人在网络论坛上评论行为的分析与建模

叶起惠，肖井华
（北京邮电大学理学院，北京 100876）

摘要：本文研究了中国最出名的在线社区“天涯社区”和“百度贴吧”上的用户的评论行为，我们发现在一个版块里，一个话题的访问量和回帖量满足幂率关系。这说明存在一部分帖子受到的关注度非常高，影响力非常大，在挖掘舆论和引导舆论的时候必须注意这些帖子。基于个人的看帖习惯，本文建立了数学模型，很好的体现了人在网络论坛上的集体访问和回帖过程，为将来认识人在网络中的集体行为规律和引导舆论发展提供了实证与理论基础。

关键词：人类行为动力学；在线网络社区；BBS；幂率分布
中图分类号：N94

Modeling and statistical properties of human

Ye Qihui, Xiao Jinghua
(School of Science, Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract: Statistical properties of human comment behavior are studied using data from 'Tianya' and 'Tieba' which are very forums on-line social systems in China. We find that both the reply number R and the view number V of a thread in a sub-forum both obey power-law distributions.Respectively, which indicate that there exist a kind of highly popular topics. These topics should be specially focused on, because they play an important role in public opinion formation and opinion control. Based on human comment habit,a model is introduced to explain human view and reply behaviors in the forum.Numerical simulations of the model fit well with the empirical results. Our findingsare helpful for discovering collective patterns of human behaviors and the evolutionof public opinions on the virtual society as well as the real one.

Keywords: human dynamics; on-line society; BBS; power-law distribution

0 引言

关于人类行为的研究已经有超过 100 年的历史了，但由于人类自身的复杂性和多样性，以及实验数据的缺乏，到目前未知，大多数的命题和结论都是定性描述的。在以往的一些对社会，经济系统的研究中，常常把单个人的行为简化为可以使用泊松过程描述的稳态随机过程[1,2]，但最近越来越多的证据，包括从通讯模式到在线娱乐，从金融活动到网页浏览[3-7]表明人的很多行为中连续两个行为间的时间间隔是一个具有厚尾的幂率分布，不是一个简单的泊松过程，而 Barabasi[8]等科学家从 2005 年开始从记录人类活动历史的数据库中挖掘出人类行为的不同于泊松过程的统计规律，开创了一个名为人类行为动力学的新的研究方向，这个方向具有理论和应用上双重价值，很快就吸引了国际上大量知名科学家的关注。

论坛网站作为一个在现实生活中越来越重要的系统，作为人类行为动力学中的一种-人在论坛上的行为特征也吸引这越来越多的科学家的关注，Anna[9]等人分析了不同的门户网站的子网页之间的访问者的迁移行为。A.Grabowski[10]通过对四个不同类型的网站的分析，发现人的行为的活跃程度的分布是幂率的，人的活跃程度与时间之间也是幂率的关系。Wu[11]通过分析“天涯”的数据，发现人在一个话题里的评论动力学是一个非泊松过程，并建立了

作者简介：叶起惠，（1986-），男，主要研究方向：复杂网络与人类行为动力学. E-mail: yeah0102@sina.com
通信联系人：肖井华，（1965-），教授,现任北京邮电大学理学院执行院长。北京邮电大学学术委员会委员，国际理论物理中心 Fellow。主要科研领域是非线性系统与混沌，在 SCI 刊物上发表论文 40 余篇，应邀在国际论著上发表综述一篇，在物理学最著名的刊物 Phys.Rev.Lett.上发表论文 5 篇。论文在 SCI 刊物被他人引用 200 多次，出版专著一部，主编教材一本。主持国家自然科学基金 2 项，参加自然科学基金重点项目 2 项。E-mail: jhxiao@bupt.edu.cn

数学模型来解释。这些研究表明人在虚拟网络里的一些行为不是一个简单的泊松过程，而具有一些有趣的标度率在里面，而更多的科学家把论坛当作一个虚拟社会，来研究在这个社会中的各种人际关系网络统计规律以及其演化。

我们通过对中国最大的网络评论社区“天涯社区”和“百度贴吧”的帖子访问记录分析，发现论坛中版块每个帖子访问量和回复量是个幂率分布，同时访问量和回复量之间也满足幂率关系，说明在论坛里，存在着很多帖子的影响力非常大，这些帖子在舆论演化中起到很大的作用，必须加以关注。进一步，我们基于个人的看帖习惯，建立了人在论坛中的访问和回复的数学模型，仿真的结果表明我们的模型能提现个人在论坛中的评论行为动力学，希望为网络舆论分析提供理论基础。

1 数据

本文的数据采集于天涯社区和百度贴吧，天涯社区和百度贴吧里的帖子以及舆论可以说是中国互联网舆论的方向标，是灌水雇佣军的一个主要作战阵地。本文采集天涯社区和百度贴吧的版块数据，版块中的话题一般称为帖子，帖子构成了网络论坛中的信息单元。两个论坛版块的主体内容从政治杂谈到娱乐八卦，从个人生活到国际新闻，包含方方面面的不同话题。其参与人有白领，大学生，中学生等等，具有很大的广泛代表性。

表 1　天涯和百度论坛的数据格式

标题	*作者*	*查看量*	*回复量*	*最后回复时间*
话题 1	ID1	23124	123	2009/02/12,12:11:05
话题 2	ID3	323532	3243	2009/02/12,12:10:05
话题 3	ID32	42421	323	2009/02/12,12:03:05
. . .	...			
话题 N	ID28	232	43	2009/02/12,12:01:05

具体数据格式如表 1 所示。第一列是话题题目，第二列是话题的发起人，第三列是这个话题的访问量，第四列是回帖的数目，最后一列是这个话题里的最后一个回帖的时间。表 2 是四个版面的其他统计特征，包括了版面的总帖子数，总查看数，总回复数还有抓到的数据一共持续的天数。

表 2 四个版面的其他特征

版面	*天涯杂谈*	*天涯娱乐*	*百度魔兽世界吧*	*百度李毅吧*
总帖子数	19,492	19,479	27,359	5,302
总查看数	10.5 亿	16.2 亿	2.33 亿	519 万
总回复数	1.11 亿	1.43 亿	789 万	17.7 万
持续时间（天）	247	2569	13	15

2 数据结果

图 1 为不同的版面里的访问量和回帖数的统计分布，从表 2 可以知道，尽管版面的内容各不相同，涉及生活中的方方面面，参与的人也是各不相同，版面的热门成都也各不相同，但从图 1 可以看出，访问量和回帖数的分布都遵从幂率分布，斜率有点不同，即 $P(R) \propto R^{\alpha}$，$P(V) \propto V^{\beta}$，，其中实线是查看数的拟合线，虚线是回复数的拟合线，各个版面的拟合斜率为(a)天涯杂谈 $\alpha = 1.40 \pm 0.01, \beta = 1.44 \pm 0.02$ (b) 天涯娱乐 $\alpha = 1.35 \pm 0.02, \beta = 1.12 \pm 0.01$（c）百度魔兽世界吧 $\alpha = 1.51 \pm 0.01, \beta = 1.68 \pm 0.01$ (d) 百度李毅吧

$\alpha = 1.12 \pm 0.02, \beta = 1.76 \pm 0.02$

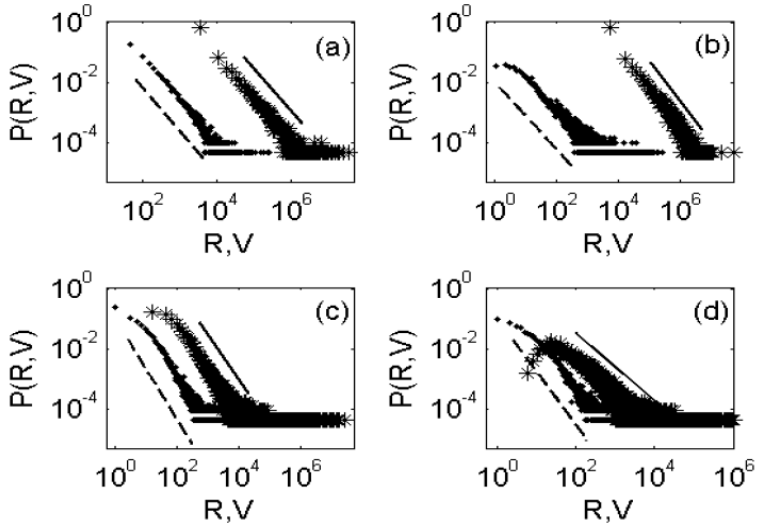

图 1 各个版面的查看数 V(实点)和回复数 R(星号)的幂率分布

由图可知这个访问和回帖的过程不是一个均匀的过程。其厚尾巴的幂率分布说明，存在着一大类帖子的影响力非常大，同时存在一大类帖子的参与互动程度很高，这类帖子的数目不可以忽略，所以必须重点关注。这类帖子一般代表着社会的舆论走向。

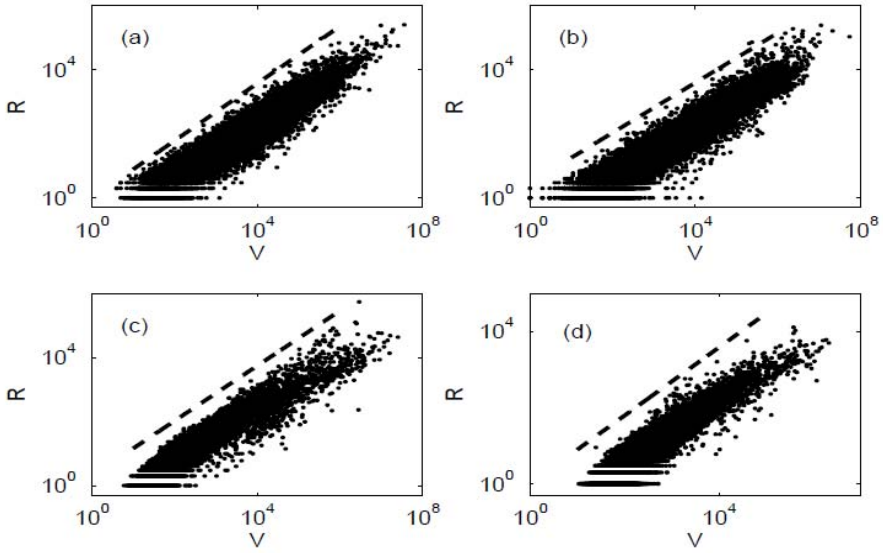

图 2 各个版面上回复数 R 和查看数 V 之间的关系

图 2 是各个版面上帖子的查看数和回复数的关系图, 其中虚线为拟合直线,各个版面的斜率为(a) 天涯杂谈 $\gamma = 0.77$ (b)天涯娱乐 $\gamma = 0.89$ (c)百度魔兽世界吧 $\gamma = 0.85$ (d)百度李毅吧 $\gamma = 0.90$

总所周知，一个帖子如果查看数越多，它的回复数也相对较多。但是他们之间的量化关系却不是很容易为人所知，而这个恰恰是一个帖子的基本的特性，因此我们关注他们之间的关系。如图 2 所示我们发现点击量和回复量在双对数坐标上可以很好的用一条直线来描述，用数学语言描述就是 $R \propto V^{\gamma}$， 这个关系很好的说明了查看数越多，回复数也越多，但是

这种非线性关系也表明，当一个帖子的查看数很大量的时候，其回复数的增长速度相对于查看数的增长速度是慢的。

3 数学模型与结果分析

根据我们自己平时的回帖习惯可以知道，帖子的查看数和回复数是逐渐增加的，虽然中间某些帖子会因为某种原因被删掉，但这部分只占很少的比例，可以忽略不计。所以我们分两步建立数学模型：

1)增长。在 t=0 论坛版面上存在少量的帖子，每个帖子都有少量的随机的 V 和 R，接下来的时间内，每个时间步会有一个新的帖子出现。而且在这个时间步中将会有 $c\times t^{\theta}$ 人次查看旧帖子，而且每个旧帖子讲会有一定概率被查看到。

2)查看喜好。每个时间步中，哪个帖子被查看到取决于改帖子的吸引度：$\Pi(i)=\frac{A_i(t)}{\sum_{i=1,N_t}A_i(t)}$，其中 $A_i(t)$ 是标记为 i 的帖子在 t 时刻的吸引度，N_t 为到 t 时刻，版面一共存在的帖子的总数目。吸引度 $A_i(t)$ 的定义如下：

$$A_i(t)=A(0)+V_i(t)$$

其中 $A(0)$ 为初始吸引度，对于不同的话题，取不同值。$V_i(t)$ 为 t 时刻标记为 i 的帖子已经拥有的查看数。

3)回复喜好。在每个时间步，当用户查看一个帖子，他将以概率

$$P(i)=L(i)\times(\frac{R(i)}{V(i)})^{\eta}$$

回复该帖子。其中 $L(i)$ 是帖子的初始吸引度

从数学上讲，这个模型和文章[12]提到的增长模型类似，所以根据文章[12]中的分析，可以知道最终的回帖数 V_i 的分布是幂率的，其斜率 $\alpha=1+(1+\theta)^{-1}$。我们拿图 1，2 中的 C 话题的参数来分析，$\alpha=1.51$，所以仿真时，我们取 $\theta=0.9$。

所得结果如图,其中图 a 是模型中查看数 V 的概率分布，折线是它的拟合线，该拟合线斜率为 β=1.51±0.02，图 b 是模型中回复数 R 的概率分布，折线也是拟合线，拟合线斜率为 α=1.41±0.04，图 c 是模型中查看数 R 和回复数 V 的关系图，折线也是拟合线，拟合线斜率为 γ=0.90±0.02，图 d 是斜率 γ 和常数 η 的关系。

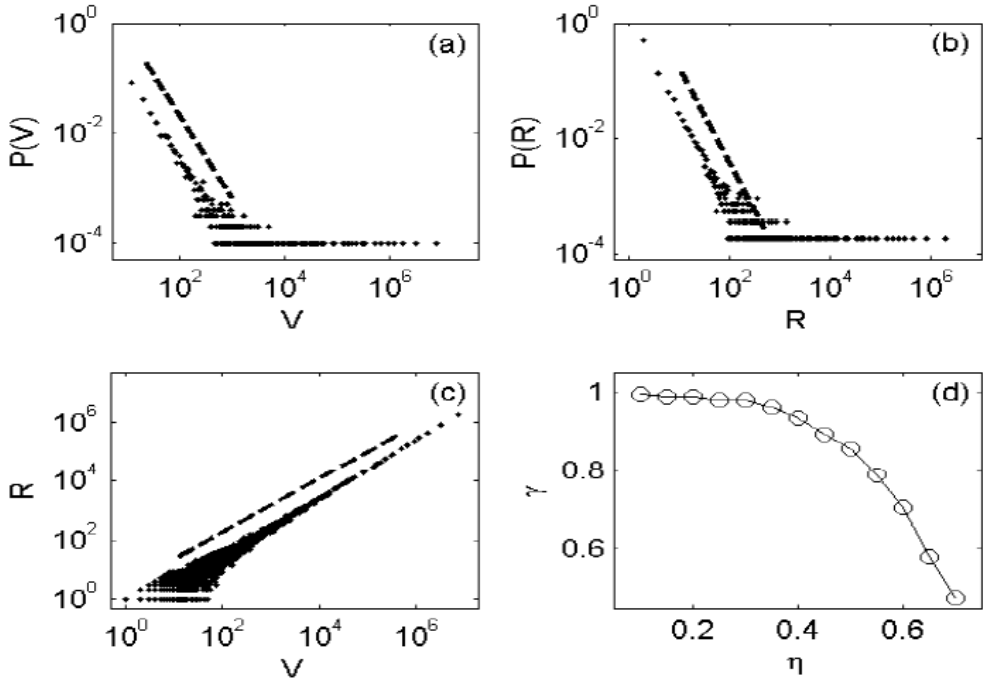

图 3 仿真结果。各个参数选取为 θ=0.9，L=0.1，η=0.5

我们是以版面 C，也就是百度魔兽世界吧的数据进行建模。模型得到的结果和实际结果进行对比，图 3 中的查看数和回复数分布明显的都服从幂率分布，而且斜率也与实际数据中的结果相近，而模型中查看数 V 和回复数 R 之间的非线性关系也明显的是一个幂率分布，最后，我们也研究了建模过程中的常数η对于斜率γ的影响，我们可以看到随着η的增大，γ是不断变小的。

4 结论

本文以“天涯社区”和“百度贴吧”的若干版面为研究对象，分析了在线社区查看和回复行为的统计特性。查看和回复行为都服从于幂率分布，而且他们之间的关系也是服从一个幂率关系。基于这个关系，还有加上模拟个人看帖回复帖子的行为，我们建立了数学模型。该模型在数值上很好的符合了实际数据，我们的模型可以很好的解释一些人们现实中的评论行为，例如人们移动电话会议中的评论行为[13]。本文的模型只体现了个人在网络上的行为规律，人在网络上的行为不是孤立的，而是具有相互作用，在度分布极度不均匀的网络上[14,15]的人的相互作用对个人行为规律的影响，是将来需要关注的方向。

[参考文献] (References)

[1] F.A.Haight.handbook of the Poisson Distribution[M]. New York:Wiley,1967.
[2] P Reynolds. Call Center Staffing[M]. Tennessee:The call Center School Press, 2003
[3] Oliveira J,Barabási A.Human dynamics:The correspondence patterns of darwin and einstein[J]. Nature, 2005, 437(7063): 1251-1251.
[4] Eckmann JP,Moses E, Sergi D. Entropy of dialogues creates coherent structures in e-mail traffic[J]. PNAS, 2004,101(40): 14333-14337
[5] Vazquez A, Oliverira, Dezso, Goh KI, etc. Modeling bursts and heavy tails in human dynamics[J]Phys. Rev. E, 2006, 73(3):036127-036145
[6] Malmgen RD, Stouffer DB, Mottter AE, Amaral LAN. A poissonian explanation for heavy tails in e-mail communication[J]. PNAS, 2008, 105 (47): 18153-18158

[7] Hong W,Han XP,Zhou T,Wang BH.Heavy-Tailed statistics in short-message communication[J].Chinese Physics Letters,2009,26(2):Article 028902.
[8] Barabási AL.The origin of bursts and heavy tails in human dynamics[J].Nature,2005,435(7039):207-211.
[9] Anna Chmiel1, Kamila Kowalska2, and Janusz A. Hołyst. Scaling of human behavior during portal browsing[J]. Phys. Rev. E, 2009, 80(6): 066122- 066128
[10] Grabowski A,Kruszewska N.Experimental study of the structure of a social network and human dynamics in a virtual society[j].Int'l Journal of Modern Physics C,2007,18(10):1527-1535.
[11] Y. Wu, C S. Zhou, M Y. Chen, J H, Xiao, J. Kurths. Human comment dynamics in on-line social systems[J]. Phys. A, 2010, 389(24): 5832-5837
[12] S. N. Dorogovtsev, J. F. F. Mendes. Effect of the accelerating growth of communications networks on their structure[J]. Phys. Rev. E, 2001, 63(2): 025101- 025104
[13] Xenikos D G.Modeling human dialogue - the case of group communications in trunked mobile telephony[J]. Phys. A, 2009, 388(23): 4910-4918
[14] 宋玉蓉，蒋国平.基于一维元胞自动机的复杂网络恶意软件传播研究[J]. 物理学报，2009，58(9): 5911-5918
[15] 吴晔，肖井华，吴智远，马宝军，杨俊忠. 手机短信网络的生长过程研究[J]. 物理学报，2007，56(4)：2037-2041