大数据与大数：解读齐夫定律

- 美国计算机协会分类 G.2 和 G.3，美国计算机协会主题分类 I.2.7，美国计算机协会主题分类 C.2

Horia-Nicolai Teodorescu，罗马尼亚雅西技术大学，
罗马尼亚雅西；以及罗马尼亚科学院计算机科学研究所，
罗马尼亚雅西/布加勒斯特

摘要—**事实证明，大数据中的一些经验事实是大数属性所产生的效应。齐夫定律中的“噪声”就是这类假象的一个例子。我们阐述幂律分布及类似分布的若干属性；当总体有限，且总体中元素的秩和计数均为自然数时，就会出现这些分布。我们尤其关注定律图中低秩端的情况、定律中的噪声可能性，以及对不同秩处对象类型数量的近似估算。本文重点关注近似解，而非精确解，并讨论了这些发现对齐夫定律解读的影响。**

关键词——幂律、齐夫定律、ζ函数、自然语言处理、网络属性、统计学

I. 预备知识

语言学、经济学、人口统计学、计算机科学（文件大小、下载时间）和网络（互联网拓扑、Web 图）等不同领域中发现的大量统计过程模型令人不安（Mitzenmacher）。Mitzenmacher 列举了“……帕累托：收入分布、股价；齐夫-奥尔巴赫：城市规模；齐夫-埃斯托夫：词频；洛特卡：文献计量学；尤尔：物种和属；曼德布罗特：经济学/信息论”。相应的分布如下：帕累托分布$p(X \geq x) = (x/k)^{-\alpha}$；幂律分布 $p(X \geq x) \sim cx^{-\alpha}$；对数补充累积分布函数 $\ln\ p(X \geq x) = \alpha(\ln k - \ln\ x)$（Mitzenmacher）。我们关注这些分布在离散变量情形下的一些性质，以及它们主要在文本中的应用。虽然我们着重讨论幂律，但对于任何概率分布函数定义在离散变量上且在无穷远处趋于零的过程变体，本文的讨论大体上仍然成立，其中包括对数正态分布和曼德布罗特分布。

本文关注秩为自然数的秩分布：$r = 1$、2、……。假设总体规模任意大但有限；总体计数记为N。总体是一个多重集，其中可能包含多个类型相同的总体元素。类型是离散的，类型数量少于总体中的元素数量。本文中，总体或子总体也称为集合。一个相关总体及其子总体的例子见(Teodorescu M.，2017)、(Teodorescu M.，2017，2018)：这些研究分析了齐夫定律在专利子语料库比较中的效用，所涉及的语料库包括“美国授权专利摘要语料库、美国授权专利标题语料库和美国授权专利权利要求语料库（所有专利均于2007年至2017年11月期间颁发）”。此外，总体可以是由城镇和城市构成的集合，类型是其人口数量；也可以是由一国人口构成的总体，类型是其收入范围；还可以是由在一国获得专利的发明者构成的总体，类型是其专利计数。其他例子包括游客流量（Blackwell等人）、（Lin和Lee）、破产企业（Fujiwara 2003）、软件组件规模（Sharma和Pendharkar，2022）、作为创新指标的专利（O’Neale 2012）、（Teodorescu M.，2018）、互联网社区参与情况的统计数据（Tenorio-Fornés等人），以及文本中的句子。本文关注指定文本中的词（或词元）总体；类型就是词。入门介绍也可参见（Corral等人，2015）；关于不同类型文本（布朗语料库、专利标题语料库、专利摘要语料库和专利权利要求语料库）中齐夫定律的比较实例，可参见该文第49页的（Teodorescu M.，2018），图3。

我们对总体 $n(t)$ 中类型为 t 的元素数量进行计数，并按计数降序排列；计数最大的类型排在第一位，该计数记为 $n(r = 1)$ 或 $n(1)$。我们承认，所研究的总体可能是更大多重集 U 的一个子多重集，而且 U 中的某些类型可能不会出现在所研究的总体中；不过，我们关注的是直到 $n(r) \geq 1$ 的秩，即总体中元素的秩，而不是 U 的秩。

文献中常将齐夫定律的两种形式混为一谈。第一种形式是：随机生成有限数量 N 个服从分布 $\sim \frac{1}{r^{\alpha}}$ 的实数，统计其中有多少个（$n(b_j)$）落在从某个值 x_j 开始划分的一组区间箱 b_j 中，然后将$n\left(b_j(x_j)\right)$ 表示为 $n(x_j)$ 的函数。尤其当 N 相对于 $\frac{1}{x_{j_m}}$ 较小时——其中 x_{j_m} 是最大的x_j——在具有较大 x_j 的区间箱中所计数的元素数量，很可能偏离根据$N \to \infty$所预期的数量，因此，使用这种方式生成齐夫定律的多数论文，其末端都会出现“噪声”。一个例子是（Newman，2005）。在该论文中，作者使用幂律生成了 10^6 个随机值，然后将它们放入范围为 0.1 的区间箱中，并将所得图形描述为：“请注意，在面板右侧朝向尾部的区域，结果变得多么嘈杂。这是因为区间箱中的样本数量变少了，因此统计波动相对于样本数量而言变得很大。”但是，通过使用区间箱，人们实际上创建了元素的“类型”，每个区间箱对应一种类型。噪声之所以出现，是因为横坐标表示的是数值，而不是秩；所得的

© 2023 作者CC BY-NC-ND：知识共享署名-非商业性使用-禁止演绎

图是按数值计数，而不是按秩计数。然而，对于 $N \to \infty$ 而言，每个数值对应的计数往往趋近于每个秩对应的计数。当这种形式的齐夫定律出现较大噪声时，这只是表明：相对于总体中的数值数量而言，所考察的数量或秩过大了。

齐夫定律的第二种形式是按秩计数定律；当问题中规定了类型时，这是一种自然的表示方式；在文本中，词或短语是自然的类型。人们发现，它们的计数与相应的秩遵循齐夫定律。我们讨论的正是齐夫定律的这种形式；当总体中的元素具有定性而非数值特征时，这种形式很自然。正如下一节所讨论的，这种定律形式似乎也存在一个嘈杂的尾端。

在语言和图统计学中，人们关注的是自然数计数，且秩也是自然数。因此，幂律的形式为

$$n(r) = \frac{A}{r^a}, a \in \mathbf{R}, a > 0, A \in \mathbf{R},\ A > 1. \qquad (1)$$

由于计数 $n(r)$ 是整数，式（1）中的等式通常无法成立；因此，应将式（1）理解为取整后的 $n(r) = \left[\frac{A}{r^a}\right]$，或者可以使用下取整函数或上取整函数将右侧转换为整数。由秩的概念和秩排序得出的另一个条件是，对于任意 h

$< j$) ≥, n(h 。这 $n(j)$ $a \geq 0$ $a = 0$ 意味着，其中这种情况（所有类型的秩都相同）并无讨论价值。

定义 (1) 令人想起黎曼函数$\zeta(s) = \sum_{r=1}^{\infty} \frac{1}{r^s}$，$s \in \mathbf{C}$。我们将使用该函数的一些基本性质，其中包括 $\zeta(2)$ 和 $\zeta(3)$的值。由于我们所关注的总体是有限的，所有秩的计数之和是 ζ 的部分和，求和至最大等级 r_M，

$$S_{r_M}(a) = \sum_{r=1}^{r_M} n(r) = A \sum_{r=1}^{r_M} \frac{1}{r^a} \qquad (2)$$

并且都是有限的；我们不关注 S_{r_M} 的收敛，尽管我们根据 $\zeta(s)$ 使用渐近近似。对于较大的偶数指数，部分级数 S_{r_M} 趋近于按 (Ciaurri 等人 , 2015) 中的方法计算的值。注意，
$S_{r_M}(a) = A \cdot H_{r_M}^{(a)}$，其中 $H_m^{(a)}$ 表示指数为 a 的调和数；标准调和数为 $H_m^{(1)}$。在示例中，我们将使用一些通过 Mathematica

® $a = 3/2$ 计算 $\zeta(3/2) = 2\,61237$
的值，例如对于 , ... (Wolfram)。

本文基于一组可能存在争议的假设。最重要的一点是，对象的类型已被明确定义（不是随机变量）。这一点可能并不成立。例如，如果类型是根据人们在某一年中的消费金额范围划分的，那么大多数人可能并不清楚具体数额，因此类型界限会变得模糊。其他涉及排名不精确的应用包括偏好理论和心理学（瑟斯顿），（Li 等，2022）（ 等）。因此，在统计学中，各类型的计数是估计值，而非确定的数值；用随机变量来描述会更合适。这也使得秩变得不精确。研究秩不精确情况下的秩分布，是通信学及其他领域的重要研究方向，参见例如（Li 等 2009）。尽管如此，本文中的结果仍基于自然数计数和秩。我们还假设该分布服从幂律。不过，我们允许幂律中存在噪声；此外，我们还简要讨论了以 $1/r^{(a+\mu/r)}$ 形式表示的依赖于秩的指数。其他类似的规律，例如对数正态分布，在偏离齐夫定律的情况下，也会在不同领域产生与实证结果相似的结果。

II. 由整数计数和秩所产生的性质

实变量的实值幂律与变量和秩为自然数的情况之间存在本质差异。在后一种情况下，自然数施加了限制，并产生了特定性质。其结果之一是，幂律总体图中的“噪声端”并非真正的噪声。

考虑元素数量为N 的总体所对应的幂律基本模型（1），其中每个计数对应一种不同的类型。元素总数为

$$N = \sum_{r=1}^{r_M} \frac{A}{r^a} = A\big(\zeta(a) - \psi(a, r_M)\big), \qquad (3)$$

其中 A 为常数，$A = N/S_{r_M} = N/\sum_{r=1}^{r_M} \frac{1}{r^a}$ 和$\psi(a,\ r_M)$ 将称为泽塔误差函数

$$\psi(a, r_M) = \zeta(a) - S_{r_M}(a) \qquad (4)$$

并用于便于说明。不过，该模型与条件

$n(r) \in \mathbf{N}\ r \in \mathbf{N}\ a > 0\ a \in , , ,$, 不 **R**
兼容，因为部分和不是自然数（即使对于 $a \in \mathbf{N}$ 也不是）。因此，实际计数可以建模为舍入后的数值（如 Teodorescu HN（2023）针对 $a = 1$ 的讨论），也可以建模为附带概率的变量。例如，秩 r 对应的计数可能有两个取值：$\left\lfloor \frac{A}{r^a} \right\rfloor$ 的概率为 p_1，或 $\left[\frac{A}{r^a}\right]$ 的概率为 p_2。在实际应用中，可以选择 $p_1 = p_2 = 0.5$，也可以选择 $p_1 = \int_r^{r+0.5} \frac{B}{r^a} dr$，$p_2 = \int_{r+0\,5}^{r+1} \frac{B}{r^a} dr$。，$B = 1/\int_r^{r+1} \frac{1}{r^a} dr$。目前，我们将使用基于舍入的调整后模型，

$$n(r) = \left[\frac{A}{r^a}\right]. \qquad (5)$$

接下来的问题是，对于足够大的 r，我们会得到

$$n(r) = \left[\frac{A}{r^a}\right] = \left[\frac{A}{(r+1)^a}\right] = n(r+1). \qquad (6)$$

但是，这个等式与统计学中秩的唯一性相矛盾。在这类统计中，只计数“完整”的元素，例如完整词语，而不计数它们的部分。因此，必须允许两个或多个连续等级对应同一个计数。这就是为什么我们在预备知识部分允许满足以下条件：对于任意两个秩 h $< j, n(\text{h}) \geq n(j)$，而不是采用严格条件 n(h) > $n(j)$。

我们可能会关注是否存在一个阈值秩 r_0，使得对于秩 $r < r_0$，模型会针对给定的计数预测出一个单一秩；而对于秩 $r > r_0$，模型则预测

同一计数可能对应两个或更多秩的情况。等价地，定义 r_0 的条件为

$$r < r_0 \to \frac{A}{r^a} \geq \frac{A}{(r+1)^a} + 1, \tag{7}$$

$$\forall r \geq r_0, j \geq 1: \ \frac{A}{(r+j)^a} - \frac{A}{(r+j+1)^a} < 1. \tag{8}$$

第一个条件 (7) 导出了方程 $(A, r \in \mathrm{N})$

$$A(r+1)^a \geq r^a(A + (r+1)^a), \tag{9}$$

其中，r_{01} 是满足上述条件的最小值。第二个条件（8）中的 r 的最小值在 $r_{02} < r_M$ 时导出 r_{02}。然而，这两个条件可能产生不同的值，即 $r_{01} \neq r_{02}$；我们选择了 $r_0 = r_{01}$。这两个条件都必须得到满足，因为 r_0 表示具有保证不同于下一个计数的计数时的最大范围。

为简化讨论，对于秩为 r 的计数，本文后续部分采用基于向下取整函数的定义，

$$n(r) = \left\lfloor \frac{A}{r^a} \right\rfloor. \tag{10}$$

图 1 展示了针对 $M = 10^6, a = -2$ 得到的一条模拟（仿真）类齐夫定律；红色曲线表示未考虑元素数量舍入的定律。

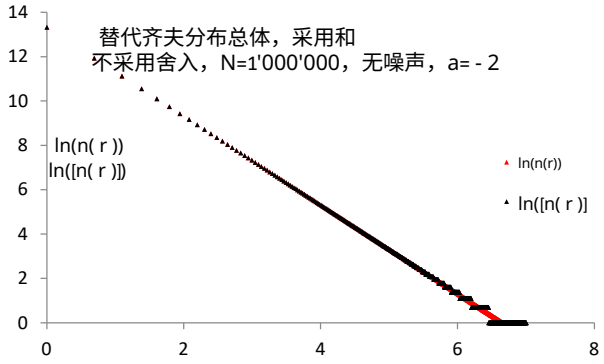


图 1。 齐夫定律的替代形式。

图 2 展示了对于 M 的两个取值，具有相同元素数量的秩的数量因舍入而产生的变化。

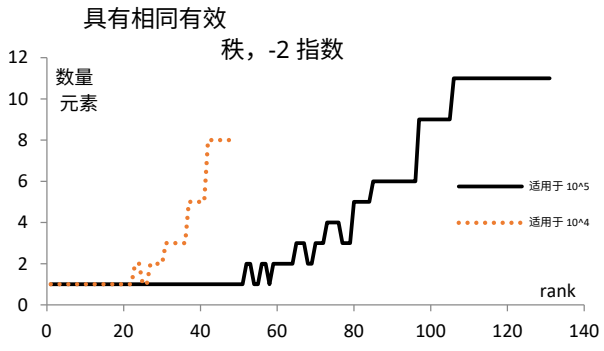


图 2。 具有相同计数的秩的数量。

III。 幂律总体的合并 R ANKS

设 C_1, C_2 是取自同一集合的两个离散值总体（例如，来自同一种语言的词）。设 $m_j = m(r_j)$ $n_h = n(r_h)$ 和 分别为两个总体中秩为 r_j 和 r_h 的元素数量。用 w 表示其中一个总体的一个元素。那么，可能存在 $w \neq w'$，使得 $m(w) = m(w')$。在这种情况下， w, w' 将被赋予连续等级（任一者都可以排在前面）；元素数量相同的情况无论有多少个，也同样如此。例如，如果“是”和“否”都出现 2 次，而“也许”是唯一出现 3 次且秩为 8 的词，那么所赋予的秩可以是 $r(yes) = 4, r(no) = 5$。

$m_{jM} =$ 将满足 1 和 $n_{hM} = 1$ 的最大秩分别记为 r_{jM} 和 r_{hM}。则：

如果 $m_{jM} = m_{hM}$，则 $C_1 \cup C_2$ 中仅出现一次的元素数量（其中，C_1、C_2 和 $C_1 \cup C_2$ 是多重集，$\cup$ 表示多重集的合并）至多等于初始集合 C_1、C_2 中仅出现一次的元素数量之和。在语言学中，合并后的多重集中的一次词数量至多等于初始集合 C_1、C_2 中单现词数量之和。这是因为 $C_1 \cup C_2$ 中没有出现新的词（除 C_1、C_2 中的词之外），而其中一些单现词可能同时出现在两个集合中，因此在合并后的集合中不再是单现词。

假设（不切实际地）C_1 和 C_2 包含完全相同的元素，并满足以下条件：

$$n_{11} > n_{12} > n_{21} > n_{22} > \cdots > n_{1h} > n_{2h} > n_{1h+1} = n_{1h+2} = n_{2h+1} = n_{2h+2} > n_{1h+3} = \cdots = n_{2h+3} = \cdots = n_{2h+u} > n_{1h+u} = \cdots \tag{11}$$

换言之，在这两个集合中，对于从 1 到 h 的每个秩，都有一个词具有相同的计数，共涉及若干个元素。此外，这些计数满足：C_2 中元素的计数介于 C_1 中连续的计数之间。对于高于 h 的秩，在每个集合中至少有两个词具有相同的计数，这意味着在合并后的集合中，至少会有四个不同的元素具有相同的计数，因而拥有连续的秩。由于我们假设 C_1 和 C_2 中的元素彼此不同，因此 $C_1 \cup C_2$ 中将有 2h 个元素没有相同秩的竞争者（不存在具有相同计数的元素，因为我们假设 $n_{12} > n_{21} > n_{22}$ 等）。假设尽管不同，$n_{1h} > n_{2h}$、$n_{1h} \approx n_{2h}$，$\ln r_{2\times h}(C_1 \cup C_2) - \ln r_1(C_1 \cup C_2) = \ln 2\mathrm{h} - \ln 1 = \ln 2 + \ln \mathrm{h}$ 且 $\ln n_{11} - \ln n_{2\times h}(C_1 \cup C_2) \approx \ln n_{11} - \ln n_{1h}(C_1)$。我们明确使用 $(C_1 \cup C_2)$ 这一记号，以表明该秩或计数属于合并后的集合。因此，线段 $y = \ln(count) = a + b\ln(rank)$ 的斜率从 $\frac{\ln n_{11} - \ln n_{1h}}{\ln h}$ 变为 $\frac{\ln n_{11} - \ln n_{1h}}{\ln 2 + \ln h}$。明确地说，合并两个理想集合可能会改变齐夫定律第一部分的斜率。

根据上述假设，在 $C_1 \cup C_2$ 和两个集合 C_1、C_2 中，秩 h 之后的秩数量相同。因此，图的这一部分斜率不会改变。不过，对于每个不同的计数，具有相同计数的元素数量都会翻倍，这意味着对应齐夫定律图中低秩端，也就是“扫帚”端，会变得宽一倍。

因此，将两个具有相同指数的幂律集合合并后，可以得到一个噪声齐夫定律，其扫帚末端可能更粗，且起始尾部的斜率略小。

指数不同时，“扫帚末端”的宽度会受到影响，而起始尾部的斜率可能有所不同。

幂律还有一种变体我们在此不作讨论，即广义幂律，

$$n(r) = \frac{A}{r^a} \cdot q(r) \tag{12}$$

其中 $q(r)$ 是一个“肥队列”函数。

IV. 齐夫定律的“扫帚末端”性质

对于“秩为 r的组中有多少个元素？”这个问题，方程（10）给出了答案。

接下来，考虑这个问题：有多少个秩 r $r+$ 1 $r+$ 、……、 h 拥有相同数量的元素？（也就是说，实际上有多少个秩 r 拥有相同的计数？）当 r 较小时，任意两个连续等级的计数都不会相同；但当 r 较大时，我们预计会出现计数相同的秩；这意味着，对这些元素而言，秩之间的区别是人为约定的。我们可以将这些元素组视为拥有相同的有效排名，其中有效排名由计数决定。若如此，对应于这些有效排名的元素必须属于不同类型；否则，它们的计数相加后，合并起来的元素就会形成一个有效排名不同的组。

考虑根据（10）有两个或多个连续等级产生相同计数的情况。这会使齐夫定律图中的点处于同一水平 $\ln(count)$。用 $\omega(r)$ 表示秩 r 处扫帚末端的宽度。该宽度由具有相同元素数量的秩数目决定。如果元素数量为 1（也就是说，在语言学应用中，这些元素是一次词），那么条件就变为确定最大 r ，使得 $A/r^a < 2$ （这里使用小于而非小于等于，因为我们将在确定给定秩的元素数量时使用向下取整函数），此外，还要确定满足 $1 \le \frac{A}{(r+h)^a}$ 的最大 h，即，

$$1 \le \frac{A}{(r+h)^a} < \frac{A}{r^a} < 2. \tag{13}$$

非平凡条件为

$$(r+h)^a \le A,\ A < 2r^a \tag{14}$$

或 $(r+\mathrm{h})^a < 2r^a$，因此 $r+\mathrm{h} < 2^{1/a}r$，$\mathrm{h} < \left(2^{\frac{1}{a}}-1\right)r$。对于 $a=1$，$\mathrm{h} < r$。一般而言，对于计数 k，条件为

$$k \le \frac{A}{(r+h)^a} < \frac{A}{r^a} < k+1. \tag{15}$$

or, $k(r+h)^a < (k+1)r^a$, or $r+h < \left(\frac{k}{k+1}\right)^{1/a} \cdot r$, or $h < r\left(\left(\frac{k}{k+1}\right)^{\frac{1}{a}}-1\right)$.

取对数，

$$\ln h < \ln r + \ln\left(\left(\frac{k}{k+1}\right)^{\frac{1}{a}}-1\right), \tag{16}$$

若在秩 r 处计数为 k，即 $k = \frac{A}{r^a}$，$\ln k = \ln A - a\ln r$。则可得 h 与计数 k 或秩 r之间的关系，取决于已知哪一项，

$$\ln h < \frac{\ln A - \ln k}{a} + \ln\left(\left(\frac{k}{k+1}\right)^{\frac{1}{a}}-1\right). \tag{17}$$

接下来，我们关注一种定律，它近似描述秩取较大值时“扫帚”尾部的宽度（即趋近于“扫帚”的末端）。

设 M 和 r 为自然数，且 $a \ge 1, a \in \mathbf{R}$。假设 $M > r^a$，在 M/r^a 和 $M/(r+1)^a$之间有多少个自然数？设 $a \in \mathbf{N}$，$n_1 = \frac{M}{r^a} - mod\left(\frac{M}{r^a}\right) \in \mathbf{N}$ 和$n_2 = \frac{M}{(r+1)^a} - mod\left(\frac{M}{(r+1)^a}\right) \in \mathbf{N}$。那么，所求的数为 $n_1 - n_2$。

对于指数为 -1 的幂律，连续等级 n 和 $n+1$ 之间的整数数量为$\nu(n;\ M;\ a=1) \le \frac{M}{n\cdot(n+1)} \to M/n^2$。因此，最大等级为$r_M \le \frac{M}{n^2}$，因为 $r_M \ge 1$。由 $\frac{M}{n^2} \ge 1$ 可得 $n^2 \le M$，因此，$r_M \le \sqrt{M}$。

对于指数 $a = (-)2, \frac{1}{n^2} - \frac{1}{(n+1)^2} = \frac{2n+1}{n^2\cdot(n+1)^2}$。条件为 $1 \le \frac{2n+1}{n^2\cdot(n+1)^2} < 2$ $n^2 \cdot (n+1)^2 < 2n+1 \le$，或】【。$2 \cdot n^2 \cdot (n+1)^2$

对于 $a=2$ 的幂律，连续等级 n 和 $n+1$ 之间的整数数量为 $\nu(n;\ M;\ a=2) \le \frac{(2n+1)M}{n^2\cdot(n+1)^2} \to \frac{2M}{n\cdot(n+1)^2} \sim \frac{2M}{n^3}$。因此，最大等级为 $r_M \le \frac{2M}{n^3}$，因为 $r_M \ge 1$。可得 $n^3 \le M$，因此 $r_M \le \sqrt[3]{M}$。在极限情况下，$r_M = \sqrt[3]{M}$，因此单现词的数量为 $\frac{2M}{n^3}$

我们猜想，对于任意的 $a > 0$，都存在两个常数 λ、Λ，使得函数 $\nu(n;\ M;a) \xrightarrow[n\to\infty]{} \lambda n^{\Lambda}.M\to\infty$

近似公式

设 $M\ r\ s\ q \in \mathbf{N}\ \ a \in \mathbf{R}\ \ s \le \frac{M}{r^a} \le s+1\ \ M = \lfloor sr^a \rfloor +,,,,,,,,。$ $q+\eta\ \ 0 \le \eta < 1$ $M = 16\ r = 3\ a = 1\ 2$ 例如，,,/；那么 $s = 9 < \frac{16}{\sqrt{3}} = 9\ 2376\ \ < 10$ $q = 0$ $\eta =$。 0 2376 ……，并且，,。……那么，

$$\frac{M}{r^a} - \frac{M}{(r+1)^a} = M\frac{(r+1)^a - r^a}{r^a(r+1)^a} \tag{18}$$

$$\frac{M}{(r+1)^a} + c = \frac{M}{r^a},\ c = M\frac{(r+1)^a - r^a}{r^a(r+1)^a} \tag{19}$$

$$\ln c = \ln M + \ln((r+1)^a - r^a) - \ln r^a - \ln(r+1)^a \tag{20}$$

For large r, $\ln r \approx \ln(r+1)$. Then,

$$\ln c \approx \ln M + a\ln r + \ln\left(\frac{(r+1)^a}{r^a} - 1\right) - a\ln r - a\ln(r+1)$$

$$\ln c \approx \ln\left(M\left(\frac{(r+1)^a}{r^a} - 1\right)\frac{1}{(r+1)^a}\right) \qquad (21)$$

对于较大的 r 和 a（$a > 0.5$ 且不太小的），$(r+1)^a - r^a \approx a(r+1)^{a-1}$，例如 . ，. 然后，

$$\ln c \approx \ln\left(M\left(\frac{a}{r}\right)\frac{1}{(r+1)^a}\right) \approx \ln\left(\frac{Ma}{(r+1)^a}\right) \qquad (22)$$

因此

$$\nu(r; M; a) = \lfloor c \rfloor \approx \left\lfloor \frac{Ma}{(r+1)^{a+1}} \right\rfloor \qquad (23)$$

而且 $\frac{M}{r^a}$ 与 $\frac{M}{(r+1)^a}$ 之间的整数数量约为 $\left\lfloor \frac{Ma}{(r+1)^{a+1}} \right\rfloor$，采用近似变体 $\left\lfloor \frac{Ma}{r^{a+1}} \right\rfloor$ 。这证明了上述猜想的一个界。

例如，图 3 显示了上述计算中的误差。

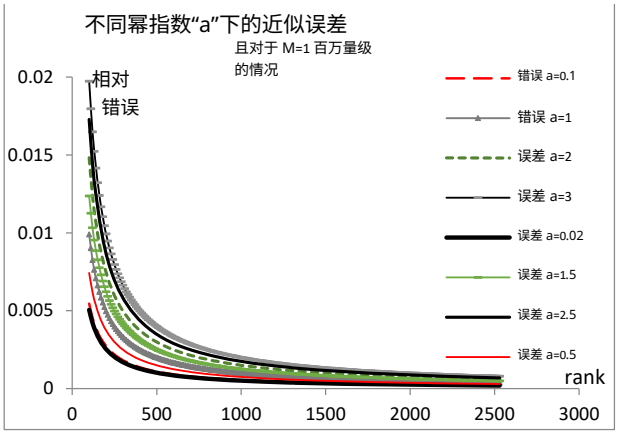
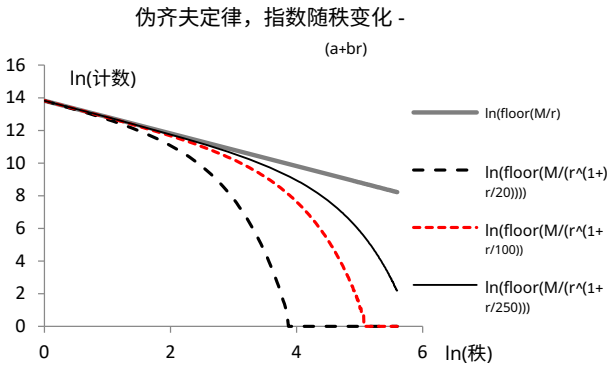


图 3。 根据公式 (23) 近似计算扫帚末端宽度时的误差。

V. 依赖秩的指数

我们以 $1/r^{(a+b/r)}$ 的形式展示秩依赖指数的情况。这或许可以作为一种总体模型，其中齐夫定律似乎由两个线性区段组成。许多文本以及经济领域的案例都呈现出这种行为。图 4 展示了不同 b 常数对应的图；它们类似于文献中针对各种过程根据经验确定的“齐夫定律”，例如 (Williams et al.) 中的图 1 和图 3、(Neuwman 2006) 中的图 4，以及 (Teodorescu M.H.M. 2017) 中的图 3。



VI. 幂律中的噪声

首先请注意，基本幂律中自然会加入“噪声”，因为我们只能考虑整数计数；这种“舍入”噪声为 $\left\lfloor \frac{N}{r_M^a} \right\rfloor - \frac{N}{r_M^a}$，取决于语料库中的词总数和最大等级（而最大等级又由 $\sum_{r=1}^{r_M} \frac{A}{r^a} = N$ 给出）。应用向上取整或向下取整函数时，也会得到类似的版本。我们不关注这种“噪声”。

向秩添加噪声没有意义，因为秩是整数，因此噪声也应当是整数；但是，这样一来，添加噪声只会改变范围，并相应地改变计数 $n(r)$。只要不改变范围，就可以轻松地向计数添加整数噪声 $\eta(r)$ ，也就是说，

$$n(r-1) + \eta(r-1) \geq n(r) + \eta(r) \geq n(r+1) + \eta(r+1) \qquad (24)$$

这些条件与幂律结合后导致了

$$\frac{A}{r^{a-1}} + \eta(r-1) \geq \frac{A}{r^a} + \eta(r) \geq A/r^{a-1} + \eta(r+1) \qquad (25)$$

或者 $\frac{A(r-1)}{r^{a-1}} \geq \eta(r) - \eta(r-1)$ 和 $\frac{A(r-1)}{r^a} \geq \eta(r+1) - \eta(r)$。由于噪声可能为正，也可能为负，因此一个方便的简化条件是

$$\frac{1}{2} \cdot \frac{A(r-1)}{r^a} \geq \eta(r). \qquad (26)$$

本质上，满足上述条件的噪声不会改变秩，但计数 $n(r)$ 仍可处于 $n(r-1)$ 和 $n(r+1)$之间。

接下来，考虑一种更直观的幂律噪声形式，即指数中的噪声。噪声 $\varepsilon(r)$ 应满足某些限制，即 $\frac{1}{r^{a+\varepsilon}} > \frac{1}{(r+1)^a}$，以保持秩不变。对于 $r > 1$ a $(r+1) > (a+\varepsilon)$ r $\frac{a+\varepsilon}{a} <$ ，$\frac{\ln(r+1)}{\ln r}$ 这会得到 ln ln ，或 ，进而得到

$$\varepsilon(r) = \varepsilon_0 \cdot a \cdot \left(\frac{\ln(r+1)}{\ln r} - 1\right) \qquad (27)$$

其中，ε_0 是一个取值小于 1 的随机变量；一个简单的例子是均匀噪声 ε_0 。（注意，$\frac{\ln(r+1)}{\ln r} - 1 =$

$\frac{\ln(r+1)-\ln r}{\ln r} \xrightarrow[r\to\infty]{} 0$）。例如，$\varepsilon_0$ 可以是最大绝对值小于 1 的均匀噪声。）

可以通过乘法选择其他函数 $g(r;\ a)$ 来进行噪声整形，前提是它们满足条件$g(r;$ 对于所有秩 $a) \le a \cdot \left(\frac{\ln(r+1)}{\ln r} - 1\right)$ $r < r_M$。

在计算语言学的语境中，文本或作品全集在齐夫定律中的噪声量可能反映作者风格。

VII. 关于单现词以及 Honoré 指数和西歇尔指数的说明

假设将一篇文本添加到另一篇文本时，词语出现的概率不会改变，那么任何文本的添加都会增加不同词语的数量，而不会改变它们的秩（基本上如此）。然而，原本为一次词的词可能会变成二次词，并且会出现新的单现词——也就是说，在其中一篇文本中是一次词、但在另一篇文本中不出现的词，在合并后的文本中仍然是一次词。然而，合并后的文本中不会有任何原先不是一次词的新词变成一次词。相同的单现词在两篇文本中（合并之前）同时出现的概率非常低，因为一个词在任一文本中成为一次词的概率本来就很低。对于给定的 N，一次词的概率是多少？数量显著不同的一次词能告诉我们有关作者、语料库等什么信息？

Brooks（Brooks 2009），一位来自美国联邦调查局的作者，在她对单现词的研究中得出结论：“类型/词例比、丰富度分数和只出现一次词百分比这些作者归属判定技术，未能区分嫌疑人作品（SPs）与语料库其余部分的表现。”正如我们所证明的，这是正常现象，因为除了作者身份或风格之外，还有多个因素会决定单现词是什么以及有多少个，尤其是数字属性。

了解单现词如何出现在大型文本中，对于“个人语言习惯和话语独特性这一概念”以及法庭上的作者身份与剽窃分析、法庭语言学都很重要（Coulthard，2004）。作为剽窃有力指标的“共享词汇比例”，用于短文本时可能效果良好，正如 has to be 的分析所示（Coulthard，2004）；但当文本规模较大，且证据基于罕见词（单现词、二次词或类似词）时，这种方法就有其局限，因为这些词的出现可能由概率和数论解释。文体计量指标也会受到前几节讨论的数值伪影影响，例如 Honoré 指数 H = 100 × log N /(1 − (V1/V))，其中“N 是目标文本中的词元数量，V1 是单现词（仅出现一次的词）的数量，V 是词类的数量。”（Baayen 等，1996），（Ishihara）；Brunettte指数（Chandrika 和 Kallimani）；以及考虑了二次词的 Sichel 测度（指数）。

对于呈现一致幂律分布且指数不同的两位作者的大型文本而言，其文本中单现词和二次词数量存在差异这一事实，并不能作为作者身份或风格的指标，因为幂律预示着这种差异本应存在。（我们在（Teodorescu & Bolea, 2019）以及其他研究中都曾陷入这一陷阱。）只有词汇方面的差异（即不同的单现词和二次词）可能显示出风格差异。当然，文本的齐夫定律偏离程度不同（文本长度相同时），或幂律指数不同，可能表明它们的风格和作者身份不同。我们建议不使用 Honoré 指数（它主要由 (23) 决定），而采用一个有意义的指数：$I_{hp} = \frac{V_{hp} - \nu(r_M;\ M\ a);;;}{\nu(r_M\ M\ a)}$，其中 V_{hp} 是文本中单现词的数量。这个指数需要计算 a 和 r_M；如果文本不符合恰当的齐夫定律，则无法应用该指数。尤其对于较短的文本，最好精确计算 $\nu(r_M;\ M;\ a)$，而不是使用 (23)。所提出的指数在词汇贫乏时为负，在词汇丰富时为正。对西歇尔指数也建议作类似修改。请注意，所建议的指数衡量的是文本偏离理论齐夫定律的程度。进一步而言，总噪声（所有秩的偏离之和）指数是单现词和二次词指数的推广。

VIII. 讨论、进一步的问题与结论

仍有许多相关问题和潜在发展方向尚未探讨。考虑一个严格递减至零的函数 $\mathrm{h}(n): N \to R$。定义一个类似于 ζ 函数的函数 $\xi(\mathrm{h}) = \sum \mathrm{h}_{k=1}^{\infty}\ {}^{n}(k)$；特别是，$\mathrm{h}(n) = \frac{1}{n^a}, a > 0$（或 $a \in \boldsymbol{C}$）。问题在于确定 $\nu(n; M; \mathrm{h})$ 在 n 的两个连续取值之间的整数函数，这是一个与针对 ζ 函数所讨论的问题类似的问题。

对于来自各种领域、长度各异的文本，语言模型可能表现出符合齐夫定律的可预测行为。目前尚不清楚，能否通过相关的齐夫定律或其衍生指标（单现词、西歇尔指数、Honoré 指数等），将 ChatGPT 或类似的 AI 机器识别为“作者”。不过，在对抗性场景中，可以训练或调整这些模型，例如通过规则进行教导，使其随时间改变特定的幂律；比如，加入噪声使其偏离基本幂律，或让其在一组指数之间随机跳转。

对于概率由幂律 $p(r) \sim 1/r^a$ 给出的过程，秩为 r 的元素数量的估计值为 M/r^a，M 为常数，平均而言会得到前文已讨论的情况。

对于涉及其他类似幂律的分布（例如对数正态分布）的模型，未来应进一步开展本文所述的分析。

关于齐夫定律的更多应用示例，详见（Teodorescu M., 2023）。附录 1 中讨论了一个具体主题。

此外，本文的讨论表明，通常用于根据经验数据确定近似幂律的方法（即通过最小化平方误差来确定）并不恰当；这一问题将在另一篇文章中探讨。

致谢

特别感谢：Mike HM Teodorescu 教授指出，齐夫定律图表的“扫帚”端需要更深入的研究，包括探究其“噪声”机制；他协助推导了大部分公式，提出在分析不同语料库的风格差异时使用自助法，并提出使用杰卡德、余弦距离以及距离来

比较单现词和二次词的词汇表，并确定含噪幂律序列在“扫帚”端方面与理想序列的差异程度；此外，他还提出了贯穿全文、尤其是在第一节和最后一节中出现的若干想法。

参考文献

[1] Baayen H, Van Halteren H, Tweedie F.《走出阴影洞穴：使用句法标注增强作者归属判定》。《文学与语言计算》。1996年9月1日；11(3)：121-32。[2] Barrientos A.F., Sen D., G.L. Page, and D.B. Dunson，《不确定排名与排序的贝叶斯推断：在球员与阵容排名中的应用》。《贝叶斯分析》（2022）第1卷，第1期，1-30。[3] Blackwell, C., Pan, B., Li, X., and Smith, W.，《旅游流中的幂律》（2016）。旅游与旅行研究协会：推动全球旅游研究。63。https://scholarworks.umass.edu/ttra/2011/Oral/63。[4] Brooks R. L. 请用自己的话作答：利用作者归属判定识别翻译测试中的作弊行为。载于：Lynda Taylor 和 Cyril J Weir（编），《语言测试的重要性——考察评估更广泛的社会与教育影响》。ALTE剑桥会议论文集，2008年4月。剑桥大学出版社，英国剑桥。[5] Chandrika CP, Kallimani JS. 使用基于特征档案的方法进行卡纳达语文本作者归属判定。载于《第二届机器学习、物联网、智慧城市与应用最新趋势国际会议论文集：ICMISC 2021》，2022年（第679-688页）。施普林格新加坡。[6] Ciaurri O., Navas L.M., Francisco J. Ruiz, & Juan L. Varona.（2015）。ζ (2k)的简单计算。《美国数学月刊》，122(5)，444–451。https://doi.org/10.4169/amer.math.monthly.122.5.444。Conrad B., Mitzenmacher M.，《猴子随机打字的幂律：不等概率情形》。[7] Corral Á, Boleda G, Ferrer-i-Cancho R（2015）《文本中的词频齐夫定律：词形与词元》。《公共科学图书馆·综合》10(7)：e0129031。https://doi.org/10.1371/journal.pone.0129031。[8] Coulthard M.，《作者识别、个人语言习惯与语言独特性》，《应用语言学》，第25卷，第4期，2004年12月，第431–447页，https://doi.org/10.1093/applin/25.4.431。[9] Fujiwara Y.，《企业破产中的齐夫定律》。arXiv:cond-mat/0310062v1 [cond-mat.stat-mech] 3 2003年10月。[10] Ishihara S.，《语言文本证据的强度：融合式法证文本比对系统》，《国际法证科学》。http://dx.doi.org/10.1016/j.forsciint.2017.06.040。[11] Lee, WC.，《考虑位置参数与尺度参数的秩统计量联合分布及其检验效能研究。《统计分布与应用期刊》1，6（2014）。https://doi.org/10.1186/2195-5832-1-6。[12] Li F., Yi K., Jestes J.，《分布式概率数据排名》。SIGMOD’09，2009年6月29日至7月2日，美国罗得岛州普罗维登斯。2009 ACM 9781605585511。[13] Li X., Yi D. and Liu J.S.，《含协变量及异质排名者的排名数据贝叶斯分析》。《统计科学》，第37卷，第1期，2022年2月。[14] Lin, JH., Lee, WC. 用于流行病学数据充分原因建模的互补对数回归。《科学报告》6，39023（2016）。https://doi.org/10.1038/srep39023。[15] Lu L., Zhang Z-K, Zhou T（2010）《齐夫定律导出希普斯定律：分析有限规模系统中的二者关系》。《公共科学图书馆·综合》5(12)：e14139。doi:10.1371/journal.pone.0014139。[16] Mitzenmacher M.，《幂律研究的新方向》。Radcliffe.ppt。哈佛大学，https://www.eecs.harvard.edu, https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.eecs.harvard.edu%2F~michaelm%2FTALKS%2FRadcliffe.pp t&wdOrigin=BROWSELINK。[17] O’Neale DRJ, Hendy SC（2012）《作为创新指标的专利幂律分布》。《公共科学图书馆·综合》7(12)：e49501。https://doi.org/10.1371/journal.pone.0049501。

[18] Newman MEJ (2005) 幂律、帕累托分布和齐夫定律，《当代物理学》，46:5，323-351，DOI: 10.1080/00107510500052444。另见 arXiv:cond-mat/0412004v3[凝聚态物理.统计力学] 29 2006年5月。[19] Sharma, S, Pendharkar, PC. 关于软件组件规模中幂律分布的分析。《软件演化与过程杂志》。2022；34(2):e2417。doi:10.1002/smr.2417。[20] Tenorio-Fornés Á, Arroyo J, Hassan S. 维基社区中的参与：重新审视其统计特征。《PeerJ 计算机科学》。2022年1月3日；8:e792。doi: 10.7717/peerj-cs.792。[21] Teodorescu H.N.，齐夫定律参数及相关语料库属性。提交至第27届国际知识型与智能信息及工程系统会议（KES 2023）。[22] Teodorescu HN, Bolea SC. 基于文体距离的文本分段。载于：2019年国际语音技术与人机对话会议（SpeD），2019年10月10日（第1-6页）。IEEE。[23] Teodorescu, M., 2018. 企业内部及企业间的知识流与知识产权——经济学与机器学习方法。哈佛商学院博士学位论文。永久链接 http://nrs.harvard.edu/urn-3:HUL.InstRepos:41940978。[24] Teodorescu M.H.M.，《战略研究中的机器学习方法》，2017/7/31，哈佛商学院研究论文系列，报告编号18-011，见 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3012524。[25] Teodorescu M.H.M.，2023，《创新文献中的自然语言处理方法》。即将刊出。[26] Thurstone LL. 比较判断定律。《心理学评论》。1927年7月；34(4):273。[27] Williams JR., Lessard P., Desu, S., 等。齐夫定律适用于短语，而非词。《科学报告》5, 12209 (2015). https://doi.org/10.1038/srep12209[28] Wolfram Alpha，Zeta[3/2]。https://www.wolframalpha.com/input?i=Zeta%5B3%2F2%5D。

附录 1

我们关注函数 $\zeta_n(a,M)=M\sum_{r=1}^{n}\frac{1}{r^a}$, M 一个较大的正数, $a>0$ 的性质。具体而言，我们关注 n 的取值，其中对于某个 $m<M$ 的整数，$m<\zeta_n(a,M)\le m+1$。确切地说，我们记作

$$\varphi(m;M)=|\{n|m<\zeta_n(a,M)\le m+1\}|$$

并且关注确定这个函数，其中对于某个 m，函数值表示部分和的数量$\zeta_n(a,M)=M\sum_{r=1}^{n}\frac{1}{r^a}$，其取值介于 m 和 $m+1$ 之间。函数 φ 表示具有幂律分布的一组 M 个对象的尾部宽度。还可以根据其他概率密度函数或累积分布函数的级数定义类似的函数。

更一般地，对于一个函数 $\eta(n,a,M)$，它对 $\zeta_n(a,M)$ 函数中的噪声进行建模，我们可能会关注

$$\varphi_\eta(m;M)=|\{n|m<\zeta_n(a,M)+\eta(n,a,M)\le m+1\}|$$

函数 $\varphi(m;M)$ 和 $\varphi_\eta(m;M)$ 具有以下性质：

$$\sum_{m=0}^{M}\varphi(m;M)=\sum_{m=0}^{M}\varphi_\eta(m;M)=M.$$