ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 2,Mar. 2021,pp. 296－303
DOI:10. 16337/j. 1004-9037. 2021. 02. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于时间序列的人工智能在线翻译网络分析

冯吉芳[1]，田德红[2]，孙海信[3]

(1. 三江学院外国语学院,南京 210012;2. 南京宇天万维信息技术有限公司,南京 210019;3. 厦门大学信息学院,厦门 361005)

摘　要:从复杂网络角度出发,基于时间序列数据构建了人工智能在线翻译搜索指数的网络模型,并根据我国实际数据分析其网络结构特征。研究结果表明:在线翻译搜索指数虽然呈现出显著的波动特征,但大部分时间仍以小波动为主;在线翻译网络的最短路径长度分布近似呈偏态分布,网络中从一个符号到另一个符号的转换平均需要3个中间符号;波动性较小的符号具有较大的聚类系数;在线翻译整体呈下降趋势,经历了从早期不成熟到逐渐成熟的过程。

关键词:在线翻译;复杂网络;符号化;时间序列;人工智能

中图分类号:TPP393;TP301.5　　**文献标志码**:A

Network Analysis of Artificial Intelligence Online Translation Based on Time Series

FENG Jifang[1], TIAN Dehong[2], SUN Haixin[3]

(1. School of Foreign Languages, Sanjiang University, Nanjing 210012, China; 2. Nanjing Yutian Wanwei Information Technology Co Ltd, Nanjing 210019, China; 3. School of Informatics, Xiamen University, Xiamen 361005, China)

Abstract: From the perspective of complex networks, this paper constructs a network model of the search index of artificial intelligence online translation based on the time series data, and then analyzes its structure characteristics based on actual data from China. The results show that: although the search index of online translation shows significant fluctuation characteristics, it is still dominated by small fluctuations in most of the time; the distribution of the shortest path lengths of the online translation network is approximately skewed distribution, and the conversion from one symbol to another in the network requires three intermediate symbols on average; the symbols with less volatility have larger clustering coefficient; online translation shows a downward trend as a whole, and has experienced a process from early immature to gradually mature.

Key words: online translation; complex network; symbolic; time series; artificial intelligence

引　　言

复杂网络科学为研究现实世界中的复杂系统提供了一种有效的方法,例如互联网、万维网和金融系统等[1-6]。而语言作为一种高级符号系统,具有复杂性的本质。语言是一个复杂网络的观点已被广泛接受,来自复杂网络的方法也越来越多地用于语言系统的定量分析[7]。复杂网络可以为人类语言作为

基金项目:国家自然科学基金(61971362)资助项目;江苏省社会科学基金(19YYB003)资助项目。
收稿日期:2020-10-10;**修订日期**:2021-01-10

一个复杂系统进行适当的建模,并在系统层面上为其复杂性提供强有力的量化指标[8]。

Cancho和Solé[9]发现语言中连接单词的图与其他复杂网络具有相同的统计特征。在Cancho和Solé的研究之后,不同语言单元组成的语言网络及其在不同语言中的关系受到了学者们的关注[10-14]。复杂网络的构成要素主要是网络节点和节点间边,而如何确定语言网络的节点和边呢?学者们提出了不同的构建语言网络的方法[11, 15],主要包括:可以根据同义词表确定原始词与其同义词之间的网络连接;可以根据词汇表进行语义连接;根据词在句子中的共现情况,可以构建语言的共现网络;通过标注依存句法的语料库,可以得到语言网络连接。如Steyvers和Tenenbaum[16]研究了3类语义网络的大规模结构,发现它们都具有小世界网络结构特征。Gao等[17]基于语料库构建6个加权定向词共现网络,考察了不同语言之间的共性和差异。

现有关于语言系统的复杂网络研究,主要是通过语言文本数据挖掘语言系统中各种关联性,进而构建复杂网络模型进行分析。随着人工智能技术快速发展与应用,各个领域产生了大量复杂形式的数据。在语言领域亦是如此,特别是人工智能技术在语言翻译领域中应用,导致人工智能在线翻译技术得到广泛开发与应用,而此产生了各种各样数据,如采用人工智能在线翻译的搜索指数时间序列数据。复杂网络为从整体和局部挖掘语言规律提供了一条途径。近年来,学者们发现复杂网络方法也非常适合于复杂时间序列内部重要信息的挖掘[18-22]。通过符号化网络研究时间序列的优势主要有[23-24]:基于复杂网络拓扑结构特征可以有效区分噪声过程与复杂过程;过程复杂性的强弱可以通过复杂网络的统计特征进行测度,而其他一些典型的复杂性测度方法,易受噪声和数据样本量等因素影响;可以部分解决时间序列分析中其他传统方法未能解决的一些问题,如网络分析方法计算分形维数更加便捷。

然而,鲜有关于复杂网络在人工智能在线翻译中的研究。因此,本文基于人工智能在线翻译时间序列数据,构建人工智能在线翻译搜索指数复杂网络模型,对人工智能在线翻译趋势进行网络可视化分析,旨在揭示人工智能在线翻译趋势特征。本文的主要贡献在于以下3个方面:(1)鉴于时间序列符号化网络方法在时间序列分析中的优势,本文将其引入研究人工智能在线翻译指数变化内在规律;(2)相比已有语言系统复杂网络研究不同,本文基于时间序列数据展开分析,扩展了语言系统领域的相关复杂网络研究;(3)本文将时间序列符号化网络分析方法扩展到人工智能在线翻译方面研究,丰富了时间序列符号化网络应用研究。

1 研究方法

基于符号动力学和随机过程的思想,可以将人工智能在线翻译搜索指数时间序列数据转化为符号表示。时间序列的符号化是指将原始连续的时间序列划分为有限个离散区间,并将不同的符号分配给不同的区间[25-26]。人工智能在线翻译搜索指数原始连续时间序列数据分成不同间隔的部分对应于相应的符号。通过符号化,从而将人工智能在线翻译搜索指数时间序列转化为复杂网络。在这个网络中,符号是网络的节点,不同节点间的边为不同符号间的转移,边的方向为符号的转移方向,边的强度为不同符号间转移的次数。因此,构建的人工智能在线翻译网络是有向加权网络。

1.1 人工智能在线翻译网络模型

本文采用符号时间序列构建人工智能在线翻译搜索指数网络模型,网络模型的构建过程如下。

步骤1 设定人工智能在线翻译搜索指数时间序列为$G=(g_1, g_2, \cdots, g_n)$,计算其波动序列$G^1=(g_1^1, g_2^1, \cdots, g_{n-1}^1)$,其中,$g_i^1=g_{i+1}-g_i, i=1, 2, \cdots, n-1$。

步骤2 将人工智能在线翻译搜索指数波动序列转换成符号序列$S=(S_1, S_2, \cdots, S_{n-1})$,$S_i \in \{R, r, e, d, D\}$,其中

$$S_i=\begin{cases}R & S_i>P_{80}\\ r & P_{60}<S_i\leqslant P_{80}\\ e & P_{40}<S_i\leqslant P_{60}\\ d & P_{20}<S_i\leqslant P_{40}\\ D & S_i\leqslant P_{20}\end{cases} \tag{1}$$

式中：$S_i\in\{R,r,e,d,D\}$表示在线翻译搜索指数波动程度；P_α表示指数波动序列的α分位数，α取值为{20, 40, 60,80}。依据分位数将在线翻译搜索指数波动序列分为5类，分别表示大幅上升(R)、小幅上升(r)、平稳(e)、小幅下降(d)和大幅下降(D)。

步骤3 将在线翻译搜索指数波动序列转化为符号化序列后，以每3天作为一个符号。将不同的符号作为网络的节点，前一个符号向后一个符号的转变作为两个网络节点的有向边。两个符号之间的转换数是整个符号序列中两个节点间有向边的相应权重，进而得到在线翻译搜索指数网络。

1.2 网络结构指标

在本文中，主要使用以下结构指标来考察在线翻译搜索指数网络。

(1) 度和点强度

节点的度是度量网络节点重要性最基本的指标之一[27]。在无向网络中，节点i的度表示与节点i相连接的边的数量，则节点i的度k_i表示为

$$k_i=\sum_{j=1}^{N}a_{ij} \tag{2}$$

式中：当节点i和节点j之间存在边时，$a_{ij}=1$，否则，$a_{ij}=0$；N为网络节点数目。在有向网络中，度分为入度和出度。节点i的入度表示由网络中其他节点发出指向节点i的边的数量，节点i的出度表示由节点i发出指向网络中其他节点的边的数量，则节点i的入度k_i^{in}和出度k_i^{out}分别表示为

$$k_i^{\text{in}}=\sum_{j=1}^{N}a_{ji} \tag{3}$$

$$k_i^{\text{out}}=\sum_{j=1}^{N}a_{ij} \tag{4}$$

在加权网络中，节点的重要性程度不仅要考虑节点连接边的数量，同时也要考虑节点连接边的权重，将节点连接边的权重之和定义为点强度，则节点i的点强度表示为

$$N_{S_i}=\sum_{j=1}^{N}w_{ij} \tag{5}$$

式中w_{ij}表示节点i和节点j之间连接边的权重。边的权重为在线翻译指数网络中相邻两个节点间转换的次数。当节点i和节点j之间不存在连接边时$w_{ij}=0$。

(2) 平均路径长度和聚类系数

网络的平均路径长度和聚类系数是用来刻画网络紧密程度的指标[28-30]。网络中节点i和节点j之间的最短路径d_{ij}为连接这两个节点的边数目最少的路径。网络的平均路径长度L定义为任意两个节点间距离的平均值，即

$$L=\frac{1}{\frac{1}{2}N(N-1)}\sum_{i\geqslant j}d_{ij} \tag{6}$$

对于网络中的节点i，节点i的度k_i为与节点i相连的节点的数量，这k_i个与节点i相连的邻居节点之间存在的边数为E_i，则节点i的聚类系数为

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{7}$$

聚类系数表示与某一节点相邻的任意两个节点间存在连接的概率。

(3) 社团结构

社团是指复杂网络中连接较为紧密的节点集,社团内部节点间连接紧密,而社团间节点连接较为稀疏。本文采用Blondel等[31]提出的社团划分方法对在线翻译搜索指数网络进行社团划分。模块度被用作度量社团划分结果的指标,如下所示

$$Q = \frac{1}{2m}\sum_{i,j}\left[w_{ij} - \frac{A_iA_j}{2m}\right]\delta(D_i, D_j) \tag{8}$$

式中:w_{ij}表示节点i和节点j间连接边的权重,并满足$A_i = \sum_j w_{ij}$,$A_j = \sum_i w_{ij}$和$m = \frac{1}{2}\sum_{i,j} w_{ij}$;$D_i$和$D_j$分别表示节点$i$和节点$j$所属的社团。当社团$D_i$和社团$D_j$拥有共同节点时,$\delta(C_i, C_j) = 1$;反之,则$\delta(C_i, C_j) = 0$。

2 结果分析

在中国,英语是最为主要的第二语言,因此本文选择英语在线翻译搜索指数时间序列为研究对象。而研究对象的数据来源于百度,分析的样本时间为2011年1月1日至2020年10月26日,共有3 587个观测值。样本数据的最小值为1 885,最大值为64 237,平均值为25 077,标准方差为15 471。其中,在线翻译搜索指数由计算机终端和移动终端组成。

2.1 网络拓扑结构特征

表1给出了点强度排名前20的符号对应的点强度、入度、出度和聚类系数。可以看出符号*DRR*和

表1 网络节点的拓扑性质

Table 1 Topological properties of network nodes

排名	符号	点强度	入度	出度	聚类系数
1	*DRR*	111	41	32	0.062
2	*RRD*	106	27	38	0.069
3	*eee*	69	21	19	0.143
4	*eed*	58	19	17	0.146
5	*edd*	53	20	14	0.098
6	*ree*	52	20	16	0.144
7	*dDR*	50	17	9	0.065
8	*drr*	48	15	17	0.142
9	*edD*	46	17	13	0.078
10	*rre*	46	13	15	0.163
11	*ddD*	44	17	14	0.055
12	*RDr*	44	10	16	0.117
13	*RDd*	40	7	16	0.111
14	*DeR*	34	14	17	0.082
15	*DrR*	32	15	15	0.082
16	*ere*	30	11	10	0.155
17	*Ddr*	30	11	14	0.121
18	*rRr*	30	12	12	0.055
19	*DdR*	30	10	12	0.138
20	*ddd*	28	12	12	0.095

RRD 的点强度超过了 100，说明在线翻译搜索指数波动表现出显著的大涨大跌的特征，即在大幅下跌之后会出现连续的大幅上涨，在连续的上涨之后会出现大幅下跌。符号 *DRR* 的转入路径和转出路径分别为 41 条和 32 条，这意味着在线翻译搜索指数波动更倾向于转入大幅下跌后连续上涨。符号 *RRD* 的转入路径和转出路径分别为 27 条和 38 条，可见在线翻译搜索指数波动更倾向于从连续上涨变为大幅下跌，然后转换成其他状态。

符号 *DRR* 和 *RRD* 的聚类系数相对较小，说明与 *DRR* 和 *RRD* 相连的其他符号间存在连接的概率较低。此外，发现聚类系数较大的符号主要集中在波动较小的状态中，如 *eee*、*eed*、*ree*、*drr*、*rre* 和 *ere* 等。这反映了波动较小的符号之间的联系更为紧密。上述结果表明，我国在线翻译搜索指数虽然表现出显著的波动特征，但在大部分时间段内，仍以小幅波动为主。

图 1 为在线翻译网络最短路径长度的分布。通过对最短路径长度的研究可以揭示任意两个符号之间的转变距离问题。图 1 中不同节点间的最短和最长路径长度分别为 1 和 5，整个网络的最短路径长度分布近似为偏态分布。最短路径长度为 2 和 3 的数量占总数的比例超过 80%。数量最多的最短路径长度和整个网络的平均路径长度均为 3，说明由一个符号转换成另一个符号平均需要经过 3 个中间符号。

图 2 是对在线翻译网络的社团结构划分，符号大小表示相应的点强度，符号间的连线和箭头表示符号间具有转换以及转换方向。可以看出，将在线翻译网络划分为 5 个社团结构。点强度最大的 4 个符号 *DRR*、*RRD*、*eee* 和 *eed* 分别位于 3 个社团中。从图 2 可以看出，波动较小的符号间的联系更为紧密，这与波动较小的符号具有较大的聚类系数相对应。其他社团中则是大幅波动和小幅波动组成的符号占主体。

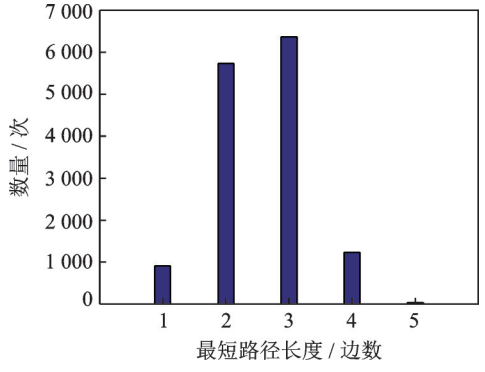

图 1 最短路径长度的分布

Fig.1 Distribution of the shortest path length

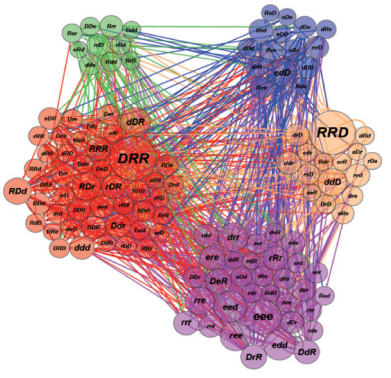

图 2 在线翻译网络社团结构

Fig.2 Community structure of the online translation network

2.2 在线翻译网络的引领模式

图 3 为出现频率最高的 *RRD*、*DRR*、*eed* 和 *eee* 四种符号在 2011—2020 年间的分布特征，其中图 3 (a～d)分别为符号 *RRD*、*DRR*、*eed* 和 *eee* 相应的结果。从图 3 可以看出，在线翻译搜索指数波动整体呈现下降趋势。其中，在 2011—2016 年出现了两次较大上升和下降过程，在 2017 年之后则逐渐下降。符号 *RRD* 和 *DDR* 主要分布在指数波动较大的 2011—2016 年间，而符号 *eed* 和 *eee* 则主要分布在 2017—2020 年间，在 2011—2016 年间也有少量分布，这在一定程度上说明了搜索指数在转折时期的剧烈波动特征，同时也反映了在线翻译从早期的不成熟到逐渐发展成熟的过程。

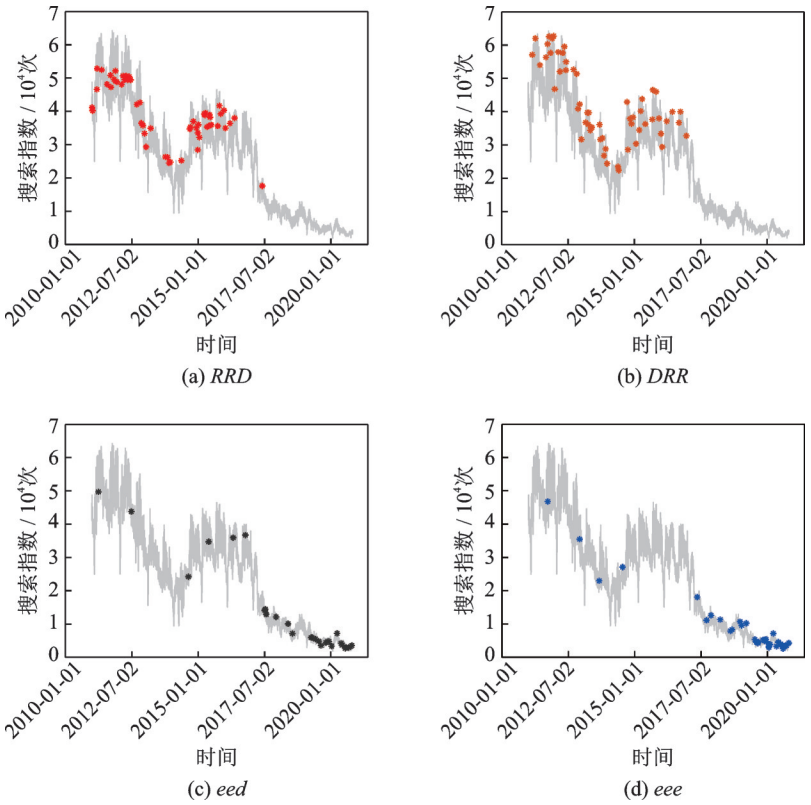

图3 4种引领模式的分布特征

Fig.3 Distribution characteristics of four leading modes

3 结束语

基于符号动力学和随机过程的思想,采用复杂网络理论方法,本文构建了时间序列数据驱动的人工智能在线翻译搜索指数的网络模型,并基于中国2011—2020年的数据分析了其网络结构特征。首先,根据节点强度最高的20个符号对应的节点强度、入度、出度和聚类系数的结果,发现尽管在线翻译搜索指数呈现出显著的波动特征,在大多数时间里仍然以小波动为主。其次,本文发现网络的最短路径长度分布近似为一个偏态分布,网络中从一个符号到另一个符号的转换平均需要3个中间符号。此外,本文还发现,波动性较小的符号具有较大的聚类系数。最后,根据引领模式的分布特征,发现在线翻译的发展过程是从早期的不成熟到逐渐成熟。本文基于时间序列符号化网络分析方法,揭示了人工智能在线翻译搜索指数变化规律。一方面,有助于了解中国人工智能在线翻译使用现状,为人工智能在线翻译系统产品设计开发等提供市场数据,为人工智能在线翻译相关企业提供决策参考。另一方面,为语言领域研究者开展相关问题研究提供现实依据。

参考文献:

[1] BARABÁSI A L, ALBERT R, JEONG H. Scale-free characteristics of random networks: The topology of the world-wide web[J]. Physica A: Statistical Mechanics and Its Applications, 2000, 281(1/2/3/4): 69-77.

[2] 王辉,赵文会,施佺. 复杂网络中节点重要性Damage度量分析[J]. 南京理工大学学报,2012,36(6): 926-931.
WANG Hui, ZHAO Wenhui, SHI Quan. Analysis on Damage measure of vertex importance in complex networks[J]. Journal of Nanjing University of Science and Technology, 2012, 36(6): 926-931.

[3] 张琨,沈海波,张宏,等. 基于灰色关联分析的复杂网络节点重要性综合评价方法[J]. 南京理工大学学报,2012,36(4): 579-586.
ZHANG Kun, SHEN Haibo, ZHANG Hong, et al. Synthesis evaluation method for node importance in complex networks based on grey relational analysis[J]. Journal of Nanjing University of Science and Technology, 2012, 36(4): 579-586.

[4] ANAND K, GAI P, KAPADIA S, et al. A network model of financial system resilience[J]. Journal of Economic Behavior & Organization, 2013, 85: 219-235.

[5] 王鑫,陈喜,钱付兰,等. 结合共同邻居贡献度的节点相似性链路预测算法[J]. 数据采集与处理,2018,33(5): 900-910.
WANG Xin, CHEN Xi, QIAN Fulan, et al. Node-similarity link prediction algorithm combined common neighbor contribution [J]. Journal of Data Acquisition and Processing, 2018, 33(5): 900-910.

[6] 刘晋霞,孙丽萍,杜静,等. 基于物理场论的探测复杂网络社团结构的分布估计算法[J]. 数据采集与处理,2017,32(1): 126-133.
LIU Jinxia, SUN Liping, DU Jing, et al. Estimation of distribution algorithm for detecting community structure of complex networks based on field theory model[J]. Journal of Data Acquisition and Processing, 2017, 32(1): 126-133.

[7] AMANCIO D R, ALUISIO S M, OLIVEIRA O N, et al. Complex networks analysis of language complexity[J]. EPL (Europhysics Letters), 2012, 100(5): 58002.

[8] CONG J, LIU H. Approaching human language with complex networks[J]. Physics of Life Reviews, 2014, 11(4): 598-618.

[9] CANCHO R F I, SOLÉ R V. The small world of human language[J]. Proceedings of the Royal Society of London. Series B: Biological Sciences, 2001, 268(1482): 2261-2265.

[10] ČECH R, MACUTEK J. Word form and lemma syntactic dependency networks in Czech: A comparative study[J]. Glottometrics, 2009, 19: 85-98.

[11] SOLÉ R V, COROMINAS-MURTRA B, VALVERDE S, et al. Language networks: Their structure, function, and evolution[J]. Complexity, 2010, 15(6): 20-26.

[12] 刘海涛,丛进. 基于平行词同现网络的语言聚类[J]. 科学通报,2013,58: 432-437.
LIU Haitao, CONG Jin. Language clustering with word co-occurrence networks based on parallel texts[J]. Chinese Science Bulletin, 2013, 58: 432-437.

[13] 赵怿怡,刘海涛. 语言网络研究的数学模型——从复杂网络、社会网络到语言网络[J]. 中文信息学报,2015,29(6): 46-53.
ZHAO Yiyi, LIU Haitao. Mathematical modeling in language networks research—from complex networks to social networks and language networks[J]. Journal of Chinese Information Processing, 2015, 29(6): 46-53.

[14] SIEW C S Q, VITEVITCH M S . The phonographic language network: Using network science to investigate the phonological and orthographic similarity structure of language[J]. Journal of Experimental Psychology: General, 2019, 148(3): 475-500.

[15] LIU H. Linguistic complex networks: A new approach to language exploration[J]. Grundlagenstud Kybern Geisteswiss, 2011, 52(4): 151-170.

[16] STEYVERS M, TENENBAUM J B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth[J]. Cognitive Science, 2005, 29(1): 41-78.

[17] GAO Y, LIANG W, SHI Y, et al. Comparison of directed and weighted co-occurrence networks of six languages[J]. Physica A Statal Mechanics & Its Applications, 2014, 393: 579-589.

[18] LACASA L, TORAL R. Description of stochastic and chaotic series using visibility graphs[J]. Physical Review E, 2010, 82 (3): 036120.

[19] 曾明,王二红,赵明愿,等. 基于时间序列符号化模式表征的有向加权复杂网络[J]. 物理学报,2017,66(21): 292-302.
ZENG Ming, WANG Erhong, ZHAO Mingyuan, et al. Directed weighted complex networks based on time series symbolic pattern representation[J]. Acta Physica Sinica, 2017, 66(21): 292-302.

[20] TIAN Z. Chaotic characteristic analysis of network traffic time series at different time scales[J]. Chaos, Solitons & Fractals, 2020, 130: 109412.

[21] 汪丽娜,成媛媛,臧臣瑞. 基于seasonal-trend-loess方法的符号化时间序列网络[J]. 物理学报,2019,68(23): 320-328.
WANG Lina, CHENG Yuanyuan, ZANG Chenrui. A symbolized time series network based on seasonal-trend-loess method [J]. Acta Physica Sinica, 2019, 68(23): 320-328.

[22] 孙延风,王朝勇.一种基于文本互信息的金融复杂网络模型[J].物理学报,2018,67(14):270-280.
SUN Yanfeng, WANG Chaoyong. Financial complex network model based on textual mutual information[J]. Acta Physica Sinica, 2018, 67(14): 270-280.

[23] 袁铭.基于符号化的时间序列复杂网络构造及其拓扑结构研究[J].计算机应用研究,2015,32(4):1044-1047.
YUAN Ming. Research on building complex network based on symbolization of time series and its topological properties[J]. Application Research of Computers, 2015, 32(4): 1044-1047.

[24] 邹勇,DONNER R V, MARWAN N, 等.非线性时间序列的复杂网络分析[J].中国科学:物理学 力学 天文学,2020,50(1):133-147.
ZOU Yong, DONNER R V, MARWAN N, et al. Nonlinear time series analysis by means of complex networks[J]. Scientia Sinica Physica, Mechanica & Astronomica, 2020, 50(1): 133-147.

[25] TANG X Z, TRACY E R, BROWN R, et al. Symbol statistics and spatio-temporal systems[J]. Physica D, 1997, 102(3): 253-261.

[26] NICOLIS G, CANTU A G A, NICOLIS C. Dynamical aspects of interaction networks[J]. International Journal of Bifurcation and Chaos, 2005, 15: 3467-3480.

[27] BRIDA J G, PUNZO L F. Symbolic time series analysis and dynamic regimes[J]. Structural Change & Economic Dynamics, 2003, 14(2): 159-183.

[28] ZHANG Q, XUE Y, WANG D, et al. Asymptotic formula on average path length in a hierarchical scale-free network with fractal structure[J]. Chaos, Solitons & Fractals, 2019, 122: 196-201.

[29] ALBERT R, BARABÁSI A L. Statistical mechanics of complex networks[J]. Review of Modern Physics, 2002, 74: 47-97.

[30] ZOU Y, DONNER R V, MARWAN N, et al. Complex network approaches to nonlinear time series analysis[J]. Physics Reports, 2019, 787: 1-97.

[31] BLONDEL V D, GUILLAUME J, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 10: 155-168.

作者简介:

冯吉芳(1981-),**通信作者**,女,博士,讲师,研究方向:语言复杂性、语言经济学等,E-mail: oasishon001@126.com。

田德红(1979-),男,博士,高级工程师,研究方向:复杂网络分析。

孙海信(1977-),男,博士,教授,研究方向:阵列信号处理。

(编辑:夏道家)