Библиографический список

1. Лебедев, А.В. Максимумы рекуррентных случайных последовательностей [Текст] // Вестник МГУ. – Сер. 1. Матем.-Мех. – 2001. – № 1. – С. 10–14.
2. Лебедев, А.В. Максимумы рекуррентных случайных последовательностей. Случай тяжелых хвостов [Текст] // Вестник МГУ. – Сер. 1. Матем.-Мех. – 2001. – № 3.– С. 63–66.
3. De Haan L., Resnick S.I., Rootzen H., de Vries G.C. Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. // Stoch. Proc. Appl. 1989. V. 32. № 1. P. 213–224.
4. Kozubowski T.J., Podgorski K. Log-Laplace distributions // Int. Math. J. 2003. V. 3. № 4. P. 467–495.
5. Embrechts P., Kluppelberg C.P., Mikosh T. Modelling extremal events for insurance and finance. Springer, 2003.
6. Новицкая, О.С., Яцало, Е.Б. Экстремумы рекуррентных случайных последовательностей // Вестник МГУ. – Сер. 1. Матем.-Мех. – 2008. № 5, С. 6-10.
7. Alpuim M.T., Catkan N.A., Husler J. Extremes and clustering of nonstationary max-AR(1) sequences. // Stoch. Proc. Appl. 1995. V. 56. № 1. P. 174–184.
8. Лебедев, А.В. Степенные хвосты и кластеры в линейных рекуррентных случайных последовательностях // Труды VI Колмогоровских чтений. – Ярославль: Изд-во ЯГПУ, – 2008 – С. 126–130.

А.В. Лебедев

МОДЕЛЬ ТОПА НОВОСТЕЙ НА ОСНОВЕ ЭКСТРЕМАЛЬНОГО ДРОБОВОГО ШУМА

1. Введение. В современных средствах массовой информации и сети Интернет часто составляют списки наиболее популярных новостей в порядке убывания их популярности. Такой список называют "топом". Топ-k состоит из k наиболее популярных новостей (например, топ-10). Отметим, что аналогичные списки составляют из песен, музыкальных групп, знаменитостей и т.п.; их называют также чартами, рейтингами и др. Популярность чего-либо в общем случае может оцениваться интуитивно (по мнению экспертов) или статистически (по частоте упоминаний, результатам голосований и др.). Будем считать, что она выражается произвольной неотрицательной величиной.

Рассмотрим следующую модель. Пусть новости поступают пуассоновским потоком с интенсивностью λ в случайные моменты $t_1, \dots t_n, \dots$ и популярность n новости в момент времени $t > 0$ описывается случайным процессом $\eta_n = \xi_n f(t-t_n)$, где ξ_n, $n \geq 1$ — независимые неотрицательные случайные величины с одинаковым непрерывным распределением F, $f(t)$ — неотрицательная функция, непрерывная справа, интегрируемая по Риману на любом отрезке $[0, T]$ и равная нулю при $t < 0$.

В наборе случайных величин $\{\eta_n(t)\}_{n=1}^{\infty}$ в каждый момент отлично от нуля лишь конечное их число (почти наверное). Поэтому для данного набора можно определить максимум $M(t)$ и k максимумы $M^{(k)}(t)$ (т.е. значения k в порядке убывания) [1, гл. 2]. Эти величины описывают текущие популярности новостей в топе.

Отметим, что $M(t)$ представляет собой процесс экстремального дробового шума. Этот термин употребляется в отношении процессов, получаемых из классического (аддитивного) дробового шума заменой операции суммирования на максимум или минимум [2]. Процесы $M^{(k)}(t)$ также можно рассматривать как модификации дробового шума с заменой суммы на экстремальные порядковые статистики.

2. Одномерные распределения популярностей. Введем сумму

$$S(t, u) = \sum_{n=1}^{\infty} \mathbf{I}(\eta_n(t) > u), \quad t, u > 0,$$

[1]Работа выполнена при поддержке РФФИ, проекты № 07-01-00077, № 07-01-00373.

которая описывает, популярность скольких новостей в момент t превышает уровень u, и содержит лишь конечное число слагаемых (почти наверное).

Лемма 1. *Случайная величина $S(t,u)$ имеет пуассоновское распределение с параметром*

$$\mu(t,u) = \lambda \int_0^t \bar{F}\left(\frac{u}{f(s)}\right) ds.$$

Доказательство. Воспользуемся известным свойством пуассоновского потока: при условии, что число точек на отрезке $[0,t]$ известно, их можно считать независимо и равномерно распределенными на этом отрезке (без учета порядка). Новость, поступившая в момент $t-s$, $s \in [0,t]$, дает единичный вклад в сумму с вероятностью $\bar{F}(u/f(s))$. Усредняя по моменту поступления, получаем, что каждая точка, независимо от других, дает вклад с вероятностью

$$r_t(u) = \frac{1}{t} \int_0^t \bar{F}\left(\frac{u}{f(s)}\right) ds.$$

Переходя к производящим функциям, получаем

$$\mathrm{M} z^{S(t,u)} = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} (r_t(u)z + (1 - r_t(u)))^n = \exp\{\lambda t r_t(u)(z-1)\},$$

что как раз соответствует пуассоновскому распределению с нужным параметром.

Теорема 1.

$$\mathrm{P}(M(t) \le u) = e^{-\mu(t,u)}, \quad \mathrm{P}(M^{(k)}(t) \le u) = \sum_{l=0}^{k-1} \frac{\mu(t,u)^l}{l!} e^{-\mu(t,u)}.$$

Доказательство. Используем естественные соотношения:

$$\mathrm{P}(M(t) \le u) = \mathrm{P}(S(t,u) = 0); \quad \mathrm{P}(M^{(k)}(t) \le u) = \mathrm{P}(S(t,u) \le k-1).$$

Подобный метод применялся в [1, §2.2] для исследования максимумов независимых одинаково распределенных случайных величин.

Следствие 1. *Пусть интеграл*

$$I(u) = \int_0^{\infty} \bar{F}\left(\frac{u}{f(t)}\right) dt$$

сходится при $u > u_0 \ge 0$ и $\tilde{\mu}(u) = \lambda I(u)$, тогда

$$\lim_{t\to\infty} \mathrm{P}(M(t) \le u) = e^{-\tilde{\mu}(u)}, \quad \lim_{t\to\infty} \mathrm{P}(M^{(k)}(t) \le u) = \sum_{l=0}^{k-1} \frac{\tilde{\mu}(u)^l}{l!} e^{-\tilde{\mu}(u)}$$

при $u > u_0$.

В данном случае предельный переход очевиден. Заметим, что если существует $u_0 > 0$ такое, что $I(u)$ расходится при всех $u < u_0$, то $\mu(t,u) \to +\infty$, $t \to \infty$, так что $\mathrm{P}(M(t) \le u) \to 0$, $\mathrm{P}(M^{(k)}(t) \le u) \to 0$, т.е. предельные функции распределения равны нулю при $u < u_0$. Таким образом, следствием можно пользоваться и в этом случае, формально полагая $\tilde{\mu}(u) = +\infty$ и $\tilde{\mu}(u)^l e^{-\tilde{\mu}(u)} = 0$, $l \ge 0$. Далее будем предполагать, что условие следствия 1 выполнено, т.е. интеграл сходится при всех достаточно больших u, а значит, предельные распределения существуют.

Обозначим предельные распределения $M(t)$ и $M^{(k)}(t)$ при $t \to \infty$ через Ψ и $\Psi^{(k)}$ и введем случайные величины с такими распределениями $\tilde{M}$, $\tilde{M}^{(k)}$, а также независимую от них величину ξ с распределением F.

2. Многомерные распределения популярностей. Рассмотрим $S(t,u)$ как случайный процесс по $u > 0$ при фиксированном $t > 0$.

Теорема 2. *Пусть заданы числа $u_1 > u_2 > \ldots > u_m > 0$, $m \geq 2$. Тогда случайные величины $S(t,u_1)$, $S(t,u_2) - S(t,u_1)$, ..., $S(t,u_m) - S(t,u_{m-1})$ независимы и имеют пуассоновские распределения с параметрами $\mu(t,u_1)$, $\mu(t,u_2) - \mu(t,u_1)$,..., $\mu(t,u_m) - \mu(t,u_{m-1})$ соответственно.*

Обозначим указанные параметры через $\Delta\mu_i(t)$, $1 \leq i \leq m$.

Доказательство. Используем те же соображения, что и при доказательстве теоремы 1. Вклад отдельно взятой новости, поступившей в равномерно распределенный на отрезке $[0,t]$ момент времени, в вектор $(S(t,u_1), S(t,u_2) - S(t,u_1), \ldots, S(t,u_m) - S(t,u_{m-1}))$ принимает значения $(1,0,\ldots,0)$, $(0,1,\ldots,0)$,..., $(0,\ldots,0,1)$ с вероятностями $r_t(u_1)$, $r_t(u_2) - r_t(u_1)$,..., $r_t(u_m) - r_t(u_{m-1})$, которые обозначим через $\Delta r_{t,i}$, $1 \leq i \leq m$. Его производящая функция равна

$$\sum_{i=1}^{m} \Delta r_{t,i} z_i + (1 - r_t(u_m)) = \sum_{i=1}^{m} \Delta r_{t,i}(z_i - 1) + 1.$$

Таким образом, получаем

$$\mathrm{M} z_1^{S(t,u_1)} z_2^{S(t,u_2)-S(t,u_1)} \ldots z_m^{S(t,u_m)-S(t,u_{m-1})} = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \left(\sum_{i=1}^{m} \Delta r_{t,i}(z_i - 1) + 1 \right)^n = \exp\left\{ \sum_{i=1}^{m} \Delta\mu_i(t)(z_i - 1) \right\},$$

что и доказывает утверждение теоремы.

Таким образом, процесс $S(t,u)$ по $u > 0$ имеет независимые пуассоновские приращения (со знаком минус). Множество точек популярностей текущих новостей образует на оси $(0, +\infty)$ неоднородный пуассоновский поток.

Зная совместное распределение сумм $S(t,u_1)$,..., $S(t,u_m)$, можно найти совместное распределение k максимумов, используя соотношение

$$\begin{aligned} &\mathrm{P}(M(t) \leq u_1, M^{(2)}(t) \leq u_2, \ldots, M^{(m)}(t) \leq u_m) = \\ &= \mathrm{P}(S(t,u_1) = 0, S(t,u_2) \leq 1, \ldots, S(t,u_m) \leq m-1), \end{aligned}$$

однако в общем виде формулы оказываются очень громоздкими.

Рассмотрим случай m=2.

Следствие 2. *Для любых $u_1 > u_2 > 0$ верно*

$$\mathrm{P}(M(t) \leq u_1, M^{(2)}(t) \leq u_2) = e^{-\mu(t,u_2)}(\mu(t,u_2) - \mu(t,u_1) + 1).$$

Доказательство. Имеем

$$\begin{aligned} &\mathrm{P}(M(t) \leq u_1, M^{(2)}(t) \leq u_2) = \mathrm{P}(S(t,u_1) = 0, S(t,u_2) \leq 1) = \\ &= e^{-\mu(t,u_1)}(e^{-\Delta\mu_2(t)} + \Delta\mu_2(t)e^{-\Delta\mu_2(t)}) = e^{-\mu(t,u_2)}(\Delta\mu_2(t) + 1). \end{aligned}$$

Следствие 3. *Для любых $u_1 > u_2 > 0$ верно*

$$\begin{aligned} &\lim_{t\to\infty} \mathrm{P}(M(t) \leq u_1, M^{(2)}(t) \leq u_2) = \\ &= e^{-\tilde{\mu}(u_2)}(\tilde{\mu}(u_2) - \tilde{\mu}(u_1) + 1) = \Psi(u_2)(\ln\Psi(u_1) - \ln\Psi(u_2) + 1). \end{aligned}$$

3. Распределения мест и времен пребывания. Пусть время "раскрутки"новости от нуля до максимальной популярности мало по сравнению с общим временем, когда она представляет интерес (в отношении песен и знаменитостей это уже нельзя считать верным). Тогда можем предположить, что функция f имеет скачок в нуле и $f(0) = 1$, а далее функция не возрастает.

Предположим также, что система находится в стационарном (предельном) режиме, что эквивалентно ситуации, когда поток новостей начинается в $-\infty$. В этом случае можно найти вероятности того, что очередная новость в момент поступления:

1) займет место лидера

$$p_1 = \mathrm{P}(\tilde{M} < \xi) = \int_0^\infty e^{-\tilde{\mu}(u)}\, dF(u);$$

2) попадет в топ-k

$$P_k = \mathrm{P}(\tilde{M}^{(k)} < \xi) = \int_0^\infty \sum_{l=0}^{k-1} \frac{\tilde{\mu}(u)^l}{l!} e^{-\tilde{\mu}(u)}\, dF(u);$$

3) попадет на k место

$$p_k = P_k - P_{k-1} = \int_0^\infty \frac{\tilde{\mu}(u)^{k-1}}{(k-1)!} e^{-\tilde{\mu}(u)}\, dF(u).$$

Вероятности p_k, $k \geq 1$, описывают распределение случайного места ν, которое занимает очередная новость в момент поступления (без учета ограничений на размер топа).

Вычисление по указанным формулам в общем случае затруднительно. Предположим, что начальная популярность новости имеет распределение Парето $F(x) = 1 - x^{-\alpha}$, $x > 1$, $\alpha > 0$. Такое предположение, вообще говоря, находится в русле современных представлений о распространенности степенных хвостов в природе, технике и человеческой деятельности [3]. В данном случае интеграл $I(u)$ при $u \geq 1$ считается очень просто:

$$I(u) = \int_0^\infty \left(\frac{u}{f(t)}\right)^{-\alpha} dt = u^{-\alpha} \int_0^\infty f(t)^\alpha\, dt.$$

Введем обозначение

$$\rho = \lambda \int_0^\infty f(t)^\alpha\, dt,$$

тогда

$$\Psi(u) = e^{-\rho u^{-\alpha}}, \quad \Psi^{(k)}(u) = \sum_{l=0}^{k-1} \frac{(\rho u^{-\alpha})^l}{l!} e^{-\rho u^{-\alpha}}, \quad u \geq 1.$$

При $0 \leq u < 1$ имеем

$$I(u) = f^{-1}(u) + \int_{f^{-1}(u)}^\infty f(t)^\alpha\, dt.$$

Например, если $f(t) = e^{-\gamma t}$, $t \geq 0$, получаем $\rho = \lambda/(\alpha\gamma)$ и $\Psi(u) = e^{-\rho} u^\lambda$, $0 \leq u < 1$.

Однако для вычисления вероятностей p_k, $k \geq 1$, значения функций распределения при $0 \leq u < 1$ нам не нужны, поскольку в этой области $F(u) = 0$. Интегрируя, получаем

$$\begin{aligned}
p_1 &= \frac{1 - e^{-\rho}}{\rho}, \\
p_k &= \frac{1}{\rho} \int_0^\rho \frac{v^{k-1}}{(k-1)!} e^{-v}\, dv = \frac{1}{\rho}\left(1 - e^{-\rho} \sum_{l=0}^{k-1} \frac{\rho^l}{l!}\right), \\
P_k &= \sum_{j=1}^{k} p_j = \frac{1}{\rho}\left(k - e^{-\rho} \sum_{l=0}^{k-1} (k-l) \frac{\rho^l}{l!}\right).
\end{aligned}$$

В вычислительных целях вероятности p_k, $k \geq 2$ удобнее считать рекуррентно, по формуле

$$p_k = p_{k-1} - e^{-\rho}\frac{\rho^{k-2}}{(k-1)!}.$$

Примечательно, что все зависит только от одного параметра ρ.

Можно заметить, что p_k, $k \geq 1$, представляют собой рандомизированные пуассоновские вероятности (со сдвигом на единицу). А именно, случайное место новости ν можно представить как пуассоновскую случайную величину, параметр которой равномерно распределен на отрезке $[0, \rho]$, плюс единица. Отсюда, в частности, следует соотношение $\mathrm{M}\nu = \rho/2 + 1$.

Кроме того, имеет место асимптотика $p_k \sim 1/\rho$, $P_k \sim k/\rho$, $k \geq 1$, при $\rho \to \infty$. Отсюда следует, что условное распределение места новости в топе-m при условии ее попадания в этот топ асимптотически равномерно при $\rho \to \infty$, для любого $m > 1$.

В общем случае порядок новостей в топе может меняться и в периоды между поступлениями. А именно, если f убывает быстрее экспоненты, то "старые"новости могут опускаться вниз, а "свежие"подниматься вверх; если медленнее, возможен обратный процесс. Единственный случай, когда порядок новостей не меняется, это если f — показательная функция. В этом случае популярность всех новостей меняется со временем пропорционально.

Тогда легко исследовать времена, которые новости проводят на своих местах в топе. Новость может покинуть k место только в момент поступления очередной новости, с вероятностью P_k, независимо от предыдущих событий. Поэтому время пребывания на k месте имеет показательное распределение со средним $T_k = (\lambda P_k)^{-1}$. Среднее время пребывания в топе-m для новости, занявшей в момент поступления k место, $k \leq m$, получается равным $T_{k,m} = \lambda^{-1}(P_k^{-1} + \ldots + P_m^{-1})$.

Следующие таблицы рассчитаны при $\lambda = 1$, $\rho = 4$; здесь $\mathrm{M}\nu = 3$.

Вероятности попадания на k место:

k	1	2	3	4	5	6	7	8	9	10
p_k	0,245	0,227	0,19	0,142	0,093	0,054	0,028	0,013	0,005	0,002

Вероятности попадания в топ-k:

k	1	2	3	4	5	6	7	8	9	10
P_k	0,245	0,473	0,663	0,805	0,897	0,951	0,979	0,992	0,997	0,999

Средние времена пребывания на k месте:

k	1	2	3	4	5	6	7	8	9	10
T_k	4,075	2,116	1,508	1,243	1,114	1,051	1,022	1,008	1,003	1,001

Средние времена пребывания в топе-10 для занявших k-е место при поступлении:

k	1	2	3	4	5	6	7	8	9	10
$T_{k,10}$	15,142	11,067	8,951	7,443	6,2	5,086	4,034	3,013	2,004	1,001