计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(11):271−277 [doi: 10.15888/j.cnki.csa.006656]

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

基于话单数据的移动通信用户画像研究①

张海旭[1], 胡访宇[1], 赵家辉[2]

[1](中国科学技术大学 信息科学技术学院, 合肥 230027)
[2](安徽省公安厅 科技信息化处, 合肥 230061)
通讯作者: 张海旭, E-mail: hxz2015@mail.ustc.edu.cn

摘　要: 用户通话产生的详细话单数据具有丰富的时空信息和社交信息, 这些信息在一定程度上反映了用户的生活习惯和社交模式, 对于移动通信用户画像研究具有重要意义. 我们的研究是基于中国某运营商提供的 10 000 名用户一个月详细话单数据, 本文从用户日常移动模式方面提取移动距离、回旋半径、访问点个数和移动方向熵特征, 从用户社交生活方面提取通话时长、联系人数量、主叫比率和社交熵特征, 利用上述特征对用户进行群体划分和构建用户词云名片, 从而完成对移动通信用户的画像研究. 本文使用用户话单数据为推测用户属性、理解用户特征提供了新的视角.

关键词: 话单数据; 移动模式; 社交生活; 用户画像

引用格式: 张海旭,胡访宇,赵家辉.基于话单数据的移动通信用户画像研究.计算机系统应用,2018,27(11):271–277. http://www.c-s-a.org.cn/1003-3254/6656.html

Mobile Communication User Profiling Based on Call Detail Records

ZHANG Hai-Xu[1], HU Fang-Yu[1], ZHAO Jia-Hui[2]

[1](School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China)
[2](Science and Technology Informatization Office, Public Security Department, Anhui Province, Hefei 230061, China)

Abstract: Call detail records contain rich spatio-temporal information and social information, which partly reflect users' habits and social pattern. It is of great significance for the study of mobile communication user profiling. Our study is based on a monthly call detail records of 10 000 subscribers provided by a Chinese telecom operator. In this study, on the one hand, we extract the moving distance, the radius of gyration, the number of access points, and the entropy of moving directions to characterize user's mobile pattern. On the other hand, we extract the call duration, the number of contact, the ratio of calling, and the entropy of sociality to characterize user's social life. Then users are divided into groups and each user gets a word cloud card based on these features. So the portrait study of mobile communication users is completed. Our work is a promising step towards inferring user attributes and understanding user characteristics using call detail records.

Key words: call detail records; mobile pattern; social life; user profiling

1　引言

随着我国移动通信市场的迅速发展, 手机已成为人们日常生活中不可或缺的一部分. 用户在使用手机的过程中产生了大量的个人历史数据, 这些数据可以概括为以下几种: 1) 位置信息, 通过全球定位装置(Global Positioning System, GPS)、手机信号塔等方式

① 基金项目: 安徽省科技计划 (1201b0403021)
Foundation item: Science and Technology Project of Anhui Province (1201b0403021)
收稿时间: 2018-04-23; 修改时间: 2018-05-21; 采用时间: 2018-05-23; csa 在线出版时间: 2018-10-24

获取的地理位置信息; 2) 使用信息, 记录了用户在何时使用了手机做了什么; 3) 社交信息, 隐含在话单数据, GPS 以及通讯录等数据里. 这些历史数据隐含了与用户相关的个性化信息, 反映了用户的生活习惯和社交模式. 这些数据为研究用户属性和特征提供了新的渠道.

话单数据是运营商计费所产生的. 话单数据有被动产生、覆盖范围广、成本低、分析周期短等优点, 已经在了解人们的移动模式[1], 理解人类行为动力学特征[2,3], 感知用户所在地区的地理环境、生活方式、交通状况和发展水平等[4,5]方面广泛地使用. 例如 Etienne Thuillier 等[6]使用话单数据, 根据用户每天与预设区域的关系, 将用户划分为 6 类, 在此基础上, 对用户进行以一周时间为周期的聚类分析, 发现了 12 种类型的周活动模式. 杨喜平、方志祥[7]等利用移动电话位置数据, 理解人类时空聚散模式. Schneider等[8]借鉴复杂网络中模体的概念, 发现人们日常生活中存在的 17 中网络结构, 然后使用模体来概括来自不同国家人们的时空移动模式. Jiang 等[9]以新加坡为例, 演示了如何使用手机通话详细记录 (CDR) 数据, 其中包含数百万匿名用户, 以提取可与基于活动的方法相媲美的个人移动网络.

手机话单数据中含有丰富的时空信息和社交信息, 目前基于话单数据的研究多集中在分析数据中的时空信息. 本文同时利用话单数所包含的时空信息和社交信息, 提取用户特征, 发现特征相似的用户群体和为用户创建个性化词云名片, 完成对用户画像. 文本研究, 为理解用户特征提供新的视角, 为生产生活的提高、相关政策的制定提供了参考.

2 数据集与研究方法

2.1 实验数据集

本文手机通话数据由合作单位某运营商提供, 为保护用户隐私, 用户号码已作匿名化处理. 数据分为两部分: 手机通话话单数据, 由 10 000 名用户在 2013 年 6 月一个月期间通话产生的话单数据, 数据格式如表 1 所示; 基站小区位置信息数据, 14 549 个基站小区的 GPS 坐标、行政划分、道路等信息, 数据格式如表 2 所示. 其中手机用户选取条件如下:

1) 用户号码注册于一个匿名的高科技工业区注册;

2) 用户在 2013 年 6 月 1 日~6 月 30 日一个月内的通话总时长大于 100 分钟.

表 1 话单数据格式

字段	样例
用户手机号码	“139****9132”
对端用户号码	“138****3151”
通话发起时刻	“2013-06-25-11.10.06.000000”
通话时长	18(min)
呼叫类型	“0”(主叫) 或“1”(被叫)
位置区 ID	“ORZR”
小区 ID	“183D”

表 2 通信小区信息格式

字段	样例
位置区 ID	“ORZR”
小区 ID	“183D”
经度	“116.3294”
纬度	“39.9783”

2.2 研究方法

本文同时利用话单数所包含的时空信息和社交信息, 从用户日常移动模式和社交生活两个方面来刻画用户特征. 在提取特征时, 提出衡量用户移动随机程度的移动方向熵特征和衡量用户社交集中程度的社交熵特征. 对用户一个月内的特征进行分析, 然后使用 K-MEANS 聚类算法[10]用户进行聚类分析, 完成用户群体划分. 接着时间窗口设为一周, 利用每周内特征的均值与均方差, 给用户打上标签, 完成对用户个性化特征的刻画, 构建用户词云名片.

(1) 用户特征定义

为了描述用户的移动模式, 本文从移动强度、活动范围、移动随机程度以及出行的随机性等角度提出定义用户移动模式的特征; 从用户社交圈的规模、主动程度、社交上的精力以及会交往集中程度等角度提出定义用户社交生活的特征.

与朋友发生的相互通话是一个人社交生活中的重要表现形式. 通过对用户的通话时长、联系人数量、主叫比率和社交熵进行提取, 以得到反映用户的社交能力的特征.

定义 1. 移动距离特征定义为在一定时间内用户移动轨迹的长度, 是用户移动强度的体现, 公式为:

$$d(t)=\sum_{i=1}^{n_c(t)}\sqrt{(\overrightarrow{x_i}-\overrightarrow{x_{i-1}})^2} \quad (1)$$

其中, $n_c(t)$表示用户在 t 时间内的通话次数, $\overrightarrow{x_i}, i=1,2,\cdots,n_c(t)$表示用户通话发生时刻的位置.

定义 2. 回旋半径特征定义为在一定时间内用户通话发生时刻所在地点偏离移动轨迹重心距离的标准差,

可以表示用户的移动范围, 公式为:

$$r_g(t)=\sqrt{\frac{1}{n_c(t)}\sum_{i=1}^{n_c(t)}(\overrightarrow{x_i}-\overrightarrow{x_{cg}})^2} \tag{2}$$

其中, $\overrightarrow{x_{cg}}=\frac{1}{n_c(t)}\sum_{i=1}^{n_c(t)}\overrightarrow{x_i}$表示用户在 t 时间内所有位置的重心.

定义 3. 访问点个数特征定义为用户的所有发起通话地点的个数, 可以反映用户活动的规律, 公式为:

$$n_{ap}(t)=\frac{1}{n_c(t)}\sum_{i=1}^{n_c(t)}f(\overrightarrow{x_i}) \tag{3}$$

其中仅当首次计算到位置$\overrightarrow{x_i}$时, $f(\overrightarrow{x_i})=1$, 否则 $f(\overrightarrow{x_i})=0$.

定义 4. 将以东西方向为横坐标轴, 南北方向为纵坐标轴组成的坐标系均分成 12 个方向区间$\theta_1,\theta_2,\cdots,\theta_{12}$. 计算出用户每次出行方向, 然后统计用户出行方向位于各方向区间的概率: $p(\theta_1),p(\theta_2),\cdots,p(\theta_{12})$, 计算其信息熵作为用户的移动方向熵特征, 反映用户出行方向的随机性, 公式为:

$$E(direction)=\sum_{i=1}^{12}-p(\theta_i)*\log(p(\theta_i)) \tag{4}$$

定义 5. 通话时长特征定义为指用户在一段时间内所有通话时间的总和, 可以反映用户在“电话社交”中的活跃程度, 公式为:

$$c_d(t)=\sum_{i=1}^{n_c(t)}c_i \tag{5}$$

其中, c_i表示用户第 i 次通话的通话时间.

定义 6. 联系人数量特征定义为所有和用户发生通话行为的人数总和, 可以体现用户社交圈的规模, 公式为:

$$n_m(t)=\sum_{i=1}^{n_c(t)}g(m_i) \tag{6}$$

其中仅当首次计算到对端用户m_i时, $g(m_i)=1$, 否则 $g(m_i)=0$.

定义 7. 主叫比率特征定义为在一定时间内用户主叫通话次数与总的通话次数的比率, 可以体现用户在社交中的主动程度, 公式为:

$$r_c(t)=\frac{\sum_{i=1}^{n_c(t)}h(v_i)}{n_c(t)} \tag{7}$$

其中仅当呼叫类型v_i为主叫时, $h(v_i)=1$, 否则$h(v_i)=0$.

定义 8. 在一段时间内用户与 n 个用户发生总共 N 次通话, 其中与 n 个用户的通话次数分别为$u_1,u_2,\cdots,u_n$, 计算熵值作为用户的社交熵特征. 社交熵特征可以反映社会交往集中程度, 公式为:

$$E(sociality)=\sum_{i=n}^{n}-(\frac{u_i}{N})*\log(\frac{u_i}{N}) \tag{8}$$

(2) 特征相关性分析

为了从整体上了解用户, 将时间窗口 T 设定为一个月, 计算用户在一个月时间内, 在移动模式和社交生活两方面的特征向量 F^T, F^T 的定义如下:

$$F^T=(f_1^T,f_2^T,f_3^T,f_4^T,f_5^T,f_6^T,f_7^T,f_8^T) \tag{9}$$

其中$f_i^T,i=1,2,\cdots,8$分别代表移动距离 (DD)、访问点个数 (AP)、回旋半径 (RG)、移动方向熵 (DE)、通话时长 (CD)、联系人数量 (CC)、主叫比率 (CR) 和社交熵 (SE) 特征.

为了进一步了解代表移动模式和社交生活的特征, 为了消除特征之间的差异性, 对每一维特征进行 z-score 标准化:

$$\chi^*=\frac{\chi-\mu}{\sigma} \tag{10}$$

式中, μ代表所有用户特征数据的均值, σ为所有用户特征数据的标准差.

通过计算标准化后特征之间的相关系数, 分析本文提取特征之间的相关性.

(3) 用户群体发现

本文选择使用多特征对用户进行聚类, 根据话单数据发现移动模式和社交模式类似的用户群体. 首先将代表用户将时间窗口 T 设为一个月, 提取用户一个月内的八个特征. 考虑到本文提取的八个特征间可能存在一定的相关性并且可能存在冗余和噪声, 本文对八个特征进行主成分分析, 提取主要特征成分. 选择保留 90% 以上的方差信息, 来确定主成分的个数. 在此基础上根据提取的主成分使用 K-MEANS 聚类算法对用户进行聚类, 发现用户群体. 因为 K-MEANS 聚类算法是一种简单、快速的算法, 并且当处理大数据集时, 也可保持伸缩性和高效性, 所有选择它作为本文的距离算法.

(4) 用户词云名片生成

词云图一种基于信息文本词频的可视化形式, 是对文本信息中出现频率较高的“关键词”予以视觉化的展现. 词云图可以将重点内容突出, 过滤掉的低频低质的内容, 使得浏览者只要一眼扫过便可领略主旨. 词云

图被广泛的使用在艺术、新闻学、社交网络等不同的领域. 生成词云图的方法有很多, 如 Wordle、WordItOut 还有 Python 库 wordcould, 本文采用 WordItOut 工具, 为用户生成词云名片.

本文借助词云图方式为用户制作词云名片, 使用户特点被清晰地呈现. 构建用户词云名片, 关键是要找到用户与众不同的特点, 利用一定的规则生成用户标签. 本文根据用户特征值的均值和均方差, 将特征值位处于整体分布两端的用户打上标签, 为生成词云名片提供数据. 然后将用户的标签数据送入 WordItOut 工具, 为用户生成个性化的词云名片.

3 实验和分析

3.1 数据预处理

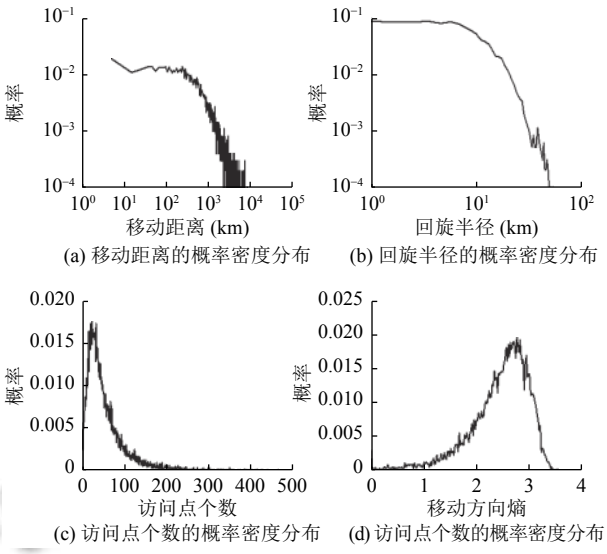
由于 CDR 数据需要关联了小区的位置信息才能用于对用户定位, 而二者主要通过位置区 ID 和小区 ID 建立起关联. 统计发现, 数据集中的小区 ID 已经具有唯一性, 故删除了 CDR 数据中小区 ID 缺失或未被包含在小区信息数据集里的记录, 最终共得到 9514 位用户的 2380 598 条话单数据.

3.2 特征提取

将时间窗口 T 设为一个月, 提取用户一个月内的八个特征. 用户移动模式特征的概率密度分布如图 1 所示, 用户社交生活特征的概率密度分布如图 2 所示. 移动距离、回旋半径、通话时长和联系人数量特征值主要集中在一定范围内, 超过一定值后, 概率会迅速下降且出现重尾现象, 特征值较大的用户稀疏的存在. 主叫比率和社交熵概率密度函数服从正态分布. 访问点个数的峰值处于较小数值段, 概率密度函数在达到峰值前增长较快, 达到峰值后下降比较缓慢. 和访问点个数特征的概率密度函数相反, 用户的移动方向熵的峰值处于较大的数值段, 在达到峰值前增长缓慢, 达到峰值后下降很快, 说明存在少量出行方向随机性很强的用户.

3.3 特征相关性分析

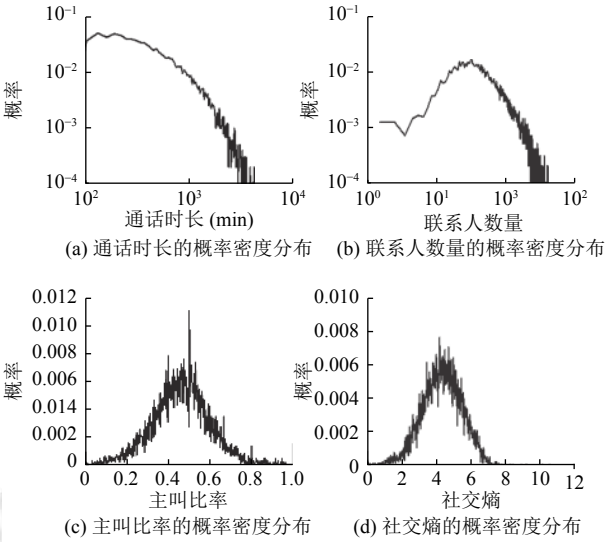
为了进一步了解代表移动模式和社交生活的特征, 计算标准化后特征之间的相关系数, 结果如表 3 所示. 由表 3 可知, 特征间存在 6 对显著相关 ($0.5<r<0.8$) 的特征, 不存在高度相关 ($r>0.8$) 的特征对. 对显著相关的特征对解释如下:



(a) 移动距离的概率密度分布 (b) 回旋半径的概率密度分布

(c) 访问点个数的概率密度分布 (d) 访问点个数的概率密度分布

图 1 四种移动模式特征的概率密度分布



(a) 通话时长的概率密度分布 (b) 联系人数量的概率密度分布

(c) 主叫比率的概率密度分布 (d) 社交熵的概率密度分布

图 2 四种社交生活特征的概率密度分布

表 3 不同特征之间的相关性

特征	DD	RG	AP	DE	CD	CC	CR	SE
DD	1.000	-	-	-	-	-	-	-
RG	0.511	1.000	-	-	-	-	-	-
AP	0.644	0.284	1.000	-	-	-	-	-
DE	0.204	0.288	0.268	1.000	-	-	-	-
CD	0.441	0.085	0.616	–0.062	1.000	-	-	-
CC	0.412	0.055	0.570	–0.038	0.577	1.000	-	-
CR	0.042	0.025	0.018	–0.144	0.082	–0.074	1.000	-
SE	0.338	0.104	0.485	0.045	0.345	0.706	–0.155	1.000

(1) 移动距离和回旋半径 (r=0.551)、移动距离和访问点个数 (r=0.644) 存在显著的相关性. 这个不难理解, 用户移动距离越大, 可能伴随着活动范围越大、发生通话的地点越多.

(2) 访问点个数和通话时长 (r=0.616)、访问点个数和联系人数量 (r=0.570) 存在显著的相关性. 因为本实验中的社交信息是由话单数据体现, 所以通话时间长、联系人比较多的用户记录的话单数据越详细, 导致他们的访问点数目也比较多.

(3) 联系人数量和通话时长 (r=0.577)、联系人数量和社交熵 (r=0.706) 存在显著的相关性. 用户联系人数量越多, 总的通话时长也有很大概率越大, 同样由于社交熵的定义, 用户的社交熵也很大概率越大.

3.4 用户群体发现

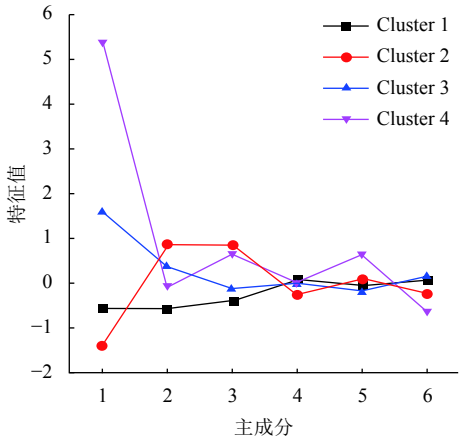
将时间窗口 T 设为一个月, 提取用户一个月内的八个特征. 对用户特征值进行主成分分析, 选择保留 90% 以上的方差信息, 保留了六个主成分. 对保留的特征主成分使用 K-MEANS 聚类算法对用户进行群体划分, 参考轮廓系数, 通过测试和调整, 最终确定 k=4. 将每一类的聚类中心点作图如图 3(a) 所示. 为了对聚类结果有清楚的认识, 使用每一类用户的原始八个特征对聚类结果进行展示. 计算每一类用户的原特征的平均值, 将每一类用户的特征平均值作图如图 3(b) 所示.

从图 3(a) 中可以看到用户在特征主成分上被很好地分离开了, 尤其是在占主导作用的前 3 个主成分方面. 接下来根据图 3(b) 对用户群体发现结果进行解释说明.

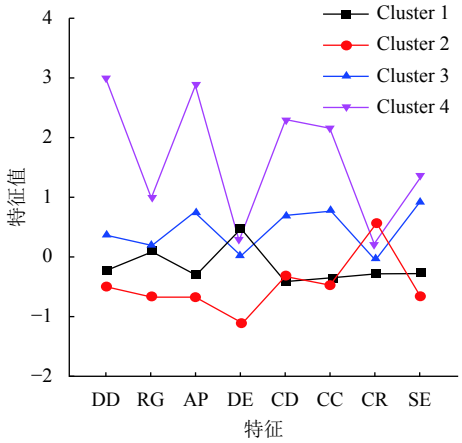
Cluster 1 共有 4735 人, 占比为 49.8%. 这部分用户最多, 他们的日常移动模式特征和社交生活特征值均在平均值上下 0.5 左右, 反映了数据集中大部分用户的移动模式和社交生活的特点.

Cluster 2 共有 2227 人, 占比为 23.4%. 他们日常移动模式特征值均是四类用户中最小的, 在社交生活特征方面, 在通话时长特征与大部分用户相仿的前提下, 社交熵特征和联系人数量特征值比大部分用户小, 主叫比率特征值却最大, 说明这类用户日常移动性较差, 社交圈相对集中, 并且通话多数都是主动.

Cluster 3 共有 2119 人, 占比为 22.3%. 在日常移动模式特征方面, 回旋半径特征和大部分用户相同, 访问点个数特征和移动距离特征比大部分用户大, 移动方向熵特征却比大部分用户小; 在社交生活特征方面, 四种特征值都比大部分用户大. 这代表这类用户的活动范围虽然和大部分用户差不多, 但但移动距离更大, 活动地点更多并且移动更有规律, 平时通话时间长, 联系人多, 社交圈也比较广, 与朋友联系一般为主动联系.



(a) 聚类结果在主成分特征表示



(b) 聚类结果在原始特征表示

图 3 用户聚类结果

Cluster 4 共有 433 人, 占比为 4.5%. 这类用户最少, 他们除了移动方向熵特征、主叫比率特征外的其他特征都远大于其他用户, 他们活动范围广, 移动距离长, 访问点多, 通话时间长, 社交圈也广, 是数据集中最活跃的那一部分群体.

3.5 用户词云名片生成

构建用户词云名片的关键是制订规则发现用户与众不同的特点并生成标签数据. 本文标签制订规则如表 4 所示, 首先计算每一维特征整体均值 $mean$ 和均方差 std. 将特征值 f^T 落在区间$[mean-std, mean+std]$外的用户按照表 4 所示规则添加标签.

人们的工作生活多数以星期作为周期, 因此将时间窗口 T 设为一个星期, 这样样可以获得更多的用户标签, 以对用户进行更详细的分析. 计算用户的特征向量 f^T, 然后根据表 4 所示规则计算用户标签, 最后将每位用户获得的标签分别送入 WordItOut 工具, 就生成了用户的词云名片.

表 4 标签制订规则

特征	标签 (f^T<(mean-std))	标签 (f^T>(mean-std))
移动距离	DD-Ⅰ	DD-Ⅱ
回旋半径	RG-Ⅰ	RG-Ⅱ
访问点个数	AP-Ⅰ	AP-Ⅱ
移动方向熵	DE-Ⅰ	DE-Ⅱ
通话时长	CD-Ⅰ	CD-Ⅱ
联系人数量	CC-Ⅰ	CC-Ⅱ
主叫比率	CR-Ⅰ	CR-Ⅱ
社交熵	SE-Ⅰ	SE-Ⅱ

取实验中两名用户的用户词云名片展示如图 4, 可以发现用户 1 的词云名片中 DD-Ⅱ、AP-Ⅱ、CD-Ⅱ和 CC-Ⅱ比较突出, 它们表示用户 1 的移动距离特征、访问点个数特征、通话时长特征和联系人数量特征位于区间$(mean+std,+\infty)$中, 其它特征处于正常水平. 这表明用户 1 移动距离大, 访问地点多, 同时通话时间长, 联系人比较多. 基于此可以推测用户 1 可能是在较大城市区域内从事联系交流工作的室外工作者; 而用户 2 的词云名片中 DE-Ⅰ、CR-Ⅱ和 AP-Ⅰ比较突出, 它们表示用户 2 的移动方向熵特征和访问点个数特征位于区间$(-\infty,mean-std)$中, 主叫比率特征位于区间$(mean+std,+\infty)$中, 其它特征处于正常水平. 这表明用户 2 活动地点少且移动具有规律性, 通话多为主叫, 基于此用户 2 可能是喜欢宅在某些地点, 用电话处理日常生活的人.

(a) 用户1的词云名片 (b) 用户2的词云名片

图 4 用户词云名片

4 总结

本文利用用户话单数据提取出多个反映用户时空信息和社交信息的特征, 在真实的数据上通过对特征的综合分析, 完成了对移动通信用户的画像研究. 基于用户的多方面特征, 发现了四类移动模式和社交生活相似性的用户群体, 创建了用户词云名片的使得用户个体的特点可以被清晰地呈现. 以本文研究为基础, 移动通信运营商可以针对用户特点制订相应的套餐并向用户推荐, 其他利益相关企业可以针对用户特点推荐相关的商品, 实现精准营销; 在城市治理方面, 可以通过对用户的移动性和行为模式的分析, 识别非法营运车辆的从业人员.

由于话单数据是由通话事件触发采样的, 因此用户移动行为、社交行为只有在通话行为发生的情况下才能被记录, 所以本文结果具有一定的局限性. 受实验话单数据获取途径的限制, 不能在更大数据集下对本文提出的方法和分析结果进行进一步地研究. 今后的工作将主要从两个方向进行展开: 第一, 挖掘话单数据中隐含更多的特征, 从多角度对用户间的差异性进行表达; 第二, 获得信息更加丰富的实验数据, 增加数据种类, 通过多种数据对比、融合来刻画用户画像.

参考文献

1 Calabrese F, Diao M, Di Lorenzo G, *et al*. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies, 2013, 26: 301–313. [doi: 10.1016/j.trc.2012.09.009]

2 Becker R, Cáceres R, Hanson K, *et al*. Human mobility characterization from cellular network data. Communications of the ACM, 2013, 56(1): 74–82. [doi: 10.1145/2398356]

3 González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature, 2008, 453(7196): 779–782. [doi: 10.1038/nature06958]

4 Amini A, Kung K, Kang CG, *et al*. The impact of social segregation on human mobility in developing and industrialized regions. EPJ Data Science, 2014, 3: 6. [doi: 10.1140/epjds31]

5 Järv O, Ahas R, Saluveer E, *et al*. Mobile phones in a traffic flow: A geographical perspective to evening rush hour traffic analysis using call detail records. PLoS One, 2012, 7(11): e49171. [doi: 10.1371/journal.pone.0049171]

6 Thuillier E, Moalic L, Lamrous S, *et al*. Clustering weekly patterns of human mobility through mobile phone data. IEEE Transactions on Mobile Computing, 2018, 17(4): 817–830. [doi: 10.1109/TMC.2017.2742953]

7 Yang XP, Fang ZX, Xu Y, *et al*. Understanding spatiotemporal patterns of human convergence and divergence using mobile phone location data. ISPRS International Journal of Geo-Information, 2016, 5(10): 177. [doi: 10.

3390/ijgi5100177]

8 Schneider CM, Belik V, Couronné T, *et al*. Unravelling daily human mobility motifs. Journal of the Royal Society Interface, 2013, 10(84): 20130246. [doi: 10.1098/rsif.2013.0246]

9 Jiang S, Ferreira J, Gonzalez MC. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. IEEE Transactions on Big Data, 2017, 3(2): 208–219. [doi: 10.1109/TBDATA.2016.2631141]

10 Kanungo T, Mount DM, Netanyahu NS, *et al*. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881–892. [doi: 10.1109/TPAMI.2002.1017616]