计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(8):265−269 [doi: 10.15888/j.cnki.csa.006424]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

移动端非显式用户身份信息的隐私问题研究①

张　颖, 代遵志, 滕彩峰, 张路桥

(成都信息工程大学 网络空间安全学院, 成都 610000)
通讯作者: 张　颖, E-mail: 2284846803@qq.com

摘　要: 智能设备给人们带来方便的同时也记录了大量使用者的使用习惯、位置、访问内容等隐私信息. 文章综合考虑用户行为数据的采集方式, 数据处理方式以及存储方式,自主设计用户行为数据的采集系统, 并在智能终端设备上采集用户的大量非显式身份信息数据, 包括网络流量信息、屏幕状态信息等, 通过对这些数据进行处理和分析, 发现利用这些非显式的用户身份信息可以有效对用户身份进行识别, 并能推测出部分用户间的社会关系. 实验表明, 基于非显式身份信息数据的研究对保护用户隐私有重要的现实意义和很大的应用价值.

关键词: 移动互联网; 隐私保护; 用户行为数据; 用户行为模式

引用格式: 张颖,代遵志,滕彩峰,张路桥.移动端非显式用户身份信息的隐私问题研究.计算机系统应用,2018,27(8):265–269. http://www.c-s-a.org.cn/1003-3254/6424.html

Research on Privacy Analysis of Mobile Implicit User Behavior Data

ZHANG Ying, DAI Zun-Zhi, TENG Cai-Feng, ZHANG Lu-Qiao

(Chengdu University of Information Technology, College of Cyberspace Security, Chengdu 610000, China)

Abstract: Smart devices bring convenience to people and also record a large number of user habits, location, access content, and other private information. In this study, considering the way of user behavior data collection, data processing and storage, we design a data acquisition system of user behavior data, and collect a lot of non-explicit identity data on the smart terminal equipment, including network traffic information and screen status information. Through analyzing and processing the data, we found that this non-explicit user identity information can effectively identify the user identity, and can figure out social relationships between the users. The experiment show that, based on the explicit non-identifying information to protect user privacy data has important significance and great value.

Key words: mobile internet; privacy protection; user behavior data; user behavior patterns

随着社会的发展, 人们对隐私保护的重视程度越来越高, 大家对隐私问题越来越敏感, 移动互联网隐私安全研究逐渐成为一个热点. 传统的隐私保护技术主要分为 3 类: 数据扰动技术、数据加密技术和数据匿名化技术[1]. 通过这些技术, 我们能有效保护个人隐私, 但是在日常生活中不经意间隐私信息可能会就被泄漏, 从用户的角度看, 微量隐私[2]的泄露也许并无严重后果, 但难以证实的是, 通过获取大量用户的数据, 利用机器学习分析用户行为习惯可能会得到用户其他隐私信息, 用户的身份信息, 以及社会关系, 现阶段这方面的研究是很少的.

为了更好的保护我们的隐私, 我们得知道隐私泄露的途径有哪些, 一个应用可以通过哪些信息得到我们的隐私数据. 显而易见, 隐私数据包括通信录, 通话记录, 聊天记录, 短信[3]等等这些有明显身份信息的显式信息, 而另一方面一个应用不仅可以收集上传信息,

① 收稿时间: 2017-11-02; 修改时间: 2017-11-27; 采用时间: 2017-12-07; csa 在线出版时间: 2018-07-28

也可以在不需要额外的权限下收集一些非显式的用户身份信息, 这方面是经常被忽略的, 而本文的研究目的就是去分析这些非显式的用户身份信息是否对用户的隐私产生威胁, 如果产生威胁的程度有多大, 这个衡量的标准取决于通过这些信息挖掘出用户身份的准确率[4]. 准确率越高那么产生的威胁越大, 也叫告诉我们在使用手机的过程中, 为了保护我们的隐私[5]也要额外警惕不良应用收集这些非显式的身份信息, 有需要的话可以利用 Android 系统权限限制, 禁止应用使用某些权限.

本文通过志愿者来采集智能手机使用过程中产生的不包含任何显式用户身份信息的用户行为数据, 通过分析数据对用户身份进行识别, 推测他们的社会关系, 并研究哪些数据会对用户隐私产生威胁, 而这些数据造成的隐私泄露不容小觑, 应该对非显式用户身份信息加以保护, 使用过程中加以限制, 帮助普通用户更好的保护隐私.

1 国内外研究现状

对于移动互联网隐私安全的研究很早就开始了, 并且对于移动用户信息隐私的泄露和保护的研究也逐渐成为一个热点. 文献[6]是利用手机传感器数据加上蜂窝网及 Wi-Fi 网络信息强度等参数实现对用户连接 AP 时长的预测. 文献[7]利用用户与 AP 关联数据实现对用户身份预测, 并证明了仅仅通过对设备编号的哈希等匿名化处理是不能够有效保护用户隐私的. 文献[8]利用手机使用过程中所产生的网络流量对用户网络使用行为进行统计学分析. 文献[9]利用了用户在不同类型地点连接网络时体现出来的网络使用方式的差异, 实现了用户所在地点类别的识别. 现阶段大多是利用 Wi-Fi 网络信息、AP 关联数据、用户手机使用行为等直接对个体或群体层面进行用户行为分析预测, 而关于非显式的用户身份信息对用户的隐私保护的研究还不完善, 本文就此进行研究.

2 用户行为数据处理流程

人们在智能手机使用的过程中会产生各种各样的数据, 绝大部分的数据本身是不包括和携带任何用户身份信息的[10,11], 这些原始信息如果没有经过特殊处理是不会对用户隐私产生威胁的. 但是通过采集大量用户数据建立用户行为特征库, 可以实现对用户身份信息进行识别, 甚至推测用户的社会关系.

2.1 移动端非显示身份信息数据采集阶段

数据采集阶段主要根据需求采集相应的非显式身份信息, 即这些信息是不直接标识用户身份的. 因为这些数据均与用户使用习惯和行为存在着直接或者间接的联系, 可以通过分析对用户身份进行识别, 并且一个正常的应用收集这些信息也不需要申请其他敏感权限. 这些信息详情见基础数据表 1.

表 1 基础数据表

序号	数据类别	数据名称	数据类型
1	网络流量信息	Wi-Fi 网络发送、接收数据包数量	int
2	网络流量信息	Wi-Fi 网络发送、接收的字节数量	int
3	网络流量信息	移动网络发送、接收的数据包数量	int
4	网络流量信息	移动网络发送、接收的字节数量	int
5	Wi-Fi 网络信息	Wi-Fi 网络连接状态	boolean
6	Wi-Fi 网络信息	SSID 名称	string
7	移动网络信息	移动网络类型	string
8	移动网络信息	移动网络连接状态	boolean
9	电池电量信息	电池剩余电量	int
10	电池电量信息	充电与否	boolean
11	电池电量信息	充电方式	string
12	屏幕状态信息	屏幕点亮与否	boolean
13	手机陀螺仪	手机所处的位置	float
14	光敏传感器	手机是否被遮挡	boolean

① 网络流量信息、Wi-Fi 网络信息以及移动蜂窝网信息: 这些信息反映了用户网络使用习惯, 网络连接的偏好, 智能手机安装应用产生的网络流量, 推测用户选择的通信运营商等等. 另外, Wi-Fi 网络的 SSID 信息和蜂窝网基站信息都间接隐藏着用户的地理位置信息, 这种对应关系可以经过推测能得到用户访问的历史位置信息, 可以通过构建每位用户连接过的 Wi-Fi 网络集合, 对用户社会关系进行推测.

② 屏幕状态信息: 通过每一次屏幕的点亮熄灭的时间间隔、一段时间的点亮次数, 知道用户各个时段手机使用时长等习惯, 推测用户使用手机强度和频率.

③ 电池电量信息: 用户使用手机的充电方式, 充电时间规律等行为习惯的体现.

④ 手机陀螺仪、光敏传感器: 是对用户使用手机时的环境、拿握手机的姿势 (睡姿或者坐姿), 手机携带方式 (随身携带或者放置在桌面上等) 行为的体现.

除此之外, 还会采集设备的国际移动设备标识 (International Mobile Equipment Identity, IMEI)[12]为用户身份标定, 以便进行实验结果的准确率的验证.

本次采集的数据未包含任何显式的用户身份信息,

与数据相对应的时间戳也同步记录. 为此我们开发了一个数据采集系统安装在志愿者的手机上去采集这些数据, 此系统包括 Android 手机客户端, Python 自动化脚本, 用户行为特征数据库.

2.2 移动端非显示身份信息数据预处理和特征选取阶段

数据预处理阶段主要是从采集到的数据中提取与用户行为有关的数据, 通过脚本导入数据库, 剔除无效数据, 再根据采集数据类型进行分类, 以便于后期进一步处理[13].

对于数据包大小、数据量等数值数据, 将根据其采集时间进行分箱, 数据分箱后可以进一步得到其最大值, 最小值、均值等统计学特征. 网络连接状态, 充电状态等布尔型的数据可以通过计算转化为网络连接时长, 充电时长等数据, 再计算其统计学特征. 最后, 去除部分相关性较高的数据, 例如: 电池电量消耗速度与屏幕点亮熄灭的频率等, 以降低后续数据处理的复杂度.

2.3 移动端非显示身份信息数据分析阶段

数据分析阶段结合预处理后的数据和数据分析模型实现用户行为的分析和身份的识别, 推测出其存在的社会关系. 为降低难度和提高识别率, 前期可通过数据可视化技术得到一些统计学特征和趋势图, 使用 Weka 分类算法中的 J48(决策树 C4.5) 将数据预处理后充电时长, 充电间隔, 屏幕点亮时长, 屏幕点亮间隔, 网络流量大小, 并以天为单位进行分箱, 得到输入的样本数据, 构建决策树从而对用户分类识别[14], 再根据这些信息选取适当的数据使用皮尔森相关系数等方式描述用户行为, 推测其社会关系.

3 数据分析

3.1 用户识别

目前小规模收集的有效数据有八万多条, 包括 5 名用户对象, 采集的数据的用户社会关系包括了情侣关系, 同年级不同寝室同实验室的同学关系, 不同年级不同寝室同实验室的同学关系以不同年级不同寝室不同实验室的同学关系.

以手机的 IMEI 号为唯一标识确定他们的对应关系如表 2 所示, 为保护采集数据的用户隐私, 本文用户名字使用字母代替.

表 2 用户关系图

数据采集对象	A	B	C	D	E
A					
B	同年级不同寝室同实验室				
C	不同年级不同寝室不同实验室	不同年级不同寝室不同实验室			
D	不同年级不同寝室不同实验	不同年级不同寝室不同实验室	不同年级不同寝室不同实验室	不同年级不同寝室不同实验室	
E	不同年级不同寝室同实验室	不同年级不同寝室同实验室	情侣		

我们对目前收集的数据进行分析, 发现用户的行为不论是网络行为, 还是手机的使用习惯都存在着明显的差异. 如表 3 所示, 用户每天屏幕点亮的次数就存在这明显的差异

表 3 用户每天屏幕点亮数据统计表

统计数据	A	B	C	D	E
有效数据 (天)	34	34	34	34	34
平均数 (次)	74.76	122.15	61.29	110.79	54.26
中位数 (次)	75.50	117.00	63.50	104.00	55.50
标准偏差	18.916	22.194	24.761	31.765	21.794
范围	85	157	96	134	74
最小值 (次)	35	87	20	52	21
最大值 (次)	120	214	116	186	95

分析表中的数据可以发现, 用户 BD 相对而言每天使用手机的频率要远高于其他三人, 与事实相吻合, 用户 BD 经常使用手机刷微博微信等社交软件. 用户 AC 使用手机的频率相对较少, 原因是 A 为研究生一年级的同学, 课时任务比较多, C 同学在上班, 使用手机的频率自然要少一些. 这说明屏幕点亮信息与个人使用手机的习惯是正相关的. 同时也对手机充电次数也做了类似的分析, 其充电次数统计信息如表 4 所示

表 4 每天充电次数

统计数据	A	B	C	D	E
有效数据	34	34	34	34	34
平均数	1.94	3.18	1.74	4.06	3.38
中位数	2.00	2.50	1.00	4.00	3.00
标准偏差	1.013	2.504	0.931	2.436	2.323
范围	3	14	3	12	9
最小值	1	1	1	1	1
最大值	4	15	4	13	10

从表中可以看出用户 BDE 每天充电的次数要高于其他两者, 与表 3 每天屏幕状态统计信息的数据基本相符, 一般而言, 手机使用频繁度与充电次数是正相关的. 其中不同的地方在于用户 E, 因为 E 从事 Android 开发, 常常会在真机上测试程序, 但平时使用手机并不频繁, 这也是用户 E 充电次数会偏高的原因.

这些数据表明不同的用户的手机行为或者习惯是不一样的, 正是因为每个人都有自己的行为特点, 所以我们可以利用屏幕状态信息和充电次数信息对用户进行识别. 本文使用 Weka 分类算法中的 J48(决策树 C4.5) 对用户进行识别, J48 是对 ID3 算法的扩展, 其主要区别在于可以容忍缺失数据, 这一点也是本文选择这个算法的主要原因. 由于手机上收集数据的特殊性, 数据会存在一部分的缺失, 通过 J48 这一特性可以很好的弥补数据上的缺陷. J48 的主要思想是以信息熵的增益为依据, 从原始样本中提取最有利于区分类别的属性, 逐渐的由根节点到叶子节点构建决策树, 其流程如图 1 所示.

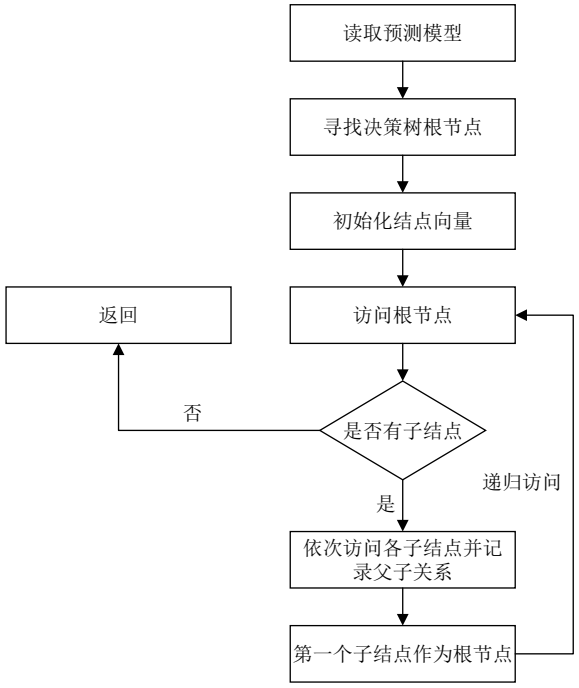

图 1 决策树算法流程图

我们通过对数据预处理后, 得到充电时长, 充电间隔, 屏幕点亮时长, 屏幕点亮间隔, 网络流量大小, 并以天为单位进行分箱, 这样得到输入的样本数据, 而构建决策树的过程中以上文计算得到的各种统计值作为分支条件, 算法性能数据详见表 5.

表 5 使用 Fast decision tree 进行用户识别的准确率

User	True Positive	False Positive	Precision	Recall	F-Measure	ROC Area
A	0.758	0.116	0.641	0.758	0.694	0.924
B	0.484	0.057	0.682	0.484	0.566	0.811
C	1	0.008	0.964	1	0.982	0.993
D	0.875	0.057	0.8	0.875	0.836	0.961
E	0.613	0.089	0.633	0.613	0.623	0.849
Weighted Avg.	0.74	0.068	0.737	0.74	0.734	0.906

从上表中 ROC Area 的值均在 0.9 左右, 其准确率已远高于随机猜测, 说明我们的分类算法能有效的对用户身份进行识别.

3.2 社会关系推测

社会关系推测采用的方法如下:

首先, 因 Wi-Fi 网络的 SSID 不一样, 而每个 SSID 代表一个地理位置, 若两人连接的 SSID 相同说明两人在同一区域出现过, 越经常在同一区域出现, 两人认识的机率越大, 故可通过用户访问过的 Wi-Fi 网络重合度、相似度进行社会关系紧密程度推测. 其 SSID 统计数据如表 6 所示, 横轴代表用户, 纵轴代表连接 Wi-Fi 的 SSID, 数字代表本文收集数据期间用户与该 SSID 的 Wi-Fi 连接次数. 由于每个用户连接过的 SSID 数据较大, 但常连接的一般只有三四个, 因此只保留了用户连接次数最多的前四个 SSID, 具体统计数据如下.

表 6 SSID 连接次数

SSID	A	B	C	D	E
FiveMeters	0	0	7212	0	7173
TP-LINK_9058B0	3943	0	0	0	0
LAIRMEY-1	0	0	1208	0	0
CUIT_YangBo	0	0	0	1447	0
FAST_584D92	0	0	0	953	0
bigWIN	148	2757	0	86	763
APC-20160107PHG	806	0	0	0	0
joyou2	0	812	0	0	0

其中连接 bigWIN(SSID 名称) 的人最多, 说明用户 ABDE 经常出现在同一个地点. 还可以明显看出用户 CE 社会关系紧密程度很高, 其共同连接 FiveMeters 的次数均在 7200 左右, 而其他人均为了 0, 用户 CE 除去 FiveMeters 外, 并没有相同连接的 SSID, 可以明确推测出 CE 用户的关系十分紧密, 实际上,

CE 为情侣关系, 一人上班, 一人在学校, 所以其相同的 FiveMeters 只有一个. 其中连接 bigWIN 的次数也可以明显分析出他们的社会关系. 用户 ABDE 在经常出现在同一地点, 而用户 C 除了与 E 关系亲近, 与其他人并不熟悉, 与实际相符合.

在判断数据的相关性, 将采用皮尔森相关系数去计算两两用户间访问过 SSID 集合的相似度. Pearson 相关系数也称为皮尔森积矩相关系数, 其计算公式如下:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{S_X}\right)\left(\frac{Y_i-\overline{Y}}{S_Y}\right) \quad (1)$$

Pearson 相关系数的取值在–1 到 1, 值越接近于正负 1, 相关度越大, 值为 0 代表两个数据完全不相关. 双尾显著性检测就是双侧检验, 举例说明若双尾显著性为 0.05, 则说明有 95% 的把握确认相关性的存在.

我们使用表 6 的样本计算 Pearson 相关系数, 其结果如表 7 所示:

表 7 SSID 的 Pearson 相关系数表

		A	B	C	D	E
A	皮尔森相关	1	-.189	-.212	-.277	-.196
	双尾显著性		.654	.614	.506	.641
B	皮尔森相关	-.189	1	-.218	-.226	-.084
	双尾显著性	.654		.604	.591	.844
C	皮尔森相关	-.212	-.218	1	-.262	.977
	双尾显著性	.614	.604		.531	.000
D	皮尔森相关	-.277	-.226	-.262	1	-.242
	双尾显著性	.506	.591	.531		.564
E	皮尔森相关	-.196	-.084	.977	-.242	1
	双尾显著性	.641	.844	.000	.564	

综上所述, 我们的小规模实验结果表明, 用户 CE 的关系十分紧密, 与实际情况相一致, 其中 CE 的皮尔森相关系数为 0.977, 双尾显著性为 0.000029.

4 结语

通过我们的研究发现即便是非显式用户身份信息, 通过大规模数据分析, 对用户身份进行识别, 并推测出部分用户间的社会关系等结论也会对用户隐私造成威胁, 为了保护我们的隐私要额外警惕不良应用收集这些非显式的身份信息, 有需要的话可以利用 Android 系统权限限制, 禁止应用使用某些权限来加以保护. 随着个人的隐私越来越受到人们的重视, 隐私保护逐渐成了当前迫在眉睫的研究课题.

参考文献

1 刘雅辉, 张铁赢, 靳小龙, 等. 大数据时代的个人隐私保护. 计算机研究与发展, 2015, 52(1): 229–247.

2 Bartsch M, Dienlin T. Control your Facebook: An analysis of online privacy literacy. Computers in Human Behavior, 2016, 56: 147–154. [doi: 10.1016/j.chb.2015.11.022]

3 沈薇薇, 熊金波, 黄阳群, 等. 面向手机短信的隐私保护方案. 计算机系统应用, 2016, 25(4): 118–122.

4 罗宇凡. 第 35 次《中国互联网络发展状况统计报告》发布. 青年记者, 2015, (6): 17.

5 王妮娜. 移动互联网时代个人隐私保护研究. 现代电信科技, 2015, 45(2): 45–49. [doi: 10.3969/j.issn.1002-5316.2015.02.010]

6 Manweiler J, Santhapuri N, Choudhury RR, *et al*. Predicting length of stay at WiFi hotspots. 2013 Proceedings IEEE INFOCOM. Turin, Italy. 2013. 3102–3110.

7 Tan KR, Yan GH, Yeo J, *et al*. Privacy analysis of user association logs in a large-scale wireless LAN. 2011 Proceedings IEEE INFOCOM. Shanghai, China. 2011. 31–35.

8 Falaki H, Lymberopoulos D, Mahajan R, *et al*. A first look at traffic on smartphones. Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement. Melbourne, Australia. 2010. 281–287.

9 陆悠, 罗军舟, 李伟, 等. 面向网络状态的自适应用户行为评估方法. 通信学报, 2013, 34(7): 71–80. [doi: 10.3969/j.issn.1000-436x.2013.07.008]

10 王浩. 移动通信网络社会行为关联优化研究[博士学位论文]. 武汉: 华中科技大学, 2010.

11 王丽文. 基于社交网络的数据挖掘研究[硕士学位论文]. 西安: 西安电子科技大学, 2014.

12 卢凤英, 张燕萍. 浅谈手机串号 IMEI 的应用. 信息通信, 2011, (2): 108–109. [doi: 10.3969/j.issn.1673-1131.2011.02.054]

13 林冠洲. 网络流量识别关键技术研究[博士学位论文]. 北京: 北京邮电大学, 2011.

14 黄炜. 基于数据挖掘的学习者身份识别[硕士学位论文]. 杭州: 杭州电子科技大学, 2011.