一种基于复杂网络模型的作者身份识别方法*

李晓军 刘怀亮 杜坤
西安电子科技大学经济管理学院 西安 710126

摘要: [目的/意义] 作者身份识别是语言文体学的重要研究方向,利用文本特征的身份识别也是文本挖掘的重要任务。在开放和虚拟网络环境下海量信息的作者身份或发布者的识别难题和传统作者身份识别方法在处理效率和成本等方面存在的问题有待解决。[方法/过程] 将复杂网络理论引入该研究领域,在利用传统文体学特征识别作者身份方法的基础上结合文本词共现网络模型及其指标特征改进相关算法,使用文本文体学特征和文本网络模型度量指标构建作者风格特征集合,通过计算文本间风格相似度进行作者识别。[结果/结论] 基于复杂网络模型的作者身份识别方法可以有效的利用作者风格特征,提高识别的精度,与其他算法的对比试验表明其识别结果的准确性更高。

关键词: 作者识别 文本分类 复杂网络 特征提取 词共现 文体学
分类号: TP391
DOI: 10.13266/j.issn.0252-3116.2015.18.016

作者身份识别[1]作为一项应用广泛的研究,是文本挖掘的重要探索方向,文献或文本信息资源作为当前数量最大、利用率最高的信息资源,是图书情报相关研究领域的重要客体。在海量文本数据处理背景下,基于传统语言学研究的文体分析越来越多地借鉴自然语言处理[2]的一些技术和方法[3],利用不同作者文章在语法、词汇、修辞以及句型结构等方面的差异,结合计算机处理技术识别作者身份已成为研究的热点,具体成果在名著考证、用户识别和信息安全领域受到广泛关注和引用,在开放和虚拟网络环境下识别信息的作者或发布者将成为图书情报领域相关研究的重要内容。

作者身份识别是以作者为标志的文本分类[4],即作者身份识别为文本分类的一种特例。作者身份识别的关键问题是从文本中提取出代表作者风格的识别特征,在评估不同文本之间风格特征相似度基础上,综合各项特征参数比较结果,识别作者身份,其中以基于标点符号和最常见功能词频数的分析方法受到较为普遍的认同。文本分类是文本挖掘[5]的一个子集,故作者身份识别也是文本挖掘的一个子集。近年来,国内外有关作者身份识别的研究成果显著,相关研究主要集中于利用文本表示模型[6]提取用于识别作者身份的特征集合等。相关研究中文献[7]使用自组织模型定义了从属某一作者的文体学[8]特征空间从而进行作者身份的识别;文献[9]则证明了在大样本条件下运用概率主题识别作者身份会有较高的正确率,将概率主题模型引入作者身份识别领域并取得较好的效果;国内对该主题的研究有基于传统文体学统计特征的网络生成内容用户身份识别[10],该方法综合文体学研究多种风格特征进行作者识别,但未对文本内容进行分析处理;基于语义分析作者身份识别[11],将语义技术引入身份识别,仅进行简单的主题聚类;基于语句节奏特征的作者身份识别[12]、基于 VSM 模型的文学作品[13]或典籍作者身份识别[14]等侧重于文本的局部特征,未能从文本内容特征和段落篇幅特征综合考虑,在一定程度上影响了身份识别的准确度和精度。

复杂网络作为复杂性科学研究的有力工具,受到了来自不同研究领域学者们的普遍重视。人类语言也是一个动态的复杂系统[15],其在词语、语法、语义各个层面上都显示出复杂网络结构,可以采用复杂网络技

* 本文系国家自然科学基金“基于复杂网络的中文文本语义相似度研究”(项目编号:71373200)研究成果之一。
作者简介: 李晓军(ORCID: 0000-0003-0701-9348),硕士研究生,E-mail: xidianlixiaojun@126.com; 刘怀亮,教授,博士生导师; 杜坤,硕士研究生。
收稿日期: 2015-08-16 修回日期: 2015-09-02 本文起止页码: 102-107 本文责任编辑: 王善军

术对语言进行分析和研究,小世界、无标度等网络特性被验证存在于中文语言网络中[16],因此复杂网络常被运用到文本挖掘领域[17],为语言学、自然语言处理等学科提供了新的研究视角,用计算复杂网络参数的方法来分析语言网络的综合特性,对语言类相关学科的研究具有重要的意义。

传统的文本作者识别在海量文本数据处理过程中存在效率与成本的问题,而复杂网络已经被运用于海量文本挖掘的具体领域,本文将复杂网络引入作者身份识别研究领域,将作者作品文本表示为复杂网络模型,并在此基础上提出一种基于复杂网络的作者身份识别方法。首先提取著作者作品中的功能词等构建文体学风格特征集合,然后通过计算文本网络模型中相关指标参数,同时在利用功能词特征集合识别出候选作者的基础上,计算文本间风格相似度进行作者识别,最后通过实验分析与评估具体算法。

1 相关理论

1.1 复杂网络与文本

语言网络作为一种典型的复杂网络,语言组成要素如字、词或短语及语义、句法关系等可以被表示为网络的节点或边等,例如功能语言学对作者文体风格的研究[18]借鉴一些复杂网络的方法和技术。对单篇文本来讲,可以将其表示为网络形态,文献[19]提出的一种英文文本网络表示模型,通过连接邻接词构建网络,去除停用词和合并同义词、近义词形成节点,使用词在句子中具体的位置关系连边,通过这种方法组建的网络模型被称为文本邻接矩阵模型或文体图模型。如何将中文文本表示为复杂网络的模型,国内学者已经有一定的探索,基于图论的文本表示模型已经被应用于文本分类[20]及聚类[21]研究中。

不同的文本表示模型因其组建网络时依据的原则不同,模型对文本信息的表征内容与表征程度也有较大差异,本文通过借鉴复杂网络的方法和技术来改进文体学研究中识别作者风格传统算法,文体学研究中一般认为作者风格是独立于作品内容的,因此本文选择使用词共现关系来组建文本复杂网络模型,可以更加恰当地表示作者的用词习惯和语言特色。

文本词同现网络模型的特点是,两个出现在文档同一窗口单元(例如同一句话)的词在网络中具有相同的位置及作用,这种同现信息对自然语言处理有重要的作用。在这种模型下,节点数量与文本长度正相关,且边的数量将远小于文本长度。图1中,列举了无向词同现网络模型的图表形式和矩阵形式(以句子"我爱北京天安门,天安门上太阳升"为例)。由于每篇文本的结构特征和写作风格可以通过网络模型的统计参数进行分析,本文中选取了平均最短路径长度、平均邻近节点度、聚类系数和α指数来进行文本作品的作者风格特征提取。

节点	我	爱	北京	天安门	上	太阳	升
标签	1	2	3	4	5	6	7

$$M=\begin{bmatrix} Node & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 4 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 5 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 6 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 7 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

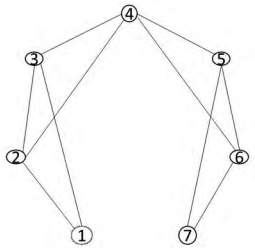

图1 词共现网络邻接矩阵模型示意

1.2 文本网络的幂律指标及其计算方法

Heaps定律[22]是一个语言学中词汇增长的经验法则,它描述的是一个由各种词汇组成的、不断增长的文本或文本集合中特殊词汇所占的比例,定律表明随着文本长度的增加,文本内容不断地生成,文本词汇量的增加率则随之递减。Heaps定律可以用公式表示为$N(s)=cs^{\theta}$,其中$N(s)$表示文本长度为s时文本的词汇量。c和θ都是经验系数。在英语语言环境下,c通常在10到100的范围内,θ一般在0.4到0.6的范围内[23]。Heaps定律已广泛地应用到了语言学、经济学、社会学、计算机科学、生物学、地理学乃至整个生产应用中。

在文本网络模型条件下,Heaps定律通常表现为节点和边之间的指数关系。本文中,通过$N_{link}\approx N_{node}{}^{\alpha}$公式来简要说明中文文本词共现复杂网络中边跟节点之间的关系。N_{link}、N_{node}分别为文本网络的边的数目和节点的数目,这样就可以得到一个度量文本词汇量与文本长度直径的指标α,这个指标在一定程度上能反映写作该文本的作者的某些特性或风格,该指标在结合文本网络其他指标的情况下以用来识别作者身份。

2 基于复杂网络文本模型的特征提取与作者身份识别方法

作者身份识别的关键问题是通过分析特定作者独

特的语言形式和作品风格形成作者风格特征集合，然后通过匹配从已知作者作品中识别出的特征跟未署名作者作品的识别特征，即计算风格特征相似度来识别作者身份。

2.1 复杂网络指标计算

复杂网络是复杂系统研究的有利工具，网络中的节点是复杂系统中的个体，节点之间的边则是系统中个体之间按照某种规则而自然形成或人为构造的一种关系。路径长度指从某一节点到另外一个节点时中间途经边的数目，平均最短路径长度是网络模型一个重要的特征，其定义公式如式(1)所示:

$$\bar{l}=\frac{1}{n(n-1)}\sum_{i,j=1}^{n}l_{ij} \quad (1)$$

其中 l_{ij} 被定义为节点 i 和 j 之间的最短路径，n 为所有节点的数目。节点 i 的度则被定义为其与该节点相连接的所有边的数目，即 $k_i=\sum_{j=1}^{n}a_{ij}$。度分布是描述网络全局性特征的重要标志之一，p(k) 是指某一节点有度 K 的概率，那么整个网络的平均度的计算方法如公式(2)所示，其中 k_{min} 与 k_{max} 是网络节点的中最大、最小度。

$$\bar{k}=\frac{1}{n}\sum_{i=1}^{n}k_i=\sum_{k_{min}}^{k_{max}}kp(k) \quad (2)$$

网络中邻接节点的平均度的计算公式为 $k_{nn,i}=\sum_{i\in V(i)}k_j/k_i$，其中 V(i) 集合中包涵节点的最邻近节点，网络中全部节点的临近节点度的计算公式如式(3)所示:

$$\bar{k}_{nn}=\frac{1}{n}\sum_{i=1}^{n}k_{nn,i} \quad (3)$$

网络中聚类系数是指网络中节点聚集的倾向性概率，对于一个给定节点其聚类系数定义公式为 $c_i=[\sum_{j,l}a_{ij}a_{jl}a_{li}]/[k_i(k_i-1)]$，其中 $1\geq ci\geq 0$，根据此公式，网络中全部节点的平均聚类系数的计算公式如式(4)所示:

$$\bar{c}=\frac{1}{n}\sum_{i=1}^{n}c_i \quad (4)$$

2.2 特征提取

本文在使用复杂网络模型分析特定作者作品的识别特征基础上，通过利用作品文本网络的统计指标特性，即平均最短路径长度、平均邻近节点度、聚类系数和 α 指数，结合文体学研究方法提取文献或文学作品中的标点符号频数、功能词频数等[24]特征构建文体学作者风格特征集合识别作者身份，从而确定未署名作品的作者。

不同文体的差异使其在功能词的使用上有显著的不同[25]，文体类型在功能词的确定中发挥着重要的作用，对照学者王自强归纳的现代汉语功能词表，本文选择新闻报道作为实验数据类型，统计每个作者使用频率较高的功能词，组成作者常用表功能词表，然后对比未知署名作品中的功能词，选择出现频率较高的作者作为该篇文献的候选作者，结合下文提出的网络模型指标参数匹配法在候选作者集合中识别文献作者。

对某一特定作者，统计其全部作品中功能词和与之对应的的词频，根据每个功能词在风格特征集合中的频率 $f_1,f_2,\cdots,f_m$，重新计算该词在特征集合中的 Zscore 标准得分，计算公式如式(5)所示:

$$z_i=\frac{f_i-mean(f_i)}{sd(f_i)}，其中\ m\geq i\geq 1 \quad (5)$$

然后选取风格特征集合中的功能词标准得分 z_i 超过规定阈值的特征词作为该作者的风格特征集合，设其由 m 个功能词构成，用集合 $F=\{F_1,F_2,\cdots,F_m\}$ 表示。这样每一作者的风格特征可以由一个 m 维向量来表示，然后求得未署名作者作品的功能词特征向量，通过向量间的相似度来计算未署名作者的可能身份，使用最大相异系数算法[26]选择相似度较高的作者为候选作者，结合下文提出的方法进一步识别作者身份。

2.3 作者识别方法

基于复杂网络的作者身份识别方法是建立在评估未署名作品与已知作者作品集之间文体风格特征相似度基础上的，通过统计每个作者所有选定作品的文本复杂网络的模型参数，形成包含作者独特语言形式和作品风格的作者列表集合，然后计算未知署名作品网络模型的参数，最后，综合所有作者列表集合中的特征参数比较结果，识别作者身份。

设由 i 个作者组成的列表集合为 $N=\{N_1,N_2,N_3,\cdots,N_i\}$，对于其中某个任意作者 N_a，设其有 j 个指标特征，即作者 N_a 的选定作品中用以表征该作者语言形式和作品风格的参数数目。对特定的某一指标，计算未知署名作品在表征作者写作风格的列表集合中的概率，其计算公式如式(6)所示:

$$p_a(b)=1-\frac{|M_x(b)-M_a(b)|}{\sum_{i=1}^{i}|M_x(b)-M_a(b)|} \quad (6)$$

其中 a 代表某一特定作者,b 代表特定的模型指标,且 $1 \leq b \leq j$,$|M_x(b) - M_a(b)|$ 为未知署名作品与特定作者作品在模型指标参数 b 上计算产生的距离 b,$M_x(b)$ 是未知署名作品网络模型的 b 类指标参数,$M_a(b)$ 为特定作者选定作品网络模型指标参数 b 的平均值。公式表明,就某一类指标来讲,距离越小,差异越小,概率值越大,则说明该作品属于该作者的可能性越大。

事实上,由于不同作品具体内容的差异,通过模型计算出的指标也不完全相同,不同的指标在计算方法上也有一定的差异,指标数值的不同说明作者在作品中表现出的语言形式、作品风格在词汇、句法、语篇和语义方面的差异。需要联合多个指标综合计算,以便准确定义于识别作者身份,计算公式如式(7)所示:

$$p_a = N\prod_{j=1}^{j} p_a(b) \qquad (7)$$

其中 j 为作者作品文本网络模型的指标参数数目,N 为基于 $\sum_{i=1}^{i} p_i = 1$ 的可调节的参数,对判断与识别作者身份无影响。

综上所述,这种混合功能词词频和复杂网络文本模型指标特性的作者身份识别方法,充分考虑了文献或文学作品著作者的作品风格的语言使用特色,能更好地表征作者的识别特征,下文将通过实验对该种方法进行检验。

3 实验与结果分析

本文选取新闻机构的知名记者、编辑的新闻报道文章共 25 542 篇作为测试文本数据,数据集包含不同的主题或板块,下文将使用 java 语言编程进行实验比较,分析中文环境下基于复杂网络的作者身份识别算法的有效性问题,图 2 为实验的流程图。

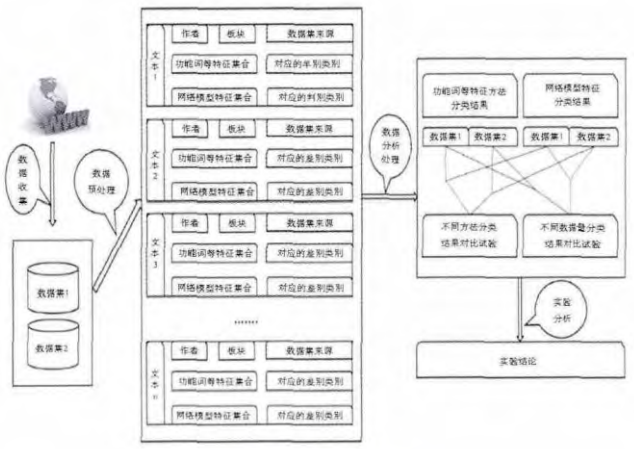

图 2 实验流程

表 1 列出了数据集的一些基本概况,其中左侧为从某一晚报选取的作者及其文本数据概况,右侧为从某一都市报选取的作者及其文本数据概况。为方便起见,称前者为扩展数据集(作者 1-20)、后者为普通数据集(作者 21-40),划分原因在于两个数据集的某一作者文本数据容量存在差异。在数据的选取过程中,忽略了由多位作者合著的文章文本或单篇字数过少的文章文本。经过筛选后,实验共选取 25 542 篇文本,平均每位作者 638 篇,且不同数据集在容量上有较大差距。

表 1 测试数据概况

编号	作者	板块	文章数	编号	作者	板块	文章数
1	刘珊	财经	983	21	蔡建国	娱乐	40
2	戴曼曼	财经	942	22	陈奋翔	娱乐	183
3	吕楠芳	民生	1 173	23	晁阳	时事	162
4	赵燕华	财经	1 217	24	丁聘	民生	24
5	何伟杰	民生	979	25	葛兰	科技	66
6	赫子仪	文化	645	26	胡琳	教育	85
7	董柳	文化	1 167	27	黄美茹	财经	44
8	杨辉	民生	847	28	李海涛	财经	297
9	安颖	教育	2 450	29	李红	教育	177
10	李雯洁	时事	781	30	刘成峰	时事	106
11	张爱丽	财经	773	31	马秀红	民生	118
12	凌越	时事	893	32	阮班慧	民生	58
13	赵映光	民生	868	33	唐明军	财经	225
14	余宝珠	科技	1 346	34	王宝存	民生	262
15	黄蔚山	时事	379	35	王梅	文化	533
16	全良波	娱乐	846	36	夏明勤	文化	473
17	陈泳强	体育	1 219	37	宜茜	娱乐	597
18	王俊	家居	1 473	38	张艳芳	科技	964
19	梁怿韬	时事	1 209	39	赵明	时事	261
20	郝婧羽	娱乐	568	40	周书养	文化	109

对实验数据预处理之后,将实验步骤按实验目的设置为 3 步:第一步对每位作者随机抽取一定数量的作品作为训练集语料库,剩余作品为测试语料库;第二步测试结合网络模型的指标参数的综合方法对仅使用功能词的传统作者身份识别方法的改进效果;第三步测试不同的作者文本数据容量对识别结果的影响程度。

判断一个作品的归属作者的方法是,比较这个作品与每一位作者所有训练作品的平均相似度,以可能性最高者为最终结果。在测试方法的选取上,采用五折交叉验证法,将实验语料分成 5 份,轮流将其中 4 份作为训练数据,1 份作为测试数据,分别进行实验得到具体数据,对每位作者求其多次实验的平均查全率和

平均查准率，并在此基础上求出调和平均值，即 F 值，最后求取 40 位作者的宏平均 F_1 值，作为最终的评估标准。

实验中识别不同作者的主要方法就是通过区分不同作者所写文章在文本特征方面的相似性，利用本文提出的模型指标参数，对相似性的精度进行进一步的提升。需要指出的是，文本模型指标特征选择问题中，需要对一些指标特征进行精确的计算。本实验由以下两个实验共同构成。

(1) 基于复杂网络指标特征的作者身份识别方法的有效性检验实验。本实验在使用功能词特征矩阵识别作品作者身份的基础上，对比分析使用网络模型的特征参数综合判断前后的实验结果，检验本文提出的方法是否可行有效。实验过程中对数据集按相关板块或类型进行分类处理，实验结果如表 2 所示:

表 2 实验结果的 F_1(%) 值比较

类型	传统功能词向量法		网络模型参数法	
	普通数据集	扩展数据集	普通数据集	扩展数据集
财经	46.98	57.21	76.37	84.50
民生	49.02	57.98	77.81	85.62
文化	51.44	53.94	80.31	83.03
教育	47.98	59.28	78.93	85.64
时事	52.56	56.40	82.46	85.69
娱乐	50.77	59.89	79.98	85.53
科技	51.29	55.99	80.56	82.25
平均	50.00	57.24	79.49	84.61

由表 2 平均 F_1 值可以看出，基于复杂网络文本模型的作者身份识别方法比传统的功能词向量法在识别与分类的精度与准确度方面有很大的提升，其主要原因在于利用文本网络模型得到的指标参数在一定程度上代表了作者的写作风格和语言特色。

(2) 作者身份识别精度与准确度影响因素分析实验。本实验通过对比不同作者文本数量在明显差异情况下文本分类的精度与准确度，分析比较实验结果与选取作者的文本数目的关系，相关实验结果如图 3 所示:

结合表 2 和图 3，从实验结果可以发现，作者的文本数量的不同对分类效果有影响，较大作者文本数据容量可以提升识别与分类算法的稳定性，在一定程度上影响识别与分类的正确率，但程度不大，即选取某一作者更多的文本作品参与模型的构建与分析处理能提升该作者类别文本判定的准确程度。综上所述，在作

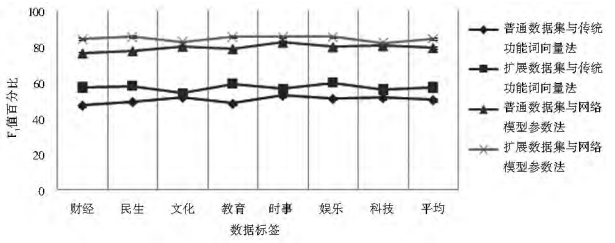

图 3 F_1 值对比

者文本数据容量较大的条件下，本文提出的基于复杂网络模型参数的作者身份识别方法较之传统的基于文体学词频统计的方法，在识别精度和准确度方面有较大的提升，是一种综合性能较好的方法。

4 结语

本文所提出的基于复杂网络的作者身份识别方法，运用复杂网络理论提取文本特征用于作者身份识别。通过构建复杂网络模型对文本进行特征提取，在文体学研究的基础上，利用文体学统计特征与网络模型指标构建代表作者写作风格的特征集合，利用文本间的风格相似度识别作者身份。实验结果表明，在综合运用网络模型参数条件下，本文提出的方法能够取得较好的实验效果。结合文本其他特征的网络模型方法可以进一步提升分析的准确度和精度，同时由于不同文体的文本具有不同网格，其文本特征的提取存在差异，故针对不同文体类型还需要对相关算法进行深入研究。

参考文献:

[1] Bozkurt I N, Baghoglu O, Uyar E. Authorship attribution [C] // 22nd International Symposium on Computer & Information Sciences. Piscataway: IEEE, 2007: 1 - 5.

[2] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning [C] // Proceedings of the 25th International Conference on Machine Learning. New York: ACM, 2008: 160 - 167.

[3] Stamatatos E, Fakotakis N, Kokkinakis G. Computer-based authorship attribution without lexical measures [J]. Computers & the Humanities, 2001, 35(2): 193 - 214.

[4] Sebastiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(2): 1 - 47.

[5] Klein A, Riazanov A, Hindle M M, et al. Benchmarking infrastructure for mutation text mining [J]. Journal of Biomedical Semantics, 2014, 5(1): 11.

[6] Liu Wenyin, Quan Xiaojun, Feng Min, et al. A short text modeling method combining semantic and statistical information [J]. Infor-

mation Sciences,2010,180:4031-4041.

[7] Neme A,Pulido J R G,Abril Muñoz,et al. Stylistics analysis and authorship attribution algorithms based on self-organizing maps[J]. Neurocomputing,2015,147:147-159.

[8] Parasher S V. Indian English: Certain grammatical,lexical and stylistic features[J]. English World-Wide,1983,4(1):27-42.

[9] Savoy J. Authorship attribution based on a probabilistic topic model[J]. Information Processing & Management,2013,49(1):341-354.

[10] 吕英杰,范静,刘景方. 基于文体学的中文UGC作者身份识别研究[J]. 现代图书情报技术,2013(9):45-49.

[11] 武晓春,黄萱菁,吴立德. 基于语义分析的作者身份识别方法研究[J]. 中文信息学报,2006,20(6):61-68.

[12] 王少康,董科军,阎保平. 基于语句节奏特征的作者身份识别研究[J]. 计算机工程,2011,37(9):4-5.

[13] 年洪东,陈小荷,王东波. 现当代文学作品的作者身份识别研究[J]. 计算机工程与应用,2010,46(4):226-229.

[14] 祁瑞华,霍跃红,郭旭,等. 典籍英译作者身份识别研究[J]. 现代图书情报技术,2015,31(1):31-37.

[15] 刘海涛. 语言是一种复杂网络[J]. 山西大学学报:哲学社会科学版,2013,36(5):65-77.

[16] Li Yong,Wei Luoxia,Li Wei,et al. Small-world patterns in Chinese phrase networks[J]. Science Bulletin,2005,50(3):287-289.

[17] Pavelec D,Oliveira L S,Justino E,et al. Author identification using compression models[C]//10th International Conference on Document Analysis and Recognition. Los Alamitos: IEEE,2009:936-940.

[18] 李菁菁. 功能语言学视角下的文体风格研究[J]. 吉林化工学院学报,2012,29(10):46-48.

[19] Antiqueira L,Pardo T A S,Nunes MG V,et al. Some issues on complex networks for author characterization[J]. Revista Iberoamericana de Inteligencia Artificial,2007,11(36):51-58.

[20] 孟海东,张炼,吕海林,等. 基于图模型的文本分类方法的研究[J]. 计算机与现代化,2010(9):38-40,44.

[21] 刘巧凤. 基于图结构的中文文本聚类方法研究[D]. 大连:大连理工大学,2009.

[22] Heaps H C. Information retrieval: Computational and theoretical aspects[M]. New York: Academic Press,1978.

[23] Egghe L. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments[J]. Journal of the American Society for Information Science & Technology,2007,58(5):702-709.

[24] Toolan M J. Language in literature: An introduction to stylistics[M]. London: Arnold,2009:3-10.

[25] 徐燕文. 以功能词为文体标识符:对小说、新闻、诗歌和学术写作的分析[D]. 杭州:浙江大学,2014.

[26] 张宇,刘雨东,计钊. 向量相似度测度方法[J]. 声学技术,2009,28(4):532-536.

作者贡献说明:

李晓军:提出研究思路,设计研究方案,撰写论文;

刘怀亮:负责论文及实验的处理工作,参与论文撰写;

杜坤:负责实验数据的收集与分析工作。

An Authorship Attribution Algorithm Based on Complex Network

Li Xiaojun Liu Huailiang Du Kun

School of Economy and Management,Xidian University,Xi'an 710126

Abstract: [Purpose/significance] Authorship analysis by means of textual features is an important task in text mining and linguistic studies. Tosolve the problem of low efficiency and high costs in authorship attribution using traditional method,complex networks theory has been employedto tackle this disputed problem. [Method/process] In this paper, some measurable quantities of word co-occurrence complex network of text has been for used for authorship characterization. Based on stylistics and the network features,the approach is defined for authorship identification bycomputing theauthors´stylefeatures similarity. [Result/conclusion] The authorship attribution algorithm based on complex network can use authors´style featureseffectively. The experimental results show high accuracy rate in authorship attribution and prove the validity of this method.

Keywords: authorship attribution text classification complex network feature selection word co-occurrence linguistic