在线协同写作的人类动力学分析*

赵 飞[1)2)] 刘金虎[1)3)] 查一龙[1)4)] 周 涛[1)†]

1)(电子科技大学计算机科学与工程学院,互联网科学中心,成都 610054)

2)(电子科技大学经济与管理学院,信息经济与互联网实验室,成都 610054)

3)(电子科技大学数学科学学院,成都 610054)

4)(电子科技大学示范性软件学院,国际化软件人才实验班,成都 610054)

(2011 年 1 月 22 日收到, 2011 年 3 月 10 日收到修订稿)

对人类在线行为模式的探讨是近年来人类动力学研究的热点. 基于维基百科数据,文章针对一类重要且普遍的在线行为——在线协同写作,进行时间统计特性分析和内容更新统计分析. 实证显示在线协同写作时间间隔分布呈多尺度特征,1 min 到 30 min 和 30 min 到 24 h 两个时间段上时间间隔分别服从指数为 1.62 和 1.16 的幂律分布,而大于 24 h 的时间间隔服从形如 $F(\tau) \propto \tau^{-b-a\log(\tau)}$ 的累积分布. 分析表明,连续提交行为和交互提交行为共同导致时间间隔的多尺度分布,其中连续提交行为主导了 30 min 之内幂指数为 1.62 的间隔时间分布,而交互提交行为主导了 30 min 到 24 h 范围内幂指数为 1.16 的时间间隔分布. 进一步地,文章作者发现反向更新是协同写作过程中普遍存在的现象,反向更新的比例与更新量的关系反映出更新量与相应内容被保留的概率具有极强的关联,很大的更新量得以存活的概率较低. 统计分析暗示维基百科编辑行为中存在“看门狗”和“编辑战争”. 文章的结果有助于加深我们对人类集群行为,特别是协同合作开发行为的认识.

关键词: 在线协同写作, 人类动力学, 多尺度特征, 维基百科

PACS: 89.75.Da, 05.45.Tp, 02.50.Ey

1. 引 言

网络化、多媒体化将人们带入了第二次信息革命时代,基于互联网的在线活动已经成为推动整个经济、社会发展不可缺少的动力[1]. 通过诸如博客、维基、分类标签、搜索引擎、社交网络、微博、即时通信等服务,人们可以免费查找、浏览、发布或评论在线内容,从而形成影响民意、文化、政策、广告利益等方面的巨大草根力量[2]. 这些新兴的 Web2.0 和社会媒体不仅改变了传统通信模式,同时产生并记录下了大量的带时间标记的数据,为我们定量研究个体乃至群体的行为统计特征创造了条件.

传统研究认为,人类的行为可以用泊松过程近似刻画,即相关事件发生速率被近似假设为一个常数. 在这个假设下,两个相继行为时间间隔 τ 的分布是指数的. 然而,这一假设受到了来自大量人类在线活动实证的挑战[3—5]. 从电子邮件接收与发送[6]、市场交易活动[7,8]、网站浏览[9,10]、电影与网络音乐点播[11,12]、手机短信通信[13,14]、在线游戏[15]、虚拟社区活动[16]、无线射频识别[17]等活动的实证结果看,人类活动总是在短时间内密集发生,然后紧接着很长时间的静默等待,行为的间隔时间或等待时间表现出很强烈的偏离泊松过程的胖尾特性. 实证证实泊松分布描述人类行为极不准确,采用幂律分布[18,19]、广延指数分布[20—22]、带截断的幂律分布[23,24],或者指数分布与幂律分布形成的双模态分布[14]等来描述的人类行为会更加贴切.

解释人类复杂行为背后的动力学机理具有非常重大的理论意义和经济社会价值. Barabási[6]提出基于优先级的排队决策过程可以解释单个个体的等待时间的胖尾现象,Vázquez 等[25]完善了该理论,得到幂指数分别为 1 和 3/2 的两大普适类(最近的实证倾向于否定普适类的存在性). 之后,基于含时优先权列表[26—29]、人类自适应兴趣[30]、行为的周期性及季节性[31]、任务本身的关联性[32]、多重泊松分

* 国家自然科学基金重点项目(批准号:10635040)和国家自然科学基金面上项目(批准号:70871082,10975126,70971089)资助的课题.

† 通讯联系人. E-mail:zhutou@ustc.edu

©2011 中国物理学会 Chinese Physical Society

布[33]、记忆性[34]等的研究,进一步多视角地解释了单个个体的行为模式. 值得强调的是,Zhou 等[11],Radicchi[35]分别发现人在线行为的活跃程度对分布的幂指数有非常重大的影响. 交互行为方面,Oliveira 等[36]基于优先级排队提出了的两人交互影响模型,Wu 等[14]进一步完善了该模型并应用来刻画短消息通信时的反馈现象. 然而,至今未见对多人集群行为方面的研究报道.

除了对在线行为的时间特性的关注,研究者还对在线内容流量的演化规律进行了探索. Crane 等[37]采用 Hawkes 条件泊松过程对 Youtobe 视频点击率及关键字的查询量变化进行了建模,Ratkiewicz 等[38]提出带随机跳跃的优先链接模型对外部因素所导致的维基百科文章的点击率、链接入度等指标的爆发性增长进行了刻画,Chmiel 等[39]采用在含权网络上带短记忆的自吸引的随机游走模型描述了门户网站访问流量在其子页面中的分流情况,等等.

本文将以维基百科(www. wikipedia. org)为对象,研究在线协同写作中多人合作表现出来的统计规律. 与以往研究关注特定个体或者群体的时间统计特性不同,本文关注维基百科中针对特定文章(编辑对象)的群体行为展现出来的性质——这是从另外一个视角观察人类协同行为的性质. 在线协同写作方式是伴随着 web2.0 技术的维基模式的提出和实践而诞生的,是一种典型的多人集群行为. 与传统印刷文本的知识信息生产方式比起来,在线协同写作允许任何用户编辑任何一篇文章,且每次编辑产生的新文章版本会被完整地记录下来,这样就形成了维基模式下写作特有的版本提交现象. 英文维基百科是在线协同写作方式最成功的典范之一,目前已有大量研究关注其网络结构演化特征,文献[40]从复杂网络视角实证研究了维基百科文章-超链接结构,文献[41]提出三种基于网络分析的方法对文章进行层次和类别划分,文献[42]详细讨论了文章和超链接的增长模式,文献[43]考察了文章所属类别的自组织特性,文献[44]提出二维排序算法对文章进行排序和检索,等. 关于维基百科的复杂网络特性,以及从复杂性科学的视角透视维基百科,请参考综述论文[45]. 这些研究揭示了维基百科一些非常重大的性质:文章-超链接结构如同万维网一样的蝶形结构;超链接出入度服从指数分别为 2.15 和 2.57 的幂律分布;文章-超链接增长可以近似用"优先连接"机理来描述;两篇文章间平均距离仅为 4 等. 尽管以上研究内容十分丰富,但都未对维基百科的协同写作行为特征进行分析.

2. 数据来源

本文分析的数据皆来自英文版的维基百科网站(http://en. wikipedia. org). 我们首先以文章"Complex Network"为种子,沿着文章中的超链接指向采用宽度优先搜索算法(BFS)抓取了 4 层,一共获取了 9 万个文章的标题,然后从中随机选取 2000 个标题抓取对应文章的完整编辑历史数据. 每个提交的版本元素数据包含版本号、编辑者、提交时间以及版本字节数这几个信息. 以"Complex Network"这篇文章为例,前 5 个版本的元素数据见表 1. 2000 篇文章共提交了 2775385 个版本. 我们分别用提交时间进行时间统计特性分析,用版本字节数进行编辑内容更新量统计特性分析.

表 1 文章"Complex Network"前 5 个提交版本的元素信息

版本号	编辑者	提交时间	版本字节数
1	JFromm	14:05, 7 April 2005	411
2	JFromm	14:06, 7 April 2005	866
3	JFromm	14:11, 7 April 2005	1254
4	JFromm	14:18, 7 April 2005	1324
5	JFromm	14:20, 7 April 2005	1322

3. 协同写作的时间统计特性

3.1. 时间间隔分布

我们首先从单篇文章编辑行为来分析相继提交的版本(可能是同一个用户连续提交,也可能来源于两个不同用户)之间的时间间隔分布. 图 1 显示了从收集的数据集中选取的编辑版本数从少到多的六篇文章的相继版本提交时间间隔分布 $p(\tau)$,其中 τ 表示间隔时间(单位:min). 6 篇文章提交的版本总数从 373 到 30057 不等. 对比发现,不管文章的版本数是多少,时间间隔 $\tau \leqslant \tau^* \equiv 100$ 部分都近似服从幂指数 λ 为 1.50 的幂律分布(幂指数的计算也只涵盖从 1 到 100 这两个数量级),即 $p(\tau) \propto \tau^{-1.50}$. 但它们的尾部($\tau > \tau^*$ 部分)却表现出极大地差异,对版本数较少的文章来说,分布尾部统计规律不明显,但胖尾特征明显,随着版本数的增

多,分布尾部逐渐表现出近似幂律的分布形式. 我们进一步考察数据集中每篇文章 $\tau \leqslant \tau^*$ 的分布情况,从图 2 计算得到的幂指数 λ 的频数直方图可以看出, λ 近似服从均值为 1.51 的正态分布. 这说明所有文章版本提交的时间间隔具有非常相似的统计特征.

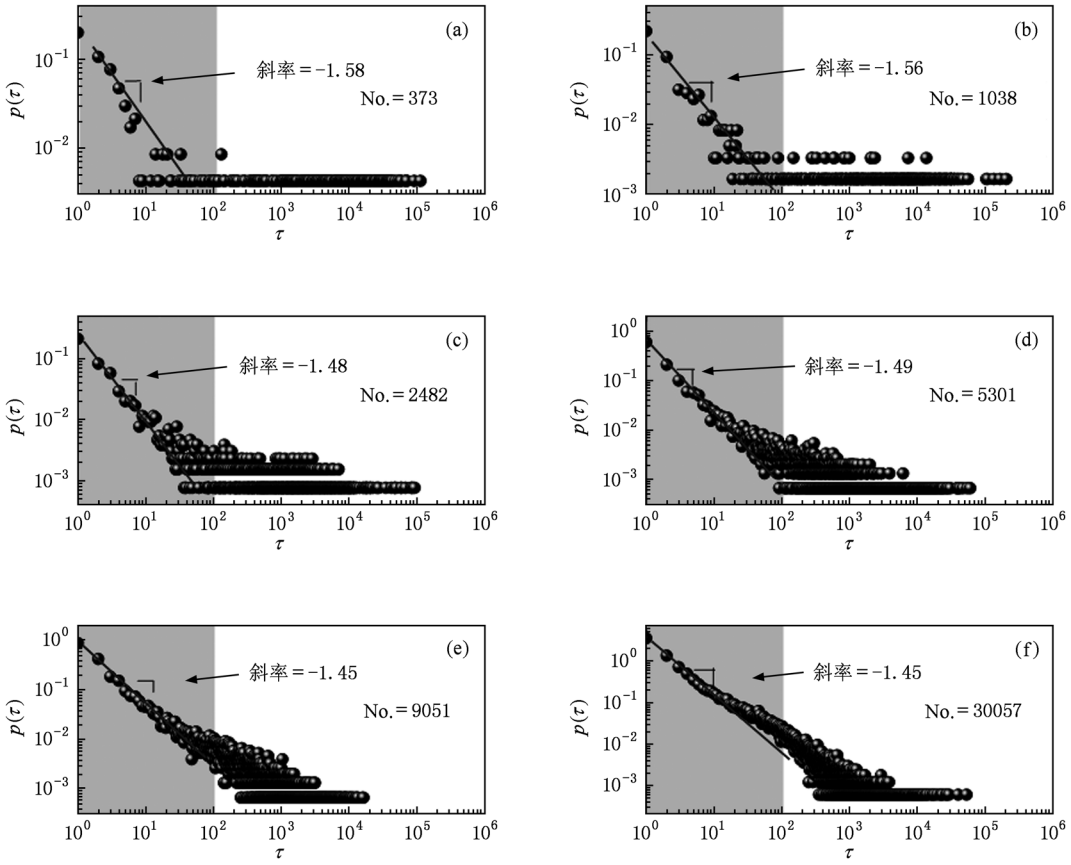

图 1 6 篇随机选取文章版本提交时间间隔频率分布图,6 篇文章依次是(a) Complex Network, (b) WebPage, (c) City, (d) Music, (e) Google, (f) Wikipedia. 图中标记为 No. 的数字表示相应文章提交的版本总数. 所有幂指数的计算均采用极大似然法估计(仅考虑从 1 到 100 这两个数量级),(b)—(f)通过了阈值为 0.9 的 Kolmogonov-Sminov 检验[19,46]

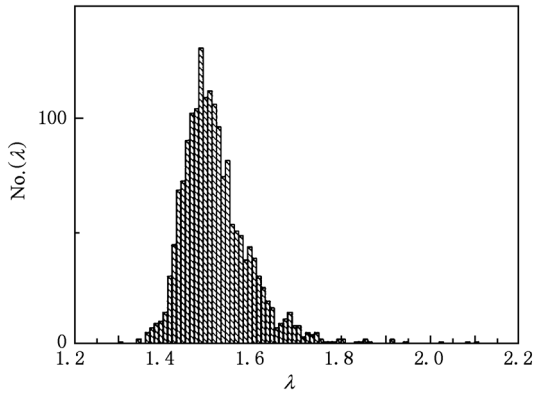

图 2 文章 $\tau \leqslant \tau^*$ 时间间隔内的幂指数频数直方图

我们进一步将所有文章的时间间隔聚集在一起从总体上统计时间间隔分布特性. 图 3(a)是所有文章版本提交时间间隔在双对数坐标下频率分布情况,分布曲线明显表现出 3 个尺度:1) $1 \leqslant \tau \leqslant \tau^0$:服从 $\lambda_1 = 1.62$ 的幂律分布,即 $p(\tau) \propto \tau^{-1.62}$,分布的放大图见图 3(b);2) $\tau^0 < \tau \leqslant \tau^1$:服从 $\lambda_2 = 1.16$ 的幂律分布,即 $p(\tau) \propto \tau^{-1.16}$,分布的放大图见图 3(c);3) $\tau > \tau^1$:此部分样本数量相对较少, $p(\tau)$ 的尾部发散,不便于直接采用幂律函数拟合,于是我们分析其累计分布 $F(\tau) = \int_{\tau}^{+\infty} p(\tau)\mathrm{d}\tau$ 以排除掉尾部发散造成的误差. 图 3(d)中外图的实线给出了 $F(\tau)$ 与 τ 在双对数坐标下的关系,是一条弯曲的弧形;子图中实线显示了 $-\log F(\tau)$ 和 τ 在 x 轴取对数的半对数坐标上的关系,曲线明显具有二次函数的特征,说明无论采用广延指数分布[20—22]还是尾部带截断的幂律分布[23,24]拟合都会有较大的误差. 因此我们直接采用 $-\log F(\tau)$ 关于 $-\log(\tau)$

的二次函数形式 $-\log F(\tau) = a\log^2(\tau) + b\log(\tau) + c$ 来进行拟合,由这个二次函数可以推导出 $F(\tau) \propto \tau^{-b-a\log(\tau)}$. 采用最小二乘法拟合得到系数的估计值 $\hat{a} = 0.50$, $\hat{b} = -3.10$, $\hat{c} = 4.96$, 从而 $\hat{F}(\tau) \propto \tau^{3.10-0.50\log(\tau)}$. 如图 3(d)所示实际数据与拟合曲线符合得很好.

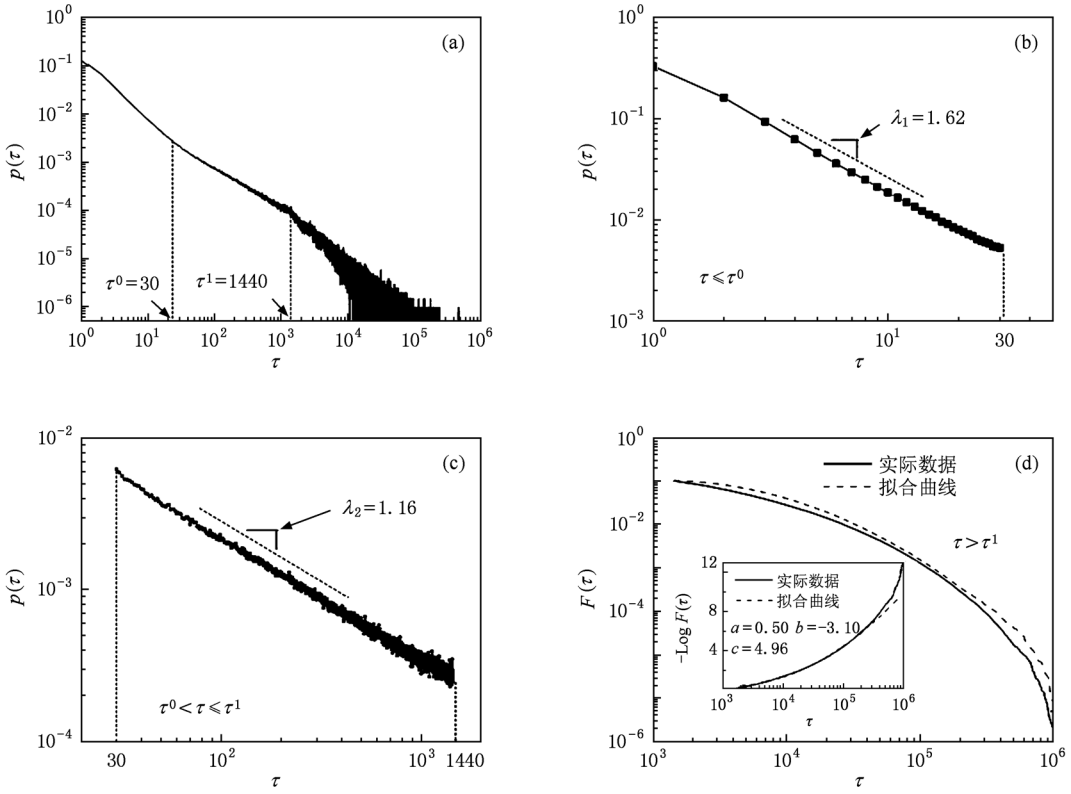

图 3 所有文章时间间隔聚集起来的分布规律 (a)总体时间间隔频率分布;(b) $1 \leqslant \tau \leqslant 30$ 的时间间隔频率分布;(c) $30 \leqslant \tau \leqslant 1440$ 的时间间隔频率分布;(d)外图实线是累积频率分布 $F(\tau)$, 插图实线是 $-\log(F(\tau))$ 与 τ 在 x 轴取对数的半对数坐标上的关系,虚线是拟合曲线. a, b, c 参数的计算采用最小二乘法. 所有指数的计算都采用极大似然法估计

注意到 $\tau^0 = 30$ min $= 0.5$ h, $\tau^1 = 1440$ min $= 24$ h, 说明时间间隔 τ 在分钟、小时、天三个不同时间尺度表现出了不同的统计特性,与 Wang 等人[47,48]分析单个人多尺度的时间统计特性结论相似(他们并未对此现象作出解释). 如何解释这种集群行为在不同时间尺度上表现出的不同统计规律呢? 我们注意到人们在协同写作时表现出两种不同的提交行为:一种行为是前后两个版本提交由同一用户完成,换句话说用户会在自己上次编辑的内容上继续作修改,我们称这种提交为用户连续提交,对应产生的时间间隔为连续提交时间间隔,记为 τ^{C};另一种行为是连续连个版本由不同用户提交,我们称这种提交为用户交互提交, 对应产生的时间间隔为交互提交时间间隔,记为 τ^{I}. 显然,所有版本提交行为都可以归为这两类行为中的一类. 图 4 是连续提交时间间隔 τ^{C} 和交互提交时间间隔 τ^{I} 的在同一坐标下的频率分布图. 对比发现,分钟时间尺度上连续提交 τ^{C} 的比重较大,而小时、天时间尺度上交互提交 τ^{I} 的比重较大. 耐人寻味的是,比重切换点 τ_0 正好对应于总体时间间隔多尺度分布的第一个拐点 $\tau^0 = 30$ min. 连续提交 τ^{C} (图中实线)整个分布几乎都服从 $\lambda^{\mathrm{C}} = 1.63$ 的幂律分布. 注意到 λ^{C} 与总体分布中分钟时间尺度的指数 $\lambda_1 = 1.62$, 两者几乎相等—这样的巧合只有一个解释,即分钟时间尺度上的时间间隔主要由单个用户连续提交行为驱动形成.

交互提交 τ^{I} (图中虚线)的频率分布曲线明显可以分为两段,分界点 τ_1 也正好对应于总体时间间隔多尺度分布的分界点 τ^1, 见图 3(a). 中间小时段 $\tau_0 < \tau < \tau_1$ 的时间间隔服从指数为 $\lambda^{\mathrm{I}}_{\tau_0<\tau<\tau_1} = 1.16$ 的幂律分布,该指数也正好等于总体分布中小时时间尺度分布的指数 $\lambda_2 = 1.16$, 这说明这一段时间间隔主要由用户交互提交形成. 最后我们考察交互提交在天时间尺度上 $\tau > \tau_1$ 的时间间隔,图 5 展示

了频率累积分布,采用与图3(d)一致的分析方法得到 $\hat{F}(\tau^{\mathrm{I}})$ 的拟合结果: $\hat{F}(\tau^{\mathrm{I}}) \propto \tau^{\mathrm{I}3.06-0.48\log(\tau^{\mathrm{I}})}$. 不难看出此结果与总体时间间隔中天时间尺度上的分布结果也几乎一致,从而这段时间间隔也可以认为主要是由多人用户交互提交形成的. 通过以上分析我们可以认为,不同时间尺度下对应的不同的起主导作用的提交行为模式是导致在线协同写作时间间隔分布呈多尺度特征的主要原因.

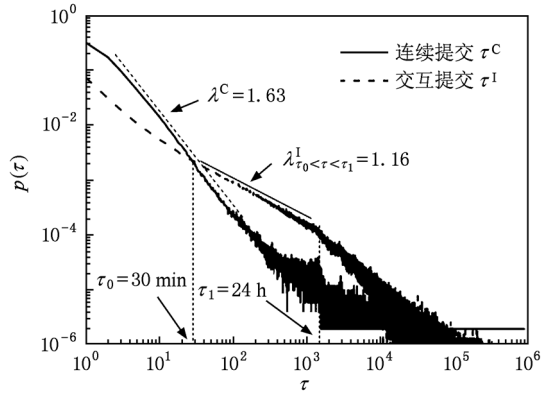

图4 连续提交时间间隔 τ^{C} (实线)和交互提交时间间隔 τ^{I} (虚线)的频率分布图. 幂指数用极大似然法估计而得

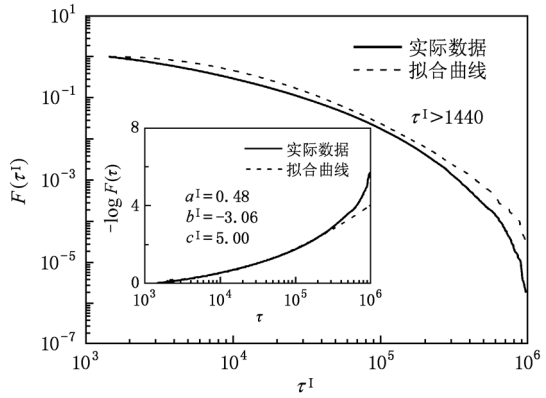

图5 交互提交时间间隔在 $\tau^{\mathrm{I}} > \tau_1 \equiv 1440$ 部分的分布:外图实线是累积频率分布 $\hat{F}(\tau^{\mathrm{I}})$, 子图实线是 $-\log(\hat{F}(\tau^{\mathrm{I}}))$ 与 τ^{I} 的关系,虚线是拟合的曲线

3.2. 活跃度及时间间隔分布宽度

前面的分析证实了所有文章的编辑具有相似的统计特性,但不同文章的区别主要体现在哪儿呢? 为了回答这个问题,我们采用文献[11]中研究在线电影点播、网页浏览等活动时的方法,观察活跃度 A 的作用. 某篇文章 i 的活跃度定义为

$$A_i = \frac{\mathrm{No}_i}{T_i}. \tag{1}$$

其中 No_i 是该文章已提交的版本数, T_i 是最后提交的版本与首次提交的版本间的时间差. 我们考察3.1节统计得到的幂指数 λ 与 A 的关系,如图6(a)所示,在对数坐标上把 A 等距离划分,将 λ_i 分成6组后统计每组的均值 $E(\lambda_k)$ 和标准差 $D(\lambda_k)$ (用误差棒表示), $k = 1,2,\cdots,6$. 每组的 $E(\lambda_k)$ 都约等于1.5,但随着 A 的增大, $E(\lambda_k)$ 有微弱的减小;而标准差随着 A 越大而显著地降低. 幂指数标准差的降低一方面是因为活跃度大的组文章数较少,更重要的是活跃度越大的文章其时间间隔分布情况也越来越趋于一致.

二阶矩 $\langle\tau^2\rangle = \int p(\tau)\tau^2 \mathrm{d}\tau$ 是衡量时间间隔 τ 分布宽度的指标[11]. 如图6(b)所示,我们发现 $\langle\tau^2\rangle$ 无一例外地会随着 A 幂律下降,且 $\langle\tau^2\rangle \propto A^{-1.28}$ (计算幂指数时排除掉了图形左上方 $\langle\tau^2\rangle \geqslant 10^{11}$ 的少量异常样本). 与文献[11]类似,活跃度大的文章其编辑时间间隔分布比较窄. 但是,正如我们在图2

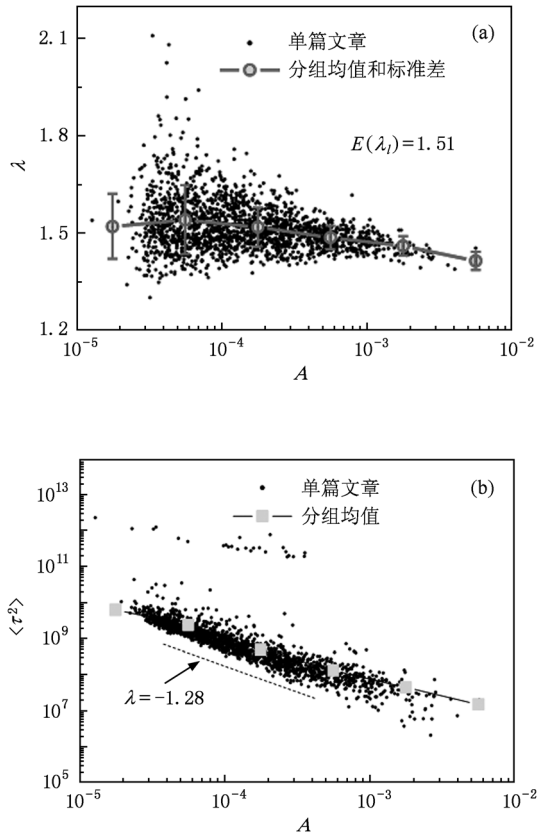

图6 (a) 2000篇文章幂指数 λ 与活跃度关系;(b)分布宽度 $\langle\tau^2\rangle$ 与文章活跃度 A 的关系

中观察到的,不同文章的统计性质相差并不大,至少要明显小于在线电影观看[11]. 编辑行为中的相互作用有可能是增加不同个体统计一致性的原因.

3.3. 阵发性和记忆性

阵发性(Burstiness)和记忆性(Memory)是用来描述包括人类行为在内的很多复杂系统的两个重要物理量[48]. 其中阵法性 B 刻画人类活动存在的短期内密集活动和长时间静默的程度,采用公式

$$B = \frac{\sigma_\tau - m_\tau}{\sigma_\tau + m_\tau} \tag{2}$$

来计算,其中 σ_τ 和 m_τ 分别指时间间隔 τ 的标准差和均值, B 取值在 $(-1,1)$ 之间. $B \approx 1$ 对应于胖尾最严重的时间间隔; $B = 0$ 是泊松分布时间间隔;而 $B = -1$ 则是周期性很强的,如心脏跳动、睡眠等时间间隔(周期时间序列, $\sigma_\tau = 0$). 记忆性 M 刻画人类活动短(长)时间间隔之后出现短(长)时间间隔的趋势,采用公式

$$M = \frac{1}{N_\tau - 1}\sum_{i=1}^{N_\tau - 1}\frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1\sigma_2} \tag{3}$$

计算,其中 N_τ 指时间间隔的总数, $m_1(\sigma_1)$ 和 $m_2(\sigma_2)$ 是相邻两个时间间隔 τ_i 和 τ_{i+1} 的均值和标准差,取值也在 $(-1,1)$ 间. $M > 0$ 意味着短(长)时间间隔后面跟随短(长)时间间隔可能性更大; $M = 0$ 意味着跟随短(长)时间间隔可能性一样大; $M < 0$ 则意味着跟随长(短)时间间隔可能性更大. M 越大,显示出时间间隔的可预测性越强.

之前的研究表明诸如电子邮件发送、图书馆借阅、校园打印等单个个体的活动具有较强的阵发性($0 < B < 0.5$),但记忆性却比较微弱($M \approx 0$)[49]. 我们计算每篇文章版本提交时间间隔的记忆性和阵发性. 图 7 是所有文章的 B 和 M 投影在 (B,M) 二维空间上的视图,不难看出,几乎所有文章的 B 为正值,说明所有文章的阵发性都强于泊松分布,且均值 $E(B) = 0.44$ 比之前研究中单个个体行为的阵发性要强;另外,绝大多数文章的 M 大于 0,均值 $E(M) = 0.19$,这与单个个体行为微弱的记忆性显著不同,说明版本提交行为可能具有较强的可预测性. 考虑到版本提交时间间隔的产生是多人协同完成的结果,其较高的阵发性及记忆性暗示在线协同写作这种集群行为可能存在某种具有高预测性的编辑模式.

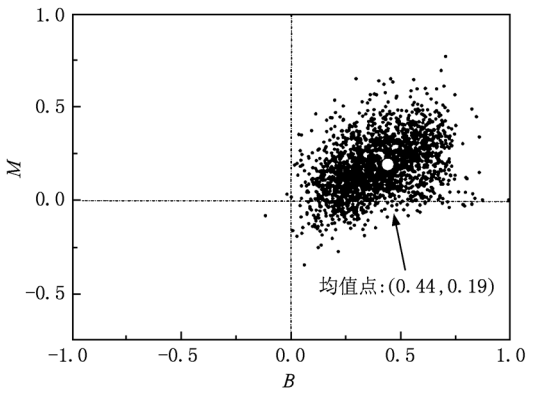

图 7 文章版本提交时间间隔 τ 的阵发性 B 和记忆性 M

4. 文章内容更新量统计分析

近年来,针对维基百科内容的数据挖掘是 web 语义研究及自然语言处理(NPL)的一个热点[50—56]. 这些研究深层次地探索了维基百科的内容结构,但没有反映出人们是如何合作编辑完成这些内容的. 较不关注维基文章的具体内容,仅仅通过对文章内容更新量的统计来挖掘协同写作中的编辑模式. 一个大小为 S_l 的文章版本相对于大小为 S_{l-1} 的上一个版本的更新量定义为 $s_l = S_l - S_{l-1}$, s_l 可正可负. 分别图 8(a)是文章内容更新量 s 的频率分布图(正负字节更新分开计算后画在同一图上),正负字节都服从幂率分布, $p(s) \propto |s|^{-\lambda_\pm}$, 其中, $\lambda_+ = 1.23$, $\lambda_- = 1.22$. 正负字节分布的对称性暗示添加和删除操作之间存在一定的关联性. 事实上,我们观察图 8(b)中连续两个更新量 s_l 和 s_{l+1} 的关系,会发现无论是正字节更新还是负字节更新,总有相当部分的点落在直线 $y = -x$ 上—这些点对应的操作是在第 l 个版本时做了 s_l 的更新,紧接着第 $l+1$ 次做了 $s_{l+1} = -s_l$ 的更新,我们将这样的第 $l+1$ 次操作称为针对上一次的"反向更新"操作. 当 s_l 较小时,反向更新有可能与前一次完全无关. 例如,如果上一次添加了"is"这个单词,这次删除的可能是"be",但当 s_l 很大时,前后两次更新量相反的操作往往表明后一次操作就是针对前一次做的版本恢复操作. 用户为什么经常采取这样的操作呢? 可能的原因是很多用户钟情于自己的版本,认为已臻完善,一旦有其他用户在这个版本上进行修改,他立刻会改回来. 这就是维基百科特有的"看门狗"(watching dog)现象. 另外,有些用户的更新是不负责任的,甚至完全是

试验性质的,这种修改也会很快被改回来.

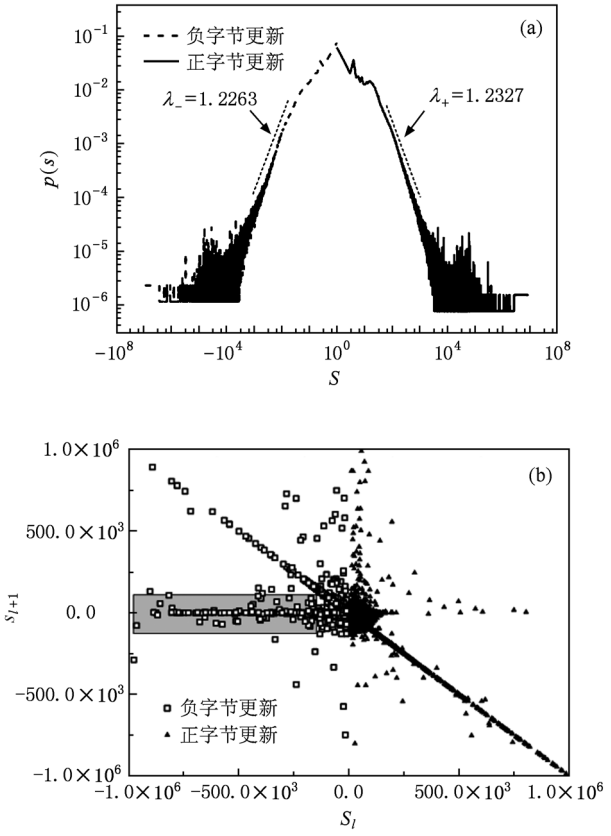

图 8 (a) 文章内容更新量 S 的频率分布图;(b) 连续两次更新量 s_l 和 s_{l+1} 的关系

为了测量哪些内容更容易被反向更新,我们考察在不同更新量 s 下反向更新的比例

$$P_{\mathrm{R}}(s) = \frac{N(-s \mid s)}{N(s)}. \tag{4}$$

其中 $N(s)$ 表示这次更新量为 s 的数量; $N(-s \mid s)$ 表示在这次更新量为 s 的前提下, 下一次更新量为 $-s$ 的次数. 如图 9 所示,在更新量 s 较小时(图中阴影部分), $P_{\mathrm{R}}(s)$ 都小于 40%,正字节更新的 P_{R} 要大于负字节更新的 P_{R}, 而在更新量 s 很大时,反向更新比例 $P_{\mathrm{R}}(s)$ 随 $|s|$ 的增大而增长,甚至达到了 100%,这点容易理解,因为大字节更新时,无论是添加还是删除,都极易被用户识别出来并当成是破坏性的操作,从而快速通过版本恢复功能修正回去. 最后,添加 1000 字节左右的更新和删除 100 字节左右的更新对应的 P_{R} 值最低,说明这两种情况下相应内容被保留的概率最大(超过 90%).

反向更新不仅仅只针对上一次,有些隐蔽的错误要间隔多步之后才可能被发现(或者当用户希望恢复某种更新时已经有若干其他更新间插其中). 我们计算在不同版本间隔 D 下反向更新比例 $\psi_{\mathrm{R}}(D)$ 来衡量这种效应

$$\psi_{\mathrm{R}}(D) = \frac{\sum_{|s| \geqslant \delta} N(D, -s \mid s)}{\sum_{|s| \geqslant \delta} N(s)}. \tag{5}$$

其中 $N(s)$ 表示这次更新量为 s 的数量; $N(D, -s \mid s)$ 表示在这次更新量为 s 的情况下, D 次后更新量为 $-s$ 的次数; δ 取不同的阈值,能反映出更新量在不同范围的情况. 图 10 显示了反向更新比例 $\psi_{\mathrm{R}}(D)$ 与版本间隔 D 的关系,不难发现 $D = 1$ 时比例最高, D 取奇数时 $\psi_{\mathrm{R}}(D)$ 要高于 D 取偶数时的值, $\psi_{\mathrm{R}}(D)$ 随 D 震荡下降, δ 越大震荡越大. 这一现象能够用维基百科编辑中普遍存在的编辑战争(Edit War)[50] 来解释,即在协同写作过程中人们常常对某些内容有较大的争议,在编辑时你来我往地持续反向更新多步,这在大字节的更新情况下表现得非常明显.

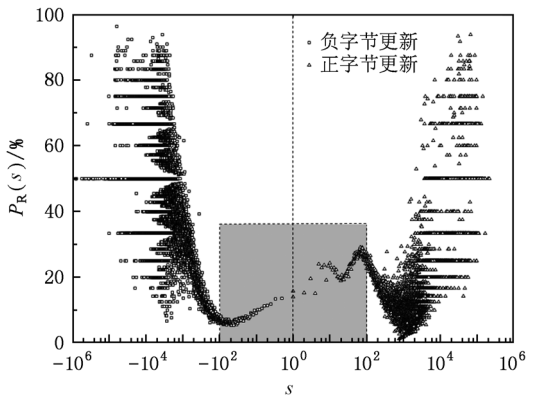

图 9 反向更新比例 $P_{\mathrm{R}}(s)$ 与更新量 s 的关系

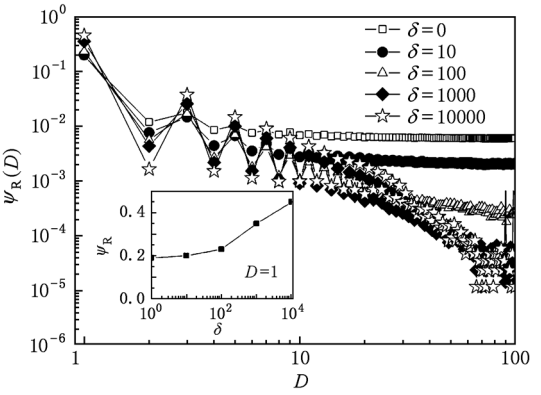

图 10 反向更新比例 $\psi_{\mathrm{R}}(D)$ 与版本间隔 D 的关系. 插图是 D = 1 时, ψ_{R} 与 δ 的关系

5. 结论与讨论

在第二次信息革命时代,人们广泛通过互联网进行多种形式的合作生产. 在线协同写作是其中一种十分普遍的人类集群行为. 本文采用维基百科中文章的编辑数据,对这类行为进行了实证探讨. 时间统计分析发现了在线协同写作特殊的多尺度分布,即 1 min 到 30 min 和 30 min 到 24 h 两个时间段上时间间隔分别服从指数为 1.62 和 1.16 的幂律分布,而大于 24 h 的时间间隔分布服从形如 $F(\tau) \propto \tau^{-b-a\log(\tau)}$ 的累积分布. 进一步分析发现不同时间尺度下对应的不同的起主导作用的提交行为是导致多尺度特征的主要原因. 对编辑内容更新量的分析发现反向更新行为是合作过程中普遍存在的现象,更新内容被保留的概率与更新量的大小有极强的关联性. 我们的研究支持了关于维基百科协同写作中存在"看门狗"和"编辑战争"的说法.

本文的研究厘清了网络协同写作行为的模式与特征,有助于了解网民是如何在虚拟环境下发挥民主、草根以及多元的主动性的. 我们发现的多尺度分布显示,用户在 30 min 内常常会对一篇文章做连续多次的编辑,而在 30 min 到一天时间内会同其他编辑者在线"讨论"共同编辑. 这一行为特点可能有助于指导网络软件设计、在线投放广告等方面的商业应用. 协同写作中的"看门狗"和"编辑战争"等现象说明尽管草根创作虽无监督机构,也无质量评价的标准,但创作并非随意,系统存在自组织的管理方式. 本文的研究还有助于推动可信计算研究. 根据文章编辑的活跃度、时间间隔编辑宽度、时间间隔幂指数等统计指标,建立对文章的质量以及编辑者的信誉进行评估的名声系统,有望解决长期存在的信誉评估问题. 最后,本文的研究对其他人类群体行为如短信[57,58]、博客[59]等的深入探讨有一定的借鉴作用.

在获得丰富实证结果的同时,也产生了以下几个亟待深层次研究的问题:

1. 如何对版本提交时间间隔多尺度分布特征进行建模. 尽管通过分析知道不同时间尺度下对应的不同提交模式是导致这种多尺度分布的主要原因,但对各模式的统计规律缺乏定量理解,因此缺乏反映各个尺度特征的模型. 比如从 1 min 到 30 min 中的连续提交行为,这种在短时间内的发生的事件显然不适合用优先级列表[6]、兴趣驱动[11]等机制来解释,我们尚不了解何种机制导致了这样的分布.

2. 版本提交时间间隔在天尺度下对应的累积分布形式为 $F(\tau) \propto \tau^{-b-a\log(\tau)}$, 对应的频率分布为 $p(\tau) \propto (b + 2a\log(\tau))\tau^{-b-1-a\log(\tau)}$. 这是一种有别于幂律分布[15]、广延指数分布[18,19]、带截断的幂律分布[20—22]、以及指数分布与幂律分布形成的双模态分布[14]等一系列分布的新分布形式. 这个新的分布形式是否有一定的物理意义,分布本身具有什么样的特殊性质,是否也适合用来描述其他人类行为特征,是值得我们进一步思考的问题.

3. 反向更新的普遍存在性,极有可能是导致版本提交行为高阵发性和高记忆性的原因. 因为反向更新行为,特别是针对大字节修改的反向更新行为,具有很高的可预测性. 如何利用反向更新来对协同写作进行预测,这对于从系统层面上维护维基百科内容,甚至进行网络舆情监测[60]将有重大的意义.

4. 如何理解反向更新在更新量较小时正负字节的不同表现(见图 9 的阴影部分).

[1] Gerald W 2003 *The Second Information Revolution* (Cambridge: Harvard University Press) p279

[2] Liu X H 2007 *The Success of the Internet Grass-Roots revolution in Web2.0 era* (Beijing: Tsinghua University Press) p10 (in Chinese) [刘向辉 2007 互联网草根革命 Web2.0 时代的成功方略 (北京:清华大学出版社) 第 10 页]

[3] Li N N, Zhou T, Zhang N 2008 *Complex Syst. & Complexity Sci.* **5** 2 (in Chinese) [李楠楠、周 涛、张 宁 2008 复杂系统与复杂性科学 **5** 2]

[4] Wang B H, Han X P 2008 *Physics* **39** 1 (in Chinese) [汪秉宏、韩筱璞 2010 物理 **39** 1]

[5] Han X P, Wang B H, Zhou T 2010 *Complex Syst. & Complexity Sci.* **5** 2 (in Chinsese) [韩筱璞、汪秉宏、周 涛 2010 复杂系统与复杂性科学 **5** 2]

[6] Barabási A L 2005 *Nature* **435** 207

[7] Mainardi F, Raberto M, Gorenflo R 2000 *Physica* A **287** 468

[8] Plerou V, Gopikrishnan P, Amaral N 2000 *Phys. Rev.* E **62** 3023

[9] Dezsö Z, Almaas E, Lukács A 2006 *Phys. Rev.* E **73** 066132

[10] Gonçalves B, Ramasco J J 2008 *Phys. Rev.* E **78** 026123

[11] Zhou T, Kiet H A T, Kim B J 2008 *Europhys. Lett.* **82** 28002

[12] Hu H B, Han D Y 2008 *Physica* A **387** 5916

[13] Hong W, Han X P, Zhou T, Wang B H 2009 *Chin. Phys. Lett.* **26** 028902

[14] Wu Y, Zhou C S, Xiao J H 2010 *Proc. Natl. Acad. Sci. U. S. A.* **107** 18803

[15] Henderson T, Bhatti S 2001 *in Proceedings of the ninth ACM international conference on Multimedia ACM: Ottawa*, Canada 212

[16] Grabowski A, Kruszewska N, Kosinacuteski R A 2008 *Phys. Rev.* E **78** 066110

[17] Cattuto C, Broeck W, Barrat A 2010 *PLoS ONE* **5** e11596

[18] Clauset A, Shalizi C R, Newman M E J 2009 *SIAM Rev.* **51** 661

[19] Goldstein M L, Morris S A, Yen G G 2004 *Eur. Phys. J.* B **41** 255

[20] Sornette D, Laherrère J 1998 *Eur. Phys. J.* B **2** 525

[21] Shang M S, Lü L Y, Zhang Y C 2001 *Europhys. Lett.* **90** 48006

[22] Zhou T, Wang B H, Jin Y D 2007 *Int. J. Mod. Phys.* C **2** 297

[23] Newman M E J 2000 *Proc. Natl. Acad. Sci. U. S. A.* **98** 405

[24] Boguñá M, Pastor S R, Vespignani A 2004 *Eur. Phys. J.* B **38** 205

[25] Vázquez A, Oliveria J G, Dezsö Z 2006 *Phys. Rev.* E **73** 036127

[26] Gabrielli A, Caldarelli G 2007 *Phys. Rev. Lett.* **98** 208701

[27] Min B, Goh K I, Kim I M 2009 *Phys. Rev.* E **79** 056110

[28] Blanchard P, Hongler M O 2007 *Phys. Rev.* E **75** 026102

[29] Guo J L 2010 *Acta Phys. Sin.* **59** 6(in Chinese)[郭进利 2010 物理学报 **59** 6]

[30] Han X P, Zhou T, Wang B H 2008 *New J. Phys.* **10** 073010

[31] Hidalgo R, César A 2006 *Physica* A **369** 877

[32] Jiang Z Q 2008 *Physica* A **387** 5818

[33] Malmgren D, Stoufer D B, Motter A E, Amaral L A N 2008 *Proc. Natl. Acad. Sci. U. S. A.* **105** 18153

[34] Vazquez A 2006 *Physica* A **373** 747

[35] Radicchi F 2009 *Phys. Rev.* E **80** 026118

[36] Oliveira J G, Vazquez A 2008 *Physica* A **388** 187

[37] Crane R, Sornette D 2008 *Proc. Natl. Acad. Sci. U. S. A.* **105** 15649

[38] Ratkiewicz J, Fortunato S, Flammini A, Menczer F, Vespignani A 2010 *Phys. Rev. Lett.* **105** 158701

[39] Chmiel A, Kowalska K, Holstrokyst J A 2009 *Phys. Rev.* E **80** 066122

[40] Zlatić V, Božičević M, Štefančić H, Domazet M 2006 *Phys. Rev.* E **74** 016115

[41] Muchnik L, Itzhak R, Solomon S, Louzon Y 2007 *Phys. Rev.* E **76** 016106

[42] Capocci A, Servedio V D P, Colaiori F, Buriol L S, Donata D, Leonardi S, Caldarelli G, 2006 *Phys. Rev.* E **74** 036116

[43] Capocci A, Rao F, Caldarelli G 2008 *Europhys. Lett.* **81** 28006

[44] Zhirov A O, Zhirov O V, Shepelyansky D L 2010 *Eur. Phys. J.* B **77** 523

[45] Zhao F, Zhou T, Zhang L, Ma M H, Liu J H, Yu F, Zha Y L, Li R Q 2010 *J. UESTC* **39** 3 (in Chinese)[赵 飞、周 涛、张良、马鸣卉、刘金虎、余 飞、查一龙、李瑞琪 2010 电子科技大学学报 **39** 3]

[46] Bauke H 2007 *Eur. Phys. J.* B **58** 167

[47] Wang P, Xie X Y, Yeung C H, Wang B H 2011 *Physica* A **390** 12

[48] Wang P, Lei T, Yeung C H, Wang B H 2011 *Europhys. Lett.* **94** 18005

[49] Goh K I, Barabási A L 2008 *Europhys. Lett.* **81** 48002

[50] Krötzsch M D, Vrandecić M, Völkel 2006 *in The Semantic Web - ISWC* 2006 Springer Berlin Heidelberg 935

[51] Chernov S, Iofciu T, Nejdl W, Zhou X 2006 *in Proceedings of the First Workshop on Semantic Wikis—From Wiki To Semantics* ESWC 2006

[52] Yeh E, Ramage D, Manning C D, Agirre E, Soroa A 2009 *in TextGraphs-4: Proceedings of the* 2009 *Workshop on Graph-based Methods for Natural Language Processing* 41

[53] Strube M, Ponzetto S P 2006 *Proceedings of the Twenty-First National Conference on Artificial Intelligence*

[54] Holloway T, Božievi M, Börner K 2007 *Special Issue on Understanding Complex Systems* **12** 30

[55] Suh B, Chi E H, Pendleton B, Kittur A 2007 *IEEE Symposium on Visual Analytics Science and Technology Sacramento*

[56] Broughton J 2008 *Wikipedia: the Missing manual* p182 ISBN: 10:0-596-51616-2

[57] Wu Y, Xiao J H, Wu Z Y, Yang J Z 2007 *Acta Phys. Sin.* **56** 4 (in Chinese)[吴 晔、肖井华、吴智远、杨俊忠 2007 物理学报 **56** 4]

[58] Li M J, Wu Y, Liu W Q, Xiao J H 2009 *Acta Phys. Sin.* **58** 8 (in Chinese)[李明杰、吴 晔、刘维清、肖井华 2009 物理学报 **58** 8]

[59] Xiong F, Liu Y, Si X M, Ding F 2010 *Acta Phys. Sin.* **59** 10 (in Chinese)[熊 菲、刘 云、司夏萌、丁 飞 2010 物理学报 **59** 10]

[60] Liu Y 2007 *Introduction to Network Option Research* (Tianjing: Tianjing Renmin Press) p84 (in Chinese) [刘毅 2007 网络舆情研究概论 (天津:天津人民出版社) p84]

Human dynamics analysis in online collaborative writing*

Zhao Fei[1)2)] Liu Jin-Hu[1)3)] Zha Yi-Long[1)4)] Zhou Tao[1)†]

1) (*Web Sciences Center, University of Electronic Science and Technology of China, Chengdu* 610054, *China*)

2) (*School of Economy and Management, University of Electronic Science and Technology of China, Chengdu* 610054, *China*)

3) (*School of Applied Mathematics, University of Electronic Science and Technology of China, Chengdu* 610054, *China*)

4) (*Experimental Class of International Software Professionals, University of Electronic Science and Technology of China, Chengdu* 610054, *China*)

(Received 22 January 2011; revised manuscript received 10 March 2011)

Abstract

Investigating the human online behavior has become a central issue for understanding human dynamics in recent years. In this paper we analyze the temporal and content-updating statistical properties of online collaborative writing based on Wikipedia data. Online collaborative writing is one of the important and widespread human online behaviors, which is of great apphication. Empirical result shows that the distribution of inter-event time in collaborative writing is on the multi-scale. That is to say, two time intervals that range from 1 min to 30 min and 30 min to 24 h both obey power-law distribution with exponents equal to 1.62 and 1.16 respectively, while the interval larger than 24 h obeys a distribution whose cumulative form is $F(\tau) \propto \tau^{-b-a\log(\tau)}$. More investigatons show successive updating behavior and mutual updating behavior working together to lead to the multi-scale distribution of inter-event time. Successive updating behavior leads to the power-law distribution with an exponent 1.62 of interval within 30 min while mutual updating behavior leads to the power-law distribution with an exponent 1.16 of interval ranging from 30 min to 24 h. Furthermore, we find that reverse updating repeats frequently in collaborative writing. The proportions of reversing updating and the updating size are strongly relatively reflect that the updating size is a main reason leading to the relevant content to be preserved. The bigger the updating size, the harder it would be preserved. More statistical analyses imply that "watching dog" and "edit war" exist in Wikipedia editing. Those results are very helpful to deepen the understanding of the human collective behavior, especially of the collaborative developing behavior.

Keywords: online collaborative writing, human dynamics, multi-scale property, Wikipedia

PACS: 89.75.Da, 05.45.Tp, 02.50.Ey

* Project supported by the Key Program of the National Natural Science Foundation of China (Grant No. 10635040) and the Major Research Plan of the National Natural Science Foundation of China (Grant Nos. 70871082, 10975126, 70971089).

† Corresponding author. E-mail: zhutou@ustc.edu