

基于移动终端的无线局域网用户行为研究

商正仪,梁羽燕,薛建宇,陈伟

(南京邮电大学 计算机学院,江苏 南京 210046)

摘要:由于无线空间传播信道所独具的开放性以及特殊辐射性,产生了许多隐私泄露的隐患。实验在公共环境下对无线局域网通信数据进行了还原,基于当前多类市场主流终端应用充分挖掘了无线网用户行为。在对用户及其行为建模之后,采用基于加权相似度的非监督聚类算法,进一步研究用户的社会性特征与潜在兴趣趋向。目的是明确用户的真正需求,增加用户之间的直接关联,便于设计更加贴近用户的协议与服务,优化改进无线网络机制。实验结果表明,该方法能够有效分析用户行为,聚类用户群体,有助于改善无线网用户使用体验,防范用户隐私泄露以及定制个性化网络服务。

关键词:无线局域网;用户行为分析;聚类;隐私分析;机器学习

中图分类号:TP393.1

文献标识码:A

文章编号:1673-629X(2018)03-0132-05

doi:10.3969/j.issn.1673-629X.2018.03.028

Research on User Behavior of WLAN Based on Mobile Terminals

SHANG Zheng-yi, LIANG Yu-yan, XUE Jian-yu, CHEN Wei

(School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210046, China)

Abstract: There are many hidden dangers of privacy leakage due to the openness and special radiation of wireless communication channel. The communication data from WLAN is restored in the public environment, and the wireless network users behavior is fully exploited based on the mainstream terminal application in current market. After modeling the users and their behavior, we adopt the non-supervised clustering algorithm based on weighed similarity for further study of the social characteristics and potential interest trend of users. We aim to clarify the user's real demand and increase the direct relation between users, in order to design the agreement and service which are more close to users, and optimize the wireless network mechanism. Experiment shows that this method can effectively analyze user's behavior and cluster user groups, which is help improve the user experience, prevent user's privacy disclosure and customize personalized web service.

Key words: WLAN; user behavior analysis; cluster; privacy analysis; machine learning

0 引言

如今智能终端设备对于有线网络基础架构的依赖逐渐减少,政府与企业也在大力推动建设城市公共场所的无线网络,使得无线局域网成为用户在固定场所下的最优网络解决方案。但公共场所的无线局域网保密措施较差,终端设备接入时可能引发的安全问题日益增多。密码攻击、脚本注入、会话劫持等恶意攻击方式均可以截获网络中的数据流量,导致用户个人信息遭到泄露^[1-2]。因此,通过监控网络流量,对用户行为进行分析,可以在提高第三方服务商个性化服务质量的同时,有效改善网络运行效率,为网络提供更加优质的管理。

在任何无线网络环境下,终端都扮演了一个重要

的角色,其对防范用户隐私泄露有重要的作用。已有研究表明,通过监控智能终端定期发送的广播探针请求可追踪用户^[3]并分析用户历史行为^[4];利用 Android 终端采集无线局域网资源并与后台服务器进行数据交互,可实现对无线网络的检测与分析^[5]。考虑到当前终端应用市场所呈现出的多样化、细分化的局面,无线网络中用户的社会关系相对于传统网络社会关系更加具有高维复杂性、环境感知性、关系隐藏深等诸多特性。因此,实验在现有研究基础上,引入近百种主流手机应用的特征参数数据库,实现还原用户基于终端应用的网络行为,统计常规网络信息,并分析用户群体特征与行为特征,有助于改善用户体验,创造用户与第三方服务商的共赢局面。

收稿日期:2017-03-07

修回日期:2017-07-20

网络出版时间:2017-11-15

基金项目:国家自然科学基金(61202353);大学生创新创业训练计划项目(SZD2016008)

作者简介:商正仪(1996-),女,研究方向为网络安全;陈伟,博士,教授,硕导,CCF 会员(E200025391M),研究方向为网络安全。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20171115.1438.062.html>

文中首先分析了无线网络机制下的数据采集技术和超文本传输协议,介绍了基于加权相似度的用户行为分析方法;然后设计实验,给出实验数据并做出分析,验证方法的有效性;最后提出了对所用方法的改进并总结了无线网络的使用特点,为防范用户隐私泄露提出了几点建设性意见。

1 相关工作

1.1 无线局域网数据采集技术

无线局域网使用无线通信技术将计算机等多种网络设备互联起来,构成可以实现数据通信与资源共享的网络体系。通常由无线站点(station,STA)、无线接入点(access point,AP)以及一些相关网络设备构成。STA一般是指智能终端或者配备无线网卡的PC机等,AP是指无线路由器、无线网桥或者无线网关,在无线网络中主要充当交换机的角色。

无线局域网采用单元结构,整个网络系统被划分成多个单元,每个单元被称为基本服务集(basic service set,BSS)。每个BSS由一个AP控制,此AP负责该BSS下所有STA的接入认证、网络通信以及流量控制。每个BSS的AP通过分布式系统(distribution system,DS)相连,组成扩展服务集(extended service set,ESS),使得STA可以在ESS内不同的BSS之间进行漫游。

当STA接入无线网络后,AP会以广播的形式进行数据消息的交换,这使得BSS范围内所有的STA均可以收到数据报文。在正常工作模式下,网卡虽然能收到网络中所有的数据报文,但需要将数据包的目的MAC地址与自身MAC地址进行比对,相同才接收并进行相应处理,不相同则直接丢弃。而无线局域网下的数据采集技术是将网卡设置工作在射频监听模式(混杂模式),使得网卡可以接收网络内所有正在传输的数据包,而不能发送数据包^[6]。为了获得较好的移植性,文中采用WinPcap进行数据采集^[7]。WinPcap可以在内核态直接对数据包进行预处理,与把数据包从内核中复制到用户空间中再处理相比较,提高了数据采集的运行性能。

1.2 超文本传输协议

现阶段手机应用与后台服务在进行数据通信时大多会选择超文本传输协议(hypertext transfer protocol,HTTP)。HTTP协议作为应用层的主要协议,采用请求/响应模型来传输包括文本信息与多媒体信息在内的所有资源。当用户向网站服务器请求服务时,只需要传送请求方法以及资源的路径,便可以获得相应的资源。

应用层协议规定的唯一资源定位符(uniform re-

source locator,URL)格式为HTTP://主机[“:”端口][路径],其中HTTP是表示通过HTTP协议定位网络资源,主机是表示因特网的主机域名或IP地址,端口是表示终端所使用的端口号,路径则表示指定资源的路径^[8]。因此,分析数据包中的URL,对资源进行重组,可以实现还原用户搜索词、用户浏览信息、用户访问网站、用户历史记录等多种用户行为^[9]。

1.3 机器学习

机器学习是指使用计算机程序来模拟人类学习的方式,获取新知识、新规则或者新技能,如今已经成为人工智能领域的一个重要组成部分。机器学习按照其学习形式,即数据集中经验包含的情况,可以分为监督学习和非监督学习两种。监督学习,需要对训练样本集中的给定样本提供确切的输出结果,根据训练样本进行学习,通常包含分类问题和回归问题。非监督学习,也成归纳学习,直接对无类别标签的数据样本进行识别,预测样本类型,其中最典型的一类问题就是聚类问题。

聚类是将数据集分成若干个簇,要求在某种度量标准下同簇内的相似度足够大,而不同簇间的相似度足够小。聚类根据数据间具有的某种潜在联系或者相关性,对数据集进行合理的组织以及摘要,便于发现数据集中的隐含信息。由于其具有的灵活性和自动化处理能力,广泛应用于搜索引擎、数字图书馆、数据分析等多个领域。文中采用高精度的K-means聚类算法,对用户行为关键词进行文本聚类,实现对用户行为的分析。

2 基于加权相似度的用户行为分析方法

互联网具有用户群体广泛、用户行为活跃、用户记录完整等多种社会特性,这为研究社会网络中的社会群体提供了一个较为理想的环境。网络用户行为包括用户在网络上发生的所有行为,如浏览、点评、搜索、社交媒体上的交流、购物趋向、收藏等^[10]。本实验主要是从搜索词、应用使用类型和网站浏览信息三个维度来描述用户行为,搜索词是指用户在搜索引擎中搜过的词句,应用使用类型是指用户使用各类型应用所占的比重,网站浏览信息是指用户所访问的网站及其浏览的信息。

为了从上述三个维度描述用户,实验归纳得到了不同应用对不同行为下URL的编码规则,并将这些规则导入特征参数数据库,使其作为匹配、解码并提取用户行为关键词的工具。特征参数数据库涵盖市场主流的近百种应用,文中将其分为表1所示的8类。这8类应用从性质、关注人群和侧重点上均有所不同,所还原的用户网络行为、各类资源的关注程度、使用流量类

型具有代表性,能够反映出个性用户群体的不同需求,从而有效地掌控全局用户的宏观行为。

表 1 应用类型及示例

应用类型	示例应用
社交类	QQ、微信
浏览器类	UC 浏览器、QQ 浏览器
电商类	一号店、淘宝
新闻类	今日头条、新浪新闻
娱乐类	网易云音乐、优酷、节奏大师
社区类	豆瓣、知乎
教育类	有道词典、作业帮
生活服务类	美团、大众点评

特征参数数据库中的每一条记录由应用名、行为类型、主机地址、路径格式、特征参数以及编码类型构成。通过在不同类型终端下的多次测试,数据库准确记录下同种网络行为的多种参数。表 2 给出了部分浏览器类应用的特征参数记录。以百度引擎为例,当用户使用百度引擎搜索某个关键词时,可以将 URL 与数据库中记录进行匹配,一旦匹配成功,便可用对应编码类型来解码特征参数后字符串,从而实现还原用户搜索词。

表 2 浏览器类常用特征参数

搜索引擎	主机地址	路径格式	特征参数	编码类型
360	so. com	/s?	q =	UTF-8
		/baidu?	word =	GB2312
百度	baidu. com	/s? word	word =	UTF-8
		/s?	wd =	UTF-8
UC	so. m.	/s?	q =	UTF-8
搜狗	sogou. com	/web?	query =	GB2312

基于加权相似度的用户行为分析方法的核心是挖掘同一网络下的用户之间的隐性连接关系和潜在兴趣趋向。此方法将改进后的词频-逆文档频率(term frequency-inverse document frequency, TF-IDF)与高精度的 K-means 聚类算法相结合,可将用户划分为联系更加紧密的团体。

TF-IDF 用来评估每个用户行为关键词对于整个数据集(全体用户行为关键词)的重要性,其核心思想是:关键词的重要性随着它在单个用户文本矩阵中(对单个用户行为关键词切分、去除停顿词后形成的矩阵)出现的次数呈正比增加,但会随着它在整个数据集中出现的频率呈反比下降。通过计算用户文本矩阵中所有关键词的 TF,可形成代表用户的词频向量,从而将对用户文本矩阵相似度的计算转化为对用户词频向量的计算。数学中常用余弦相似度来测量两个向量之间的夹角,因此改进后的余弦 IDF 值可以表示为:

$$\text{Cosine-IDF}(X,Y)=\frac{\sum_{x\in X\cap Y}\text{IDF}_x^2}{\sqrt{\sum_{x\in X}\text{IDF}_x^2}\sqrt{\sum_{y\in Y}\text{IDF}_y^2}}$$

其中, $\text{IDF}_i=\log\frac{1}{f_i}$ 表示关键字 i 在整体数据集中出现的频率^[11]。

实验聚类部分选择了基于划分的非监督 K-means 聚类算法。该算法先将数据集划分成若干个分组并初始化每个分组的簇中心,然后通过计算同一分组内每个点到簇中心的距离,不断改变分组直至方差达到最小标准,实现将数据集划分为 K 组具有相似实例的簇^[12]。作为非监督聚类算法,尽管 K-means 算法对大规模文本处理的精度比其他聚类算法较高,但是其关于 K 值以及初始化聚类中心点的选取仍会直接影响到聚类的优劣程度。针对 K 值的选取问题,文中使用肘部法则,而对初始化聚类中心点的选取,则采用了 K-Center 算法,使所有球型聚类簇的最大半径最小化,以获得更优的初始中心。

综上,基于加权相似度用户行为分析方法可以归纳为三步:第一步,匹配特征参数数据库,对用户应用层网络行为进行还原;第二步,使用改进后的加权相似度 TF-IDF 计算用户间相似度;第三步,对用户进行聚类,进行用户簇内与簇间的综合比较。该方法符合当前用户依赖终端应用上网的现状,可获得对单一用户、用户群体与全局用户较为贴切的分析。

3 实验设计及数据分析

3.1 实验设计

实验在南京邮电大学公共无线网环境下,利用 Winpcap 网络开发包启动无线网卡的混杂模式进行监听,捕获网络中应用层 HTTP 协议数据包。由于用户行为存在偶然性,不能单纯地使用某天某时的数据片面地定位用户,实验设定在学校固定时段固定地点进行长期性的数据采集。对于采集得到的数据包,提取其有效信息写成一条记录存入网络日志,供后续程序读取。日志格式设置如下:源 IP 地址@ #目的 IP 地址@ #源 MAC 地址@ #目的 MAC 地址@ #源端口@ #目的端口@ #Url@ #Cookie@ #转移地址 Refer@ #时间@ #数据长度。

考虑到每个 AP 都会使用动态主机配置协议(dynamic host configuration protocol, DHCP) 来为用户分配 IP,即当终端接入无线局域网时,AP 会从固有 IP 地址池中分配一个 IP 地址供用户使用,当用户退出网络时,AP 会回收此 IP 并重新分配给其他用户。因此,实验使用终端唯一的 MAC 地址来过滤网络日志,从而达到了标识用户的目标。

文中通过 Python 语言实现基于加权相似度的用户行为分析方法。使用 jieba 分词库提供的精确分词模式合理切分词句并去除停顿词,完成用户文本矩阵的构造;使用 sklearn 库完成 TF-IDF 相似度计算、K-means 聚类以及 PCA 降维工作;使用 matplotlib 绘图库完成对实验结果可视化的绘图工作。

3.2 数据分析

通过过滤和筛选,实验的数据集共涵盖 129 名用户在内的 10 492 条日志记录,每名用户的数据记录均超过 60 条,并涉及多类应用,能够从三个维度上较为准确地定位用户。

作为非监督聚类算法,K-means 算法的聚类效果直接取决于 K 值的选取,文中使用了肘部法则(elbow method)来解决这一问题^[13]。肘部法则是在 K 值依次确定($K=1,2,3\cdots$)的条件下,计算所有样本的畸变函数(样本点到其所在簇中心距离的标准平方和),然后将这些值连成一条曲线,如图 1 所示。随着 K 值的增多,簇数的增加会导致曲线总体呈下降趋势,但从某个位置开始下降得较为缓慢,如图中 $K=4$ 的位置,此处代表曲线的“肘”点,意味着达到最佳聚类。因此,实验选取 $K=4$,将整个用户群体划分为 4 类。

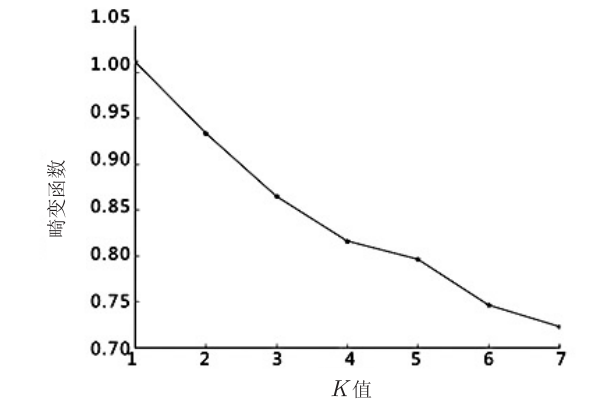


图 1 K 值的选取曲线

为便于将分析结果可视化,实验程序中使用 PCA 将多维数据降至二维,得到如图 2 所示的用户簇分布图。图中每一个点均代表一位用户,每个点的位置是根据用户间相似度得出的相对位置。观察可得,图中有 4 个较为集中的簇,簇与簇之间较为独立,用不同符号代表的每个簇均可代表一类具有相似兴趣爱好的用户群体。

表 3 给出了 4 位簇中心用户的部分文本矩阵,将表 3 与图 2 共同进行分析,可得到以下结论:

(1)就网络用户兴趣爱好进行分类,该数据集将所有用户划分成四类,分别是:影视娱乐类用户、综合类用户、时事新闻类用户和学术研究类用户。

5. 实验数据

(2)在各类用户文本矩阵中均存在类似于南京邮

电大学主页、查询课程、教务处一类的关键词,这类关键词属于整体用户集的共同特性。从聚类图中可以看出,这些关键词对聚类影响不大,所占权重较低,符合 TF-IDF 核心思想。

(3)图中综合类用户簇的位置处于其余三类用户簇的中间,比较综合类用户与其余三类用户的文本矩阵可发现均存在部分重叠,因此聚类图基本体现出数据集全体用户之间的隐形关系与潜在兴趣趋向,能够对网络进行综合性掌控。

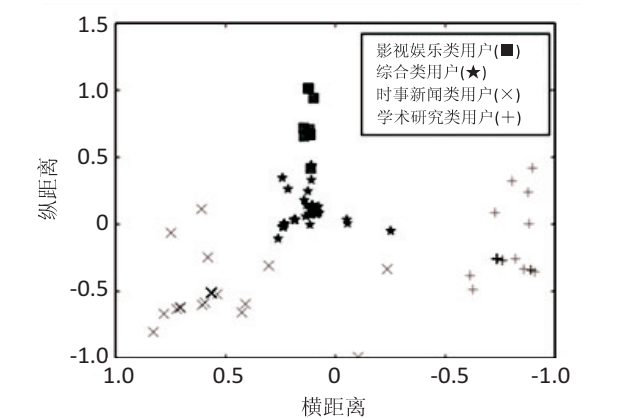


图 2 K-means 聚类

表 3 簇中心用户部分文本矩阵

用户 MAC 地址	标记符号	用户文本矩阵
SamsungE_bc:0f:XX	■	腾讯网 腾讯视频 热门 影视 发布会 摇一摇 欢乐颂 刘涛 海瑟薇 计算机 六级 南邮 教务处
97:65:c6:03:49:XX	★	百度 山水图片 百度文库 课程格子 零食 音悦台 习题 答案 高数 创新 阿里巴巴 导航 耳钉 淘宝
Apple_1b:15:XX	×	新浪新闻 腾讯体育 NBA 全明星 今日头条 ofo 课程格子 川普 总统 收购 Google 银行 柴静
Apple_a9:c9:XX	+	微信 当当 人间失格 百年孤独 试卷 追书神器 左耳 摇一摇 有机电子 建模 复试 实验 MATLAB

为评判聚类结果,实验根据用户文本矩阵人工对用户进行划分,并将所得数据与实验数据进行比较。对比发现,在上述四类用户簇中存在聚类的偏差,各类用户簇的准确度分别为 81.25%、85.71%、78.13% 以及 91.67%。综上,整体聚类实验结果的准确性可达 84.49%。

文中将实验结果通过友好的可视化图形展现出来,明确了不同网络用户人群的使用习惯及特征,这对未来管理、优化网络打下了良好的基础。同时,实验结论与所用方法的核心相一致,使得该方法的合理性以及有效性得到验证。

4 结束语

文中基于加权相似度的用户行为分析方法能够对当前网络环境进行全面监控以及分析,但同时其还原用户网络行为部分仍存在数据库不够完善,以至于不能成功匹配的情况。该部分内容可进一步通过长期的测试配合解密技术,扩充数据库内特征参数记录,解决还原失败的问题。实验还可以引入用户轨迹、用户定位等其他行为因素,加强用户群体的社会性分析。

此外,实验分析结果表明,在公共无线网络环境下用户的隐私安全并不能得到保障。由于恶意用户在公共网络下更容易实施攻击,用户需注意所接入网络是否为钓鱼接入点并注意在公共网络下不要对陌生软件进行授权,必要时用户可减少公共网络的使用频率。随着无线网络的不断发展,该方法不仅可以为用户提供更加贴切的个性化服务,还可以扩展至网络定位^[14]、网络监控^[15]以及网络取证^[16]等方面,对防范用户隐私的泄露和构建更加安全的局域网具有实际意义。

参考文献:

- [1] 任伟. 无线网络安全问题初探[J]. 信息网络安全, 2012(1): 10-13.
- [2] 朱建明, 马建峰. 无线局域网安全: 方法与技术[M]. 第2版. 北京: 机械工业出版社, 2009.
- [3] MUSA A B, ERIKSSON J. Tracking unmodified smartphones using Wi-Fi monitors[C]//ACM conference on embedded network sensor systems. Toronto: ACM, 2012: 281-294.
- [4] CUNCHE M, KAAFAR M A, BORELI R. I know who you will meet this evening! Linking wireless devices using Wi-Fi probe requests[C]//World of wireless, mobile and multi-

dia networks. [s. l.]: IEEE, 2012: 1-9.

- [5] 赵世功. 基于移动终端的无线局域网资源监测与分析系统的设计与实现[D]. 北京: 北京邮电大学, 2015.
- [6] 王智明. 无线局域网数据监听系统设计[D]. 北京: 北京邮电大学, 2012.
- [7] AN X G, LU X F. Packet capture and protocol analysis based on Winpcap[C]//International conference on robots & intelligent system. [s. l.]: IEEE, 2016: 272-275.
- [8] RESCHKE J. Use of the content-disposition header field in the hypertext transfer protocol (HTTP) [R]. [s. l.]: [s. n.], 2011.
- [9] 董志安, 吕学强. 基于百度搜索日志的用户行为分析[J]. 计算机应用与软件, 2013, 30(7): 17-20.
- [10] 毛佳昕, 刘奕群, 张敏, 等. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报, 2014, 37(4): 791-800.
- [11] CHENG N N, MOHAPATRA P, CUNCHE M. Inferring user relationship from hidden information in WLANs[C]//Military communications conference. [s. l.]: IEEE, 2012: 1-6.
- [12] 翟东海, 鱼江, 高飞, 等. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究[J]. 计算机应用研究, 2014, 31(3): 713-715.
- [13] BHOLOWALIA P, KUMAR A. EBK-means: a clustering technique based on elbow method and k-means in WSN [J]. International Journal of Computer Application, 2014, 105: 17-24.
- [14] 孙善武, 王楠, 陈坚. 一种改进的基于信号强度的 WLAN 定位方法[J]. 计算机科学, 2014, 41(6): 99-103.
- [15] 胡晓娅, 曹连杰. 无线局域网背景下的电子邮件监控系统研究[J]. 计算机工程与科学, 2010, 32(2): 63-66.
- [16] 胡东辉, 夏东冉, 史昕岭, 等. 网络取证技术研究[J]. 计算机科学, 2015, 42(10A): 1-22.

(上接第 131 页)

- 特征检测方法[J]. 计算机集成制造系统, 2011, 17(11): 2333-2342.
- [7] HASAN M M, MISHRA P K, MISHRA P K. Real time fingers and palm locating using dynamic circle templates[J]. International Journal of Computer Applications, 2012, 41(6): 33-34.
- [8] HUANG H, JU Z, LIU H. Real-time hand gesture feature extraction using depth data[C]//2014 international conference on machine learning and cybernetics. [s. l.]: IEEE, 2014: 206-213.
- [9] GANAPATHYRAJU S. Hand gesture recognition using convexity hull defects to control an industrial robot[C]//2013 3rd international conference on instrumentation control and automation. [s. l.]: IEEE, 2013: 63-67.
- [10] 李博男, 林凡. 基于曲率的指尖检测方法[J]. 南京航空航天大学学报, 2012, 44(4): 587-591.

- [11] PATHAK B, JALAL A S, AGRAWAL S C, et al. A framework for dynamic hand gesture recognition using key frames extraction[C]//2015 fifth national conference on computer vision, pattern recognition, image processing and graphics. [s. l.]: IEEE, 2015: 1-4.
- [12] TAN W, WU C, ZHAO S, et al. Dynamic hand gesture recognition using motion trajectories and key frames[C]//2010 2nd international conference on advanced computer control. [s. l.]: IEEE, 2010: 163-167.
- [13] 曹昕燕, 赵继印, 李敏. 基于肤色和运动检测技术的单目视觉手势分割[J]. 湖南大学学报: 自然科学版, 2011, 38(1): 78-83.
- [14] 王小俊, 刘旭敏, 关永. 基于改进 Canny 算子的图像边缘检测算法[J]. 计算机工程, 2012, 38(14): 196-198.
- [15] 王西颖, 戴国忠, 张习文, 等. 基于 HMM-FNN 模型的复杂动态手势识别[J]. 软件学报, 2008, 19(9): 2302-2312.