

从图 1 可以看出, 对 W、G、V、T 四种行为, 其数据包长度的分布具有显著的差异: 对于用户聊天产生的数据(图 1-1), 由用户端产生的数据包长度普遍较短, 而且此长度有一定的波动, 这主要是因为在即时聊天环境中, 用户普遍以短文进行交流, 但另一方面, 我们注意到服务器端始发的包长度明显大于客户端的包长度, 对于这一现象, 我们猜测这可能是服务端除了转发一般的聊天信息之外, 也可能向用户端主动推送了新闻与广告信息; 对于网页浏览的数据包, 从图 1-2 可以看出, 由服务端发生的数据包大部分都是满包状态, 这是由于当前大部分网页的页面元素都非常丰富, 除了正文内容外, 页面上通常还有大量的广告

与链接信息, 因而会产生大量的满载数据包, 相比之下, 由用户端产生的数据包的长度则普遍较小, 这是因为用户在浏览过程中, 主要的行为是请求打开新页面, 这一动作通常不会产生大的数据包; 对于用户在玩在线游戏产生的数据包(图 1-3), 可以看出, 由用户端产生的数据包的长度比较规则, 而服务器端产生的数据包长则有显著的“两极分化”的现象: 在 0～90 和 120～170 这两个编号区间内, 数据包的长度非常低, 但在 90～120 与 170～200 这两个区间内, 数据包的长度则有显著的增大, 我们猜测这是由于游戏地图被缓存的原因, 当玩家重复调用同一地图时, 该地图已被缓存到本地, 因而在网络上只传输了玩家的指令数据包, 但若玩家进入了一块新的区域, 则用户端需向服务端申请这一区域的地图, 此时网络的流量会有显著的变化; 对于用户观看在线视频的行为, 从图 1-4 可以看出, 几乎所有由服务器端发出的数据包都处于满载状态, 而用户端的包长则保持不变, 这一现象可以很容易由网络视频的特性获得解释: 对于前者, 这是由于在线视频为获得良好的用户体验, 因而它需要在每一次传输中传输尽可能多的内容, 对于后者, 由于用户在观看视频过程中不需要与其有过多的交互, 因而只需定期向服务端发送固定的状态信息即可.

除了数据包的长度之外, 我们还注意到, 对不同的网络应用, 它们的数据发送的时间间隔也有显著的差异, 如用户使用即时信息应用进行文字交流的数据包发生间隔应远大于他们观看在线视频时的产生的数据包间隔. 由此, 我们还分析了在 W、G、V、T 四种行为中, 各源端向目标端发送数据包的时间间隔, 对此, 我们仍以前文对包长度分析中所使用的数据包进行说明.

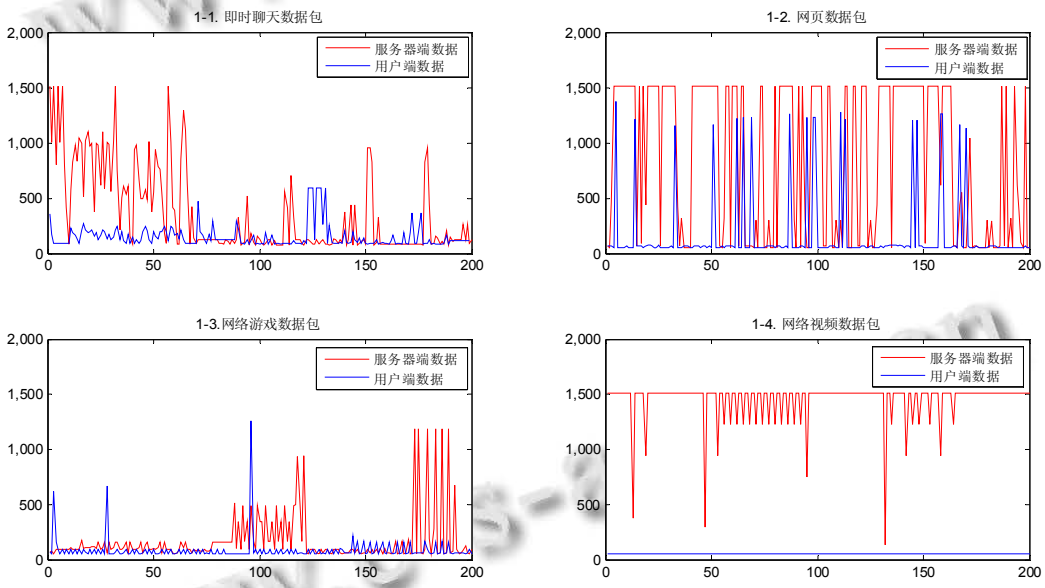

图 1 数据包的长度分析

我们的计算结果如图 2 所示.

与图 1 相类似, 图 2 中横坐标对应的是数据包间隔的编号, 纵坐标对应的是此间隔的长度, 由于各数据包间的间隔都非常小, 所以这里我们都取了它们的对数. 红色折线对应的服务器端的数据包间隔, 蓝色折线对应的则是用户端的包间隔. 图 2-1、2-2、2-3 与 2-4 分别与在线聊天、浏览网页、在线游戏与网络视频四种行为相对应.

从图 2 可以看出, 对于在线聊天与浏览网页这两种行为, 用户端的数据间隔与服务器端的间隔基本一致; 但对在线游戏与网络视频这两种行为, 这两者的数据间隔则出现了显著的区别, 特别的, 这一现象在图 2-3 中的表现更为明显. 对此, 我们认为, 对前两种行为(在线聊天与浏览网页), 都属于用户与服务器双方同步的交互行为, 因而它们的数据间隔也趋于相同; 而对后两种行为(在线游戏与网络视频), 为保证用户的体验, 服务器端需要向用户端频繁的发送大量的控制指令与视频数据, 因而其数据包间的时间间隔也更小.

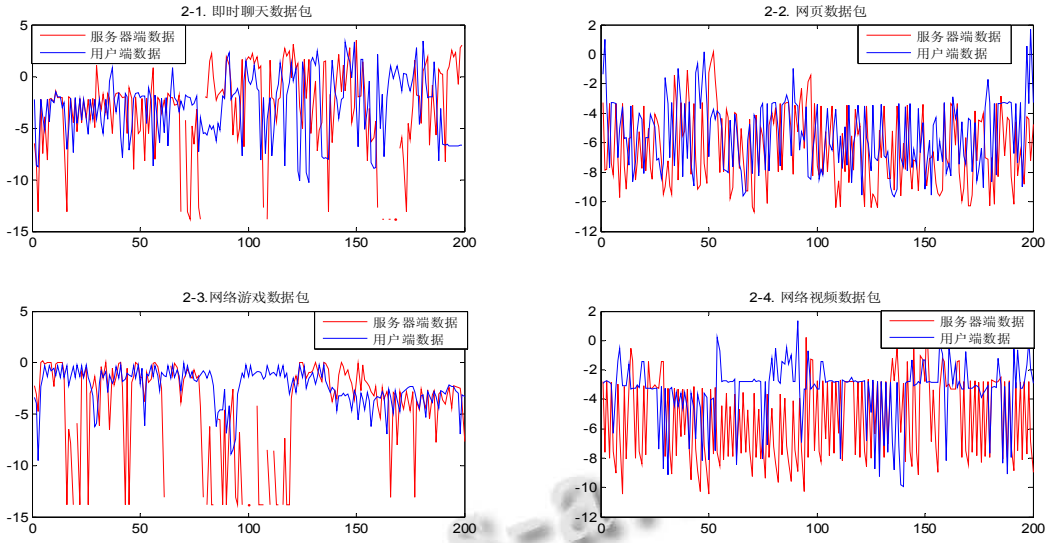

图 2　数据包产生的时间间隔分析

3　算法

根据前节的分析，我们可以知道，对接入互联网的用户，他的网络行为与数据包的长度和数据包的产生间隔都有非常密切的关系，因此，在我们的识别算法中，我们采用了表 2 的特征设计:

表 2　特征设计

特征	描述
S_Length	服务器端数据包的平均长度.
S_Length_Var	服务器端数据包的长度方差.
S_Interval	服务器端数据包的平均间隔.
S_ Interval _Var	服务器端数据包的间隔方差.
U_Length	用户端数据包的平均长度.
U_Length_Var	用户端数据包的长度方差.
U_Interval	用户端数据包的平均间隔.
U_ Interval _Var	用户端数据包的间隔方差.

其中，为计算表 2 中的各特征，我们设置了一个窗口阈值 Win，对每一对“用户—服务器”对<U, S>，我们都缓存它们间的连续通信数据包进行统计，每当缓存的数据包的数目达到 Win 时，则分别针对其中的由 U 和 S 发出的数据进行统计，并根据统计的结果生成一条特征记录.

根据以上的特征设计，我们首先使用已有的标注数据训练生成行为分类的决策树 TR: TR 的每一个非叶结点对应一个数据特征，树枝对应这个特征的取值，一个叶结点则代表了从树根到叶结点之间的路径对应的记录所属分类. 在决策树的构造过程中，选择合适的分裂结点是最为重要的步骤，这里我们采纳的是 Quinlan 在文献[16]中提出的基于信息增益(Information Gain)的策略.

基于训练获得的 TR，我们对每一对 “用户—服务器”地址对，使用表 3 的算法来识别其中的用户端的网络行为.

表 3　算法 1

00	输入：决策树 TR，训练 TR 的数据集 D，用户端 IP 地址，服务器段 IP 地址，窗口阈值 Win，记录数目 T，用户的行为类型集合{1, 2, 3, 4} 输出：用户端的网络行为类型.
01	初始化训练数据数目记录计数器 c=0，用户行为类型计数器 $t_1=t_2=t_3=t_4=0$;
02	初始化待标注数据集 D'={};
03	while t<T
04	缓存用户端与服务器端的数据包，当数据包数目达到 Win 时，生成一条无标记的特征记录 f;
05	令 D'=D'∪{f}, t=t+1;
06	end while
07	对 D'中的每一条无标记记录，使用 TR 对其分类和标注，若分类结果为第 i(i=1、2、3、4)类，则 t_i++;
08	记 m=max{t_1, t_2, t_3, t_4}，若 m 唯一，则转步骤 10;
09	缓存用户端与服务器端的数据包，当缓存数据包的总数达到 4W 时，生成 4 条无标记记录 f_1, f_2, f_3, f_4，令 D'=D'∪{f_1, f_2, f_3, f_4}，转步骤 07;
10	把 D'中所有数据的标记记为 m 所对应的行为类型;
11	若 t_1~t_4 中有>1 个非零值，令 D=D∪D'，并使用 D 重新训练 TR.

4　实验

本文实验所使用的数据在前文已有介绍. 根据算法 1 的描述，本文算法的关键在于决策树 TR 是否有足

够高的识别准确率. 因此, 在实验中, 我们首先检验决策树对独立的特征数据的识别准确率. 由于注意每条特征数据由 Win 个数据包统计获得, 所以这里分别令 Win 取不同的值以观察决策树的识别效果. 在这部分的实验中, 我们采用了 Weka[17]中 J48 的决策树实现, 算法的准确率是汇总了 10 折交叉检验的结果后得到的. 实验的结果如表 4 所示.

表 4 决策树识别准确率

Win	50	100	200
准确率	95.55%	95.88%	95.12%

从表 4 可以看出, 即使在单条特征记录的情况下, 决策树已能获得相当高的识别准确率. 进一步的, 当我们使用数据集中的一半数据用于训练决策树作为表 3 算法的输入, 而使用另一半数据作为测试时, 在以上三种 Win 值的设置下(Win=50, 100, 200), 识别的准确率都达到了 100%.

5 讨论及工作展望

在本文的工作中, 我们提出了一种与内容无关的用户在互联网上的访问行为识别算法, 我们采集了 20 多万个用户互联网访问的数据包, 通过对这些包的分析, 我们把用户的访问行为总结为四种不同的类型, 并分析了在这些类型的行为中, 相应的网络数据包的特征, 进而设计了用户行为的识别算法, 实验显示本文的算法具有非常高的准确率.

另一方面, 注意到在表 3 的算法中, 对每一对“用户—服务器”地址对, 我们都需要使用决策树来识别相应的用户行为类型, 显然在工程实现中, 这是一种非常低效的做法. 因而在我们的实现中, 我们还引入了一个服务器地址数据库, 这一数据库中包含了服务器地址与它所提供的服务类型间的对应信息, 对每一对 “用户—服务器”地址对, 若其服务器地址已经存在于地址数据库中, 则我们直接从数据库中读取它的服务类型作为用户的行为信息, 否则我们才调用表 3 算法来对用户行为进行识别, 并把识别的结果与服务地址添加到数据库中.

以本文算法为依据, 我们已经开发出了用户网络行为识别的原型系统, 在后续的工作中, 我们将对用户行为的类型作更进一步的细分, 以获得更细粒度的分析结果.

参考文献

1 Nguyen TTT, Armitage G. A survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys & Tutorials, 2008, 10(4): 56–76.

2 刘颖秋,李薇,李云春.网络流量分类与应用识别的研究.计算机应用研究,2008,5.

3 彭芸,刘琼.Internet 流分类方法的比较研究.计算机科学, 2007,34(8):58–61.

4 Moore A, Papagiannaki K. Toward the accurate identification of network applications. Proc. of Passive and Active Measurement Workshop (PAM2005). Boston (USA). 2005.

5 Sun AX, Lim EP, Ng WK. Web Classification Using Support Vector Machines. Proc. ACM Workshop on Web Information and Data Management (WIDM’02) with ACM CIKM. Virginia, USA. Nov. 2002. 96–99.

6 Zander S, Nguyenh, Armitage G. Automated traffic classification and application identification using machine learning. Proc. of the 30th IEEE Conference on Local Computer Networks Anniversary. Washington DC. IEEE Computer Society. 2005. 250–257.

7 Martin S, Sewani A, Nelson B, Chen K. Analyzing behavioral features for email classification. Email and Anti-Spam, 2005.

8 Moore A, Zuev D. Internet traffic classification using Bayesian analysis techniques. Proc. of SIGMETRICS’05. New York. ACM Press. 2005. 50–60.

9 Fraleigh C, Moon S, Lyles B, et al. Packet-level traffic measurements from the sprint IP backbone. IEEE Trans. on Networks, 2003, 17(6): 6–16.

10 Barakat C, Thiran P, Iannaccone G, et al. Modeling internet backbone traffic at the flow level. IEEE Trans. on Signal Processing (Special Issue on Networking), 2003,51(8): 2111–2124.

11 He T, Zhang H, Li X, et al. A methodology for analyzing backbone network traffic at stream-level. International Conference on Communication Technology Proceedings. 2003.

12 Roughan M, Sen S, Spatscheck O, et al. Class of service mapping for QoS: a statistical signature-based approach to IP traffic classification. Proc of IMC’04. Italy, Taormina. 2004. 5–27.

13 Erman J, Arlitt M, Anirban M. Traffic classification using clustering algorithms. Proc of SIGCOMM Workshop on Mining Network Data. New York. ACM Press. 2006. 11–15.

14 Bernaille L, Teixeira R, Akodjenou I. Traffic classification on the fly. Proc of ACM SIGCOMM Computer Communication Review. New York. ACM Press. 2006. 23–26.

15 www.wireshark.org

16 JR Quinlan. C4.5: programs for machine learning. 1993.

17 http://www.cs.waikato.ac.nz/ml/weka/