文章编号:1007-6735(2012)01-0001-05

在线人类行为动力学中的肥尾特征

王 澎[1,2], 汪秉宏[3,4]

(1.杭州师范大学 信息经济研究所,杭州 310036; 2.杭州师范大学 阿里巴巴商学院,杭州 310036;
3.中国科学技术大学 近代物理系,合肥 230026; 4.上海理工大学 复杂系统科学研究中心,上海 200093)

摘要:回顾了最近有关在线人类行为动力学重要的实证结果:在线行为的时间间隔分布的肥尾特征.通过博客发文与维基修改记录等在线行为数据的实证结果,阐述了时间间隔分布的幂指数随着活跃性减小而递减的关系;进一步强调了在不同时间尺度上人类行为的时间间隔分布表现出来的异质性;最后总结和展望了这些特性对于未来研究在线群体演化的重要意义.

关键词:在线用户行为;人类动力学;时间间隔分布;肥尾
中图分类号:N 94 **文献标志码**:A

The Heavy-tails in On-line Human Dynamic

WANG Peng[1,2], WANG Bing-hong[3,4]

(1. *Insititute of Information Economy, Hangzhou Normal University, Hangzhou 310036, China*;
2. *Alibaba Business College, Hangzhou Normal University, Hangzhou 310036, China*;
3. *Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China*;
4. *The Research Center for Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: The main result arising from recent studies on on-line human dynamics was reviewed. It concluded that the human activity has the heavy-tailed nature. Based on the empirical evidence from Blog-posting and Wiki-revising, it could be found that the distributions of the interevent time τ decay powerlike as τ increase at both individual and population levels. Furthermore, time scales and obtain heterogeneous decay exponents in the intra- and inter-day range were pointed out. In the end, the significance of these features for the further research on the evolution of its group was emphasized.

Key words: *online user behavior*; *human dynamic*; *inter-event time*; *heavy-tails*

收稿日期:2012-01-21
基金项目:国家重点基础研究发展计划资助项目(2006CB705500);国家自然科学基金重大研究计划资助项目(91024026);国家自然科学基金资助项目(10975126,10635040);高校博士点基金资助项目(20093402110032);浙江省自然科学基金资助项目(Y6110317)
作者简介:王 澎(1981-),男,讲师.研究方向:人类动力学、复杂系统.E-mail:wangpenge@gmail.com
汪秉宏(联系人),男,教授.研究方向:统计物理、非线性科学、复杂系统理论.E-mail:bhwang@ustc.edu.cn

1 互联网与海量人类行为数据

对人的行为模式和特征的研究有着重要的社会和经济价值.人是构成庞大而复杂社会的基本单元,因此,当我们希望了解和模拟任何这个群体中发生的现象和过程的时候,有关人这个基本单元本身的动态特性的知识则是永远无法回避的.然而,人本身又是极其复杂的系统.早期的金融物理、经济和社会学建模,还有流行病模型中,作为过程执行者的人往往只是被简单的假设为服从某种泊松行为.在社会学和心理学方面,对人的研究已经有悠久的历史和一套完整的方法,如通过问卷调查抽样了解样本人群的特性,以及实验心理学中通过对志愿者的实验来了解个体行为中的心理细节.虽然这些传统的研究成果从某种程度上反映了个人的某些行为特征,但是一方面其数据仍然太少,很难对人的行为给出一个比较全面和普遍性的特征描述;另一方面是其结果过于细节,对于由人构成的复杂系统的建模研究似乎可以利用之处甚少.

然而计算机的出现,信息技术的发展,则让我们看到了实现的希望.首先是个人电脑的出现.人们通过电脑的每一个操作理论上都是可以记录和收集的,如鼠标的移动和点击、链接的访问、文件的修改、命令的输入都能够被准确地记录下来,成为很好的研究对象.同时电脑的大量普及,互联网的诞生则导致了大量信息的数字化.Web2.0 网站的出现则进一步推动了这个过程.不同于传统的网站信息的单向发布模式,Web2.0 网站上的内容通常都是用户自己发布的.也正因为有了用户的大量参与,使得这些网站记录了丰富的用户行为数据,比如说发文信息、回帖信息、商品买卖、用户登录、照片上传等.通过对这些数据的研究,可以极大地丰富有关人类动力学行为的知识.

2 从静态分析到动态分析

这些丰富的数据吸引了从物理学、计算机学到社会学、心理学等广阔学科领域的学者.早期,对于这些数据的分析,学者们大量使用了传统的数理统计方法,并且在此基础上慢慢进化为一门专门的学科——“数据挖掘”[1].同时近 10 年来,图论作为另一种数据分析方法被大量应用,并让“复杂网络”成为一门新兴的交叉学科飞速发展起来[2-5].这种方法的核心就是把数据里不同个体看成节点,而个体之间的相互作用数据看成连边,这样形成了一个网络的图景,并利用传统图论中的一些特征量来研究这个“网络”.在研究在线社区的时候,社区里的每个人的主页往往被看成一个节点,而主页上的好友链接则看成连边[6].研究手机用户行为时,每个手机用户则被看成节点,而彼此的通话则看成连边[7].不过这里网络的连边数往往是各节点相互作用的累积量,这样的处理本质上把动态的网络静态化了,这主要是因为图论中的各种测量方法都只针对静态网络,有关动态网络的刻画实际上没有现成的理论.比如,在线社会关系网中友情链接的建立其实是一个动态过程,用户往往逐渐增加一些好友,有时还会删除一些许久不联系的好友,但是由于文献中往往不考虑这个动态过程,直接以现有好友数作为节点的度,这其实只是反映了一个经历了长期演化后的累积量.早期,Holme 就注意到了这一点,通过研究在线用户的 email 通信间隔,他发现这里的时间间隔分布呈幂指数分布[8].这种分布说明在线用户之间的交流存在很强的突发性.

3 人类动力学的实证

3.1 时间间隔分布的肥尾特征

不过真正对时间间隔(两次连续行为之间的间隔)分布的肥尾特征的普遍证实和关注则是之后 Barabási 等的集大成的工作[9-10].图 1 显示了指数和幂率分布的间隔分布的不同.τ 表示等待时间间隔,$p(\tau)$为其出现的概率.其中,图(a),(b),(c)分别为以泊松序列的时序和分布图,图(d),(e),(f)分别为以幂律序列的时序和分布图[9].可以看到幂指数分布存在着非常多长间隔,这是指数分布中不可能有的,这对应着的行为特点就是人往往会在经历了很长时间的暂停后又会突然密集地从事某事,因此又称这样的行为特征为突发性(burstiness).他们研究了 email、通信、网站访问、图书借阅、股票交易等 5 种行为,并在这些行为里都发现了呈幂律的时间间隔(或等待时间)分布[10].结合早前零星的证据,显示出这种特征是人类行为中的普遍规律.

在很多商业和社会实际问题中,如交通流模型、交通事故发生频率、呼叫中心的呼叫、存货控制问题等,学者们往往都是假设人的行为间隔分布为泊松分布.因此,Barabási 以及之前的研究完全打破了这个传统观点.尔后的研究,包括打印间隔[11]、短信收发间隔[12]、手机通信间隔[13]、在线游戏登录[14-15],

拍卖等待[16]、论坛回复[17]、网上冲浪[18]等都进一步证实了人类行为突发性这一基本特征，这意味着之前那些问题的研究需要重新加以考量.

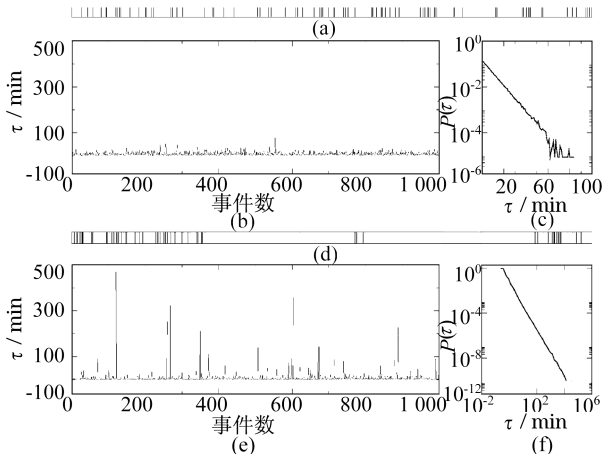

图 1　泊松序列和肥尾序列的不同[9]

Fig.1　The difference between the poisson activity patterns and the one in human dynamics[9]

3.2　变化的幂指数

更进一步的研究显示，用户的分布幂指数实际上和其活跃性（个人用户的平均间隔时间的倒数）呈正比[19]，这一结论最早来自周涛等的研究. 他们首先根据活跃性对不同的用户进行分类，每一类用户产生一个集体间隔分布，通过拟合这个幂指数发现，分布的幂指数与活跃性呈现单调的递增关系. 很快，Filippo 在 AOL 在线查询、Ebay 信息发送、维基用户登录中证实了同样的相关性[20]. 与周涛等的不同之处在于这里是根据行为数量来进行分类的（见图 2）. 图 2(a)表示 Ebay 留言间隔的全局时间间隔分布，这里的每幅小图对应于一组有相似留言数的用户. 全部用户被分为 11 组，这里分别显示了其中的 3,5,9,11 组. 虚线为其幂律拟合，对应的幂指数分别为 1.1,1.2,1.8,2.3. 图 2(b)为不同组里能被以上全局分布很好描述的比例，这里的比例值 $R(Q)$ 是通过 Kolmogorov-Smirnov 测试计算出来的，Q 是对分布幂指数的偏差. 图 2(c)表示不同组的 $R(Q)$ 值（$Q=0.5$）以及幂指数与行为数 n 的关系[20]. 不过，由于数量和活跃性是正比关系，因此得到的结果和周涛等本质上是没有区别的. 从图 2 可以看到，当用小时为单位记录间隔时，在天以外有非常强的震荡，这实际上对拟合以及分布规律的展现有一定影响. 另外，Filippo 通过重新标度的方法（取 $\tau/\langle\tau\rangle$，其中 τ 为间隔分布，$\langle\tau\rangle$为其平均量），发现在相对尺度下，原来不同幂指数的分布都塌缩到一起[20]. 因此，他强调利用相对时间间隔来描述人类行为会更加合适. 并且这一发现也说明不同活跃性的用户虽然间隔分布幂指数不同，但是其内在机制很可能仍然是一致的. 不过一个有意思的地方是，他用同样的分析方法来研究 Ebay 的留言等待时间，却发现不同活跃性用户的等待时间间隔分布幂指数却没有明显的变化.

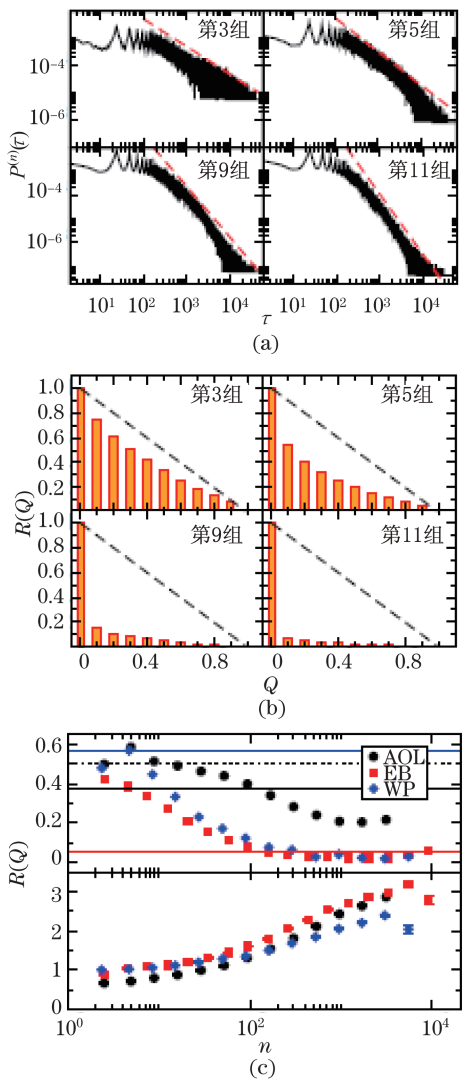

图 2　分布幂指数与活跃性的正比关系

Fig.2　Exponents of interevent time distributions of users decreases as n increases[20]

3.3　不同时间尺度下的幂指数

自从 Barabási 等首次提出人类动力学这个概念，并发现肥尾特征在人类行为间隔分布中的普遍性以来，几乎所有的模型都似乎默认了一个关键的假设，那就是驱动人类行为的机制在所有时间尺度上都是一致的[21-27]. 然而，如果仔细分析实证结论，却能得到一个非常不同的结果.

表 1（见下页）罗列了最近有关人类行为的实证

结果，主要包括其间隔分布指数以及其所用的单位和分布所在的主要时间范围.这里特别注明了每个数据的单位以及分布所在的范围. * 对应于所有个人幂指数的平均值；Δ 对应于个人分布的幂指数；其它的则为全局分布的幂指数.表中简单地把时间范围分为两个区域：天以内和天以外.如表所见，当数据的单位为秒或者分钟时，研究往往只是集中在天以内的部分；当数据的单位为小时或者天的时候，呈现的间隔分布又往往是天以外的部分.没有任何研究同时关注过天以内和天以外的行为，尽管有人也注意到在 24 h 左右的地方分布会有轻微的隆起[28].一个被广泛研究的例子就是电子邮件和通信记录.一些人根据这两个行为的间隔分布的指数不同，认为这两种行为分别属于不同的普遍类[10]；而另外一些人则认为这两个行为的内在机制是相同的，因为通过尺度缩放(rescaling)，他们把一些看起来幂指数完全不同的分布塌缩到了一条分布上[20].但是，没有人注意到这两个行为的分布实际上是处于两个迥异的时间范围.得益于电脑的自动记录，电子邮件的收发时间都是以秒或分钟为单位，而通信的时间记录习惯往往只是精确到天，这导致之前的研究对电子邮件只是注意天以内收发间隔分布，而通信则是天以外.仔细比较表 1 中行为的间隔分布，可以看到天以外分布的幂指数普遍要高于天以内分布的幂指数值.5 个天以外的幂指数中有 4 个约等于或者大于 2；而所有 6 个天以内的幂指数都只是等于 1 或者稍微高于 1.

表 1　不同人类行为的分布幂指数比较

Tab.1　Comparison of the exponents from different human activities

行为种类	单位	范围	幂指数
电子邮件[9,10,25]	s	天以内	1*,0.9
通信[25,29]	d	天以外	2.37Δ,2.1Δ
图书借阅[10]	min	天以内	1.0*
打印[30]	s	天以内	1.3Δ
路由访问[10]	s	天以内	1.0*
点击同一链接[29]	s	天以内	1.0
点击任何链接[29]	s	天以内	1.25
AOL 在线查询[20]	h	天以外	1.9
Ebay 信息[20]	h	天以外	1.9
维基登录[20]	h	天以外	1.2
电影评价[19]	d	天以外	2.08

如果对于不同行为不同时间尺度下的分布比较还并不能完全说明这种异质性，进一步地，同一行为不同时间尺度下的分布幂率更完整地证实了这个结论[31-32].通过对博客发帖和维基词条修改两种行为的调查，无论是在集体层面还是在个人层面，都显示了分布幂指数在两个时间尺度上的不同(见图 3).图 3 中 N 为间隔的累积数，图 3(a)，(b)中的用户来自维基，图 3(c)，(d) 中的用户来自博客，其对应的幂指数分别为：$\beta_{\min}\approx 0.38$，$\beta_{\text{hour}}\approx 0.11$，$\beta_{\text{day}}\approx 1.23$(图(a))；$\beta_{\text{hour}}\approx 0.19$，$\beta_{\text{day}}\approx 1.57$(图(b))；$\beta\approx 1.22$(图(c))；$\beta\approx 1.13$(图(d))[32].这种不同不仅反映在分布幂指数上，同时也反映在相关性、活跃性以及与幂指数的依赖关系上[31].

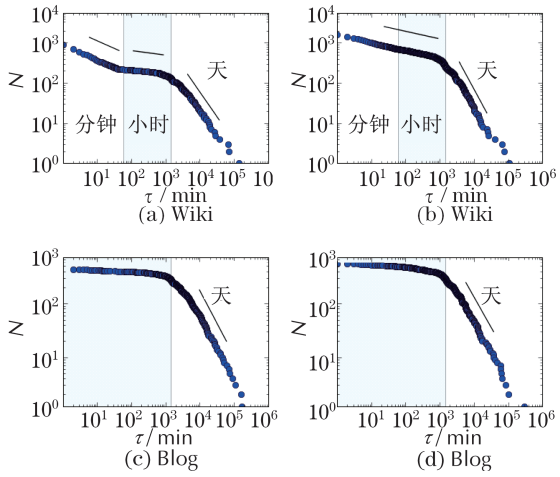

图 3　个人行为的时间间隔累积分布图[32]

Fig.3　The cumulative distribution of interevent times of individuals from Wiki and Blog[32]

4　问题与讨论

由于该领域还处于发展初期，因此，目前仍然存在着大量问题有待研究.第一，肥尾特征的普遍性.这里的普遍性不仅指对于不同人类行为是否都存在间隔分布的肥尾特性，更指这个特性在不同的时间尺度上的差异，以及是否在个体和群体层次上一致.第二，仅使用时间间隔(等待间隔)分布来描述和区分人类在线行为是远远不够的，对人类行为的其它特性的调研才有可能对人类行为进行更完整的刻画.这方面的研究才刚刚开始，比如用相关性系数来描述行为间隔序列[33].第三，行为的区分和模型的鉴别.现在绝大部分人类动力学模型只是用来解释肥尾这一特征，很少涉及到更多的特征.

更重要的一个方向和可能的应用是研究个体行为在在线群体演化中的角色.尽管几乎所有的在线社会网络演化模型都会假设个体行为规则,然后通过这个规则来拟合相应的群体结构特征(比如度分布)[3,33-34].然而,很少有研究会同时考虑到这样的个体行为规则是否符合真实的个人行为特征的实证结果,真正合理的演化模型是要必须同时考虑这两个方面的.

参考文献:

[1] Fayyad U M, Gregory P S, Padhraic S, et al. Advances in knowledge discovery and data mining[M]. Cambridge: MIT Press, 1996.

[2] Albert R, Barabási A L. Statistical mechanics of complex networks[J]. Rev Mod Phys, 2002, 74(1): 47-97.

[3] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: structure and dynamics[J]. Phys Rep, 2006, 424(4/5): 175-308.

[4] Watts D J, Strogatz S H. Collective dynamics of small-world networks[J]. Nature, 1998, 393(6684): 440-442.

[5] Barabási A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[6] Kumar R, Novak J, Tomkins A. Structure and evolution of online social networks[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 611-617.

[7] Wang P, González M C, Hidalgo C A, et al. Understanding the spreading patterns of mobile phone viruses[J]. Science, 2009, 324(5930): 1071-1076.

[8] Holme P. Network dynamics of ongoing social relationships[J]. Europhysics Letters, 2003, 64(3): 427.

[9] Barabási A L, The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039): 207-211.

[10] Vázquez A, Oliveira J G, Dezsö Z, et al. Modeling bursts and heavy tails in human dynamics[J]. Phys Rev E, 2006, 73(3): 036127.

[11] Harder U, Paczuski M. Correlated dynamics in human printing behavior[J]. Physica A, 2006, 361(1): 329-336.

[12] Hong W, Han X P, Zhou T, et al. Heavy-tailed statistics in short-message communication[J]. Chinese Physics Letters, 2009, 26(2): 028902.

[13] Candia J, González M C, Wang P, et al. Uncovering individual and collective human dynamics from mobile phone records[J]. Journal of Physics A, 2008, 41(22): 224015.

[14] Grabowski A, Kruszewska N, Kosiński R A. Dynamic phenomena and human activity in an artificial society[J]. Phys Rev E, 2008, 78(6): 066110.

[15] Jiang Z Q, Zhou W X, Tan Q Z. Online-offline activities and game-playing behaviors of avatars in a massive multiplayer online role-playing game[J]. Europhysics Letters, 2009, 88(4): 48007.

[16] Scalas E, Kaizoji T, Kirchler M, et al. Waiting times between orders and trades in double-auction markets[J]. Physica A, 2006, 366: 463-471.

[17] Yu J F, Hu Y Q, Yu M, et al. Analyzing netizens' view and reply behaviors on the forum[J]. Physica A, 2010, 389(16): 3267-3273.

[18] Dezsö Z, Almaas E, Lukács A, et al. Dynamics of information access on the web[J]. Phys Rev E, 2006, 73(6): 066132.

[19] Zhou T, Kiet H A T, Kim B J, et al. Role of activity in human dynamics[J]. Europhysics Letters, 2008, 82(2): 28002.

[20] Radicchi F. Human activity in the web[J]. Phys Rev E, 2009, 80(2): 026118.

[21] Malmgren R D, Stouffer D B, Motter A E, et al. A poissonian explanation for heavy tails in e-mail communication[J]. PNAS, 2008, 105(47): 18153-18158.

[22] Malmgren R D, Stouffer D B, Campanharo A L O, et al. On universality in human correspondence activity[J]. Science, 2009, 325(5948): 1696-1700.

[23] Han X P, Zhou T, Wang B H. Modeling human dynamics with adaptive interest[J]. New J Phys, 2008, 10(7): 073010.

[24] Oliveira J G, Vázquez A. Impact of interactions on human dynamics[J]. Physica A, 2009, 388(2/3): 187-192.

[25] Vázquez A. Impact of memory on human dynamics[J]. Physica A, 2007, 373: 747-752.

[26] Shang M S, Chen G X, Dai S X, et al. Interest-driven model for human dynamics[J]. Chin Phys Lett, 2010, 27(4): 048701.

[27] Wang P, Zhou T, Han X P, et al. Modeling correlated human dynamics[DB/OL]. [2010-08-03]. http://arxiv.org/abs/1007.4440.

[28] Baek S K, Kim T Y, Kim B J. Testing a priority-based queue model with Linux command histories[J]. Physica A, 2008, 387(14): 3660-3668.

(下转第 17 页)