В. В. Нешитой,
доктор технических наук, профессор,
профессор кафедры информационных ресурсов

О ВЗАИМОСВЯЗИ ЗАКОНОВ СТАРЕНИЯ И РАССЕЯНИЯ ИНФОРМАЦИИ

Рассмотрим первую и вторую системы непрерывных распределений [2], которые заданы четырехпараметрическими плотностями вида:

$$p(x) = Ne^{k\beta x}(1-\alpha u e^{\beta x})^{\frac{1}{u}-1}, \quad -\infty < x < \infty; \qquad (1)$$

$$p(t) = Nt^{k\beta-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1}, \qquad 0 < t < \infty, \qquad (2)$$

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

Плотность (1) обладает тем замечательным свойством, что при u<1/2 кривые распределения, заданные этой плотностью, имеют моду и две точки перегиба, расположенные на равных расстояниях по обе стороны от моды. Эта плотность является универсальным законом старения публикаций. Она не может аппроксимировать статистические ранговые распределения. Для этих задач наиболее подходит плотность (2), т. е. вторая система непрерывных распределений. Отметим, что обе плотности взаимосвязаны. Так вторая система непрерывных распределений может быть получена из первой при $x = lnt$ по известной формуле $p(t) = p(x)dx/dt,$ на основании которой и равенства $x = lnt$ имеем плотность (2).

Отсюда следует, что при известном универсальном законе старения легко находится формула (в виде распределения) для универсального закона рассеяния. Эта взаимосвязь значительно облегчает поиск обоих законов, но несмотря на это обстоятельство, они так и не были открыты. Даже для аппроксимации ранговых распределений не было предложено закона, с удовлетворительной точностью описывающего все статистические ранговые распределения. То же касается и закона старения публикаций.

Изучению потоков научно-технической информации посвящено большое количество работ. При этом особое внимание уделяется ранговым распределениям и закону рассеяния публикаций.

Если упорядочить журналы по убыванию числа помещенных в них статей по данной тематике, то получим ранговое распределение, для описания которого Дж. К. Ципф предложил формулу (закон Дж. К. Ципфа[11])

$$p_r = \frac{k}{r^{\gamma}}, \qquad (3)$$

из которого при γ=1 следует, что произведение ранга слова (журнала) на относительную частоту есть величина постоянная, равная k, т. е. $rp_r = k$. Однако последнее равенство никогда не выполняется даже на частотном словаре текста Джойса Улисс, на базе которого Ципф предложил свой «закон». Но даже в этом случае кривая распределения $\mathrm{rp_r} = f(\ln r)$ имеет вид не прямой, а колоколообразной кривой с модой и двумя точками перегиба, которые расположены на равных расстояниях по обе стороны от моды. Это значит, что закона Ципфа в виде формулы (3) **не существует**!!! По данным автора этого «закона» *к = 0.1, γ=1; r* – ранг журнала, т. е. его порядковый номер от начала частотного списка; $\mathrm{p_r}$ – доля статей из общего их числа (по данной тематике) в журнале с рангом *r*. Тогда накопленная доля статей в *r* первых журналах будет равна (при *γ=1*)

$$F(r) = \sum_{i=1}^{r} p_i = \sum_{i=1}^{r} \frac{k}{i} \approx k(\ln r + C). \qquad (4)$$

где С – постоянная Эйлера (С=0,577216).

Выражение (4) по форме совпадает с формулировкой закона рассеяния публикаций С. Бредфорда в ранговой интегральной форме

$$X(r) = a + b \log r, \qquad (5)$$

где *X(r)* – накопленное число статей в *r* первых журналах; *a, b* – параметры.

Таким образом, закон Бредфорда представляет собой не что иное, как закон Ципфа, примененный к периодическим изданиям. Другими словами, закон рассеяния в общем случае задается ранговым распределением. Отсюда следует вывод: для отыскания точной формулировки закона рассеяния публикаций

необходимо найти такое теоретическое распределение, которое с высокой точностью описывает все разнообразие статистических ранговых распределений.

С. Бредфорд проиллюстрировал свой закон кривой рассеяния $X(r) = f(\ln r)$, которая на границе зоны ядра переходит в прямую (правда, неизвестно, как эту границу вычислять), и сформулировал его в виде (цитируется по [2, с. 93]): «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 $1 : N : N^2$». В исследованиях Бредфорда величина $N \approx 5$.

Последователи Бредфорда понимали, что закон рассеяния публикаций нуждается в уточнении, но не могли найти то ранговое распределение, которое могло бы с высокой точностью описывать статистические ранговые распределения журналов. Поэтому их «уточнения» не вносили ничего нового в закон рассеяния или даже искажали его.

Так, Б. Викери [10] предложил записать модель Бредфорда в виде $T_x : T_{2x} : T_{3x:} = 1 : N : N^2$, где T_x – число журналов, содержащих x статей по данному предмету, а величина N зависит от x. Однако формула Викери не уточняет модель Бредфорда.

Б. Брукс [8, 9], пытаясь учесть кривизну графика статистической зависимости $X(r) = f(\ln r)$, предложил две формулы:

$$X(r) = \alpha r^{\beta}, \quad X(r) = b \ln (r/s),$$

первая из которых описывает зону ядра, а вторая – зоны рассеяния. Но здесь также не приводятся формулы для расчета границ ядра и зон рассеяния.

А. Т. Мицевич аппроксимирует зависимость $X(r) = f(\ln r)$ кривой третьего порядка, поскольку «несмотря на поправки, внесенные в модель Бредфорда, она не отражает разнообразие изучаемых реальных распределений» [1, с. 5]. Но кубическая парабола не является законом распределения.

Из приведенных примеров видно, насколько беден набор моделей, которыми многие исследователи пытались аппрокси-

мировать распределение Бредфорда. В монографии С. Д. Хайтуна [7] приводится большое количество примеров аппроксимации распределения Бредфорда и большинство из них связано с законом Ципфа. Это одна из причин, не позволившая получить математически точную формулировку закона рассеяния.

Вторая причина заключается в том, что большинство исследователей вслед за Бредфордом принимали число статей в ядре журналов и зонах рассеяния одинаковым, что не согласуется с опытными данными. Так, Т. Викери предлагал вводить любое количество зон рассеяния с равным числом статей в ядре и зонах рассеяния. Ясно, что такой модели не соответствует ни одно теоретическое распределение.

Распределение Вейбулла как частный случай закона рассеяния публикаций

Для описания статистических ранговых распределений необходимо использовать теоретические законы с убывающей плотностью, заданной на положительной полуоси, поскольку ранг журнала – положительное число.

Одним из простейших и подходящих законов для описания статистических ранговых распределений является закон Вейбулла, который впервые использовал профессор Г. Г. Белоногов [1] для описания рангового распределения слов частотного словаря. Функция распределения и плотность вероятности этого закона задаются формулами $F(t) = 1 - e^{-\alpha t^{\beta}}$; $p(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}$. При $\beta<1$ плотность $p(t)$ с ростом t убывает. Величина t в данном случае может обозначать ранг журнала; $F(t)$ – накопленную долю статей по заданной теме в t первых журналах; плотность $p(t)$ численно равна доле статей (из общего их числа) в журнале с рангом t.

Закон Вейбулла (так же, как и закон Ципфа) является частным случаем более общего закона рассеяния публикаций (2). Исследования показали, что закон Вейбулла в отличие от закона Ципфа может с высокой точностью описывать, по крайней мере, многие статистические ранговые распределения. Это позволяет выполнять все необходимые расчеты, касающееся рассеяния публикаций. Более того, на основе закона Вейбулла можно получить математически точную формулировку закона рассеяния публикаций в смысле Бредфорда.

Кривая распределения Вейбулла при $0 < \beta < 1$ является убывающей и не имеет никаких особых точек, которые можно было бы использовать в качестве границ ядра и зон рассеяния. Для нахождения этих границ преобразуем ранговое распределение Вейбулла в форме *tp(t) = f(lnt):*

$$tp(t) = \frac{\alpha\beta e^{\beta \ln t}}{e^{\alpha e^{\beta \ln t}}}.$$

Произведение *tp(t)* представляет собой плотность распределения случайной величины *lnT*.

Приняв обозначения *lnt = x, tp(t) = p(lnt) = p(x),* последнюю формулу перепишем в виде

$$p(x) = \frac{\alpha\beta e^{\beta x}}{e^{\alpha e^{\beta x}}}.$$

Приведенная плотность *p(x)* является частным случаем обобщенной плотности (1) при $u \to 0$, к=1 и обладает тем замечательным свойством, что кривая распределения (т. е. график плотности *p(x)*) имеет три характерные точки: моду *C* и две точки перегиба *A* и *B* с абсциссами соответственно x_A , x_C, x_B.

График функции распределения, представленный в полулогарифмическом масштабе, имеет точку перегиба *C* (на кривой распределения она соответствует моде) и две точки, в которых третья производная равна нулю. Они соответствуют точкам перегиба на кривой распределения (в этих точках вторая производная равна нулю). Точки перегиба находятся на равных расстояниях от моды, т. е. $x_C - x_A = x_B - x_C$. Переходя к плотности p(t), с учетом равенства $x = lnt$ можем записать

$lnt_C - lnt_A = lnt_B - lnt_C$, откуда имеем $\frac{t_C}{t_A} = \frac{t_B}{t_C} = n,$

из последней формулы следует математически точная формулировка закона рассеяния в смысле Бредфорда

$$t_A : t_C : t_B = t_A(1 : n : n^2). \qquad (6)$$

Эта формула отличается от закона Бредфорда тем, что величины t_A , t_C, t_B в левой ее части обозначают число журналов от

начала частотного списка соответственно до точек *A, C, B*, а не в ядре и зонах рассеяния.

Если по закону Вейбулла рассчитать число журналов в ядре и зонах рассеяния, то соответствующая формула будет иметь вид

$$t_A : t_I : t_{II} = t_A(1 : (n\text{-}1) : (n\text{-}1)n). \qquad (7)$$

Здесь t_A - число журналов в ядре; t_I, t_{II} – число журналов соответственно в первой и второй зонах рассеяния.

Формулировка закона рассеяния в смысле Бредфорда в виде формул (6), (7) была предложена мною совместно с Б. В. Петренко еще в 1974 г. [6].

Закон Вейбулла позволяет получить формулы для расчета координат точек *A, C, B*, т. е. для расчета границ ядра и зон рассеяния и доли статей в них. Эти формулы имеют вид

$$t_C = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}}; \quad t_A = \frac{t_C}{n}; \quad t_B = t_C \cdot n; \quad n = \left(\frac{3+\sqrt{5}}{2}\right)^{\frac{1}{\beta}}. \qquad (8)$$

$$F(t_A) = 0{,}3175;\ F(t_C) = 0{,}6321;\ F(t_B) = 0{,}9271. \qquad (9)$$

Итак, в случае справедливости закона Вейбулла в ядро журналов входит примерно 32 % от всех статей по данному предмету, в ядро и первую зону – 63 %, а в ядро и первые две зоны – 93 %. Следовательно, на первую зону приходится 31 % статей, на вторую – 30 %, а на третью лишь 7 % статей. Как видно из приведенных формул, они зависят от двух параметров закона Вейбулла – α и β.

Величина t_B может характеризировать оптимальный объем фонда по заданному профилю с точки зрения полноты комплектования. Чтобы увеличить полноту комплектования статей, например, на 5 % выше оптимального значения $F(t_B) \approx 0{,}93$, необходимо увеличить количество наименований журналов примерно в 2 раза (при $\alpha = 0{,}1$; $\beta = 0{,}5$), в то время как уменьшение полноты комплектования по статьям на 5 % ниже значения $F(t_B)$ приводит к уменьшению количества наименований журналов в 1,5 раза.

Отсюда следует, что для более полного удовлетворения информационных потребностей специалистов в справочно-информационном фонде должны комплектоваться по крайней

мере те журналы по данному профилю, которые образуют ядро и первые две зоны рассеяния, где содержится около 93 % статей.

Закон Вейбулла позволяет также находить величину t при любых значениях $F(t)$

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

Таким образом, распределение Вейбулла позволило получить математически точную формулировку закона рассеяния в смысле Бредфорда, а также формулы для вычисления доли статей в ядре и зонах рассеяния. Однако это распределение представляет собой частный случай некоторого более общего семейства распределений. Если ранговое распределение журналов не подчиняется закону Вейбулла, то для вычисления границ ядра, зон рассеяния и доли статей в них полученные формулы оказываются непригодными. Вот почему усилия многих исследователей не привели к нахождению общих формул для вычисления границ ядра и зон рассеяния, а также вычисления доли статей в них. Для этого требовалась разработка универсальных распределений, которые и были успешно разработаны автором настоящей статьи [4].

1. Белоногов, Г. Г. О некоторых статистических закономерностях в русской письменной речи / Г. Белоногов // Вопросы языкознания. – 1962. – № 1 – С. 100–101.

2. Михайлов, А. И. Основы информатики / А. И. Михайлов, А. И. Черный, Р. С. Гиляревский. – М. : Наука, 1968. – 756 с.

3. Мицевич, Т. А. Исследование структуры потоков научно-технической информации по машиностроению / Т. А. Мицевич // НТИ. Серия 2. – 1975. – № 5. – С. 3–16.

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

5. Нешитой, В. В. Универсальные законы рассеяния и старения публикаций / В. В. Нешитой // Весн. Беларус. дзярж. ун-та культуры і мастацтваў. – 2007. – № 8. – С. 128–133.

6. Петренко, Б. В. Применение закона Вейбулла для расчета полноты комплектования справочно-информационного фонда / Б. В. Петренко, В. В. Нешитой // Проблемы оптимального комплектования и использования справочно-информационного фонда для принятия решений / Общество «Знание» Украинской ССР. – Киев, 1974. – С. 6–8.

7. Хайтун, С. Д. Наукометрия. Состояние и перспективы / С. Д. Хайтун. – М. : Наука, 1983. – 344 с.

8. Bradford, S. C. Documentation / S. C. Bradford. – London : Crosly Lockwood, 1948. – 156 p.

9. Brookes, B. C. The Derivation and Application of the Bradford-Zipf Distribution / B. C. Brookes // Journal of Documentation. – 1968. – V. 24. – № 4. – P. 247–265.

10. Brookes, B. C. Bradford's law and the bibliography of science / B. C. Brookes // Nature. – 1969. – № 9. – P. 953–956.

11. Vickeri, B. C. Bradford's law of scattering / B. C. Vickeri // Journal of Documentation. – 1948. – V. 4. – № 3. – P. 198–203.

12. Zipf, G. K. Human behaviour and the principle of least effort / G. K. Zipf. – Cambridge, Mass. : Addison-Wesley, 1949. – 573 p.

Т. Д. Орешко,
старший преподаватель кафедры
информационных технологий в культуре

ФОРМИРОВАНИЕ ИНФОРМАЦИОННОЙ КУЛЬТУРЫ – НЕОБХОДИМЫЙ КОМПОНЕНТ ПОДГОТОВКИ СПЕЦИАЛИСТОВ-КУЛЬТУРОЛОГОВ

Информационная культура является важным аспектом развития информационного общества. Сам термин «информационная культура» состоит из двух компонентов, каждый из которых играет роль в развитии и становлении личности и может рассматриваться с нескольких точек зрения: социологической и технологической.

Слово «информация» происходит от латинского слова informatio, что в переводе означает сведение, разъяснение, ознакомление. Данное понятие используется в различных науках, при этом в каждой науке понятие «информация» связано с различными системами понятий. К фундаментальным свойствам информации относятся новизна, актуальность, достоверность, объективность, полнота, ценность и др. К. К. Колин отмечал, что «…роль информации при изучении как природных явлений, так и социальных процессов является определяющей и ранее явно недооценивалась. … Информация является таким же фундаментальным и всеобщим свойством мироздания, как вещество и энергия» [1]. Определяющая роль в увеличении значимости информации во всех процессах жизнедеятельности