doi:10.3772/j.issn.1002-0470.2019.09.004

无线漫游场景下的轨迹相似度计算方法①

常祎祎②* ** *** 王劲松③* ** ***

(*天津理工大学计算机科学与工程学院 天津 300384)
(**天津市智能计算及软件新技术重点实验室 天津 300384)
(***计算机病毒防治技术国家工程实验室 天津 300457)

摘 要 用户在无线网络间漫游时产生了大量的行为数据。这些数据蕴含着用户的生活轨迹,轨迹越相似的用户具备亲密社会关系的可能性越大。传统方法通过比较两条语义轨迹中的最长公共子序列来挖掘用户之间轨迹的相似程度。但这种算法忽视了轨迹的时序性和轨迹点的连续性。为此,提出了一种基于时间特征和空间特征的轨迹相似度计算方法,从时空两个维度计算用户的轨迹距离,并依据轨迹相似度对用户聚类,挖掘不同时间切片下的聚类结果,对亲密度更高的用户对进行“共同漫游行为”的画像。实验结果表明,在无线漫游场景下,该方法可以较为准确地衡量用户之间的相似度,在找出具备社会关系的用户方面具有较好的效果,并能可视化用户间的共同漫游行为。

关键词 校园网, 无线漫游, 时空数据, 轨迹相似度, 聚类

0 引 言

随着无线网络的迅速发展、可连接无线网络设备的层出不穷,用户对无线网络的依赖性日益加深。无线网络的使用在高校中也非常广泛。高校用户使用自己的账号密码,就可以在本校、全市、全国高校组成的无线联盟中通过身份认证使用网络[1]。无线联盟中的无线访问接入点(accesspoint,AP)具有相同的服务集标识(service set identifier,SSID)。用户移动时,保持 WLAN 连接不间断,终端设备通过周围信号最强的 AP 收发数据,这样的过程即为无线漫游[2]。

无线漫游过程中,产生的数据包括上网位置、时长、MAC 地址等信息。这些信息映射着用户的真实漫游轨迹,轨迹相似度越高的用户越可能具有亲密的社交关系[3]。在轨迹相似度的研究中,Zheng 等人[4]对每个用户轨迹中的停留点进行基于密度的聚类,建立停留位置的序列模型,分别计算序列中停留位置的 TF-IDF 值作为相似度衡量用户间的相似性。Ying 等人[5]从 GPS 轨迹中提取用户的频繁轨迹,计算用户间频繁轨迹中的语义相似度。王冠男[6]提出了寻找相似用户,计算分段轨迹的几何相似度,并将时间属性作为向量,分段相似度的时间加权之和即为整体相似度。但 Wi-Fi 数据的轨迹点与 GPS 数据相比,较为稀疏。Hsu 等人[7]利用用户接入和使用无线网络的数据,计算用户间的关联矩阵。周昌令等人[8]引入了终端的稀疏链接区间(sparse linked intervals,SLI)表示方法,采用社交网络分析方法来寻找无线网络中的终端聚集关系。Miao 等人[9]考虑到无线用户的行为习惯较为规律,计算用户轨迹的最长公共子序列相似度,将动态规划的思想加入相似度的计算中。

在高校无线漫游场景下,轨迹相似度可以很好

① 国家重点研发计划(2018YFC0831405)和天津市自然科学基金重点(18JCZDJC30700)资助项目。
② 女,1993 年生,硕士生;研究方向:数据可视化;E-mail: 474765197@qq.com
③ 通信作者,E-mail: jswang@tjut.edu.cn
(收稿日期:2018-11-29)

地衡量用户之间的行为相似度。用户的到达、离开行为趋于一致,且带有明显的时间特征。然而,分别计算轨迹中停留点的相似度,而忽略了时序性,间接地降低了轨迹数据的维度,可能会降低用户相似度衡量的准确性。

本文提出一种基于时空特征的轨迹相似度计算方法。首先,对用户的轨迹按照时间顺序排序。然后,计算相邻时间特征下的用户轨迹空间距离作为实际距离。再对这段实际距离求时间的积分获得分段距离。最后,分段距离之和作为衡量用户相似性的轨迹相似度,并基于轨迹相似度,采用层次化聚类算法找出轨迹较为相似的用户。另外,本文还对相似用户共同的停留位置、在线时长、轨迹进行了可视化展示。

1 轨迹相似度计算方法

基于轨迹相似度,寻找可能具有社会关系用户群体的方法,如图 1 所示。下面将对各部分的功能和实现做具体说明。

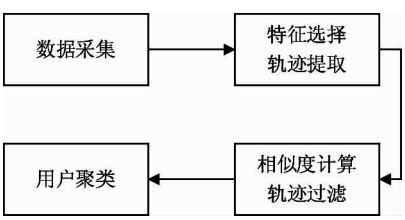

图 1 基于轨迹相似度的用户聚类方法

1.1 数据采集

本文所使用的原始数据是由 Cisco 系统提供的 RESTful 接口定时获取的。RESTful 接口是 HTTP 接口的实现和表现,它将不同资源封装成一组 HTTP 的请求方法。接入控制器(accesspoint controller, AC)收集到数据后,每一种数据资源都存入一条唯一的 URL 中。因此,使用 RESTful 接口的 GET 方法可以从 AC 获取用户使用无线网络的数据。

为避免 Cisco 系统负担过大、影响校园无线网络的正常使用,需设置采集的轮询周期。经测试,在保证数据完整的条件下,每秒取 100 条数据。

1.2 特征选择及轨迹提取

采集到的数据包含大量信息。当用户连接到无线网络时,产生一条连接状态为“ASSOCIATED”的数据。若用户的移动设备超过 10 min 未产生流量,连接状态变为“DISASSOCIATED”,并产生一条下线数据。挖掘用户轨迹需要获取用户与无线网络设备通讯时的连接时间、位置、状态等数据。所以,首先对原始数据中的信息进行特征选择。本文选择了 4 个维度的时空信息,并将其定义为切片数据,形式为四元组:

$$< MAC\_ADDRESS, STATUS, ASSOCIATION\_TIME, AP\_LOCATION >$$

其含义如表 1 所示。

表 1 无线漫游切片数据

漫游数据信息	含义
MAC _ ADDRESS	用户 MAC 地址
STATUS	用户连接状态
ASSOCIATION _ TIME	STATUS 发生改变的时间
AP-LOCATION	AP 所安装的位置

利用切片数据提取轨迹时,MAC 地址是区分不同用户的唯一标识。基于 MAC 地址,将同一用户的切片数据提取为用户轨迹数据,并按照“*ASSOCIATION _ TIME*”增大的顺序对轨迹数据中的切片数据排序,排序后的数据即为用户的漫游轨迹数据。

1.3 相似度计算及轨迹过滤

基于轨迹的相似度计算方法,大多应用于 GPS 轨迹数据的挖掘[10]。在现有的研究中,大部分的轨迹相似度计算方法只考虑轨迹的空间特征。但在挖掘用户行为模式的过程中,若只考虑空间相似度,则有可能把方向完全相反的两条轨迹计算出极高的相似度。若考虑了时空相似度,而把空间相似度和时间相似度分开来看,又难以调整出合理的权重比例。

因此,在计算用户漫游轨迹的相似度时,应当结合空间特征和时间特征。本文中轨迹相似度的计算思想是以漫游轨迹数据中的时间特征为分隔点,计算不同切片数据下用户的空间距离。然后,对空间距离求时间的积分获得分段距离,分段距离之和,即为整体相似度。计算方法如下:

比较两条轨迹 A、B 的相似度时，令 A、B 两个用户的两条轨迹分别为

$$Tr_A = << t_1, d_1, location_1 >, < t_2, d_2, location_2 >, \cdots, < t_i, d_i, location_i >, \cdots < t_m, d_m, location_m >>$$

$$Tr_B = << t_1, d_1, location_1 >, < t_2, d_2, location_2 >, \cdots, < t_i, d_i, location_i >, \cdots < t_n, d_n, location_n >>$$

其中，称 $< t_n, d_n, location_n >$ 为轨迹切片数据；t_n 表示用户停留在该轨迹点的时间点；d_n 表示用户停留在该轨迹点的停留时长，是用户两个相邻切片数据的 *ASSOCITATION _ TIME* 之差；$location_n$ 表示用户停留在该轨迹点的位置；m、n 分别表示两条轨迹的长度。

1.3.1 分段相似度度量

首先将切片数据按照时间增大的顺序排序，形成一条融合轨迹 Tr_S，Tr_S 可表示为

$$Tr_S = << MAC\_A, t_1, d_1, location_1 >, < MAC\_B, t_1, d_1, location_1 >, \cdots, < MAC\_B, t_n, d_n, location_n >>$$

第一次进入计算时，采用临时标记 M_A、M_B 分别标记 Tr_A、Tr_B 的初始切片数据：

$< MAC\_A, t_1, d_1, location_1 >$

$< MAC\_B, t_1, d_1, location_1 >$

比较 M_A、M_B 所在的时刻 t，则 t_i 和 t_{i+1} 有如下 4 种情况：

(1) $t_i \in A$，$t_{i+1} \in B$，且 $t_i < t_{i+1}$，则：

$M_A = t_i$，$M_B = t_{i+1}$ 计算 A、B 之间的空间距离 $\Delta L(M_A, M_B)$，时间距离 $\Delta T(t_i, t_{i+1})$；

(2) $t_i \in A$，$t_{i+1} \in B$，且 $t_i = t_{i+1}$，则：

$M_A = Tr_{S_i}$；

(3) $t_i \in A$，$t_{i+1} \in A$，且 $t_i < t_{i+1}$，则：

计算 A、B 之间的空间距离 $\Delta L(M_A, M_B)$，时间距离 $\Delta T(t_i, t_{i+1})$，再令 $M_A = Tr_{S_i}$；

(4) $t_i \in A$，$t_{i+1} \in A$，且 $t_i = t_{i+1}$，则：

$M_A = Tr_{S_i}$。

循环进行上述计算，并不断更新 M_A、M_B，直到达到两条轨迹中较长轨迹的终点。计算空间距离时，将学校地图的西南角视为坐标原点，建立二维坐标轴模型，计算方法为

$$\Delta L(t_i, t_{i+1}) = \sqrt{(X_{A_i} - X_{B_i})^2 + (Y_{A_i} - Y_{B_i})^2} \quad (t_i \in A, t_{i+1} \in B)$$

时间距离 ΔT 的计算方法为

$$\Delta T(t_i, t_{i+1}) = t_{i+1} - t_i (t_i \in A, t_{i+1} \in B, \text{且 } t_i < t_{i+1})$$

1.3.2 整体相似度度量

得到分段的轨迹相似度之后，整体的用户轨迹相似度 $Sim(A,B)$ 计算公式如下：

$$Sim(A,B) = \frac{1}{\sum_{d=0}^{n} \sum_{h=7}^{19} \int_{i}^{i+1} \Delta L(t_i, t_{i+1}) + 1}$$

其中，n 为采集数据的时间跨度，以“日”为单位。h 代表着用户在不同位置漫游时的连接时间，以“小时”为单位。这是由于用户的活动时间范围通常在白天的 7:00AM – 7:00PM。而生活中无线网络覆盖的范围可能包括居住区，居住区通常具有较大的地理位置分布差异。故为避免 h 受两用户居住位置距离较远的影响，而导致计算出的轨迹相似度误差过大，h 的取值为[7,19]。

整体相似度度量是分段相似度度量之和。$Sim(A,B)$ 的值越小，相似度越低。$Sim(A,B)$ 值越大，相似度越高。

计算出用户两两之间的相似度后，获得一个相似度矩阵，用于下一步用户聚类。

例如，A、B 两位同学 8:00 ~ 10:00 的轨迹，如图 2所示。

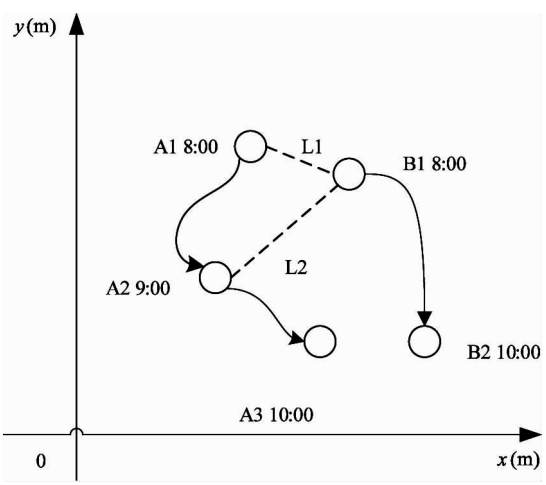

图 2 A、B 两位同学的 Wi-Fi 使用轨迹

他们之间的分段相似度:

$$L1 = \sqrt{(X_{A1} - X_{B1})^2 + (Y_{A1} - Y_{B1})^2}$$

$$\Delta T_1 = T_{A2} - T_{B1}$$

$$L2 = \sqrt{(X_{A2} - X_{B2})^2 + (Y_{A2} - Y_{B2})^2}$$

他们之间的整体相似度:

$$Sim(A, B) = \frac{1}{(L1 \times \Delta T_1 + L2) + 1}$$

获得用户的相似度矩阵后,为了降低用户聚类的复杂度,需要对相似度进行排序,过滤掉相似度过低的用户。

1.4 用户聚类

时空数据的聚类方法包括基于距离的聚类方法、基于密度的聚类方法、基于层次的聚类方法等[11,12]。在无线漫游场景下,聚类的目的是找寻轨迹相似度高的用户群。层次化聚类的思想是将每个用户都视为一个孤立点,寻找与该用户相似度最高的用户进行聚类,更适合这种场景[11]。故本文采用层次化聚类方法对用户聚类。

层次化聚类的步骤为:开始聚类时,每个用户都作为一个孤立点,单独为一个类。从相似度矩阵中,找出与它时空距离最近的类进行连接,生成新的类。然后,重新计算新的类与其他类的时空距离,并更新相似度矩阵。自底向上的不断凝聚两个不同的类,直到达到停止阈值,得到的聚类结果即为轨迹相似的用户。聚类的停止阈值是经过反复实验后,得出的最优阈值。

2 实验及结果分析

2.1 数据集

采集天津理工大学无线网络 Cisco 系统 6 周用户无线网络连接数据作为数据源,数据的具体参数如表 2 所示。

表 2 天津理工大学无线网络数据

参数	无线网络数据
采集时间	2017.11.18 ~ 2017.12.31
AP 类型	Cisco
AP 数量	329
用户数量	24 382

2.2 用户相似度计算

经统计,天津理工大学 24 382 个用户使用无线网络的访问次数如图 3 所示。其中,约有 46% 的用户访问次数高于 7 次,也有部分用户的访问次数较少。用户访问次数较少的原因可能是:这些用户的身份为教师,其常驻的办公室安装有非校园网的独立 Wi-Fi 设备,也可能是这些用户的身份为高年级学生,在校时间较短。这样的用户本身与其他用户的轨迹相似度较低,对本课题的研究意义不大。故过滤掉这些登录次数较低的用户,保留登录次数较高的用户进行相似度计算。

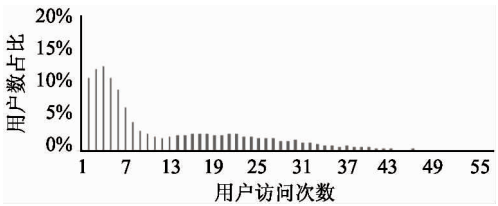

图 3 无线网络用户访问次数

经过实验,我们计算相似度时选取的时间粒度为“天”。计算出每 2 个用户之间的相似度,组成相似度矩阵。以某一天为例,不同用户之间相似度的累积分布函数(cumulative distribution function,CDF)分布如图 4 所示,横轴表示相似度,纵轴代表不同相似度下的 CDF 值。从图 4 可以看出,只有极少数用户的相似度为 0,大部分用户都有一定的相似度。这是因为在计算相似度时,本文选取的时间范围是 7:00 ~ 19:00,用户在食堂、图书馆、公共教学楼等区域相遇的几率较大。另外,存在少数用户之间的相似度趋近于 1,即校园中存在着相似度极高的群体,符合校园网用户的行为模式。

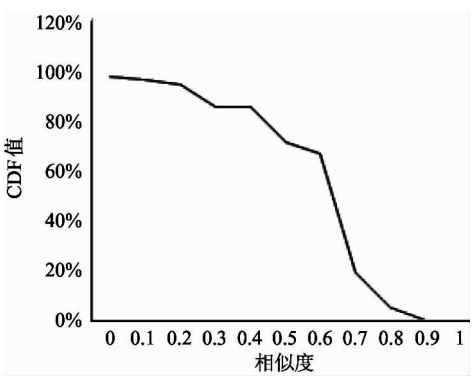

图 4 用户相似度 CDF 分布

2.3　用户聚类

本实验的用户聚类方法采用层次化聚类方法，类与类之间的距离使用 Average Linkage 方法计算。经过实验，将聚类时的相似度阈值设为 0.8。

以某一天为例，类簇内用户数量的分布如图 5 所示。从类簇大小的分布来看，存在类簇内用户数量大于 1 的情况。这说明在校园中存在着轨迹极为相似、关系较为紧密的用户对或群体，符合高校用户群体的行为模式。另外，存在类簇大小为 1 的情况。一方面，可能的原因是用户的选课方案不同，上课地点之间的相对距离较大，导致用户之间的相似度过低。另一方面，可能的原因与无线网络管理系统的设置规则有关。用户快速经过的轨迹点被系统忽略未能生成数据，从而降低了用户的轨迹相似度。

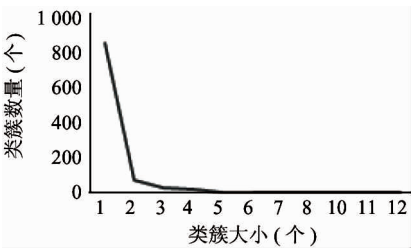

图 5　类簇内用户数量统计图

本文对实验数据进行聚类。如果两个或多个用户多日的轨迹相似度都很高，这样的用户具备亲密社会关系的可能性更大。通过比对每天的聚类结果可以发现，一些用户始终分布在同一类簇中。以某 2 位同学为例，他们的轨迹在 5 周内都被分在了同一类簇中。对他们的行为进行可视化展示，结果如图 6 所示。图 6（a）是这 2 位同学的共同轨迹。图 6（b）是这 2 位同学每周的共同在线时长，以小时为单位。可以发现，这 2 位同学每周在同一个 AP 下的共同在线时长的波动情况，第 2 周明显少于第 1 周，第 3 周又有回升。图 6（c）是 2 位同学共同登录位置的分析图，统计了 2 位同学在不同日期登录同一位置的次数。可以发现，2 位同学共同出现在 27B 教学楼的次数最多。图 6（d）是 2 位同学各自的词云图，可以发现他们共同漫游的位置与各自的漫游位置相比占有较大比重。

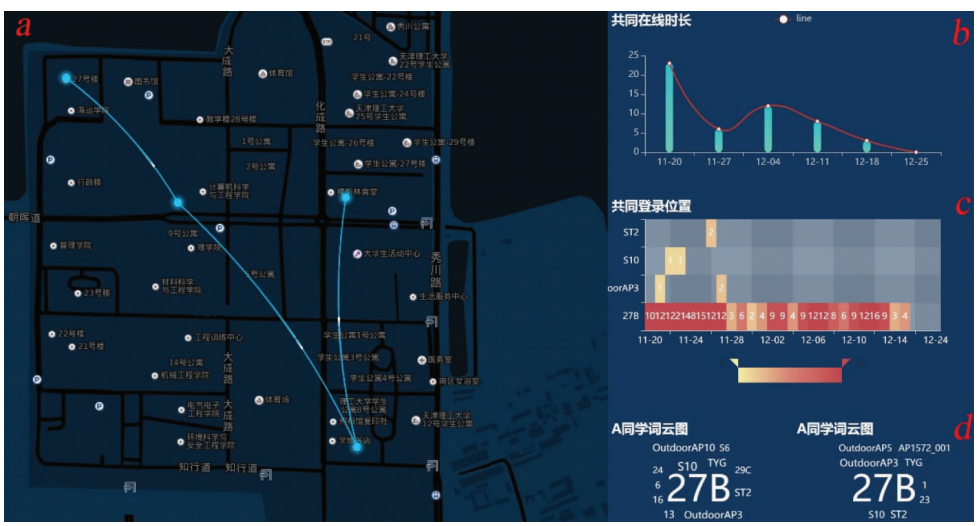

图 6　相似用户漫游行为可视化

2.4　实验验证

为了验证本文算法的有效性，本文以 6 周的无线网络数据为数据源，比较 4 种算法的准确率。4 种算法的简称如表 3 所示。

表 3　轨迹相似度算法及其简称

算法名称	算法简称
本文所提轨迹相似度算法	ST-TS
时间轨迹相似度算法	T-TS
空间轨迹相似度算法	S-TS
最长公共子序列算法	LCS

第 1 种算法为本文提出的轨迹相似度计算方法，结合了时间特征和空间特征计算轨迹相似度。第 2 种算法只使用时间特征计算轨迹相似度。第 3

种算法只使用空间特征计算轨迹相似度。这2种算法分别称为时间轨迹相似度算法、空间轨迹相似度算法。第4种算法为最长公共子序列(longest common subsequence, LCS)算法。LCS是传统轨迹相似度算法,首先将轨迹转换为语义轨迹,并依据时间维度进行动态规划。算法定义:

长度为 m、n 的两条轨迹 A、B:

$A = < x_1, x_2, \cdots, x_n >$

$B = < y_1, y_2, \cdots, y_n >$

设 c 为这 A、B 的最长公共子序列,则:

$$c[i,j] = \begin{cases} 0 & i = 0 \text{ 或 } j = 0 \\ c[i-1, j-1] = 1 & i, j > 0 \text{ 且 } x_i = y_i \\ \max(c[i, j-1], c[i-1, j]) & i, j > 0 \text{ 且 } x_i \neq y_i \end{cases}$$

A、B 轨迹相似度为

$$Sim(A,B) = \frac{c(A,B)}{\min(m,n)}$$

计算出4组轨迹相似度后,采用层次化聚类方法对4组相似度进行聚类分析。将聚类结果交由无线网络管理人员进行评估。管理人员进行评估时,如果计算结果与现实情况相符,则证明相似度计算正确。反之,则证明结果错误。

经验证,以上4种算法的正确率如图7所示。与其他算法相比,本文提出的算法在不同相似度阈值下均具有较好的准确率。对用户聚类时,采用的最优相似度阈值应为0.8。当阈值大于0.8后,孤立点迅速增多,原本属于同一个类的用户未被分入同一类中,造成准确率降低。

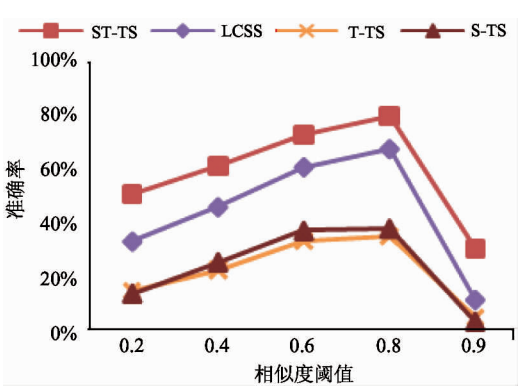

图7 轨迹相似度算法准确率

3 结 论

本文结合无线漫游场景下轨迹数据的时序特征,首先,对传统轨迹相似度的计算方法进行了改进,计算了时序特征下的时空距离来衡量不同用户漫游轨迹之间的差异。然后,利用聚类方法找出了轨迹相似的用户,并进一步挖掘这些用户在不同的时间切片上其轨迹是否高度相似。最后,对多日内轨迹相似的用户对,进行可视分析,刻画用户对的轨迹及时空特征上的画像。本文的研究对用户群体行为的研究、预测具有重要的意义。

参考文献

[1] Wang D, Wang N, Wang P, et al. Preserving privacy for free: efficient and provably secure two-factor authentication scheme with user anonymity[J]. *Information Sciences*, 2015, 321:162-178

[2] He D B, Wang D, Xie Q, et al. Anonymous handover authentication protocol for mobile wireless networks with conditional privacy preservation[J]. *Science China Information Sciences*, 2017, 60(5):109-125

[3] 郑宇,谢幸. 基于用户轨迹挖掘的智能位置服务[J]. CCF通讯(中文版),2010, 6(6): 23-29

[4] Zheng Y, Zhang L Z, Ma Z X, et al. Recommending friends and locations based on individual location history [J]. *ACM Transaction on the Web*, 2011, 5(1):1-44

[5] Ying J J, Lee W C, Weng T C, et al. Semantic trajectory mining for location prediction[C]. In: ACMSIG Spatial International Conference on Advances in Geographic Information Systems, Chicago, USA, 2011. 34-43

[6] 王冠男. 基于GPS轨迹和照片轨迹的时空数据挖掘[D]. 长沙:中南大学数学与统计学院, 2013. 20-25

[7] Hsu W, Dutta D, Helmy A. Structural analysis of user association patterns in university campus wireless LANs [J]. *IEEE Transactions on Mobile Computing (TMC)*, 2012,11(11):1734-1748

[8] 周昌令, 钱群, 赵伊秋, 等. 校园无线网用户群体的移动行为聚集分析[J]. 通信学报, 2013,34(z2):111-116

[9] Miao C, Zhu X, Miao J. The analysis of student grades based on collected data of their Wi-Fi behaviors on cam-

pus[C]. In: IEEE International Conference on Computer and Communications, Chengdu, China, 2017. 130-134

[10] Magdy N, Sakr M A, Mostafa T, et al. Review on trajectory similarity measures[C]. In: IEEE 7th International Conference on Intelligent Computing and Information Systems, Cairo, Egypt , 2016. 613-619

[11] Zhang J, Lin Y, Lin M, et al. An effective collaborative filtering algorithm based on user preference clustering [J]. *Applied Intelligence*, 2016, 45(2):230-240

[12] Craenendonck T V, Blockeel H. Constraint-based clustering selection[J]. *Machine Learning*, 2017, 106(9-10): 1497-1521

Algorithm for trajectory similarity in wireless roaming scene

Chang Yiyi[*,**,***], Wang Jinsong[*,**,***]

([*]School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384)

([**]Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin 300384)

([***]National Engineering Laboratory for Computer Virus Prevention and Control Technology, Tianjin 300457)

Abstract

Users generate a lot of behavior data when roaming under wireless network. These data contain users' trajectories. The more similar the trajectory is, the more likely users have intimate social relationship. Traditionally, the similarity of trajectory between users is mined by comparing the longest common subsequences of two semantic trajectories. However, these kinds of algorithms do not focus on sequential properties and continuity of trajectories. In this paper, a method of trajectory similarity computation based on spatio-temporal character is proposed. The trajectory similarity of users is calculated from two dimensions of time and space. Users are clustered based on the trajectory similarity. Then clustering results are mined under different time slices, drawing a persona for users who have higher intimacy as a "common roaming behavior" persona. Experimental results show that this method can accurately measure the similarity between users, performs better to find social relationship, and can visualize the common roaming behavior among users in wireless roaming scenarios.

Key words: campus network, wireless roaming, trajectory similarity, spatio-temporal data, clustering