

基于社交网络的用户行为记忆性研究

李凯凯, 宋礼鹏

(中北大学 计算机与控制工程学院, 山西 太原 030051)

摘 要: 为了研究用户访问社交网络时的记忆效应, 分析了某高校师生用户对 QQ 空间和新浪微博这两个目前较主流的社交平台的访问记录, 采用自相关函数来量化记忆性, 发现在阵发周期内, 个体用户和群体用户的行为数都服从幂律分布, 因此, 记忆性取决于同一阵发周期内相邻时间间隔的关联性, 并从相邻时间间隔相关性的角度分析了用户在线行为的记忆性. 研究结果表明, 在同一个阵发周期内用户在线行为相邻时间间隔的比率为双帕累托分布, 用户根据该比值的大小来增加或减少之后的登陆频率, 实证数据研究与理论分析相吻合, 证实了时间间隔中记忆性的存在.

关键词: 社交网络; 时间间隔; 记忆性; 双帕累托分布; 相关性

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2017)03-0133-03

The Memory Effect Based on Online Users Behavior of Social Network

LI Kai-kai, SONG Li-peng

(School of Computer Science and Control Engineering, North University of China, Taiyuan 030051, China)

Abstract: To research the influence of memory effect on online uses' behavior based on social network, the paper studied the records of some university users' access to social network and used the auto-correlation coefficient method to quantify the memory effect. Empirical research shows that the number of events obeys power-law distribution in the individual and group level. Therefore, the memory effect depends on the adjacent time interval. The continuous time interval less than some threshold are defined as burst periods. The result shows that the ratio of two adjacent time interval in a same active period shows the double Pareto distribution. According to the ratio, users increase or decrease the frequency of landing. The empirical data is consistent with the theoretical analysis. Thus, it is proved that there exists memory in the time interval.

Key words: social network; time interval; memory effect; double Pareto distribution; correlation

1 引言

随着互联网技术的发展, 社交网络逐渐成为人们信息分享、交流、交互不可或缺的重要工具, 而大数据的发展又使得获取用户的各种行为数据成为可能. 通过对海量的用户行为数据进行分析, 可以更准确地了解和分析用户的行为, 研究用户在线行为的记忆效应, 能更加准确地预测用户的行为, 对基于用户行为的信息推荐、舆论引导、谣言传播与抑制以及

网络安全防护等都有重要的指导意义^[1-2]. 针对人类的各种行为, 近年来出现了大量实证研究, 众多的学者也提出了多种模型来解释其内在的机制. 有基于优先级的任务队列模型^[3], 兴趣驱动模型^[4-5], 基于周期性和个体交互模型^[6], 以及由价值驱动的动力学模型^[7]等. 这些模型从不同的角度揭示了人类行为的内在动力学机制. Hou 等^[8]研究发现用户在进行选择和评价行为时, 都表现出了较强的记忆效应. 由于人类行为的复杂性和多样性, 在揭示人类记忆

规律方面仍需进一步研究。

本文通过对某高校师生用户在十个月内对 QQ 空间和新浪微博的访问记录进行统计分析,用自相关函数对记忆性进行量化,并研究了用户在阵发周期内事件数的分布,发现用户对社交网络的访问存在较强的记忆效应,之后,分析了用户在线行为时间间隔序列之间的关联性,从相邻时间间隔的相关性方面解释了用户在线行为的记忆性。

2 实验数据分析

本次研究的实验数据来源于国内某高校信息中心,记录了该校师生在十个月的时间内对 QQ 空间和新浪微博的访问记录.对于这两种社交媒体来说,操作简单和即时传播是其最大的特点,随着平台功能的不断完善,用户迅猛增加,同时两者的用户都具有普遍性以及高忠诚度的特性^[9],因此,研究的对象相对比较稳定.为了提高实验的准确性,实证研究时去掉了登陆次数少于 10 次的用户.若用户 i 访问某一网站的时间为 $t_1, t_2, \dots, t_n$,则时间间隔定义为 $\tau_1 = t_2 - t_1, \tau_2 = t_3 - t_2, \dots, \tau_{n-1} = t_n - t_{n-1}$.

记忆性作为人类行为中的一个重要属性,对人的行为习惯有着至关重要的影响.如果用户当前的某一行为会对之后的行为产生一定的影响,则认为用户在该行为中记忆性是存在的.对人类行为记忆性进行量化分析时,常用自相关函数来表示用户行为的记忆性^[10],即

$$M = \frac{1}{n_\tau - 1} \sum_{i=1}^{n_\tau-1} \frac{(n_i - m)(n_{i+1} - m)}{\sigma^2} \quad (1)$$

式中, n_τ 为用户时间间隔的个数, m, σ 分别为时间序列 $\{n_\tau, n_\tau, \dots, n_\tau\}$ 的均值和标准差.从上述表达式可知,用户在线行为的记忆性依赖于相邻时间间隔之间的关联性.当连续两次事件发生的时间间隔小于某一阈值 Δt 时,可认为这两个事件是相关的.若将所有连续相关事件序列发生的时间段定义为阵发周期,则在观测区间内,每个用户都可能存在多个阵发周期.如图 1 所示为某一用户访问特定网站的时间序列,其中 τ_1, τ_2, τ_3 均小于 Δt ,而 $\tau_4 > \Delta t$,因此, τ_1, τ_2, τ_3 的集合构成了一个阵发周期 T_1 ,将一个阵发周期内所发生的事件数表示为 E , T_1 周期内的事件数为 4, T_2 周期内的事件数为 3.每个阵发周期事件数 E 分布为^[11]

$$P(E = n) = \left(\int_0^{\Delta t} P(\tau) d\tau \right)^{n-1} \left(1 - \int_0^{\Delta t} P(\tau) d\tau \right) \quad (2)$$

式中, $P(\tau)$ 为时间间隔 τ 的分布.令 $a = \int_0^{\Delta t} P(\tau) d\tau$,对任意 $n > 0$,由于时间间隔的阈值 Δt 为一具体数值,

故 $a < 1$,从而 $P(E = n) \sim a^{n-1}$.因此,在事件发生的整个序列中, E 的分布为

$$P(E) \sim E^{-\beta} \quad (3)$$

图 2 所示分别为个体和群体层面上用户事件数 E 的概率分布.

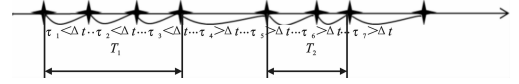
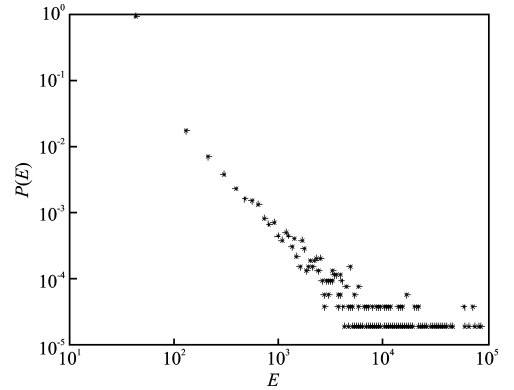
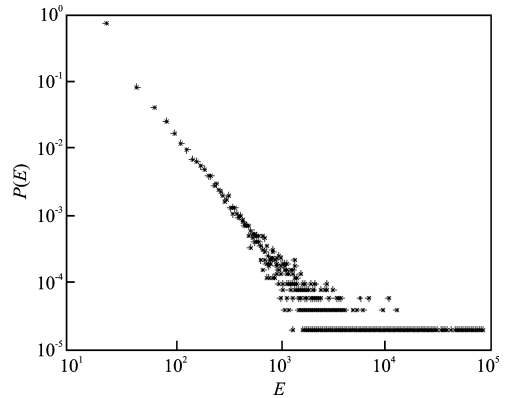


图 1 阵发周期



(a) 个体用户



(b) 群体用户

图 2 阵发周期内用户事件数的分布

3 统计模型及模型检验

由以上分析可知,阵发周期内相邻时间间隔之间的相关性可以表征用户在线行为的记忆效应.因此,可将相关性系数表示为

$$k_i = \frac{\tau_{i+1}}{\tau_i} \quad (4)$$

式中, τ_i 是一个活跃周期内第 i 个时间间隔.则对于每一个活跃周期内的第 i 个时间间隔 τ_i 都满足:

$$\tau_i = k_i \tau_{i-1}, i \geq 1 \quad (5)$$

对等式(4)两边取对数,可得:

$$\ln k_i = \ln \tau_i - \ln \tau_{i-1} \quad (6)$$

时间间隔 τ_i 之间相互独立,根据中心极限定理

可知, k 的密度函数为

$$f(k) = \frac{1}{\sqrt{2\pi}\sigma k} e^{-\frac{(\ln k - \mu)^2}{2\sigma^2}}$$

(7)

由式(5) 递推可以得到:

$$\tau_i = k_i k_{i-1} \cdots k_1 \tau_0$$

(8)

τ_0 为每一个活跃周期的初始值. 而从对数正态分布的性质可知,

$$\begin{aligned} \ln f(K) &= \ln f(k_1) + \ln f(k_2) + \cdots \ln f(k_n) \\ &= \ln f(k_1) f(k_2) \cdots f(k_n) \end{aligned}$$

(9)

因此, $\frac{\tau_i}{\tau_0}$ 服从参数为 $(i\mu, i\sigma^2)$ 的对数正态分布.

即:

$$f_{\tau_i/\tau_0}(k) = \frac{1}{\sqrt{2i\pi}\sigma k} e^{-\frac{(\ln k - i\mu)^2}{2i\sigma^2}}$$

(10)

若一个用户有 n 个活跃周期,在每一时刻用户结束一个活跃周期的概率为常数 λ ,一个用户在所有活跃周期内的时间间隔分布为

$$g(k) = \int_{i=0}^{\infty} \frac{\lambda}{\sqrt{2\pi i}\sigma k} e^{-\lambda i - \frac{(\ln k - i\mu)^2}{2i\sigma^2}} di$$

(11)

令 $i = u^2$,式(11) 可变形为

$$g(k) = \frac{2\lambda}{\sqrt{2\pi}\sigma k} \int_{u=0}^{\infty} e^{-\lambda u^2 - \frac{(\ln k - u^2\mu)^2}{2u^2\sigma^2}} du$$

(12)

进一步化简可得:

$$g(k) = \frac{2\lambda k^{\mu/\sigma^2}}{\sqrt{2\pi}\sigma k} \int_{u=0}^{\infty} e^{-(\lambda + \frac{\mu^2}{2\sigma^2})u^2 - \frac{\ln^2 k}{2\sigma^2 u^2}} du$$

(13)

令 $m = \lambda + \frac{\mu^2}{2\sigma^2}$, $n = \frac{\ln^2 k}{2\sigma^2}$,已知 $\int_{u=0}^{\infty} e^{-mu^2 - \frac{n}{u^2}} du = \frac{1}{2}$

$\sqrt{\frac{\pi}{m}} e^{-2\sqrt{nm}}$,带入表达式(13),可将式(11) 化为

$$g(k) = \begin{cases} \frac{\lambda}{\sqrt{\mu^2 + 2\lambda\sigma^2}} k^{-1+\mu+\frac{\sqrt{\mu^2+2\lambda\sigma^2}}{\sigma^2}}, & 0 < k < 1 \\ \frac{\lambda}{\sqrt{\mu^2 + 2\lambda\sigma^2}} k^{-1+\mu-\frac{\sqrt{\mu^2+2\lambda\sigma^2}}{\sigma^2}}, & k \geq 1 \end{cases}$$

(14)

变量替换之后可得:

$$g(k) = \begin{cases} \frac{ab}{a-b} k^{-1+a}, & 0 < k < 1 \\ \frac{ab}{a-b} k^{-1-b}, & k \geq 1 \end{cases}$$

(15)

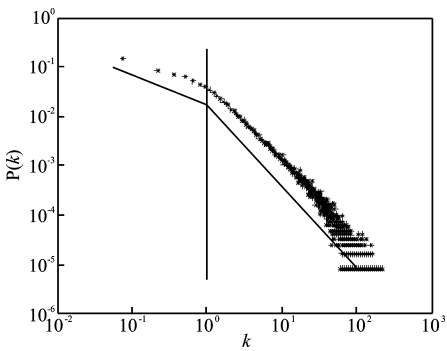
式中, $a = \mu + \frac{\sqrt{\mu^2 + 2\lambda\sigma^2}}{\sigma^2}$, $b = \mu - \frac{\sqrt{\mu^2 + 2\lambda\sigma^2}}{\sigma^2}$. 式

(15) 即为双帕累托分布的概率密度函数.

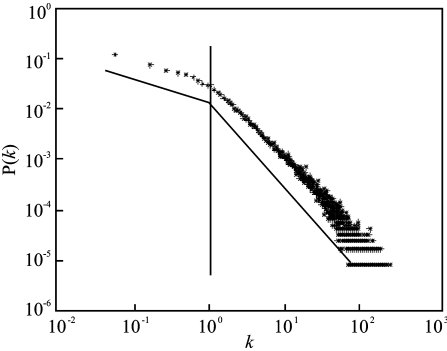
取时间间隔的阈值为 1 小时,对 QQ 空间和新浪微博中用户相邻时间间隔的比值进行统计分析,实证研究结果如图 3 所示,从图中可以看出,时间间隔比值 k

的分布满足双帕累托分布,用户根据当前的行为来调整之后的行为. $k = 1$ 时为分布的临界点:

- (1) $0 < k < 1$ 时,用户登陆网站的频率增大;
- (2) $k > 1$ 时,用户登陆网站的频率减小.



(a) QQ 空间



(b) 新浪微博

图 3 相邻时间间隔比值的分布

4 结束语

本文通过对某高校师生对 QQ 空间和新浪微博的访问记录进行分析,研究了用户在线行为的记忆效应,结果表明,用户的上网行为表现出了较强的记忆性. 并从时间间隔相关性方面分析了用户的在线行为,相邻时间间隔之间的比率满足双帕累托分布,表明当前时间间隔会对之后的时间间隔的大小产生影响,从而证实了在时间间隔方面记忆性的存在.

参考文献:

[1] 周涛,韩筱璇,闫小勇,等. 人类行为时空特性的统计力学[J]. 电子科技大学学报,2013,15(4):481-540.

[2] 余林琛,沈江波,何仲彻,等. 一种基于用户行为习惯的主机安全性检测方法[J]. 微电子学与计算机,2011,28(9):158-159,164.

[3] Albert-László B. The origin of bursts and heavy tails in human dynamics. [J]. Nature, 2005, 435(7039): 207-211.

是选择全复用增益来传输数据包. 所以, 如果发射器不增加发送功率的话, 它的分集增益很小, 数据包误码率就会很高. 因此, 当区域内总的中继节点数量很少时, MIMO 全复用算法就会消耗更多的电量.

6 结束语

MIMO 全复用算法具有更多选择下一跳的机会, 可以选择临近它的节点作为它的下一跳, 并以此减少发射功率. 和 MIMO 全分集相比, 它具有更高的复用增益. 相反的, UMIMO 全分集可以在发送端和接收端之间的较大距离内选择它的下一跳, 和 MIMO 全复用相比, 它具有较高的分集增益, 这样可以减少包错误率.

本文提出权衡复用和分集的 MIMO 算法实现了分集和复用的折衷, 实现了分集增益和复用增益的权衡. 在保证了传输可靠性的基础上实现了节点能耗降低, 延长了整个网络的生存时间, 同时实现了传输速率、误码率等水下传感器网络的性能提升.

参考文献:

[1] 郭瑛, 张震. 大规模水下传感器网络时间同步研究[J]. 电子与信息学报, 2014, 36(6): 1498-1503.

[2] 夏娜, 郑语晨, 杜华争, 等. 刚性驱动水下传感器节点自组织布置[J]. 计算机学报, 2013, 36(3): 494-505.

[3] 赵旦峰, 王杨, 廖希, 等. 级联编码的 MIMO-OFDM 水声通信系统[J]. 吉林大学学报(工学版), 2014, 44(4): 1171-1176.

[4] Kilfoyle D B, Preisig J C, Baggeroer A B. Spatial modulation experiments in the underwater acoustic channel[J]. IEEE Journal of Oceanic Engineering, 2005, 30(2): 406-415.

[5] Gaultier Real, Pierre-Philippe Beaujean, Pierre-Jean Bouvet. A channel model and estimation technique for MIMO underwater acoustic communications in ports and very shallow waters at very high frequencies[C]// OCEANS 2011. Waikoloa; IEEE, 2001: 1-9.

[6] Kuo Li-Chung. Distributed medium access control strategies for MIMO underwater acoustic networking [J]. IEEE Transactions on Wireless Communications, 2011, 10(8): 2523-2533.

[7] Cai Lifeng , Pan Xiang , Wen X U, et al. Underwater acoustic MIMO communication based on active time reversal[C]//Asia Pacific Conference on Postgraduate Research in Microelectronics & Electronics. Shanghai; IEEE, 2009: 45-48.

[8] Kuo Li-Chung, Tommaso Melodia. Cross-layer routing on MIMO-OFDM underwater acoustic links[C]// IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON). IEEE, 2012: 227-235.

作者简介:

王建平 男, (1981-), 博士研究生, 讲师. 研究方向为水下传感器网络、SDN 架构设计.

陈伟(通讯作者) 男, (1964-), 教授, 博士生导师. 研究方向为无线传感器网络、认知无线电、软件无线电技术.

E-mail: greatchen@whut.edu.cn.

(上接第 135 页)

[4] Shang M S, Chen G X, Dai S X, et al. Interest-driven model for human dDynamics[J]. Chinese Physics Letters, 2010, 27(4): 48701-48703.

[5] 刘建粉, 史永昌. 基于用户兴趣分类优化的聚类模型仿真[J]. 微电子学与计算机, 2014, 31(5): 171-174.

[6] 罗卫国. 基于用户交互行为的建模方法研究[D]. 济南: 山东大学, 2007.

[7] 符丁, 李明江, 黎路. 基于价值驱动的人类行为动力学实证研究和建模[J]. 电子科技大学学报, 2015, 17(5): 652-656.

[8] Hou L, Xue P, Qiang G, et al. Memory effect of the online user preference[R]. Scientific Reports, 2014.

[9] 赵飞飞. 网络口碑与微博用户忠诚度的关系研究[D].

大连: 大连理工大学, 2013.

[10] Jo H H, Perotti J I, Kaski K, et al. Correlated bursts and the role of memory range[J]. Physical Review E, 2015(5): 2-9.

[11] Karsai M, Kaski K, Barabási A L, et al. Universal features of correlated bursty behaviour[J]. Scientific Reports, 2011, 2(5): 539-539.

作者简介:

李凯凯 女, (1991-), 硕士研究生. 研究方向为复杂网络.

E-mail: 18734922510@163.com.

宋礼鹏 男, (1975-), 副教授, 博士生导师. 研究方向为复杂网络与网络安全、大数据.