БІБЛІЯТЭЧНА-ІНФАРМАЦЫЙНЫЯ РЭСУРСЫ Ў АДУКАЦЫЙНЫМ ПРАЦЭСЕ

В. В. Нешитой, *д-р технических наук, проф.,*
проф. каф. информационных ресурсов БГУКИ

НАУКОМЕТРИЧЕСКИЙ АНАЛИЗ ДОКУМЕНТНЫХ ПОТОКОВ НА БАЗЕ РАНГОВЫХ МОДЕЛЕЙ

Наукометрия – это совокупность количественных методов исследования структуры и динамики массивов и потоков научной информации. Результаты этих исследований весьма важны в библиотечно-информационной деятельности, например, при решении задач оптимизации комплектования библиотечного фонда, вычислении ядра и зон рассеяния публикаций, определении вероятности удовлетворения информационных потребностей пользователей фондом заданного объема и др.

Исследование структуры документных потоков осуществляется обычно на базе законов Дж. Ципфа (для описания статистических ранговых распределений), эмпирического закона рассеяния публикаций С. Бредфорда, законов Лотки, Парето и др. При их отыскании используется метод выдвижения гипотез и проверки каждой из них по критериям согласия. Как правило, такие исследования не дают существенных результатов, потому что предпринимаются попытки решать частные задачи без предварительного решения общей задачи – разработки методов вычисления закона распределения, в том числе рангового, по статистическому ряду. Но для этого требуется решить несколько достаточно сложных проблем.

Во-первых, необходимо дать классификацию непрерывных случайных величин в зависимости от их свойств. Поскольку число классов случайных величин ограничено, то по традиционному методу можно будет выдвинуть не более двух-трех гипотез, хотя в большинстве случаев достаточно и одной.

Во-вторых, для каждого класса случайных величин необходимо разработать свою систему непрерывных распределений,

способную с высокой точностью аппроксимировать все многообразие статистических распределений своего класса. Свойства случайной величины определяют систему распределений.

В-третьих, необходимо разработать метод вычисления закона распределения и оценок параметров по статистическому ряду, единый для всех систем непрерывных распределений.

В-четвертых, довести этот метод до программной реализации.

Все эти задачи автором решены и в итоге разработана теория обобщенных распределений [1]. Использование этой теории позволяет вычислять закон распределения, в том числе рангового, по статистическому ряду. На базе найденного закона легко решаются многие задачи.

Рассмотрим первые плотности первой и второй систем непрерывных распределений:

$$p(x) = Ne^{kbx}(1 - aue^{bx})^{\frac{1}{u}-1}, \quad -\infty < x < \infty; \qquad (1)$$

$$p(t) = Nt^{kb-1}(1 - aut^{b})^{\frac{1}{u}-1}, \quad 0 < t < \infty, \qquad (2)$$

где α, β, k, u – параметры распределений; N – нормирующий множитель.

Плотность (1) предназначена для описания статистических распределений таких случайных величин, последующие значения которых образуются из предыдущих путем прибавления некоторой величины C, а плотность (2) – путем умножения предыдущих значений на величину C. Отсюда следует, что без выдвижения гипотез об аппроксимирующем распределении можно безошибочно выбрать систему непрерывных распределений на основании свойств случайной величины. Например, статистическое распределение сотрудников некоторой организации по возрасту должно описываться первой системой непрерывных распределений (1), а распределение тех же сотрудников по размеру заработной платы – второй системой (2). Этой же системой должно описываться, например, распределение межгалактического расстояния, что не противоречит закону всеобщего разбегания галактик – закону Хаббла.

Следует отметить, что кривые распределения, представляющие собой графики плотности (1), при значениях параметра $u<1/2$ всегда имеют моду (это такое значение случайной величины X, при котором плотность максимальна) и две точки перегиба, расположенные на равных расстояниях по обе стороны от моды (они отделяют выпуклую часть кривой от вогну-

той). Поэтому эта плотность может описывать только одновершинные статистические распределения, в том числе распределение числа ссылок в зависимости от года издания публикаций. Поскольку закон распределения наиболее полно характеризует случайную величину, плотность (1) по праву может считаться универсальным законом старения публикаций.

Плотность (2) может описывать не только одновершинные распределения, но и ранговые (убывающие). Она является универсальным законом рассеяния публикаций. На базе этой плотности можно вычислить координаты трех характерных точек, т.е. границ ядра и зон рассеяния публикаций. Для этого плотность (2) необходимо привести к форме плотности (1). Если умножить левую и правую части формулы (2) на t, а выражение t^b представить в виде $e^{b \ln t}$, то получится плотность (1), так как tp(t)=p(x), lnt=x. Следовательно, график статистической зависимости rp(r)=f(lnr), где r – ранг (журнала, книги, термина и т.д.) должен иметь моду C и две точки перегиба A и B, расположенные на равных расстояниях от моды. Они делят кривую распределения на четыре части с разными долями статей: ядро и три зоны рассеяния. При 1/2<u<1 третьей зоны не существует.

Методы вычисления законов распределения по статистическим данным, а также границ ядра и зон рассеяния публикаций изложены в работах [1, 2]. При вычисленных оценках параметров теоретического закона мода t_C находится из условия $dtp(t)/dlnt = 0$, а точки перегиба t_A, t_B – из условия $d^2tp(t)/d(\ln t)^2$. Доли статей в каждой зоне и для любого другого интервала рангов вычисляются с помощью функции распределения или по статистическому ранговому распределению.

В некоторых случаях статистическое ранговое распределение может с высокой точностью описываться законом Вейбулла, функция распределения и плотность вероятностей которого задаются формулами

$$F(t) = 1 - e^{-at^b}, \quad p(t) = abt^{b-1}e^{-at^b}.$$

Поскольку этот закон весьма простой, его целесообразно проверять, в первую очередь, при отыскании подходящего рангового распределения. Для этого функцию распределения необходимо преобразовать к линейному виду

$$\ln\ln(1/(1-\mathrm{F(t)})) = \ln a + b\ln t. \qquad (3)$$

Принимая $Y = \ln\ln(1/(1-\mathrm{F(t)})),\ X = \ln t,$ из (3) получим $Y=\ln\alpha + \beta X$.

Для проверки применимости закона Вейбулла необходимо по статистической функции распределения вычислить значения X, Y и построить график зависимости Y=f(X). Если эмпирические точки расположатся вдоль прямой, то далее по методу наименьших квадратов вычисляются оценки параметров α и β этой прямой:

$$b = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \left(\overline{X}\right)^2},\quad a = \exp\left(\overline{Y} - b\,\overline{X}\right). \qquad (4)$$

Абсциссы точек A, C, B для закона Вейбулла вычисляются по формулам

$$t_C = \left(\frac{1}{a}\right)^{\frac{1}{b}};\quad t_A = \frac{t_C}{n};\quad t_B = t_C \cdot n;\quad n = \left(\frac{3+\sqrt{5}}{2}\right)^{\frac{1}{b}}. \qquad (5)$$

Значения функции распределения в этих точках при любых значениях параметров α и β соответственно равны:

$$F(t_A) = 0{,}31748;\ F(t_C) = 0{,}63212;\ F(t_B) = 0{,}92705. \qquad (6)$$

Рассмотрим в качестве примера кумулятивное число статей в сериальных изданиях, отраженных в выпуске РЖ «Математика» (1997–2010 г.), в зависимости от ранга [3]. Общее число источников равно 3799, число статей в них – 257 960.

Автор отмеченной статьи не нашел теоретического рангового закона распределения. Использование второй системы непрерывных распределений привело к однозначному решению – эти статистические данные с высокой точностью описываются законом Вейбулла, который следует из плотности (2) при u→0, k=1 (см. рис.). Параметры закона Вейбулла равны: α= 0,023349, β=0,687331. Далее по формулам (5) вычисляются абсциссы характерных точек: n=4,56; $[t_C = 236{,}598]$; $[t_A = 58{,}333]$; $[t_B = 959{,}641]$.

Накопленные доли статей в этих точках задаются формулами (6). Таким образом, в ядро входят 58 журналов, которые содержат 31,7% статей. Оптимальный объем фонда, т.е. ядро и первые две зоны рассеяния, составляют 960 журналов, на которые приходится 92,7% статей. На остальные 2839 журналов, т.е. самую обширную третью зону рассеяния, приходится лишь 7,3% статей.

Ранговое распределение сериальных изданий по числу опубликованных в них статей по математике (прямая линия – расчетная).

1. Нешитой, В. В. Элементы теории обобщенных распределений: монография / В. В. Нешитой. – Минск : РИВШ, 2009. – 204 с.

2. Нешитой, В. В. Математико-статистические методы анализа в библиотечно-информационной деятельности : учеб.-метод. пособие / В. В. Нешитой. – Минск : БГУ культуры и искусств, 2009. – 203 с.

3. Шамаев, В. Г. Инфометрическое исследование документального потока по физико-математическим и некоторым другим наукам, отраженным в РЖ ВИНИТИ РАН / В. Г. Шамаев. – НТИ. Сер. 2. – 2011. – № 1. – С. 24–30.