УДК 801.73:519.2

РАНЖИРОВАНИЕ СЛОВ ПО СТЕПЕНИ СЕМАНТИЧЕСКОЙ НАГРУЗКИ

В. В. Нешитой

Разобьем связный текст достаточно большой длины на равные отрезки и соберем статистику употребления отдельных слов в каждом таком отрезке. Выберем любое из часто употребляемых слов и проследим, сколько раз (в каком количестве отрезков) оно не встретилось ни разу, встретилось один раз, два раза и т. д., т. е. рассмотрим его *частотный спектр*. Если бы выбранное нами слово употреблялось в тексте независимо и случайно, то частота его употребления в отрезках равной длины подчинялась бы закону Пуассона [1, с. 111—119].

Известно, что в процессе Пуассона вероятность появления некоторого слова от отрезка к отрезку не изменяется. Отсюда следует вывод, что частотные спектры слов, несущих определенную семантическую нагрузку, не должны подчиняться закону Пуассона, так как вероятности употребления таких слов не постоянны и изменяются от отрезка к отрезку в зависимости от ситуации.

В настоящей статье ставится задача — по заданному частотному спектру слова научиться находить дискретный (спектровый) закон распределения и на его основе оценивать степень неравномерности (неслучайности) употребления данного слова в генеральной совокупности. Решение этой задачи позволит ранжировать разные слова по степени неравномерности употребления, или, другими словами, по удаленности от распределения Пуассона и, следовательно, в некоторой мере — по степени семантической нагрузки.

Для решения поставленной задачи требуется рассмотреть случайный процесс появления разных событий с ростом числа испытаний, найти функциональную зависимость между числом произведенных испытаний и математическим ожиданием числа наступивших при этом разных событий, т. е. найти выражение для кривой роста разных событий, исследовать свойства данной кривой и на основе полученных результатов построить систему дискретных (спектровых) распределений, включающую, как частный случай, закон Пуассона.

Хотя в настоящей статье ставится частная задача, при ее решении получаются побочные результаты, которые представляют для специалиста самостоятельный интерес и могут быть использованы в информационной теории и практике.

1. КРИВАЯ РОСТА ЧИСЛА РАЗНЫХ СОБЫТИЙ И НЕКОТОРЫЕ ЕЕ СВОЙСТВА

Рассмотрим n разных событий $A_1, A_2, \ldots, A_n$, составляющих полную группу. Вероятности их равны, соответственно, $p_1, p_2, \ldots, p_n$, причем $\sum_{k=1}^{n} p_k = 1$. Пусть далее производятся независимые испытания, в каждом из которых может наступить любое из n событий. Введем понятие *нового* события $A_k^{\text{н}}$. *Под новым будем понимать любое из n разных событий*, составляющих полную группу, *при первом его появлении от начала испытаний*. Найдем математическое ожидание числа разных (новых) событий, наступающих при X испытаниях. Оно задается формулой [2]:

$$M[Y(X)] = \sum_{k=1}^{n} [1-(1-p_k)^X], \tag{1}$$

где $(1-p_k)^X$ — вероятность непоявления k-го события при X испытаниях, $1-(1-p_k)^X$ — вероятность появления k-го события хотя бы один раз при X испытаниях.

Математическое ожидание числа разных событий, наступивших ровно m раз при X испытаниях (частотный спектр), определится по формуле

$$M[Y_m(X)] = \sum_{k=1}^{n} C_X^m p_k^m (1-p_k)^{X-m}. \tag{2}$$

Важной величиной, характеризующей скорость роста числа новых событий, является вероятность появления какого-нибудь нового события $A^{\text{н}} = \sum_{k=1}^{n} A_k^{\text{н}}$ в $(X+1)$-м испытании (которая равна математическому ожиданию числа новых событий, наступающих при одном $(X+1)$-м испытании):

$$P(A^{\text{н}}, X+1) = \sum_{k=1}^{n} (1-p_k)^X p_k = M[Y(X+1)] - M[Y(X)]. \tag{3}$$

Если вероятности отдельных событий малы, а число испытаний X достаточно большое, то вероятности p_k можно аппроксимировать непрерывной плотностью $p(t)$, удовлетворяющей условию $\int_{k-1}^{k} p(t)\,dt = p_k$, а формулы (1)—(3)—представить в виде

$$y = \int_0^n \left(1 - \frac{1}{e^{xp(t)}}\right) dt, \tag{4}$$

$$y_m = C_x^m \int_0^n \frac{[p(t)]^m}{e^{(x-m)p(t)}}\,dt, \tag{5}$$

$$P(A^{\text{н}}, x) = \int_0^n \frac{p(t)}{e^{xp(t)}}\,dt = \frac{dy}{dx}. \tag{6}$$

Как видно из формул (4) и (6), вероятность появления нового события при x произведенных испытаниях равна значению первой производной в точке (x, y) кривой роста новых событий $y = f(x)$, где y — среднее число новых событий, наступающих при x испытаниях.

Отметим, что при больших x и малых m между зависимостями (4) и (5) существует связь, установленная В. М. Калининым [2]:

$$y_m \approx (-1)^{m+1} \frac{x^m}{m!} \frac{d^m y}{dx^m}. \tag{7}$$

Формула (7) позволяет найти частотный спектр, т. е. вычислить количество событий с частотой $0, 1, \ldots, m$ по кривой роста числа новых событий $y=f(x)$.

Установим связь между кривой $y=f(x)$ и законами распределения вероятностей новых событий.

Пусть производится несколько серий испытаний. Тогда порядок появления новых событий в каждой серии будет разный. Это значит, что j-м по порядку новым событием в каждой серии может наступить любое из n событий, составляющих полную группу. Обозначим через $\bar{p}_j$ среднюю вероятность новых событий, которые наступают j-ми от начала испытаний во всех сериях, а через $\bar{p}(y)$ — среднюю плотность распределения вероятностей новых событий, аппроксимирующую вероятности $\bar{p}_j$. Обозначим далее через $\bar{p}_i$ среднюю вероятность новых событий, которые наступают во всех сериях в i-х испытаниях, а через $\bar{p}(x)$ — среднюю плотность распределения вероятностей новых событий, аппроксимирующую вероятности $\bar{p}_i$.

Понятия средних плотностей, $\bar{p}(y)$, $\bar{p}(x)$, применимы также к одной серии независимых испытаний, поскольку в обоих случаях они являются выравнивающими распределениями статистических законов распределения вероятностей новых событий.

Так как в ходе испытаний новые события должны наступать, в среднем, в порядке убывания их вероятностей, то средние плотности $\bar{p}(y)$, $\bar{p}(x)$ являются невозрастающими. Тогда вероятность появления нового (еще не наступившего) события в точке (x, y) кривой $y=f(x)$ можно представить в виде

$$P(A^{H}, x)=\int_{y}^{n}\bar{p}(y)\,dy=\int_{x}^{\infty}\bar{p}(x)\,dx=\frac{dy}{dx},$$

а накопленная вероятность y разных (новых) событий, уже наступивших при x испытаниях, будет равна

$$\bar{F}(y)=\bar{F}(x)=\int_{0}^{y}\bar{p}(y)\,dy=\int_{0}^{x}\bar{p}(x)\,dx=1-\frac{dy}{dx}, \tag{8}$$

где $\bar{F}(y)$, $\bar{F}(x)$ — функции распределения вероятностей новых событий.

Из формулы (8) дифференцированием по y и по x найдем выражения для средних плотностей распределения вероятностей новых событий [3, с. 7—9]:

$$\bar{p}(x)=-\frac{d^2y}{dx^2}, \tag{9}$$

$$\bar{p}(y)=-\frac{d}{dy}\left(\frac{dy}{dx}\right)=-\frac{y_x''}{y_x'}. \tag{10}$$

Из выражений (9), (10) и (4) установим связь между плотностью $p(t)$, аппроксимирующей вероятности p_k, и средними плотностями $\bar{p}(x)$, $\bar{p}(y)$. Дифференцируя дважды по x кривую роста новых событий (4) и подставляя полученные результаты в (9) и (10), получим, соответственно:

$$\bar{p}(x)=\int_{0}^{n}\frac{[p(t)]^2}{e^{xp(t)}}\,dt, \tag{11}$$

$$\bar{p}(y)=\frac{\int_{0}^{n}\frac{[p(t)]^2}{e^{xp(t)}}\,dt}{\int_{0}^{n}\frac{p(t)}{e^{xp(t)}}\,dt}. \tag{12}$$

При $x=0$, $y=0$ формулы (11), (12) дают

$$\bar{p}(x=0)=\bar{p}(y=0)=\int_{0}^{n}[p(t)]^2\,dt=M[p(t)],$$

т. е. значения средних плотностей $\bar{p}(x)$, $\bar{p}(y)$ при $x=0$, $y=0$ равны среднему значению (математическому ожиданию) плотности $p(t)$, что и оправдывает их название.

Формула (8) позволяет находить кривую роста новых событий по заданному закону распределения вероятностей новых событий, а формулы (9), (10) — решать обратную задачу. Так из формулы (8) при заданной функции распределения $\bar{F}(x)$ имеем

$$y=\int[1-\bar{F}(x)]\,dx+C, \tag{13}$$

где постоянная интегрирования C находится из условия: $y=0$ при $x=0$.

При заданной функции распределения $\bar{F}(y)$ из формулы (8) находим

$$x=\int\frac{dy}{1-\bar{F}(y)}+C. \tag{14}$$

Формулы (13) и (7) дают возможность связать частотный спектр со средней плотностью $\bar{p}(x)$. Найдем m-ю производную по x от выражения (13). Она имеет вид

$$\frac{d^m y}{dx^m}=-\frac{d^{m-2}\bar{p}(x)}{dx^{m-2}}. \tag{15}$$

Подставляя выражение (15) в формулу (7), получим

$$y_m=(-1)^m\frac{x^m}{m!}\,\frac{d^{m-2}\bar{p}(x)}{dx^{m-2}}. \tag{16}$$

Как следует из (11), средняя плотность $\bar{p}(x)$ при всех x является невозрастающей функцией $(d\bar{p}(x)/dx<0)$ и не имеет точек перегиба $(d^2\bar{p}(x)/dx^2>0)$.

Полученные формулы (9), (10), (13), (14) позволяют построить систему непрерывных распределений новых событий, а также систему кривых роста числа новых событий, т. е. систему зависимостей $y=f(x)$, и на основе последних — систему дискретных (спектровых) распределений (по формулам (7), (16)).

Для этого на первом шаге достаточно задать одно (например, равномерное) распределение вероятностей новых событий $\bar{p}(y)$. Далее отыскиваются кривая роста числа новых событий $y=f(x)$ и средняя плотность распределения вероятностей новых событий $\bar{p}(x)$, выраженная через номер испытания x. На втором шаге средняя плотность $\bar{p}(y)$ задается новой формулой, полученной на первом шаге для средней плотности $\bar{p}(x)$, и опять находятся зависимости $y=f(x)$, $\bar{p}(x)$ и т. д.

Выражение для частотного спектра на каждом шаге может быть получено либо по кривой роста числа новых событий $y=f(x)$ (с помощью формулы (7)), либо по средней плотности $\bar{p}(x)$ (с помощью формулы (16)).

Рассмотрим три первых шага (три частных случая) на пути к получению общих закономерностей.

2. ПОСТРОЕНИЕ СИСТЕМЫ ДИСКРЕТНЫХ (СПЕКТРАЛЬНЫХ) РАСПРЕДЕЛЕНИЙ

Случай 1 (процесс Пуассона). Пусть все n событий, составляющих полную группу, имеют равные вероятности $p_k=\frac{1}{n}=\alpha$. Следовательно, плотность $p(t)$,

аппроксимирующая вероятности p_k, также постоянна $p(t)=\alpha$.

По формулам (11) и (12) найдем средние плотности $\bar{p}(x)$ и $\bar{p}(y)$. При $p(t)=\alpha$ они оказываются равными

$$\bar{p}(x)=\frac{\alpha}{e^{\alpha x}}, \tag{17}$$

$$\bar{p}(y)=\alpha. \tag{18}$$

Таким образом, в процессе Пуассона средняя плотность распределения вероятностей новых событий постоянна ($\bar{p}(y)=\alpha$), а интервалы между новыми событиями (число испытаний x) распределены по показательному закону ($\bar{p}(x)=\alpha/e^{\alpha x}$).

Найдем выражение для кривой роста числа новых событий в процессе Пуассона. Функция распределения $\bar{F}(y)$ в данном случае равна

$$\bar{F}(y)=\int_0^y \bar{p}(y)\,dy=\int_0^y \alpha dy=\alpha y.$$

Тогда из формулы (14) найдем

$$x=\int\frac{dy}{1-\bar{F}(y)}+C=\int\frac{dy}{1-\alpha y}+C=-\frac{1}{\alpha}\ln(1-\alpha y),$$

откуда

$$y=\frac{1}{\alpha}\left(1-\frac{1}{e^{\alpha x}}\right), \tag{19}$$

причем, $0<y<\frac{1}{\alpha}=n$.

Дифференцируя m раз по x выражение (19) и подставляя m-ю производную в (7), найдем формулу для частотного спектра в процессе Пуассона

$$y_m=\frac{(\alpha x)^m}{m!\,e^{\alpha x}},\quad m=0,1,2,\ldots. \tag{20}$$

Из этой формулы можно найти вероятность появления событий ровно m раз при x испытаниях. Она равна отношению числа событий, наступивших ровно m раз (y_m), к общему числу разных событий, $n=1/\alpha$:

$$p_m=\frac{y_m}{n}=\frac{(\alpha x)^m}{m!\,e^{\alpha x}},\quad m=0,1,2,\ldots, \tag{21}$$

где αx есть средняя частота $\bar{m}=x/n$.

Случай 2. Пусть теперь новые события распределены по показательному закону, найденному на предыдущем этапе (случай 1) для распределения интервалов между новыми событиями с тем же параметром $\alpha-\bar{p}(y)=\alpha/e^{\alpha y}$. Тогда

$$\bar{F}(y)=1-\frac{1}{e^{\alpha y}}, \tag{22}$$

$$x=\int\frac{dy}{1-\bar{F}(y)}+C=\int e^{\alpha y}dy+C=\frac{1}{\alpha}(e^{\alpha y}-1),$$

откуда

$$y=\frac{1}{\alpha}\ln(1+\alpha x); \tag{23}$$

$$\bar{p}(x)=-\frac{d^2y}{dx^2}=\frac{\alpha}{(1+\alpha x)^2}. \tag{24}$$

Далее, из формулы (23) и (7) находим

$$y_m=\left(\frac{\alpha x}{1+\alpha x}\right)^m\frac{1}{\alpha m},\quad m=1,2,\ldots. \tag{25}$$

Разделив (25) на (23), будем иметь

$$\frac{y_m}{y}=p_m=\left(\frac{\alpha x}{1+\alpha x}\right)^m\frac{1}{m\ln(1+\alpha x)}. \tag{26}$$

Последнее распределение известно как распределение Фишера по логарифмическому ряду.

Случай 3. Пусть средняя плотность $\bar{p}(y)$ задается формулой (24), т. е.

$$\bar{p}(y)=\frac{\alpha}{(1+\alpha y)^2}.$$

Тогда, выполняя последовательность операций, описанных при рассмотрении второго случая, найдем

$$\bar{F}(y)=1-\frac{1}{1+\alpha y},\quad \bar{p}(x)=\frac{\alpha}{(1+2\alpha x)^{3/2}},$$

$$y=\frac{1}{\alpha}(\sqrt{1+2\alpha x}-1),$$

$$y_m=\left(\frac{\alpha x}{1+2\alpha x}\right)^{m-1}\frac{1\cdot3\ldots(2m-3)}{m!}y_{m=1},$$

$$\frac{y_m}{y}=p_m=\left(\frac{\alpha x}{1+2\alpha x}\right)^{m-1}\frac{1\cdot3\ldots(2m-3)}{m!}p_{m=1},$$

$$y_{m=1}=\frac{x}{\sqrt{1+2\alpha x}},\quad p_{m=1}=\frac{\alpha x}{1+2\alpha x-\sqrt{1+2\alpha x}},$$

$m=2,3,\ldots$.

Сравнив полученные результаты для трех частных случаев, нетрудно сделать обобщение. Оно достигается путем введения нового параметра u, при определенных значениях которого из общих формул будут следовать рассмотренные выше частные случаи.

Итак, в общем случае имеем

$$\bar{p}(y)=\alpha(1-\alpha u y)^{\frac{1}{u}-1}, \tag{27}$$

$$\bar{F}(y)=1-(1-\alpha u y)^{\frac{1}{u}}, \tag{28}$$

$$\bar{p}(x)=\alpha[1-\alpha(u-1)x]^{\frac{1}{u-1}-1}, \tag{29}$$

$$\bar{F}(x)=1-[1-\alpha(u-1)x]^{\frac{1}{u-1}}, \tag{30}$$

$$y=\frac{1}{\alpha u}\left[1-\frac{1}{(1+\alpha(1-u)x)^{\frac{u}{1-u}}}\right]. \tag{31}$$

Рассмотрим обобщенную плотность (27). Будем относить распределения к I типу при $0<u<\infty$; ко II типу — при $u\to0$; к III типу — при $-\infty<u<0$.

При $u=1$ формула (27) дает $\bar{p}(y)=\alpha$, т. е. имеем равномерное распределение (случай 1).

При $u\to0$ $\bar{p}(y)=\alpha/e^{\alpha y}$, т. е. плотность убывает по показательному закону (случай 2).

При $u=-1$ $\bar{p}(y)=\alpha/(1+\alpha y)^2$, т. е. случай 3. Плотность $\bar{p}(y)$ является невозрастающей при $u\leqslant1$.

Плотность $\bar{p}(x)$ отличается от плотности $\bar{p}(y)$ тем, что значения параметра u здесь на единицу меньше. Плотность $p(x)$ является невозрастающей при $u\leqslant2$.

Теперь осталось найти частотные спектры для распределений I—III типов, заданных обобщенной плотностью (27). Для этого используем формулы (7) и (31).

Распределения I типа $(0<u<\infty)$.

Из формулы обобщенной плотности (27) следует, что в данном случае число разных (новых) событий, на-

ступающих при x испытаниях, ограничено: $0<y<1/\alpha u=n$. Кривая роста числа новых событий задается формулой (31). Дифференцируя m раз по x выражение (31) и подставляя m-ю производную в выражение (7), найдем

$$y_m=\left(\frac{\alpha u x}{1+\alpha(1-u)x}\right)^m \frac{\prod_{i=0}^{m-1}\left[1+i\left(\frac{1}{u}-1\right)\right]}{m!}\, y_{m=0},$$

$$m=1, 2, \ldots, \qquad (32)$$

где

$$y_{m=0}=\frac{1}{\alpha u\left[1+\alpha(1-u)x\right]^{\frac{u}{1-u}}}.$$

Разделив, далее, выражение (32) на $n=1/\alpha u$, получим выражение для вероятности наступления событий ровно m раз при x испытаниях:

$$\frac{y_m}{n}=p_m=\left(\frac{\alpha u x}{1+\alpha(1-u)x}\right)^m \times$$

$$\times\frac{\prod_{i=0}^{m-1}\left[1+i\left(\frac{1}{u}-1\right)\right]}{m!}\, p_{m=0}, \quad m=1, 2, \ldots, \qquad (33)$$

где

$$p_{m=0}=\frac{y_{m=0}}{n}=\frac{1}{\left[1+\alpha(1-u)x\right]^{\frac{u}{1-u}}}.$$

Исследования показывают, что частными случаями распределения I типа (33) являются:

— биномиальное распределение при $u>1$;

— распределение Пуассона (21) при $u\to 1$;

— отрицательное биномиальное распределение при $0<u<1$ (в том числе геометрическое распределение при $u=1/2$).

Распределения II типа ($u\to 0$).

В данном случае из выражения (27) имеем: $0<y<\infty$. Кривая роста новых событий задается формулой (23), на основании которой получаются выражения (25) и (26).

Распределения III типа ($-\infty<u<0$).

Кривая роста числа новых событий задается формулой (31). При $-\infty<u<0$ из формул (31) и (7) имеем

$$y_m=\left(\frac{-\alpha u x}{1+\alpha(1-u)x}\right)^{m-1} \frac{\prod_{i=1}^{m-1}\left[i\left(1-\frac{1}{u}\right)-1\right]}{m!}\, y_{m=1},$$

$$m=2, 3, \ldots, \qquad (34)$$

где

$$y_{m=1}=\frac{x}{\left[1+\alpha(1-u)x\right]^{\frac{1}{1-u}}}.$$

Разделив (34) на (31), получим

$$\frac{y_m}{y}=p_m=\left(\frac{-\alpha u x}{1+\alpha(1-u)x}\right)^{m-1} \times$$

$$\times\frac{\prod_{i=1}^{m-1}\left[i\left(1-\frac{1}{u}\right)-1\right]}{m!}\, p_{m=1}, \quad m=2, 3, \ldots, \qquad (35)$$

где

$$p_{m=1}=\frac{y_{m=1}}{y}=\frac{-\alpha u x}{(1+\alpha(1-u)x)\left[1-(1+\alpha(1-u)x)^{\frac{u}{1-u}}\right]}.$$

Как и следовало ожидать, из общего распределения III типа (34), (35) при $u=-1$ имеем частный случай 3, рассмотренный выше.

Сделаем некоторые выводы.

Если новые события распределены по обобщенному закону, заданному средней плотностью распределения вероятностей новых событий (27) с масштабным параметром α и параметром формы u, то интервалы между новыми событиями распределены по закону (29) с тем же масштабным параметром α, но с параметром формы на единицу меньше ($u-1$); частота появления новых событий в зависимости от параметра u задается формулами (32), (25), (34), а кривая роста числа новых событий описывается общим уравнением (31).

Описанный в настоящей работе случайный процесс появления новых событий включает как частный случай процесс Пуассона (при $u\to 1$), для которого $\bar{p}(y)=\alpha$. С уменьшением параметра u процесс появления новых событий все более отличается от пуассоновского, а средняя плотность $\bar{p}(y)$ становится все более неравномерной. Таким образом, признаком, по которому можно классифицировать события на основании их частотных спектров, является параметр формы u выравнивающего распределения. Например, события можно ранжировать по убыванию параметра u.

Поскольку законы распределения, заданные плотностями $\bar{p}(y)$, $\bar{p}(x)$, и дискретный закон p_m связаны между собой, а также с кривой роста числа новых событий $y=f(x)$, то безразлично, по какой из этих закономерностей искать оценки параметров α, u. Наиболее просто они могут быть найдены по дискретному статистическому закону распределения.

Непрерывные плотности $p(t)$, $\bar{p}(x)$, $\bar{p}(y)$ и кривая роста числа новых событий $y=f(x)$ являются характеристиками генеральной совокупности, в то время как дискретный закон, задающий распределение разных событий по частоте m при x независимых испытаниях, — характеристикой выборки.

Следует подчеркнуть, что полученные формулы оказываются справедливыми не только для n разных несовместимых событий, составляющих полную группу, но и при рассмотрении одного и того же события, наступающего в n разных подвыборках.

3. ОЦЕНКА ПАРАМЕТРОВ ВЫРАВНИВАЮЩИХ РАСПРЕДЕЛЕНИЙ

Приведем без вывода порядок установления типа выравнивающего распределения и нахождения оценок параметров α, u.

Для установления типа выравнивающего распределения находим оценку параметра α в предположении, что распределение относится ко II типу ($u\to 0$). Оценку находим дважды, по формулам

$$\alpha_1=\frac{1}{y_{m=1}}-\frac{1}{x}; \quad \alpha_2=\frac{1}{y}\ln\frac{x}{y_{m=1}}. \qquad (36)$$

При $\alpha_1>\alpha_2$ распределение относится к типу I ($u>0$); при $\alpha_1\approx\alpha_2$ — ко II типу ($u\to 0$); при $\alpha_1<\alpha_2$ — к III типу ($u<0$).

Отметим, что для распределений трех типов спра-

ведливо отношение

$$\frac{y_{m=1}}{y_{m=2}} = 2\left(\frac{1}{\alpha x} + 1 - u\right),$$

с помощью которого можно не только установить тип выравнивающего распределения, но также приближенно оценить значение параметра u (при условии, что величина $1/\alpha x$ близка к нулю)

$$u = \frac{1}{\alpha x} + 1 - \frac{y_{m=1}}{2y_{m=2}}.$$

Распределения с модой $Mo > 0$ или при $y_{m=0} > 0$ всегда относятся к I типу.

Для распределений I типа по опытным данным находим оценки параметров α, u по формулам

$$\alpha = \frac{1}{x^2}\sum_{m\geqslant 1} m^2 y_m - \frac{1}{x}, \qquad (37)$$

$$u = \frac{1}{\alpha n}, \qquad (38)$$

где

$$x = \sum_{m\geqslant 1} m y_m, \quad n = \sum_{m\geqslant 0} y_m.$$

Для распределений II типа методом итераций уточняем полученную ранее по формулам (36) оценку α

$$\alpha_{i+1} = \frac{\ln(1+\alpha_i x)}{y}, \qquad (39)$$

где α_i — любое из ранее вычисленных значений α (α_1 или α_2); $y = \sum_{m\geqslant 1} y_m$.

Для распределений III типа методом итераций находим оценку параметра u:

$$u_{i+1} = -(1-u_i)\frac{x}{y}\frac{\left(\frac{x}{y_{m=1}}\right)^{-u_i} - 1}{\left(\frac{x}{y_{m=1}}\right)^{1-u_i} - 1}. \qquad (40)$$

Далее рассчитываем оценку параметра α:

$$\alpha = \frac{1}{-uy}\left[\left(\frac{x}{y_{m=1}}\right)^{-u} - 1\right]. \qquad (41)$$

Последние две формулы справедливы также для распределений I типа. Из них, в частности, следует, что параметры α, u могут быть оценены по трем известным из опыта величинам: $x, y, y_{m=1}$. Далее по соответствующим формулам восстанавливаются кривая роста числа новых событий, средние плотности $\bar{p}(y)$, $\bar{p}(x)$, а также дискретный (спектровый) закон распределения разных событий.

4. РАНЖИРОВАНИЕ СЛОВ ПО СТЕПЕНИ СЕМАНТИЧЕСКОЙ НАГРУЗКИ

В книге [4, с. 18—20] приведены частотные спектры белорусских слов *і, у, не, але, ты*.

Найдем в качестве примера дискретное выравнивающее распределение для частотного спектра слова *але*. В табл. 1 приведены частоты m и количество подвыборок y_m (каждая длиной 1000 словоупотреблений), в которых данное слово встретилось ровно m раз [4]. Общее количество испытаний в данном случае составляет $x = \sum_{m\geqslant 1} m y_m = 1559$; количество наступивших при x испытаниях разных событий — $y = \sum_{m\geqslant 1} y_m = 284$, а общее количество подвыборок, в которых слово *але* употребилось от 0 до $m_{\max} = 18$ раз, составило 290 $\left(n = \sum_{m\geqslant 0} y_m = 290\right)$.

Из табл. 1 видно, что статистическое распределение имеет моду $Mo = 5$, т. е. $Mo > 0$. Следовательно, выравнивающее распределение относится к I типу. Найдем по формулам (37) и (38) оценки параметров α, u. В результате вычислений получим: $\alpha = 0{,}003968$, $u = 0{,}86903$.

Таблица 1

Распределение употреблений слова *але* по фрагментам длиной 1000 словоупотреблений

Частота слова в тысяче, m	Количество тысяч (фрагментов), y_m	Теоретические значения y_m	$\frac{(y_m^{\text{эмп}} - y_m)^2}{y_m}$
0	5	5,65	0,07
1	19	16,79	0,61
2	30	28,69	0,06
3	30	36,96	1,30
4	36	39,85	0,37
5	47	37,93	2,16
6	32	32,93	0,03
7	30	26,60	0,16
8	24	20,29	0,33
9	9	14,77	2,25
10	8	10,34	0,53
11	7	7,00	0
12	4	4,60	0,08
13	3	2,95	0
14	3	1,85	0,71
15	1	1,14	0,02
16	0	0,69	0,14
17	1	0,41	0,85
18	1	0,24	2,40
Итого:	290	289,68	12,07

Чтобы оценить степень приближения выравнивающего распределения к статистическому, рассчитаем значения y_m при $m = 0, 1, \ldots, 18$ по формулам

$$y_{m+1} = y_m \frac{\alpha u x\left[1 + m\left(\frac{1}{u} - 1\right)\right]}{[1+\alpha(1-u)x](m+1)},$$

$$y_{m=0} = \frac{1}{\alpha u [1+\alpha(1-u)x]^{\frac{u}{1-u}}},$$

которые следуют из формулы (32), и вычислим критерий «хи-квадрат».

Он оказался равным $\chi^2 = 12{,}07$, что при 16 степенях свободы соответствует вероятности $P(\chi^2) = 0{,}744$ (число степеней свободы вычислялось по формуле: $\nu = k - r - 1$, где $k = 19$ — число строк в табл. 1, $r = 2$ — число параметров выравнивающего распределения). Поскольку эта вероятность достаточно высока, отклонения статистического распределения от выравнивающего следует признать случайными.

Таким образом, слово *але* характеризуется показателем $u=0{,}86903<1$, т. е. о данном слове нельзя сказать, что оно употребляется в текстах независимо и случайно. Его употребление в некоторой степени связано с ситуацией.

Таким же путем были найдены оценки параметров α, u выравнивающих распределений для других слов. Теперь осталось упорядочить слова по убыванию оценки параметра u (табл. 2).

Таблица 2

Ранжирование слов по убыванию оценки параметра u

Слово	Ранг	Оценки параметров	
		u	α
у	1	0,97077	0,0035521
і	2	0,96825	0,0035491
не	3	0,92216	0,0037653
але	4	0,86903	0,0039680
ты	5	0,56519	0,0061222

Распределение слов *у*, *і* по частоте встречаемости в подвыборках весьма близко к закону Пуассона, для которого $u\to1$. Эти слова употребляются в текстах равномерно.

Слово *ты*, для которого $u\approx0{,}565$, употребляется в текстах весьма неравномерно. Его употребление в наибольшей степени, по сравнению с другими словами, связано с ситуацией.

Таким образом, с помощью критерия u слова могут быть ранжированы по степени неравномерности употребления в текстах, которая связана с семантической нагрузкой слов.

Задавая некоторое пороговое значение параметра u, с помощью ЭВМ можно автоматически выделять ключевые слова текста (при достаточно большой его длине), а также служебные слова.

Полученные в настоящей работе результаты могут быть использованы при описании потоков НТИ, при установлении закона распределения дескрипторов по частоте их употребления в поисковых образах документов, описании динамики новых дескрипторов, при оценке степени неравномерности использования абонентами запросов в режиме ИРИ, а также во всех тех случаях, когда речь идет о последовательности независимых испытаний и когда частота появлений разных событий в выборке подчиняется одному из дискретных законов: биномиальному, закону Пуассона, отрицательному биномиальному, распределению Фишера по логарифмическому ряду и некоторым другим.

ЛИТЕРАТУРА

1. Пиотровский Р. Г. Тест, машина, человек. — Л.: Наука, 1975.
2. Калинин В. М. Некоторые статистические законы математической лингвистики. — В кн.: Проблемы кибернетики. Вып. II. М., 1964.
3. Нешитой В. В. Система непрерывных распределений для построения информационных моделей. — Минск: БелНИИНТИ, 1976.
4. Можейко Н. С., Супрун А. Е. Частотный словарь белорусского языка. Художественная проза. — Минск: БГУ им. В. И. Ленина, 1976.

Статья поступила в редакцию 16 мая 1985 г.

ОПЕЧАТКА

к сб. «Научно-техническая информация», серия 2, 1986, 2

Страница	Колонка	Строка	Напечатано	Следует читать
13	левая	9 сверху	поиска информации о классах соединений	поиска информации об индивидуальных соединениях