

文章编号:1007-6735(2013)04-0345-06

基于有限匹配与择优选取的大众生产系统演化研究

李霞¹, 李守伟²

(1. 滨州医学院 网络中心, 滨州 256603; 2. 江苏大学 管理学院, 镇江 212013)

摘要: 基于维基百科作为典型大众生产系统以及众多用户参与编辑的现实背景, 考虑到用户编辑行为的“择优”和“有限匹配”的特征, 给出大众生产系统的“择优选取”和“有限匹配”规则以及演化模型, 采用计算实验的方法, 对大众生产系统演化中的页面编辑次数、页面与用户状态变化、匹配度影响等进行全面分析. 实验结果表明: 在“择优选取”和“有限匹配”两个原则的作用下, 维基百科系统逐渐演化到稳定状态. 在稳定状态下, 页面被编辑次数服从幂律分布; 用户状态与页面状态之差 (即匹配度) 趋向于零; 用户与页面的匹配度越大, 幂律分布的幂指数就越小, 从而幂律分布的“尾巴”就越长.

关键词: 大众生产系统; 有限匹配; 择优选取; 维基百科; 计算实验
中图分类号: N 941 **文献标志码:** A

Evolution of Peer Production System Based on Limited Matching and Preferential Selection

Li Xia¹, Li Shou-wei²

(1. Network Center, Binzhou Medical University, Binzhou 256603, China;
2. School of Management, Jiangsu University, Zhenjiang 212013, China)

Abstract: Based on the real background of Wikipedia adopted as a classic peer production system and many users taking part in its editing, the two characteristics of preferential selection and limited matching during the editing process were considered. Two rules for “preferential selection” and “limited matching” and the evolving model of peer production system were presented. The analysis was based on computational experiments on the times of page editing, the status variation of pages and users, the affection of matching degree on page editing times, etc. The computational experiments show that the Wikipedia system evolves to a stable status under the action of the two rules. In the stable status, the times of page editing follow power-law distribution; the difference between user’s status and page status (i. e. the matching degree) is toward to zero; the larger the matching degree of user and page, the smaller the power index of power-law distribution, so the longer the tail of power-law distribution.

收稿日期: 2012-10-25

基金项目: 滨州市科技计划资助项目(2011ZC1002); 国家社会科学基金资助项目(11BJL074); 教育部人文社会科学研究规划基金资助项目(10YJAZH042).

第一作者: 李霞(1977-), 女, 实验师. 研究方向: 复杂网络与网络应用. Email: lix770429@126.com

通讯作者: 李守伟(1970-), 男, 副教授. 研究方向: 复杂网络、知识管理与技术创新. Email: shouweili1996@163.com

Key words: *peer production system; limited matching; preferential selection; Wikipedia; computational experiments*

在过去十年里,大规模协作平台已在多个生产环境中频繁出现,特别是基于互联网的知识创造系统.如著名的Linux操作系统是由众多程序员共同参与开发的开源软件系统;Youtube用户相互之间在线分享和评价视频,形成了大规模视频知识库;Wikipedia吸引了数量众多的用户参与词条的编撰,进而组建了大型的知识库——维基百科.维基百科是一个基于Wiki技术的全球性多语言百科全书协作计划,同时也是一部用不同语言写成的网络百科全书,其目标及宗旨是为全人类提供自由的百科全书,是一个动态的、可以自由访问和编辑的全球性知识体.Benkler^[1]将这种由大量分散在世界各地的互联网用户共同参与生产与共享知识产品的知识生产模式称为大众生产(peer production).随着互联网的普及和技术进步,大众生产方式越来越成为流行的知识生产方式^[2].

基于普通用户的大众生产已创造了广泛的协作系统,用于生产软件、共享数字内容、并组织大型知识库.这种现象引起了各种学者的关注.经济学家主要关注如何有效地激励大众生产者的参与行为^[3];社会学家分析影响大众生产者行为的因素及其参与动机^[4];法学家主要关注大众生产所带来的知识产权问题^[5];计算机学家则研究如何从技术层面提高大众生产方式的效率等问题^[6].随着大众生产者规模的不断扩大,物理学家开始关注大众生产者所构成的网络虚拟社会的整体规律性特征^[7].

在大众生产系统中,知识产品的质量与大众生产者的合作密切相关,存在着“质量与合作”的悖论.即知识产品质量的提高需要大规模生产者的合作,然而,随着生产者规模的增加,合作过程中的冲突也随之增加,又直接导致了知识产品质量的下降.对于小型的团队,由于生产者的规模有限,这似乎不是问题.然而,对于较大规模的大众生产系统,这是一个不可回避的问题.维基百科是典型的大规模的大众生产产品.Wilkinson等^[8]认为在维基中词条质量和用户合作之间是有联系的.但是,由于生产者是基于非金钱的激励而自愿参加词条的编撰工作.因此,在大规模维基百科中,必须设置较低的访问障碍,以便吸引数量众多的贡献者(生产者).其中,专家用户在维持和监督方面扮演至关重要的角色,以提高词条

的质量.

大规模协作系统中的用户行为已经被一些学者所关注,如Leskovec等^[9]和Yang等^[10]分别研究了社交网站和知识共享平台上的用户行为.Wilkinson^[11]基于对在线大众生产系统的分析,提出了用户寿命分布的法则.Grabowski等^[12]研究发现用户账户的寿命分布按照厚尾进行衰变,并进而提出了幂律模型.然而,通过Wikipedia的实证数据,Ciampaglia等^[13]发现用户寿命并不简单地服从幂律分布.

事实上,网上大众生产者的参与行为是有“择优性”的,即大众生产者按照一定的规则,有选择性地、以不同的概率参与不同知识的生产.因此,一个新加入者的长期参与性可以由其早期的互动结果来预测.这可以由社会化理论来解释,即用户评估网络社区对其接纳的意愿,反之亦然.另外,生产内容的质量评估以及个体与社团目标的比较,在决定用户参与上也是非常重要的,在维基百科上表现为词条编辑者之间的日常协作形式与群组的一致性.

以维基百科为例,在一个用户“社区”中,多个用户从事编辑页面的合作.页面由一系列特征所标识,如页面的编写风格.用户可以发现他们是否同意这些特征,用户试图根据他们的目标编辑页面,即使用他们自己的风格来编辑页面.同时,通过页面编辑行为的互动,用户也被其他用户的风格所影响.然而,这种相互影响仅仅对于某些情况才发生,也就是说,只有当用户和页面的风格足够的相似,他们之间才相互影响.在社会心理学的背景下,这种现象被称为“有限的信任”,作为组内尝试达成共识的人类沟通的一般特征.它也可以看作是人们在社会环境中由社会背景影响所形成的一种从众形式.在维基百科的编辑中,“有限信任”表现为编辑者(用户)与页面风格(或内容)之间的匹配,即“有限匹配”.

无疑,有限匹配与择优选取对于大众生产系统的演化有着重要的影响.探究大众生产系统在有限匹配和择优选取机制作用下的演化规律,是本文研究的主要内容.笔者还给出了有限匹配与择优选取的演化规则、基于有限匹配与择优选取的演化模型,并对演化模型做了计算实验分析.大众生产系统的动态演化分析对于提高生产效率、分析用户行为等

有着重要的指导意义.

1 有限匹配与择优选取

为了分析有限匹配与择优选取机制对于大众生产系统演化的影响,以维基百科的页面编辑为例,分析大众生产者合作下的随机页面编辑行为.

1.1 基本假设

在维基百科中,用户与页面是两类 Agent. 为了研究有限匹配度的影响,用户与页面的状态分别用他们的撰写风格来表示为 $[0,1]$ 区间上的数值. 分别用 $x(t)$ 表示 t 时刻用户的状态,用 $y(t)$ 表示 t 时刻页面的状态. 状态数值的大小与页面内容的正确与否无关,只关注用户与页面状态(风格)的匹配情况.

1.2 页面新建与择优选取机制

在用户编辑特定页面之前,必须要作出决策:要么新建一个页面,要么编辑已有的页面,两者必选其一.

规则 1(页面新建与择优选取机制) 用户以一定的概率创建新页面或者编辑已有页面. 在选取已有页面时,由于不同的页面有不同的主题或风格,用户会根据页面的知名度(或流行度)作出选择并编辑.

对于用户选择已有页面进行编辑的情况,假设 $c_p \geq 0$ 是一个常数, m_t 是直到时刻 t 该页面所受到编辑的次数(也即是其流行度或知名度),则其在时刻 t 被选中的概率正比于 $(m_t + c_p)$. 若用 $\pi(t)$ 表示页面选取的概率,则满足等式

$$\pi(t) = \frac{m_t + c_p}{\sum (m_t + c_p)} = \frac{m_t + c_p}{\sum (m_t) + Mc_p} \quad (1)$$

式中, M 为大众生产系统中的页面总数. 当 $c_p \rightarrow \infty$, 则 $\pi \rightarrow 1/M$, 页面将以均匀分布的方式被选中编辑,而忽略其已经被编辑的次数. 因此,可以通过设置 c_p 的数值大小来研究内容流行度的影响.

当然,用户并不经常选择去编辑一个存在的页面. 有时候,一个用户可以决定创建一个新页面. 通过设置新页面的创建概率来对其建模. 无论何时一个页面被创建,其状态等于创建者的状态. 创建者是从已有的用户中随机选中或者是新添加的用户.

若 t 时刻某个新页面被某个用户创建,则用户的状态不变,而新页面的状态满足

$$y(t) = x(t) \quad (2)$$

若 t 时刻某个用户择优选取某个页面进行编辑,则该页面和用户的状态按照规则 2 进行变化.

1.3 状态有限匹配机制

在时刻 t , 用户编辑一个页面. 假设 μ 为发生编辑行为后用户或者页面状态更新的速率,其数值大小表明用户或页面状态更新幅度的不确定性大小, $\mu \in [0, 0.5]$; ϵ 为用户与页面之间的匹配度, $\epsilon \in [0, 1]$.

规则 2(用户与页面的状态有限匹配机制) 在大众生产系统(维基百科)中,用户与页面通常依据他们之间的匹配度相互影响(编辑与被编辑),并更新其状态.

在 t 时刻,如果 $|x(t) - y(t)| < \epsilon$, 则同时更新用户和页面的状态

$$x(t) = x(t) + \mu(y(t) - x(t)) \quad (3)$$

$$y(t) = y(t) + \mu(x(t) - y(t)) \quad (4)$$

式(3)和式(4)表明,当用户与页面状态相互匹配时,用户与页面的状态(风格)都将作出调整. 如果 $x(t) > y(t)$, 则 $x(t)$ 减少, $y(t)$ 增大; 如果 $x(t) < y(t)$, 则 $x(t)$ 增大, $y(t)$ 减少.

在 t 时刻,如果 $|x(t) - y(t)| \geq \epsilon$, 则以概率 ρ_w 只更新页面的状态,而用户状态保持不变

$$y(t) = y(t) + \mu(x(t) - y(t)) \quad (5)$$

式(5)说明,当用户与页面状态相差较大(即不匹配)时,用户对页面的编辑与否是不确定的,这里用概率 ρ_w 表示对页面编辑的可能性. 当用户编辑页面时,用户按照其风格编辑页面内容,并影响页面风格(状态),但不会被页面状态(风格)所影响. 所编辑的页面将会以给定概率增大或减小页面状态,以向用户状态靠拢.

有限的匹配度反应了这样的事实,在大众生产系统中,用户与页面的风格不但相互影响,而且用户也可能处理他们不认同的页面风格,而不被其影响. 在随机概率的作用下,用户与页面之间的择优选取与有限匹配机制抓住了大众生产系统中社会影响的动态性.

2 演化模型

基于择优选取和有限匹配的规则,本文构建了大众生产系统的演化模型. 演化模型共有 5 个步骤:

步骤 1 初始状态

假设大众生产系统(维基百科)在初始时刻有 n_0 个用户, m_0 个页面. 用户和页面的状态均为 $[0, 1]$ 之间的随机数. 每个页面的初始编辑次数为 1.

步骤 2 添加新用户或选取老用户

考虑到大众生产系统的开放性,必须以一定的

概率向系统添加新用户,才能更好地符合实际情况.假设以概率 ρ_u 向系统添加一个新用户,新用户的状态为 $[0,1]$ 之间的随机数;以概率 $1 - \rho_u$ 从已有用户中随机选取一个.概率 ρ_u 的大小也可以反映出系统对用户的吸引力.

步骤3 页面新建或选取

在选定用户(包括新建的用户)后,用户的行为有新建页面或者编辑已有的页面.假设以概率 ρ_p 新建一个页面,则以概率 $1 - \rho_p$ 择优选取一个页面,择优选取的概率满足式(1).概率 ρ_p 的大小反映出用户对于已有页面的兴趣度.

步骤4 用户和页面状态更新

在用户与页面建立联系后,要进行各自状态的更新.

对于新建的页面,按照式(2)将用户的状态赋予新页面,而用户的状态不变.

对于选定的旧页面,分别按照有限匹配度规则中的式(3)~(5)进行用户和页面状态的更新.

步骤5 终止条件判断

转向步骤2,当达到指定用户规模数、或者页面规模数、或者一定时间时,系统演化停止.

通过对演化步骤的分析,本文给出了大众生产系统(维基百科)的演化流程图如图1所示.

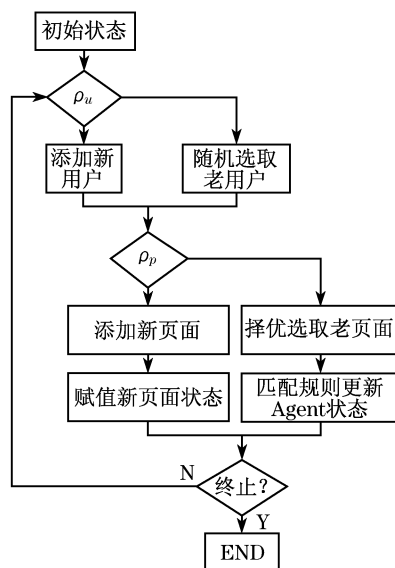


图1 大众生产系统的演化流程图

Fig.1 Evolution flow chart of peer production system

3 计算实验模拟结果与讨论

为了检测与验证择优选取与有限匹配对于大众

生产系统演化的影响,采用基于多 Agent 的计算实验方法,来模拟维基百科的演化过程.演化程序采用 Matlab 编写的.

在初始时刻,假设维基百科的初始状态由 $n_0 = 5$ 个用户和 $m_0 = 20$ 个页面构成.页面新建概率为 $\rho_p = 0.5$,即新建或编辑页面的概率各占一半.新添加用户的概率 $\rho_u = 0.5$,即每个时刻以 0.5 的概率添加新用户,使系统规模不断扩大.

有限匹配度为 $\epsilon = 0.5$,用户或页面状态的更新速率为 $\mu = 0.8$,用户与页面状态不匹配时,用户对页面编辑的概率为 $\rho_w = 0.5$.整个演化采用时间控制法,当系统运行到 10 000 个时间步后,系统停止演化.

3.1 页面编辑次数的幂律分布

对系统中的页面编辑次数进行统计分析,并采用双对数坐标绘制了频率分布图,如图2所示.

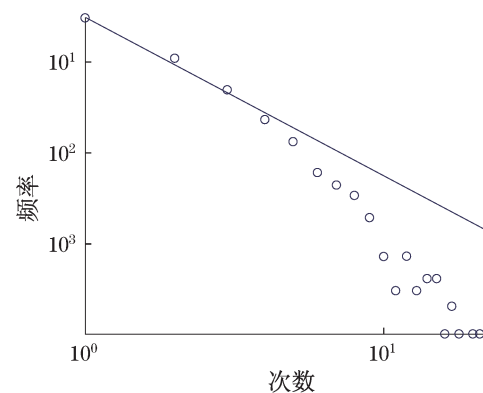


图2 页面编辑次数的幂律分布图

Fig.2 Power-law distribution of page editing times

从图2可以看出,页面编辑次数服从幂律分布, $p(m) \propto m^{-\tau}$,其中幂指数为 $\tau = 1.72$.从图2还可以看出,新建页面(即编辑次数 $m = 1$ 的页面)占据了系统页面总数的较高比例,同时编辑次数较多的页面则大多是系统初创时期或发展早期的创建的.

通过对演化过程的分析可知,要使大众生产系统成功实现,首先要保证系统的开放性,降低用户登录和编辑的门槛,以吸引更多的贡献者.其次,系统在初创或发展早期的内容质量(或风格)与目标定位也是非常重要的.另外,研究发现,专家型用户不但对于系统知识质量的监控有着重要的作用,而且在系统初创时期也影响着大众生产系统的发展方向.

3.2 用户与页面的状态变化

在大众生产系统的演化中,用户与页面的状态也随机地发生变化.当用户与页面状态匹配时,由式(3)和式(4)可知,用户与页面状态差的绝对值落在

图3所示的用户与页面状态演化示意图中阴影部分,此时,用户与页面状态都发生变化,在对角线上方, $x(t)$ 增大, $y(t)$ 减小;在对角线下方, $x(t)$ 减小, $y(t)$ 增大,其变化方向如图3中箭头所示。

当用户与页面状态不匹配时,由公式(5)可知,用户与页面状态差的绝对值落在图的左上角和右下角部分。此时,只有页面状态发生改变,在左上角部分, $y(t)$ 减小;在右下角部分 $y(t)$ 增大,其变化方向如图3中箭头所示。

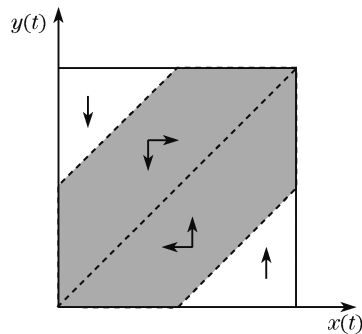


图3 用户与页面状态演化示意图

Fig.3 Evolving schematic diagram of user and page status

选定用户,分别取其初始状态为0.8和0.3,统计其状态更新次数,如图4所示。从图4中可以看出,初始状态较大时,其落在对角线上方的概率就越大,状态减小的可能性越大;初始状态较小时,其落在对角线下方的概率就越大,状态增加的可能性越大。

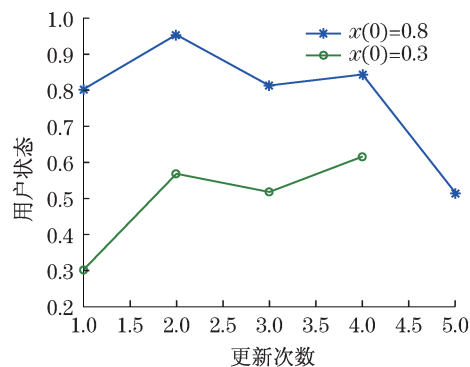


图4 用户状态更新图

Fig.4 Update diagram of user status

选定页面,分别取其初始状态为0.8和0.3,统计其状态更新次数,如图5所示。从图5中可以看出,初始状态较大时,其落在对角线上方的概率就越大,状态减小的可能性越大;初始状态较小时,其落

在对角线下方的概率就越大,状态增加的可能性越大。

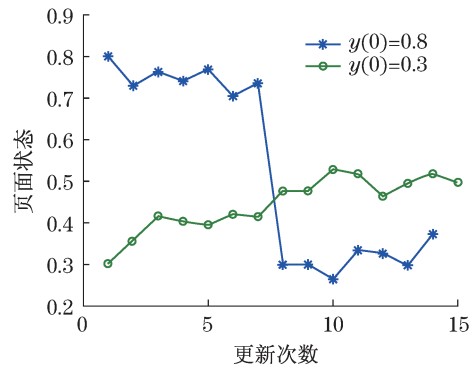


图5 页面状态更新图

Fig.5 Update diagram of page status

当然,由于系统的开放性以及演化中的随机性,用户和页面的状态并不是严格按照上述规律变化的,但总体上或者局部时间段内是符合上述演化规律的。

3.3 匹配度与页面编辑次数的关系

在大众生产系统中,用户与页面的匹配度不但影响着用户的行为,而且影响着页面被编辑的次数。在其它参数不变的情况下,分别对匹配度赋值0.2, 0.4, 0.6和0.8,然后进行系统演化,最后统计出页面的编辑次数。图6给出了4种情况下页面编辑次数的分布图。从图6中可以看出,页面编辑次数服从幂律分布, $p(m) \propto m^{-\tau}$, 其中,匹配度为 $\epsilon_1 = 0.2$ 、 $\epsilon_2 = 0.4$ 、 $\epsilon_3 = 0.6$ 和 $\epsilon_4 = 0.8$ 所对应的页面编辑次数分布的幂指数分别为 $\tau_1 = 1.876$ 、 $\tau_2 = 1.762$ 、 $\tau_3 = 1.695$ 和 $\tau_4 = 1.674$ 。从而说明,随着匹配度的增加,幂律分布的幂指数也在增加($-\tau < 0$),幂律分布的“尾巴”更长,更多的已有页面被编辑。

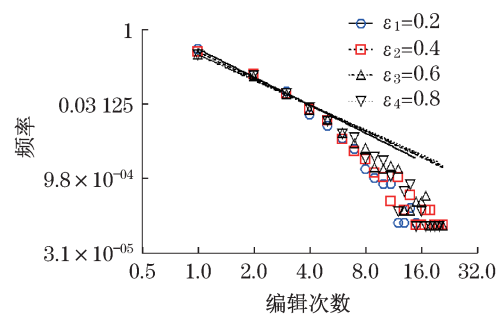


图6 匹配度与页面编辑次数的关系图

Fig.6 Relationship between match degree and page editing times

4 结 论

随着互联网的普及和大规模协作平台技术的发展,大众生产方式已成为未来知识生产的主要方式之一.大众生产系统的动态演化是众多研究者关注的话题,其中维基百科是备受关注的较成功的大众生产系统之一.

择优选取影响着用户对页面的认同度,而有限匹配不但影响用户的行为,而且影响着用户与页面的状态.用户与页面的状态朝着相互匹配的方向发展,即用户与页面的状态差的绝对值 $|x(t) - y(t)|$ 趋向于零.在有限匹配和择优选取机制的共同作用下,页面编辑次数服从幂律分布.同时,有限匹配度越大,页面被编辑次数的幂律分布越“水平”,出现较长的“尾巴”,表明更多的已有页面被用户编辑.

大众生产系统中的用户行为的相互影响是非常值得研究的内容,进一步的研究将在用户寿命分布、页面内容质量提高等方面展开.

参考文献:

- [1] Benkler Y. Coase's penguin, or, linux and the nature of the firm[J]. The Yale Law Journal, 2002, 112(3): 369 - 446.
- [2] 李志宏,王娜.大众生产系统发展动力机制建模——基于系统动力学方法[J].科学学研究, 2012, 30(2): 232 - 240.
- [3] Ariel Y. Webuser's gratifications expectations[R]. Presented at the Seventh International Conference General Online Research-GOR05, Zurich: University of Zurich, 2005.
- [4] Ling K, Beenen G, Ludford P, et al. Using social psychology to motive contributions to online communities[C] // Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work. Chicago, 2005, 212 - 221.
- [5] Roberts J, Hann I, Slaughter S. Understanding the motivations, participation, and performance of open source software developers: a longitudinal study of the apache projects[J]. Management Science, 2006, 52(7): 984 - 999.
- [6] Bruns A, Humphreys S. Wikis in teaching and assessment: the m/cyclopedia project [C] // Proceedings of the 2005 international Symposium on Wikis. San Diego, 2005.
- [7] Crowston K, Howison J. The social structure of free and open source software development [J]. First Monday, 2004, 10(2): 1 - 21.
- [8] Wilkinson D M, Huberman B A. Cooperation and quality in wikipedia[C] // Proceedings of the 2007 International Symposium on Wikis. Quebec, 2007.
- [9] Leskovec J, Backstrom L, Kumar R, et al. Microscopic evolution of social networks[C] // Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, 2008.
- [10] Yang J, Wei X, Ackerman M, et al. Activity lifespan: an analysis of user survival patterns in online knowledge sharing communities[C] // Proceedings of the 2010 International AAAI Conference on Weblogs and Social Media. Washington, 2010.
- [11] Wilkinson D M. Strong regularities in online peer production [C] // Proceedings of the 9th ACM Conference on Electronic Commerce. Chicago, 2008.
- [12] Grabowski A, Kosinski R A. Life span in online communities [J]. Physics Review E, 2010, 82(6): 066108.
- [13] Ciampaglia G L, Vancheri A. Empirical analysis of user participation in online communities: the case of Wikipedia [C] // Proceedings of ICWSM 2010. Washington, 2010.

(编辑:金虹)