Osiem i pół + ε

(czyli dodatek z ostatniej chwili do artykułu „Osiem i pół" (MSN nr 21))

Andrzej DĄBROWSKI, Wrocław

Na dwa tygodnie przed zamknięciem tego numeru MSN przypadkowo wpadł mi do ręki artykuł Shraggi Irmaya, z 1997 roku, opublikowany w niedostępnym, przynajmniej we Wrocławiu, czasopiśmie *Journal of Applied Statistics* [3]. Moją uwagę zwrócił tytuł „Związek pomiędzy prawem Zipfa i rozkładem pierwszej cyfry". Z obu zagadnieniami, wymienionymi w tytule pracy Irmaya spotkałem się w różnych okresach mojej działalności. O prawie Zipfa wspominałem w swojej książeczce o teorii informacji, wydanej 24 lata temu [2], zaś o rozkładzie pierwszej cyfry pisałem całkiem niedawno [1].

Rozpocznę od przedstawienia pojęć, występujących w pracy Irmaya w tak rozumianym porządku chronologicznym.

Prawo Zipfa i okolice

Profesor Uniwersytetu Harvarda G.K. Zipf w latach 30. badał częstość pojawiania się różnych słów w tekstach. Zwrócił uwagę na zadziwiającą, jego zdaniem, właściwość języków:

jeżeli uporządkujemy słowa w miarę zmniejszania się ich częstości pojawiania się w tekście, to częstość słowa o numerze n będzie w przybliżeniu równa $f_n = c/n$, gdzie stała c związana jest z liczbą różnych słów danego języka.

Czym jest słowo, badacz musi zdecydować. Jedni za różne uważają słowa, różne znaczeniowo (*koło (figura matematyczna)* i *koło (gospodyń wiejskich)*), ale utożsamiają słowa w różnych formach gramatycznych (*idzie*, utożsamiają z *poszedł*); inni uwzględniają tylko rzeczowniki i czasowniki w ich podstawowych formach słownikowych (a więc nie uwzględniają takich jednostek, jak *lub*, *więc*, *biały*) itp.

Zipf starał się objaśnić mechanizm używania słów i swoje prawo wydedukował z nazwanej przez siebie zasady najmniejszego wysiłku [5]. Wzór empiryczny Zipfa był jednak w wielu przypadkach bardzo niedokładny. Benoit Mandelbrot, 30 lat później, w latach 60., bazując na teorii informacji sformułował prawo Zipfa w nieco innej postaci:

częstość słowa o numerze n będzie w przybliżeniu równa $f_n = c/n^b$, gdzie b ($b \geq 1$) charakteryzuje bogactwo użytego słownictwa.

W tekstach, w których b jest znacząco większe od 1 używa się niewielkiego zasobu słów, częstość pojawiania się słów spoza tego wydzielonego repertuaru jest bardzo mała. Później Mandelbrot dołożył w swoim wzorze jeszcze jeden parametr, uzyskując zadowalającą elastyczność wzoru:

$$f_n = c/(n+a)^b.$$

Dla różnych języków dopasowywano wzór Mandelbrota i uzyskiwano ciekawe, ilościowe opisy bogactwa słów: w potocznym języku angielskim parametr b jest równy 1,1, w norweskim jest bardzo bliski 1, a w chińskim daleki od tej wartości. Język Jamesa Joyce'a w jego *Ulissesie* jest niezwykle bogaty (b praktycznie równe 1). Parametr b dla języka rozwijającego się dziecka maleje od wartości 1,6 do 1,15.

Wzór Mandelbrota można uzasadnić, przyjmując założenie, że długość wypowiadania (zapisu) danego słowa jest taka, aby najłatwiej można było je odnaleźć wśród wszystkich tekstów języka. System identyfikacji polega na dzieleniu zbioru tekstów na d części na każdym etapie identyfikacji. Optymalny system identyfikacji to taki, w którym oczekiwana (średnia) liczba etapów systemu identyfikacji jest najmniejsza. Twierdzenie 1.2 z [2] pokazuje, że taki optymalny system identyfikacji istnieje, gdy prawdopodobieństwa pojawienia się słowa o numerze n wynosi $p_n = d^{-ut_n}$, gdzie t_n jest długością wypowiadania n-tego słowa, a u jest stałą normującą. Z kolei, z badań eksperymentalnych wynika, że pomiędzy numerem słowa na liście słów uporządkowanych od najczęściej do najrzadziej występujących a długością jego wypowiadania zachodzi relacja logarytmiczna. Gdy przyjmiemy, że $t_n = v_1 + v_2 \log_d n$, to otrzymamy wzór Mandelbrota $p_n = c/n^b$.

Powrót do prawa Benforda

W artykułach [4] i [1] zajmowaliśmy się pokazaniem, jak można poprawnie zdefiniować odpowiedź na pytanie o prawdopodobieństwo, że liczba zaczyna się cyfrą p. Przyjęliśmy model liczb rzeczywistych jednorodny względem skali, oznaczający z grubsza, że prawdopodobieństwo zauważenia liczby α razy większej od tej, przy której stoimy zależy tylko od α. Otrzymane wzory można bez trudu uogólnić na dowolną podstawę liczenia.

Cyfra „p" w systemie liczenia o podstawie większej niż 10 jest równa *liczbie* p.

Prawdopodobieństwo, że liczba zacznie się cyfrą „p" w systemie liczenia o podstawie N wynosi

$$P(D_p|[1,\infty)) = \frac{p^{-q} - (p+1)^{-q}}{1 - N^{-q}}.$$

Uogólnienie prawa Benforda dla takiego systemu liczenia, czyli granica powyższego wzoru przy q zmierzającym do 0, przybierze postać

$$P(D_p|[u_0,\infty)) = \log_N\left(\frac{p+1}{p}\right).$$

Możemy w ten sposób obliczyć, na przykład, prawdopodobieństwo, że początkowymi cyframi pewnej liczby jest 314. Wystarczy bowiem wziąć $N = 1000$ i $p = 314$. Prawdopodobieństwo to wynosi (dla $q = 0$)

$$\log_N\left(\frac{p+1}{p}\right) = \log_1 000\left(\frac{315}{314}\right) \approx 0{,}00046.$$

Gdy $q > 0$, to rozwijając $P(D_p|[1,\infty))$ w szereg Taylora, otrzymamy, dla dość dużego a (chodzi o to, by $1/(p+a)$ było niedaleko od zera), biorąc wyrazy do pierwszego rzędu, wzór przybliżony

$$P(D_p|[1,\infty)) \approx \frac{q}{1 - N^{-q}} \frac{1}{(p+a)^{q+1}}.$$

Pamiętając o tym, że $\sum_{p=1}^{N} P(D_p|[1,\infty)) = 1$, wprowadzimy współczynnik c, taki że

$$P(D_p|[1,\infty)) \approx \frac{c}{(p+a)^{(q+1)}}$$

i taki, by

$$\sum_{p=1}^{N} \frac{c}{(p+a)^{q+1}} = 1.$$

Podobnie można postąpić dla $q = 0$ (prawo Benforda), otrzymując przybliżenie

$$P(D_p|[1,\infty)) = \frac{1/\log(N)}{p+a}$$

(można ten wzór otrzymać również przechodząc z q do 0 we wzorze przybliżonym).

Stała $1/\log(N)$ jest przybliżoną wartością prawdopodobieństwa, że wybierając losowo liczbę ze zbioru $\{1, 2, \ldots, N\}$ otrzymamy liczbę pierwszą.

Prawo Benforda a prawo Zipfa

Irmay w artykule [3] zauważył podobieństwo między prawem Benforda: $P(D_p|[1,\infty)) = \frac{1/\log(N)}{p+a}$ a prawem Zipfa: $f_n = c/n$.

Zaproponował on ponumerowanie słów liczbami zapisanymi w systemie o podstawie liczenia N, gdzie N jest liczbą różnych słów w języku. Słowo o numerze 1 (najczęściej używane) ma zapis, zaczynający się cyfrą „1", zaś adresy w tekście, gdzie to słowo występuje, kodowane są w dowolny sposób tworząc kolejne cyfry zapisu, słowo o numerze 2 ma zapis, zaczynający się cyfrą „2", itd. Wtedy prawdopodobieństwo występowania słowa o numerze p jest równe prawdopodobieństwu wylosowania liczby, zaczynającej się cyfrą p w systemie o podstawie liczenia N.

Podobnie można zauważyć podobieństwo między wzorem na prawdopodobieństwo występowania pierwszej cyfry i prawem Mandelbrota dla przypadku $q > 0$. Zarówno we wzorze Zipfa jak i Mandelbrota parametr b jest równy $q + 1$. Interpretacja współczynnika q jest więc taka, jak parametru b. Oznacza to, że współczynnik q może być interpretowany jako mierzący bogactwo słownictwa, a jednocześnie oceniający zasięg naszej listy słów. Prawdopodobieństwo, że spotkamy słowo o numerze α razy dalszym jest α^p razy mniejsze od prawdopodobieństwa naszego słowa [1]. Prawo Zipfa odpowiada przypadkowi $q = 0$, a więc przypadkowi pełnej jednorodności wyboru słów, czyli równemu dostępowi do słów całego słownika. Przypadek ten oznacza, że tekst o współczynniku $q = 0$ ma najbogatsze możliwe słownictwo.

Korzystając z informacji o statystyce słów w niektórych językach [3] spróbujmy opisać je parametrami, o których była mowa w artykule.

Język	Wielkość próby	Liczba różnych słów	Procent pierwszych n słów				
			$n = 10$	$n = 50$	$n = 100$	$n = 1000$	$n = 2000$
angielski	5 088 721	86 741 (1,7%)	25	50	59	82	95,4
arabski	136 089	5 981 (4,4%)	17,4	29	35,1	75,4	88,6

Bogactwo różnych słów jest zdumiewające: popularny słownik języka angielskiego zawiera 80 000 słów, *Le Petit Larousse* ma ich ponad 50 000. Przypuszcza się, że w języku angielskim występuje ponad 1 000 000 słów. Z danych, które odnalazłem w sieci Internet, w bazie danych TREC o objętości 1 gigabajta występuje ponad 120 milionów słów, w tym pół miliona różnych (0,4 %).

W tabeli wielkość próby oznacza łączną liczbę słów w tekstach podlegających analizie, liczba różnych słów jest uzupełniona informacją, jaką część słów w tych tekstach stanowią różne słowa. Procent pierwszych n słów oznacza, jaką część zestawu różnych słów stanowi n najczęściej używanych.

W poniższej tabeli są parametry prawa Mandelbrota, oszacowane tak, by mniej więcej odpowiadały tabeli, zawierającej rzeczywiste dane. Dla kontrastu, umieściłem oszacowanie, wynikające z prawa Zipfa i wzoru Benforda.

Język	c	a	q	Przybliżony procent pierwszych n słów ze wzoru Mandelbrota (prawo Zipfa; prawo Benforda)				
				$n = 10$	$n = 50$	$n = 100$	$n = 1000$	$n = 2000$
angielski	0,25	5	0,1	21 (26;21)	43 (40;35)	54 (46;41)	86 (66;61)	94 (72;67)
arabski	0,18	5	0,06	16 (34;28)	35 (52;45)	44 (60;53)	73 (86;79)	81 (94;87)

Dla języka angielskiego, prawdopodobieństwo według prawa Benforda jest dalekie od rzeczywistych wartości, bo q jest istotnie różne od 0. Jak można zauważyć, bogactwo języka arabskiego w analizowanych tekstach jest znacznie większe, niż bogactwo języka angielskiego.

Literatura

[1] A. Dąbrowski, *Osiem i pół*, MSN 21(1998), 9–14.
[2] A. Dąbrowski, *O teorii informacji*, Warszawa 1974, WSiP.
[3] S. Irmay, *The relationship between Zipf's law and the distribution of first digits*, J. of Applied Statistics, 4 (1997), 383–393.
[4] K. Omiljanowski, *Osiem*, MSN 21(1998), 1–8.
[5] G.K. Zipf, *Human Behavior and the Principle of Least Effort*, Reading 1949, Addison-Wesley.