Mayo 13, 2010

Monos al teclado, la ley del menor esfuerzo y los buscadores Web

Categoría: Sin categoría — dccuchile - 2:54 pm

Por Gonzalo Navarro, profesor del Departamento de Ciencias de la Computación, FCFM, Universidad de Chile

Se ha preguntado alguna vez **¿cómo es posible que los grandes buscadores de la Web encuentren, en centésimas de segundo, las mejores páginas para su consulta entre los miles de terabytes de información que almacenan?** Intentaré explicar que su éxito se debe a que somos, al menos en un sentido estadístico, monos al teclado.

Esencialmente, **para cada palabra los buscadores almacenan una lista de las páginas Web donde dicha palabra aparece. Y frente a una consulta formada por varias palabras, buscan en las listas de cada una de éstas y calculan las páginas que contienen el mayor número de palabras que aparecen en la consulta realizada**. (Esto es una simplificación, pues hay otros factores involucrados como la importancia global de la palabra, de la palabra en la página, de la página en sí y un largo etcétera).

Pero esta estrategia ¿funcionaría en cualquier escenario? Pues no. Resulta que si los textos que la humanidad pone en la Web fueran los que típicamente usa la naturaleza para codificar información en nuestros genes, o las amplitudes de las ondas electromagnéticas que emiten las estrellas u otros tipos de secuencias con las que convivimos, las cosas serían mucho más difíciles. El esquema de listas recién descrito colapsaría por diversas razones.

Pero los textos que escribimos los humanos son muy distintos. En 1949, George Zipf publicó un libro titulado *Human Behavior and the Principle of Least-Effort* ("El Comportamiento Humano y el Principio del Menor Esfuerzo"). En él, Zipf observaba que si uno toma un texto generado por humanos, cuenta las frecuencias de las palabras y las ordena de la más a la menos frecuente, resulta que la i -ésima palabra ocurre con una frecuencia proporcional a $1/i^{\theta}$, donde $\theta \geq 1$ es un parámetro que depende del tipo de texto.

Esto significa que la frecuencia de las palabras va decreciendo rápidamente en esta lista. Y como consecuencia unas pocas palabras concentran la mayor parte de las ocurrencias en un texto. Dicho de otro modo, **hay unas pocas palabras que aparecen muchas veces y muchas palabras que aparecen pocas veces**.

Las palabras que aparecen muchas veces son las llamadas palabras vacías, como los artículos, conjunciones y preposiciones. Estas no sirven para discriminar qué páginas son relevantes a una consulta, porque aparecen en casi todas. Nadie consultaría en un buscador por "el y pero" pretendiendo encontrar algo útil. Por ello los buscadores pueden ahorrarse procesar listas gigantes de ocurrencias de palabras que no ayudarán a discriminar entre una buena y una mala respuesta.

Las palabras que aparecen en pocas páginas Web son las que sirven para distinguir aquellas relevantes a una consulta. Sus listas de ocurrencias son más cortas y por ello más manejables. Esto permite a los buscadores encontrar las páginas más adecuadas sin considerar realmente todas las páginas donde aparece cada término de la consulta.

Zipf propuso una explicación a este fenómeno basada en la "ley del menor esfuerzo", según la cual este tipo de distribución aparecería cuando tanto el hablante como el oyente intentan minimizar el esfuerzo para entenderse mutuamente. Pero la verdad es

Archivos

- Monos al teclado, la ley del menor esfuerzo y los buscadores Web
- El futuro de la Web: ¿nuestro futuro?
- China ¿en guerra contra Internet?
- Un computador (digital) por niño
- El retraso en el cambio de hora: ¿acierto o desacierto?
- Codd: ¿Cómo darle un buen diseño a los datos?
- Igual se entiende, ¿no?
- ¿Programación de computadores en la educación media? Reflexiones al calor de una Escuela de Verano
- Terremoto 2010: ¿Internet resistió bien la prueba?
- Ciencia y Tecnología: las propuestas del próximo gobierno

Otros Blogueros

- Belisario Iturra Peralta (Noticias)
- Claudio Uson (Tecnología)
- Juan Guillermo Tejeda (Noticias)
- Tomás Flores Economista (Invertia)
- Ximena Torres Cautivo (Libros)

que la Ley de Zipf (y su versión más refinada, la Ley de Mandelbrot) no se aplica sólo a la frecuencia de las palabras, sino a muchos otros fenómenos humanos y sociales: tamaño de las ciudades, cantidades de amigos en un grupo social o de links que reciben los sitios Web, distribución de la riqueza, tamaño de empresas, etc.

El esquema de las listas tampoco funcionaría si existieran miles de millones de palabras distintas, pues habría innumerables listas demasiado cortas y esto generaría problemas serios de eficiencia. Nuevamente las estadísticas vienen al rescate. En 1978, Harold Heaps, en su libro *Information Retrieval - Computational and Theoretical Aspects* ("Recuperación de Información - Aspectos Computacionales y Teóricos") propuso otra ley empírica, según la cual el número de palabras distintas en un texto de *n* palabras crece como $K \cdot n^{\beta}$, para constantes $K > 0$ y $0 < \beta < 1$ que dependen del lenguaje (o tipo de texto). La ley se cumple con sorprendente fidelidad, para valores de β cercanos a 1/2. Esto significa que **el vocabulario de palabras distintas en un texto crece aproximadamente en razón de la raíz cuadrada del tamaño del texto, gracias a lo cual no hay tantas palabras diferentes. Y las máquinas de búsqueda Web pueden ser eficientes al mantener la información sobre cada palabra distinta en la memoria principal (y no yendo al disco, que es mucho más lento)**.

¿De dónde provienen estas leyes tan curiosas, que se cumplen con tanta fidelidad y a las que debemos tanto? La explicación es sorprendentemente sencilla y quizás desalentadora. Siente a un mono frente a un teclado (o, si prefiere, use un generador de letras aleatorio). El mono apretará la barra espaciadora, que supondremos será la única tecla que corta las palabras, con una cierta probabilidad p, y todas las demás teclas sumarán $1 - p$. Es fácil ver que la probabilidad de que el mono tecleé una palabra de largo i es $(1-p)^i \cdot p$, es decir, decrece exponencialmente con i. Esto significa que las palabras más cortas son las más frecuentes (¡precisamente! artículos, conjunciones, preposiciones…). Más aún, de este modelo tan simple se deducen las leyes de Heaps y de Zipf, cumpliéndose idealmente $\beta \cdot \theta = 1$; no tan lejos de lo que sucede en los textos reales.

Es un conocido cliché que si pusiéramos a un mono a escribir por suficientes millones de años, alguna vez terminarían apareciendo todas las obras de Shakespeare. Peor aún. Ahora vemos que al menos en lo que respecta a las estadísticas del texto que se genera, el mono haría un trabajo muy parecido al nuestro desde el primer minuto. Por suerte hay mucho más en *Cien Años de Soledad*, *La Casa de los Espíritus* o *Ficciones* que lo que capturan estas simples estadísticas. Y de lo que aún pueden capturar los computadores que tratan de comprender el lenguaje natural. Suerte para lo que va quedando de nuestro aporreado orgullo y quién sabe por cuánto tiempo. Nuevamente, la Computación nos enseña mucho sobre nosotros mismos.

permalink trackback
Comentarios (0)
« Older Posts