基于拓扑势熵的维基百科词条编辑演化研究

赵东杰[1,2],王 华[3,4],李德毅[5],李 智[1],杨海涛[1],陈桂生[5]

1. 装备学院重点实验室,北京 101416
2. 中国人民解放军 63628 部队,北京 101601
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

摘要 网络群体智能涌现问题是信息科学和社会科学的多学科交叉研究问题。基于映射思想将其映射为维基百科词条编辑演化问题,提出词条编辑演化研究方法框架;以维基百科高质量词条编辑历史数据为数据集,以编辑者为节点,以编辑者间编辑交互关系为连边,构建词条编辑交互网络,通过建立网络演化测度拓扑势熵实现对词条编辑演化研究。实验分析表明,网络结构和词条总体上朝着有序方向演化,演化呈现出"从注重完整性到注重准确性再到注重可读性"的由低到高 3 个阶段,直至群体结构趋于稳定,结构具有无标度性,词条质量和群体智能达到很高水平;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。方法框架有效,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;词条编辑演化;拓扑势熵;网络化数据挖掘;群体智能;社会计算

中图分类号 TP39,N94,O415,C81　　**文献标识码** A　　**doi** 10.3981/j.issn.1000-7857.2012.04.011

Article Edit Evolution in Wikipedia Based on Topology Potential Entropy

ZHAO Dongjie[1,2], WANG Hua[3,4], LI Deyi[5], LI Zhi[1], YANG Haitao[1], CHEN Guisheng[5]

1. *The Key Laboratory, Academy of Equipment, Beijing 101416, China*
2. *No. 63628 Troop of PLA, Beijing 101601, 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*

Abstract The emergence of web collective intelligence is an interdisciplinary research topic involving information science and social science. The topic is mapped to the problem of article edit evolution in Wikipedia based on mapping idea, then the framework of article edit evolution is proposed, the article edit interaction networks are constructed based on the history data of the featured articles in Wikipedia; in the networks, a node is editor and a link is the edit interaction connection between editors. The topology potential entropy is developed to study the article edit evolution. Results show that networks structure and articles evolve toward the orderly direction; the evolution has experienced three development stages from low to high with the emphasis on integrity, accuracy, and readability. The collective structure gradually becomes stable, and it has a scale-free property, both article quality and collective intelligence reach at a high level. There is a process from seesaw-like complementarity to dynamic balance between word quantity and word meaning, the critical point of balance is closed to the golden section point; article edit evolution follows the golden section law. The framework is effective, and the research deepens the knowledge of article edit evolution, web collective intelligence, and social computing.

Keywords Wikipedia; article edit evolution; topology potential entropy; networked data mining; collective intelligence; social computing

收稿日期:2011-12-31;修回日期:2012-01-16

基金项目:国家自然科学基金项目(69120912,61035004)

作者简介:赵东杰,博士研究生,研究方向为信息系统设计优化和智能信息处理等,电子信箱:zhaodj2006@126.com;李德毅(通信作者),研究员,中国工程院院士,研究方向为数据挖掘和人工智能等,电子信箱:leedeyi@126.com

0　引言

随着信息技术发展和互联网普及,21 世纪初互联网逐渐由 Web l.0 单纯通过网络浏览器浏览 html 网页模式向内容更丰富、联系性更强、工具性更强的 Web 2.0 互联网模式发展,互联网已进入 Web 2.0 时代。Web 2.0 将互联网和社会网进一步结合,注重大众用户的参与以及用户之间的交互作用。2011 年 2 月,美国哈佛大学公布了当前及未来需要解决的十大社会科学问题,其中"人类如何增加自身集体智慧"、"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"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网络群体智能演化过程中所涌现出来的规律,仍缺乏有效研究方法,有待深入研究。本文以维基百科高质量词条编辑历史数据为数据集,通过建立网络演化测度拓扑势熵实现对维基百科词条编辑演化研究,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　词条编辑演化研究方法框架

图 1 为词条编辑演化研究方法框架。为深入研究大众交互的互联网环境下群体智能(以下称网络群体智能)涌现问题,以 Wikipedia 历史数据为目标数据。作为研究载体,由于词条从初始阶段(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,因此将问题一级映射为词条编辑演化问题。由于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(体现为词条版本增加,词条内容质量提高),因此将此问题二级映射为词条编辑群体结构演化问题,可通过构建词条编辑交互网络研究词条编辑群体的结构演化。

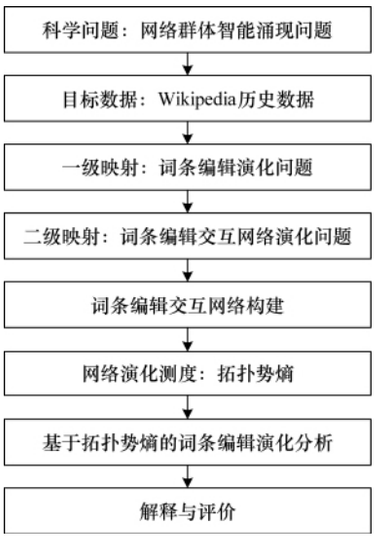

图 1　词条编辑演化研究方法框架

Fig. 1　Research framework of article edit evolution

近年来,熵作为描述复杂系统结构的物理量,在复杂系统理论中受到越来越多的关注,成为研究复杂系统的一个重要工具。在信息科学中,熵可以用来表示事物的不确定性。熵作为系统状态的一种定量描述,能够用来表征系统的复杂性、有序程度和系统变化的方向或趋势。一般而言,熵值愈小,对应的宏观态愈加有序。从系统动力学的角度看,词条编辑群体可以看作是一类非线性动力系统,系统拓扑结构的综合性质决定了整体的组织效率和行为特征。拓扑势熵综合考虑了网络节点的不确定性因素,可以比较准确地显示网络拓扑中节点间的位置差异和节点本身的连接特性,刻画网络中心节点特征[7,13]。本文通过拓扑势熵刻画和度量系统结构的有序性,研究词条编辑演化,深化对网络群体智能涌现的认识。

2　拓扑势熵定义

给定网络 $G=(V,E)$,其中 $V=\{v_1,v_2,\cdots,v_n\}$为节点集,网络规模$|V|=n$,$E\subseteq V\times V$ 为边的集合。将网络 G 看作是一个包含 n 个节点及其相互作用的系统,则每个节点周围存在一个虚拟的作用场,网络中的任何节点都将受到其他节点的联合作用,由此在整个网络拓扑上确定了一个数据场,称之为拓扑势场(Topology Potential Field)。为刻画节点间相互作用的局域性以及影响随着网络距离而快速衰减的特性,可采用代表短程场作用的高斯势函数来描述节点间的相互作用,据此网络 G 中任一节点 v_i 的拓扑势可表示为[14]

$$\varphi(v_i)=\sum_{j=1}^{n}\left[m_j\times e^{-\left(\frac{d_{ij}}{\sigma}\right)^2}\right]=\varphi_i \qquad (1)$$

其中,d_{ij} 表示节点 v_i 和 v_j 间的距离,以最短路径长度来度量;影响因子 σ 用于控制每个节点的影响范围;$m_j\geqslant 0$ 表示节点 $v_j(j=1,2,\cdots,n)$的质量,用于描述每个节点的固有属性。这里,假设每个节点的质量相等,每条边的属性值均为 1,则节点的重要性仅取决于节点在网络中拓扑位置的差异性,由此可得到简化的拓扑势公式

$$\varphi(v_i)=\sum_{j=1}^{n}e^{-\left(\frac{d_{ij}}{\sigma}\right)^2}=\varphi_i \qquad (2)$$

σ 用于调节一个节点影响力衰减的速度与范围,其具体取值可以参考数据场中的方法[14],则拓扑势熵(Topology Potential Entropy)定义为

$$H=-\sum_{i=1}^{n}I_i\ln I_i \qquad (3)$$

其中,$I_i=\frac{\varphi_i}{Z}$,$Z=\sum_{i=1}^{n}\varphi_i$。

为了消除节点数目对拓扑势熵的影响,对拓扑势熵进行归一化处理,定义

$$H'=\frac{H-H_{\min}}{H_{\max}-H_{\min}}=\frac{-2\sum_{i=1}^{n}I_i\ln I_i-\ln 4(n-1)}{\ln n^2-\ln 4(n-1)} \quad (4)$$

为网络的标准拓扑势熵(Standard Topology Potential Entropy),其中 n 为网络的节点数目。当网络完全均匀,即 $I_i=1/n$ 时,H 和 H' 分别取最大值 $H_{\max}$ 和 1;当网络中所有节点都与某一个中心节点相连时,网络最不均匀,H 和 H' 分别取最小值 $H_{\min}$ 和 0。

3　词条编辑交互网络演化分析

本文随机抽取 60 个高质量词条(Featured Articles)作为数据集,利用 API 下载词条从开始创建时刻起的 600 个历史版本,包括作者信息和版本内容。以网络化数据挖掘思想方法为指导,以高质量词条编辑历史数据为目标数据,利用文本分析方法从句子的粒度上分析相邻版本的文本差异,以句子的作者(以下称为编辑者)为节点,将编辑者间编辑关系(修改、删除和添加等)为连边,对词条编辑交互进行网络化表示,构建词条编辑交互网络,具体方法参见文献[8]。

词条的一个版本对应一个编辑交互网络,从初始版本演化到第 600 个版本,对应网络从初始时刻演化到第 600 个时刻。由于 60 个词条的拓扑势熵特性相似及篇幅所限,本文以 4 个词条"Sheep"、"Yao Ming"、"Turing Machine"和"Microsoft"的拓扑势熵特性曲线为例进行说明,如图 2 所示。

根据熵值变化发现,随着时间推移,网络拓扑结构总体

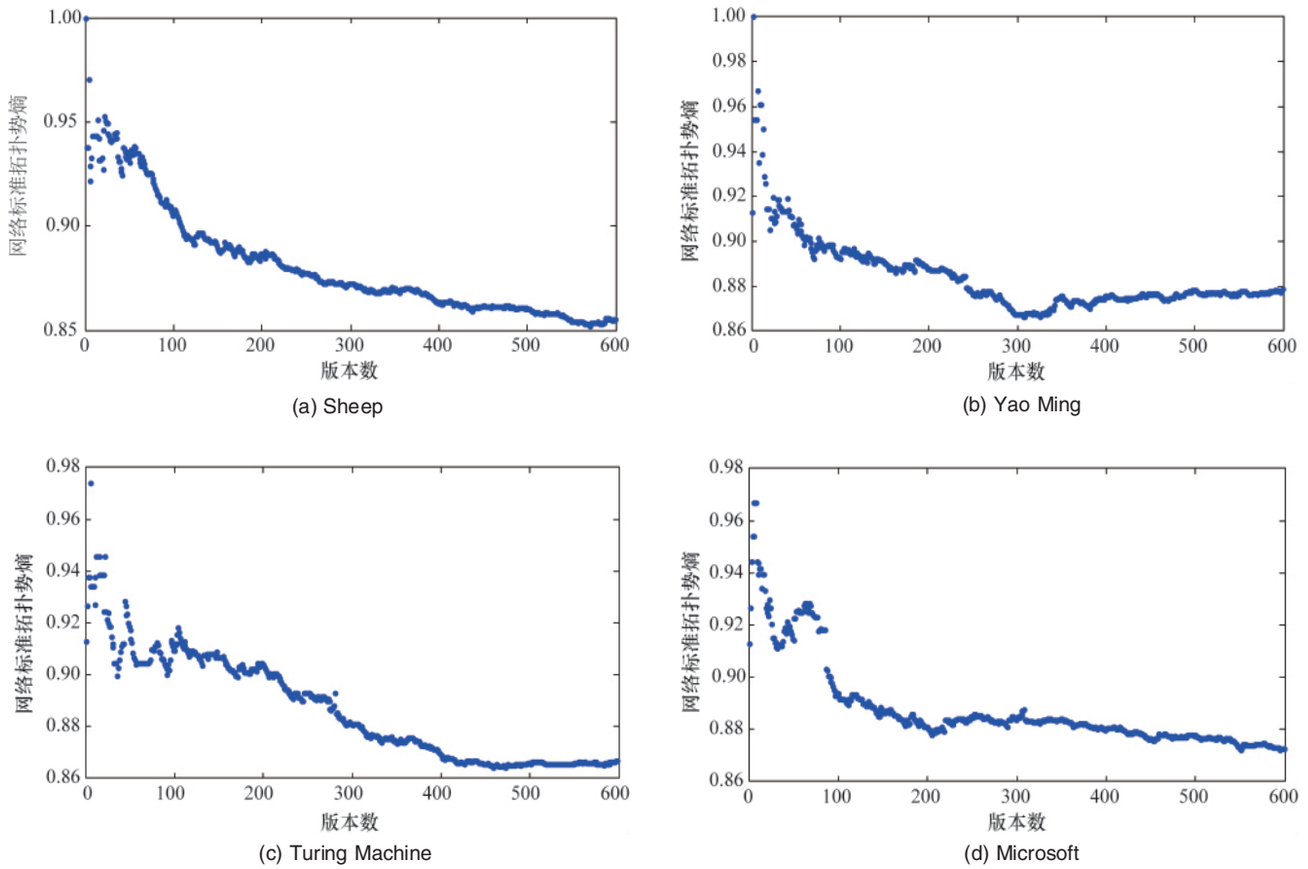

图 2　词条编辑交互网络标准拓扑势熵演化过程

Fig. 2　Standard topology potential entropy evolutions for article edit interaction networks

上朝着更有序的方向演化(熵值总体上逐渐减小,节点的度逐渐呈现不均匀分布,累积度分布符合漂移幂律分布[7],呈现出无标度特性[8]),表现为 3 个演化阶段,如表 1 所示。

(1) 当版本数<100 时,是网络拓扑演化的初期,熵值较大,总体下降,但变化比较迅猛,网络分布较均匀;编辑群体异质性较低,编辑者间交互不多,个体间相互影响不强,交流争论有限,词条内容存在较大冗余,词条质量较低;词条知名

表 1　词条不同演化阶段特性描述

Table 1　Characteristic descriptions for the different evolution stages of article

阶段	主要矛盾	熵值变化	互动程度	群智水平
1	完整性	迅猛	较弱	较低
2	准确性	缓慢	较强	较高
3	可读性	平缓	减弱	很高

度不高,浏览者较少,编辑者数量不多,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。

(2) 当 100≤版本数<400 时,是网络拓扑演化的中期,熵值下降、变化存在波动,但较前一阶段变得缓慢,网络逐渐显现出小世界特性;编辑群体异质性变大,词条质量和知名度较高,浏览者较多,编辑者数量较多,编辑行为注重准确性,使词条内容更正确、可信;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,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,达到了"理越辩越明"的效果,群体智能水平较高。

(3) 当版本数≥400 时,是网络拓扑演化的后期,熵值变化逐渐趋于平缓,网络逐渐显现出无标度特性;编辑群体异质性较大,词条质量和知名度很高,浏览者很多,编辑者数量很多,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,拓扑势熵能够有效表征词条演化过程。词条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[15]的思想。统计分析可知,这个平衡的临界点大概介于 360—480 之间,与总版本数 600 的比例介于 0.6—0.8 之间,与黄金分割率 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。

随着词条不断演化,编辑群体逐渐具有偏好依附性,在兴趣爱好等驱使下会针对词条不同段落内容进行选择性编辑,形成多个关注度较高的局域性编辑区域;群体结构也逐渐演化为由不同小社区组成的网络,具有抱团性、层次性[7],群体结构趋于稳定,词条质量达到很高水平;同时群体对词条的认知水平是螺旋式上升的,群体智能水平也是逐渐提高的,具有层次性。

4　结论

本文提出词条编辑演化研究方法框架,以维基百科高质量词条为数据集,构建词条编辑交互网络,建立网络演化测度拓扑势熵实现对维基百科词条编辑演化研究。实验分析表明网络结构和词条总体上朝着有序方向演化,演化呈现出由低到高 3 个阶段,演化遵循着黄金分割律;方法框架有效,为网络群体智能和社会计算的建模仿真提供了理论支撑,在知识管理创造、群体协作决策和组织绩效管理等领域具有推广应用价值。

参考文献(References)

[1] Dennis W, Bernardo H. Cooperation and quality in Wikipedia [C]. WikiSym 2007, Montreal, Canada, October 21–24, 2007.

[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]. 18th International World Wide Web Conference (WWW2009), Madrid, Spain, April 20–24, 2009.

[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, China, August 7–8, 2010.

[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.
Zhao Dongjie, Zhang Haisu, Yang Haitao, *et al*. *Computer Science*, 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, China, December 4–5, 2010.

[7] 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, China, September 24–25, 2011: 363–366.

[8] 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.
Zhao Dongjie, Hao Li, Li Deyi, *et al*. *Computer Science*, 2011, 38(10A): 153–156.

[9] Ponzetto S, Strube M. Deriving a large scaletaxonomy from Wikipedia[C]. 22nd National Conference on Artificial Intelligence (AAAI -07), Vancouver, British Columbia, July 22–26, 2007.

[10] Yeh E, Ramage D, Christopher D M, *et al*. WikiWalk: Random walks on Wikipedia for semantic relatedness[C]. 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs -4), Suntec, Singapore, August 7–8, 2009.

[11] Weld D S, Wu F, Adar E, *et al*. Intelligence in Wikipedia [C]. 23rd National Conference on Artificial Intelligence, Chicago, USA, July 13–17, 2008: 1609–1614.

[12] Adler B T, Alfaro L D. A content-driven reputation system for the Wikipedia [C]. 16th International Conference on World Wide Web Conference (WWW2007), Banff, Canada, May 8–12, 2007.

[13] 肖俐平, 孟晖, 李德毅. 基于拓扑势的网络节点重要性排序及评价方法[J]. 武汉大学学报: 信息科学版, 2008, 33(4): 379–383.
Xiao Liping, Meng Hui, Li Deyi. *Geomatics and Information Science of Wuhan University*, 2008, 33(4): 379–383.

[14] He N, Gan W Y, Li D Y. Evaluate nodes importance in the network based on data field theory [C]. 2nd International Conference on Convergence Information Technology (ICCIT 2007). Gyeongju, Republic of Korea, November 21–23, 2007.

[15] 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.
Zhang Yinghua, Jiang Lihua. *Journal of Tianjin Normal University: Social Science Edition*, 2011(3): 72–76.

(责任编辑　马宇红,代丽)