

文章编号: 1003-0077(2018)11-0097-06

从高频词等级相关角度探析《红楼梦》作者

马创新¹, 陈小荷²

(1. 江苏师范大学 语言科学与艺术学院, 江苏 徐州 221009;

2. 南京师范大学 文学院, 江苏 南京 210097)

摘要: 该文提出一种“基于高频词等级相关度的方法”来探析存疑文献的作者信息, 把各份语料中的词型均按照出现频次递减排列并确定等级, 然后通过计算出语料之间高频词等级的相关度, 来推断语料之间语言风格的相似度, 并且把这种方法与“基于词型共现率的方法”和“基于词例共现率的方法”相比较。把《红楼梦》的 120 回均分为 12 份语料, 使用“基于高频词等级相关度的方法”计算这 12 份语料两两之间的相关度。研究发现《红楼梦》的前 8 份语料两两之间相关度高, 后 4 份语料两两之间相关度也高, 而前 8 份语料与后 4 份语料这两部分语料之间相关度低。推断《红楼梦》前 80 回应是同一人所写, 后 40 回应是另一人所写。

关键词: 高频词; 等级; 相关度; 作者信息

中图分类号: TP391

文献标识码: A

Author Identification of *The Dream of Red Mansions* Based on the Rank Correlation of the High Frequency Words

MA Chuangxin¹, CHEN Xiaohe²

(1. Linguistic Sciences and Arts School of Jiangsu Normal University, Xuzhou, Jiangsu 221009, China;

2. College of Liberal Arts, Nanjing Normal University, Nanjing, Jiangsu 210097, China)

Abstract: This paper puts forward an author identification method based on rank correlation of high frequency word types. Words in each corpus are arranged according to the frequency of occurrence and the rank is determined, then the correlation degree between the high frequency word types among the corpus is calculated, which is applied as the similarity of the language style between corpus. This method is compared the word intersection based method and token intersection based method on 12 sub-divisions of total 120 chapters from *The dream of Red mansions*. It is revealed that the correlation is rather high either between the former 8 corpus or between the latter 4 corpus, while the correlation significantly decreases between the former and the latter chapters. It is inferred that the former 80 chapters of *The dream of Red mansions* were written by one author, and the latter 40 chapters by another one.

Keywords: high frequency word types; rank; correlation; author identification

0 引言

古今中外存在着很多作者存疑的文献, 具体情况包括: 有些文献本来就没有作者署名; 有些文献署的是作者笔名, 而世人无法确定该笔名在现实世界中的所指人物对象; 有些文献有具体可查的署名作者, 但世人对该文献作者的真实性和真实性产生怀疑或有争议。比如, 俄裔作家索尔仁尼对于《静静的顿河》

是否为肖洛霍夫所写表示公开质疑, 他认为《静静的顿河》这样的鸿篇巨著, 不是当时只有 20 多岁的年轻人——肖洛霍夫所能写出的, 还有人怀疑肖洛霍夫抄袭了已故作家克鲁乌可夫的作品^[1]。狄更斯和马克·吐温对于《罗密欧和朱丽叶》是否为莎士比亚所写也表示过怀疑, 因为他们觉得莎士比亚的出身是英国平民, 而《罗密欧和朱丽叶》描写的是意大利上流社会的生活^[2]。中国古典小说《红楼梦》的作者也有悬疑, 有些学者认为《红楼梦》全书 120 回为同

收稿日期: 2018-03-13 定稿日期: 2018-05-10

基金项目: 江苏省社会科学基金(15YYC001)

一人所作,而有些学者认为前 80 回与后 40 回并非同一人所作^[3]。

对于如何确定存疑文献的真实作者,我们可以从高频词的等级相关度方面来分析这个问题。相对于中低频词型来说,文献中出现的高频词中,连词、介词和副词占有更大的比例。如果把写文章比作盖房子的话,名词、动词、形容词等实词就相当于砖瓦等建筑材料,连词、介词和副词等虚词就相当于水泥、黄沙等黏合材料。同一作者在写作两部题材不同的作品时,两部作品中所使用的名词重合度会比较低,但所用的连词、介词和副词等虚词重合度会较高^[4-5]。我们所提出的方法是基于这样的考虑:两部文献语言风格的差异不仅体现在词型的重合度上,还更细微地体现在高频词的等级相关度上。如果两部作品是同一作者所写,那么它们的相关系数就会比较高;如果两部作品是不同作者所写,那么它们的相关系数就会比较低。

1 相关研究

1984 年,挪威奥斯陆大学的一个统计学家领导一个小组统计三组文献中的词语特征,三组文献分别是肖洛霍夫的确切作品、存疑作品《静静的顿河》、克鲁乌可夫的作品。他们先是统计不同词汇量与总词汇量的比值,三组分别是 65.5%、64.6%、58.9%;再选择最常见 20 个俄语单词,统计它们出现的频率,分别是 22.8%、23.3%、26.2%;然后统计出现多于一次的词语所占百分比,分别是 80.9%、81.9%、76.9%。上述三种统计结果都显示,肖洛霍夫比克鲁乌可夫更有可能是《静静的顿河》的真正作者^[6]。

在《红楼梦》作者信息的研究方面,最早使用统计方法展开研究的是瑞典汉学家高本汉。高本汉(1952 年)选取了 32 种语法、词汇现象,统计它们在《红楼梦》等五部作品中的出现频率。高本汉根据统计结果,认为《红楼梦》全书 120 回为同一人所作^[7]。1980 年,在美国威斯康星大学举行的《红楼梦》研讨会上,陈炳藻发表论文“从词汇上的统计论《红楼梦》的作者问题”,他把《红楼梦》分为三组,分别是 1~40 回、41~80 回、81~120 回,另外还配上了《儿女英雄传》。他按一定比例从各组中抽选特定词类,再统计各组词语之间的相关系数,计算出《红楼梦》前 80 回和后 40 回的词汇相关度为 78.57%,而《红楼梦》与《儿女英雄传》的词汇相关度仅为 32.14%。

由此认为《红楼梦》前 80 回和后 40 回为一人所作^[8]。

刘钧杰在《红楼梦》前 80 回中选取 40 回,和后 40 回进行比较,对六项语言材料在前部和后部的出现进行统计比较,结论是前、后的语言风格存在明显差异^[9]。陈大康选取 27 个词、46 个字,考察它们在《红楼梦》前后出版的情况,并且分析 89 758 个句子的句长分布及平均句长,认为《红楼梦》前 80 回和后 40 回并非一人所作^[10]。

李贤平从《红楼梦》中抽取了 47 个虚字,统计其在各回中的使用频率,用统计学方法探索各回写作风格的接近程度,并用聚类方法对 120 回进行分析,认为《红楼梦》各个部分是由不同的作者在不同的时期撰写的^[11]。

徐秉铮等从词的相关性和上下文的相关性、字符数的统计、字符串的统计等三方面判断《红楼梦》前 80 回与后 40 回的语言风格有明显的不同^[12]。张运良等将《红楼梦》120 回平均分成 1~40 回、41~80 回、81~120 回等三个集合,然后以句类为特征向量,采用 K 近邻算法作为分类算法构建分类器,实验发现集合 1 和集合 2 句类风格相近,集合 3 句类风格和前两个集合差距较大^[13]。施建军使用支持向量机技术,以 44 个文言虚字频率为特征向量,对《红楼梦》120 回进行分类研究,结果发现,前 80 回与后 40 回在写作风格上存在明显差别^[14]。

2 基于高频词等级相关度的方法

2.1 理论依据

布拉德福提出了频次—等级排序法,这种方法在社会科学领域中被广泛应用^[15],例如,把某部文献中的词型按照其出现频次递减排列,就会呈现出布拉德福分布。布拉德福分布的特点显示:我们考察的具体对象的大多数集中于少数主体来源。例如,人们写文章时总是倾向于选择自己常用的词语。Zipf 发现了词型的出现频率与等级序号之间的关系,任何一篇文章中词型的频次和频次等级的乘积总为一个常数^[16]。

人们在表达一个观点或者描述一个事物时,会有多个同类词语可供选择,有的词语会被经常用到,而有的词语不常被使用。这种选择上的频度不均现象致使被选词语的特征信息变得越来越突出,这又

会反过来作为再次被选的影响因素。如果把个体在表达一个观点或者描述一个事物时选用某词语看作这个词语的一次成功,那么这种成功的累积必然会产生新的成功,这就使得个体在语言运用方面会形成思维定势。文献之间的语言风格差异不仅体现在使用的高频词上,还更加细微地体现在高频词的使用频率等级上^[17]。

2.2 计算方法

为了能够给鉴定作者存疑的文献提供更多的参考信息,我们提出了一种“基于高频词等级相关度的方法”,测量各份语料之间在词型等级方面的相关度,推断“存疑文献”的作者信息。这种算法分为三个步骤:

(1) 首先,对于各份语料,词型均按照出现频次(即词型的词例数)递减顺序排列;

(2) 然后,对于已经排序的词型按照“频率法”确定等级,把出现频次最高的词型等级定为 1,次高的词型等级定为 2,……依次类推,频次相等的词型为一个等级,以其在语料中词频序值为等级^[18]。

(3) 接下来,计算各份语料之间高频词等级的相关度。相关度的计算方法采用“斯皮尔曼等级相关”,如式(1)所示。

$$R_s = 1 - \frac{6 \sum_{i=1}^n D_i^2}{n(n^2 - 1)} \quad (1)$$

其中, D_i 表示每一对数据相应的两个等级之差, n 表示样本数。

斯皮尔曼等级相关适用于研究数据是具有等级性质的成对数据,并且变量之间呈线性关系^[19-20]。但是,两份语料中出现的词型数据并不是成对的,所以采用这种计算方法所得到的相关系数是一个近似值。我们用 ARs 来表示“以语料 A 中特定数量词型为样本”与语料 B 中全部词型比较所得到的相关系数,对于在语料 A 中出现而语料 B 中没有出现的词型,不放在计算范围内。同样,以 BRs 来表示“以语料 B 中特定数量词型为样本”与语料 A 中全部词型比较所得到的相关系数,对于在语料 B 中出现而语料 A 中没有出现的词型,也不在计算范围内。通常选取在语料中出现频次排在前 100、200、300 位的高频词作为样本。语料 A 与 B 的相关度用 ABRs 来表示,ABRs 等于 ARs 与 BRs 的均值,即: ABRs = (ARs + BRs) / 2。也就是说,语料 A 与 B 的相关度就等于:“以语料 A 中特定数量词型为样本”与语料

B 的全部词型比较所得到的相关系数,加上“以语料 B 中特定数量词型为样本”与语料 A 的全部词型比较所得到的相关系数,两个系数之和再除以 2 所得到的商。

2.3 实验与分析

为了验证此方法的效果,我们选取《孟子》《荀子》这两部先秦文献作为实验语料,对这两部文献做人工分词处理。这两部文献都是儒家经典,在主题内容上有着很大的相关性。学术界对于这两部文献的作者,也无异议。把《孟子》语料均分为两部分,两部分语料用“《孟子》一”和“《孟子》二”表示;把《荀子》语料均分为四部分,四部分语料用“《荀子》一”、“《荀子》二”、“《荀子》三”和“《荀子》四”表示。采用“频率法”确定词型等级,选取频次排在前 100 位的词型作为样本,分别测量这七份语料两两之间的相关度,形成如表 1 所示的相似度矩阵。

将表 1、表 2 和表 3 中的数据分别划分为三个区,第一区位于表格左上角,是《孟子》两份语料之间的相关度数据,在表中都以黑色字体显示;第二区位于表格右下角,是《荀子》四份语料相互之间的相关度数据,在表中都以黑色斜体字显示;第三区位于右上角和左下角,是《孟子》两份语料与《荀子》四份语料之间的相关度数据,都以常规字体显示。

表 1 使用“基于高频词等级相关度的方法”得到的相关度矩阵(%)

	《孟子》 一	《孟子》 二	《荀子》 一	《荀子》 二	《荀子》 三	《荀子》 四
《孟子》一		90.84	75.84	75.29	73.29	79.02
《孟子》二	90.84		80.45	73.56	75.36	84.39
《荀子》一	75.84	80.45		82.06	88.17	89.15
《荀子》二	75.29	73.56	82.06		83.12	77.27
《荀子》三	73.29	75.36	88.17	83.12		82.85
《荀子》四	79.02	84.39	89.15	77.27	82.85	

为了评估“基于高频词等级相关度方法”的有效性,我们使用另外两种常用的分析文献相似度的方法与之相比较^[21]。一种是“基于词型共现率的方法”。其计算方法如式(2)所示。

语料 A 与语料 B 的相关度 = (A 与 B 的共现词型数) / (A 与 B 的词型数) (2)

式(2)中,“A 与 B 的词型数”并不等于“A 的词

型数+B 的词型数”,因为语料 A 与语料 B 中有一些共现词型,这些共现词型既出现在语料 A 中,又出现在语料 B 中,不能重复计算,所以“A 与 B 的词型数”等于“A 的词型数+B 的词型数-A 与 B 的共现词型数”。

另一种是“基于词例共现率的方法”。其计算方法如式(3)所示。

语料 A 与语料 B 的相关度=(A 与 B 的共现词型的词例数)/(A 与 B 的词例数) (3)

式(3)中,“A 与 B 的词例数”等于“A 的词例数+B 的词例数”。

表 2 是使用“基于词型共现率的方法”所得到的七份语料相互之间的相关度矩阵,表 3 是使用“基于词例共现率的方法”所得到的相关度矩阵。

表 2 使用“基于词型共现率的方法”得到的相关度矩阵(%)

	《孟子》 一	《孟子》 二	《荀子》 一	《荀子》 二	《荀子》 三	《荀子》 四
《孟子》一		33.90	30.62	29.69	32.13	34.16
《孟子》二	33.90		28.65	27.57	29.12	31.27
《荀子》一	30.62	28.65		33.90	35.62	36.57
《荀子》二	29.69	27.57	33.90		34.87	34.72
《荀子》三	32.13	29.12	35.62	34.87		37.75
《荀子》四	34.16	31.27	36.57	34.72	37.75	

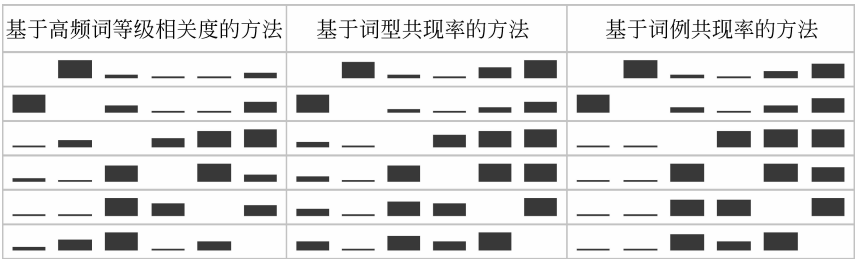


图 1 三种方法的数据柱形图

为了分析使用三种方法分别得到的数据的集中与离散情况,我们计算了每种方法所得到数据的各个区的标准差,把计算结果汇总起来,形成表 4。通过分析表 4,我们发现:(1)使用“基于高频词等级相关度的方法”所得到的数据三个区的标准差分别为 0、3.97%、3.59%,均略大于使用另外两种方法所得到数据标准差,这说明使用“基于高频词等级相关度的方法”所得到的数据波动性略大;(2)使用“基于词型共现率的方法”与“基于词例共现率的方法”所得到的标准差数值都很小,这两种方法所得到的标准差数值差异也很小。

表 3 使用“基于词例共现率的方法”得到的相关度矩阵(%)

	《孟子》 一	《孟子》 二	《荀子》 一	《荀子》 二	《荀子》 三	《荀子》 四
《孟子》一		90.22	87.15	86.54	87.95	89.37
《孟子》二	90.22		87.27	86.09	87.63	89.21
《荀子》一	87.15	87.27		91.08	91.48	91.63
《荀子》二	86.54	86.09	91.08		91.31	90.46
《荀子》三	87.95	87.63	91.48	91.31		91.80
《荀子》四	89.37	89.21	91.63	90.46	91.80	

为了能够直观地观察到使用这三种方法所得到的数据在“量”上的特征,我们使用 Excel 2016 把表 1、表 2、表 3 中的数据转化为柱形图,如图 1 所示。观察图 1 能够发现:

(1)使用“基于高频词等级相关度的方法”所得到的数据三个区之间的区别明显,左上角第一区数据的柱形高度显著高于第三区,右下角第二区的柱形高度也显著高于第三区;

(2)使用“基于词型共现率的方法”和“基于词例共现率的方法”所得到数据三个区之间也有区别,但不如使用“基于高频词等级相关度的方法”所得到数据区别度大,左上角第一区数据的柱形高度显著高于第三区,右下角第二区的柱形高度与第三区左上角柱形高度相关差并不大,区分度较小。

表 4 三种方法的标准差对比(%)

	第一区 标准差	第二区 标准差	第三区 标准差
基于高频词等级相关度的方法	0	3.97	3.59
基于词型共现率的方法	0	1.28	1.97
基于词例共现率的方法	0	0.44	1.09

接下来,计算每种方法所得到数据的各个区的均值,并且计算了各区之间的均值之差,把计算结果

汇总起来,形成表 5。通过分析表 5,我们发现:
(1)使用“基于高频词等级相关度的方法”所得到的数据三个区的均值分别为 90.84%、83.77%、77.15%,介于使用另外两种方法所得到的均值之间;
(2)使用“基于词型共现率的方法”和“基于词例共现率的方法”所得到的数据三个区之间的均值差异比较小;
(3)使用“基于高频词等级相关度的方法”所得到的数据三个区之间的均值差异比较大,第一、三区均值之差为 13.69%,第二、三区均值之差为 6.62%,显著高于使用另外两种方法所得到的相应数据。

表 5 三种方法的均值对比(%)

	第一区 均值	第二区 均值	第三区 均值	第一、三区 均值之差	第二、三区 均值之差
基于高频词 等级相关度 的方法	90.84	83.77	77.15	13.69	6.62
基于词型共 现率的方法	33.90	35.57	30.40	3.50	5.17
基于词例共 现率的方法	90.22	91.29	87.65	2.57	3.64

分析上述数据,能够得出以下结论:(1)“基于高频词等级相关度的方法”所生成的数据,在“第一、三区均值之差”和“第二、三区均值之差”方面均显著高于另两种方法所生成的数据,证明这种方法区分语言风格的能力最强。(2)“基于词型共现率的方法”和“基于词例共现率的方法”所产生的数据波动较小,而“基于高频词等级相关度的方法”所产生的

数据波动略大,离散度略高。

3 探析《红楼梦》的作者信息

以《红楼梦》作为实验语料,使用哈工大社会计算与信息检索研究中心研发的“语言技术平台”对语料作分词处理,把《红楼梦》的 120 回分为 12 份语料,每份语料包含 10 回,这样第一份语料就包含第 1 至第 10 回,第二份语料包含第 11 回至第 20 回,……,依次类推,简写为:一(第 1~10 回)、二(第 11~20 回)、三(第 21~30 回)、四(第 31~40 回)、五(第 41~50 回)、六(第 51~60 回)、七(第 61~70 回)、八(第 71~80 回)、九(第 81~90 回)、十(第 91~100 回)、十一(第 101~110 回)、十二(第 111~120 回)^[22]。

使用“基于高频词等级相关度的方法”计算这 12 份语料相互之间的相关度,均取出现频次排在前 100 位的词型作为样本语料。把相关数据汇总起来,形成表 6 所示的相关度矩阵。为了便于发现前 80 回与后 40 回之间的区别,把表 6 中的数据也划分为三个区,第一区位于表格左上角,是前八份语料相互之间的相关度数据,在表中都以黑色字体显示;第二区位于表格右下角,是后四份语料相互之间的相关度数据,在表中都以黑色斜体字显示;第三区位于右上角和左下角,是前 8 份语料与后 4 份语料两部分语料之间的相关度数据,都以常规字体显示。

表 6 使用“基于高频词等级相关度的方法”得到的相关度矩阵(%)

	一	二	三	四	五	六	七	八	九	十	十一	十二
一		78.96	61.95	55.03	53.41	57.15	59.63	64.10	42.35	37.76	33.92	25.73
二	78.96		71.05	68.18	66.01	65.92	70.92	69.98	54.50	51.51	56.73	47.19
三	61.95	71.05		86.08	75.42	64.95	70.23	59.20	75.83	68.09	59.00	53.30
四	55.03	68.18	86.08		86.15	73.23	70.25	67.08	68.45	62.82	57.86	51.60
五	53.41	66.01	75.42	86.15		71.65	72.24	64.76	62.65	56.61	56.93	46.41
六	57.15	65.92	64.95	73.23	71.65		71.97	76.12	56.55	48.02	49.60	35.66
七	59.63	70.92	70.23	70.25	72.24	71.97		66.63	50.17	49.11	51.26	41.84
八	64.10	69.98	59.20	67.08	64.76	76.12	66.63		42.20	36.82	54.67	38.68
九	42.35	54.50	75.83	68.45	62.65	56.55	50.17	42.20		86.47	65.80	62.37
十	37.76	51.51	68.09	62.82	56.61	48.02	49.11	36.82	86.47		74.95	70.36
十一	33.92	56.73	59.00	57.86	56.93	49.60	51.26	54.67	65.80	74.95		82.21
十二	25.73	47.19	53.30	51.60	46.41	35.66	41.84	38.68	62.37	70.36	82.21	

计算出使用这种方法所得到数据的各个区均值,并且计算出各区之间的均值之差,把结果汇总起来,形成表7。通过分析表7,我们发现:使用“基于高频词等级相关度的方法”所得到的数据三

个区的均值分别为68.51%、73.69%、50.74%,三个区之间的均值差异比较大,第一、三区均值之差为17.77%,第二、三区均值之差为22.95%,差异明显。

表7 各区均值及区间均值之差(%)

	第一区均值	第二区均值	第三区均值	第一、二区 均值之差	第一、三区 均值之差	第二、三区 均值之差
基于高频词等级相关度的方法	68.51	73.69	50.74	-5.18	17.77	22.95

分析上述数据,能够得到以下结论:(1)《红楼梦》的前8份语料相互之间的相关度要高,后四份语料相互之间的相关度也高,即语言风格相似度高;(2)前8份语料与后4份语料之间的相关度要低,即语言风格差异度大。

4 结语

我们把《红楼梦》的120回均分为12份语料,每10回作为一份语料,然后使用“基于高频词等级相关度的方法”,计算这12份语料两两之间的相关度,得到结论:“《红楼梦》的前8份语料两两之间相关度高,后4份语料两两之间相关度也高,而前8份语料与后4份语料这两部分语料之间相关度低。”也就是说,前80回之间语言风格相似度高,后40回之间的语言风格相似度高,而前80回与后40回的语言风格差异很大。由此推断《红楼梦》前80回是同一人所写,后40回是另一人所写。

参考文献

- [1] 余杰. 谁是《静静的顿河》的作者? [J]. 出版广角, 1999, (5): 67-69.
- [2] 霍思温. 是拉辛, 还是莎士比亚? [J]. 中国图书评论, 2007, (3): 106-107.
- [3] 汪维辉. 《红楼梦》前80回和后40回的词汇差异[J]. 古汉语研究, 2010, (3): 35-40.
- [4] 史存直. 汉语词汇史纲要[M]. 上海: 华东师范大学出版社, 1989: 79-96.
- [5] 段磊, 韩芳, 宋继华. 古汉语双字词自动获取方法的比较与分析[J]. 中文信息学报, 2012, 26(4): 34-42.
- [6] 李启虎, 尹力, 张全. 信息时代的人文计算[J]. 科学, 2015, (1): 35-39.
- [7] 蒋绍愚. 近代汉语研究概要[M]. 北京: 北京大学出版社, 2005: 306-307.
- [8] 陈炳藻. 关于《红楼梦》后四十回[J]. 红楼梦学刊, 2002, (3): 267-282.
- [9] 刘钧杰. 红楼梦前八十回和后四十回语言差异考察[J]. 语言研究, 1986, (1): 172-181.
- [10] 陈大康. 从数理语言学看后四十回的作者——与陈炳藻先生商榷[J]. 红楼梦学刊, 1987, (1): 293-318.
- [11] 李贤平. 《红楼梦》成书新说[J]. 复旦学报(社会科学版), 1987, (5): 3-16.
- [12] 徐秉铮, 蔡伟鸿. 从信息论角度探讨《红楼梦》的作者[J]. 中文信息学报, 1990, 4(2): 1-5.
- [13] 张运良, 等. 基于句类特征的作者写作风格分类研究[J]. 计算机工程与应用, 2009, 45(22): 129-131, 223.
- [14] 施建军. 基于支持向量机技术的《红楼梦》作者研究[J]. 红楼梦学刊, 2011, (5): 35-52.
- [15] 靖继鹏, 马费成, 张向矢. 情报科学理论[M]. 北京: 科学出版社, 2009: 33-50.
- [16] G K Zipf, Human behavior and the principle of least effort[M]. Cambridge: Addison-Wesley. 1949: 5-12.
- [17] 马创新, 陈小荷. 文献中的词语分布、词型等级和风格计算[J]. 中文信息学报, 2017, 31(4): 20-27.
- [18] 俞士汶, 朱学锋. 词汇计量研究与常用词知识库建设[J]. 中文信息学报, 2015, 29(3): 16-20.
- [19] Michel J B, et al. Quantitative analysis of culture using millions of digitized books[J]. Science, 2011, 331(6014): 176-182.
- [20] Booth A D. A law of occurrences for words of low Frequency[J]. Information and Control, 1967, 10(4): 386-393.
- [21] 孙清兰. 高频、低频词的界分及词频估计方法[J]. 情报科学, 1992, 13(2): 28-32.
- [22] 马创新, 陈小荷. 基于引文分析的古籍文献影响力评估[J]. 大学图书馆学报, 2016, (1): 16-24.



马创新(1980—), 博士, 讲师, 主要研究领域为计算语言学、知识组织。
E-mail: machxin@126.com



陈小荷(1952—), 博士, 教授, 博士生导师, 主要研究领域为计算语言学、汉语语法学。
E-mail: chenxiaohu5209@126.com