计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20148336
52(4):779-788,2015

# 基于小数据的在线用户兴趣长程演化研究

李 勇[1,2] 孟小峰[1] 刘 继[3] 王常青[4]
[1](中国人民大学信息学院 北京 100872)
[2](西北师范大学计算机科学与工程学院 兰州 730070)
[3](新疆财经大学统计与信息学院 乌鲁木齐 830012)
[4](中国互联网络信息中心互联网基础技术开放实验室 北京 100190)
(facingworld@126.com)

# Study of The Long-Range Evolution of Online Human-Interest Based on Small Data

Li Yong[1,2], Meng Xiaofeng[1], Liu Ji[3], and Wang Changqing[4]
[1](*School of Information, Renmin University of China, Beijing* 100872)
[2](*College of Computer Science and Engineering, Northwest Normal University, Lanzhou* 730070)
[3](*School of Statistics and Information, Xinjiang University of Finance and Economics, Urumqi* 830012)
[4](*DNSLAB, China Internet Network Information Center, Beijing* 100190)

**Abstract** The availability of network big data, such as those from online human surfing log, e-commerce and communication log, makes it possible to probe into and quantify the dynamics of human-interest. These online behavioral data is called "small data" in the era of big data, which can help explaining many complex socio-economic phenomena. A fundamental assumption of Web user behavioral modeling is that the user's behavior is consistent with the Markov process and the user's next behavior only depends on his current behavior regardless of the historical behaviors of the past. However, Web user's behavior is a complex process and often driven by human interests. We know little about regular pattern of human-interest. In this paper, using more than 30 000 online users behavioral log dataset from CNNIC, we explore the use of block entropy as a dynamics classifier for human-interest behaviors. We synthesize several entropy-based approaches to apply information theoretic measures of randomness and memory to the stochastic and deterministic processes of human-interests by using discrete derivatives and integrals of the entropy growth curve. Our results are, however preliminary, that the Web user's behavior is not a Markov process, but a aperiodic infinitary long-range memory power-law process. Further analysis finds that the predictability gain can exceed 95.3 percent when users click 7 consecutive points online, which can provide theoretical guidance for accurate prediction of online user's interests in the era of big data.

**Key words** small data; block entropy; excess entropy; evolution of interest; predictability gain

**摘 要** 网络大数据中与Web用户行为相关的数据,例如在线点击数据和通讯记录等,为人们深度挖掘和定量分析人类兴趣动力学带来了机遇,这些在线行为数据被称为大数据时代的"小数据",有助于揭示

收稿日期:2014-12-09;修回日期:2015-02-27
基金项目:国家自然科学基金项目(61379050,91224008,71261025);国家"八六三"高技术研究发展计划基金项目(2013AA013204);高等学校博士学科点专项科研基金项目(20130004130001);中国人民大学科学研究基金项目(11XNL010)
通信作者:孟小峰(xfmeng@ruc.edu.cn)

许多复杂的人类社会与经济现象. Web 用户行为建模时常见的前提假设就是人的行为符合 Markov 过程,用户下一行为仅依赖于当前行为,与过去的历史行为无关. 然而,在线用户行为是一个复杂过程,常常依赖于人的兴趣,对于人类兴趣动力学的本质规律目前知之甚少. 利用中国互联网络信息中心提供的 30 000 多名在线用户行为记录数据,基于块熵理论对在线用户行为进行分类研究,通过信息论分析方法,结合熵增曲线的离散导数和积分理论,分析在线用户点击行为的随机性和记忆性特征. 研究表明,与常见的假设不同,Web 用户的行为并不是一个简单的 Markov 过程,而是一个符合幂率的非周期无限长程记忆过程;进一步还发现,用户在线连续点击 7 个兴趣点,其行为的平均预测增益就可达到 95.3%以上,可为大数据时代在线用户兴趣精准预测提供理论指导.

关键词 小数据;块熵;超熵;兴趣演化;预测增益

中图法分类号 TP393.4; TP202$^{+}$.4; TP311.13

互联网是人类最伟大的发明之一,已成为影响经济社会发展、改变人类文明形态的重要载体. 在线查询、浏览、标记、购物、娱乐等行为已成为在线用户最重要的生活常态. 人的许多行为由兴趣所驱动,兴趣随时间在不断变化,有些兴趣伴其一生,有些兴趣只能持续短暂的时间,在线用户的许多行为都被网络日志详细地记录下来,为分析人类兴趣演化规律提供了可能. 兴趣的演化规律在商业、医学、群体事件预防等领域有着广泛的应用前景,对用户兴趣的了解可以促进定向广告的设计和产品精准营销,了解精神病患者的兴趣变化有助于做到准确的诊断和治疗,对个人及群体的兴趣了解有助于预测非常规突发事件.

2013 年神经信息处理系统国际会议上(NIPS 2013),美国康奈尔大学 Estrin 教授作了题为"Small, $n$=me, Data"的报告,她指出用户上网和使用各种移动设备过程中产生了大量用户行为轨迹数据,从这些广泛并具有个性化的数据中提取个体数据,为揭示人类行为模式规律提供了可能. 这些在线行为数据被称为大数据时代的"小数据"[1],基于小数据分析人类行为规律已逐渐成为当前研究的热点,涌现出了社会计算[2]、计算社会科学[3]等跨学科研究.

Web 用户建模时一个常见的前提假设就是用户的行为符合 Markov 过程,即用户下一行为仅依赖于当前行为,与过去的历史行为无关[4]. 这一假设是 Web 页面 PageRank 算法以及在线广告推送算法的基石. 人类在线行为是一个复杂过程,伴随着确定性和随机性 2 种过程,由于记忆的影响,过去的行为对未来的行为有着较大的影响,将人的行为看作完全随机过程显然不恰当,但人类兴趣的演化有没有统一的模式?演化规律是怎样的一个过程?这是本文将要探讨的问题.

根据热力学第二定律,信息、物质、能量是主导生命过程的核心要素[5],近年来从宇宙学到计算机科学的研究都显示出智力与熵最大化之间有某种深层次的关联[6]. 信息与负熵相当,信息的失去由负熵的增加所补偿,因此使系统的熵减少,若要不做功而使系统熵减少,就必须获得信息,吸收外界的负熵. 物理学家薛定谔在《生命是什么?》一书中指出[7],生命赖负熵为生,一个生命有机体在不断地增加着它的熵,并趋于接近最大熵的危险状态——死亡状态,要摆脱死亡,只能从环境中不断汲取负熵,有机体就是赖负熵即信息为生.

受此启发,本文基于熵理论,根据中国互联网络信息中心提供的 30 000 多名志愿者用户在线点击行为数据,通过计算行为变化的块熵(block entropy)、超熵(excess entropy)、熵增等量,分析用户兴趣点在特定时间段内演化的过程,从而揭示人类兴趣变化的潜在规律. 主要贡献如下:

1) 对 Web 用户的在线点击行为进行分析建模;

2) 根据块熵、熵率、超熵等理论,对在线点击行为数据进行定量分析,应用随机过程理论进行对比分析并确定用户的兴趣演化规律;

3) 实验分析表明,用户在线行为既不符合 Markov 过程,也不同于随机游走过程,而是一个非周期无限长程记忆过程;

4) 通过预测增益定量分析,确定在线行为精准预测所需的连续点击数据量的理论下界,并在大量真实数据实验的基础上,验证了兴趣行为规律的普遍性.

## 1 相关工作

人类行为的研究已有很长的历史,被认为是经

济学的基础[8],但由于缺乏定量分析,这方面的研究一直未能引起学术界的广泛认可. 2005 年,《Nature》上的一篇论文揭示了人类行为在时间上对泊松过程的偏离,提出了一个基于任务优先级的排队论模型[9];此后又有研究发现人类行为在空间上的标度律,揭示了人类在空间上的行为也不同于随机游走[10]. 受这 2 篇开创性研究的影响,大量文章出现在《Nature》,《Science》,《PNAS》等期刊,掀起了人类行为动力学研究的热潮,提出了采用多个模型对人类行为的重尾分布进行解释,例如泊松概率模型[11-12]、兴趣变化模型[13]、记忆效应和人际交互模型[14-15]等.

近年来,计算机领域的研究人员基于小数据对人类兴趣的研究主要针对行为定向(behavioral targeting)和兴趣点挖掘[16-18],通过用户的历史行为数据分析在线行为模式,应用机器学习理论预测用户潜在的兴趣和需求,从而有针对性地推荐产品或投放广告. 虽然这些研究推动了在线企业的发展,带来了可观的经济效益,但对人类行为本质规律探索的贡献却很小.

人类兴趣动力学作为人类行为动力学的一部分还鲜有研究. 文献[19]是第一个对人类兴趣动力学所做的系统研究工作,基于豆瓣、淘宝、移动手机阅读 3 种数据,分析了人类兴趣的时间统计分布以及兴趣的转变规律,认为喜好返回、惯性效应、探究新兴趣点这 3 个因素是人类兴趣动力学的最基本特征. 该研究的出发点是经验性地假定人类兴趣演化是不完全随机的,并通过一个有偏随机游走模型对兴趣的演化进行建模,但对于人类兴趣变化为什么是不完全随机的以及使用的建模方法的依据未能给出令人信服的分析论证.

熵理论常用于时间序列数据的分析,文献[20]用块熵定量分析了气象数据,证明降雨时间序列数据是一个有限阶的 Markov 过程. 文献[21]通过理论分析与实验验证,证明熵理论在数据分析中具有简单性、鲁棒性、稳定性、快速计算等特点. 文献[22]系统地介绍了熵、块熵、熵率、超熵、预测增益等理论,并提出了暂态信息概念,基于这些理论分析了独立同分布过程、周期性过程、Markov 过程等,证明熵理论是分析并区分不同随机过程的强大理论工具,这为本文定量分析人类在线兴趣行为动力学奠定了理论基础.

## 2　数据及问题描述

本文采用中国互联网络信息中心提供的用户在线行为数据①,该数据由 30 000 多名志愿者用户在个人计算机上安装数据采集程序在线获取. 在保证在线用户个人隐私的前提下,数据采集程序以每 2 s 一次的频率扫描用户计算机的当前焦点窗口,若焦点窗口发生变化,则会在日志中追加一条记录,详细记录了用户开关机时间、窗口进程名、浏览器地址栏内容(已部分截断)、焦点窗口对应的程序版本号、程序所属公司名以及用户属性等信息.

该数据集已累积了数以 TB 量级的数据,为分析方便,本文首先随机抽取 1 000 个样本用户一个月内约 1.2 亿条数据记录. 假定每个用户在这一个月的短期行为具有稳定性,分析并得出结论,然后将结论推广到所有用户以及所有时间进行验证,以证实这种行为规律的普遍性.

个性化推荐研究中对用户在线兴趣作了比较细致的分析和研究[23-24],用户在线很多行为都能暗示其喜好,如查询、浏览页面、标记书签、反馈信息、点击鼠标等,用户访问时的停留时间、访问次数等动作也能揭示用户兴趣. 本文将用户兴趣定位于点击鼠标、停留时间和访问次数等行为上.

对在线点击数据统计分析发现,用户在长时间的停歇状态后会产生密集的点击行为,之后又是长时间的停歇状态,表现出很强的随机性和阵发性,如图 1(a)所示;同时也表现出很强的幂率特性,短时间的密集点击行为所占时间比例较小,长时间的停歇状态呈现出长尾特征,如图 1(b)所示,幂指数 $\alpha \approx -1.7$,图 1(b)中嵌入的小图是点击时间间隔的概率在双对数坐标下的分布图.

用户点击行为展现出一定程度的随机性,主要有以下 3 方面原因:1)人们还不知道该行为过程的规则,一个观察者只拥有一个行为过程的不完全知识;2)可能有一种机制放大了宏观行为波动性过程的不可预测性;3)存在着多种多样的明显的观察所引起的随机性[22].

为分析方便,首先将用户在线点击行为时间序列看作一个随机过程,根据随机过程已有理论进行对比分析,例如周期性过程、Markov 过程、$R$ 阶 Markov 过程等,以准确界定在线用户兴趣行为的演化规律.

① http://h.cnnicresearch.cn/

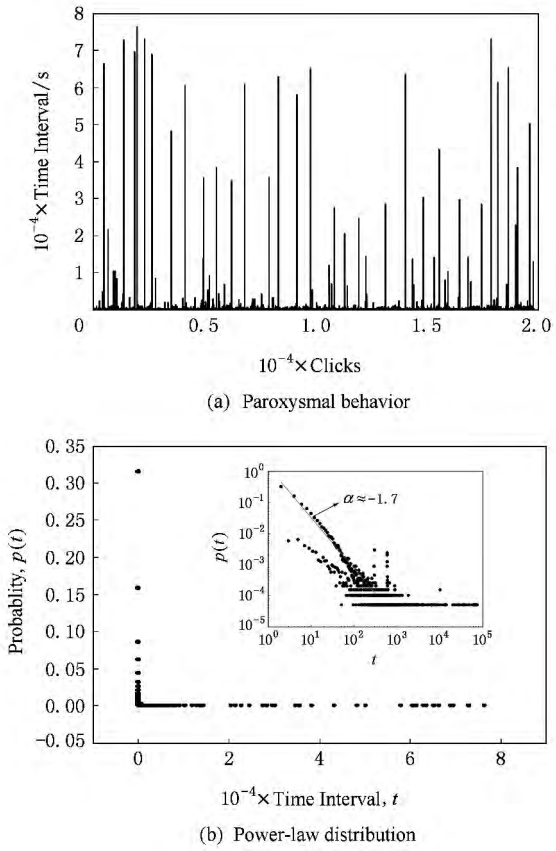



Fig. 1 Distribution of paroxysmal behavior of online human-interest.

图 1 在线用户兴趣行为阵发性统计分布

## 3 理论与方法

信息论由 Shannon 于 1948 年提出,基于 Boltzmann 的热力学熵理论[25],最初用于通信信道的分析建模. 信息论可用于分析随机变量的结构属性,也为比较离散变量及连续变量的概率量提供了可能. 首先假定在线点击行为是一个随机过程,根据熵理论分析随机过程的结构特征,根据连续兴趣点随机变量的块熵分析其随机性和记忆性特征,根据熵率、超熵、离散导数与积分等方法,分析随时间演化过程中时间序列数据表现出的随机性和有序性信息,从而定量分析在线兴趣的演化规律.

### 3.1 与在线行为相关的随机变量定义

设随机过程$\{X(t), t\in T, t\geqslant 0\}$,其中 $t$ 表示时间,$X(t)$是一个随机变量,如果 $T$ 是一个可数集,则称 $X$ 为一个离散时间的随机过程. 随机变量 $X(t)$ 表示在时间 $t$ 的兴趣点.

定义 1. 随机变量块. 设 $L$ 个时间上连续的兴趣点表示为 $X^L\equiv X_1\cdots X_L$,其中大写字母 $X$ 表示随机过程,用小写字母 $x$ 表示随机变量的特定值,用 $x^L\equiv x_1\cdots x_L$ 表示长度为 $L$ 的连续兴趣点形成的随机变量块.

定义 2. 随机变量的词频. 设用户在特定时间段内的兴趣数量有限,则随机变量 $X(t)$的取值范围可由一个有限的字符集 $A$ 表示. 根据符号动力学理论[26],将随机过程 $X$ 表示为一个按时间递增排列的符号序列 $\zeta=X(t_0),X(t_1),\cdots X(t_l)$,设字符集 $A$ 的长度为$|A|$,$|A|$表示时间段 $T$ 内不同兴趣点的总量. 根据符号序列中的特定词频可统计随机过程中随机变量的变化规律,这里的词是指一个长度为 $L$ 的连续兴趣点形成的随机变量块.

### 3.2 块 熵

设随机过程 $X$ 有 $n$ 个离散状态:$\{x_1, x_2\cdots, x_n\}$,表示在线用户在连续时间段 $T$ 内 $n$ 个不同的兴趣点,每个兴趣点对应的概率为 $p\equiv\{p_1, p_2\cdots, p_n\}$,满足条件 $\forall i\in\{1,2,\cdots,n\}$,有 $0\leqslant p_i\leqslant 1$,且 $\sum_{i=1}^{n} p_i=1$,则 $X$ 的信息熵定义为

$$H(X)=-\sum_{x\in X}p(x)\,\mathrm{lb}\,p(x). \tag{1}$$

根据 3.1 节随机变量块及词频的定义,将随机过程 $X$ 符号化后,依据式(1)可计算 $L$ 个连续兴趣点组成的块熵,如式(2)所示. 其中,$L=\{1,2,\cdots, Max\}$,$Max$ 表示选取的最大块长,$L$ 由 4.3 节所述原则确定.

$$H(L)\equiv-\sum_{x^L\in X}p(x^L)\,\mathrm{lb}\,p(x^L). \tag{2}$$

信息熵 $H(X)$用来度量一个随机过程 $X$ 的不确定性,同时也可度量随机过程的信息平均保留量. 块熵 $H(L)$依据信息熵原理,将连续 $L$ 个兴趣点视为一个词,统计点击序列中所有长度为 $L$ 的词频并计算信息熵,将连续的兴趣点看作一个块,通过计算块熵可以分析用户点击序列中的随机性和记忆性等特征.

根据块熵原理,随着块长 $L$ 的不断增加产生的块熵的变化可得到兴趣演化的熵增曲线,熵增曲线是一个单调非减曲线,如图 2 所示. 块熵的算法如算法 1 所示.

算法 1. 在线点击序列的块熵算法.

输入:用户点击的时间序列的符号表示:*clicklist*;

输出:点击序列块长从 1 到 $Max$ 的块熵.

① $Pr=Counter(clicklist)$; /* 统计单个符号的出现频数 */

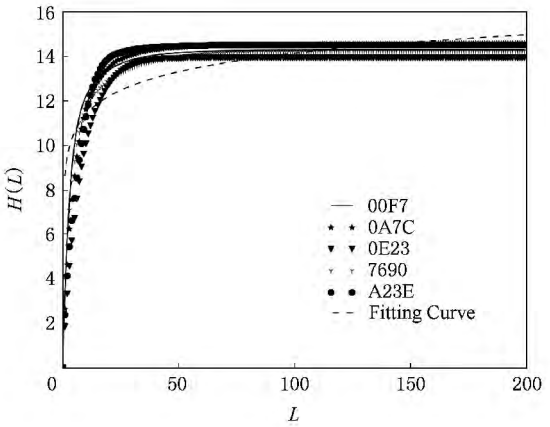



Fig. 2　Block entropy curves.

**图 2　在线用户兴趣行为的块熵变化曲线**

② $HL=\{\}$; $HL[0]=0$; $HL[1]=Entropy(Pr)$;
③ for($i=2$; $i<Max$; $i++$)
④　$P=\{\}$; $m=length(clicklist)-i$; $j=0$;
⑤　while ($j<m$)
⑥　　$a=0$; $str1=$''';
⑦　　while ($a<i$)
⑧　　$str1=str1+clicklist[a+j]$;
⑨　　$a+=1$;
⑩　　endwhile
⑪　　if $str1$ in $p.keys()$ /* 已在字典中的块 */
⑫　　$P[str1]=p[str1]+1$;
⑬　　else /* 不在字典中的块 */
⑭　　$P[str1]=1$;
⑮　　endif
⑯　　$j+=1$;
⑰　endwhile
⑱　$Pr=Counter(P)$; /* $Pr$ 是字典,$key$ 为块,$value$ 为频数 */
⑲　$HL[i]=Entropy(Pr)$; /* 计算块熵 */
⑳ endfor
㉑ return $HL$.

算法 1 中,在线点击序列 $\zeta$ 中长度为 $L$($L\ll\ell$)的词可用以下方式统计:用一个长度为 $L$ 的滑动窗口每次滑动一个符号穿过序列 $\zeta$,每滑动一次就得到一个连续兴趣点的符号序列构成的词,共可得到 $\ell-L+1$ 个词(对应算法 1 中行③～行⑰).每个词出现的频率由 $p_i=n_i/(\ell-L+1)$ 计算得到(通过算法 1 中的计数器函数 $Counter()$ 计算),根据词频即可得到块熵(熵由 $Entropy()$ 函数计算,$Entropy()$ 函数的算法可由式(2)计算得到,本文因篇幅所限略去).

根据块熵定义,长度为 0 的符号串块熵为 0,长度为 1 的块熵由单个符号出现的频次计算(对应算法 1 中行②).

### 3.3　熵　率

给定一个块长为 $L$ 的随机变量序列,序列的熵随块长 $L$ 的增加而增加,将这个增长率称之为熵率,定义为[22]

$$h\equiv\lim_{L\to\infty}\frac{H(L)}{L}. \tag{3}$$

熵率表示当连续兴趣点形成的块长 $L$ 增大到极限的情况下熵的增加率.熵率用以刻画一个随机过程的随机性和有序性,可用来度量在线点击序列中不能减少的随机性,随着数据块越来越长,其相关性和结构被加以考虑而使随机性的程度增加.

### 3.4　离散导数与积分

对于任一函数 $F:Z\to R$,定义其离散导数为[22]

$$(\Delta F)(x)\equiv F(x)-F(x-1), \tag{4}$$

其中,$x>1$.

高阶离散导数为 $\Delta^{(n)}F\equiv\Delta\circ\Delta^{(n-1)}F(x)$.

定义离散函数 $\Delta F(x)$ 的离散积分为

$$\sum_{x=a}^{b}\Delta F(x)=F(b)-F(a-1). \tag{5}$$

根据离散导数定义,块熵 $H(L)$ 的一阶离散导数 $\Delta H(L)=H(L)-H(L-1)$,二阶离散导数 $\Delta^{(2)}H(L)=\Delta H(L)-\Delta H(L-1)$,其中 $L>0$.设 $\Delta H(0)=\mathrm{lb}\,|A|$,又因为 $H(0)=0$,所以 $\Delta H(1)=H(1)$.

块熵 $H(L)$ 的离散导数和积分可用于度量在线兴趣点序列的复杂性、可预测性以及记忆性等特征.一阶离散导数 $\Delta H(L)$ 为块熵增益,也称为信息增益,可以估量兴趣点序列长度由 $L-1$ 增加到 $L$ 时有多少量的更多信息被展现出来,也可看作在长度 $L$ 范围内近似的熵率 $h$.二阶离散导数 $\Delta^{(2)}H(L)$ 称为预测增益,可以估量连续的兴趣点序列长度由 $L-1$ 增加到 $L$ 时熵率有多快接近其渐近值.

### 3.5　块熵的积分量

超熵、总体可预测性(global predictability)以及暂态信息(transient information)是度量随机过程的 3 个常用的离散积分量,其定义如表 1 所示.

超熵又称为"预测信息"、"记忆信息"、"复杂性"等[22],用来度量在线兴趣行为的历史信息在当前保留的成份以及对未来的影响.有学者认为超熵无穷大即为长程关联.

Table 1 Three Integral Quantities of Entropy Rates

表 1 3 个有关熵率的积分量

| Quantity | Definition |
|---|---|
| Excess Entropy $E$ | $\sum_{L=1}^{\infty}[\Delta H(L)-h]$ |
| Global Predictability $G$ | $\sum_{L=1}^{\infty}\Delta^2 H(L)$ |
| Transient Information $T$ | $\sum_{L=0}^{\infty}[E+hL-H(L)]$ |

总体可预测性可以估量在线点击行为序列的非随机行为,满足条件 $\mathrm{lb}|A|=|G|+h$,其中 $\mathrm{lb}|A|$ 表示只知道被测量点击行为序列的总长度,对数据其他方面信息不了解的情况下序列的最大熵;熵率 $h$ 表示在已知数据分布情况下的熵,用以度量兴趣序列的不可预测性. 如图 6 中嵌入的小图所示,总体可预测性是 $\Delta^{(2)}H(L)$ 曲线与纵轴、横轴形成的阴影部分的面积.

暂态信息度量兴趣行为序列的结构属性,这是超熵无法做到的. 如果在线兴趣行为是一个混沌过程,既包含周期性成份,又包含随机性成份,则通过暂态信息可检测到其同步有序成份.

本文主要应用熵率、超熵和暂态信息 3 个量分析在线用户兴趣行为点击数据,根据已有研究结论,这 3 个量已足以区分已知的多个随机过程,如表 2 所示:

Table 2 Sample Values about Entropy Rates of Some Processes

表 2 一些常见随机过程的熵率及相关值

| Process Type | $h$ | $E$ | $T$ |
|---|---|---|---|
| I. I. D. : | | | |
| Fair Coin $p=0.5$ | 1 | 0 | 0 |
| Biased Coin $p=0.7$ | 0.8813 | 0 | 0 |
| Period-16 | 0 | $\mathrm{lb}\,16$ | 16.6135 |
| Order-R-Markov | $>0$ | $H(R)-Rh$ | $T=E$ |
| Thue-Morse | 0 | $\infty\mathrm{lb}\,L$ | $\infty\ L$ |
| Finitary(Exponential-Decay) | $\Delta H(L)-c2^{-\gamma L}$ | $\frac{H(1)-h}{1-2^{-\lambda}}$ | $\frac{H(1)-h}{(1-2^{-\lambda})^2}$ |
| Aperiodic Infinitary | $\Delta H(L)-L^{-\alpha}$ | $c_1\,\mathrm{lb}\,L+c_2$ | $\sum_{L=0}^{\infty}[E+hL-H(L)]$ |

## 4 在线用户兴趣演化分析

### 4.1 兴趣的演化规律

表 3 记录了第 2 节描述的数据集中 5 名样本用户在一个月内兴趣点的总数 $\ell$ 以及兴趣演化的熵率 $h$、超熵 $E$、暂态信息 $T$ 的分析值. 所有 30 000 多名在线用户的熵率、超熵分析结果与 5 名样本用户的结果相近.

图 2 是 5 名样本用户的熵增曲线 $H(L)$ 的演化情况,尽管他们在性别、年龄、地域等人口属性方面截然不同,但其兴趣演化的熵增曲线以及超熵值却接近一致性,这也证实了人作为高级生物具有“生物是物,生物有理”的规律.

从图 2 可看出,块熵是一个非减上凸曲线,块长 $L=11$ 是块熵的相变点;当 $L\leqslant 11$ 时,块熵增长非常迅速;当 $L>11$ 时,块熵在 $14\pm 2$ 范围内平稳缓慢渐增. 块熵曲线近似服从式(6)所示的拟和曲线,由图 2 中的虚线所示,其中 $x$ 表示块长度,$y$ 表示块熵.

$$y=1.2\ln x+8.58. \tag{6}$$

这说明在线用户的兴趣变化具有一定的随机性,但同时又是一个长程记忆过程,当兴趣量达到一定数量时,对新兴趣点的探索趋近饱和. 将表 3 的实验数据与表 2 的已有研究结论对比可发现,在线用户兴趣行为与 Markov 过程差别较大.

Table 3 Descriptions and Results for Human Interests

表 3 人类兴趣行为分析结果

| ID | Sex | Age | Province | $\ell$ | $h$ | $E$ | $T$ |
|---|---|---|---|---|---|---|---|
| 00F7 | F | 24 | Jilin | 19 859 | 0 | 14.26 | 88.84 |
| 0E23 | F | 26 | Yunnan | 15 700 | 0 | 13.92 | 114.75 |
| A23E | M | 38 | Guangdong | 23 498 | 0 | 14.50 | 102.01 |
| 7690 | M | 50 | Gansu | 19 813 | 0 | 14.25 | 112.18 |
| 0A7C | F | 60 | Shanghai | 24 790 | 0 | 14.57 | 114.12 |

暂态信息 $T$ 度量在线用户的兴趣变化过程的结构属性,估量与一个过程进行同步时的难度大小,

即需要多少信息量才能达到其渐近形式.一个过程有较大的 $T$ 值,其内部状态的不确定性也较高.从表 3 可以看出,5 名样本用户尽管在行为变化的熵增曲线 $H(L)$以及超熵 $E$ 上具有一致性,但每个人的暂态信息 $T$ 值却差别较大,符合"人人相似,人人不同"的客观现实.在线用户行为尽管相似,但每个人的点击过程却各不相同,根据暂态信息 $T$ 值即可将每个人的不同特征区分开来.

图 3 展示了熵率演化曲线 $h(L)$与块熵增益 $\Delta H(L)$的演化曲线,嵌入的小图是将纵轴与横轴接近 0 的部分放大后的结果.熵率度量一个过程中不可减少的随机性,显示随着块长度 $L$ 的逐渐增大,序列中的相关性和结构的随机性在序列中的变化程度.一个过程有较高的随机性则熵率较大,较小的熵率说明过程中各行为之间的相关性较强.在线用户点击行为的熵率随 $L$ 的增加其极限趋近于 0,即$h=\lim\limits_{L\to\infty} h(L)=0$,说明兴趣演化过程的随机性很小、规律性很强.这表明,在经济学、心理学等学科的科学研究中预先假设人的行为具有规律性,进而通过实验抽样验证其最基本的前提假设和研究范式是可信的.

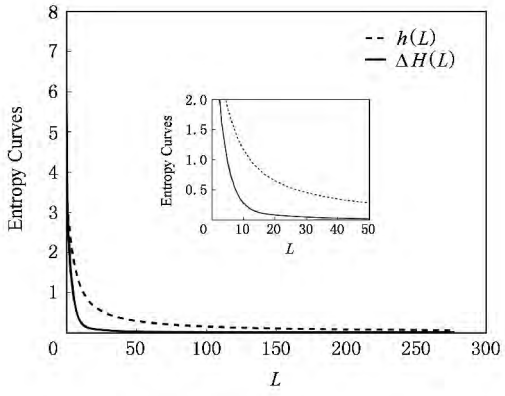



Fig. 3　Entropy curves for $h(L)$ and $\Delta H(L)$.

图 3　熵率与块熵增益曲线

从图 3 可以看出,块熵增益 $\Delta H(L)$与熵率变化曲线 $h(L)$都近似服从幂率形式,块熵增益与熵率及块长之间服从式(7)所示的关系:

$$\Delta H(L)-h=\Delta H(L)\approx L^{-\alpha}. \tag{7}$$

其中,$\alpha$ 值在 1.3～2.4 之间,与人类其他行为的研究结论基本一致,例如人类撰写书籍中的文字序列 $\alpha$ 值在 0.4～0.6 之间[27],贝多芬的音乐作品 $\alpha\approx 0.75$[28].为验证在线用户兴趣演化过程 $\alpha$ 值的一致性,对所有用户 $\alpha$ 值的分布进行统计,如图 4 所示,$\alpha$ 值近似服从 $N(1.5109,0.1598^2)$的正态分布,对 $\alpha$ 值分布的正态性进行统计检验如图 5 所示:

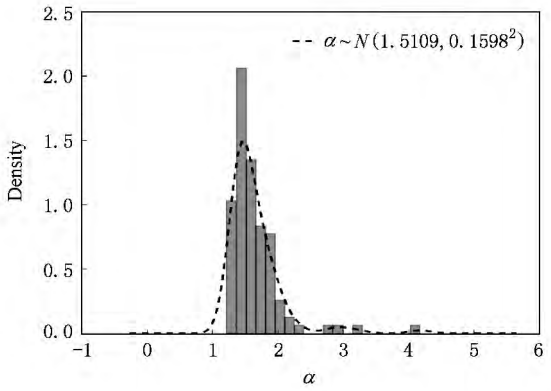



Fig. 4　The distribution of alpha.

图 4　$\alpha$ 值的分布

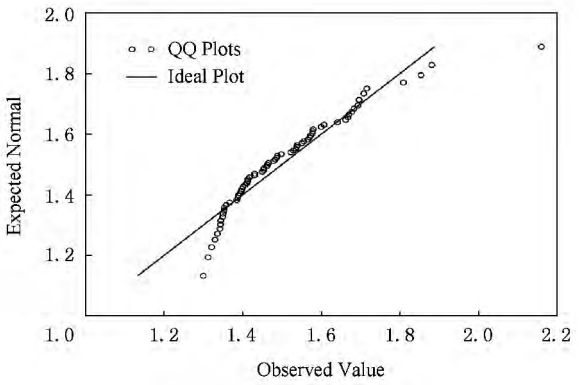



Fig. 5　Expected normal test of alpha.

图 5　$\alpha$ 值的正态性检验

由于超熵近似服从 $E(L)\propto \mathrm{lb}\,L$,说明用户的兴趣演化过程是一个无限长程记忆过程,超熵 $E$ 的值与块熵曲线 $H(L)$高度相关.综合表 3 数据以及 4.1 节分析可知,在线用户的兴趣演化过程是一个与 Thue-Morse 过程相似的无限非周期性长程记忆过程.

### 4.2　在线用户兴趣演化的可预测性分析

可预测性是大数据研究的核心价值之一,是体现大数据"4V"特征之"Value"属性的最重要方面[29],对个体行为的精准预测可促进定向广告的投放,对群体行为的预测可预防非常规突发事件,研究人员还试图通过大数据"预测流感"[30]等.但是,传统大数据分析技术大多采用机器学习算法从海量的数据中挖掘模式,对于"用多少数据量才能达到精准预测?"这样的问题考虑较少,使得学习效率低下且计算资源浪费严重,块熵理论可为此提供借鉴思路.

由 3.4 节可知,块熵的一阶离散导数 $\Delta H(L)$即为信息增益,用来区分兴趣演化过程中点击序列分布特征的不同,可以度量点击序列分布之间的距离,

也可对兴趣变化的不可预测性进行度量. 由于兴趣演化的信息增益极限 $\lim_{L\to\infty}\Delta H(L)=0$, 说明随 $L$ 变大, 不同时间段点击行为之间的分布距离在减小, 因此不可预测性也随块长度 $L$ 的渐增而减小.

图 6 展示了块熵的二阶离散导数 $\Delta^2 H(L)$ 的演化曲线, $\Delta^2 H(L)$ 用来度量兴趣演化过程中点击序列随块长 $L$ 渐增而不可预测性减小的程度, 称为预测增益. $|\Delta^2 H(L)|$ 的值越大, 表明块长从 $L-1$ 增加到 $L$ 后减小的不确定性量越大. 图 6 中的小图是将原图局部放大后的结果, 可以看出兴趣演化过程的可预测性较强, 当块长 $L$ 达到 7 时, 即在线兴趣点的连续点击达到 7 个时, 其行为的平均预测精确度就可达到 95.3% 以上. 从图 6 中的大图可以看出, 当连续点击数达到 14 个时, 精确度平均可达到 99% 以上, 表明人类在线行为与心理学中"神奇的数字 7"[31] 这一现象相符.

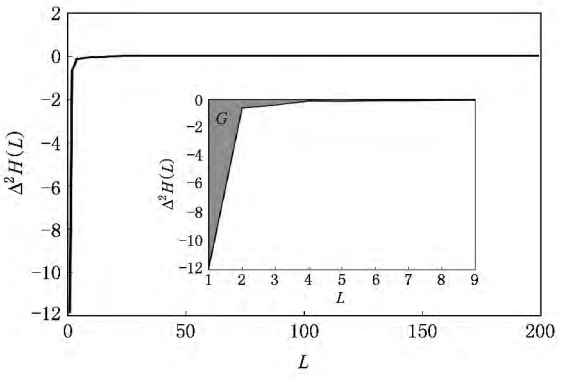



Fig. 6 Predictability gain vs sequence length.

图 6 预测增益随块长度变化曲线

### 4.3 块长 $L$ 最大长度的选择

在线用户每时每秒都可能发生大量的点击行为, 一个总长度为 $\ell$ 的点击序列 $[s]$, 可抽取出块长为 $L=\{1,2,\cdots,Max\}$ 的块共 $\ell-L+1$ 个, 这样的块序列组合成的词 $[w]=(w_i)_{1\leqslant i\leqslant \ell-L+1}$ 就构成了对长度为 $L$ 的块抽样. 这种有重叠的抽样不仅是为了度量连续兴趣点的重叠性, 还为了度量连续兴趣点序列内部的统计依赖关系.

通常认为大数据时代可以分析更多的数据, 甚至是与事物相关的所有数据, 可以带来更全面的认识, 可以更清楚地发现样本无法揭示的细节信息[2]. 对于数据分析而言, 数据块长 $L$ 的最大长度 $Max$ 应该越长越好; 然而, 本文针对在线用户兴趣演化过程的研究发现, 当前(present)兴趣点的信息熵是一个次广延量(sub-extensive), 不仅是过去(past)与未来(future)之间的或(or)信息, 还是过去与未来之间的与(and)信息, 兴趣演化过程具有较强的相关性, 因此分析所有数据既浪费计算资源还会对预测结果产生干扰.

在实际应用中, 也不可能按数据的总长度 $\ell$ 来选择块长, 因为在线点击数据是一个动态增长的时间序列数据. 实验发现, 当块长 $L$ 渐增到一定程度时, 块熵就达到最大值并且不再保持单调递增, 而是处于平稳或缓慢递减状态, 如图 2 所示, 因此选择 $L$ 过大对于数据分析意义不大.

文献[32-34]对熵率 $h$、总兴趣数 $|A|$、点击总长度 $\ell$ 与块熵 $H(L)$ 之间的关系作了一定探讨, 但结论并不统一. 文献[35]指出了选择块长 $L$ 最大长度应遵循的原则, 对于一个总长度为 $\ell$ 的数据序列, 随序列长度 $L$ 的增加而计算块熵 $H(L)$ 应满足以下 3 个条件: 1)块长 $L$ 的取值应满足 $\ell h\geqslant L|A|^L\ln|A|$; 2)使块熵增益 $\Delta H(L)$ 接近 0; 3)使超熵 $E$ 接近常量 $\mathrm{lb}\,\ell$, 其中 $\ell$ 为点击总量.

本文在实验中发现, 因为在线点击序列的熵率 $h=0$, 条件 1)无法适用, 利用条件 2), 3)虽然可经验性确定最大块长 $Max$ 值, 但如何准确确定这些量之间的统计关系却没有定论, 科学选择块长 $L$ 的最大长度将是未来需要进一步开展的研究工作.

## 5 结 论

根据熵理论定量分析了在线用户兴趣演化过程, 研究发现, 兴趣的演化过程既不是随机游走过程, 也不是 Markov 过程, 而是一个兴趣点序列服从幂率的无限非周期性长程记忆过程. 根据块熵的二阶离散导数——预测增益可发现, 当兴趣点序列块长达到 7 以上时其行为有较强的可预测性, 由此可确定在线行为精准预测所需的数据长度理论下界. 人类行为是一门复杂性科学, 是经济学、心理学等学科的基础, 本文的研究是对用户在线兴趣演化规律所作的初步探索, 将对人类在线行为分析、建模、预测等提供理论指导. 同时, 基于该研究结论可对在线产品推荐、机器学习等技术预测, 以及对大数据时代人类在线行为数据分析中所需的数据量提供指导. 对在线行为精准预测方法的研究是今后将要进一步开展的工作.

致谢 感谢中国科学院武汉物理与数学研究所丁义明研究员为本文工作提供的参考资料!

## 参 考 文 献


[1] Estrin D. Small data, where $n=$me [J]. Communications of the ACM, 2014, 57(4): 32-34

[2] Meng Xiaofeng, Li Yong, Zhu Jonathan. Social computing in the era of big data: Opportunities and challenges [J]. Journal of Computer Research and Development, 2013, 50(12): 2483-2491 (in Chinese)
(孟小峰, 李勇, 祝建华. 社会计算: 大数据时代的机遇与挑战[J]. 计算机研究与发展, 2013, 50(12): 2483-2491)

[3] Lazer D, Pentland A S, Adamic L, et al. Computational social science [J]. Science, 2009, 323(5915): 721-723

[4] Chierichetti F, Kumar R, Raghavan P, et al. Are web users really Markovian? [C] // Proc of the 21st World Wide Web Conf (WWW 2012). New York: ACM, 2012: 609-618

[5] Martyushev L M, Seleznev V D. Maximum entropy production principle in physics, chemistry and biology [J]. Physics Reports-Review Section of Physics Letters, 2006, 426(1): 1-45

[6] Wissner-Gross A D, Freer C E. Causal entropic forces [J]. Physical Review Letters, 2013, 110(16): 168702-1-168702-4

[7] Schrödinger E. What is Life? The Physical Aspect of the Living Cell and Mind [M]. Cambridge, UK: Cambridge University Press, 1967: 67-75

[8] Mises L. Human Action [M]. Auburn, Alabama: Ludwig von Mises Institute, 1998: 11-71

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

[10] Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel [J]. Nature, 2006, 439(7075): 462-465

[11] Malmgren R D, Stouffer D B, Motter A E, et al. A poissonian explanation for heavy tails in e-mail communication [J]. Proceedings of the National Academy of Sciences, 2008, 105(47): 18153-18158

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

[13] Han X P, Zhou T, Wang B H. Modeling human dynamics with adaptive interest [J]. New Journal of Physics, 2008, 10(7): 073010(1-8)

[14] 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

[15] Oliveira J G, Vazquez A. Impact of interactions on human dynamics [J]. Physica A: Statistical Mechanics and its Applications, 2009, 388(2): 187-192

[16] Aly M, Hatch A, Josifovski V, et al. Web-scale user modeling for targeting [C] //Proc of the 21st World Wide Web Conf (WWW 2012). New York: ACM, 2012: 3-12

[17] Aly M, Pandey S, Josifovski V, et al. Towards a robust modeling of temporal interest change patterns for behavioral targeting [C] //Proc of the 22nd World Wide Web Conf (WWW 2013). New York: ACM, 2013: 71-82

[18] Zheng N, Jin X, Li L. Cross-region collaborative filtering for new point-of-interest recommendation [C] // Proc of the 22nd World Wide Web Conf (WWW 2013). New York: ACM, 2013: 45-46

[19] Zhao Z, Yang Z, Zhang Z, et al. Emergence of scaling in human-interest dynamics [J]. Scientific Reports, 2013, 3(3472): 1-7

[20] Larson J W, Briggs P R, Tobis M. Block-entropy analysis of climate data [J]. Procedia Computer Science, 2011, 4(1): 1592-1601

[21] Christoph B, Bernd P. Permutation entropy: A natural complexity measure for time series [J]. Physical Review Letters, 2002, 88(17): 174102-1-174102-4

[22] James P C, David P F. Regularities unseen, randomness observed: Levels of entropy convergence [J]. Chaos, 2003, 13(1): 25-54

[23] Claypool M, Le P, Wased M, et al. Implicit interest indicators [C] //Proc of the 6th Int Conf on Intelligent User Interfaces. New York: ACM, 2001: 33-40

[24] Zhu Xia, Song Aibo, Dong Fang, et al. A collaborative filtering recommendation mechanism for cloud computing [J]. Journal of Computer Research and Development, 2014, 51(10): 2255-2269 (in Chinese)
(朱夏, 宋爱波, 东方, 等. 云计算环境下基于协同过滤的个性化推荐机制[J]. 计算机研究与发展, 2014, 51(10): 2255-2269)

[25] Shannon C E. A mathematical theory of communication [J]. Bell System Technical Journal, 1948, 27(3): 379-423

[26] Hao Bolin. Symbolic dynamics and characterization of complexity [J]. Physica D: Nonlinear Phenomena, 1991, 51(1): 161-176

[27] Ebeling W, Pöschel T. Entropy and long-range correlations in literary English [J]. Europhysics Letters, 1994, 26(4): 241-246

[28] Anishchenko V S, Ebeling W, Neiman A B. Power law distributions of spectral density and higher order entropies [J]. Chaos, Solitons & Fractals, 1994, 4(1): 69-81

[29] Meng Xiaofeng, Ci Xiang. Big data management: Concepts, techniques and challenges [J]. Journal of Computer Research and Development, 2013, 50(1): 146-169 (in Chinese)
(孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169)

[30] Lazer D, Kennedy R, King G, et al. The parable of Google flu: Traps in big data analysis [J]. Science, 2014, 343(6176): 1203-1205

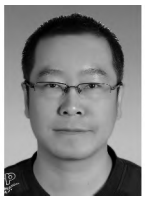
[31] Miller G A. The magical number seven, plus or minus two: Some limits on our capacity for processing information [J]. Psychological Review, 1956, 63(1): 81-97

[32] Wyner A D, Ziv J. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression [J]. IEEE Trans on Information Theory, 1989, 35(6): 1250-1258

[33] Schurmann T, Grassberger P. Entropy estimation of symbol sequences [J]. Chaos, 1996, 6(3): 414-427

[34] Ebeling W, Nicolis G. Entropy of symbolic sequences: The role of correlations [J]. Europhysics Letters, 1991, 14(3): 191-196

[35] Lesne A, Blanc J -L, Pezard L. Entropy estimation of very short symbolic sequences [J]. Physical Review E, 2009, 79(4): 0462081-1-0462081-10



**Li Yong**, born in 1979. PhD candidate at Renmin University of China. Lecturer at Northwest Normal University of China. Member of China Computer Federation. His main research interests include social computing, data analytics etc.

**Meng Xiaofeng**, born in 1964. Professor and PhD supervisor at Renmin University of China. Executive member of China Computer Federation. His main research interests include cloud data management, Web data management, flash-based databases, privacy protection etc.

**Liu Ji**, born in 1974. Professor at Xin Jiang University of Finance and Economics. His main research interests include social computing, network opinion management (liuji5000@126.com).

**Wang Changqing**, born in 1978. PhD and Senior Engineer at China Internet Network Information Center. Member of China Computer Federation. His main research interests include data analytics, online behavior analytics, human computer interaction etc (wangchangqing@cnnic.cn).