

文章编号:1007-6735(2012)05-0441-06

基于 CAS 理论的群体协作维基词条编辑建模仿真

赵东杰^{1,2}, 王 华^{3,4}, 李德毅⁵, 李 智², 赵洪利⁶, 杨海涛²

(1. 63628 部队, 北京 101601; 2. 装备学院, 北京 101416; 3. 中国航天员科研训练中心, 北京 100094;
4. 军事医学科学院, 北京 100850; 5. 中国电子系统工程研究所, 北京 100840; 6. 总装备部, 北京 100720)

摘要: 基于复杂适应性系统理论将维基百科用户抽象为内容添加者、内容修改者、内容删除者、多样编辑者和内容浏览者 5 种主体, 建立了群体协作词条编辑模型; 基于不同质量词条主体出现的概率配置, 利用 Netlogo 仿真软件构建仿真平台, 实现了对群体协作词条编辑的多主体建模仿真. 仿真实验表明, 多样编辑者是词条质量提升的重要驱动力, 其出现概率越大, 生成词条质量越高; 编辑者“自我修改”行为对提升词条质量起到重要促进作用; 主体出现概率配置遵循黄金分割律时, 可使群体绩效趋于最大化. 词条演化存在语量与语义之间此消彼长、最终达到动态平衡的过程, 语量与语义平衡临界点近似为黄金分割点, 词条编辑演化遵循着黄金分割律. 研究深化了对词条编辑演化、网络群体智能和社会计算的认识.

关键词: 维基百科; 群体协作词条编辑; 复杂适应性系统; 建模仿真; 黄金分割; 网络群体智能

中图分类号: TP 391; N 94; G 20 **文献标志码:** A

Model and Simulation of Collective Collaboration Article Edit in Wikipedia Based on CAS Theory

ZHAO Dong-jie^{1,2}, WANG Hua^{3,4}, LI De-yi⁵, LI Zhi², ZHAO Hong-li⁶, YANG Hai-tao²

(1. Troop No. 63628 of PLA, Beijing 101601, China; 2. The Academy of Equipment, Beijing 101416, China;
3. China Astronaut Training and Research Center, Beijing 100094, China;
4. Academy of Military Medical Sciences, Beijing 100850, China;
5. Institute of Electronic System Engineering, Beijing 100840, China;
6. General Equipment Department, Beijing 100720, China)

Abstract: Wikipedia users were classified into five agents, including content creator, content modifier, content cleaner, diverse editor and content visitor, and a collective collaboration article edit model was established based on the complex adaptive system theory. The multi-agent simulation of collective collaboration article edit was achieved by using Netlogo software based on the configuration of agent appearance probabilities of different quality articles. Simulation results show that diverse editor is an important driving force for the improvement of article quality; the bigger the appearance probability of diverse editors is, the higher the article quality is. The self-modifying behavior of editors plays an important role in promoting article quality. When the configuration of agent appearance probabilities follows the golden section law, the collective

收稿日期: 2012-04-27
基金项目: 国家自然科学基金资助项目(69120912, 61035004)
作者简介: 赵东杰(1980-), 男, 博士研究生. 研究方向: 复杂系统设计优化、智能信息处理. E-mail: zhaodj2006@126.com

performance can tend to be maximum. There exists a process from seesaw-like complementarity to dynamic balance between word quantity and word meaning, the critical point of balance is close to the golden section point, and the article edit evolution follows the golden section law. The research deepens the knowledge of article edit evolution, web collective intelligence and social computing.

Key words: Wikipedia; collective collaboration article edit; complex adaptive system (CAS); model and simulation; golden section; web collective intelligence

Wiki 是一种“允许”互联网上多个不同的用户以浏览器作为客户端,来直接修改网页内容的机制.采用这种机制的站点,可被多个用户修改,形成面向社区的、由大众参与的对等生产,体现了人人参与创造 Web 2.0 这一理念.2011 年 2 月,美国哈佛大学公布了当前及未来亟需解决的 10 大社会科学问题,其中“人类如何增加自身群体智能”、“我们如何才能集合每个人所拥有的信息来作出最佳决定”和“怎样理解人类创造和表达知识的能力”这 3 个问题位列其中.维基百科利用互联网上大众用户的集体参与来创作百科知识,是利用大众普遍参与、编辑交互形成群体智能的典型应用,为研究以上问题提供了高价值数据资源.目前,一些研究者已对大众交互的互联网环境下人的群体行为展开研究^[1-8],对维基百科的研究,主要集中在语义知识挖掘^[9-11]和优良条目的自动发现与挖掘方面^[1,12],对词条演化研究不足.同时,对于大众不断参与的在线群体协作演化过程,仍缺乏有效研究方法,有待深入研究.本文基于复杂适应性系统 (complex adaptive system, CAS) 理论建立群体协作词条编辑模型,利用 Netlogo 仿真软件实现对群体协作词条编辑的建模仿真,是对大众交互的互联网环境下群体智能(以下称为网络群体智能)和社会计算这门新兴学科研究的有益探索,是信息科学与社会科学的交叉研究,可深化对网络群体智能和以上 3 个社会科学问题的认识.

1 群体协作词条编辑特性分析

在维基百科中,词条相当于一个针对某特定主题的“黑板”,任何人遵循一定规则都可以对词条进行增、删、改等编辑,但对浏览者而言词条仍是针对特定主题的完整体.如图 1,词条 I 的 m 个版本 $\{v_1, v_2, \dots, v_m\}$ 按编辑时间先后顺序均被 Wiki 系统保存下来, v_1 是由编辑者 u_1 创建的最初版本,版本 $v_i (1 \leq i \leq m)$ 是由编辑者 $u_j (1 \leq j \leq n, i \leq j)$ 通过编辑动作 e_i 所形成的版本. e_i 包括多种类型,如创建词条、修

改词条、编辑词条链接和词条分类等,其中,编辑者 u_n 对版本 v_m 的贡献(如修改量)可用 o_{nm} 表示.维基中众人对公共词条的共同编辑相当于他们同时在一空白纸上作画,得到大家共识的部分才能因为笔墨线条的不断加重而“涌现”出来^[13].词条演化的主要驱动力来源于编辑者对词条的持续编辑及其交互协作(体现为词条版本增加,词条内容质量提高).持续不断地编辑交互协作使频繁发生的破坏行为和错误内容被快速而高效地纠正,依靠群体协作使词条质量不断改善提高,达到了“真理越辩越明”的效果,实现词条从初始阶段(低质量词条)到高级阶段(高质量词条)的演化,直至达到至善至美的水平.

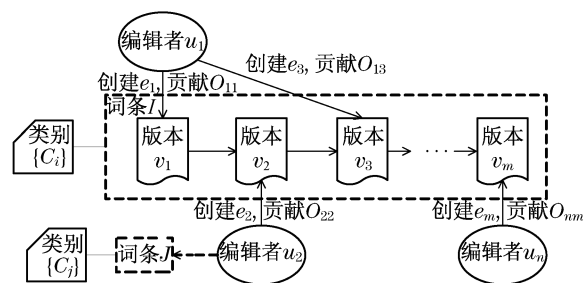


图 1 Wiki 基本机制

Fig.1 Basic mechanism of Wiki

维基允许互联网上不同用户以浏览器作为客户端直接修改网页产生内容,是无集中控制下群体协作的一种典型形态.采用维基机制的站点形成面向社区、大众参与的对等生产^[14],可视为人类计算^[15]和群体智能的重要形式.维基百科是有人参与的知识生产系统,具有开放共享、互动协作、平等中立、简单快捷等特点.由于人的智能性、主动性和适应性,能够了解其所处的周围环境,预测其变化,按照设定的目标行动,因此,维基百科词条编辑者和浏览者(称其为主体)对词条演化起着决定性作用,是维基百科词条“适应性、自组织性”的根本来源.随着词条不断演化,编辑群体结构也逐渐演化为由不同小社区组成的网络,群体结构趋于稳定,具有小世界性、抱团性和层次性^[7].编辑交互网络累积度分布符合

漂移幂律分布^[8],参与主体的行为、角色等呈现差异性,具有合作—竞争网络的特性.

1994 年 Holland 提出的复杂适应系统理论是研究复杂系统的重要研究成果,其核心理论“信息涌现理论”成为信息自组织研究的重要理论基础. CAS 理论的基本思想可以概括如下^[16]:系统中的成员称为具有适应性的个体(adaptive agent),简称主体. 所谓具有适应性,是指主体能够与环境以及其他主体进行交互作用. 主体在这样持续不断地交互作用过程中,不断“学习”或“积累经验”,并根据学习到的经验改变自身的结构和行为方式. 最重要的是, CAS 理论认为,正是这种主动性以及它与环境的反复相互作用,才是系统发展和进化的基本动因. 自组织过程会出现涌现现象,涌现是 CAS 的一个重要特征. 通常,将涌现用来指称这样的微—宏观效应——“因局部组分之间的交互而产生系统全局行为”^[17]或“缘起于微观的宏观效应”^[18]. Holland 归纳出了复杂适应系统通用的 4 个特性(聚集、非线性、流、多样性)和 3 个机制(标识、内部模型、积木)^[19],他认为复杂适应系统的其它共性都可以通过这 7 个基本点的适当组合“派生”出来. 利用复杂适应系统理论可论证群体协作词条编辑具有复杂适应性.

2 基于 CAS 理论的群体协作词条编辑建模

词条编辑群体为编辑出高质量词条而相互编辑交互,在模型的构建中,编辑群体中每个用户表示为具有适应能力的 Agent 主体. 基于 CAS 理论,根据维基百科用户(编辑者和浏览者)行为特点及实证分析,可将维基百科用户抽象为 5 种主体,其角色分类如表 1 所示,并构建群体协作词条编辑模型(collective collaboration article edit model, CCAEM),包括环境模型和主体模型.

表 1 主体角色分类及描述

Tab.1 Sort and description of agent roles

主体组成	主体角色描述
内容添加者	专注于词条内容的添加,很少进行内容的修改、删除等
内容修改者	专注于词条内容的修改,主要修改其他编辑者添加的内容
内容删除者	专注于词条内容的删除,很少进行内容的添加、修改等
多样编辑者	积极参与词条内容的添加、修改和删除等,既能修改其他编辑者编辑的内容,又能修改自己编辑的内容
内容浏览者	专注于词条内容的浏览,不进行词条内容编辑

2.1 环境模型

环境由 $n \times n$ 个方格组成,方格总数可表征某个词条的语量,所有方格区域是主体能够访问覆盖的区域,设定一个主体每次可以访问一个方格区域.

2.2 主体模型

主体根据环境条件及行为准则来调整自身的行为,以达到自身期望目标. 在此采用“在黑板上涂色”的形式建立主体模型.

a. 主体属性

主体属性是对其自身情况的描述,定义如下:

(X, Y): 二维网格中的坐标; r : Agent 的视野半径,所能观察到的区域半径,取值在 1~3 之间.

b. 交互规则

主体交互规则如表 2 所示,主体按各自的行为规则在屏幕方格上涂色,并根据环境的变化,调整各自行为. 黑色表示空白内容,红色表示正确内容,紫色、黄色和蓝色分别表示不完全正确内容 B、C、D. 屏幕内红色格数的多少表征词条质量的高低,当红色格数达到很高的比例时,表明词条质量很高,编辑群体趋于达成共识,群体智能水平很高.

表 2 主体交互规则

Tab.2 Agent interaction rules

主体	交互规则	规则意义
内容添加者	遇黑色,则涂为紫色	遇到空白位置,则添加 B 类不完全正确内容
内容修改者	遇黄色且有一个以上浏览者参与,则涂为红色;遇黑色和黄色以外的颜色且没有浏览者参与,则涂为黄色;遇黑色和黄色以外的颜色且有浏览者参与,则涂为红色	遇到 C 类不完全正确内容且有一个以上浏览者参与,则将其修改为正确内容 A;遇到 C 类以外的内容且没有浏览者参与,则将其修改为 C 类内容;遇到 C 类以外的内容且有浏览者参与,则将其修改为 A 类内容
内容删除者	遇红色和黑色以外的颜色,则涂为黑色	遇到已添加的不完全正确内容,则进行删除
多样编辑者	遇黑色、紫色,则涂为蓝色;遇黄色、蓝色且有浏览者参与,则涂为红色;遇黄色且没有浏览者参与、且参与的内容修改者与多样编辑者总数小于 2,则涂为黑色	遇到空白位置,则添加 D 类不完全正确内容;遇到 B 类内容,则将其修改为 D 类内容;遇到 C 类内容或 D 类内容且有浏览者参与,则将其修改为 A 类内容;遇到 C 类内容且没有浏览者参与、且参与的内容修改者与多样编辑者总数小于 2,则将其删除
内容浏览者	遇红色,则复制一个自己	遇到正确内容,则介绍一个新的浏览者加入

3 仿真实验

基于群体协作词条编辑模型,利用 Netlogo 软件构建仿真平台,从词条演化的视角实现对群体协作词条编辑建模仿真.突出的核心思想是:词条从初始阶段(低质量词条)逐渐演化到高级阶段(高质量词条),体现了群体编辑交互协作群体智能的涌现;词条演化的主要驱动力来源于主体对词条的持续编辑及其交互协作(体现为词条版本数增加,词条质量提高).

Netlogo 程序运行窗口中的黑色屏幕是各种主体的活动场所,该屏幕由 20×20 个小方格组成.将词条质量从高到低依次分为 $L1, L2, L3$ 和 $L4$,主体出现概率配置如表 3 所示(5 种主体初始状态及形状如图 2(a)所示,5 种不同颜色形状标识代表 5 种不同主体),仿真步长设为 1,仿真时刻为 t ,其最大值 T 设为 10 000.根据词条质量与词条声誉、吸引力的关系分别将质量为 $L1, L2, L3$ 和 $L4$ 的内容浏览者出现概率设为 0.8, 0.7, 0.3 和 0.2.程序开始运行后,主体向随机方向移动,每个仿真步长移动 3 步,按各自的行为规则在屏幕上涂色(根据维基的时序编辑特点,在此设置每个仿真

步长内只有一个主体进行编辑涂色),并根据环境的变化,调整各自行为.分别对不同配置的模型进行 600 次仿真,得到词条演化仿真的统计平均结果,如表 4 所示;群体协作词条编辑不同时刻仿真场景(主体编辑涂色情况)运行状态如图 2 所示.

表 3 主体参与出现概率配置
Tab.3 Configuration of agent appearance probabilities

主体	L1	L2	L3	L4
内容添加者	0.10	0.20	0.40	0.67
内容修改者	0.10	0.20	0.17	0.16
内容删除者	0.10	0.10	0.23	0.10
多样编辑者	0.70	0.50	0.20	0.07
内容浏览者	0.80	0.70	0.30	0.20

表 4 词条编辑仿真结果
Tab.4 Article edit simulation results

词条质量	红格比例	红格均值
L1	0.9~1.0	373
L2	0.7~0.9	316
L3	0.4~0.7	233
L4	0.1~0.4	132

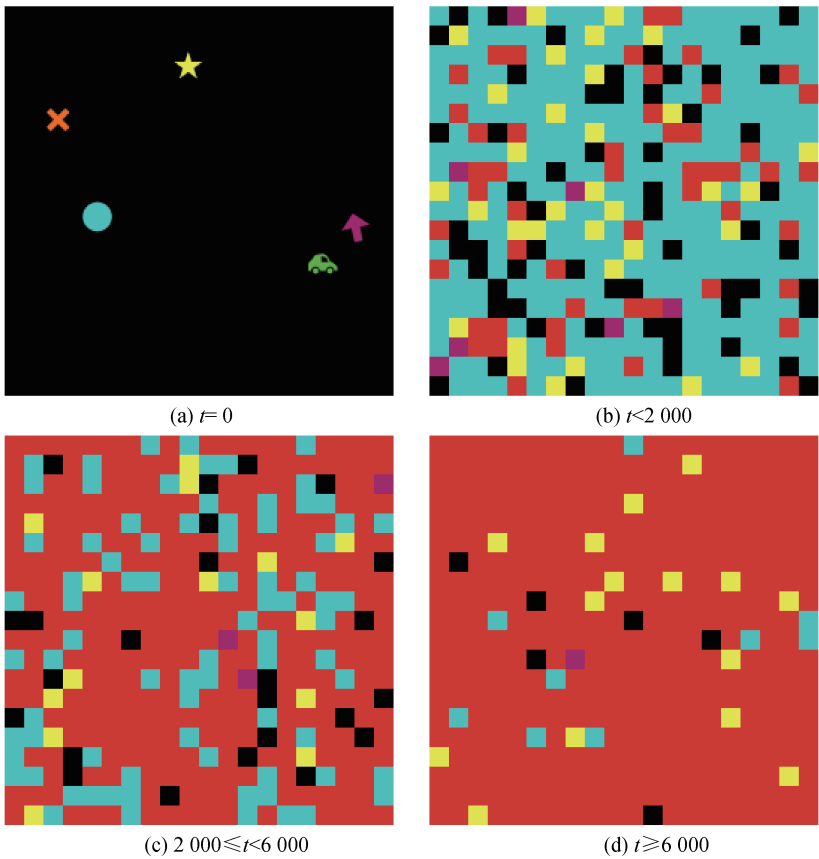


图 2 不同时刻的群体协作词条编辑仿真场景
Fig.2 Simulation scenes of collective collaboration article edit at different time

由表3、表4可知,主体出现概率配置不同,生成的词条质量也不同,得到主要结论如下:

a. 多样编辑者出现概率越大,对应生成词条质量越高. 当其出现概率达到0.7时,生成词条质量达到L1,最高. 说明多样编辑者在词条从低级到高级演化过程中扮演重要角色,是词条质量提升的主要推动者. 内容添加者出现概率越大,对应生成词条质量越低. 当其出现概率达到0.67时,生成词条质量为L4,最低. 说明内容添加者对词条质量提升作用不大,其出现概率过大会阻碍词条质量提升.

b. 从L4到L1的4种配置中,内容修改者和内容删除者的出现概率变化不大,多样编辑者和内容添加者出现概率变化较大,呈现出多样编辑者和内容添加者出现概率分别增大和减小,说明多样编辑者和内容添加者的出现概率对词条质量具有较大影响. 低质量词条演化过程中内容修改者占据统治地位,随着词条质量由L4到L3到L2再到L1的逐步提高,内容修改者和多样编辑者的角色地位逐渐发生变化,内容修改者的地位逐渐减弱,多样编辑者的地位逐渐增强,呈现出此消彼长的现象;在L4和L1中多样编辑者和内容添加者出现概率配置中存在对称性(0.07对应0.10, 0.67对应0.70).

c. 多样编辑者修改其以前自己编辑内容的概率较内容修改者要大,在高质量词条演化过程中编辑者自我修改比例明显比低质量词条要高,说明编辑者“自我修改”(对以前自己编辑内容的“否定之否定”)行为对提升词条质量起到重要促进作用,体现了编辑者对词条认知水平螺旋式上升过程.

d. 多样编辑者出现概率达到0.7左右时,生成词条质量会达到很高水平,说明多样编辑者出现概率的配置存在阈值0.7(近似符合黄金分割律比例介于0.6~0.8),当达到阈值时,即使再增大出现概率,对词条质量提升影响也不大,反而会增加主体编辑词条创作成本,投入产出率不高;在对编辑主体出现概率进行配置时可以黄金分割律为指导,提高投入产出率,使群体绩效趋于最大化,以达到“事半功倍”的效果.

由图2、图3(横坐标为仿真时间,纵坐标为红格数量)可知,根据红色块数的变化发现:随着时间的推移,词条质量逐渐升高,编辑群体朝着更有序的方向演化,表现出3个演化阶段:

a. 当 $t < 2\,000$ 时,是词条编辑演化的初期,红

色块数较少. 编辑者间交互不多,个体间相互影响不强,交流争论有限;词条内容不准确,存在较大冗余,词条质量较低;词条知名度不高,浏览者较少;编辑行为注重完整性,使词条内容更完整、全面;主要以量的积累为主,群体智能水平较低.

b. 当 $2\,000 \leq t < 6\,000$ 时,是词条编辑演化的中期,红色块数较多. 词条质量和知名度较高,浏览者较多;编辑者间交互增多,个体间相互影响增强;观点、知识不断碰撞、融合,新观点、新知识逐渐涌现;编辑行为注重准确性,使词条内容更正确、可信;词条冗余内容减少,正确内容大量增加,是量积累基础上质的提升,群体智能水平较高.

c. 当 $t \geq 6\,000$ 时,词条编辑演化的后期,红色块数很多,变化逐渐趋于平缓. 词条质量和知名度很高,浏览者很多;编辑行为注重可读性,使词条内容更精炼、易懂;编辑者间交互减少,群体逐渐达成共识,基本达到动态平衡,群体结构趋于稳定,是质提升基础上量的微调,群体智能水平很高.

由以上分析可知,词条编辑演化不同阶段的主要矛盾不同. 随着时间推移,主要矛盾发生变化,即“从注重完整性到注重准确性再到注重可读性”,存在“去冗余”过程,即存在“语量与语义之间此消彼长”的过程,最终语量与语义之间达到动态平衡,这体现了精益涌现^[20]的思想,统计分析可知这个平衡的临界点大概介于6 000~8 000(与总仿真时间10 000的比例约为0.6~0.8)之间,与黄金分割比例0.618近似,即语量与语义平衡临界点近似为黄金分割点,符合黄金分割律. 词条编辑演化似乎遵循着黄金分割律,当达到量与质的动态平衡后,词条质量会达到较高水平,令人赏心悦目,具有美学意义.

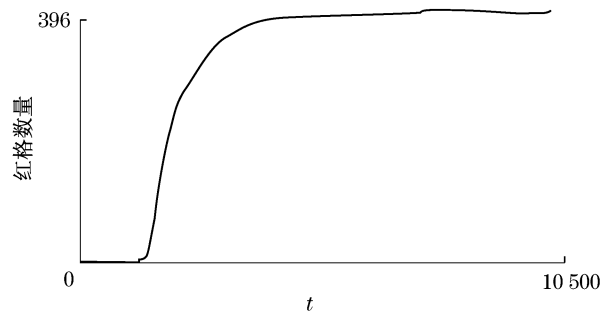


图3 L1质量词条编辑演化红色格数变化曲线图
Fig.3 Variational graph of red grid number of L1 article edit evolution

4 结 论

本文基于CAS理论建立了群体协作词条编辑模型CCAEM,利用Netlogo软件构建仿真平台实现了对群体协作词条编辑的多主体建模仿真.仿真实验表明多样编辑者是词条质量提升的重要驱动力,编辑者“自我修改”行为对提升词条质量起到重要促进作用;主体出现概率配置遵循黄金分割律时,可使群体绩效趋于最大化;词条编辑存在从低到高的“三阶段演化”,并遵循着黄金分割律.研究深化了对词条编辑演化、网络群体智能和社会计算的认识,在知识管理创造、群体协作决策和群体绩效管理等领域具有推广应用价值.

参考文献:

- [1] Dennis W, Bernardo H. Cooperation and quality in Wikipedia[C]//WikiSym 2007. Montreal, 2007: 157 - 164.
- [2] Cattuto C, Loreto V, Pietronero L. Semiotic dynamics and collaborative tagging[J]. PNAS, 2007, 104(5): 1461 - 1464.
- [3] Liu D, Hua X S, Yang L J, et al. Tag ranking[C]//Proceedings of the 18th International World Wide Web Conference (WWW2009). Madrid, 2009: 351 - 360.
- [4] Zhao D J, Jiang J, Zhang H S, et al. Research on internet evolution mode based on user behavior[C]//2010 Asia-Pacific Youth Conference on Communication Technology. Kunming, 2010: 835 - 839.
- [5] 赵东杰, 张海粟, 杨海涛, 等. 基于网络交互演化的智能涌现研究[J]. 计算机科学, 2010, 37(10A): 112 - 116.
- [6] Zhao D J, Zhang H S, Han Y N, et al. An approach to study collective intelligence based on networked data mining[C]//2010 3rd International Conference on Computational Intelligence and Industrial Application. Wuhan, 2010: 239 - 243.
- [7] 赵东杰, 郝黎, 李德毅, 等. 维基百科词条编辑特性研究[J]. 计算机科学, 2011, 38(10A): 153 - 156.
- [8] Zhao D J, Yang H T, Jiang J, et al. A research for the centrality of article edit collective in Wikipedia[C]//2011 International Conference of Information Technology, Computer Engineering and Management Sciences (ICM 2011). Nanjing, 2011: 363 - 366.
- [9] Ponzetto S, Strube M. Deriving a large scale taxonomy from Wikipedia[C]//Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI-07). Vancouver, 2007: 1440 - 1447.
- [10] Yeh E, Ramage D, Christopher D M, et al. WikiWalk: random walks on Wikipedia for semantic relatedness[C]//Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. Suntec, 2009: 41 - 49.
- [11] Weld D S, Wu F, Adar E, et al. Intelligence in Wikipedia[C]//AAAI'08 Proceedings of the 23rd National Conference on Artificial Intelligence. Chicago, 2008: 1609 - 1614.
- [12] Adler B T, Alfaro L D. A content-driven reputation system for the Wikipedia[C]//Proceedings of the 16th International Conference on World Wide Web Conference (WWW2007). Banff, 2007: 261 - 270.
- [13] 张树人. 从社会性软件, Web 2.0 到复杂适应信息系统研究[M]. 北京: 中国人民大学出版社, 2006.
- [14] Tapscott D, Williams A. Wikinomics: how mass collaboration changes everything [M]. Richmond: Portfolio Books, 2006.
- [15] von Ahn L. Human computation [D]. Pittsburgh: Carnegie Mellon University, 2005.
- [16] 胡晓峰, 罗批, 司光亚, 等. 战争复杂系统建模与仿真[M]. 北京: 国防大学出版社, 2005.
- [17] de Wolf T, Holvoet T. Emergence versus self-organization: different concepts but promising when combined[M]//Brueckner S, Di Marzo Serugendo G, Karageorgos A, et al. Engineering Self Organising Systems: Methodologies and Applications. Berlin: Springer-Verlag, 2005.
- [18] Russ A. Emergence explained: abstractions: getting epiphenomena to do real work[J]. Complexity, 2006, 12(1): 13 - 26.
- [19] Holland J H. Hidden order: how adaptation builds complexity[M]. Reading, MA: Addison-Wesley Publishing Company, 1995.
- [20] 张英华, 蒋丽华. 复杂系统“精益涌现”的形成机理研究[J]. 天津师范大学学报: 社会科学版, 2011(3): 72 - 76.