·信息计量与科学评价·

基于Zipf-Mandelbrot定律的人工智能生成内容离散分布规律探析

朱 禹[1] 蔡凌晨[2] 陆泳溶[3] 张逸勤[1,4] 叶继元[1*]
（1. 南京大学信息管理学院，江苏 南京 210023；2. 四川大学数学学院，四川 成都 610065；
3. 四川大学匹兹堡学院，四川 成都 610207；
4. 江苏省数据工程与知识服务重点实验室，江苏 南京 210023）

摘 要：［目的/意义］生成式人工智能的突破引发了人工智能生成内容（Artificial Intelligence Generated-Content，AIGC）的迅猛增长，正在迅速重塑信息资源环境。然而，目前针对AIGC内容特征及其对未来信息资源开发与利用影响的系统、定量分析仍然较薄弱。［方法/过程］本文应用Zipf-Mandelbrot定律，分析AIGC在内容单元上的离散分布规律，并与15种人类自然语料对比，探析AIGC内容分布规律以及生成式人工智能在内容生成过程中的行为模式。［结果/结论］研究发现，AIGC的分布规律符合人类交流中“最省力法则”，生成式人工智能在多种任务中实现了信息生产中多样性和统一性的平衡。人工智能时代，生成式人工智能确能成为重要的新型信息生产源，并能够将AIGC纳入信息资源体系，填补个性化信息内容的不足。本文还提出，不同大模型的Zipf-Mandelbrot参数存在差异，这些参数具有评估模型生成结果性能的潜力。基于Zipf-Mandelbrot定律研究AIGC的内容离散分布特征，本文提出了一种适用于信息资源管理学科和语言学的AIGC定量评估框架。

关键词：人工智能生成内容；生成式人工智能；信息分布；齐夫定律；Zipf-Mandelbrot定律
DOI：10.3969/j.issn.1008-0821.2025.11.015
〔中图分类号〕G250.252 〔文献标识码〕A 〔文章编号〕1008-0821（2025）11-0167-11

Analysis of the Discrete Distribution Patterns of AI-Generated Content Based on the Zipf-Mandelbrot Law

Zhu Yu[1] Cai Lingchen[2] Lu Yongrong[3] Zhang Yiqin[1,4] Ye Jiyuan[1*]
（1. School of Information Management，Nanjing University，Nanjing 210023，China；
2. School of Mathematics，Sichuan University，Chengdu 610065，China；
3. Pittsburgh Institute，Sichuan University，Chengdu 610207，China；
4. Jiangsu Province Key Laboratory of Data Engineering and Knowledge Service，Nanjing 210023，China）

Abstract：［Purpose/Significance］The breakthrough of Generative AI has triggered a surge in Artificial Intelligence Generated-Content（AIGC），rapidly reshaping the environment of information resources. However，systematic and quantitative analyses of the characteristics of AIGC or its impact on the development and usage of information resources in the future remain insufficient.［Method/Process］This study applied the Zipf-Mandelbrot Law to analyze the discrete distribution patterns of AIGC across content units. It also compared these patterns of AIGC with 15 types of natural human corpora to explore the distribution characteristics of AIGC and the behavioral patterns of generative artificial intelligence（GAI）in content creation.［Result/Conclusion］The study finds that the distribution patterns of AIGC align with the “least effort prin-

收稿日期：2024-12-04
基金项目：国家社会科学基金重大项目“学术‘全评价’视域下中国特色哲学社会科学评价体系建设研究”（项目编号：24&ZD323）；江苏省研究生科研与实践创新计划项目“生成式AI驱动的哲学社会科学‘全评价’模型构建与验证”（项目编号：KYCX25_0130）。
作者简介：朱禹（2000-），男，博士研究生，研究方向：人工智能生成内容、学术评价。蔡凌晨（2001-），男，硕士研究生，研究方向：人工智能、高级统计学。陆泳溶（2004-），女，本科生，研究方向：人工智能。张逸勤（2000-），女，博士研究生，研究方向：自然语言处理。
通信作者：叶继元（1955-），男，教授，博士生导师，研究方向：信息资源建设、文献计量学、图书情报学理论。

ciple" in human communication, suggesting that GAI achieves a balance between diversity and uniformity in information production across various tasks. In the era of artificial intelligence, GAI would serve as a significant new source of information production. Additionally, integrating AIGC into the information resources system could address the insufficiency of personalized content creation. The article also reveals that the Zipf-Mandelbrot parameters vary across different large models, demonstrating the potential for evaluating the results of models' performance. By analyzing the discrete distribution patterns of AIGC based on the Zipf-Mandelbrot Law, the paper proposes a quantitative evaluation framework for AIGC applicable to the fields of information resources management and linguistics.

Key words: artificial intelligence generated-content; generative artificial intelligence; information distribution; zipf's law; zipf-mandelbrot's law

近年来，以GPT-4等大语言模型(Large Language Models)为代表的生成式人工智能技术取得了显著进展。生成式人工智能模型凭借其远超传统自然语言处理技术的算法和算据优势，擅长生成自然流畅的类自然语言文本，展现出令人印象深刻的内容汇聚、融合与生成能力[1]。由于生成式人工智能大模型采用了大量的、广泛的各领域数据资源作为训练依据，并具备聚合、融合和推理跨领域数据的能力，其本身可以作为一种新的信息来源。因此，有学者指出，生成式人工智能的出现显著加速了信息资源管理领域中信息资源的流通与循环速率，并在信息资源生态系统中引入了一种全新的信息资源和信息生产者[2]。

然而，对于大模型所生成的内容——人工智能生成内容(Artificial Intelligence Generated-Content，AIGC)区别于人类自然语言的细节特征及其在信息资源子集中的定位，目前尚缺乏系统、科学的定量研究。尽管计算机科学领域的研究者提出了许多用于评估生成式人工智能性能的指标[3-4]，但其主要聚焦计算性能和AIGC的形式层面。在AIGC这一新兴机器信息资源迅速发展的背景下，深入到AIGC的内容维度，分析AIGC的细粒度内容分布特性及其作为新兴信息资源的载体特征显得尤为重要。

基于上述原因，本研究利用Zipf-Mandelbrot定律(以下简称为ZM定律)探讨AIGC的内容离散分布规律。研究旨在回答以下3个关键问题：①AIGC是否遵循自然人类语言中观察到的ZM定律？②如果遵循，如何解释AIGC的新参数？③AIGC在内容单元的离散分布规律对未来的信息资源管理有何启示？

1 相关研究

1.1 人工智能生成内容

自2022年底OpenAI推出ChatGPT以来，生成式人工智能(Generative AI)及其产物AIGC引发了学界和业界的广泛关注。因此，目前信息资源管理学科有关AIGC的讨论中存在两种主流的视野分野[5]。一种研究视野认为，AIGC的核心是"内容"，是一种伴随大数据和人工智能等数智技术进步而产生的新型信息，为图书馆/情报服务领域供给更丰富的信息内容提供了广阔的应用空间[6]；另一种研究视野则认为，AIGC的重点在于"生成"，将AIGC类比为用户生成内容(User-Generated Content，UGC)和专业生成内容(Professional Generated Content，PGC)等一类内容生产方式。这两种视野一是看到了将AIGC作为新型信息资源的潜力，二是认为生成式人工智能可以成为新型信息生产源，简化内容创作过程，使普通用户能够更快速地生成高质量内容[7]。此外，关于AIGC的价值判断，研究者还从工具性和技术性的角度对AIGC进行了批判性审视，主要分为先锋派、科学派和保守派3种价值观思潮[8]，对于能否将AIGC作为现有信息资源集合的补充以及能否采用生成式AI作为一种新型信息生产源来辅助内容生产存在着较大的分歧。然而，上述认知和争议多是基于经验演绎和理论阐释得出的，尚且缺乏定量实验提供的论据。

纵观当前关于AIGC的探讨，无论采取何种视角、价值观来研究AIGC，均看到了AIGC应用于未来信息资源环境的巨大潜力，生成式人工智能将显著改变人工智能时代的信息生产方式和信息资源子集。因此，有必要以定量方式探析作为生成式人工智能技术直接产物的AIGC内容层面的离散分布规律，并据此探讨其对于未来社会整体信息资源环境的意义。因此，对涉及的核心术语进行如下定义：

1）生成式人工智能：具有文本、图片、音频、

视频等内容生成能力的模型及相关技术[9]。

2）自然语言：与人工语言相对，人类日常使用的满足人类交流需要而自然演化出来的语言，如汉语、英语等[10]。

3）人工智能生成内容：由生成式人工智能技术在既有数据训练的基础上生成的有意义、可利用的多媒体信息集合的载体，本质属性是信息资源价值性[11]。

1.2 Zipf-Mandelbrot 定律

在所有信息内容单元离散分布规律相关探究中，1949 年由 Zipf G K[12] 提出的齐夫定律尤为突出。齐夫定律指出，自然语言语料库中词频与其在频率分布中的排名成反比。齐夫定律自提出以来，便持续引起语言学、信息科学及相关领域的关注与讨论，被广泛应用于信息资源管理学科以及计算语言学的自然语言规律研究之中。

尽管齐夫定律被得以广泛证实，但其数学解释仍不尽如人意[13]，这促使学者提出了一些改进版本，如提出 n-gram 版本的齐夫定律[14]。在对齐夫定律的改进研究中，以数学家 Mandelbrot 推导的该定律的广义形式最为经典。他对齐夫定律进行了三参数的修正，使其更具普适性和通用性[15-16]，被称为 Zipf-Mandelbrot 定律（以下简称“ZM 定律”）。

近年来，ZM 定律的研究主要围绕定律验证及应用探索两个方面展开。定律验证方面，研究者在多种语言和应用领域对 ZM 定律进行了验证。尽管 ZM 定律最初是基于英语提出的，但随后被证实同样适用于汉语等表意文字语言[17]。基于古登堡计划语料库的研究显示，ZM 定律的确改进了参数拟合效果[18]。此外，包括汉字甲骨文在内的古代语言[19]，甚至不同音乐流派中也同样能够较好地拟合 ZM 分布[20]。ZM 定律除了在人文研究中得以广泛验证外，还被多个学科广泛应用。例如，ZM 定律不仅在人文研究领域得到广泛验证，其应用也已拓展至多个学科。例如，研究者将 ZM 定律的参数应用于生态系统中优势度与多样性关系的数据分析[21]、天文学中的宇宙星体识别[22] 等自然科学方向，同时在人工语言与语言界面比较[23]、同行评议机制研究[24] 等社会科学领域也发挥了重要作用。上述研究表明，ZM 定律在跨语言和跨领域任务中具有较好的适用性。

基于前人的有益研究，本文尝试拟合 AIGC 的 ZM 分布，验证 ZM 定律在 AIGC 这一区别于人类自然语言的新型文本中的适用性。此外，通过分析 AIGC 中的 ZM 定律参数，解释 AIGC 参数与各类自然语言语料参数的异同，并提出 ZM 定律在生成式 AI 领域中的新应用。

2 研究过程

2.1 语料采集与预处理

语料采集过程如图 1 所示。为确保多样性，本文在文心一言、Claude 和 Gemini 等大模型产品提供的任务示例基础上，对其筛选并改编，形成了 5 种不同的使用场景——聊天（Chatting）、规划（Planning）、评估（Evaluating）、创作（Creation）和写作（Writing），每种场景包含 5 个开放式和半开放式任务。开放式任务要求具有创造性和灵活性，半开放式任务则存在约束条件，即引入了“高尔夫”这一主题的情境，在允许多种解决方案的前提下，使得内容更加聚焦。中英文双语任务总计 100 项（各 50 项），通过调用各个大模型厂商官方提供的 API 接口与提示词工程的方式，由来自美国和中国的主流生成式人工智能工具生成本文所需语料。中文语料总字符数为 218 879，英文语料总字符数为 731 786，其中按 5 个任务场景划分的中文语料平均字符数为 43 775.8 字符，英文语料平均字符数为 146 357.2 字符；按两个开放程度划分的中文语料平均字符数为 109 439.5 字符，英文语料平均字符数为 365 893 字符；按不同模型划分的中文语料平均字符数为 27 359.875 字符，英文语料平均字符数为 91 473.25 字符，满足 ZM 定律验证所要求的“较长文档”的要求。

在预处理阶段，首先，对生成式人工智能工具返回的内容进行了人工质量审查，确保返回内容紧密围绕提示词中提出的问题或任务进行展开。随后，利用 Python 工具（英文使用 NLTK 库和中文使用 Jieba 库）完成了文本预处理工作，包括对英文文本的单词级分词和词形还原以及对中文文本的分词，并对分词后的中英文文本进行词频统计，为后续分析的准确性和内容分布的拟合提供可靠的样本基础。

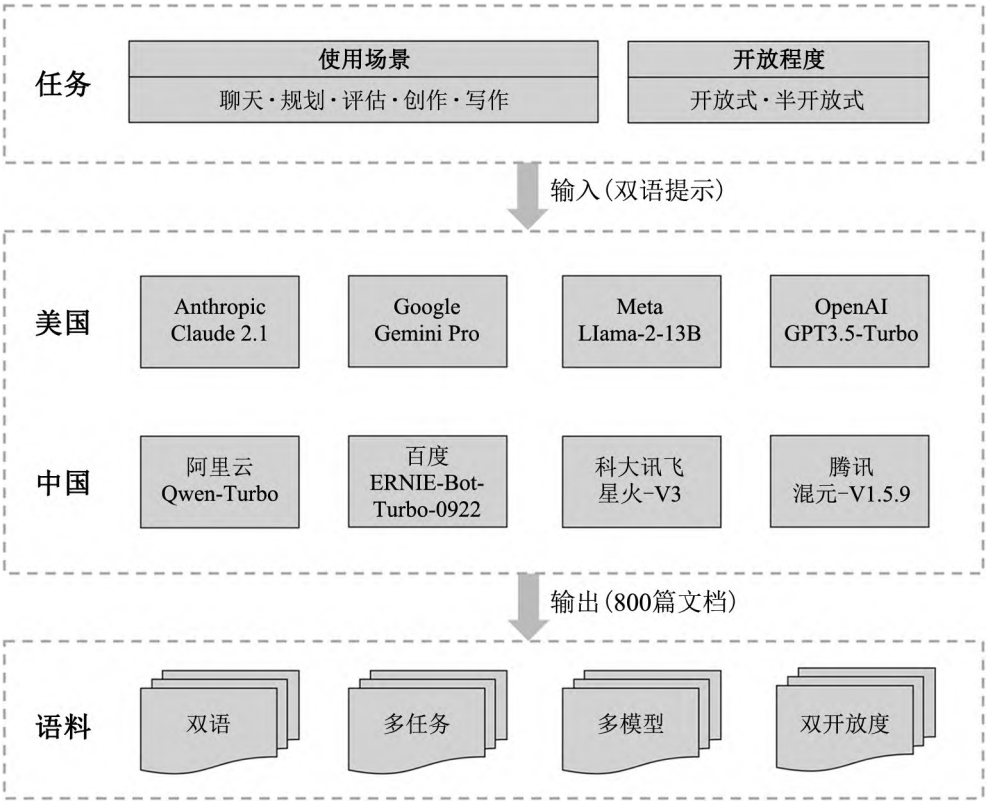

图1 语料采集流程

Fig. 1 Corpus Acquisition Process

2.2 分析方法

2.2.1 Zipf-Mandelbrot定律

Mandelbrot提出的三参数分布方程如式（1）所示：

$$f(r) = c(r + a)^{-b} \tag{1}$$

其中，$0 \leqslant a < 1$，$b > 0$，$c > 0$。Mandelbrot对参数a、b、c的含义作了如下解释：

1）参数a：与词汇数量n相关。通过对a赋予较大的选择自由度，使得该公式更加灵活，从而能够在各种条件下拟合测得的统计数据。

2）参数b:与高频词的数量相关。对于排名r<50的高频词来说，b是一个非递减函数，且其值不会随r的增大而减小。

3）参数c：与具有最高出现概率的词的概率值，即$f(1)$相关。当$a = 0$，$b = 1$时，$c = f(1)$。

该广义公式通过引入a、b、c这3个参数，使得齐夫定律的适应性更强，为本文定量描述不同来源语料的分布规律提供了更精确和灵活的工具。

2.2.2 参数估计

为了探索在不同条件下人工智能所生成内容的词频分布所对应的最优参数，本文采用非线性最小二乘法(Nonlinear Least Squares)。其目标是最小化观察数据与不同参数模型之间残差的平方和，将问题转化为求解如式（2）所示的目标函数的最小值：

$$F(a,b,c) = \sum_{i=1}^{n}\left[y_i - f(r_i \mid a,b,c)\right]^2 \tag{2}$$

其中，(r_i, y_i)代表生成文本的词频排名及其表示对应的词频。据此，最佳参数为：

$$(a,b,c) = \arg\min_{a,\,b,\,c} F(a,b,c)$$

2.2.3 拟合优度检验

部分研究认为，Kolmogorov-Smirnov(K-S)检验是评价信息计量分布的最优方法之一，其在分析大规模英文文本中齐夫定律的应用得到了验证[18]。然而，科学计量学领域的多篇权威文献指出，卡方检验在测试ZM分布模型时更加适用[24-25]。具体而言，Izsák F[26]认为，K-S检验不适用于ZM定律形式，建议使用卡方检验。Meadow C T等[23]也曾将R作为ZM定律参数检验的基准值。基于以上研究，本文认为K-S检验更适用于连续分布的累积分布变量，而不适用于ZM词频分布的离散性质。因此，选择卡方检验作为拟合优度检验方法。

在完成参数估计后，本文使用R^2和Kullback-

Leibler散度（K-LD）[27] 两项指标来评估ZM模型的拟合效果。R^2用于测量观测数据与预测数据之间的差异，其值越接近1，模型拟合效果越好，如式（3）所示：

$$R^2 = 1 - \frac{\sum_i \left(f(r_i) - y_i\right)^2}{\sum_i \left(\bar{y} - y_i\right)^2} \quad (3)$$

其中，$\bar{y}$表示生成文本中词频的均值。K-LD是用于衡量两个概率分布之间差异的指标，量化了使用一种概率分布近似另一种实际分布时的信息损失，相当于两种概率分布的信息熵之间的差异 [28]，反映了两者的不相似性或发散性。K-LD值越小，表示两个概率分布的相似性越高，若K-LD值为0，则表示两者完全相同。其计算方式如式（4）所示：

$$D_{KL}\left(y' \parallel y^*\right) = \sum_{i=1}^{n} y'(r_i)\log\left(\frac{y'(r_i)}{y^*(r_i)}\right) \quad (4)$$

其中，y'表示标准化后的词频序列，y^*表示模型预测的标准化词频序列。

若R^2和K-LD均表现良好，则可以认为所用样本遵循ZM分布。

3 研究结果

3.1 总体分布情况

本文首先将语料库按语言划分为中文和英文两部分，以获取一个粗略但具有全局意义的内容离散分布模型。中文和英文语料库的拟合结果如表1所示。

表1 中英文语料库的ZM定律拟合结果

Tab. 1 Fitting Results of the Zipf-Mandelbrot Law for Chinese and English Corpora

语料语言	R^2	K-LD	a	b	c	信息熵
中文	0.9900	0.1690	−0.6670	0.7634	0.0341	10.3343
英文	0.9872	0.0552	2.5429	1.1654	0.2212	9.7261

英文和中文模型的拟合优度统计值分别为0.9872和0.9900，均非常接近1，表明模型具有良好的拟合效果。此外，K-LD值分别为0.0552和0.1690，进一步支持了上述观察，即人工智能生成的英文和中文内容均符合自然语言中已验证的ZM定律。值得注意的是，在本文的拟合模型中，虽然英文语料R^2大于中文语料，但其差距仅为0.028，而中文语料的K-LD为0.1690，英文语料的K-LD为0.0552，二者K-LD相差3倍，因此认为英文模型的拟合效果优于中文模型。

3.2 细节分布情况

3.2.1 不同使用场景下

本文拟合了由5种生成式人工智能工具执行的五类任务的内容模型。不同使用场景下的拟合结果如表2所示。

表2 不同使用场景下的ZM定律拟合结果

Tab. 2 Fitting Results of the Zipf-Mandelbrot Law for Different Usage Scenarios

使用场景	语料语言	R^2	K-LD	a	b	c	信息熵
规划	英文	0.9887	0.0347	1.5490	0.9778	0.1233	9.1757
写作	英文	0.9888	0.0199	2.1861	1.0921	0.1899	9.0526
聊天	英文	0.9805	0.0267	2.9653	1.1421	0.2393	9.0410
评估	英文	0.9835	0.0215	1.8374	1.0628	0.1632	9.1539
创作	英文	0.9820	0.0376	2.9104	1.1837	0.2617	9.1238
规划	中文	0.9608	0.1168	−0.5040	0.7504	0.0394	9.5201
写作	中文	0.9901	0.1101	−0.6725	0.7377	0.0381	9.4048
聊天	中文	0.9883	0.0906	−0.5598	0.7483	0.0390	9.5248
评估	中文	0.9810	0.1157	−0.7162	0.7298	0.0351	9.4082
创作	中文	0.9894	0.1014	−0.6090	0.7411	0.0392	9.4100

从表2可以看出，除中文语境下规划任务的R^2稍低(0.9608)外，其余模型的拟合优度值均大于0.98。此外，K-LD的最大值为0.1168，表明所有5种场景的模型均具有良好的拟合效果。目前，已经可以观察到一些规律，如英文模型中参数a为正值，而中文模型中该参数为负值。这些结果将在后文与布朗语料库中15种类型的自然语言文本拟合结果进行对比，以探析其参数背后的潜在意义。

3.2.2 不同开放程度下

为了理解人工智能在内容生成中的创造性和发散性，本文还对开放性和半开放性任务的结果进行了测试，拟合结果如表3所示。

表3 不同开放程度下的ZM定律拟合结果

Tab. 3 Fitting Results of the Zipf-Mandelbrot Law for Different Degrees of Openness

开放程度	语料语言	R^2	K-LD	a	b	c	信息熵
开放式	英文	0.9842	0.0577	2.6177	1.1767	0.2247	9.7320
半开放式	英文	0.9842	0.0328	2.8005	1.1701	0.2459	9.3730
开放式	中文	0.9905	0.1030	−0.7011	0.7526	0.0316	10.3345
半开放式	中文	0.9945	0.1186	−0.5678	0.7818	0.0411	9.8956

3.2.3 不同来源大模型下

本文对来自中国和美国的8种主流大语言模型生成的内容进行了区分和测试，旨在识别不同模型生成内容之间的差异，拟合结果如表4所示。

表4 不同来源大模型下的ZM定律拟合结果

Tab. 4 Fitting Results of the Zipf-Mandelbrot Law for Different Large Language Models

大模型(厂商)	语料语言	R^2	K-LD	a	b	c	信息熵
百度	英文	0.9852	0.0393	0.2763*	1.1931	3.0364*	9.0068
Meta	英文	0.9836	0.0387	2.1894	1.1349	0.2030	9.2040
阿里云	英文	0.9907	0.0180	3.0316	1.1059	0.2026	9.3330
OpenAI	英文	0.9795	0.0696	2.3985	1.1715	0.2137	9.6015
谷歌	英文	0.9805	0.0888	1.5797	1.1559	0.1175*	9.5813
Anthropic	英文	0.9882	0.0121	2.4581	1.0399	0.1517	9.4146
科大讯飞	英文	0.9871	0.0269	2.6306	1.1263	0.2085	9.2456
腾讯	英文	0.9802	0.0600	3.0524	1.2212	0.2935	9.1378
QCL2	中文	0.9891	0.0452	−0.6416	0.0705*	0.0392	9.2876
百度	中文	0.9922	0.0492	−0.7271	0.7231	0.0314	9.7164
Meta	中文	0.9969	0.0369	−0.7313	0.7185	0.0311	9.7208
阿里云	中文	0.9791	0.0380	−0.5943	0.8092	0.0428*	9.6685
OpenAI	中文	0.9863	0.0396	−0.6232	0.7731	0.0325	9.9070
谷歌	中文	0.9888	0.0592	−0.7855	0.6760	0.0254	9.8796
Anthropic	中文	0.9925	0.0486	−0.7005	0.7481	0.0345	9.7108
科大讯飞	中文	0.9906	0.0478	−0.6904	0.7271	0.0307	9.9945

注：为呈现简便，仅列出大模型的开发厂商，具体型号见图1；数字右上角*表示离群值。

在表4列出的所有结果中,阿里云的QWen-Turbo在中文环境下的拟合优度最高，达到0.9969；而OpenAI公司的GPT 3.5 Turbo在中文环境下的拟合优度最低，为0.9791。所有K-LD值始终低于0.1，

因此模型的拟合效果非常优秀。

3.3 参数小结

本文最初提出了一个研究问题：ZM 定律是否仍适用于 AIGC？从多个维度对 AIGC 拟合 ZM 分布的 R^2 和 K-LD 值均表现良好，因此，认为 AIGC 依然遵循 ZM 定律，AIGC 与研究者在自然语言中观察到的最省力法则高度一致。图 2 展示了各测量维度中拟合参数的分布情况。

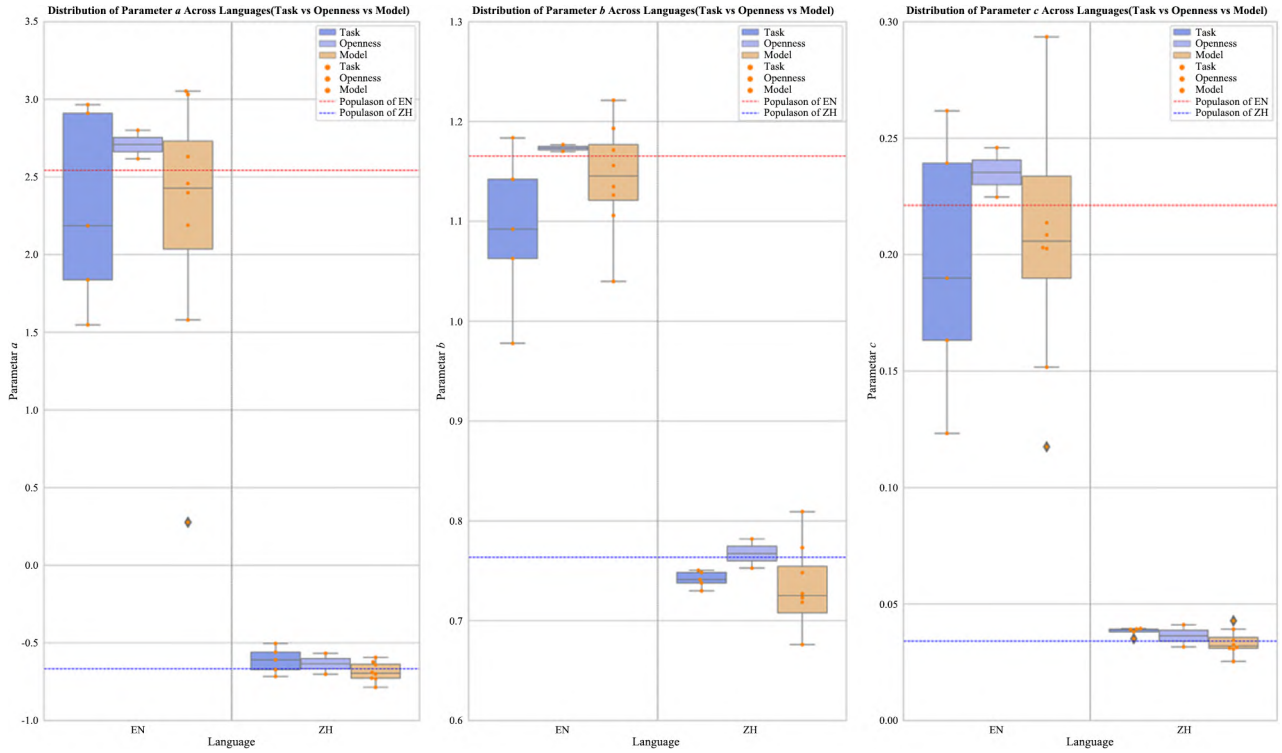

图 2 参数分布箱线图

Fig. 2 The Boxplots of the Distribution of Parameters

1）参数 a：ZM 定律中的参数 a 反映了数据中的饱和效应。从数学角度看，它表示双对数坐标中偏离原点的程度，指示了数据偏离预期分布的饱和点。引入这一偏移参数可以使排序数据更好地符合 ZM 定律，体现真实世界系统的有限性、系统规模限制以及过度采样带来的饱和效应[29]。在词汇分布研究中，参数 a 被认为是一个能够突出高频词汇对分布模式影响的重要指标，反映了词频分布的变化，其中 a 值增加与词汇多样性的降低相关，表明词汇使用更加集中。研究表明，结构化程度较低的内容（如口语或写作草稿）往往 a 值较高，经过编辑、规范的出版物则较低[23]。在本研究中，英文的 a 值为 2.5429，中文为 -0.667。基于四分位距法去除异常值后，英文的范围为 [1.5490, 3.0524]，中文的范围为 [-0.7855, -0.5040]。有趣的是，与自然语言相比，人工智能生成的中文 ZM 分布曲线向左偏移，而英文曲线向右偏移，表明英文 AIGC 的词汇多样性更高。这种差异可能源于英文任务生成的文本长度较长，同时文本长度的变化范围更大——即生成式人工智能模型在生成英文文本时表现出更大文本长度波动性。

2）参数 b：从数学上看，参数 b 表示词频收敛于零的速率。英文和中文语料的总体 b 值分别为 1.1654 和 0.7634，去除异常值后，英文的范围为 [0.9778, 1.2212]，中文为 [0.6760, 0.8092]，与 Zipf 的原始结论 $b \approx 1$ 一致。尽管两种语言的词频分布与自然语言相似，但英文的词频比中文更快趋于零，表明英文信息熵和词汇丰富度较低，这与 Mandelbrot 修正公式一致。

3）参数 c：参数 c 控制 ZM 分布曲线的垂直比例，较大的 c 值会导致曲线在垂直方向上的拉伸更长。英文和中文语料的总体 c 值分别为 0.2212 和 0.0341。去除异常值后，英文的范围为 [0.1233, 0.2935]，中文为 [0.0254, 0.0411]。由于参数 c 与最高频词的概率相关，因此人工智能生成的英文语料中高频词的比例高于中文语料。结合 c 和 b 参数来看，英文的 ZM 分布曲线更高、更细，中文的 ZM 分布曲线则更矮、更胖。

4 讨 论

ZM定律公式包含3个参数，其中参数b反映了人类语言中的通信经济成本与信息交流效率之间的平衡。b值的变化表征了这一平衡的选择性转变，并可能受到多种因素的影响，包括认知状况(如正常人或患精神分裂症)、儿童语言习得的发育阶段以及军事环境下结构化的交流需求等[30]。在理想情况下，参数b通常接近1。较大的b值表示斜率更陡，表明词汇的重复性更高，文本受到更多约束，词汇丰富度较低[31]，从而降低信息传递的效率。相反，较小的b值对应于较平坦的斜率，表明词汇更为多样化。因此，参数b是理解AIGC的内在特性并将其与人类书写的文本区分开来的一个关键分析视角。

具体而言，人工智能生成的两种语言文本的ZM分布b值都接近1，表明AIGC词频趋于零的速率与自然语言基本一致。另外，英文AIGC文本的参数b又高于中文AIGC文本。如图3所示，英文中的高频词数量明显多于中文。

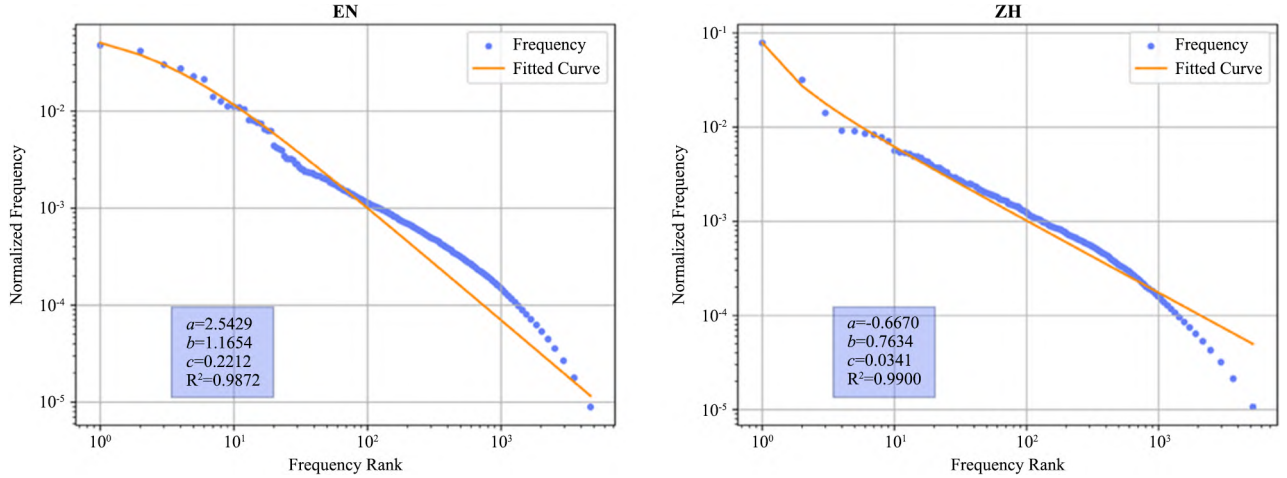

图3 中英文语料的ZM定律对数线性拟合情况

Fig. 3 Fitted Logarithmic Line of the Zipf-Mandelbrot Law for Chinese and English Corpora

为深入讨论该现象，本文利用如式（5）所示的方法计算了中英文AIGC文本各自的信息熵，以衡量词汇丰富度。

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log p(x_i) \tag{5}$$

其中，$p(x_i)$表示随机事件X为x_i的概率。研究结果显示，英文的熵值为9.7261，而中文的熵值为10.3343。因此，中文相对较高的信息熵和较低的b参数值均表明，人工智能生成的中文内容更复杂、更多样化，并可能在返回结果时传递更丰富的涵义。而英文的高频词丰富性较低、词汇量相对较小，表现出倾向于采用最小化交流熵的策略以优化交流效率，并以此节约语言资源乃至大模型的算力资源。

为了理解AIGC和不同创意类型的自然语言之间的Zipf-Mandelbrot参数差异，本文对布朗语料库进行了拟合。布朗语料库是一个代表性的英文平衡语料库，包含15种不同类型的当代美式英语文本，包括冒险类、文学评论、社论、小说、政府文件、兴趣爱好、幽默、学术、民俗、侦探小说、新闻、宗教、评论、浪漫小说和科幻小说。

如图4所示，所有AIGC模型的参数均位于自然语言模型的参数范围内，这表明AIGC在各种任务中的表现可能具有与自然语言使用类似的统计特征。分析a值最高的前10个样本发现以下顺序：幽默(布朗语料库)、聊天(AIGC)、创作(AIGC)、侦探小说(布朗语料库)、浪漫小说(布朗语料库)、写作(AIGC)、宗教(布朗语料库)、评估(AIGC)、科幻小说(布朗语料库)和政府文件(布朗语料库)。最高的10个a值中有4个来自AIGC，反映了与人类撰写的文本相比，高频词区域中的词汇多样性更低的趋势。

AIGC倾向于使用更为非正式的语言和有限的词汇范围，这可能是由于其依赖注意力机制和概率算法。这种方法优先选择高概率的词汇，而非更具创意的低概率选项。a参数值的排名显示，AIGC在聊天和创作类别中的样本与人类撰写的幽默文本具有相似性，尤其在使用非传统和不合逻辑语言方面。这表明AIGC在这些类别中生成的内容可能不适用于需要正式编辑和出版的内容生成任务。

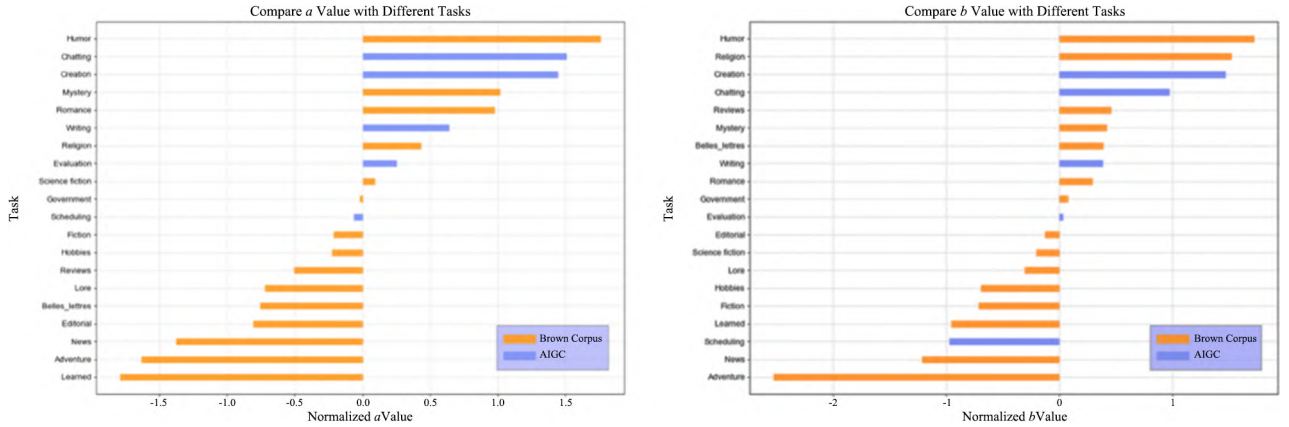

图 4　人工智能生成内容与布朗语料库的 a 和 b 参数对比

Fig. 4　Model Parameter a and b Comparison When AIGC vs. Brown Corpus

另外，AIGC 在评估和规划类别中的样本表现出较低的 a 值，表明其语言结构更加有序和组织化。

图 4 还展示了 b 参数的排名：首先是幽默和宗教（布朗语料库），其次是创作和聊天（AIGC）、评论、侦探小说和文学评论（布朗语料库）、写作（AIGC），最后是浪漫小说和政府文件（布朗语料库）。创作和聊天类别中较高的 b 值表明生成式人工智能在此类任务中采取了一种高效的交流策略，这与当前生成式人工智能工具采取的多轮对话模式和以用户为中心的响应设计保持了一致，可以通过尽量短小精悍且易于理解的相应内容让用户快速理解并获得来自用户的持续性反馈。而在规划类任务等 b 值较低的 AIGC 样本中，人工智能生成的内容则表现出更高的词汇多样性和更确定性的风格。需要说明的是，此处的“确定性”并不意味着缺乏创意，而是指一种更精确和聚焦的内容生成方式。例如，学术论文、新闻报道和政府出版物，其中准确性和审慎的语言使用至关重要。这种特性在正式文件起草或技术文档写作等需要清晰、简明和多样化交流的应用场景中尤为重要和有效。

综上所述，综合考虑参数 a 和 b 的表现，AIGC 在创作和聊天类任务中的样本类似于人类创作的幽默和富有想象力的内容。AIGC 写作的样本同诸如个人日记或宗教文本的文体相符，而规划类任务文本的确呈现出正式、结构化写作的特征。

此外，考虑齐夫分布的一个显著特征是其“长尾”，即在少数高频项之外存在大量低频项——低频词虽然稀少，但通常具有独特的意义，能够为文本注入新意。为了更好地理解这些低频词的分布，本文应用 Booth 对齐夫第二定律进行推导[32]。齐夫第二定律描述了词频的频次分布，即出现频率为 n 的词的数量 I_n 是多少。根据该模型，在长尾分布的假设下，I_n 可以表示为：

$$I_n = \frac{C \cdot N}{n(n+1)}$$

其中，N 是文本语料库的总词数，C 是一个与语料库特性相关的常数。由此，我们可以得到出现频率为 n 的词数 I_n 与仅出现一次的词数 I_1 之间的比例关系，如式（6）所示：

$$\frac{I_n}{I_1} = \frac{CN/[n(n+1)]}{CN/2} = \frac{2}{n(n+1)} \quad (n = 2, 3, 4, \cdots) \tag{6}$$

将低频词的词频序列与齐夫第二定律给出的序列进行比较。结果显示，在不同维度下，低频词的词频序列总体上符合齐夫第二定律（拟合优度 R^2 均大于 0.9，K-LD 均小于 0.06）。对比不同维度可以观察到，影响拟合效果的最显著因素是语言：中文文本的拟合效果总体上强于英文文本，而其他维度的参数差异相比语言的影响并不显著。

5　结论与启示

5.1　结论

齐夫定律是分析语言和信息内容离散分布模式的重要工具。本文应用齐夫定律的变体，即 ZM 定律，探讨了 AIGC 的内容离散分布规律和生成式人工智能在内容生产中的特征，为信息资源管理及有关学科使用定量方法深入 AIGC 内容维度开展更细粒度的内容研究提供了经验。

AIGC 在中英文不同任务中的参数与人类自然语言文本的 ZM 分布参数具有较高一致性，表明 AIGC 的计算语言特征和内容离散分布规律与人类创作的文本相似。一方面，AIGC 具备与人类文本相当的质量和实用性，AIGC 作为准知识资源载体

的信息资源价值性[5]得以确证，使其成为人工智能时代信息资源的宝贵补充；另一方面，AIGC又可以被看作是人类创作者利用新质生产力工具进行内容创作的内容生产模式[11]，未来图书馆、情报服务等单位可以采用“AIGC+”的方式开展更高效的、具有针对性的信息服务。

5.2 启示

5.2.1 生成式人工智能确能成为重要的新型信息生产源

研究者观察到，生成式AI正成为信息的重要来源之一[2]。研究结果显示，生成式人工智能工具在追求更高交流（主要是回答用户提问）效率的同时，能够针对用户提出的个性化任务生成有效内容，并表现出类似于人类语言交流中多样性与一致性的平衡。这种表现与人工智能中的“幻觉”（Hallucination）现象密切相关。与人类依赖记忆进行写作[33]不同，生成式人工智能依赖大语言模型，通过深度学习和数百亿参数生成内容。大语言模型利用神经网络将输入编码为词向量，通过位置编码（如Transformer模型中的技术）等过程来细化词语的含义，从而基于文本的结构和语义预测并生成词汇。在这一过程中，生成式人工智能能够识别用户的内容生成意图，选择适当的响应策略，并表现出一定程度的创造力。尽管幻觉可能导致误导性信息，但也为生成具有创新性的内容提供了可能。

然而，由于大语言模型的随机性，其生成的内容可能引入非事实性的“幻觉”，这也解释了本文在参数分析中观察到的多样性。尽管幻觉目前尚无法完全避免，但可以通过检测和缓解来降低其影响。研究人员正通过提高训练数据质量、进行数据清洗以及采用知识蒸馏等先进技术来减少幻觉现象。对于大语言模型来说，平衡创造力与事实性是一项重要挑战。尽管大多数研究集中于最小化幻觉的影响，但这些“错误”在故事创作、头脑风暴等需要利用非传统思维解决问题的领域中可能具有创造价值。因此，尽管准确性对于实际应用至关重要，但幻觉的创造潜力不能被直接否定[5]。

因此，未来生成式人工智能确能成为重要的新型信息生产源。由于生成式人工智能在处理、学习和生成大规模内容方面超越了人类的能力，且生成式人工智能可以提供多样化和创新的见解，因此可以改善传统信息内容供给单一、个性化服务欠缺的状况。生成式人工智能不仅是内容生产的工具，更是一种整合全领域数据的高价值信息来源。

5.2.2 ZM定律具有评估大语言模型生成内容的潜力

本文对8种生成式人工智能模型生成内容的ZM定律参数进行了拟合，并揭示了如表4所示的一些离群值。鉴于参数a、b和c的背后含义，AIGC拟合ZM分布后出现的离群值具有作为评估模型生成内容质量指标的潜力，可以为优化AIGC内容生成提供指导。例如，三参数的不一致性可能表明模型在处理词汇量、词频分布和高概率词汇时存在问题，从而影响文本的信息熵和创造力，并直接导致生成文本偏离传统自然语言模式。

三参数中，参数c是一个关键指标，因为其与高概率词的出现概率密切相关。参数c的异常可能反映了模型难以准确捕捉关键词汇，进而影响生成文本内容的意义和精准性。因此，优化参数c对于提高模型生成内容的质量至关重要。参数b与高频词的数量相关，能够为改进模型的上下文适应性提供有价值的参考基准。例如，百度千帆Chinese-Llama-2-13B模型生成的中文内容中参数b的异常，表明其处理高频词分布的方式与其他模型不同。其词频下降较慢表明模型生成的内容更为多样化，但也可能导致文本偏离现实，变得不够清晰和准确。这种现象可能来源于百度对Chinese-Llama-2的微调，通过基于大量中文数据集的预训练扩展了其中文词汇量。通过参数b的比较，有助于优化模型生成内容的词频分布，从而平衡多样性与精准性。此外，通过分析参数c的波动，可以改进模型对高概率词汇的捕捉能力。

本文存在以下局限，需要未来完成：①AIGC与自然语言语料库对比：在权衡可比性和代表性后，本文仅选取了英文AIGC和布朗语料库进行对比，未来可以考虑引入更多跨语言语料进行深入比较；②中文AIGC参数a的问题：本研究中，中文AIGC的参数a未与过去基于自然语言的ZM分布参数研究结果保持一致，其现象规律和背后原因尚需进一步的跨模型比较；③参数差异与模型分类：虽然本文观察到不同模型的ZM定律参数存在差异，但未来研究应结合现有性能指标深入探讨这些差异可以如何具体用于模型生成内容的性能评估，甚至是AIGC内容评价。

参 考 文 献

[1] 朱禹，陈关泽，陆泳溶，等. 生成式人工智能治理行动框架：基于AIGC事故报道文本的内容分析［J］. 图书情报知识，2023，40（4）：41-51.

[2] 陆伟，刘家伟，马永强，等. ChatGPT为代表的大模型对信息资源管理的影响［J］. 图书情报知识，2023，40（2）：6-9，70.

[3] Raffel C，Shazeer N，Roberts A，et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer［J］. Journal of Machine Learning Research，2020，21（1）：5485-5551.

[4] Liu Q，Allamanis M，Brockschmidt M，et al. Constrained Graph Variational Autoencoders for Molecule Design［J］. Advances in Neural Information Processing Systems，2018，31：7795-7804.

[5] 朱禹，陈关泽，叶继元. 人工智能生成内容(AIGC)的本质属性及其对信息资源管理学科的影响［J］. 信息资源管理学报，2024，14（6）：60-72.

[6] 李桂华，于泽源. 回答图书馆学的时代之问——“继学开新：图书馆与时代”学术研讨会述评［J］. 中国图书馆学报，2023，49（4）：20-33.

[7] Shi Y，Shang M Y，Qi Z Q. Intelligent Layout Generation Based on Deep Generative Models：A Comprehensive Survey［J］. Information Fusion，2023，100：101940.

[8] 朱禹，叶继元. 人工智能生成内容(AIGC)研究综述：国际进展与热点议题［J］. 信息与管理研究，2024，9（4）：13-27.

[9] 中华人民共和国中央人民政府. 生成式人工智能服务管理暂行办法［EB/OL］.［2025-02-15］. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

[10] 计算机科学技术名词审定委员会. 计算机科学技术名词 第3版［M］. 北京：科学出版社，2018.

[11] 朱禹，叶继元，贾毓洁. 图书馆学的人工智能生成内容(AIGC)：概念框架与研究进路［J］. 图书馆论坛，2025，45（2）：62-71.

[12] Zipf G K. Human Behavior and the Principle of Least Effort［M］. Oxford：Addison-Wesley Press，1949：1-573.

[13] Manin Y I. Zipf's Law and L. Levin Probability Distributions［J］. Functional Analysis and its Applications，2014，48（2）：116-127.

[14] Kim J，Lee C，Zhang B. Difference Between Spoken and Written Language Based on Zipf's Law Analysis［M］//Computational Models of Cognitive Processes，World Scientific，2013：62-71.

[15] Mandelbrot B. On the Theory of Word Frequencies and on Related Markovian Models of Discourse［J］. Structure of Language and its Mathematical Aspects，1961，12：190-219.

[16] Mandelbrot B. An Informational Theory of the Statistical Structure of Language［M］//Jackson W，ed. Communication Theory. London：Butterworths，1953：486-502.

[17] Rousseau R，Zhang Q Q. Zipf's Data on the Frequency of Chinese Words Revisited［J］. Scientometrics，1992，24（2）：201-220.

[18] Moreno-Sánchez I，Font-Clos F，Corral Á. Large-Scale Analysis of Zipf's Law in English Texts［J］. PLoS One，2016，11（1）：e0147073.

[19] Bai Y，Wang X L. Zipf's Law and the Frequency of Characters or Words of Oracles［C］//Intelligent Computing. Cham：Springer International Publishing，2019：828-835.

[20] Manaris B，Vaughan D，Wagner C，et al. Evolutionary Music and the Zipf-Mandelbrot Law：Developing Fitness Functions for Pleasant Music［C］//Applications of Evolutionary Computing. Berlin，Heidelberg：Springer，2003：522-534.

[21] Juhos S，Vörös L. Structural Changes During Eutrophication of Lake Balaton，Hungary，as Revealed by the Zipf-Mandelbrot Model［J］. Hydrobiologia，1998，369：237-242.

[22] Marzo G D，Labini F S，Pietronero L. Zipf's Law for Cosmic Structures：How Large are the Greatest Structures in the Universe?［J］. Astronomy and Astrophysics，2021，651：A114.

[23] Meadow C T，Wang J B，Stamboulie M. An Analysis of Zipf-Mandelbrot Language Measures and Their Application to Artificial Languages［J］. Journal of Information Science，1993，19（4）：247-257.

[24] Ausloos M，Nedic O，Fronczak A，et al. Quantifying the Quality of Peer Reviewers Through Zipf's Law［J］. Scientometrics，2016，106（1）：347-368.

[25] Gupta S，Singh V K. Distributional Characteristics of Dimensions Concepts：An Empirical Analysis Using Zipf's Law［J］. Scientometrics，2024，129（2）：1037-1053.

[26] Izsák F. Maximum Likelihood Estimation for Constrained Parameters of Multinomial Distributions-Application to Zipf-Mandelbrot Models［J］. Computational Statistics and Data Analysis，2006，51（3）：1575-1583.

[27] Lovričević N，Pečarić Đ，Pečarić J. Zipf-Mandelbrot Law，F-Divergences and the Jensen-Type Interpolating Inequalities［J］. Journal of Inequalities and Applications，2018（1）：36.

[28] Goodfellow I，Bengio Y，Courville A. Deep learning［M］. Cambridge，Massachusetts：The MIT Press，2016.

[29] 唐莲，王大辉. 关于Zipf-Mandelbrot律中参数ρ的一种解释［J］. 北京师范大学学报（自然科学版），2011，47（1）：97-100.

[30] Ferrer I Cancho R. The Variation of Zipf's Law in Human Language［J］. The European Physical Journal B-Condensed Matter and Complex Systems，2005，44（2）：249-257.

[31] Koplenig A. Using the Parameters of the Zipf-Mandelbrot Law to Measure Diachronic Lexical，Syntactical and Stylistic Changes-a Large-Scale Corpus Analysis［J］. Corpus Linguistics and Linguistic Theory，2018，14（1）：1-34.

[32] Booth A D. A “Law” of Occurrences for Words of Low Frequency［J］. Information and Control，1967，10（4）：386-393.

[33] Anderson J R，Schooler L J. Reflections of the Environment in Memory［J］. Psychological Science，1991，2（6）：396-408.

（责任编辑：杨丰侨）