doi:10.3969/j.issn.1000-436x.2014.z1.028

基于用户过滤的校园无线网用户聚类方法

仇一泓，尧婷娟，秦丰林，葛连升
（山东大学 信息化工作办公室，山东 济南 250100）

摘 要：随着智能终端地普及，在校园无线网用户聚类研究中采用 MAC 地址作为用户区分已不能真实反映用户的行为，为此，提出了一个基于用户过滤的校园无线网用户聚类方法，该方法基于用户活跃度对用户行为数据进行过滤，在此基础上对校园无线网用户行为做进一步地聚类分析。实验结果表明了该方法的有效性。
关键词：校园无线网；用户聚类；用户过滤；用户活跃度
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2014)Z1-0146-04

User filtering based campus WLAN user clustering method

QIU Yi-hong, YAO Ting-juan, QIN Feng-lin, GE Lian-sheng
(Information Office, Shandong University, Jinan 250100, China)

Abstract: With the widespread of smart terminals such as smart phones and smart pads, using MAC address as user identification in campus wireless local area network (WLAN) user clustering research cannot exactly represent user behavior. An user filtering based user clustering is proposed. This method filters users' behavior data by their degree of activeness, and then further conducts clustering analysis of campus WLAN user behavior. The experimental result verifies the effectiveness of the proposed method.
Key words: campus WLAN; user clustering; user filtering; user degree of activeness

1 引言

随着无线局域网（WLAN）技术的成熟和智能用户终端的普及，国内高校已逐步兴起校园 WLAN 的建设，为学生和教师提供随时随地的无线上网服务[1]。以山东大学为例，经过近几年的持续投资建设，已经部署了 3 个无线控制器（AC）和 2000 多个无线接入点（AP），WLAN 涵盖 6 个校区，覆盖区域包括公共教学楼、图书馆、实验室、学生宿舍和广场等，基本实现了校园的无缝覆盖，为师生提供方便高速的无线上网服务。自 WLAN 建设之初，学术界就开始对 WLAN 的用户行为进行测量和分析，而随着 WLAN 规模的扩展，其研究也一直在发展和深入。早期，由于受 WLAN 部署范围的限制，研究者侧重于 WLAN 用户的一般性统计分析，如用户在线时长、接收与发送字节数、认证失败次数和应用层协议等，典型的研究有 Balachandran 等对 ACM 会议 WLAN 中的用户行为研究等[2]。随着 WLAN 部署规模的扩大，研究者开始进行对用户的移动性和社会关系的研究，例如，来自加州大学的 Hsu 和 Helmy 对 WLAN 中的用户社会关系进行了研究，提出了基于用户相遇时间、相遇次数和位置这 3 个指标来衡量用户之间的社会关系，并分别比较了这 3 种指标所计算的用户相似度[3]。上海交大的吴利明博士在文献[3]的基础上做了相应的改进，其主要考虑到不同 AP 的访问频率不同，引入了位置加权参数来修正基于相遇时间的社会相似度模型[4]。微软亚洲研究院的郑宇对基于 LBS 的信息挖掘做了持续的研究，根据用户移动轨迹，挖掘出用户生活规律及热点区域，从而为用户推荐个性化、智能化的服务[5]。

收稿日期：2014-10-14
基金项目：国家自然科学基金资助项目（61170211）；山东大学自主创新基金资助项目（2012TS195, 2012TS196）
Foundation Items: The National Natural Science Foundation of China (61170211) ; Independent Innovation Foundation of Shandong University (2012TS195, 2012TS196)

在上述研究中，研究者多以 MAC 地址来区分不同用户，由于早期单个用户拥有无线终端的数量有限，这样统计出来的用户数据也较符合真实的用户行为。然而，近年来，随着智能手机、平板电脑等智能用户终端的普及，一个校园网用户可能同时拥有多个智能终端，如果依然采用 MAC 地址来区分不同的用户，则通过用户历史轨迹统计出来的用户数量可能远超过于真实的用户数量。而且，在用户数量过大时，用户之间相似性计算的复杂性也随之增大。为此，提出一个基于用户活跃度的用户过滤方法，在此基础上，对校园 WLAN 用户进行进一步的聚类分析。

2　基于用户过滤的用户聚类方法

基于用户过滤的用户聚类方法，如图 1 所示。可以看出，相对传统的用户聚类方法，基于用户过滤的用户聚类方法在数据采集、切片、用户相似度计算和用户聚类等模块以外，又增加了用户过滤模块。下面将对各模块的作用和实现做具体说明。

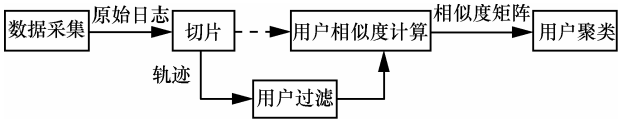


图 1　基于用户过滤的用户聚类方法

2.1　数据采集

校园 WLAN 用户数据采集主要通过 Snmp 和 Syslog 2 种方式[2,6]。其中，Snmp 方式可以定时采样 WLAN 用户数据，主要包括当前在线 AP 及在线用户的基本情况，例如 AP 的当前关联人数、用户在线的信号强度、在线用户的 MAC 地址及用户连接的 AP 等，这些记录可以有效分析网络的变化情况及当前网络的运行情况，包括用户人数随时间变化情况、用户流量随时间变化情况等，但是 Snmp 方式一般采用 5 min 的轮询周期，数据实时性较差。因此，主要采用 Syslog 方式，通过配置一个 rsyslog 服务器，将所有与用户相关的事件都从 AC 实时发送到远程的 rsyslog 服务器中。

2.2　切片

原始日志中的用户数据是由许多事件记录而成，包括关联、解关联、再次关联等事件，为了更直观地表示用户行为信息，对原始日志中的用户行为信息进行切片处理。用户切片（profile）表示从原始数据中提取出来的用户行为数据，用户切片定义成四元组列表〈user_mac, timestamp, ap_connect, status〉的形式，其中，user_mac 表示用户 MAC 地址，ap_connect 表示当前连接的 AP，timestamp 表示时间，status 表示状态。当 status 为 on 时，timestamp 表示开始连接时间，当 status 为 off 时，timestamp 则表示结束连接时间，其中时间使用 UNIX 时间戳表示。

一个典型的用户切片数据示例，如表 1 所示。

表 1　　用户切片数据示例

user_mac	timestamp	ap_connect	status
000 000 000 787	1 395 362 093	AC-2-WX6-205	on
000 000 000 787	1 395 366 728	AC-2-WX6-205	off
000 00d aff 8ba	1 395 132 961	AC-1-359	on
000 00d aff 8ba	1 395 132 964	AC-1-359	off

2.3　用户过滤

在 WLAN 发展的早期，单个用户拥有的终端数量有限，因此早期的研究者多以 MAC 地址区分不同用户，然而，近年来，随着智能手机、平板电脑等智能终端的普及，一个用户可以同时拥有多个无线智能终端，这样导致用户可以通过多种设备接入无线网络，如果依然采用 MAC 地址来区分不同的用户，则通过用户历史轨迹统计出来的用户数量可能远超过于真实的用户数量，而且，当用户数量过大时，则用户之间相似度计算复杂性也随之增大。另外，由于校园无线网在用户关联 AP 之前不需要任何认证，因此任何终端只要在 AP 的信号覆盖范围内都可以与 AP 进行关联，在关联 AP 之后，只有通过 Radius 认证后的终端才真正属于校园用户。

因此，为了去除非校园用户和统计出更真实的用户，减少用户相似度计算复杂性，提出基于用户活跃度的用户过滤策略，用户过滤步骤如下。

1) 基于用户切片文件 user_profile，统计出用户数量，并筛选出认证成功记录中的 MAC 地址，形成 MAC 地址列表文件 login_mac。

2) 对 user_profile 文件中的 user_mac 地址与 login_mac 中的 user_mac 地址进行匹配，形成新的 user_profile 文件。

3）统计出每个 user_mac 地址的访问频率及总在线时间，根据这 2 个指标来评价用户的活跃度，并滤除活跃度低的用户。

2.4　用户相似度计算

文献[3]提出了基于相遇 AP、基于相遇次数和

基于相遇时间的用户相似度计算方法。设 $ET(A,B)$ 为用户 A 与用户 B 在同一 AP 上的共同在线时间，而 $TO(A)$、$TO(B)$ 分别表示用户 A、B 在该 AP 上的总在线时间，则用户相似度 $Sim(A,B)$ 计算为

$$Sim(A,B)=\frac{1}{2}\left(\frac{ET(A,B)}{TO(A)}+\frac{ET(A,B)}{TO(B)}\right) \quad (1)$$

2.5 用户聚类

当前众多研究文献中存在各种不同的聚类算法，包括层次化聚类、划分式聚类和基于密度的聚类算法等[7]。为简单起见，采用了层次化聚类算法，在计算得到用户相似性矩阵后，使用自下向上的层次聚类对用户进行聚类分析。首先把每个用户视为一个类，通过已经计算完成的用户相似度矩阵，找到相似度最大的 2 个类，将这 2 个类进行聚合，并更新该类与其他类的相似度数据，通过不断循环迭代，从而实现聚类。需要注意的是，对用户数据集的聚类，需要选择一个合适停止聚类的阈值，这种阈值一般取决于该领域内专家经验或反复实验后的最优结果。

3 实验及结果分析

3.1 数据集

基于山东大学 WLAN 从 2 个 AC 上采集了一个星期的用户日志数据，为便于比较，从网上下载了美国 Dartmouth 学院的公开 WLAN 数据集[8]，表 2 对 2 个数据集的具体参数作了对比。

表 2　　2 个数据集比较

数据源	山东大学（SDU）	Dartmouth 学院
采集时间	2014.3.18~2014.3.25	2005.9.25~2005.10.1
数据类型	Syslog	Syslog
AP 类型	H3C	Cisco、Aruba
AP 数量	1 081	146
用户数量	41 219	2 079

3.2 用户过滤

从表 2 可以看出，山东大学的用户数量远大于 Dartmouth 学院，因此，本实验仅对山东大学数据集进行用户过滤。根据 2.3 节中描述的用户过滤步骤，经过步骤 1)，统计出不同的 MAC 地址为 41 219 个，而经过步骤 2)，匹配后的 MAC 地址数量为 20 449 个，再经过步骤 3)，统计出来的用户访问分布图和用户总在线时间分布分别如图 2 和图 3 所示。可以看出，20%的用户访问频率低于 10 次，30%的用户平均每天在线时间低于 1 h。常理来说，用户访问频率过低的用户，与其他用户相遇的概率也较低。因此，统计了总时间低于 210 min 或总访问频率低于 21 次的用户数量为 8 536 个人，占总数量的 41%左右，为了减少用户相似度计算量，本实验将这部分用户进行了过滤。

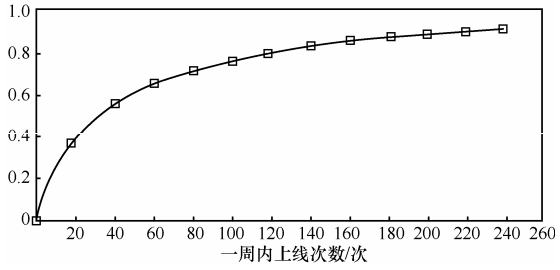


图 2　用户访问频率 CDF 分布

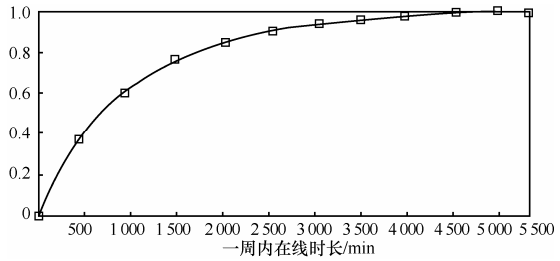


图 3　用户总在线时间 CDF 分布统计

3.3 用户相似度

图 4 显示了 2 个数据集中用户相似度的分布情况，其中横轴表示相似度，纵轴表示大于某一个相似度的用户对数占总用户对的百分比。从结果可以看出，单个用户只能和整个网络中的少数用户相遇，大部分用户之间的相似度都为 0，其中 Dartmouth 学院和山东大学分别仅有 0.7%和 5.8%的用户对拥有大于 0 的相似度，山东大学用户相似度大于 0 的比例比 Dartmouth 学院要高的原因是，相比 Dartmouth 学院，山东大学的用户数量更多，而且 AP 的部署更为密集；另外一方面，2 个数据集都有部分用户对的相似度接近于 1，即存在用户相似度非常高的用户群体。

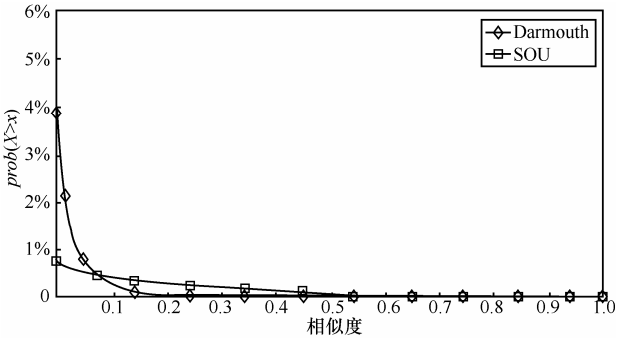


图 4　校园 WLAN 用户相似度 CDF 分布

3.4　用户聚类结果

本实验将层次聚类方法中相似度的阈值设为0.4，类内使用最大相似性度量，类与类之间的连接规则使用平均连接规则，在完成聚类后，再剔除部分单独用户的分类。

图 5 和图 6 分别显示了 2 个学校用户聚类后的类簇大小分布图。从结果可以看出，Dartmouth 学院具有 395 个社会性分组，而山东大学则具有 645 个社会性分组。其中，Dartmouth 学院的最大类中含有 13 个用户，而山东大学的最大类中则含有 31 个用户。从类簇大小分布来看，2 个学校中都存在较大的类簇，这说明在 2 个校园里都有较大的用户组，即存在较为紧密的团体或者组织，这也符合高校的学生组织团体现状，同时，2 个学校也都存在大量的较小的类簇，一方面可能是由于采集的周期过短，从而导致大部分用户的相似性计算值较小；另一方面，也可能是由于用户聚类阈值较大，使相似性较小的用户无法聚类成一类。

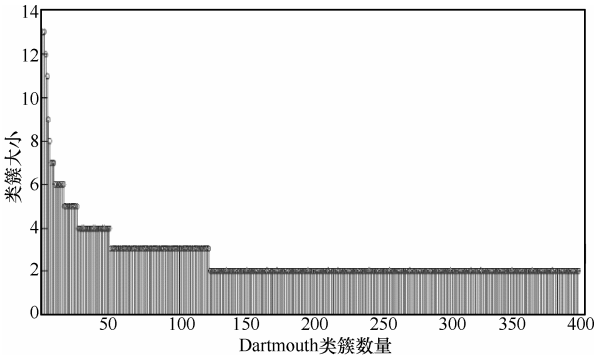


图 5　Dartmouth 学院用户类簇大小分布

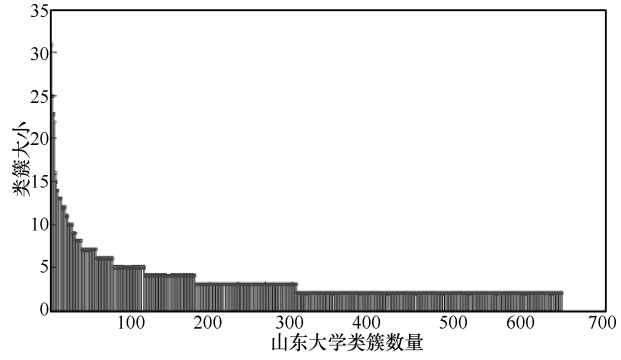


图 6　山东大学用户类簇大小分布

4　结束语

校园无线网用户行为研究对用户挖掘、个性化服务和推荐具有重要意义，用户聚类是其中的一个重要研究方向。提出了一个基于用户过滤的校园无线网用户聚类方法，说明了该方法与传统用户聚类方法的改进方案，详细阐述了方法中各个模块的作用和实现，并重点介绍了用户过滤的设计动机及其实现步骤。对山东大学和 Dartmouth 学院 2 个数据集的实验结果表明了本文方法的有效性。

参考文献：

[1] 吴颖骏. 基于“云”的智慧校园[J]. 中国教育网络,2010,(11):25-26.
WU Y J. Cloud-based wisdom campus[J]. China Education Network, 2010, (11): 25-26.

[2] BALACHANDRAN A, VOELKER G M, BAHL P, *et al*. Characterizing user behavior and network performance in a public wireless LAN[A]. ACM SIGMETRICS Performance Evaluation Review[C]. ACM, 2002.195-205.

[3] HSU W, HELMY A. Impact: Investigation of Mobile-user Patterns Across University Campuses Using Wlan Trace Analysis[R]. Technical Report, University of Southern California, 2005.

[4] 吴利明. 无线网络环境下用户行为的社会性分析[D]. 上海:上海交通大学,2012.
WU L M. Sociality Analysis of Users’ Behavior in Wireless Networks[D]. Shanghai: Shanghai Jiao Tong University, 2012.

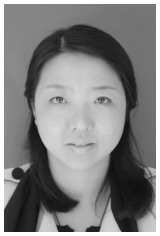
[5] 郑宇，谢幸. 基于用户轨迹挖掘的智能位置服务[J]. 中国计算机学会通信，2010, 6(6):23-30.
ZHEN Y, XIE X. User trajectory mining-based intelligent location service[J]. Communication of China Computer Federation, 2010, 6 (6): 23-30.

[6] SCHWAB D, BUNT R. Characterising the use of a campus wireless network[A]. IEEE INFOCOM[C]. 2004. 862-870.

[7] JAIN A K, MURTY M N, FLYNN P J. Data clustering: a review[J]. ACM Computing Surveys (CSUR), 1999, 31(3): 264-323.

[8] The dartmouth/campus dataset[EB/OL]. http://crawdad.cs.dartmouth.edu/dartmouth/campus/, 2007.

作者简介：



仇一泓（1977-），女，山东莱西人，山东大学工程师，主要研究方向为无线网络、网络管理、网络测量、IT 服务管理等。

尧婷娟（1988-），女，江西南昌人，山东大学硕士生，主要研究方向为无线网络、网络管理、网络测量等。

秦丰林（1978-），男，山东潍坊人，山东大学高级工程师，主要研究方向为无线网络、网络管理、网络测量等。

葛连升（1967-），男，山东莒县人，山东大学教授、硕士生导师，主要研究方向为网络测量、网络管理、智能信息处理、软件工程、IT 治理等。