Ученые записки Тартуского университета, 1987, вып. 774 ("Квантитативная лингвистика и автоматический анализ текстов")

ФОРМА ПРЕДСТАВЛЕНИЯ РАНГОВЫХ РАСПРЕДЕЛЕНИЙ

В.В.Нешитой

В статье обсуждается традиционная форма представления ранговых распределений слов в виде "прямой Ципфа" в логарифмических координатах, отмечаются ее недостатки и предлагается новая форма представления такого рода распределений. Она значительно расширяет возможности исследования статистических ранговых распределений. Рассматриваются причины, порождающие иллюзию справедливости закона Ципфа. Для описания статистических ранговых распределений однородных единиц предлагаются обобщенные распределения, одним из частных случаев которых является закон Ципфа.

Традиционная форма представления ранговых распределений. Наиболее широкое применение ранговые распределения имеют в лингвистике и информатике. Если все разные слова, которые употребились в некотором тексте (выборке), упорядочить по невозрастанию абсолютных или относительных частот и каждому слову приписать порядковый номер (ранг), то зависимость между относительной частотой слова p_z и его рангом z в первом приближении может быть описана законом Дж.К.Ципфа (Zipf G.K., 1940):

$$p_z = \frac{\kappa}{z^{\gamma}}, \qquad (1)$$

где по данным автора этого закона $\kappa \approx 0{,}1$, $\gamma \approx 1$.

В информатике этим законом описывается ранговое распределение журналов по числу опубликованных в них статей на определенную тему. Действительно, если p_z обозначает долю статей из общего их числа (по данной тематике), опубликованных в журнале с рангом z, то накопленная доля статей в z первых журналах на основании (I) будет равна (при $\gamma = 1$)

$$F(z) = \sum_{i=1}^{z} p_i = \sum_{i=1}^{z} \frac{\kappa}{i} \approx \kappa(\ln z + C). \qquad (2)$$

Выражение (2) по форме совпадает с формулировкой закона рассеяния публикаций С.Брэдфорда в ранговой интегральной форме (Bradford, 1948) $X(z) = a + b\log z$, где $X(z)$ -накопленное число статей в z первых журналах; a, b - параметры.

Выражение (I) при логарифмировании преобразуется в прямую

$$\ln p_z = \ln k - \gamma \ln z, \qquad (3)$$

которая утвердилась как одна из основных форм представления ранговых распределений. Однако график зависимости $\ln p_z$ от $\ln z$, построенный по опытным данным, близок к прямой лишь в средней части. Наличие кривизны в областях низких и высоких рангов принуждает исследователей либо вводить поправки в модель Ципфа-Брэдфорда, либо искать новые, более подходящие модели.

Однако несмотря на многочисленные попытки усовершенствовать модель Ципфа-Брэдфорда, в настоящее время не существует единого уравнения (распределения), которое с достаточной точностью описывало бы все многообразие статистических ранговых распределений. Причин неудовлетворительных аппроксимаций подобных распределений много. Рассмотрим наиболее существенные из них.

Первая причина: неудачно выбрана форма представления статистических ранговых распределений (в виде прямой (3)), не отражающая в должной мере их характерных особенностей. "Уже верное отражение природы,- писал Ф.Энгельс, - дело трудное, продукт длительной истории опыта" [К.Маркс, Ф.Энгельс, с.639] .

Принятая форма представления ранговых распределений несет слишком мало информации о статистическом распределении. На таком графике колебания частот мало заметны, поскольку последние изображены в логарифмическом масштабе. Кроме того, такое преобразование кривой распределения не имеет вероятностного смысла.

Новая форма представления ранговых распределений. В связи с вышесказанным нам представляется целесообразным перейти к другой форме представления ранговых распределений, а именно $zp_z = f(\ln z)$. По оси ординат будем откладывать произведение ранга на относительную частоту слова с данным рангом, а по оси абсцисс - натуральный логарифм ранга. При построении такого графика будем использовать координаты середин ступенек статистического дискретного распределения (см.рис.I), т.е. те точки дискретного распределения, через которые проходит выравни-

вающая непрерывная кривая распределения. Этими точками на рис.I являются:

z	0,5	I,5	2,5	...	6	I0	I8
m_z	I0	7	5	...	2,5	I,5	0,5

График зависимости $zp_z = f(\ln z)$ имеет принципиальные преимущества перед традиционной формой представления ранговых распределений. Во-первых, он представляет собой кривую распределения, что легко доказать следующим образом. Пусть $p(t)$ - невозрастающая плотность распределения, аппроксимирующая относительные частоты p_z. Тогда t будет соответствовать рангу z. Следовательно,

$$\int_{-\infty}^{\infty} tp(t)d\ln t = \int_{0}^{\infty} tp(t)\frac{dt}{t} = \int_{0}^{\infty} p(t)dt = 1,$$

т.е. площадь под кривой $tp(t) = f(\ln t)$ равна единице (условие, обязательное для кривых распределения). Во-вторых, на такой кривой видны колебания самих частот (по оси ординат), а не их логарифмов. В-третьих, статистические ранговые распределения однородных случайных величин имеют одновершинную кривую распределения. Это дает возможность устанавливать однородность или неоднородность ранговых распределений, выделять неоднородную часть, оценивать минимально необходимый объем выборки для установления типа выравнивающей кривой и нахождения оценок параметров и т.д. [Нешитой В.В., I984] .

Вторая причина неудовлетворительных аппроксимаций статистических ранговых распределений заключается в попытке описать одним уравнением распределения неоднородных случайных величин. Частотный словарь, как правило, представляет собой неоднородную совокупность элементов. Действительно, роль полнозначных и служебных слов в тексте различна, при этом служебные слова употребляются значительно чаще и поэтому находятся главным образом в начале частотного списка. Это обстоятельство позволяет весьма просто выделять неоднородную часть рангового распределения с помощью графика зависимости $zp_z = f(\ln z)$.

На рис.2 представлена кривая рангового распределения слов в романе Л.Н.Толстого "Война и мир" ["Частотный словарь романа Л.Н.Толстого "Война и мир", с.357-367] . Объем текста сос-

тавил x=409407 словоупотреблений, объем словаря y=19519 лексем. Кривая распределения имеет неправильную форму, поскольку изображает распределение неоднородных элементов, при этом границей неоднородной части является наинизшая точка впадины на кривой распределения с абсциссой $\ln z_0$=4,25 (z_0=70). Удалим из частотного словаря первые 70 слов, на которые приходится x_0=180844 словоупотребления текста, а оставшимся словам припишем новые ранги $z'=z-z_0$. Тогда количество оставшихся в словаре разных слов будет $y'=y-z_0$=19449, а их общая частота $x'=x-x_0$=228563 словоупотребления. Если теперь построить график зависимости $z'p_{z'}=f(\ln z')$, то получим весьма плавную одновершинную кривую, которая отличается закономерным характером возрастания и убывания (рис.3). Такая же одновершинная кривая получается в случае распределения любых однородных лингвистических единиц (терминов, дескрипторов и т.д.). Для описания такого рода статистических распределений существует объективная возможность нахождения выравнивающей кривой распределения, в то время как для описания неоднородных случайных величин (рис.2 такой возможности не существует.

Третья причина неудовлетворительных аппроксимаций статистических распределений заключается в том, что исследовались выборки недостаточного объема. Из рис.3 видно, что статистический закон распределения однородных единиц проявляет себя в полной мере лишь при том условии, если крайняя справа точка близка к горизонтальной оси. Только при этом условии можно подбирать выравнивающее непрерывное распределение.

Ордината крайней справа точки по построению равна

$$zp_z = \frac{y\bar{m}_z}{x} = \frac{y}{2x} ,$$

поскольку $z_{max}=y$, $m_{z=y}=1$, $m_{y\neq1}=0$, $\bar{m}_{z=y}=(1+0)/2=1/2$.
Объем выборки можно считать достаточным, если отношение ординаты крайней справа точки к наибольшей ординате $(zp_z)_c$ кривой распределения не превышает наперед заданного числа δ

$$\frac{zp_z}{(zp_z)_c} = \frac{y}{2x(zp_z)_c} \leq \delta , \qquad (4)$$

где $0 < \delta < 0,3$. Чем меньше δ, тем точнее могут быть оцене-

ны параметры выравнивающего распределения.

Статистическую кривую распределения $zp_z = f(\ln z)$ можно использовать для расчета необходимого объема выборки при построении достоверного словаря заданного объема. Пусть достоверная частота m_z и наибольший ранг z, т.е. объем словаря, заданы. Тогда из равенства

$$zp_z = z\,\frac{m_z}{x}$$

находим необходимый объем выборки x

$$x = \frac{z\,m_z}{zp_z}. \qquad (5)$$

Произведение zp_z, входящее в формулу (5), берется из графика зависимости $zp_z = f(\ln z)$.

Из последней формулы и рис.2 видно, что между объемом словаря z (при постоянной достоверной частоте m_z) и необходимым объемом выборки x нет линейной зависимости: объем выборки растет значительно быстрее объема достоверного словаря, поскольку произведение zp_z с ростом z уменьшается (см.правую ветвь кривой распределения на рис.2).

Последние две причины неудовлетворительных аппроксимаций ранговых распределений проливают свет на происхождение закона Ципфа. Иллюзия справедливости этого закона порождается двумя факторами: во-первых, неоднородностью лексического состава частотного словаря, которая влияет на форму начала кривой распределения, поскольку мы имеем композицию по крайней мере двух законов распределения слов (служебных и полнозначных); во-вторых, ограниченностью объема выборки, из-за чего последняя справа точка не успевает приблизиться к горизонтальной оси и может оказаться вблизи воображаемой прямой Ципфа $zp_z = K$.

Обратимся к фактам. На рис.4 (кривая I) изображено статистическое распределение по данным "Частотного словаря немецкого подъязыка хирургии" [Яблонская Н.Н., 1978] (x =200000, y =41041). Здесь объем выборки весьма ограничен, но композиция двух законов распределения налицо, о чем говорилось выше. Однако ни о какой (даже воображаемой) прямой Ципфа ($zp_z = K$) здесь не может быть речи.

На том же рис.4 изображена статистическая кривая 2 рангового распределения слов в литературном тексте Дж.Джойса "Улисс" [Zipf G.K ,1949] (x=260430, y=29899). На этом примере Ципф иллюстрировал свой закон прямолинейной зависимости между частотой и рангом (в логарифмических координатах), при этом угловой коэффициент прямой (3) по абсолютной величине равен единице. Однако в системе координат ($\ln z; zp_z$) четко видно, что статистическая кривая распределения на рис.4 не может быть аппроксимирована прямой $zp_z = K$ (т.е. законом Ципфа).

Итак, приведенных на рис.2-4 статистических распределений достаточно, чтобы сделать однозначный вывод: закон Ципфа не может быть использован для описания даже в первом приближении ранговых распределений слов, поскольку его не подтверждает ПРАКТИКА как КРИТЕРИЙ ИСТИНЫ.

Обобщенные распределения. Для описания распределений однородных случайных величин автором разработаны системы непрерывных распределений, заданные обобщенными плотностями. На эмпирическом материале было установлено, что ранговые распределения периодических изданий по числу помещенных в них статей на заданную тему могут быть описаны обобщенным распределением вида

$$p(t) = Nt^{\beta-1}\left(1-\alpha\mu t^{\beta}\right)^{\frac{1}{\mu}-1}, \qquad (6)$$

где $\alpha, \beta, \gamma, \mu$ - параметры распределения; N - нормирующий множитель; t - случайная величина, которая соответствует рангу z статистического распределения; $p(t)$ - непрерывная плотность, аппроксимирующая относительные частоты p_z.

Этим же законом (6) хорошо описываются статистические распределения: слов по длине (в словаре), фраз по количеству словоупотреблений, словосочетаний по длине, терминов по длине и др., а также ранговое распределение научных сотрудников по продуктивности.

Статистические ранговые распределения однородных лингвистических единиц (лексем, словоформ, терминов, дескрипторов и др.)

хорошо описываются обобщенным логарифмическим распределением вида

$$p(Y) = \frac{N(\ln Y)^{\gamma-1}}{Y}\left(1-\alpha u \ln^{\beta} Y\right)^{\frac{1}{u}-1}, \qquad (7)$$

где $Y=y+1$; величина y соответствует рангу z статистического распределения.

Выравнивающее распределение для примера на рис.3 имеет параметры: u =0,30; β =2,25105; γ =3,60168; αu =I/219,656; нормирующий множитель N =I/283,430. С учетом оценок параметров плотность распределения (7) можно представить в виде

$$Yp(Y) = \frac{(\ln Y)^{2,60168}}{283,430}\left[1-\frac{(\ln Y)^{2,25105}}{219,655}\right]^{2,33333}$$

$$1 < Y < 58226.$$

Метод нахождения оценок параметров непрерывных распределений изложен в работе [Нешитой В.В.,I985].

Место закона Ципфа в системе непрерывных распределений. Отметим, что обобщенное распределение (7) при $\alpha>0, \beta, \gamma, u$ =I переходит в закон Ципфа: $p(Y)=\alpha/Y\,(1<Y<e^{1/\alpha})$. Следовательно, закон Ципфа относится к семейству логарифмических распределений (7), является весьма частным его случаем и может быть использован для описания распределений однородных случайных величин (так же как и обобщенная плотность (7)). На практике же его пытаются применить для описания распределений неоднородных случайных величин (без всякого на то основания).

Далее, при $u\to 0$, β,γ =I, $\alpha>0$ из (7) имеем фактически формулу Б.Мандельброта

$$p(Y) = \frac{N}{Ye^{\alpha \ln Y}} = \frac{N}{Y^{1+\alpha}} = \frac{N}{(y+1)^{1+\alpha}}$$

а при $u\to 0$, β =2, γ =I, $\alpha>0$ - "четвертое приближение закона Ципфа" по П.М.Алексееву (логнормальный закон)

$$p(Y) = \frac{N}{Ye^{\alpha \ln^2 Y}} = \frac{N}{Y^{1+\alpha \ln Y}}$$

Если начало отсчета значений случайной величины Y поместить в центр распределения $Y_t = M[\ln Y]$, то последняя формула примет вид, более близкий к модели П.М.Алексеева [Алексеев П.М., 1978]

$$p(Y) = \frac{N'}{Y^{1-2\alpha Y_t + \alpha \ln Y}} .$$

При описании ранговых распределений однородных лингвистических единиц реализуется, как правило, четырехпараметрическая модель (7), а в некоторых, весьма редких случаях, логнормальный закон или четырехпараметрическая модель (6).

Заключение. Введение новой формы представления ранговых распределений позволило не только раскрыть причины, порождающие иллюзию справедливости закона Ципфа, но также решить некоторые важные практические задачи: дать графический метод выделения неоднородной части ранговых распределений z_0, метод оценки необходимого объема выборки для установления типа выравнивающего распределения и нахождения оценок параметров, а также метод расчета необходимого объема выборки x для построения частотного словаря заданного объема z с заданной наименьшей частотой слов m_z.

Показано, что законы Ципфа, Мандельброта и "четвертое приближение закона Ципфа" по П.М.Алексееву являются частными случаями обобщенного логарифмического распределения (7). Последняя модель хорошо описывает статистические ранговые распределения однородных лингвистических единиц.

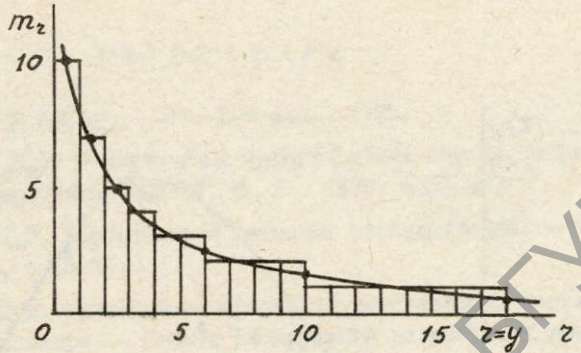

Рис.I. Выравнивание дискретного рангового распределения непрерывным

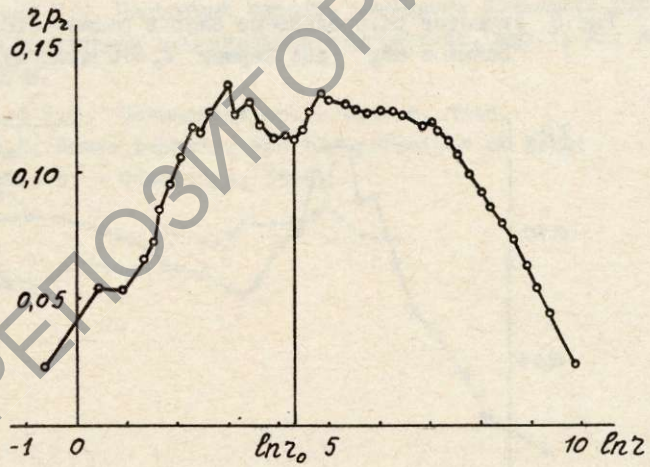

Рис.2. Ранговое распределение слов в романе Л.Н.Толстого "Война и мир"

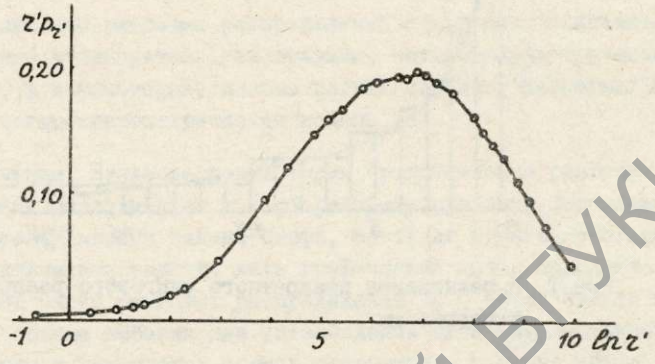

Рис.3. Ранговое распределение слов в романе Л.Н.Толстого "Война и мир" (без первых r_0=70 лексем)

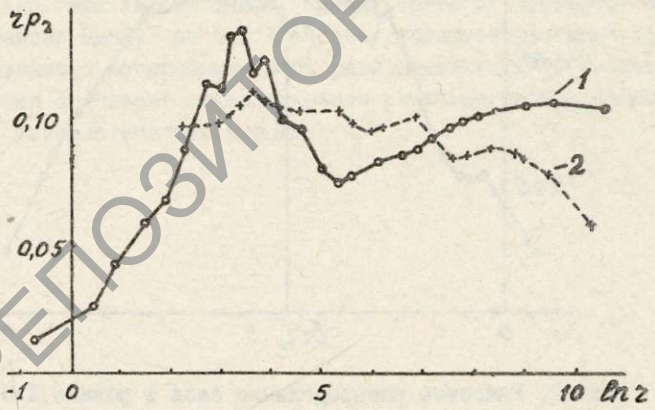

Рис.4. Ранговое распределение слов:

I - в немецком подъязыке хирургии;
2 - в тексте Дж.Джойса "Улисс".

Л И Т Е Р А Т У Р А

К.Маркс, Ф.Энгельс. Соч. 2-е изд., т.20.

Алексеев П.М. О нелинейных формулировках закона Ципфа.- Вопросы кибернетики, вып.41, М. Л., 1978, с.53-65.

Нешитой В.В. Исследование ранговых распределений. - НТИ.Сер.2, 1985, №2, с.16-20.

Нешитой В.В. Критерий однородности лексического состава частотного словаря. - Веснік Беларускага дзяржаўнага ун-та імя У.І.Леніна. Сер.4. 1984, №1, с.54-56.

Частотный словарь романа Л.Н.Толстого "Война и мир". Тула, 1978, Тульский гос.пед.инст.им.Л.Н.Толстого.

Яблонская Н.Н. Частотный словарь немецкого подъязыка хирургии.- В сб.:Структурная и прикладная лингвистика. Вып.I, Л.: изд-во ЛГУ, 1978.

Bradford S.C. Documentation. - London, 1948.

Zipf G.K. Human Behavior and the principle of least effort. - Cambridge, 1949.

ABOUT THE FORM OF REPRESENTING RANK DISTRIBUTIONS

V.V. Nešitoj

S u m m a r y

The article points out the principal demerits of the tradition of representing rank distributions in logarithmic coordinates (the Zipf straight line). A new method is put forward, with the coordinates ($\ln r$, rp_r), where r is the rank of an event and p_r - its relative frequency. It is demonstrated that in these coordinates the distribution curves of homogeneous random quantities have one peak and a regular character of increase and decrease. Generalized distributions are presented for the description of such curves.