电信数据中用户行为特征测量与分析

宋 竹，秦志光，罗嘉庆，张悦涵

(电子科技大学计算机科学与工程学院 成都 611731)

【摘要】通话和上网是电信运营商的重要业务，研究通话和上网的行为规律有助于提升电信运营商的业务规划和管理水平。现有的研究工作通常只关注于手机通话或上网行为，很少同时对两类行为进行关联的分析。该文提取了电信数据中手机通话与上网的基本特征，对通话和上网行为的频率分布进行了曲线拟合。通过比较两类行为的拟合参数与相关系数，发现了工作日与周末、以及周六与周日显著不同的用户行为特征。通过对通话和上网时间的归一化，定义了用户的使用偏好，发现54%的手机用户更多的倾向于使用手机通话，而31%的用户则倾向于使用手机上网。

关 键 词 曲线拟合； 频率分布； 测量； 电信数据； 统计分析； 用户行为

中图分类号 TP399 **文献标志码** A doi:10.3969/j.issn.1001-0548.2015.06.024

Measurement and Analysis of User Behaviors in Mobile Data

SONG Zhu, QIN Zhi-guang, LUO Jia-qing, and ZHANG Yue-han

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract Mobile calls and mobile-internet surfing are two important telecom services. The study of user behaviors on mobile calls and mobile-internet surfing is of great value to telecom operators in improving business planning and business management. Previous studies on user behaviors mainly focus on either mobile calls or mobile-internet surfing. There is little research that is conduced into the study of the interaction between mobile calls and mobile-internet surfing. In this paper, we first capture the basic features of the user behaviors in mobile calls and mobile-internet surfing. Through the curve fitting of frequency distributions of mobile calls and mobile-internet surfing, we find that there exist significant differences of user behaviors between workdays and weekends. We also normalize and compare the time that users spend on mobile-internet surfing and mobile calls. The results show that over 54% of users prefer using mobile calls, and over 31% of users prefer using mobile-internet surfing.

Key words curve fitting; frequency distribution; measurement; mobile data; statistical analysis; user behavior

如今智能手机在人们的日常生活中扮演着越来越重要的角色，随着移动网络技术的提升及智能手机的普及，移动互联网业务也逐渐成为继手机通话和短信之后的重要电信业务。通过对用户的电信通话和流量账单数据进行测量分析，可以帮助提取和挖掘手机用户的行为特征、发现用户的行为模式，对电信运营商的商业策略优化、服务水平提升也有重要的指导意义。

本文通过对中国某电信运营商4天的电信通话和流量账单数据的测量分析，着重关注用户细粒度的宏观通话行为和移动互联网行为频率分布的特点与差异，并使用相关系数、傅里叶函数和多项式函数等对频率分布进行深入研究。不同于经典的时间间隔频率分布的分析，采用在线时间频率分布来探索手机通话行为与移动互联网行为的特征与规律。

1 相关工作

现有的研究大多专注于手机通话或移动互联网的使用。近来基于手机通话的研究包括：关注手机通话间隔的分布以研究大尺度的集体行为和异常事件的发生[1]、对不同时间序列的手机用户行为模式进行研究[2]、讨论不同时间因素对手机用户移动轨迹的影响[3]以及对加权的无向移动网络的研究[4]。文献[5]使用出度、去话的比例以及通话的差异3种指标来量化个体用户的行为并对用户进行分类。目前为止，对手机用户行为特征定义的参数主要为时间间

收稿日期：2014－02－24；修回日期：2015－09－15
基金项目：国家863项目(2011AA010706)；国家自然科学基金(61170041)
作者简介：宋竹(1983－)，男，博士生，主要从事数据分析及智能交通方面的研究.

隔、时间、轨迹及通话使用等。

近来关于使用移动互联网的研究包括对移动互联网的统计分析[6]和移动互联网用户的行为分析[7]。文献[8]定性研究了活跃的手机互联网用户，并提出了一种初步框架来理解用户的动机和行为。文献[9]对移动互联网服务扩散模式进行研究。文献[10]主要关注于划分移动互联网的用户群体。文献[11]研究不同年龄群体用户使用移动互联网的行为模式。文献[12]讨论了造成日本移动互联网用户独特行为模式的原因。文献[13]研究基于不同时间的不同手机用户上网行为模式。文献[14]对移动互联网数据的时间模式，地理位置和移动性进行统计分析。文献[15]研究移动互联网与传统互联网不同的用户行为。文献[16]研究用户访问各类网页的行为。文献[17]测量分析了不同类型与不同设备的流量使用特征。文献[18]通过基于上下文的算法推测用户的3种移动互联网使用模式。文献[19]研究手机用户访问各类网页的行为随时间的变化。目前为止，对移动互联网用户行为特征定义的主要参数为时间、时间间隔、地理位置及访问行为等。

可以看出，分析手机通话行为与上网行为特征的参数有相似性，但目前并没有研究同时分析手机通话和上网行为的关系和差异。本文与上述工作不同在于本文着重于相关联的分析手机通话和手机上网的行为特征，通过基于时间的宏观频率分布来探索不同行为间的关系和差异，以及造成差异的原因。

2　统计分析

2.1　数据来源

本文使用统计分析的方法对电信数据集进行分析。该数据集是由中国某城市某电信运营商提供的匿名手机用户的实际账单数据，涵盖了该城市2011年12月1日至4日1 012个基站的所有账单记录。该运营商在该地区拥有150万活跃用户，是当地重要的电信运营商之一。

数据集包含通话账单数据和流量账单数据，其中通话账单数据包括：主叫号码、被叫号码、通话建立时间、通话持续时间和所在基站信息。流量账单数据包括：用户号码、联网建立时间、联网持续时间、下载流量、上传流量、所在基站信息。在数据预处理中排除了特殊号码、信息不完整、以及被中断的记录。其中被中断的记录是指用户使用呼叫等待等业务造成同一时间内有多个通话记录的情况。

2.2　数据特征

电信数据集中存在庞大的用户数量，其中通话账单数据中有40万用户，流量账单数据中有29.5万用户，其中用户有无通话记录或无流量记录。

用户数量随出话次数的分布如图1所示，可看出用户数量的分布与出话次数成反比。

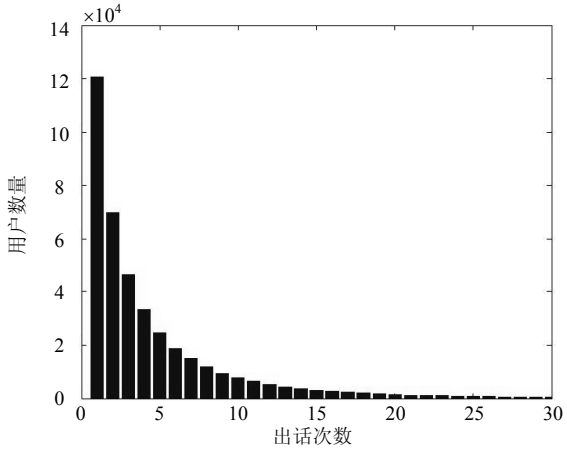


图1　用户数量随出话次数的分布

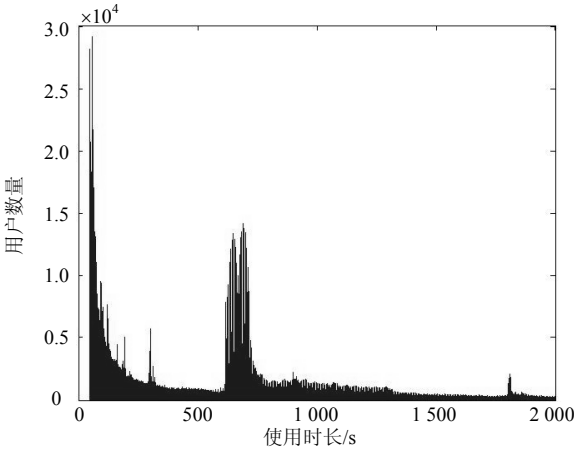


图2　95%累计百分比的上网时长频率分布

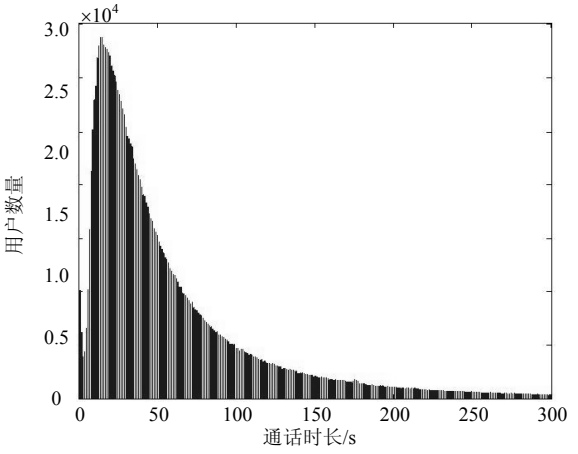


图3　95%累计百分比的通话时长频率分布

本数据集中每次通话和上网信息都以记录的形式保存，故同时对基于用户和记录的通话和上网数据特征进行统计，分别统计了通话数据和流量数据的95%累计百分比的时长频率分布，如图2和图3所

示。可以看出，两者的使用时长分布有很明显不同。图3可以看出明显的泊松分布，通话时长的众数出现在15 s；相反，图2的频率分布规律不太明显，上网时长的众数出现在59 s，同时在650～700 s处出现了使用频率的高峰。

2.3 用户偏好

为了量化用户通话和上网两种不同行为的模式和关系，定义 P_i 作为个体用户使用移动互联网的比重，用以归一化用户使用手机通话和移动互联网的时间，即用户的使用偏好。 x_i 和 y_i 分别代表示个体用户 i 的总上网时长和总通话时长， P_i 可表示为：

$$P_i = \left(nx_i \Big/ \sum_{i=1}^{n} x_i \right) \Big/ \left[\left(nx_i \Big/ \sum_{i=1}^{n} x_i \right) + \left(ny_i \Big/ \sum_{i=1}^{n} y_i \right) \right] \quad (1)$$

式中，n表示用户数量。P_i 值越高，代表用户更倾向使用移动互联网； P_i 值越低，代表用户更倾向使用手机通话。通过对用户的使用偏好进行统计，发现大量用户的使用偏好值P_i集中在0～10%和90～100%两个区间。超过86%的用户倾向于使用手机通话或使用移动互联网，且用户中倾向于使用手机通话的用户比例大于倾向使用移动互联网的用户比例。这一结论符合帕累托分布，即80/20法则。

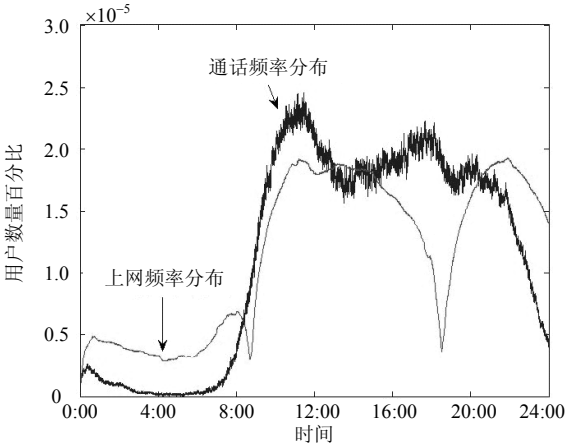
2.4 频率分布

有别于其他粗粒度的分析，在测量中X轴的精度通常精确到秒，这有利于观察细粒度下的用户行为模式和其他细微的变化，尤其是一些粗粒度下无法观测的短时间内的波动。通话频率分布与移动互联网频率分布如图4所示。

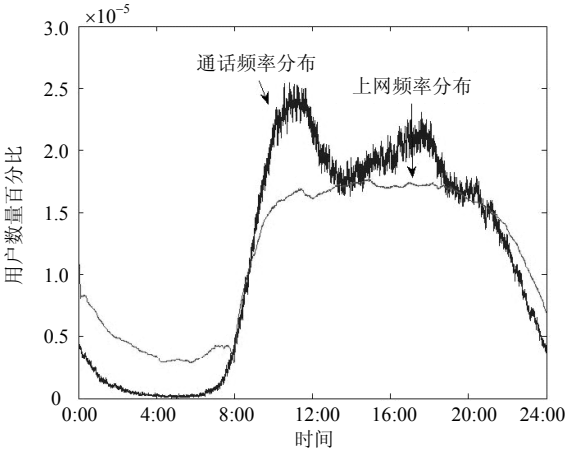
通过对通话数据频率分布的观察，可以很容易识别出通话频率分布的模式，类似于“典型人体生理节律”的双峰分布。在细粒度的分布图中，发现了一些差异(如通话频率分布在24:00−4:00的时间段中除了第一天都是递减的，其成因是由于统计的初始累加过程)。为了避免这种不确定的误差影响，所有24:00−4:00这个生理不活跃时期的数据都在之后的研究中被排除。

可以看出，通话数据的频率分布从4:00−11:00左右呈现迅速的增长直至频率的最高峰(第1天出现在10:28时；第2天出现在10:53时，第3天出现在11:00时；第4天出现在11:06时)。此后活跃用户的数量呈下降趋势直到14:00左右。在14:00−16:00时期内，工作日与周末的通话频率分布有明显不同。在这一时期周末的通话频率分布平稳降低，而工作日的通话频率分布则在上升。观察到工作日出现在18:00左右的第二次峰值在周末的频率分布图上并不明显(第1天出现在17:36时；第2天出现在17:28时，第3天出现在17:12时；第4天出现在17:05时)。在18:00−20:30这一时间段，4天的分布都呈现下降趋势，其中周末的下降趋势更加平缓。在20:30以后，4天的通话频率分布都迅速降低至一天中的最低点。

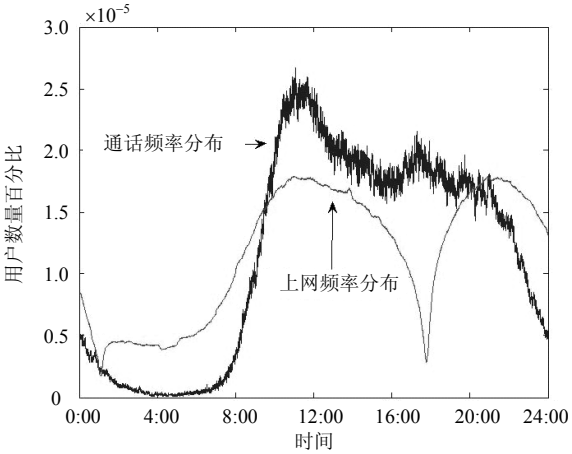
反观移动互联网的频率分布并不像通话频率分布那样有规律。最显著的差异是第1天、第3天、第4天的频率分布出现大幅度波动，而这种现象在粗粒度的频率分布图中并不明显。



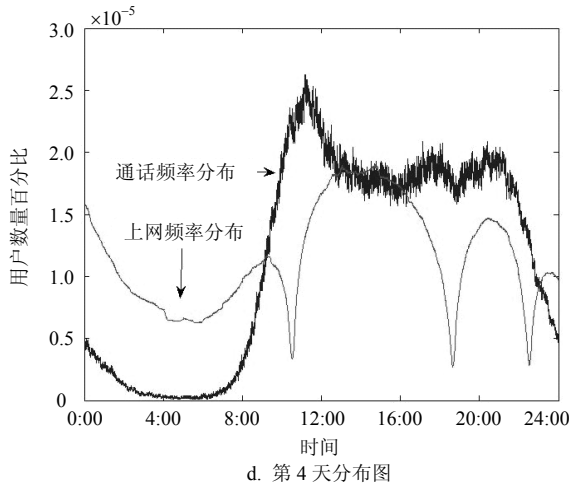
a. 第 1 天分布图



b. 第 2 天分布图



c. 第 3 天分布图



d. 第 4 天分布图

图4　通话数据频率分布与移动互联网数据频率分布图

3　特征分析

3.1　相关性分析

通过图4可发现工作日与周末的通话频率分布有显著差异。为了量化通话频率的不同，采用相关系数来评估通话频率分布的相似性。用相关系数R来描述样本x和y的相似性：

$$R(x,y)=\frac{C(x,y)}{\sqrt{C(x,x)C(y,y)}} \tag{2}$$

系数R的大小由协方差C决定：

$$C(x,y)=\frac{\sum_{i=1}^{n}\left(x_i-\frac{\sum_{i=1}^{n}x_i}{n}\right)\left(y_i-\frac{\sum_{i=1}^{n}y_i}{n}\right)}{n-1} \tag{3}$$

相关系数R越趋于1，两组样本越相似。分别计算每组通话频率分布的相关系数如表1所示。其中最高的R值分别为0.990 5与0.989 9，分别是第1天与第2天，第3天与第4天的相关系数，证明了工作日与周末的通话频率分布有明显差异，而这些差异是由用户行为的宏观差异所造成。

表1　通话频率分布的相关系数

	第1天	第2天	第3天	第4天
第1天	1	0.990 5	0.986 3	0.988 7
第2天	0.990 5	1	0.985 6	0.976 4
第3天	0.986 3	0.985 6	1	0.989 9
第4天	0.988 7	0.976 4	0.989 9	1

对于移动互联网的频率分布，同样分别计算每组分布的相关系数，如表2所示。可以看出，移动互联网频率分布的相关系数远低于通话频率分布。其最高的R值为第1天与第3天的相关系数0.859 2。且在通话频率分布图中观测到的工作日之间与周末之间的相似；工作日与周末不同的现象在移动互联网频率分布中并未发现。

表2　移动互联网频率分布的相关系数

	第1天	第2天	第3天	第4天
第1天	1	0.778 1	0.859 2	0.655 8
第2天	0.778 1	1	0.690 5	0.656 5
第3天	0.859 2	0.690 5	1	0.484 7
第4天	0.655 8	0.656 5	0.484 7	1

3.2　傅里叶拟合

从图5可以发现，所有通话数据统计点分布为一个双峰曲线，类似于“典型的人体生理节律”分布，此分布可以用5阶傅里叶函数很好的拟合，5阶傅里叶函数F_{Fourier}表示为：

$$F_{\text{Fourier}}=a_0+\sum_{i=1}^{5}\left[a_i\cos(iwx)+b_i\sin(iwx)\right] \tag{4}$$

式中，a_i和b_i代表振幅；w代表频率；a_0代表位移。

傅里叶函数通常是在信号处理中通过简单三角函数的叠加达到近似描述对象的目的。本文将通话频率分布和移动互联网频率分布近似的看作信号，即可使用傅里叶函数来描述用户通话的宏观现象。第1天为例，通话频率分布的拟合曲线如图5所示，拟合参数如表3所示。

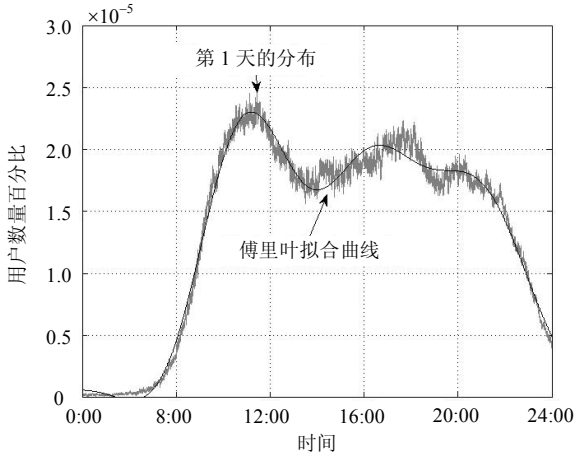


图5　第1天通话频率分布的傅里叶拟合图

式(4)中的a_0，a_i，b_i和w都是自由参数，表3中每行拟合值相似的参数都使用*号和+号标注出。可以看出第1天和第2天有7组相似参数(a_0, a_1, b_1, a_3, b_3, a_4和a_5)，第3天和第4天有5组相似参数(a_0, a_3, b_3, b_4和a_5)。除此之外，第2天和第3天有3组相似参数(b_2, b_5和w)，第1天和第4天有1组相似参数(b_2)。此外，未被标注的参数表示通话频率分布中最突出的特征(如a_0, a_2, b_5和w是第1天频率分布的显著特征)。

表3 通话频率分布的拟合参数和指标

参数	第1天	第2天	第3天	第4天
a_0	$1.15\times10^{-5*}$	$1.16\times10^{-5*}$	$1.16\times10^{-5*}$	$1.15\times10^{-5*}$
a_1	$-7.11\times10^{-6*}$	$-7.28\times10^{-6*}$	-6.83×10^{-6}	-5.95×10^{-6}
b_1	$-8.49\times10^{-6*}$	$-8.58\times10^{-6*}$	$-8.51\times10^{-6*}$	-8.81×10^{-6}
a_2	1.37×10^{-6}	1.84×10^{-6}	2.78×10^{-6}	2.45×10^{-6}
b_2	$-2.61\times10^{-6*}$	$-2.22\times10^{-6+}$	$-2.2\times10^{-6+}$	$-2.77\times10^{-6*}$
a_3	$-1.74\times10^{-6*}$	$-1.64\times10^{-6*}$	$-2.39\times10^{-6+}$	$-2.53\times10^{-6+}$
b_3	$1.84\times10^{-6*}$	$1.98\times10^{-6*}$	$1.04\times10^{-6+}$	$9.69\times10^{-7+}$
a_4	$-7.69\times10^{-7*}$	$-7.43\times10^{-8*}$	2.54×10^{-7}	-2.13×10^{-7}
b_4	$-1.77\times10^{-6*}$	-2.04×10^{-6}	$-1.54\times10^{-6*}$	$-1.63\times10^{-6*}$
a_5	$1.93\times10^{-8*}$	$1.29\times10^{-8*}$	$-1.49\times10^{-7+}$	$-2.49\times10^{-7+}$
b_5	8.99×10^{-7}	$6.59\times10^{-7*}$	$6.28\times10^{-7*}$	1.02×10^{-6}
w	7.12×10^{-5}	7.28×10^{-5}	7.27×10^{-5}	7.24×10^{-5}
SSE	4.43×10^{-8}	3.52×10^{-8}	4.58×10^{-8}	4.91×10^{-8}
R-Square	0.992 9	0.994 5	0.992 6	0.991 8
RMSE	7.16×10^{-7}	6.38×10^{-7}	7.28×10^{-7}	7.54×10^{-7}

虽然使用傅里叶函数进行曲线拟合可以用来近似的描述信号(细粒度的频率分布可以近似的看作信号)，但拟合参数所代表的客观意义很难定义，即这些参数在本文研究中只是用做测量的指标，而其具体的物理意义将在接下来的工作中使用信号分析的方法来进行研究和探讨。

3.3 多项式拟合

为了更进一步的研究细粒度下数据集的通话频率分布特征，本文选择了通话频率分布中两个重要的时间段分别进行研究，一个是8:00−10:00的频率分布增长时期，另一个是22:00−24:00的频率分布衰减时期。

在此时间段内，频率的变化呈现一种较均匀的线形增长或衰减趋势，故采用一次多项式函数对增长和衰减时期的通话频率进行拟合，一次多项式函数 $F_{\text{Polynomial}}$ 表示为：

$$F_{\text{Polynomial}} = kx + a \tag{5}$$

式中，参数k代表拟合曲线的斜率，在这里可以理解为宏观用户的活跃(衰减)速度。$|k|$值越高，代表宏观用户的活跃(衰减)速度越快，$|k|$值越低，代表宏观用户的活跃(衰减)越慢；a代表曲线位移。以第1天为例，其增长时期和衰减时期的拟合曲线如图6～图7所示，4天的拟合参数如表4和表5所示。

可以看出第1天、第2天、第3天、第4天的通话频率增长区域的k值有明显不同。第3天与第4天的k值远远小于第1天和第2天。即用户在工作日的活跃程度增长速度高于用户在周末的活跃程度增长速度。同时，衰减区域第1天和第4天的k值则远低于第2天和第4天，结果表明，当天用户的活跃程度取决于第2天是否为周末。即当第2天为周末，当天用户的活跃程度增长更慢，衰减也更慢；当第2天不为周末，当天用户的活跃程度增长更快，衰减也更快。

此外，增长区域中第3天的k值低于第4天；衰减区域中第3天的k值高于第4天，即用户在第3天的活跃程度较第4天上升更慢，衰弱也更慢。结果表明，用户在第3天(周六)较第4天(周日)更为“放松”。

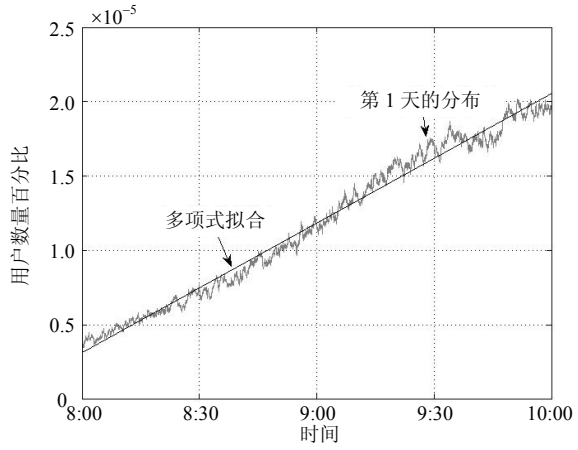


图6 第1天增长区域的多项式拟合图

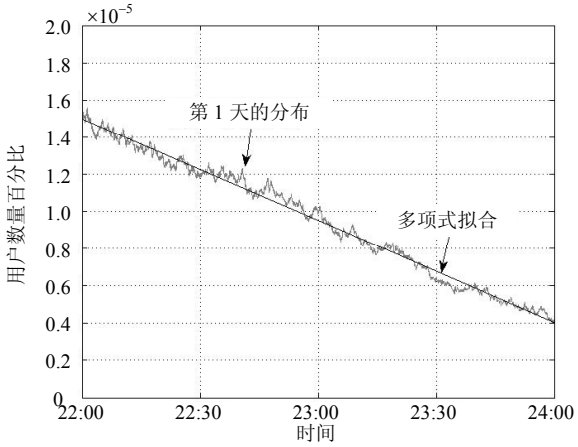


图7 第1天衰减区域的多项式拟合图

表4 拟合的参数和指标

参数	第1天	第2天	第3天	第4天
k	$8.797\times10^{-6*}$	$8.956\times10^{-6*}$	7.977×10^{-6}	8.24×10^{-6}
SSE	2.534×10^{-9}	1.655×10^{-9}	2.874×10^{-9}	2.442×10^{-9}
R-Square	0.986 3	0.991 5	0.981 5	0.985 2
RMSE	5.933×10^{-7}	4.795×10^{-7}	6.318×10^{-7}	5.824×10^{-7}

表5 拟合的参数和指标

参数	第1天	第2天	第3天	第4天
k	$-5.48\times10^{-6*}$	$-4.155\ 2\times10^{-6}$	-4.7×10^{-6}	$-5.376\times10^{-6*}$
SSE	7.288×10^{-10}	7.685×10^{-10}	1.515×10^{-9}	1.872×10^{-9}
R-Square	0.99	0.981 8	0.972 2	0.973 7
RMSE	3.182×10^{-7}	3.267×10^{-7}	4.588×10^{-7}	5.1×10^{-7}

4 结　论

通过测量与分析电信数据中用户行为的特征，定义了用户的使用偏好，发现了用户使用手机通话和使用移动互联网的倾向；通过比较频率分布的相关系数与拟合参数，发现手机通话用户在工作日和周末的活跃程度不尽相同，而且周六用户的活跃程度与周日的也有明显差异。

参 考 文 献

[1] CANDIA J, GONZALAZ M C, WANG P, et al. Uncovering individual and collective human dynamics from mobile phone records[J]. Journal of Physics A: Mathematical and Theoretical, 2008, 41(22): 1-15.

[2] JO H H, KARSAI M, KERTESZ K, et al. Circadian pattern and burstiness in mobile phone communication[J]. New Journal of Physics, 2012, 14(1): 20-37.

[3] ONNELA J P, SARAMAKI J, HYVONEN J, et al. Analysis of a large-scale weighted network of one-to-one human communication[J]. New Journal of Physics, 2007, 9(6): 179-201.

[4] MOTAHARI S, ZANG H, REUTHER P. The impact of temporal factors on mobility patterns[C]//45th International Conference on System Science(HICSS). Hawaii: IEEE, 2012.

[5] OLMEDILLA D, FRIAS-MARTINEZ E, LARA R. Mobile web profiling: a study of off-portal surfing habits of mobile users[C]//Proceedings of the 18th International Conference on UMAP. Big Island, USA: Springer Berlin Heidelberg, 2010: 339-350.

[6] DUGGAN M, SMITH A. Cell internet use 2013[EB/OL]. [2014-01-01]. http://www.pewinternet.org/2013/09/16/cell-internet-use-2013/.

[7] TAYLOR C A, ANICELLO O, SOMOHANO S, et al. A framework for understanding mobile internet motivations and behaviors[M]. New York: ACM, 2008.

[8] GHOSE A, HAN S P. An empirical analysis of user content generation and usage behavior on the mobile internet[J]. Management Science, 2011, 57(9): 1671-1691.

[9] HSU S L, DOONG H S, WANG H. Exploring diffusion patterns of 3G wireless Internet service adoption[C]//2nd International Conference on Computer Engineering and Technology(ICCET). Assisi-Perugia: IEEE, 2010.

[10] WANG C. Surfing mobile internet motivated by fashion attentiveness: an empirical study of mobile internet use in China[C]//8th Asia-Pacific Regional ITS Conference. Taipei, China: [s.n.]: 2011.

[11] PURCELL K, SMITH A, ZICKUHR K. Social media & mobile internet use among teens and young adults[M]. Washington, USA: Pew Internet & American Life Project, 2010.

[12] ISHII K. Internet use via mobile phone in Japan[J]. Telecommunications Policy, 2004, 28(1): 43-58.

[13] HALVEY M, KEANE M T, SMYTH B. Predicting navigation patterns on the mobile-internet using time of the week[C]//Special interest tracks and posters of the 14th international conference on World Wide Web. New York: ACM, 2005.

[14] DE J E, VAN P M, ROOS M. Time patterns, geospatial clustering and mobility statistics based on mobile phone network data[C]//Federal Committee on Statistical Methodology research conference. Washington, USA: Statistics Netherlands, 2012.

[15] HALVEY M, KEANE M T, SMYTH B. Mobile web surfing is the same as web surfing[J]. Communications of the ACM, 2006, 49(3): 76-81.

[16] JIANG Z Q, XIE W J, LI M X, et al. Calling patterns in human communication dynamics[J]. Proceedings of the National Academy of Sciences, 2013, 110(5): 1600-1605.

[17] CHUNG J Y, CHOI Y, PARK B, et al. Measurement analysis of mobile traffic in enterprise networks[C]// Network Operations and Management Symposium (APNOMS), 2011 13th Asia-Pacific. Taipei: China, IEEE, 2011: 1-4.

[18] VERKASALO H. Contextual patterns in mobile service usage[J]. Personal and Ubiquitous Computing, 2009, 13(5): 331-342.

[19] HALVEY M, KEANE M T, SMYTH B. Time based patterns in mobile-internet surfing[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Delft: ACM, 2006: 31-34.

编　辑　叶　芳