

微博用户的时间行为研究

张杰斌 秦 红

(北京航空航天大学电子信息工程学院 北京 100191)

摘 要 人类日常行为活动在生活的各个方面普遍存在。个体行为活动类型多样且个体之间行为差异明显,人类行为显示出高度复杂性。利用微博用户发布微博所产生的时间数据来研究用户的时间行为模式,研究发现个体用户时间间隔的统计特征主要服从幂律、指数和双模 3 种分布函数;并提出了基于任务队列的个体用户行为动力学模型,解释了用户发布微博的时间间隔分布特征。

关键词 人类行为,时间统计特性,微博,行为模式

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.004

Research on Temporal Behavior of Microblog Users

ZHANG Jie-bin QIN Hong

(School of Electronic and Information Engineering, Beihang University, Beijing 100191, China)

Abstract Human daily activities are widespread in all aspects of life. Human behavior has the characteristics of high complexity because of its various types of individual activities and individual differences. Using microblog users' microblogging time data, we studied the time behavior of users. The results show that the statistical characteristics of time interval for individual users mainly follow three distributions: power-law, exponential and bimodal distribution. Furthermore, we proposed an individual behavior dynamic model based on task queue to explain the characteristics of time interval for the microblog users.

Keywords Human behaviors, Time statistical properties, Microblogs, Behavior patterns

1 引言

人类的行为推动着社会的发展,构成了错综复杂的社会网络关系。在以往的研究中,人们往往假设人的行为在时间尺度上是随机分布的,即假设人一直以一个不变的速率 λ 执行每项任务,在一个时间区间 τ 内,事件在 $[t, t+\tau)$ 期间内发生的概率与时间 t 的取值无关,而只与时间区间的大小 τ 线性相关,时间区间范围越大,事件发生的概率越大。相邻发生的两个事件间的时间间隔分布较为均匀,服从指数分布 $p(\tau) = \lambda e^{-\lambda\tau}$,可以用泊松过程进行描述。近些年来,随着计算机和通信技术的发展,研究人员能够获取并处理人类行为在时间尺度上的海量数据。基于这些海量数据的实证研究发现,人类行为既有短时间内高频率爆发、又有长时间保持静默的特性,明显偏离了泊松分布,显示出非泊松的特性。

在传统通信模式中,Barabasi 研究了电子邮件用户发送时间、回复时间及邮件大小等数据内容^[1],结果表明,同一用户发送的两封相邻电子邮件的时间间隔分布以及邮件从接收到回复之间的等待时间间隔分布均服从幂指数为 1 的幂律分布。通过研究爱因斯坦和达尔文两位科学家的普通信件数据^[2],发现信件等待时间间隔分布也服从幂律分布,其幂指数为 1.5。还有研究运用中国运营商用户的通信语音数据^[3],提取并计算得到了每个用户的去话时间间隔,发现将全部用

户的去话时间间隔看成一个整体,其时间间隔分布服从幂律分布,尾部是一个类似指数的截断形式。利用 Kolmogorov-Smirnov (KS) 统计分布检验的方法,发现其尾部的截断更接近于韦伯分布,而非指数衰减。而以手机用户发送短信息产生的时间数据为研究对象^[4],研究者提取了一个月内收发短信多于 5 条以上的用户,将这些人群看成活跃用户。研究发现,对于一个用户而言,发送两条连续短信的时间间隔分布不服从幂律分布,而是用一个包含幂律和指数的分段函数描述。在日常活动的行为数据研究方面,有学者研究了图书馆借阅的时间数据和银行股票的交易活动行为数据^[5],发现其时间间隔分布是带有指数截断的幂律分布。而在人类网络行为研究方面,许多学者利用网络论坛、在线点播、即时通信软件和微博社交网络等在线平台^[6-10]研究了用户在网络上的时间行为特征,结果表明用户的时间行为均表现出了幂律分布的非泊松特点。

在传统通信和日常活动中,人类在时间尺度上的行为研究已渐趋成熟,人类的时间行为分布不仅有幂律分布,还有其他形式的非泊松分布,行为特征模式具有多样性。而在网络中,目前相关的用户行为模式的研究结果还主要集中于幂律分布。那么,网络中的行为模式是否也应和在传统通信和日常活动中一样具有非泊松分布多样性的特征?这个问题值得研究者进行进一步的研究,以深入挖掘人类在网络中的行为

到稿日期:2015-05-15 返修日期:2015-10-09

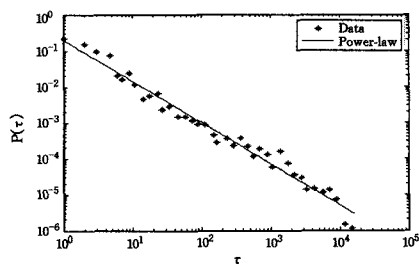
张杰斌(1986—),男,硕士,助理实验师,主要研究方向为通信与测控、复杂网络,E-mail:zjb@buaa.edu.cn;秦红(1989—),女,硕士,主要研究方向为复杂网络。

模式特征。目前,微博已经成为了人类在线社交网络沟通交流的常用工具,而新浪微博作为国内重要的微博产品,拥有数以亿计的用户群体。对于微博平台上如此庞大的用户群体,其行为模式会呈现怎样的规律?微博用户的在线网络行为特征是否也与传统通信或者日常行为等相似,具有多样的非泊松分布的行为特点?这些问题值得研究者进行进一步的深入研究。本文将利用微博用户发布微博产生的时间数据研究用户在时间上的行为模式。

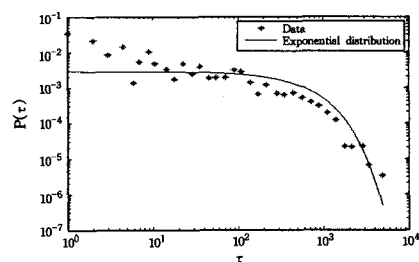
2 个体用户行为实证特征

本文采用任意抽样的方法,随机采集了 10000 名微博用户的基本信息和 2013 年 6 月至 12 月半年间发布微博的时间数据。根据每个用户半年发布微博的时间数据,可以用 Matlab 画出每个用户的微博发布的时间间隔概率密度曲线。研究发现,并不是所有用户的时间间隔分布均服从幂律分布。通过观察,一些用户的时间间隔分布比较符合指数分布,还有一些用户的时间间隔分布虽然在前一段服从幂律,但是存在一个分界点,分界点左侧的衰减速度明显快于右侧,其分布的尾部衰减缓慢,不能够用幂律分布函数完全刻画整个分布,而可以用幂律和指数的分段函数进行描述,在此将其简称为双模分布。通过每个函数拟合优度的结果判断其时间分布最终服从哪种函数^[11]。

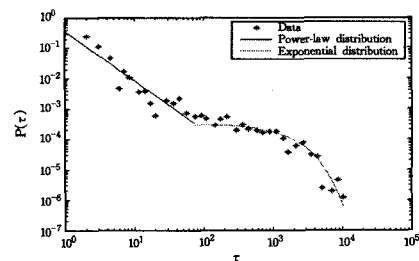
最终,得到有 6347 位用户的时间间隔服从幂律分布,151 位用户的时间间隔服从指数分布,还有 2486 位用户的时间间隔服从双模的分段函数。对不同行为特征的用户,其时间间隔分布如图 1 所示。



(a) 幂律分布



(b) 指数分布



(c) 双模分布

图 1 个体用户的时间间隔分布图

由图 1(a)知,幂律用户的时间间隔范围跨越较大,通常能够达到 3 至 4 个数量级。服从幂律分布的概率密度函数的尾部的衰减速度缓慢,说明数值较大的时间间隔数量虽然不多,但是其概率密度值不可忽略。最终,由拟合优度结果可知,6347 位用户的幂指数 γ 值分布在 0.6~2 之间,个体之间的幂律分布指数存在较大的差异。由图 1(b)可知,指数用户的时间间隔分布主要集中在 $10^0 \sim 10^2$ 的范围内,其间,时间间隔的概率密度衰减速度小。然而,超过 100 分钟的时间间隔数量少,概率密度函数尾部呈现迅速衰减的特点。最终,由拟合优度结果可知,151 位用户的指数 β 值分布在 $4.6 \times 10^{-5} \sim 1.0 \times 10^{-2}$ 的范围之间,个体之间的指数分布系数也存在较大的差异。由图 1(c)可知,双模用户的时间分布呈现分段状态,开始时间间隔的概率密度函数服从幂律分布,然而到达时间间隔中间值阶段,其衰减速度呈现出先缓慢衰减后又迅速衰减的特点。分段函数幂律系数 γ 的范围在 0.4~5.4 之间,指数 β 值的范围为 $2.3 \times 10^{-5} \sim 1.6 \times 10^{-2}$ 。

该研究结果与之前研究得到的人类在网络中的行为模式服从幂律分布的结果存在明显的差异。不同分布在大小时间间隔上的差异说明,在时间分布的背后,不同用户具有不同的行为模式特征。

3 个体用户行为动力学模型

为了解释人类行为的时间特性,研究人员将人类要进行的各种行为看成是需要完成的任务,提出了任务队列模型^[1]。在该模型中,每个用户被分配了一个任务列表。在每个单位时间步长内,用户选择执行其中的一个任务。完成该任务后,将其从任务列表中移除,同时将一个新的任务加入队列。上述模型可以较好地解释任务从进入队列到被执行的等待时间间隔分布。但是,由于在该模型中所有任务都被假设为同一个类型,如都是发送电子邮件任务或都是发布微博任务,会导致执行前后两个任务的时间间隔都等于单位时间步长,无法刻画本文发现的微博行为时间间隔分布特性。

为了解释用户发布微博的时间间隔分布,本文提出了区分任务类型的队列模型。队列中的任务类型有两种,一种是用户发布微博的行为,另一种是除了发布微博之外的其它行为。模型主要包括以下 3 部分。

队列中每个任务都有一个优先级参数 x 。此参数可以代表任务在队列中的重要程度,由给定的均匀分布函数 $\rho(x)$ 在 $[0,1]$ 之间随机生成。

以分钟为单位时间步长,在每个时间步长,按最高优先级执行任务。在实际生活中,任务被执行所用的时间可能多种多样,但为了简化研究,假设所有任务的执行时间均相同, $t_p = 1$ 。

任务执行完毕,会被移出队列,同时会有新的任务加入到队列中。进入队列的新任务是发布微博行为的概率为 v ,是其他行为的概率为 $1-v$ 。新加入队列的任务也会被随机分配一个优先级参数。

假设队列的执行总时间为 T ,这里取 $T=10^5$ 。任务队列长度用 L 表示。新任务是微博行为的概率 v 是 $[0,1]$ 之间的一个随机数。根据 v 取值的不同,会形成不同的时间间隔序列。研究中, v 的取值范围是 $[10^{-5}, 10^0]$,将不同 v 值对应产生的时间间隔分布分别用幂律、指数和双模 3 种函数形式进行拟合,其拟合优度参数 R^2 与 v 的关系如图 2 所示。

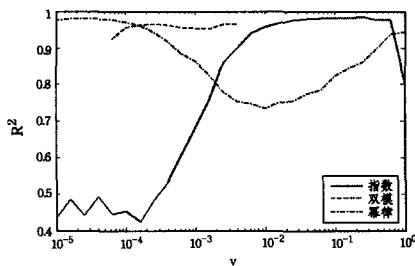


图2 v 与 R^2 的关系

由图2可知,当微博任务进入队列的概率较小时,模型产生的时间间隔分布最服从幂律分布;随着微博任务进入队列的概率增大,其时间间隔的分布形式更加符合双模分段分布;当微博任务加入队列的概率接近于1时,其时间间隔分布形式用指数函数刻画更为合适。图中双模分布的拟合优度仅在 v 的中间范围有值的原因是,按其他 v 值所产生的时间间隔分布不满足双模分段函数的函数形式,因此判定其不服从双模分布,也就不存在相应拟合优度 R^2 的计算。而在靠近 10^0 处有一段指数拟合优度数值急速下降,而幂律分布拟合优度提高,这是由于微博行为进入队列的概率较高,相邻两个微博行为发生的时间间隔较为平均,时间间隔值均在 $[10^0, 10^1]$ 之间,在进行数据拟合的步骤时,由于间隔点只有2到3个,不足以确定其函数分布形式,因此在靠近 $v=10^0$ 附近的拟合优度值并没有太大意义。

分别选取 $v=10^{-5}, 10^{-3}, 10^{-1}$ 3种情况来对模型用户的时间间隔分布情况进行研究,结果发现上述3个不同的微博行为概率 v 对应的时间间隔分布服从3种不同的函数分布形式,如图3所示。

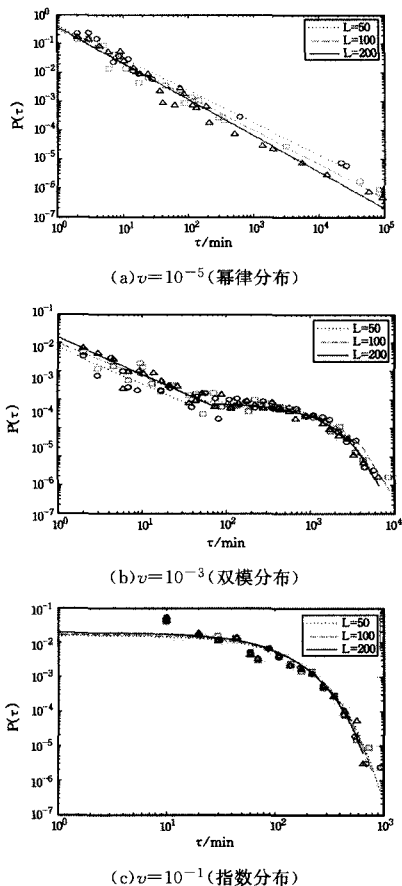


图3 模型用户时间间隔分布

由模型机制可知,新加入队列的任务是微博任务的概率等于队列执行微博事件数占总任务的比例。当 $v=10^{-5}$ 时,说明队列中发布微博任务的数量少,虽然时间间隔的数值呈现出大小不同的分布,但是由于在执行事件总数相同的条件下,执行微博事件的数量较少,造成相邻两次发布微博事件时间间隔大的情况较多,使相邻两次发布微博任务时间间隔呈幂律分布。当 $v=10^{-3}$ 时,加入队列中的任务数量有所增加,使相邻两次发布微博任务的时间间隔有大有小,因此,其时间间隔分布呈现双模的分段分布。而当 $v=10^{-1}$ 时,队列中发布微博任务数量相对较多,相邻两次发布微博事件时间间隔大的情况较少,因此发布微博任务的时间间隔分布呈指数现象。在保持微博行为概率 v 不变的情况下,分析队列长度 L 对时间间隔分布的影响。如图3所示,圆形、正方形和三角形分别表示队列长度为50、100和200时模型产生的微博事件时间间隔分布;点线、虚线和实线分别是对上述不同队列长度模型结果的拟合曲线。结果表明,在不同队列长度 L 下得到的事件时间间隔的分布形式保持不变。

我们统计得到了每位用户在2013-07-01至2013-12-31期间发布微博的总数,依据不同分布的用户群体,作出了不同分布群体发布微博数量的概率密度图,如图4所示。

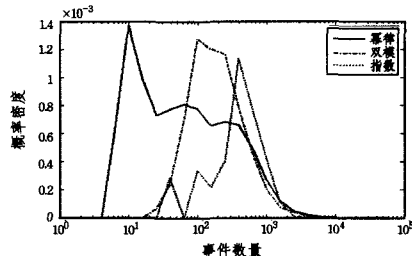


图4 半年内实际用户发布微博总数的分布

结果显示,服从幂律分布的用户在半年内发布微博的数量少;服从指数分布的用户在半年内发布微博的数量多;而服从双模分段分布的用户在半年内发布微博的数量居中。实际结果与模型结果一致。

实际生活中,一些用户发布微博行为占总行为的比例小,即发布微博行为不活跃,时间间隔数值大的情况较多,从而使发布微博时间间隔呈现幂律分布。还有一些用户发布微博的行为占总行为的比例大,即发布微博行为活跃,时间间隔数值大的情况很少,从而使发布微博的时间间隔呈现指数分布。而发布微博行为活跃程度居中的用户的时间间隔呈现双模分段分布。

结束语 随着微博平台的发展与壮大,其在人际沟通和信息传播等领域发挥的作用日渐加强,而由于微博的月活跃用户数量和日活跃用户数量已分别达到1.291亿和6140万,对微博中涉及到的人类行为模式的研究具有非常重大的意义。基于这样的研究背景,文中分析了个体用户发布微博的时间间隔分布特征,提出了用户个体时间行为的内在机制模型。根据数据拟合优度结果,发现个体用户的时间间隔在双对数坐标下的分布情况主要有3种。依据个体时间分布的行为特征,发现用户发布微博数量的多少会影响其时间间隔分布。文中提出基于优先队列研究的个体行为时间模型。该模型在队列中区分了微博发布行为和其他行为类型,在执行事

件总数相同的条件下,若新进入队列是微博任务的概率不同,则说明处理微博事件数量占处理总事件数量的比例不同,从而对应最后微博行为的时间间隔分布不同,用个体微博行为的活跃程度机制说明了用户个体行为的时间模式。

由于受到数据采集类型的限制,文中所研究的用户行为模式主要集中在时间特性上。而随着微博平台的不断发展及其功能的不断增加,用户之间的结构性差异也会日趋显现,具有不同年龄背景、身份差异和兴趣爱好的用户会不断增加。在以后的研究中,可以考虑区分个体差异性的影响,查看具有相同身份或兴趣的用户的行为模式是否呈现相同的特点。

参 考 文 献

- [1] Barabasi A L. The origin of bursts and heavy tails in human dynamics [J]. *Nature*, 2005, 435: 207-211
- [2] Oliveira J G, Barabasi A L. Darwin and Einstein correspondence patterns [J]. *Nature*, 2005, 437: 1251-1251
- [3] 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
- [4] Wu Y, Zhou C, Xiao J, et al. Evidence for a bimodal distribution in human communication [J]. *Proceedings of the National Academy of Sciences*, 2010, 107(44): 18803-18808

(上接第9页)

- 贾佳. 异构并行计算机容错技术研究[D]. 长沙: 国防科学技术大学, 2011
- [69] Dubrova E. *Fault Tolerant Design: An Introduction*[M]. Kluwer Academic Publishers, 2006
- [70] Somani A K, Trivedi K S. A cache error propagation model [C] // *Pacific Rim International Symposium on Fault-tolerant Systems*. IEEE, 1997: 15-21
- [71] Rashid L, Pattabiraman K, Gopalakrishnan S. Modeling the Propagation of Intermittent Hardware Faults in Programs [C] // *Pacific Rim International Symposium on Dependable Computing*. IEEE Computer Society, 2010: 19-26
- [72] Liu Xin-zhong. *Research on Defect Correlation and Its Application*[D]. Changchun: Jilin University, 2010 (in Chinese)
刘新忠. 关联缺陷及其应用研究[D]. 长春: 吉林大学, 2010
- [73] Jing Tao. An approach for detecting Correlation Software defects [J]. *Journal of Software*, 2005, 16(1): 17-28 (in Chinese)
景涛. 软件关联缺陷的一种检测方法[J]. *软件学报*, 2005, 16(1): 17-28
- [74] Bishop P G, Pullen F D. PODS revisited-A study of software failure behavior [C] // *Proc. of the IEEE Int'l Symp. on Fault Tolerant Computing*. Tokyo, 1988: 1-8
- [75] Katerina G P, Trivedi K S. Failure correlation in software reliability models [J]. *IEEE Trans. on Reliability*, 2000, 49(1): 37-48
- [76] Zhang Rong-hui, Jiang Nan, Gou Lang. Software Reliability Growth Model Considering Defect Correlation [J]. *Computer Engineering*. 2008, 34(8): 44-46 (in Chinese)
张荣辉, 姜楠, 勾朗. 一种考虑缺陷关联的软件可靠性增长模型 [J]. *计算机工程*, 2008, 34(8): 44-46
- [77] Le W, Soffa M L, Le W. Path-based fault correlations [C] // *Proc. of the 18th ACM SIGSOFT Int'l Symp. on Foundations of Software Engineering*. ACM Press, 2010: 307-316

- [5] Vazquez A, Oliveira J G, Dezso Z, et al. Modeling burst and heavy tails in human dynamics [J]. *Physical Review E*, 2006, 73(3): 036127
- [6] Wu Y, Zhou C, Chen M, et al. Human comment dynamics in on-line social systems [J]. *Physica A: Statistical Mechanics and its Applications*, 2010, 389(24): 5832-5837
- [7] Wu Y, Ye Q, Li L, et al. Power-law properties of human view and reply behavior in online society [J]. *Mathematical Problems in Engineering*, 2012(2): 243-253
- [8] Kim J, Lee D, Kahng B. Microscopic modelling circadian and bursty pattern of human activities [J]. *PloS one*, 2013, 8(3): e58292
- [9] Song Y, Zhang C, Wu M. The study of human behavior dynamics based on blogosphere [C] // *Proceeding of International Conference on Web Information Systems and Mining*, 2010. IEEE, 2010: 87-91
- [10] Xie J, Zhang C, Wu M. Modeling microblogging communication based on human dynamics [C] // *Proceeding of International Conference on Fuzzy Systems and Knowledge Discovery*, 2011. IEEE, 2011: 2290-2294
- [11] Choi S C, Wette R. Maximum likelihood estimation of the parameters of the gamma distribution and their bias [J]. *Technometrics*, 1969, 11(4): 683-690
- [78] Zhang Da-lin, Jin Da-hai, Gong Yun-zhan, et al. Optimizing Static Analysis Based on Defect Correlations [J]. *Journal of Software*, 2014, 25(2): 386-399 (in Chinese)
张大林, 金大海, 宫云战, 等. 基于缺陷关联的静态分析优化 [J]. *软件学报*, 2014, 25(2): 386-399
- [79] Sun Jin-shan. *Research of Regression Testing Techniques Incorporating Defect Correlation* [D]. Changchun: Jilin University, 2010 (in Chinese)
孙金珊. 引入关联缺陷的回归测试技术研究 [D]. 长春: 吉林大学, 2010
- [80] Wu K, Malaiya Y K. The Effect of Correlated Faults on Software Reliability [C] // *International Symposium on Software Reliability Engineering*. 1993: 80-89
- [81] Dai Y, Xie M, Poh K. Modeling and analysis of correlated software failures of multiple types [J]. *IEEE Transactions on Reliability*, 2005, 54(1): 100-106
- [82] Sue C, Lyu M R, Kuo S, et al. Software Reliability Growth Models Incorporating Fault Dependency with Various Debugging Time Lags [C] // *COMPSAC 2004*. 2004: 186-191
- [83] Zhao Jing, Zhang Ru-bo, Gu Guo-chang. Study on Software Reliability Growth Model Considering Failure Dependency [J]. *Chinese Journal of Computers*, 2007, 30(10): 1713-1720 (in Chinese)
赵靖, 张汝波, 顾国昌. 考虑故障相关的软件可靠性增长模型研究 [J]. *计算机学报*, 2007, 30(10): 1713-1720
- [84] Murthy D N P, Nguyen D G. Study of Two-Component System with Failure interaction [J]. *Naval Research Logistics Quarterly*, 1985, 32(2): 239-247
- [85] Li Zhi-rong, Gao Qi, Liu Shen-yang, et al. Reliability Analysis for a Multi-unit System with Failure Rate Interaction [J]. *Fire Control & Command Control*, 2014, 39(6): 104-107 (in Chinese)
栗志荣, 高崎, 刘慎洋, 等. 一种故障相关多单元系统可靠性分析 [J]. *火力与指挥控制*, 2014, 39(6): 104-107