Нешитой Василий Васильевич,
заведующий кафедрой
Белорусского государственного университета культуры и искусств, д.т.н.
Петренко Борис Васильевич,
доцент Белорусского государственного университета культуры
и искусств, д.т.н.

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ИСПОЛЬЗОВАНИЯ БИБЛИОТЕЧНОГО ФОНДА

Современные библиотеки располагают огромными фондами на традиционных носителях. Их объемы составляют десятки миллионов единиц хранения. В таком множестве документов по различным причинам значительная доля фонда в течение длительного периода времени может быть не использована. Это равносильно потере полезной информации, содержащейся в неиспользованной части фонда, часто приводит к необходимости проведения повторных исследований, дополнительным материальным затратам, труда и времени высококвалифицированных специалистов.

Для анализа степени использования библиотечного фонда и выявления неиспользуемой части целесообразно применять статистические методы, в частности, метод статистического моделирования.

Одной из главных задач этого метода является установление законов распределения отдельных видов документов, упорядоченных по уменьшению количества обращений к ним читателей.

Список документов с указанием ранга каждого наименования документа и частоты обращения к нему представляет собой статистическое ранговое распределение. Оно содержит в себе ценную

информацию об использовании фонда, которую необходимо каким-то способом извлечь. Математическая статистика решает эту задачу путем установления такого теоретического закона распределения, который наиболее точно описывает (аппроксимирует) статистическое распределение, в том числе ранговое. Задача эта весьма сложная и при попытках ее решения используются различные законы распределения, методы нахождения оценок параметров этих законов, критерии согласия, устанавливающие степень близости статистического и аппроксимирующего распределений и т. д.

Необходимость установления наиболее подходящего аппроксимирующего распределения диктуется тем, что закон распределения является наиболее полной характеристикой случайной величины. Если будет установлен такой закон, который с высокой точностью описывает статистическое распределение, то в нем отразятся все основные свойства последнего. Следовательно, о статистическом распределении можно будет узнать много полезной информации, в том числе такой, которую нельзя получить другими способами: выделение ядра библиотечного фонда, зон рассеяния, вычисление информационной полноты его комплектования и решение других задач.

Чтобы решить перечисленные задачи, необходимо иметь следующую информацию:

– статистические данные о частоте использования каждого наименования документа по видам (книги, журналы и т. д.);

– математические модели для описания ранговых распределений;

– методы установления наилучшего аппроксимирующего распределения;

– методы оценивания параметров;

– методы вычисления ядра библиотечного фонда, зон рассеяния и других характеристик.

Рассмотрим кратко каждый пункт.

1. Статистические данные о частоте использования документов

Для быстрого построения статистического распределения книг, журналов и других видов документов необходимо в составе автоматизированной библиотечно-информационной системы (АБИС) наличие соответствующего автоматизированного рабочего места (АРМ). Оно должно обеспечить выполнение в автоматизированном режиме следующих операций:

– выделение любого вида документа по любой заданной тематической рубрике за указанный период времени (при необходимости – заданных границах годов издания);

– подсчет частоты выдачи каждого наименования документа;

– сортировка документов по убыванию их выдачи;

– подсчет невыданных документов по данной тематической группе.

К сожалению, современные АБИС не решают всех этих задач. Поэтому сбор статистических данных по использованию библиотечного фонда может осуществляться лишь в полуавтоматизированном режиме, что связано с большими трудовыми затратами.

Если сравнить количество книговыдач с количеством словоупотреблений в некотором тексте, то количество разных наименований книг будет соответствовать количеству разных слов в тексте. Это значит, что моделирование и статистический анализ библиотечного фонда и текста могут осуществляться с помощью одних и тех же моделей и методов. Но если для построения частотных словарей существуют компьютерные программы, в том числе свободно распространяемые, то для составления частотных списков документов библиотечного фонда таких программ не существует.

В литературных источниках не приводятся частотные списки книг, в то время как частотных словарей имеется великое множество. Это

позволяет развивать математическое и программное обеспечение в областях информатики и математической лингвистики. В библиотечном деле этот вопрос, как правило, обходится стороной. А без наличия статистики использования каждого документа библиотечного фонда нельзя точно узнать ни информационную полноту его комплектования, ни эффективность его использования. При этом многие документы могут быть просто затеряны в огромном фонде современной библиотеки.

2. Математические модели для описания ранговых распределений

В математической лингвистике и информатике было предложено множество моделей для описания статистических распределений, в том числе ранговых которые могут быть использованы для моделирования библиотечного фонда. Одной из первых моделей для описания ранговых распределениях слов частотного словаря был закон Ципфа, устанавливающий в первом приближении зависимость между рангом слова (r) и его относительной частотой (p_r). Этот закон задается формулой $p_r = k/r$, которая содержит один параметр k. Естественно, что распределение с одним параметром не может с достаточной точностью описывать все разнообразие статистических ранговых распределений. Поэтому необходимо использовать многопараметрические распределения, включающие как частные случаи большинство известных непрерывных распределений, способных с высокой точностью описывать статистические распределения, в том числе ранговые.

Для этих целей наиболее подходящими оказываются обобщенные распределения, образующие в совокупности вторую систему непрерывных распределений В. Нешитого [1]. Они задаются тремя плотностями:

$$p(t) = Nt^{k\beta-1}(1-\alpha u t^{\beta})^{\frac{1}{u}-1};$$
$$p(\mathrm{y}) = \frac{N}{y}(\ln y - l)^{k-1}\left[1-\alpha u(\ln y - l)\right]^{\frac{1}{u}-1}; \qquad (1)$$
$$p(y) = \frac{N}{y}\left[1-\alpha u(\ln y - \overline{\ln y})^{2}\right]^{\frac{1}{u}-1}.$$

Здесь *t, y* – значения случайных величин *T, Y*. В случае ранговых распределений они обозначают ранг книги, журнала и т.д. в ранжированном ряду. Книги должны быть упорядочены по убыванию частоты их использования. Приведенные плотности содержат параметры $\alpha, \beta, \overline{\ln y}$, *k, u, l*. Их значения вычисляются по статистическому ранговому распределению. Величина $\overline{\ln y}$ представляет собой среднее значение логарифма случайной величины (ранга) *Y*.

Система распределений, заданная плотностями (1), имеет широчайшие аппроксимирующие возможности. Это позволяет ей описывать широкое разнообразие статистических распределений, в том числе ранговых, с высокой точностью. Другими словами, она позволяет вычислять закон распределения случайной величины по ее статистическому распределению. Приведенные плотности содержат как частный случай закон Ципфа. Однако многолетний опыт автора теории обобщенных распределений по обработке статистических ранговых распределений показал, что во всех случаях закон Ципфа оказался неподходящим для описания статистических ранговых распределений.

Для использования формул (1) разработаны методы вычисления типа аппроксимирующей кривой распределения, а также оценок параметров распределений.

Здесь следует отметить, что на практике довольно часто оказывается подходящим более простое распределение с двумя параметрами. Это распределение Вейбулла [2]. Его функция распределения, т.е. накопленная относительная частота, и плотность задаются формулами:

$$F(t) = 1 - e^{-\alpha t^{\beta}}, \qquad (2)$$

$$p(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}}. \qquad (3)$$

Плотность (3) входит как частный случай в обобщенную плотность *p(t)* системы распределений (1) при значениях параметров $u \to 0$, $k=1$.

3. Методы вычисления закона распределения

Для вычисления теоретического закона распределения в общем случае при наличии полных статистических данных целесообразно использовать устойчивый метод В. Нешитого. Суть его заключается в том, что по статистическому распределению вычисляются два показателя: асимметрии B^* и островершинности H^*, которые приравниваются к соответствующим теоретическим показателям B, H. Последние зависят от двух параметров формы k, u. По этим двум показателям легко установить тип аппроксимирующего распределения и найти в первом приближении оценки параметров k, u по заранее построенной бинарной сетке (номограмме) [1].

В частном случае, когда статистических данных недостаточно, целесообразно проверить применимость закона Вейбулла. Для этого функцию распределения (3) необходимо преобразовать к уравнению прямой:

$$\ln\ln\frac{1}{1-F(t)} = \ln\alpha + \beta\ln t. \qquad (4)$$

Если ввести новые обозначения:

$Y = \ln\ln\frac{1}{1-F(t)}$, $A = \ln\alpha$, $\ln t = X$, то формула (4) примет вид:

$$Y = A + \beta X. \qquad (4')$$

Далее необходимо по статистическим данным t, $F(t)$ вычислить величины X^*, Y^* и построить график зависимости $Y^* = \varphi(X^*)$. Если при

этом эмпирические точки ложатся вдоль теоретической прямой (4'), то закон Вейбулла может быть принят в качестве теоретического аппроксимирующего распределения.

4. Методы оценивания параметров

4.1. Устойчивый метод

В этом случае при известных эмпирических значениях показателей *B, H* по соответствующей бинарной сетке (номограмме) устанавливается тип аппроксимирующего распределения и находятся в первом приближении оценки параметров формы *k, u*. Оценки остальных параметров α, β вычисляются по специальным формулам. Более точные значения параметров могут быть вычислены по программе [1, с. 196–198].

4.2. Метод наименьших квадратов

Этот метод предложен К.Ф. Гауссом в 1795 г. Суть его сводится к тому, чтобы сумма квадратов отклонений эмпирических ординат от теоретических была наименьшей.

В случае распределения Вейбулла его оценки параметров по методу наименьших квадратов будут равны:

$$\beta = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - (\overline{X})^2}; \qquad (5)$$

$$\alpha = e^{\overline{Y} - \beta \overline{X}}. \qquad (6)$$

Здесь $\overline{X}, \overline{Y}$ – средние значения величин *X, Y,* которые рассчитываются по статистическому распределению; $\overline{X^2}$ – среднее квадрата случайной величины *X.* Подставляя далее оценки параметров α, β в формулы (2) и (3) можно вычислить теоретические значения плотности *p(t)* и функции распределения $F(t)$ и сравнить их с эмпирическими.

5. Вычисление ядра библиотечного фонда и других характеристик

Если ранговое распределение представить в системе координат $(r; p_r)$, то получим убывающую кривую, на которой нельзя выделить никаких характерных точек. Чтобы извлечь максимум полезной информации из статистического рангового распределения, его необходимо представить в виде зависимости $rp_r = f(\ln r)$, т. е. по горизонтальной оси откладываются логарифмы рангов (книг, журналов), а по вертикальной – произведения рангов на относительные частоты выдач [3]. Тогда получится кривая распределения с тремя характерными точками: модой $\ln r_C$ и двумя точками перегиба $\ln r_A$ и $\ln r_B$, которые расположены на равных расстояниях от моды $\ln r_C$ – и в этом, по нашему мнению, состоит суть закона рассеяния в смысле Бредфорда! Примем эти точки в качестве границ ядра и зон рассеяния [4].

Если статистическое распределение описывается законом Вейбулла, то мода t_C находится из условия $dtp(t)/d\ln t = 0$:

$$t_C = \left(\frac{1}{\alpha}\right)^{\frac{1}{\beta}} . \qquad (7)$$

Величина n, равная отношению $n = \frac{t_C}{t_A} = \frac{t_B}{t_C}$, вычисляется по формуле:

$$n = \left(\frac{3+\sqrt{5}}{2}\right)^{\frac{1}{\beta}} . \qquad (8)$$

Тогда абсциссы точек A и B равны

$$t_A = t_C / n; \quad t_B = t_C n . \qquad (9)$$

Функция распределения, т.е. накопленная доля выдач, приходящаяся на t_A книг, равна $F(t_A) = 0{,}3175$, т.е. приблизительно 0,32 (32%). Далее имеем: $F(t_C) = 0{,}6321$; $F(t_B) = 0{,}9271$ [4].

Таким образом, ядро книжного фонда (или журналов) составляет t_A разных наименований книг (или журналов). На него приходится $F(t_A)$=0,3175 книговыдач (или статей) от общего их количества (в случае закона Вейбулла).

Аналогично на ядро и первую зону рассеяния приходится t_C книг и $F(t_C)$=0,6321 книговыдач; на ядро и первые две зоны рассеяния – t_B книг и $F(t_B) = 0{,}9271$книговыдач. Остальные книги ($t > t_B$) относятся к третьей зоне рассеяния. Доля книговыдач, приходящаяся на них, составляет 0,0729.

Величину t_B можно принять в качестве оптимального объема фонда. Информационная полнота его составляет $F(t_B) = 0{,}9271$. Это значит, что такой фонд удовлетворяет информационные потребности пользователей на 92,7%. Для удовлетворения оставшейся доли информационных потребностей (0,073) требуется дополнительный фонд, по объему в несколько раз превышающий оптимальный объем t_B.

6. Анализ результатов статистической обработки книговыдач

Итак, нами вычислены границы ядра библиотечного фонда и зон рассеяния. Теперь необходимо в статистическом ранговом распределении книг отметить эти границы и исследовать все зоны, учитывая, какие книги попали в каждую зону и частоту выдач в начале и конце каждой зоны.

Особенно тщательно необходимо исследовать третью зону рассеяния, поскольку здесь содержатся редко используемые книги. Конечно, это могут быть устаревшие книги, но могут быть и такие, которые содержат новые открытия, еще не понятые и не принятые мировой общественностью. Другими словами, это зона новизны.

Чтобы ценная информация из третьей зоны не затерялась, книги, ее содержащие, должны быть представлены в электронной форме, тщательно обработаны, индексированы по всем основным аспектам.

При этом следует учитывать, что третья зона, как правило, самая обширная, и далеко не все документы оказываются запрошенными хотя бы один раз в течение достаточно длительного времени. На долю третьей зоны может приходиться до 80% книжного фонда, но число книговыдач не превышает 7–10%. Естественно, что выявить все документы третьей зоны, в том числе невыданные, можно лишь с помощью автоматизированной библиотечно-информационной системы. Автоматизированное рабочее место «Статистика» должно обеспечивать выявление документов в ядре фонда и каждой зоне рассеяния.

ЛИТЕРАТУРА

1. 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.

2. 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. – Киев, 1974. – С. 6–8.

3. Нешитой, В. В. Форма представления ранговых распределений / В. В. Нешитой // Учёные записки Тартуского гос. ун-та. – 1987.– Вып. 774. – С. 123–134.

4. Нешитой, В.В. Модели рассеяния публикаций / В. В. Нешитой, Б. В. Петренко // Библиотечный вестник. – 2010. – Вып. 2. – С. 68–74.