文章编号：2095-4980(2017)06-0928-05

面向大数据的电信宽带接入点行为特征

孙静博[1]，高　宸[2]，李　勇*[2]

(1.中国电信 北京研究院，北京 100191；2.清华大学 电子工程系，北京 100084)

摘　要：由于宽带网络接入的进一步普及，由宽带接入点构建的网络变得十分复杂。如何定义并分析宽带接入点的行为特征成为亟待解决的问题。本文研究了基于用户网络账户登录记录的宽带接入点特征，将宽带接入点下的账号记录作为数据集，定义并计算有效的特征，引入机器学习的方法，以得到宽带接入点的类型分类。通过对结果的校验，表明：本文提出的方法可以准确且高效地实现对于宽带接入点的家庭类型与非家庭类型的识别，得到其行为特征。

关键词：大数据；宽带接入点；行为特征

中图分类号：TN926⁺.1　　**文献标志码：A**　　**doi：**10.11805/TKYDA201706.0928

Behavioral features of telecom broadband access points with big data

SUN Jingbo[1]，GAO Chen[2]，LI Yong*[2]

(1.Beijing Research Institute，China Telecom，Beijing 100191，China；
2.Department of Electronic Engineering，Tsinghua University，Beijing 100084，China)

Abstract：Due to the further spread of the broadband, the network of broadband access points becomes quite complicated. It becomes a serious problem that how to define and analyze the behavioral features of large amount of access points. The features of access points are studied based on the log records of users' Internet account. Several useful features are defined and calculated, and then machine learning is introduced to classify the broadband access points. The verification of experimental results shows that it can be identified whether a broadband access point falls into residential category or non-residential category and its behavioral features can be described accurately and efficiently.

Keywords：big data；broadband access point；behavioral feature

21 世纪是互联网高速发展的时代，近年来，随着互联网的进一步普及以及人们日益增长的需求，各种网络服务层出不穷，掀起了互联网革命。随着宽带的普及率进一步提高，宽带接入点构建出一个大型的复杂网络，如何认识这样的一个网络，成为一大难点。特别的，在城市计算中，城市功能和用户画像是 2 个核心的概念，城市功能强调建立基于地理信息的模型，刻画城市不同区域、不同时间的状态，具体应用有功能区划分、城市人口分布等；而用户画像旨在对于给用户打上标签，便于理解与更深层次地研究，强调对于用户属性的判断[1]。

城市计算作为一个系统性的工程，一个典型的操作角度即是通过网络服务进行，一方面，网络账号普及，这与互联网的进一步普及，以及人们对于互联网的需求不断加大有关。其次，对于用户的移动性分析需求，当获得了海量的数据，有必要建立一个抽象的模型去认识这样一个复杂系统。再者，这样一个网络，有利于对于城市功能的网络构建，要想分析城市功能，有必要挖掘用户的特征和接入点的特征。对于宽带接入点的分析，有助于进行广告投放、商业选址、房地产估价等。从另外一个角度而言，随着互联网的飞速发展，各种各样的网络服务，包括社交网络、即时通信、电子购物等，在人们的日常生活中扮演着至关重要的角色。对于提供网络服务的公司而言，用户成为了一个个虚拟的 ID，作为在赛博空间的唯一标识。用户在使用网络服务的同时，也通过多种方式产生位置信息，包括全球定位系统 (Global Positioning System，GPS)，全球移动通信系统(Global System for Mobile，GSM)、射频识别(Radio Frequency Identification，RFID)技术等[2]。

位置信息是用户隐私信息的重要组成部分，事实上，对于基于用户“签到”信息，即带有时间戳的位置数据的隐私泄露问题，已有大量的研究与学术成果，并投入到工业界的应用中[3-6]。主要数据来源是网络服务提供商，

收稿日期：2016-09-08；**修回日期：**2016-09-23
***通信作者：**李　勇　email:liyong07@tsinghua.edu.cn

即给用户提供对应网络服务的公司，如 Twitter,Foursquare 等[7-9]。通过对于一段时间内的用户签到信息的观测，可以对用户轨迹、用户移动性进行研究，产生用户画像。

伴随网络服务的爆炸式增长，网络账号的数量成为天文数字。从 DPI(Deep Packet Inspection)获取的不同账号在不同时间、不同接入点下的登录记录，可以作为研究接入点行为特征的基础数据。

为了将用户的网络活动与物理世界对应，有必要将用户登录网络服务的位置属性进行划分，在本文的工作中，数据集中含有宽带租户接入点的唯一标识，而对于部分的宽带接入点，则含有登记信息，针对宽带接入点行为特征中的问题，通过对账号登录记录的特征提取及分析，通过数据挖掘、机器学习的方法，实现对于宽带接入点类型的基础分类，准确且高效地区分家庭类宽带接入点和非家庭类宽带接入点。

1　数据集

数据集为通过某宽带运营商 DPI 获得的某市某月的账号登录数据，见表 1。同时，训练集来自部分宽带接入点的类型登记，即已知类型的宽带接入点，见图 1。

在共计约 1 300 万的宽带接入点中，约有 300 万含有登记信息，其中家庭类与非家庭类的比例大致为 14:1。由于存在大量没有登记信息的接入点，有必要根据已有信息，构造分类器，从而得到所有宽带接入点的类型。

表 1 账号登录数据
Table1 Log-in dataset

parameter		value
accounts	shopping site	about 10 million
	instant message	about 10 million
	social network	about 1 million
	phone number	about 1 million
device		win7, IOS, android, etc.
access points		about ten million
location		latitude/longitude
timestamp granularity		hour

通过对宽带接入点的登录特性的初步观察，有大量的接入点表现很不活跃，因为登记数据时间较久，故可能有一些宽带接入点，已经无人使用，为避免部分登录记录稀少的宽带接入点影响准确性，筛选出 4 种账号均有登录记录即较为活跃的宽带接入点，其分布见图 2。

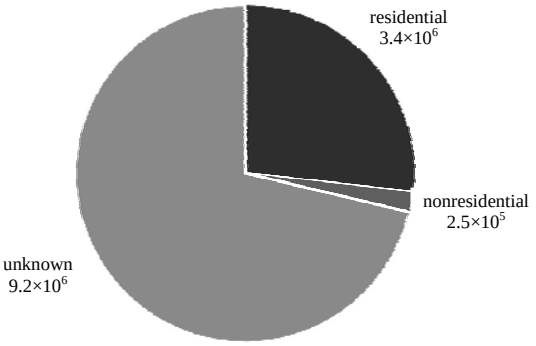

Fig1 Type distribution of all broadband access points
图 1 所有宽带接入点的类型分布

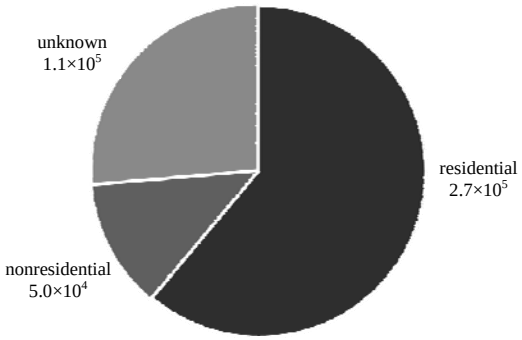

Fig.2 Type distribution of active broadband access points
图 2 活跃的宽带接入点的类型分布

2　特征分析

区分家庭类型和非家庭类型的出发点是寻找可有效区分的特征。本文主要使用 4 种特征，即：登录频次、账号熵、日夜比、工作日与周末的比值。

2.1 账号登录频次

账号登录频次的定义：对于 1 个宽带接入点，其每有 1 条登录记录，频次加 1。

图 3 为手机账号登录频次的累积概率分布。从该分布图可以看出，有 74.1%的家庭类宽带接入点的登录记录数目小于 10，而 66.4%的非家庭类宽带接入点的登录记录数目大于 10。可以看出，家庭类型与非家庭类型区别明显，非家庭类型的账号登录频次整体偏大。结果符合理论预期：非家庭多为公共热点、公司企业等，人数一般较多，与家庭相比，有较大的登录频次，此特征符合实际的情形。

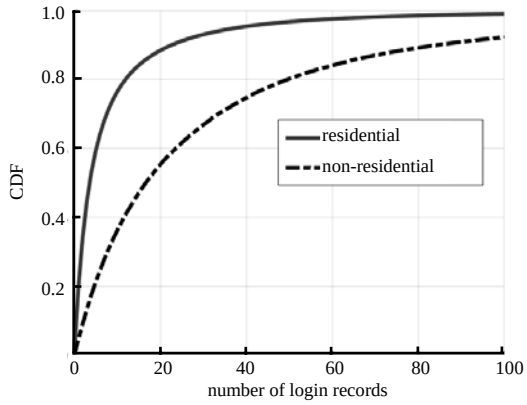

Fig.3 Log-in times cumulative distribution function
图 3 手机账号登录频次的累积概率分布

2.2 账号熵

对于某宽带接入点，统计出现的账号及其次数，对其求熵。如某一个宽带接入点出现过 3 个账号 A,B,C：A 出现的频率为 2/10，B 出现的频率为 3/10，C 出现的频率为 5/10。取熵为：$-\left(\frac{2}{10}\log_2\frac{2}{10}+\frac{3}{10}\log_2\frac{3}{10}+\frac{5}{10}\log_2\frac{5}{10}\right)$，数学表述为：$E(L)=-\sum_{u\in U_L}P_L(u)\log_2 P_L(u)$。

通过人工观察，大学校园、商场、饮食街等处具有较高的熵，而住宅区的熵则比较低。

图 4 为手机账号的熵的累积概率分布图。从图中可以看出，家庭类型与非家庭类型区别明显，非家庭类型的账号熵整体偏大。

账号的熵反映了一种分散程度，公共热点、企业公司等非家庭宽带接入点，与家庭宽带接入点相比，人流量较大，因而账号登录的分散程度也较大。

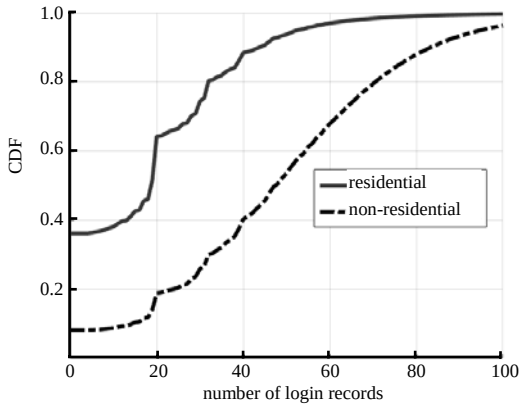

Fig.4 Entropy cumulative distribution function
图 4 手机账号熵的累积概率分布

2.3 白天与夜间登录账号数的比值

对于某接入点,统计 24 h 每一个时刻登录的账号个数。以 3:00 为例，将 30 d 中每天的 3:00 出现的账号求和。即对于一个宽带接入点而言，会得到 24 个值。

图 5 是手机账号在 1:00~24:00 登录账号数比例的分布。从图中可以看出，在 8:00~18:00 的时间段内，非家庭类有更多的登录，因为这段时间公司企业、公共热点有更多的活跃用户，符合常识；而在 18:00~24:00 的时间内，家庭类宽带接入点有更多的登录记录，这段时间通常是从人们下班回家到休息的时间。综上，家庭类和非家庭类接入点在上述两段时间内登录的用户数比例差异很大，故可以作为分类器的一项特征。在此基础上，计算 10:00~16:00 的 7 个时刻的和与 19:00~24:00 的 6 个时刻的和的比值,作为特征。通过对于特征的观察得到，这个比值对于家庭类型与非家庭类型的区分度较高。

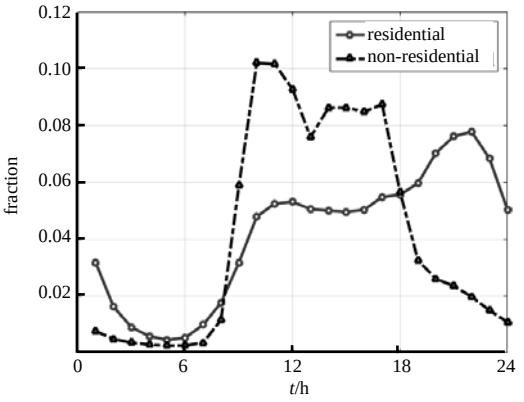

Fig.5 Fraction distribution of account number during one day
图 5 手机账号在 1:00~24:00 登录账号数比例分布

2.4 工作日与周末登录账号数的比值

对于某宽带接入点，统计 30 d 每一天登录的账号的个数。即对于一个宽带接入点而言，会得到 30 个值。

图 6 是手机账号在 1~30 日登录账号数目比例的分布。从图中可以看出，除去部分天(这部分的数据采集有误)，非家庭呈现出工作日高、周末低的规律，而家庭则较为平缓。

由于考虑部分天的数据异常，用[2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 18, 19, 20, 24, 25, 26, 27, 30]之和(工作日)与[1, 7, 8, 14, 15, 21, 22, 28, 29]之和(周末)的比值为特征。通过对于特征的观察得到，这个比值对于家庭类型与非家庭类型的区分度较高。

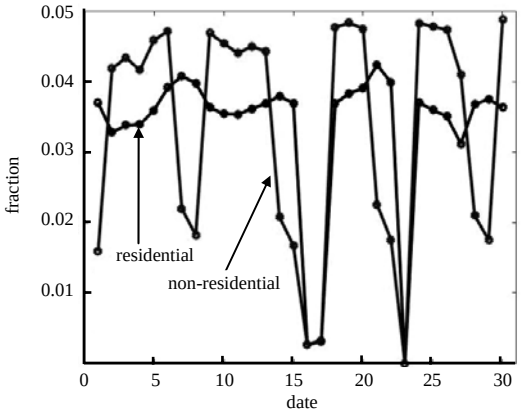

Fig.6 Fraction distribution of account number during one month
图 6 手机账号在 1~30 日登录账号数比例分布

3 分类方法

3.1 选取数据

选取手机、即时通信、社交网络、网上购物 4 种账号均有登录记录的宽带接入点，作为研究对象。

3.2 训练

对于 4 种账号，分别计算上文中介绍到的 4 种特征，即对于每一个宽带接入点而言，均有 16 种特征，包括登录频次、账号熵、日夜比、工作日与周末的比值。同时，针对已知类型的宽带接入点，将样本等分成 10 份(这里涉及到对于不良样本的筛选问题)。通过对每份样本的具体分析，观察其特征性质，均满足整体的性质，即认为样本初步可靠。

采用 3 种方法进行分类：逻辑回归、支持向量机及随机森林。逻辑回归，是一种离散选择法模型，假设类别可以用一个简单的线性函数拟合，误差较大，但其对于每一个特征，都有一个体现特征影响大小的参数，该参数可以用来衡量特征的权重。支持向量机，则用一个或多个高维乃至无穷维的超平面来进行分类，因为一个好的分类结果倾向于增大分界面与资料点的距离，支持向量机旨在寻找一个与最近资料点距离最大的分界平面。与逻辑回归相比，支持向量机的数学模型更加完善，但也提升了计算复杂度。随机森林，是一种采取多个决策树的分类器。在训练样本中，每次采取有放回的抽样，并随机选取部分特征，构造一个决策树，最终结果由构造的多个决策树的判决结果的众数决定。随机森林的优点包括平衡误差、适用于多种训练集、适用于估计遗失资料等。本文问题的需求符合随机森林理论模型，理论上有优于逻辑回归、支持向量机 2 种方法的分类效果。

3.3 输出

采取 10 次交叉验证的方法，即对于 10 份等分样本，每次选取 9 份作为训练集，进行分类器的学习，再用剩下的 1 份作为验证，评估分类器的性能。使用 F-Score 评价分类效果。对于 10 次结果，可以筛选出不良样本，在去掉不良样本的同时，重新训练，重复以上工作。F-Score 又称 f-measure，为准确率与召回率的调和平均数，在检验分类性能时避免了准确率和召回率某一项过低的问题。在实际操作的过程中，没有遇到不良样本，所以无需去除，直接对 10 次 F-Score 进行取平均操作，得到最终结果。

4　结果分析

使用逻辑回归、支持向量机和随机森林 3 种方法的结果见表 2 和表 3。

表 2 家庭类接入点不同分类方法的效果评估
Table2 Performance of different algorithms(residential)

algorithm	precision	recall	F1-Score
logistic regression	0.92	0.97	0.94
support vector machine	0.97	0.91	0.94
random forest	0.95	0.95	0.95

表 3 非家庭类接入点不同分类方法的效果评估
Table3 Performance of different algorithms(non-residential)

algorithm	precision	recall	F1-Score
logistic regression	0.79	0.58	0.67
support vector machine	0.66	0.88	0.75
random forest	0.78	0.78	0.78

从表 2 和表 3 可以看出，对于家庭类型的宽带接入点类型而言，由于其数量较多，所以有着较高的准确率和召回率，非家庭类型宽带接入点的数量较少，所以准确率和召回率较低。在 3 种分类方法中，逻辑回归的性能最差，随机森林表现出最佳性能，家庭类型的 F1-Score 达到 0.92，非家庭类型的 F1-Score 达到了 0.78~0.79。

从运行效率而言，随机森林运行时间最长，支持向量机次之，逻辑回归运行最快。

以上结果符合预期，在前文分析中，认为随机森林适合于估计遗失的数据集，与本文研究的问题较为契合。

可以用逻辑回归方法中训练得到的向量，判断每一个特征对于宽带接入点的影响。由于每个特征的数值取值范围不同，甚至差上几个数量级，所以将训练得到的向量每一维的数值，乘以对应的特征数值的标准差，以此衡量每个特征对于接入点类型的影响大小，如表 4 所示。其中 IM 代表即时通信账号，shopping 代表在线购物账号，OSN 代表社交网络账号，phone 代表手机号码。

表 4 每个特征对于接入点类型的影响大小
Table4 Influences of every feature on classification

feature	IM	shopping	OSN	phone
accounts number	-0.000 5	0.033 3	-0.013 1	-0.038 2
log-in records	-0.365 5	-0.114 5	-0.031 5	-0.044 7
entropy	-1.131 9	0.034 4	-0.161 7	-0.164 5
day night ratio	0.461 4	0.407 4	0.580 7	0.624 5

表 4 中正值表明，如果一个宽带接入点拥有更大的对应特征值，则该接入点更倾向于被判断为家庭类；而负值表明，如果一个宽带接入点拥有更大的对应特征值，则该接入点更倾向于被判断为非家庭类。表 4 中的值的绝对值越大，表明该特征对于分类的影响越大。地点的熵在区分家庭类型与非家庭类型的过程中作用最大，因为其同时包含了登录数目与次数两方面的信息。对于不同的账号类型而言，即时通信类账号的特征在区分接入点类型的过程中，起到最大的作用，而网上购物类账号的作用相对较小。同时，拥有更多网上购物类账号登录的宽带接入点，更有可能是一个家庭类型宽带接入点。这个现象表明，与公共场合相比，人们更倾向于在家使用网上购物的账号。总的来说，通过精心挑选

4 种特征，比较多种主流的分类器构造方法，成功地将几百万个位置分成 2 类，即家庭类型和非家庭类型，并达到了较好的性能。

进一步分析了那些误判的宽带接入点。通过观察具体的登录记录发现，这些误判的宽带接入点表现出一些反常的性质，即误判为家庭类型的非家庭类型接入点，实际上表现出家庭的性质，而误判为非家庭类型的家庭类型接入点，实际上表现出非家庭的性质。这一点在预期之内，因为实际上，对于一些登记为家庭类型的接入点，最终实际的用途并不是用作家庭，而可能是小餐馆、创业公司等，而登记为非家庭类型的宽带接入点，也有可能用做家庭用。

同时，本文的研究方法中，认为非家庭类型宽带接入点具有白天活跃、晚间不活跃的性质，而实际上，很多非家庭类型宽带接入点，如网吧、酒吧等，晚间仍保持着较高的活跃度，甚至超过白天的活跃度。目前的方法没有考虑到这种情况，这是本文的局限之一。此外，研究中使用的是 4 种网络账号均有登录记录的接入点的数据集，虽然这个方法可有效消除一些极不活跃的接入点的影响，但终究缩小了有效的数据集，如果降低对此处“活跃”定义的界限，势必会有更深入的结果。

5 结论

针对目前宽带接入点行为特征分析的问题，本文提出一种基于数据挖掘和机器学习的方法，通过对网络账号登录记录的特征分析，有效实现了对于宽带接入点家庭类型和非家庭类型的区分，取得了较为理想的成果。

在目前的工作中，对于非家庭类型的建模较为简单，与现实中有一定的区别，今后的工作需要继续完善模型及算法，以实现更加准确的类型判断。此外，对于家庭类型和非家庭类型各自本身，可以有更进一步的细分，如家庭类型可以从家庭的结构着手，分为年轻家庭、中年家庭和老年家庭等，非家庭类型可以细分为公司、公共热点等，在进一步的工作中，将着手引入更多的聚类、分类和聚类相结合的方法，进行更深入的研究和分析。

参考文献：

[1] ZHENG Yu,OURI L,WOLFSON O,et al. Urban computing: concepts, methodologies and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2014,5(3):38.

[2] DOMINGO-FERRER J,TRUJILLO-RASUA R. Microaggregation-and permutation-based anonymization of movement data[J]. Information Sciences, 2012,208(21):55–80.

[3] BADEN R,BENDER A,SPRING N,et al. Persona:an online social network with user-defined privacy[C]// Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication. Barcelona,Spain:ACM, 2009:135–146.

[4] FANG Lujun,LEFEVRE K. Privacy wizards for social networking sites[C]// Proceedings of the 19th International Conference on World Wide Web. Raleigh,North Carolina,USA:ACM, 2012:351–360.

[5] FENG Huan,FAWAZ K,SHIN K G. LinkDroid: reducing unregulated aggregation of app usage behaviors[C]// Proceedings of the 24th USENIX Security Symposium. Washington,D C:[s.n.], 2015:769–783.

[6] KRISHNAMURTHY B,WILLS C E. Generating a privacy footprint on the internet[C]// Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement. Rio de Janeriro,Brazil:ACM, 2006:65–70.

[7] NOULAS A,SCELLATO S,MASCOLO C,et al. An empirical study of geographic user activity patterns in foursquare[C]// Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. [S.l.]:Elsevier, 2011:570–573.

[8] PONTES T,VASCONCELOS M,ALMEIDA J,et al. We know where you live:privacy characterization of foursquare behavior[C]// Proceedings of the 2012 ACM Conference on Ubiquitous Computing. Pittsburgh,Pennsylvania:ACM, 2012:898–905.

[9] NOULAS A,SCELLATO S,LATHIA N,et al. Mining user mobility features for next place prediction in location-based services[C]// 2012 IEEE 12th international conference on Data Mining(ICDM). Brussels,Belgium:IEEE, 2012:1038–1043.

作者简介：

孙静博(1983–)，男，哈尔滨市人，博士，目前从事大数据领域的产品研发与运营.

高　宸(1996–)，男，安徽省安庆市人，在读博士研究生，主要研究方向为移动数据分析.

李　勇(1985–)，男，长沙市人，助理教授，博士生导师，主要研究方向为移动计算与社交网络、城市计算与车载网络、网络科学与未来网络.email:liyong07@tsinghua.edu.cn.