维基百科词条编辑特性研究

赵东杰[1,5] 郝 黎[2] 李德毅[3] 王 华[4] 何 宇[1]
(装备指挥技术学院 北京 101416)[1] (北京航空航天大学 北京 100191)[2]
(中国电子系统工程研究所 北京 100840)[3] (中国航天员科研训练中心 北京 100193)[4]
(63628部队 北京 101601)[5]

摘 要 针对维基百科词条编辑特性问题,以网络化数据挖掘思想方法为指导,对高质量维基百科词条进行文本分析,判断词条相邻版本间句子差异,以编辑者为节点,编辑者间编辑交互关系为连边,构建词条编辑交互网络,通过分析网络结构特征实现词条编辑特性分析。分析表明网络具有小世界特性,度分布与强度分布相似,具有较强正相关性,其累积分布与边权重分布服从幂律分布,节点度与聚集系数具有较强负相关性,最短路径长度分布与高斯分布相似,网络具有异配性和较弱的互惠性,编辑群体具有较强异质性、抱团性;深化了对词条编辑交互过程和群体智能的认识。

关键词 维基百科,词条编辑交互网络,网络化数据挖掘,群体智能
中图法分类号 TP391.9 **文献标识码** A

Research on Article Edit Characteristic in Wikipedia

ZHAO Dong-jie[1,5] HAO Li[2] LI De-yi[3] WANG Hua[4] HE Yu[1]
(Academy of Equipment Command and Technology, Beijing 101416, China)[1]
(Beihang University, Beijing 100191, China)[2] (Institute of Electronic System Engineering, Beijing 100840, China)[3]
(China Astronaut Training and Research Center, Beijing 100193, China)[4] (63628 Troops, Beijing 101601, China)[5]

Abstract Aiming at the problems of article edit characteristic in wikipedia, under the direction of the idea of networked data mining, featured articles in wikipedia were analysed by text processing to fiind the difference of sentence between adjacent versions, the article edit interaction networks were constructed, where the node is editor and the link is the edit interaction connection between editors, then the article edit characteristics in wikipedia were analysed by the empirical analysis of the nine networks. Results show that all networks have small-world properties, strong positive degree-strength correlation and negative degree-clustering coefficient distribution, their degree distributions are similar to strength distributions, their cumulate distributions and link weight distribution are power-law distribution, shortest path length distributions are similar to gauss distribution, all networks have degree and strength disassortativity and weak reciprocity, furthermore the edit collective have strong heterogeneity and community structure, which deepens the knowledge of the process of article edit interaction and collective intelligence.

Keywords Wwikipedia, Article edit interaction network, Networked data mining, Collective intelligence

1 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 1.0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互联网模式的发展,互联网已进入Web 2.0时代。Web 2.0应用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。维基百科是知名的Web 2.0应用,利用互联网上大众用户的集体参与来创作百科条目,是利用大众普遍参与、进行编辑交互形成群体智能的典型应用,为研究群体智能提供了高价值数据资源。目前,一些研究者已展开对大众交互的互联网环境下人的群体行为的研究[1-5],对维基百科的研究,主要集中在语义知识挖掘[6-8]和优良条目的自动发现与挖掘方面[1,9]。但是,对于如何有效地利用大众不断参与的群体智能演化过程中所涌现出来的规律,如何进一步挖掘其在无集中控制的大众合作编辑中所蕴含的知识,仍有待深入的研究。

本文以网络化数据挖掘[10]思想方法为指导,以维基百科

本文受国家自然科学基金项目(69120912,61035004),国家973计划项目(2007CB310804)资助。
赵东杰(1980—),男,博士生,工程师,主要研究方向为通信与信息系统、群体智能等,E-mail:zhaodj2006@126.com;郝 黎(1987—),女,硕士,主要研究方向为不确定性人工智能、数据挖掘等;李德毅(1944—),男,研究员,主要研究方向为数据挖掘、人工智能等;王 华(1980—),女,助理研究员,主要研究方向为数据统计分析与评估;何 宇(1983—),男,博士,主要研究方向为通信与信息系统。

高质量词条编辑历史数据为数据集，构建编辑者间的词条编辑交互网络，分析网络结构特性，实现对维基百科高质量词条编辑特性实证研究，是对群体智能研究的有益探索，可深化对群体智能的认识。

2 维基百科词条编辑交互的网络化表示

2.1 数据集选取

维基百科为研究者提供了 dump 和 API，dump 可以下载整个数据库。由于英文维基百科的数量庞大性与处理复杂度，最后选择用 API 来抓取所需的数据，这样的数据量较小也较容易处理。本文随机抽取 9 个高质量词条(Featured Articles)“Lion”、“Oxygen”、“Sheep”、“Microsoft”、“Australia”、“Influenza”、“Yao Ming”、“Parallel”和“Turing Machine”作为数据集，分别记为 1～9。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。

2.2 词条编辑交互的网络化表示

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，从词条开始创建时间起，描绘出词条编辑者间的编辑交互关系。本文通过开源的工具 openNLP 对文本进行分句，采用 LCS 算法[11]实现句子差异比较。

如果编辑者 A 在维基中创建了一个词条，形成版本 1，编辑者 B 编辑修改了该词条 3 个句子，形成版本 2，则从 B 指向 A 有一个连边，反之连边是从 A 指向 B，边的权重为 3；而后编辑者 C 又对版本 2 中 A 贡献的 1 个句子和 B 贡献的 2 个句子进行编辑修改，形成版本 3，则从 C 指向 A 和 B 各有一个连边，其权重分别为 1 和 2。由于具有破坏性的编辑行为(一次性恶意删除或添加大量内容)对群体编辑创作词条没有实质贡献，同时编辑者编辑自己贡献词条内容(节点自连边)，不能反映编辑交互关系，在构建词条编辑交互网络时这两种编辑行为不进行连边。

利用上述方法对随机抽取的 9 个高质量词条进行网络化表示，形成 9 个有向加权网络，边权重可定义为两节点 A 和 B 间编辑交互次数，即 A 或 B 编辑 B 或 A 所贡献的句子个数。为简化研究，按无向无权网络到无向加权网络再到有向加权网络进行研究。表 1 为 9 个词条编辑交互网络(无向无权)的基本网络特征。由表 1、图 1 可知词条编辑交互网络具有小世界特性，利于个体优秀知识或行为在群体中迅速传播，达成共识。由于 9 个网络的特性相似及篇幅所限，下面选取 4 个网络(标号分别为 3、6、7、9)的特性曲线进行显示。

表 1 词条编辑交互网络基本特征

词条	点数	边数	平均度	网络直径	最短路径	聚集系数
1	400	888	4.4400	12	4.1367	0.0873
2	374	764	4.0856	12	3.7525	0.1517
3	377	627	3.3263	20	5.0349	0.1234
4	308	848	5.5065	9	3.1620	0.2731
5	268	598	4.4627	9	3.4554	0.2224
6	331	711	4.2961	10	3.3932	0.1900
7	313	624	3.9872	13	4.1859	0.1484
8	267	530	3.9700	9	3.4185	0.2217
9	285	593	4.1614	12	3.8835	0.1510
均值	325	687	4.2484	11.8	3.8247	0.1743

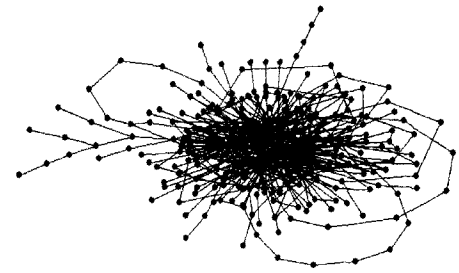

图 1 “Turing Machine”词条编辑交互网络

3 词条编辑交互网络特性分析

3.1 度分布

节点 i 的度 k_i，定义为与该节点连接的邻居节点的数目。网络中所有节点的度的平均值为网络节点的平均度，记为 $\langle k\rangle$。复杂网络拓扑分析中最基本也是最重要的一个性质就是度分布，网络中节点的度分布情况可以用分布函数 $P(k)$ 来描述，表示网络中度数为 k 的节点数与总节点数的比值。如图 3、图 4 所示，节点的度，代表了某位编辑者与其存在编辑交互关系的编辑者数目，平均度 $\langle k\rangle$ 介于 3～6，其平均值为 4.2484，在 k 取值为 2 时，$P(k)$ 会出现峰值，峰值介于 0.3～0.5，k 取值 1～3 的比例占到 60%以上。

词条编辑交互网络度分布具有一个单峰，右半部分($k\geqslant 2$)具有重尾特征并可用幂律函数近似拟合，幂律指数介于 2～3，均值为 2.4019。词条编辑交互网络累积度分布符合幂律分布，幂律指数介于 1～2，均值为 1.4295。图中散点为实际值，直线为拟合值(其它图与此相同)。

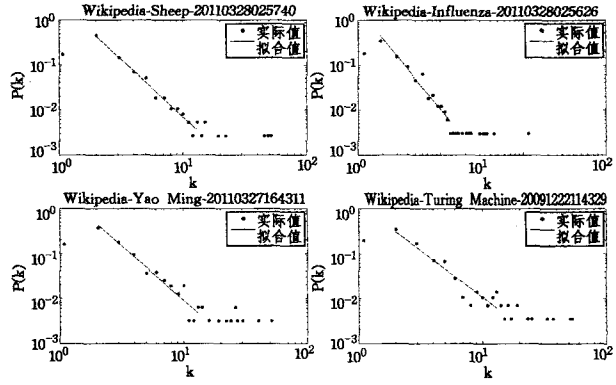

图 2 双对数坐标下词条编辑交互网络度分布

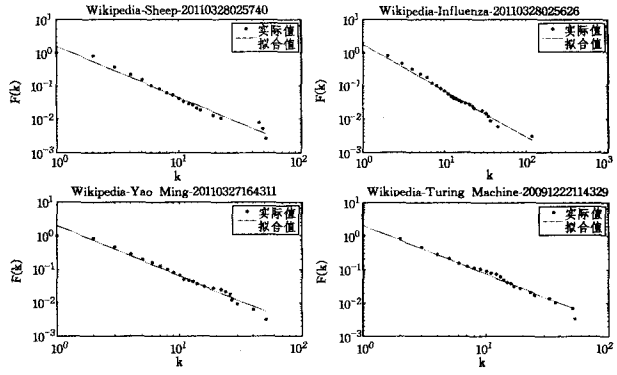

图 3 双对数坐标下词条编辑交互网络累积度分布

分布表明：编辑者在词条编辑交互过程中存在随机编辑和择优编辑行为；度数分布很不均匀，存在少数度数很大的节

点，其影响力较大，它与其他很多节点存在编辑交互关系，其与很多编辑者观点存在共鸣或冲突，这种节点的存在利于形成星型结构，利于个体知识向群体扩散；大多数节点度数较小，其影响力较弱，仅与与其他少数节点存在编辑交互关系，同时仅与少数编辑者观点存在共鸣或冲突。

3.2 边权重和强度分布

词条编辑交互网络的边权重 w_j 定义为两个节点 i 和 j 间编辑交互次数，可表征节点间编辑交互程度。如图 4 所示，边权重分布 $P(w)$具有重尾特征，符合幂律分布，幂律指数介于 2～3，均值为 2.5038，边权重平均值为 2.2469，k 取值为 1 的比例占到 70%以上。

分布表明：编辑交互程度具有较大异质性，编辑者在编辑交互过程中呈现出偏好依附性。边权重分布很不均匀，存在少数权重很大的边，其所连接的两个节点编辑交互程度很强，观点存在较大共鸣或分歧，形成频繁的合作或冲突，这种边的存在利于个人隐性知识向显性知识转移，深化对词条认识，提高其质量；大多数边权重较小，其所连接的两个节点编辑交互程度较弱，其观点不存在较大共鸣或分歧，合作或冲突不频繁。

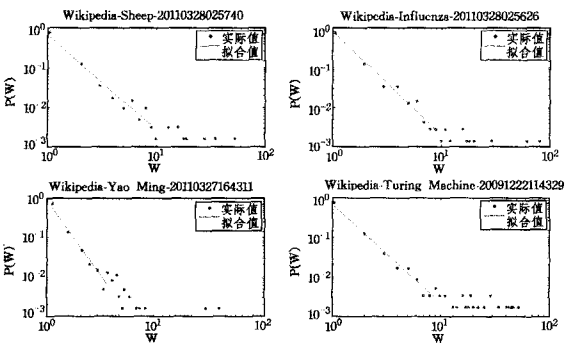

图 4 双对数坐标下词条编辑交互网络边权重分布

网络中节点 i 的强度 s_i 定义为每个节点与其邻居节点之间编辑交互次数的总和，即每个节点所连边权重之和。词条编辑交互网络强度分布 $P(s)$与度分布相似，也具有一个单峰，右半部分具有重尾特征并可用幂律函数近似拟合，幂律指数介于 2～3，均值为 2.2029，强度平均值为 9.6948，在 s 取值为 2 时，$P(s)$会出现峰值，峰值介于 0.2～0.4，s 取值 1～4 的比例占到 60%以上；如图 5 所示，词条编辑交互网络强度累积度分布符合幂律分布，幂律指数介于 0.8～1.3，均值为 1.0104。

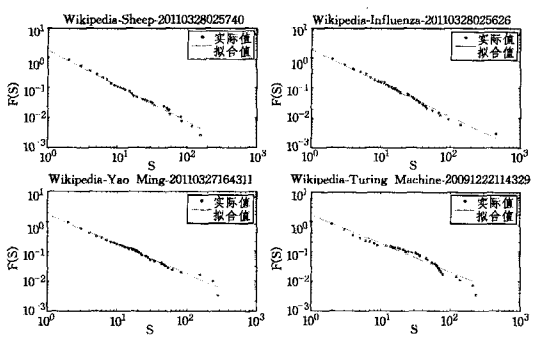

图 5 双对数坐标下词条编辑交互网络强度累积度分布

分布表明：强度分布很不均匀，编辑交互行为存在较大异质性。存在少数强度很大的节点，其影响力很大，表明其参与了很多词条编辑工作，活跃度很强，是词条编辑创造的积极贡献者，对引领词条编辑创作方向和提高词条质量起到重要作用，这种节点的存在利于词条内容更新创造和知识创新，可有效增强群体凝聚力，激发群体编辑协作，在词条形成发展的前期起到重要导向作用。大多数节点强度较小，其影响力较弱，其很少进行编辑工作，活跃度较弱，但正是这大多数所组成的群体起到了“群众的眼睛是雪亮”的作用，其能有效发现修改完善一些细小琐碎错误或不足，可使词条质量精益求精，在词条形成发展的后期起到重要的“查缺补漏”作用。可以推断在高质量词条编辑创作过程中，随着词条质量的逐步提升，会出现编辑角色、分工不同的编辑者，尽管其社会身份、知识背景等不同，但其大多都有为提升词条质量做贡献的意愿，本着“我奉献、我快乐”的精神，贡献个体知识，实现知识共享和自我价值。大多数编辑者对词条内容具有较高地认可度，其对词条内容仅进行少量编辑。

3.3 相关性

度—度相关系数定义为连在一起的节点对应的度值的 Pearson 相关系数 r[12]，$-1\leqslant r\leqslant 1$。如果一个网络的度同配指数 r 是负值，则表明中心节点倾向于跟小度节点相连，此时网络的度连接模式表现为异配特征；反之则表明度值相似的节点倾向于互相连接，$r>0$，网络是同配的。按照度—度相关性的定义可以计算网络度与度以及强度和强度之间的相关系数，9 个网络度—度和强度—强度相关系数分别介于 −0.1937～−0.0651 和 −0.1548～−0.0175，均值分别为 −0.1370和−0.0889，可知词条编辑交互网络具有异配性。

大量实证研究表明：生物和技术网络呈现异配连接模式，而大多数现实社会网络则具有同配特性[13]。说明词条编辑交互网络与实际社会网络是不同的。维基编辑机制打破了社会阶层间无形的壁垒，词条编辑交互过程中，每个人都有机会并可以很容易跟那些人气值很高的个人建立联系、进行编辑创作，这可能是词条编辑交互网络具有异配性的原因，这个特点在一定程度上体现了词条编辑创作“身份平等性”，利于编辑者畅所欲言，增强词条编辑协作的活力，也利于词条内容完善。

为了进一步研究度与强度的相关性，图 6 给出了随着节点度值的增大，拥有该度值 k 的节点的平均强度 $s(k)$的变化趋势图，其相关系数介于 0.8～1，具有较强的正相关性，分析度与强度近似满足幂律关系：$s(k)\sim k^{r}$，其幂指数介于 1～1.3，均值为 1.1497，这表明词条编辑交互网络节点度值和边权重具有特定的相关性。

在网络中，有一类现象，即节点 i 如果与节点 j 相连接，且节点 j 与节点 k 相连接，那么 i 与 k 之间也很有可能存在一条连边，聚集系数即用来描述这种概率的大小。聚集系数是社会网络中非常重要的性质之一。

词条编辑交互网络聚集系数平均值为 0.1743，如表 1 所列。图 7 给出平均聚类系数$\langle C(k)\rangle$和度值 k 之间的关系，即度值为 k 的节点的平均聚类系数$\langle C(k)\rangle$，二者的相关系数介于−0.7～−0.5 之间，直线的斜率介于−0.7～−0.4 之间，具有较强的负相关性。虽然$\langle C(k)\rangle$与不满足明显的幂律，但二者之间的关系仍是不平凡的。一般认为，$C(k)$与度值 k 的幂律关系表明网络中存在明显的层次结构，即小的节点群以层次化的方式组织成大的节点群，并维持无标度结构不变。

因而图 7 表明词条编辑交互网络拓扑表现出了某种程度的层次性和抱团性。

度数较大的节点一般会形成星型网络结构的中心，如图 7 所示，度数较大节点其聚集系数普遍较低，且度数 k 大于 30 的节点对应的$\langle C(k)\rangle$均低于 0.1，这种关系利于星型结构形成，可以推测网络中存在以少数高度值节点为中心的类星型结构。

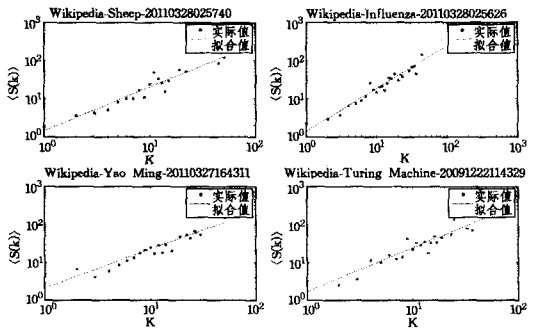

图 6　双对数坐标下节点度与强度相关函数图图

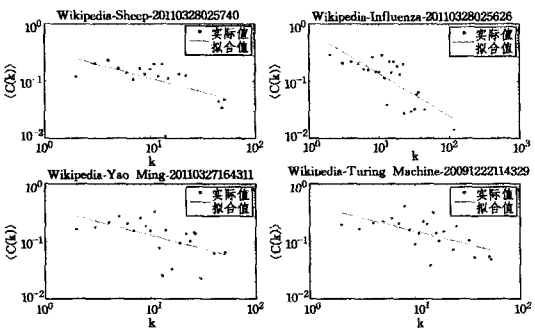

图 7　双对数坐标下平均聚类系数与节点度的关系

3.4　互惠性

文献[14]引入无偏差的互惠性系数描述有向网络中的互惠连接性。所谓互惠连接是指在网络中若点 i 指向点 j，则点 j 也指向点 i 。一个网络的互惠连接系数定义为：

$$\rho=\frac{\frac{L^{\leftrightarrow}}{L}-\alpha}{1-\alpha} \tag{1}$$

式中，$L^{\leftrightarrow}$是互惠连接的数量；L 是总连接的数量；α 是网络的连接密度。显然，ρ 越大，则网络的相互连接越多。$\rho>0$ 和$\rho<0$ 分别表明网络的互惠连接数量与随机网络比起来是多还是少。编辑交互网络互惠连接系数 ρ 介于 0.0024～0.0193 之间，平均值为 0.0100；虽然大于随机网络（$\rho=0$），但仍然较小。表明编辑交互网络中仅有少数连边是双向的，大多数连边是单向的，互惠性较弱，即词条编辑交互大多是单向的，双向交互概率较低。这种编辑交互机制更利于形成线型、环型结构。

3.5　最短路径长度

网络中两个节点 i 与 j 之间的路径长度为从 i 节点沿网络到达 j 节点所经过的距离（所经过边的数目）。最短路径长度是指两点之间最短的路径长度（经过的边数目最少），又称为两点之间的距离 d_{ij}。所有节点之间的距离最大值称为网络直径 D。词条编辑交互网络最短路径长度分布具有单峰分布，类似高斯分布，在 d 取值 3～5 时，$P(d)$会出现峰值，峰值保持在 0.3 左右，d 取值 3～5 的比例占到 60%以上，d 的均值为 3.8247，表明词条编辑者间的编辑交互距离均值约为 4 跳，网络中存在较多数量的线型结构。

线型结构的多少直接关系着群体关系的紧密程度，长串越多则意味着群体关系越紧密、群体结构越完整，词条编辑者之间的交流比较方便和频繁，也非常广泛。

结束语　本文以网络化数据挖掘思想方法为指导，对 9 个高质量维基百科词条进行文本分析，构建词条编辑交互网络，对高质量词条编辑特性进行了实证研究，得到了一些共性结论。研究表明，编辑群体经过较长时间编辑交互，逐渐形成了具有相对稳定结构的编辑交互网络，网络具有小世界和幂律特性，同时具有异配性和较弱互惠性，编辑群体具有较强异质性、抱团性，网络中存在数量较多的线型、环型结构和少量类星型结构，深化了对高质量词条编辑交互过程的认识，为进一步对群体智能涌现进行建模仿真提供了实证基础。

参 考 文 献

[1] Dennis W, Bernardo H. Cooperation and Quality in Wikipedia [C]//WikiSym 2007. 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). 2009:351-360

[4] Zhao Dong-jie, Jiang Jian, Zhang Hai-su, et al. Research on Internet Evolution Mode Based on User Behavior [C]//2010 Asia-Pacific Youth Conference on Communication Technology. Kunming, China, 2010, 8:835-839

[5] 赵东杰，张海粟，江健，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] Ponzetto S, Strube M. Deriving a large scaletaxonomy from Wikipedia[C]//Proceedings of the 22nd National Conference on Artificial Intelligence(AAAI-07). Vancouver, British Columbia. 2007:1440-1447

[7] 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(TextGraphs-4). 2009

[8] Weld D S, Wu F, Adar E, et al. Intelligence in Wikipedia[C]//AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence. 2008:1609-1614

[9] Adler B T, Alefaro L D. A Content-Driven Reputation System for the Wikipedia[C]//WWW '07: Proceedings of the 16th International Conference on World Wide Web. 2007:261-270

[10] Li D Y, Chen G S, Cao B H. Complex networks and networked data mining[C]//Advanced Data Mining and Applications. Wuhan, China, 2005:22-24

[11] Hirschberg S. Algorithms for the longest common subsequence problem[J]. Journal of the ACM, 1977, 24(4):664-675

[12] Fu F, Chen X. Liu L, et al. Social dilemmas in an online social network: the structure and evolution of cooperation[DB/OL]. http://arxiv.org/abs/physics/0701323, 2007-12-01

[13] Newman M E J, Park J. Why social networks are different from other types of networks[J]. PhysRevE, 2003, 68:036-122

[14] Garlaschelli D, Loffredo M I. Patterns of link reciprocity in directed networks[J]. Physical Review Letters, 2004, 93(6):268-701