В. В. НЕШИТОЙ

МОДЕЛИРОВАНИЕ КРИВОЙ РОСТА И СТАТИСТИЧЕСКОЙ СТРУКТУРЫ СЛОВАРЯ КЛЮЧЕВЫХ СЛОВ

Рассматривается один класс случайных функций, описывающих статистическую зависимость между количеством произведенных испытаний и количеством наступивших при этом разных событий.

При известном математическом ожидании случайной функции (кривой роста) строится система дискретных распределений, дается классификация распределений и кривых роста, разрабатывается метод оценивания параметров.

Полученные результаты могут быть использованы при обработке библиотечной статистики для выравнивания и прогнозирования кривой роста разных событий (например, разных наименований выданных книг), расчета и прогнозирования частотной структуры выборки, автоматического выделения ключевых слов, вычисления полноты словаря и полноты фонда, ранжирования слов по степени семантической нагрузки, а также во всех тех случаях, когда речь идет о последовательности независимых испытаний при этом частота появления разных событий подчиняется одному из рассмотренных дискретных законов распределения.

При свободном индексировании документов (т.е. без контроля по тезаурусу) с ростом количества заиндексированных документов с некоторой статистической закономерностью растет количество ключевых слов. Естественно, что отдельные ключевые слова при индексировании могут использоваться многократно.

Обозначим через X общее количество употреблений ключевых слов с учетом их повторяемости (объем выборки), а через Y – количество разных ключевых слов (объем словаря). Статистическая зависимость между величинами X, Y представляет собой реализацию (траекторию) случайной функции $Y(X)$. Такого рода зависимости имеются в информатике, математической лингвистике, библиотечно-информационной деятельности, технике. В качестве примеров можно привести статистические зависимости между следующими величинами:

– объемом выборки в словоупотреблениях и объемом словаря;

– количеством книговыдач и количеством разных наименований выданных книг;

– количеством абонентозапросов (т.е. запросов с учетом их повторяемости) и количеством разных запросов.

Рассмотрим далее математическое ожидание случайной функции $M[Y(X)]$. Его график представляет собой ломаную линию, около которой располагаются возможные реализации случайной функции. Эту ломаную можно аппроксимировать некоторой плавной кривой $y = f(x)$, которую назовем кривой роста разных событий.

Для аппроксимации статистических зависимостей между количеством произведенных испытаний и количеством наступивших разных событий автором [2] разработана система кривых роста, заданная двухпараметрической формулой с параметрами α, u

$$y = \frac{1}{au}\left[1-\left(1-a\left(u-1\right)x\right)^{\frac{u}{u-1}}\right], \qquad (1)$$

где y – количество наступивших разных событий (разных слов, в том числе ключевых, наименований книг, запросов и т.д.); x – количество произведенных испытаний (объем выборки в словоупотреблениях, число книговыдач, число абонентозапросов и т.д.).

Формула (1) включает систему кривых роста, которые можно разделить на типы в зависимости от значений параметра u.

При $u > 0$ имеем кривые роста I типа. Они задаются формулой (1). В частности, при $u \to 1$ из (1) следует формула

$$y=\frac{1}{a}\left(1-\frac{1}{e^{a\,x}}\right). \qquad (2)$$

При $u\to 0$ из (1) следует кривая II типа

$$y=\frac{1}{a}\ln(1+a\,x). \qquad (3)$$

При $u < 0$ имеем кривую III типа. Она задается той же формулой (1).

Между кривой роста разных событий и статистической структурой выборки существует взаимосвязь, установленная В.М.Калининым [1],

$$y_m=(-1)^{m+1}\,\frac{x^m}{m!}\,\frac{d^m y}{dx^m}. \qquad (4)$$

По формуле (4) можно рассчитать частотный спектр, или статистическую структуру выборки, т.е. количество событий с частотой появления 1, 2,…, m раз, если задана кривая роста разных событий $y = f(x)$.

Формулы (1) и (4) позволяют также построить систему дискретных распределений.

Построение системы дискретных распределений

Распределения I типа ($u > 0$). Продифференцируем выражение (1) m раз по x и подставим m-ю производную в (4). В результате получим формулу, позволяющую вычислять число событий с частотой m, т.е. y_m при числе испытаний x

$$y_m=\frac{y_{m=0}}{m!}\left(\frac{a\,u\,x}{1+a\,(1-u)\,x}\right)^m\prod_{i=0}^{m-1}\left[1+i\left(\frac{1}{u}-1\right)\right],\quad m=1,2,\ldots, \qquad (5)$$

где

$$y_{m=0}=\frac{1}{a\,u}\left[1+a\,(1-u)\,x\right]^{\frac{u}{u-1}}. \qquad (6)$$

В данном случае число разных событий, наступающих при x испытаниях, ограничено: $0<y<1/a\,u$, причем $1/a\,u = n$ (величина n – это число разных событий, составляющих полную группу; сумма вероятностей этих событий равна единице).

Разделив величину y_m на n, получим выражение для вероятности наступления событий ровно m раз при x испытаниях: $p_m = y_m/n$ (при этом в правой части формулы (5) следует разделить на n величину $y_{m=0}$):

$$p_m=\frac{p_{m=0}}{m!}\left(\frac{a\,u\,x}{1+a\,(1-u)\,x}\right)^m\prod_{i=0}^{m-1}\left[1+i\left(\frac{1}{u}-1\right)\right],\quad m=1,2,\ldots, \qquad (7)$$

где

$$p_{m=0}=\left[1+a\,(1-u)\,x\right]^{\frac{u}{u-1}}. \qquad (8)$$

Исследования показали, что частными случаями распределения I типа (7) являются: биномиальное – при $u > 1$; Пуассона – при u → 1; отрицательное биномиальное – при $0 < u < 1$ (в том числе геометрическое распределение – при $u = ½$).

Распределения II типа ($u \to 0$). В данном случае кривая роста разных событий задается формулой (3), на основании которой и формулы В.М.Калинина (4) имеем

$$y_m = \left(\frac{a\,x}{1+a\,x}\right)^m \frac{1}{a\,m}, \quad m=1,2,... \tag{9}$$

Разделив (9) на (3), получим

$$\frac{y_m}{y} = p_m = \left(\frac{a\,x}{1+a\,x}\right)^m \frac{1}{m\ln(1+a\,x)}. \tag{10}$$

Последнее распределение известно как распределение Фишера по логарифмическому ряду и находит широкое применение в биологии [3].

Распределения III типа ($-\infty < u < \infty$). Кривая роста разных событий задается общей формулой (1). Из (1) и (4) имеем

$$y_m = \frac{y_{m=1}}{m!}\left(\frac{-a\,u x}{1+a\,(1-u)x}\right)^{m-1} \prod_{i=1}^{m-1}\left[i\left(1-\frac{1}{u}\right)-1\right], \quad m=2,3,..., \tag{11}$$

где

$$y_{m=1} = x\left[1+a\,(1-u)x\right]^{\frac{1}{u-1}}. \tag{12}$$

Разделив y_m на y, получим выражение для вероятности p_m.

$$p_m = \frac{p_{m=1}}{m!}\left(\frac{-a\,u x}{1+a\,(1-u)x}\right)^{m-1} \prod_{i=1}^{m-1}\left[i\left(1-\frac{1}{u}\right)-1\right], \quad m=2,3,..., \tag{13}$$

где

$$p_{m=1} = \frac{-a\,u x}{\left(1+a\,(1-u)x\right)\left[1-\left(1+a\,(1-u)x\right)^{\frac{u}{1-u}}\right]}. \tag{14}$$

Оценивание параметров дискретных распределений

Для установления типа аппроксимирующего дискретного распределения введем критерий [2]

$$HD = \frac{\frac{x}{y}\ln\frac{x}{y_{m=1}}}{\frac{x}{y_{m=1}}-1}. \qquad (15)$$

При $HD = 1$ выравнивающее распределение относится ко II типу. При $HD < 1$ – к I типу. При $HD > 1$ – к III типу.

Далее рассчитываются оценки параметров α, u. Их можно найти по методу моментов. В случае распределений I типа

$$a = \sum_{m\geq 1}\left(\frac{m}{x}\right)^2 y_m - \frac{1}{x}, \qquad 16)$$

$$u = \frac{1}{a\,n}, \qquad (17)$$

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

В случае распределений II типа оценка единственного параметра α находится методом простых итераций по формуле, которая следует из (3)

$$a_{i+1} = \frac{1}{y}\ln(1 + a_i x), \qquad (18)$$

где $y = \sum_{m\geq 1} y_m$; a_i – значение параметра α на предыдущем шаге итерации. В качестве первого приближения можно принять $a_1 = 1/y_{m=1}$.

В случае распределений III типа (а также I типа) оценка параметра u может быть найдена методом итераций по формуле

$$u_{i+1} = (1-u_i)\frac{x}{y}\frac{1-\left(\frac{y_{m=1}}{x}\right)^{u_i}}{\left(\frac{x}{y_{m=1}}\right)^{1-u_i}-1}. \qquad (19)$$

Тогда оценка параметра α равна

$$a = \frac{1}{uy}\left[1-\left(\frac{y_{m=1}}{x}\right)^u\right] = \frac{1}{(1-u)x}\left[\left(\frac{x}{y_{m=1}}\right)^{1-u}-1\right]. \qquad (20)$$

Таким образом, для оценивания параметров α, u достаточно знать три величины: x, y, $y_{m=1}$.

Отметим, что формулы (16), (19), (20) справедливы для распределений трех типов.

При известных оценках параметров α, u вычисляются теоретические значения y_m и сравниваются со статистическими. Расчет осуществляется по рекуррентной формуле

$$y_{m+1} = y_m \frac{a\,x[u+m(1-u)]}{[1+a(1-u)x](m+1)}\ , \qquad (21)$$

которая справедлива для распределений всех трех типов. Вначале по формуле (12) вычисляется количество событий с частотой m = 1, т.е. $y_{m=1}$. Далее по формуле (21) последовательно находятся значения y_{m+1} при m = 1, 2 и т.д. В случае распределений I типа дополнительно вычисляется величина $y_{m=0}$ по формуле (6).

Кривая роста и статистическая структура словаря ключевых слов

Построенная система дискретных распределений, взаимосвязанная с системой кривых роста разных событий, позволяет легко решать многие задачи. Рассмотрим пример.

За некоторое время эксплуатации БелРАСНТИ при индексировании документов по автомобильному транспорту было употреблено y = 3786 разных ключевых слов при общей их частоте употребления x = 147644. Количество ключевых слов с частотой m = 1 составило $y_{m=1}$ = 1518. По этим трем величинам требуется рассчитать:

– тип аппроксимирующего дискретного закона распределения;

– оценки параметров u, α дискретного распределения и кривой роста разных ключевых слов;

– кривую роста разных ключевых слов.

– частотный спектр ключевых слов, т.е. количество ключевых слов с частотой употребления 1, 2, … , m раз.

Установим тип аппроксимирующего дискретного распределения. Для этого вычислим по формуле (15) критерий HD. Он оказался равным 1,854.

Поскольку HD>1, то искомое распределение относится к III типу. Далее по формулам (19), (20) вычисляем оценки параметров u, α: u = - 0,601736; α = 0,00645759. Частотный спектр описывается формулой (21).

Кривая роста разных ключевых слов описывается уравнением (1), которое при найденных оценках параметров u, α примет вид

$$y = 257.35\left[(1+0.0103435\,x)^{0.375677} - 1\right]. \qquad (22)$$

Рассчитанные по формуле (22) значения y приведены в таблице 1, графа 3. В той же таблице в графе 2 даны значения y, восстановленные по частотному спектру ключевых слов с помощью формулы В.М.Калинина [1]

$$y = y_0 - \sum_{m\geq 1}\left(1-\frac{x}{x_0}\right)^m y_m, \qquad (23)$$

где x, y – текущие значения объемов выборки и словаря ($x < x_0$; $y < y_0$; $x_0 = 147644$; $y_0 = 3786$). В таблице приводится также расчетное количество ключевых слов с частотой употребления один и два раза. Все расчеты выполнены по программе автора SDR99.

Таблица 1

Зависимость количества разных ключевых слов y от объема выборки x

Объем выборки x	Количество разных ключевых слов y		Количество ключевых слов с частотой	
	по частотному спектру	по формуле (22)	один раз $y_{m=1}$	два раза $y_{m=2}$
1	2	3	4	5
10000	1222	1218	549	170
20000	1671	1654	714	222
30000	1988	1967	833	259
40000	2240	2220	928	289
50000	2454	2436	1010	315
80000	2963	2955	1205	376
100000	3239	3236	1311	409
147644	3786	3786	1518	474
200000		4274	1702	531
500000		6135	2401	749
1000000		8037	3116	972
1500000		9401	3628	1133
2000000		10503	4042	1262

Анализ данных таблицы показывает, что максимальная относительная ошибка формулы (22) составила около 1%, и это при условии, когда оценки параметров u, α вычислялись всего лишь по трем величинам: x, y, $y_{m=1}$.

Формула (1) позволяет прогнозировать рост словаря ключевых слов в зависимости от количества заиндексированных документов D, а также полноту словаря, что важно знать при ведении информационно-поискового тезауруса. Для этого достаточно в формуле (1) заменить величину x на произведение Dh, где h – глубина индексирования. Ее можно оценить количеством ключевых слов, приходящихся в среднем на один поисковый образ документа.

Полноту словаря ключевых слов будем измерять вероятностью непоявления нового ключевого слова в точке с координатами (x;y) кривой роста, т.е. функцией

распределения вероятностей новых событий $\overline{F}(y)$, при этом $\overline{F}(y)=\overline{F}(x)$. Новым будем считать любое слово при первом его употреблении для индексирования документа.

Полнота словаря рассчитывается по формуле [3]

$$\overline{F}(y)=\overline{F}(x)=1-\frac{dy}{dx}, \qquad (24)$$

где первая производная dy/dx равна вероятности появления нового слова.

Дифференцируя выражение (1) и подставляя первую производную в (24), получим

$$\overline{F}(y)=1-\left(1-a\,u\,y\right)^{\frac{1}{u}},$$

$$\overline{F}(x)=1-\left(1-a\left(u-1\right)x\right)^{\frac{1}{u-1}}.$$

Последние две формулы позволяют вычислять значения величин y, x, при которых будет достигнута заданная полнота $\overline{F}(y)=\overline{F}(x)$:

$$y=\frac{1}{a\,u}\left[1-\left(1-\overline{F}(y)\right)^{u}\right],$$

$$x=\frac{1}{a\,(u-1)}\left[1-\left(1-\overline{F}(x)\right)^{u-1}\right].$$

Полноту словаря объемом y можно также выразить через число ключевых слов с частотой употребления один раз [2]

$$\overline{F}(y)=1-\frac{y_{m=1}}{x},$$

где $y_{m=1}$ можно взять из частотного словаря ключевых слов. Эта формула следует из (24) и (4) при m = 1.

В таблице 2 приведены частоты употребления и количество ключевых слов с указанной частотой. Статистические и расчетные данные, вычисленные по программе SDR99, достаточно близки между собой.

Система дискретных распределений, взаимосвязанная с системой кривых роста разных событий, может быть использована во всех тех случаях, когда речь идет о последовательности независимых испытаний и при этом частота появления разных событий подчиняется одному из дискретных законов, описанных в настоящей работе.

Использование системы непрерывных распределений [4] наряду с системой дискретных распределений, а также кривых роста и компьютерных программ, т.е. использование теории обобщенных распределений в целом, позволяет описать все многообразие статистических распределений и кривых роста, которые встречаются в библиотечно-информационной деятельности.

С помощью математико-статистических моделей, наиболее точно аппроксимирующих статистические закономерности, из библиотечной (и любой другой) статистики может быть извлечена наиболее полная, объективная и ценная информация. При этом теория требует наличия определенных статистических данных. Так, при статистическом учете количества книговыдач, количества абонентозапросов совершенно необходимо вести учет количества разных наименований выданных книг, разных запросов.

Таблица 2

Количество ключевых слов y_m с заданной частотой m

Частота m	Количество слов по факту		Количество слов по расчету	
	y_m	сумма y_m	y_m	сумма y_m
1	2	3	4	5
1	1518	1518	1518	1518
2	450	1968	473,6	1991,6
3	229	2197	256,2	2247,8
4	144	2341	168,0	2415,8
5	136	2477	121,7	2537,5
6	119	2596	93,7	2631,2
7	77	2673	75,3	2706,5
8	66	2739	62,3	2768,8
9	72	2811	52,7	2821,5
10	55	2866	45,4	2866,9
11	47	2913	39,7	2906,6
12	38	2951	35,2	2941,8
13	41	2992	31,4	2973,2
14	27	3019	28,3	3001,5
15	25	3044	25,7	3027,2
16 и >	742	3786	758,8	3786

Только в этом случае может быть построена статистическая кривая роста разных событий. Анализ такой кривой дает объективную информацию, необходимую при решении различных задач. Это оптимизация комплектования фонда, оценка его полноты, анализ использования, оценка состояния и прогнозирование.

Использование системы непрерывных распределений позволяет вычислять наилучшее аппроксимирующее распределение, в том числе ранговое, находить универсальные законы рассеяния и старения публикаций. На базе универсального закона рассеяния можно дать математически точную формулировку закона Бредфорда, вычислить границы ядра и зон рассеяния, доли статей в каждой зоне.

По всем разделам теории обобщенных распределений автором созданы компьютерные программы, которые апробированы на большом статистическом материале начиная с 1990 г.

Использование теории обобщенных распределений гарантирует высокую экономическую эффективность статистических методов во всех практических приложениях, в том числе в библиотечно-информационной деятельности, в системах управления качеством, в научных исследованиях.

---

1. *Калинин*, В.М. Некоторые статистические законы математической лингвистики / В.М.Калинин // Проблемы кибернетики. – 1964. – Вып. 11. – С. 246–255.

2. *Нешитой*, В.В. Исследование статистических закономерностей текста и информационных потоков: дис. … докт. техн. наук / В.В.Нешитой. – Мн., 1987. – 505 с.

3. *Нешитой*, В.В. Статистические модели в биологии / В.В.Нешитой // Кибернетика. – 1987. – № 6. – С. 91–96.

4. *Нешитой*, В.В. Методы статистического анализа на базе обобщенных распределений: учеб.-метод. пособие / В.В.Нешитой. – Мн.: Веды, 2001. – 168 с.