Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos

La ley de Zipf en la Biblioteca Miguel de Cervantes

Rafael C. Carrasco

Copyright © 2000
Actualizado el: 3 de septiembre de 2003

La ley de Zipf, llamada así por el profesor de lingüística de la Universidad de Harvard George Kingsley Zipf (1902-1950), es una curiosidad matemática que explica algunas de las dificultades que aparecen en las bibliotecas digitales.

Supongamos que hacemos una prelación ("ranking") de las palabras que aparecen en la biblioteca de forma que asignamos el número uno a la palabra más frecuente, el dos a la segunda más frecuente, etc. Por ejemplo, en la biblioteca Miguel de Cervantes, las 10 palabras más frecuentes y sus frecuencias de aparición $f(n)$ son las siguientes:

n	palabra	$f(n)$
1	de	5952871
2	que	4294496
3	y	3887331
4	la	3473934
5	en	2521954
6	el	2463429
7	a	2348470
8	los	1689770
9	se	1305932
10	no	1261456

La ley Zipf establece que el número de apariciones de una palabra es inversamente proporcional a su número de orden, es decir,

$$f(n) \simeq \frac{C}{n}$$

donde C es una constante que se fija experimentalmente.

La siguiente figura ilustra el nivel de aproximación con se cumple la ley de Zipf en la biblioteca Miguel de Cervantes (C se eligió igual a 10 millones). La línea verde representa el comportamiento ideal y los puntos rojos los valores reales.

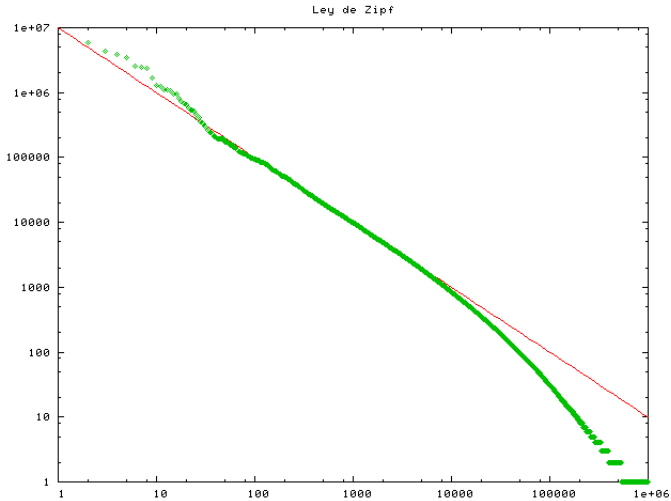

Figura 1: La ley de Zip en la biblioteca Miguel de Cervantes.

La siguiente tabla ilustra también el nivel de aproximación para algunas palabras en particular:

n	palabra	$f(n)$	C/n
10	no	1261456	1000000
100	día	93619	100000
1000	penas	9837	10000
1000	francamente	841	1000

Pese a tratarse de un resultado aproximado, una de las virtudes de la ley de Zipf es que explica lo difícil que es construir buenos diccionarios. En primer lugar, sea cual sea el tamaño de la biblioteca, la adición de nuevos documentos añade algunas palabras nuevas. En segundo lugar, un razonamiento matemático simple nos dice que la biblioteca contiene del orden de C palabras distintas y que el número de palabras N_f con frecuencia f es aproximadamente

$$N_f \simeq \frac{C}{f(f+1)}$$

Por tanto, si construimos un diccionario que contiene todas las palabras de la biblioteca (esto es, con C entradas), aproximadamente la mitad de las palabras del diccionario ($C/2$) aparecen sólo una vez en la biblioteca: este es, en efecto, el resultado si hacemos $f = 1$ en la fórmula anterior. Dicho de

otra manera, si eliminamos los "hapax legomena", el tamaño del diccionario se reduce a la mitad.

Además, las palabras que aparecen sólo dos veces ($f = 2$) constituyen otra parte considerable cerca del diccionario (sobre un sexto), y así sucesivamente. Es evidente que la probabilidad de cometer errores en la incorporación al diccionario de estas palabras infrecuentes es muy elevada y requiere una tarea de supervisión fabulosa. Por otro lado, es sabido que

$$\sum_{n=1}^{N} \frac{1}{n} \simeq \log 2N$$

Por tanto, si llamamos cobertura r del diccionario a la tasa de palabras de la biblioteca presentes en el diccionario formado por las n palabras más frecuentes, tenemos que

$$r \simeq \frac{\log 2n}{\log 2N}$$

siendo N el número de palabras distintas en la biblioteca y n el número de entradas del diccionario. En nuestra biblioteca $N \simeq 1000000$ y con un diccionario con sólo una décima parte de las palabras, esto es $n = N/10$, la cobertura se acerca al 85 %. Mejorar la cobertura en un 10 % adicional requiere incluir la mitad de las palabras ($n = N/2$), esto es, multiplicar por cinco el tamaño del diccionario.