УДК [002+801]:519.766

СИСТЕМА НЕПРЕРЫВНЫХ РАСПРЕДЕЛЕНИЙ В ИНФОРМАТИКЕ И ЛИНГВИСТИКЕ

В. В. Нешитой

При выявлении и исследовании статистических закономерностей, имеющих место в информатике и лингвистике, могут быть использованы известные из теории вероятностей и математической статистики непрерывные и дискретные распределения, а также различные функции, задаваемые уравнениями прямой, параболы, гиперболы, экспоненты, логисты и т. д. [1—3]. При этом задача подбора наилучшей математической модели для описания исследуемой зависимости довольно сложна, поскольку из опытных данных не всегда может быть предсказан вид аппроксимирующей кривой.

Решение подобных задач можно значительно облегчить, разработав обобщенные математические модели, справедливые по крайней мере для группы родственных кривых.

Цель настоящей работы заключается в построении системы непрерывных распределений и использовании ее в качестве общей модели для описания статистических закономерностей в информатике и лингвистике.

1. ПОСТРОЕНИЕ СИСТЕМЫ НЕПРЕРЫВНЫХ РАСПРЕДЕЛЕНИЙ МЕТОДОМ ОБОБЩЕНИЯ

Рассмотрим два простейших непрерывных распределения: равномерное и треугольное, плотность которого убывает.

В первом случае плотность распределения вероятностей $p(t)$ случайной величины T и функция распределения $F(t)=S^t p(x)dx$ задаются формулами

$$p(t)=\alpha, \tag{1}$$

$$F(t)=\alpha t=1-(1-\alpha t), \quad 0<t<1/\alpha \tag{2}$$

во втором случае;

$$p(t)=\alpha\left(1-\frac{\alpha}{2}t\right), \tag{3}$$

$$F(t)=1-\left(1-\frac{\alpha}{2}t\right)^2, \quad 0<t<2/\alpha. \tag{4}$$

Из приведенных формул видно, что для обоих случаев функция распределения может быть представлена в общем виде

$$F(t)=1-(1-\alpha u t)^{\frac{1}{u}}. \tag{5}$$

Рассмотрим далее совместно с равномерным распределением треугольное, плотность которого возрастает

$$p(t)=2\alpha t. \tag{6}$$

Функция распределения в данном случае равна

$$F(t)=\alpha t^2=1-(1-\alpha t^2), \quad 0<t<\sqrt{1/\alpha}. \tag{7}$$

Обобщая формулы (2)—для равномерного распре-

деления и (7)—для треугольного возрастающего распределения, получим

$$F(t)=\alpha t^{\beta}=1-(1-\alpha t^{\beta}). \qquad (8)$$

Теперь осталось обобщить функции распределения (5) и (8). Это можно сделать единственным способом

$$F(t)=1-(1-\alpha u t^{\beta})^{\frac{1}{u}}, \qquad (9)$$

откуда

$$p(t)=\alpha\beta t^{\beta-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1}. \qquad (10)$$

Функция $p(t)$ при $u>0$ определена на интервале $0<t<(1/au)^{1/\beta}$, при $u<0$, а также $u\to 0$—на интервале $0<t<\infty$.

Семейство кривых, задаваемое формулами (9) и (10), можно еще более расширить, если в уравнение (10) ввести дополнительный параметр.

Структура формулы (10) позволяет это сделать следующим образом:

$$p(t)=Nt^{\gamma-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1}. \qquad (11)$$

Здесь N—нормирующий множитель, α—масштабный параметр, β, γ, u—параметры формы.

На основе обобщенной плотности (11) можно получать другие плотности как функции случайного аргумента T. Пусть, например, $x=\ln T$. Тогда $T=e^{X}$, $\frac{dt}{dx}=e^{x}$, $p(x)=p(t)\left|\frac{dt}{dx}\right|$, что с учетом (11) дает

$$p(x)=Ne^{\gamma x}(1-\alpha u e^{\beta x})^{\frac{1}{u}-1}. \qquad (12)$$

Плотность (12) можно получить и другим способом. Преобразуем выражение (11) к виду

$$tp(t)=Ne^{\gamma\ln t}(1-\alpha u e^{\beta\ln t})^{\frac{1}{u}-1}. \qquad (13)$$

Обозначая теперь $\ln t=x$, $tp(t)=p(x)$, вместо (13) будем иметь (12).

При $y=e^{T}$ найдем $p(y)=N\frac{1}{y}(\ln y)^{\gamma-1}\times$ $\times[1-\alpha u(\ln y)^{\beta}]^{\frac{1}{u}-1}$ и т. д.

2. КЛАССИФИКАЦИЯ КРИВЫХ

В зависимости от значений параметров u, α кривые, заданные общим уравнением (11), можно разделить на типы (рис. 1).

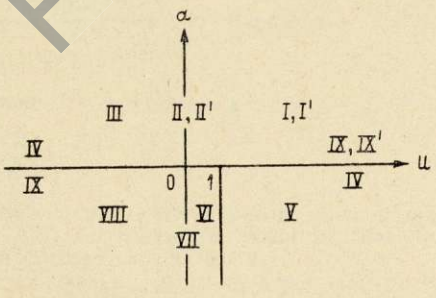

Рис. 1. Классификация кривых

Типы I—IX получаются при $\beta, \gamma>0$, типы I′, II′, IX′—при $\beta, \gamma<0$.

Для кривых типа IV $u\to\pm\infty$, $\alpha\to\mp 0$ ($\alpha u=-1/h^{\beta}$, $h>0$), типа IX — $u\to\pm\infty$, $\alpha\to\pm 0$ ($\alpha u=1/h^{\beta}$, $h>0$) для типов II и II′ $u\to 0$, $\alpha>0$, для типа VII — $u\to 0$, $\alpha<0$. Значения параметров α, u для кривых остальных типов ясны из рисунка.

Кривые типов I—V, I′, II′ могут быть кривыми распределения, причем, среди кривых типов I—V, заданных плотностью 11, можно выделить две группы симметричных распределений (обозначим их Ic—Vc).

В первую группу входят распределения типа Ic с параметрами $\beta=1$, $u=1/\gamma$, во вторую группу — распределения Ic—Vc с параметрами $\beta=2$, $\gamma=1$.

При $u\leqslant 1$, $\alpha<0$ функция (11) не может быть плотностью распределения вероятностей, так как интеграл $\int_0^{\infty} p(t)\,dt$ в данном случае расходится.

Среди распределений III—V, заданных плотностью (12), также имеется группа симметричных распределений. Параметры последних удовлетворяют условию $\frac{\gamma}{\beta}=\frac{1}{2}\left(1-\frac{1}{u}\right)$.

В зависимости от значений параметров β, γ все распределения, заданные общим уравнением (11), можно разделить на две большие группы *А* и *Б*.

Для распределений группы *А* параметр β равен параметру γ, а интегралы от плотностей распределения вероятностей всегда выражаются через элементарные функции. Функция распределения в этом случае задается формулой (9). Из последней формулы, в частности, следует, что распределений группы *А*, относящихся к типам IV и V, не существует, так как функция распределения определена при $\alpha>0$, $-\infty<u<\infty$.

Для распределений группы *Б* $\gamma\neq\beta$. В этом случае интегралы от плотностей распределения вероятностей, как правило, не выражаются через элементарные функции. Функция распределения задается интегралом

$$F(t)=N\int_0^{t} x^{\gamma-1}(1-\alpha u x^{\beta})^{\frac{1}{u}-1}dx. \qquad (14)$$

Здесь подынтегральное выражение есть не что иное, как дифференциальный бином. Из математического анализа известно, что интегралы вида (14), где β, γ, u—рациональные числа, выражаются через элементарные функции лишь в том случае, если одно из чисел $\frac{1}{u}$, $\frac{\gamma}{\beta}$, $\frac{1}{u}+\frac{\gamma}{\beta}$—целое (положительное, отрицательное или нуль) [4].

3. ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ ГРУППЫ А

К группе *А* нами отнесены распределения с параметрами $\gamma=\beta$. Следовательно, эти распределения в общем случае содержат три параметра: один масштабный параметр (α) и два параметра формы (β, u). Оценки параметров распределений группы *А* могут быть найдены по функции распределения. Для этого необходимо преобразовать ее к такому виду, чтобы получилось уравнение прямой. Тогда оценки параметров распределения можно найти по параметрам прямой.

Рассмотрим существующие типы распределений, относящиеся к группе *А*.

Тип I. Параметры: $\alpha>0$, $\beta>0$, $u>0$.

Функция распределения и плотность распределения вероятностей задаются формулами

$$F(t) = 1 - (1 - \alpha u t^{\beta})^{\frac{1}{u}}, \tag{15}$$

$$p(t) = \alpha\beta t^{\beta-1} (1 - \alpha u t^{\beta})^{\frac{1}{u}-1}, \tag{16}$$

$$0 < t < (1/\alpha u)^{1/\beta}.$$

Выражение (15) можно преобразовать так

$$\ln \{1 - [1 - F(t)]^u\} = \ln \alpha u + \beta \ln t. \tag{17}$$

Это — уравнение прямой с начальной ординатой $\ln \alpha u$ и угловым коэффициентом β. К сожалению, неизвестный параметр u содержится не только в правой части данного уравнения, но и в левой. Следовательно, для нахождения оценок параметров α, β сначала необходимо как-нибудь найти оценку параметра u (этот параметр является также критерием, с помощью которого устанавливается тип кривой распределения группы A). Оценку параметра u можно найти путем подбора. Принимая различные значения u (например, 0, 1, 0, 2 и т. д.) и строя на основе опытных данных графики зависимости $\ln\{1-[1-F(t)]^u\}$ от $\ln t$, при определенном значении u (которое принимается в качестве его оценки) мы можем получить график прямой, из которого легко найти значения параметров α, β.

Если прямая не получается, то это означает, что выравнивающее распределение не относится к распределению типа I группы A. В этом случае в качестве выравнивающего можно попытаться использовать другое распределение, например, типа II группы A.

Тип II (распределение Вейбулла). Параметры: $\alpha > 0$, $\beta > 0$, $u \to 0$. Выражения для $F(t)$ и $p(t)$ имеют вид

$$F(t) = 1 - \frac{1}{e^{\alpha t^{\beta}}}, \tag{18}$$

$$p(t) = \frac{\alpha\beta t^{\beta-1}}{e^{\alpha t^{\beta}}}, \quad 0 < t < \infty. \tag{19}$$

Выражение (18) приводится к виду

$$\ln \ln \frac{1}{1 - F(t)} = \ln \alpha + \beta \ln t. \tag{20}$$

Построив по опытным данным график зависимости $\ln \ln \frac{1}{1-F(t)}$ от $\ln t$ и убедившись, что точки рассеиваются около прямой (20), находим оценки параметров α и β распределения Вейбулла.

Тип III. Параметры: $\alpha > 0$, $\beta > 0$, $u < 0$. Выражения для $F(t)$ и $p(t)$ имеют вид

$$F(t) = 1 - \frac{1}{(1 + \alpha |u| t^{\beta})^{\frac{1}{|u|}}}, \tag{21}$$

$$p(t) = \frac{\alpha\beta t^{\beta-1}}{(1 + \alpha |u| t^{\beta})^{\frac{1}{|u|}+1}}, \quad 0 < t < \infty. \tag{22}$$

Выражение (21) приводится к виду

$$\ln \left\{ \left[\frac{1}{1 - F(t)} \right]^{|u|} - 1 \right\} = \ln \alpha |u| + \beta \ln t. \tag{23}$$

Нахождение оценок параметров α, β, u с помощью уравнения (23) производится таким же способом, как и в случае распределений типа I.

Рассмотрим далее распределения типов I′—III′ группы A.

Тип I′ получается из (11) при β, $\gamma < 0$ ($\gamma = \beta$), $\alpha > 0$, $u > 0$. Выражения для $F(t)$ и $p(t)$ имеют вид

$$F(t) = \left(1 - \frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}}, \tag{24}$$

$$p(t) = \frac{\alpha\beta}{t^{\beta+1}} \left(1 - \frac{\alpha u}{t^{\beta}}\right)^{\frac{1}{u}-1}, \tag{24}$$

$$(\alpha u)^{1/\beta} < t < \infty, \quad \alpha, \ \beta, \ u > 0.$$

Из (24) находим

$$\ln \{1 - [F(t)]^u\} = \ln \alpha u - \beta \ln t. \tag{26}$$

Тип II′ получается из типа I′ при $u \to 0$. Выражения для $F(t)$ и $p(t)$ имеют вид

$$F(t) = \frac{1}{e^{\alpha/t^{\beta}}}, \tag{27}$$

$$p(t) = \frac{\alpha\beta}{t^{\beta+1} e^{\alpha/t^{\beta}}}, \quad 0 < t < \infty. \tag{28}$$

Из (27) находим

$$\ln \ln \frac{1}{F(t)} = \ln \alpha - \beta \ln t. \tag{29}$$

Тип III′ получается из типа I′ при $u < 0$. Выражения для $F(t)$ и $p(t)$ имеют вид

$$F(t) = \frac{1}{\left(1 + \frac{\alpha |u|}{t^{\beta}}\right)^{\frac{1}{|u|}}}, \tag{30}$$

$$p(t) = \frac{\alpha\beta}{t^{\beta+1} \left(1 + \frac{\alpha |u|}{t^{\beta}}\right)^{\frac{1}{|u|}+1}}, \quad 0 < t < \infty. \tag{31}$$

Из (30) находим

$$\ln \left\{ \left[\frac{1}{F(t)} \right]^{|u|} - 1 \right\} = \ln \alpha |u| - \beta \ln t. \tag{32}$$

С помощью полученных уравнений на основе опытных данных могут быть найдены оценки параметров выравнивающих распределений типов I—III, I—III, группы A.

Если из некоторых соображений предполагается, что выравнивающее распределение может относиться к одному из рассмотренных шести типов, то в этом случае целесообразно, в первую очередь, проверить, не относится ли оно к типу II или типу II′, так как эти распределения имеют только два параметра, и для нахождения их оценок достаточно построить соответствующие графики. Если после построения этих графиков окажется, что опытные точки не ложатся на прямые, то необходимо опробовать остальные типы распределений.

Если в качестве выравнивающих принимаются распределения группы B, для которых $\gamma \neq \beta$, то в этом случае оценки параметров могут быть найдены другими, более универсальными и сложными методами (например, методом наибольшего правдоподобия).

Полученные формулы могут быть использованы для выравнивания не только эмпирических распределений разных типов, но и других кривых, не обладающих свойствами кривых распределения.

4. РАСПРЕДЕЛЕНИЕ ВЕЙБУЛЛА — ОСНОВНОЙ СТАТИСТИЧЕСКИЙ ЗАКОН НАУКОВЕДЕНИЯ И ИНФОРМАТИКИ

При статистическом описании документальных потоков, а также описании распределения слов по частоте их употребления в тексте используются так называемые ранговые распределения (ранг — это порядковый номер слова или какого-нибудь другого события в списке, где все события упорядочены по возрастанию относительных частот). Для описания таких распределений построенная система предоставляет широкие возможности. Исследования показали, что кривые распределения типов I—V, заданные плотностью (11), при $0<\gamma\leqslant 1$ и определенных значениях параметров β, u, а также кривые типа I′ при $u\geqslant 1$ являются невозрастающими.

В литературе по информатике и математической лингвистике для описания эмпирических ранговых распределений, наряду с другими, широко используется распределение Вейбулла. По нашей классификации оно относится к типу II группы *A*. В прикладной лингвистике распределение Вейбулла впервые применил Г. Г. Белоногов для описания распределения слов в русской письменной речи [5].

Проверка показала, что законом Вейбулла хорошо описываются многие эмпирические ранговые распределения, в том числе: распределение разных слов (по крайней мере полнозначных) по частоте употребления их в тексте, распределение периодических и продолжающихся изданий по числу помещенных в них статей данного профиля, распределение научных сотрудников по продуктивности, распределение публикаций по числу ссылок на них и т. д. [6, 7].

На рис. 2 приведены графики зависимости (20),

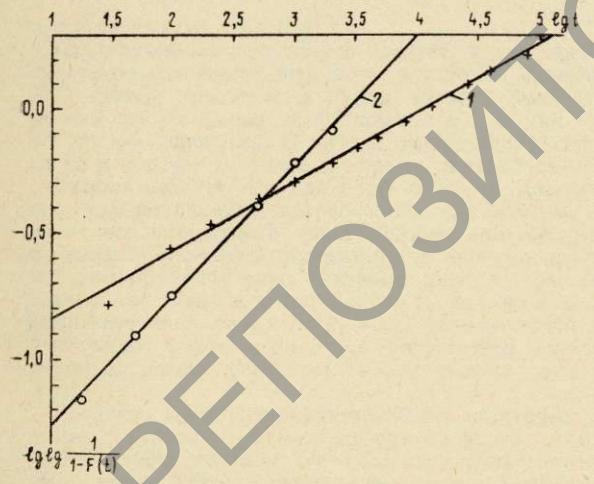

Рис. 2. Распределение разных событий по закону Вейбулла

справедливой в случае распределения Вейбулла. Отдельные точки построены по опытным значениям накопленной относительной частоты событий с рангами $\leqslant t$, т. е. по опытной функции распределения $F^*(t)$ (сплошные прямые — теоретические). Прямая 1 отражает распределение разных словоформ в смешанных текстах на немецком языке [8]. Объем выборки $X=10910777$ словоупотреблений, число разных слов (объем словаря) $Y=258173$ словоформы. Параметры прямой 1 оказались равными: $\alpha=0{,}178$, $\beta=0{,}274$. Прямая 2 отражает рассеяние 187911 публикаций по химии и химической технологии [9] по 10850 журналам. Параметры прямой 2 равны: $\alpha=0{,}036$, $\beta=0{,}53$ [10].

5. ЗАКОН РАССЕЯНИЯ ИНФОРМАЦИИ БРЕДФОРДА КАК СЛЕДСТВИЕ СВОЙСТВ РАСПРЕДЕЛЕНИЯ ВЕЙБУЛЛА

Если построить график функции распределения (18) в полулогарифмических координатах $F(t)=f(\ln t)$, то на нем можно отметить три характерные точки: *C* — точка перегиба, *A* и *B* — точки, в которых скорость изменения тангенса угла наклона касательной к данной кривой принимает экстремальные значения [10]. На графике кривой распределения $tp(t)=\varphi(\ln t)$ (поскольку $dF(t)/d\ln t=tdF(t)/dt=tp(t)$) точка *C* обозначает моду, *A* и *B* — точки перегиба. Найдем координаты этих точек.

Обозначая $\ln t=x$, $tp(t)=p(x)$, приведем выражения (18) и (19) к виду

$$F(x)=1-\frac{1}{e^{\alpha e^{\beta x}}}, \tag{33}$$

$$p(x)=\frac{\alpha\beta e^{\beta x}}{e^{\alpha e^{\beta x}}}, \quad -\infty<x<\infty. \tag{34}$$

Дифференцируя уравнение (34) по x и приравнивая первую производную нулю, будем иметь

$$x_C=\frac{1}{\beta}\ln\frac{1}{\alpha}. \tag{35}$$

Подставив значение x_C в (33), получим

$$F(x_C)=1-\frac{1}{e}\approx 0{,}6322. \tag{36}$$

Решая далее уравнение $p''(x)=0$ относительно переменной x, найдем

$$x_{A,B}=\frac{1}{\beta}\ln\frac{3\mp\sqrt{5}}{2\alpha}. \tag{37}$$

Тогда

$$F(x_{A,B})=1-\frac{1}{e^{\frac{3\mp\sqrt{5}}{2}}}. \tag{38}$$

Учитывая, что $x=\ln t$, на основании формул (35)—(38) можем записать

$$\left.\begin{aligned} &t_A=\left(\frac{3-\sqrt{5}}{2\alpha}\right)^{1/\beta}\approx\left(\frac{0{,}382}{\alpha}\right)^{1/\beta},\\ &F(t_A)\approx 0{,}3175,\\ &t_C=\left(\frac{1}{\alpha}\right)^{1/\beta},\quad F(t_C)\approx 0{,}6321,\\ &t_B=\left(\frac{3+\sqrt{5}}{2\alpha}\right)^{1/\beta}\approx\left(\frac{2{,}618}{\alpha}\right)^{1/\beta},\\ &F(t_B)\approx 0{,}9271. \end{aligned}\right\} \tag{39}$$

Из (39) следует, что между величинами t_A, t_C, t_B имеется соотношение

$$\frac{t_B}{t_C}=\frac{t_C}{t_A}=N,$$

или

$$t_A:t_C:t_B=1:N:N^2, \tag{40}$$

где

$$N\left(\frac{3+\sqrt{5}}{2}\right)^{1/\beta}\approx 2{,}618^{1/\beta}.$$

Выражение (40) по форме совпадает с законом рассеяния информации Бредфорда. Суть его заключается в следующем: «Если научные журналы расположить в порядке уменьшения числа помещенных в них статей по какому-либо заданному предмету, то в полученном списке можно выделить ядро журналов, посвященных непосредственно этому предмету, и несколько групп, или зон, каждая из которых содержит столько же статей, что и ядро. Тогда числа журналов в ядре и последующих зонах будут относиться как $1 : N : N^2$» [9].

Однако из этой формулировки неясно, как определяется число журналов, образующих ядро, какая доля статей содержится в нем, сколько может быть зон рассеяния, чему равна величина N.

Анализ свойств распределения Вейбулла дает возможность ответить на эти вопросы.

Выше отмечалось, что распределение периодических изданий по числу помещенных в них статей по данному предмету также подчиняется закону Вейбулла (см. рис. 2). Применительно к периодическим изданиям точки A, C, B являются границами так называемых зон рассеяния публикаций и делят все периодические издания, содержащие статьи по данному предмету, на четыре части: ядро, зону I, зону II и зону III.

Количество журналов, входящих в ядро, определяется равенством: $t_я = t_A$; количество журналов в первой зоне равно $t_I = t_C - t_A$, во II зоне $t_{II} = t_B - t_C$. Остальные журналы относятся к III зоне. При этом число статей, содержащихся в ядре и в каждой из первых двух зон, примерно одинаково (в ядро входит около 32% от всех статей по данному предмету, в I зону — 31%, во II зону — 30%), в то время как на III зону приходится лишь 7% статей.

Между числом наименований журналов в ядре и последующих зонах имеется соотношение

$$t_я : t_I : t_{II} = 1 : (N-1) : (N-1)\,N,$$

где по-прежнему $N \approx 2{,}618^{1/\beta}$.

Из полученных результатов следует, что в ядро входит такое количество журналов, в которых содержится 32% статей, при этом журналы должны быть упорядочены по убыванию числа помещенных в них статей по данному предмету. В ядре и первой зоне содержится 63% статей, а в ядре и первых двух зонах — 93%.

Отсюда следует, что для более полного удовлетворения информационных потребностей специалистов справочно-информационный фонд должен комплектоваться по крайней мере теми журналами по данному профилю, которые образуют ядро и первые две зоны рассеяния. Количество таких журналов равно t_B, при этом полнота комплектования фонда $F(t_B) \approx 0{,}93$ (под полнотой комплектования фонда понимается вероятность удовлетворения запросов потребителей информации этим фондом).

Величина t_B может характеризовать некоторый оптимальный объем справочно-информационного фонда с точки зрения полноты его комплектования статьями. Этот вывод обосновывается тем, что для увеличения полноты комплектования, например, на 5% выше оптимального, необходимо увеличить объем фонда в два раза (при $\alpha = 0{,}1$, $\beta = 0{,}5$), в то время как для уменьшения полноты комплектования по статьям на 5% ниже оптимального достаточно уменьшить его объем лишь в 1,5 раза в (основном за счет наименее запрашиваемых журналов).

Оптимальный объем фонда t_B, в который входят наиболее часто запрашиваемые документы, можно считать активной его частью. Таким же образом можно выделить активную часть частотного словаря. Она покрывает около 93% текстов по данной тематике, или, иначе, имеет полноту 0,93. Активный словарь содержит t_B наиболее часто употребляемых слов. Слова с рангами $t > t_B$ образуют пассивную часть словаря.

Распределение Вейбулла позволяет находить величину t по заданному значению $F(t)$. Решая уравнение (15) относительно t, получим

$$t = \left(\frac{1}{\alpha} \ln \frac{1}{1 - F(t)}\right)^{1/\beta}.$$

Свойства функции распределения Вейбулла можно использовать при построении словаря ключевых слов. Если построить на полных текстах частотный словарь разных слов по некоторой отрасли, то в нем можно выделить следующие зоны: 1 — зона наиболее частых слов с рангами от 1 до t_A, куда входят главным образом служебные слова; 2 — зона общеупотребительных слов с рангами $t_A < t < t_C$; 3 — зона отраслевой лексики с рангами $t_C < t < t_B$ (среднечастотные слова); 4 — зона межотраслевой лексики (редкоупотребляемые слова). В эту зону входят слова с рангами $t > t_B$.

Плотность распределения ключевых слов в третьей зоне ($t_C < t < t_B$) должна быть наибольшей, так как здесь сосредоточена главным образом отраслевая лексика.

6. КРИВАЯ РОСТА РАЗНЫХ СЛОВ

При известном законе распределения разных слов в тексте, заданном например плотностью (19), по формуле:

$$y = \int_0^\infty \left(1 - \frac{1}{e^{xp(t)}}\right) dt \qquad (41)$$

может быть рассчитана зависимость между объемом выборки x и средним числом разных слов y. Однако пользоваться формулой (41) неудобно и поэтому для описания зависимости $y = f(x)$ желательно иметь простую приближенную формулу. Ее можно получить из общей формулы (11), которую перепишем в виде

$$y = \frac{N x^{\gamma - 1}}{(1 - \alpha u x^{\beta})^{1 - \frac{1}{u}}}. \qquad (42)$$

Исходя из условия, что в начале координат кривая $y = f(x)$ удовлетворяет условию $\frac{dy}{dx} = 1$, находим, что уравнение (42) удовлетворяет этому условию при $\gamma = 2$, $N = 1$. Следовательно, кривая роста разных слов в выборке приближенно может быть описана уравнением

$$y = \frac{x}{(1 - \alpha u x^{\beta})^{1 - \frac{1}{u}}} \qquad (43)$$

с тремя параметрами: α, β, u. Дальнейшая проверка показывает, что весьма хорошее приближение к опытным данным (при $10^3 \div 10^4 < x < 10^7 \div 10^8$) последняя формула дает при $u = -1$, т. е. формулу (43) можно еще более упростить:

$$y = \frac{x}{(1 + \alpha x^{\beta})^2}. \qquad (43')$$

Преобразуем (43′) к виду

$$\ln\left(\sqrt{\frac{x}{y}} - 1\right) = \ln \alpha + \beta \ln x. \qquad (44)$$

Построив по опытным значения x, y график зависимости $\ln\left(\sqrt{\frac{x}{y}} - 1\right)$ от $\ln x$, по полученной прямой найдем оценки параметров α, β.

Параметры выборки α, β могут быть оценены также по количеству одноразовых слов $y_{m=1}$ (при изве-

стных значениях x, y). Для решения этой задачи воспользуемся формулой В. М. Калинина [11].

$$y_m = (-1)^{m+1} \frac{x^m}{m!} \frac{d^m y}{dx^m}, \qquad (45)$$

которая связывает количество разных слов с частотой m (y_m) с m-ой производной от кривой роста разных слов. Из формулы (45) при $m=1$ имеем

$$\frac{dy}{dx} = \frac{y_{m=1}}{x}. \qquad (46)$$

С другой стороны, дифференцируя выражение (43′) по x, получим

$$\frac{dy}{dx} = \frac{1+\alpha(1-2\beta)\,x^{\beta}}{(1+\alpha x^{\beta})^2} = \frac{1+(1-2\beta)\left(\sqrt{\frac{x}{y}}-1\right)}{\frac{x}{y}\sqrt{\frac{x}{y}}}. \qquad (47)$$

Приравнивая правые части в выражениях (46) и и (47), найдем

$$\beta = \frac{\sqrt{\frac{x}{y}}\left(1-\frac{y_{m=1}}{y}\right)}{2\left(\sqrt{\frac{x}{y}}-1\right)}, \qquad (48)$$

причем, как показывают исследования, $0{,}2<\beta<0{,}5$. Тогда параметр α (при известном значении β) определится по формуле

$$\alpha = \frac{\sqrt{\frac{x}{y}}-1}{x^{\beta}}, \qquad (49)$$

которая следует из (43′).

Параметр выборки β, входящий в формулу (43′), связан с параметром формы β распределения Вейбулла (обозначим его через B) следующей эмпирической зависимостью

$$2\beta = 1-\frac{1}{\sqrt[3]{1+2300B^{3,75}}}, \qquad (50)$$

откуда

$$B = 0{,}127\left[\left(\frac{1}{1-2\beta}\right)^3-1\right]^{0,267}. \qquad (51)$$

В качестве модели зависимости (50) была использована функция распределения (21).

Формула (43′) хорошо описывает не только кривую роста разных слов в выборке, но и в связном лексически однородном тексте, при этом между параметрами выборки (α, β) и параметрами связного текста (обозначим их α_T, β_T) существует приближенная зависимость

$$\alpha_T \approx 2\alpha, \quad \beta_T \approx \beta - \frac{\ln 2}{\ln x}.$$

Отметим, что формула (43′) при описании кривой роста разных слов в случайной выборке может использоваться только для интерполяции данной кривой, поскольку параметр β зависит от объема выборки x.

При описании кривой роста разных слов в связном лексически однородном тексте формула (43′) может использоваться также и для экстраполяции, так как параметры α_T, β_T в данном случае не зависят от x.

Формула (43′) позволяет вычислять полноту словаря при любом заданном его объеме y. Здесь имеется в виду не частотный словарь, а словарь разных слов, употребленных в тексте объемом x словоупотреблений. Полнота такого словаря равна

$$F(y) = 1-\frac{dy}{dx},$$

где $\frac{du}{dx}$ — вероятность появления нового слова (вычисляется по формуле (47)).

В случае простейших законов распределения вероятностей разных событий формула (41) позволяет находить заданные в явном виде уравнения для описания кривых роста разных событий.

Так, в случае равномерной плотности ($p(t)=\alpha$, $0<t<1/\alpha$) формула (41) дает

$$y = \frac{1}{\alpha}\left(1-\frac{1}{e^{\alpha x}}\right).$$

При показательном законе распределения ($p(t)=\alpha/e^{\alpha t}$, $0<t<\infty$) имеем асимптотическую формулу

$$y \approx \frac{1}{\alpha}(\ln \alpha x + C), \qquad (52)$$

где $C=0{,}5772\ldots$ — постоянная Эйлера.

Показательным законом описывается, например, распределение разных запросов по частоте их использования абонентами. Следовательно, формула (52) может описывать кривую роста разных запросов, при этом x обозначает количество всех запросов с учетом их повторяемости (т. е. количество абоненто-запросов), y — количество разных запросов.

Используя обобщенную плотность (11) в качестве общей модели, можно успешно решать многие задачи по выявлению и описанию статистических закономерностей в информатике и лингвистике.

В рамках настоящей статьи были показаны лишь некоторые примеры решения подобных задач.

ЛИТЕРАТУРА

1. Венецкий И. Г., Венецкая В. И. Основные математико-статистические понятия и формулы в экономическом анализе. — (М.: Статистика, 1979 разделы 5, 6).
2. Добров Г. М. Прогнозирование науки и техники. — М.: Наука, 1977.
3. Четыркин Е. М. Статистические методы прогнозирования. — М., 1975.
4. Пискунов Н. С. Дифференциальное и интегральное исчисления. — т. 1. М., 1968.
5. Белоногов Г. Г. О некоторых статистических закономерностях в русской письменной речи. — Вопросы языкознания, 1962, № 1
6. Белоногов Г. Г., Богатырев В. И. Автоматизированные информационные системы. — М.: Советское радио, 1973.
7. Петров Ю. Г. Исследование и разработка автоматизированных ИПС, на основе заглавий научно-технических публикаций: Дис.... канд. техн. наук. — М., 1972.
8. Meier H. Deutsche Sprachstatistik. — Hildesheim. 1964.
9. Михайлов А. И., Черный А. И., Гиляревский Р. С. Основы информатики. — М.: Наука, 1968.
10. Горькова В. И., Петренко Б. В., Нешитой В. В. Расчет полноты комплектования справочно-информационного фонда (СИФ). — Минск, 1974 — Информ. листок БелНИИНТИ; № 176.
11. Калинин В. М. Некоторые статистические законы математической лингвистики. — В кн.: Проблемы кибернетики, вып. 11. М., 1964.

Статья поступила в редакцию 20 мая 1983 г.