JUEGOS MATEMÁTICOS

Juan M. R. Parrondo

Cita a ciegas

Aunque al lector ajeno al mundo de la investigación le cueste creerlo, una de las partes más importantes —y más leídas— de un artículo científico es la llamada lista de referencias, en la que se encuentran todos los libros, artículos y trabajos que se citan en el texto principal del artículo. Renombrados científicos, al llegar a sus manos el trabajo de un colega, pasan páginas con rapidez convulsa para llegar a la mencionada lista y comprobar si aparecen citados o no. ¿Por qué tanto interés? ¿Pura vanidad? Puede que sea así en buena parte, pero existe otra razón: hoy en día, acertada o desacertadamente, la calidad científica de un trabajo científico se mide por el número de veces que ha sido citado. Aunque cueste reconocerlo, muchos tribunales que juzgan puestos o becas de investigación recurren al número de citas de un artículo para determinar su relevancia, en lugar de leerlo.

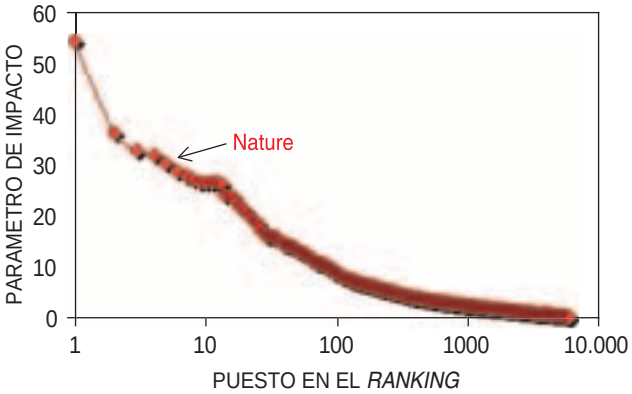
La empresa privada Institute of Scientific Information (ISI) nació en 1958 con el objetivo de recopilar los títulos y resúmenes de todas las publicaciones científicas. En 1961 decidió recopilar también información acerca de las citas. Disponer de un índice en donde se pueden buscar todos los artículos que han citado alguno dado es una herramienta extremadamente útil. Suponga que usted comienza a trabajar en un cierto tema y que encuentra un artículo de 1973 con ideas parecidas a las suyas. Rastrear todos los artículos que, desde 1973 hasta hoy, han citado aquel trabajo le permitirá conocer toda la investigación que se ha hecho en estos últimos 30 años sobre el tema y las ideas que le interesan, algo que es bastante más difícil de obtener sólo mediante búsquedas que hacen uso de las palabras que aparecen en el título o en el resumen de los artículos.

Lo que quizá no esperaban los creadores del ISI es que la recopilación de citas se convertiría con el tiempo en uno de los métodos más utilizados para evaluar, no sólo la calidad de trabajos de investigación, sino también el impacto de las revistas científicas. En efecto, lo que hoy en día determina el éxito de una revista científica es sobre todo su *parámetro de impacto*, que no es más que el número medio de citas que recibe cada artículo publicado en la revista. Una de las revistas con mayor impacto es, por ejemplo, *Nature*. Su parámetro de impacto en 2002 es 30,432 porque los 1952 artículos publicados en los años 2000 y 2001 han sido citados, durante 2002, 59.403 veces, es decir, 30,432 veces por artículo. Este es un número medio de citas bastante alto. De las 5830 revistas registradas en el ISI, *Nature* es la quinta con mayor parámetro de impacto. En la figura 1 se puede ver el parámetro de impacto de estas 5830 revistas, en función del puesto que ocupan en una lista en la que se ordenan de mayor a menor parámetro de impacto.

Si las citas son una buena medida de la importancia de un artículo o una revista es un asunto que está permanentemente en cuestión. En primer lugar, existe una cierta picaresca en torno a las citas, ya que editores y censores de revistas tienden a "recomendar" que se citen sus artículos. En segundo lugar, son muy citados los artículos que contienen alguna idea fácil de generalizar, independientemente de que sea relevante o no, mientras que un artículo que resuelva de una vez por todas un viejo problema puede que se cite muy pocas veces.

A esta polémica se ha sumado una sorprendente investigación realizada por M. V. Simkin y V. P. Roychowdhury, de la Universidad de California en Los Angeles, y que demuestra que los científicos no somos especialmente cuidadosos ni críticos a la hora de citar otros trabajos. En un primer artículo, titulado "¡Leed antes de citar!", investigaron cómo se propagan los errores en las citas: erratas en los nombres de los autores o en la página del artículo citado. Simkin y Roychowdhury analizaron estadísticamente estas erratas en el caso de un artículo citado 4300 veces y comprobaron que los errores se "propagan" a lo largo de las citas. Por ejemplo, de las 45 erratas que detectaron, una de ellas, las más frecuente, aparecía 78 veces, lo cual indica que muchos de los autores que citan ese trabajo han copiado la cita de la lista de referencias de otros artículos sin comprobar los datos del trabajo citado y, lo que es peor, probablemente sin leerlo.

Un argumento muy simple nos puede servir para estimar la fracción de "citadores" que realmente han leído el artículo. Simkin y Roychowdhury detectaron 196 erratas en las 4300 citas. De estas 196, sólo 45 eran diferentes. Es evidente que, de los 196 autores que han citado erróneamente, 196 – 45 = 151 han copiado,



1. Parámetro de impacto de 5830 revistas en función del puesto que ocupan en una lista en donde las revistas se ordenan de mayor a menor parámetro de impacto

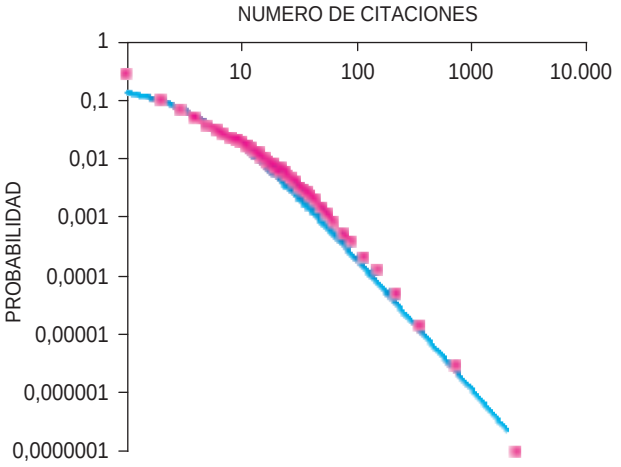
puesto que su errata no es original. Admitiendo la "presunción de inocencia" de los otros 45, es decir, suponiendo que ellos sí han leído el trabajo citado, obtenemos que la fracción de lectores frente a citadores es $R = 45/196 = 0{,}23$. Es decir, según esta deducción, los científicos leen menos de la cuarta parte de los artículos que citan. Un argumento más elaborado basado en un modelo probabilístico conduce a un valor casi idéntico, $R = 0{,}22$.

Si los científicos citan copiando la lista de referencias de otros artículos en lugar de leer las referencias originales, entonces puede que tenga lugar un efecto "bola de nieve" en las citas que recibe un artículo. Simkin y Roychowdhury, en un segundo trabajo, han demostrado que esto es posible. Han diseñado un modelo muy simple que reproduce la estadística de las citas. El modelo podría llamarse de "cita a ciegas", ya que supone que cada científico elige las citas al azar, de acuerdo con las reglas siguientes. Supongamos que se han escrito N artículos y que un científico escribe uno nuevo. El científico escoge al azar m artículos de los N existentes, los cita y cita también, con una probabilidad p, las referencias que aparecen en esos m artículos.

El modelo es tan simple que se puede simular fácilmente con un programa de ordenador, comenzando con m artículos sin lista de referencias. Puede incluso resolverse de modo aproximado, aunque la forma de hacerlo es demasiado complicada para esta sección. En la figura 2 se muestra la solución del modelo con $m = 3$ y $p = 1/4$. Estos son los valores que mejor se ajustan a los datos reales, extraídos del ISI, y que también se muestran en la figura. La gráfica representa la probabilidad de que un artículo haya sido citado k veces, o equivalentemente, la fracción de artículos citados k veces, en función de k. Como se ve en la figura, el modelo reproduce bastante bien los datos reales. Podemos también comprobar el efecto "bola de nieve" mencionado antes. Supongamos que hay N artículos y que uno de ellos ha sido citado k veces. ¿Cuál es la probabilidad de que lo cite el artículo $N + 1$? Recordemos que en el modelo un nuevo artículo elige m al azar entre los N existentes, los cita y cita con probabilidad p los artículos que aparecen en las listas de referencias de los m elegidos. Por tanto, hay dos posibilidades para ser citado: o bien a) el artículo es elegido entre los m primeros, o b) está en la lista de referencias de algunos de estos m primeros artículos. La probabilidad de que ocurra a) es, aproximadamente, m/N y la de que ocurra b) es km/N. Pero, en el caso b) el artículo es sólo citado con una probabilidad p. Por tanto, la probabilidad de que nuestro artículo con k citas sea citado de nuevo es:

$$\frac{m + pkm}{N}$$

y es mayor cuanto mayor sea k. De hecho, si k es pequeño, será muy poco probable que el artículo sea citado de nuevo. En el modelo de cita a ciegas habrá muchos artículos apenas citados y unos pocos muy citados, tal y como muestra la figura 2. Y es ésta la distribución a la que se ajustan los datos reales.



2. La gráfica muestra la probabilidad de que un artículo tenga un número de citas determinado. Sólo aproximadamente uno de cada un millón de artículos es citado más de 1000 veces. La línea azul corresponde al modelo de "cita a ciegas", mientras que los puntos color magenta son los datos reales tomados del ISI. [Figura cedida por M. V. Simkin]

Otra cantidad que puede extraerse fácilmente del modelo es el número medio de citas en cada artículo. Supongamos que, cuando hay N artículos publicados, la lista de referencias de cada uno ellos tiene, de media, $c(N)$ referencias. Un nuevo artículo aparece y, en media, tendrá un número de referencias:

$$c(N + 1) = m + mpc(N)$$

como se deduce fácilmente de las reglas del modelo de cita a ciegas. Cuando el número total de artículos N es muy grande, el tamaño medio de la lista de referencias apenas cambia. Si c es este valor casi constante, entonces tiene que verificar $c = m + mpc$, lo que da un valor para el tamaño medio de la lista de referencias:

$$c = \frac{m}{1 - mp}$$

Esta fórmula es sólo válida si mp es menor que uno. Si no fuera así, las listas de referencias crecerían indefinidamente. Los valores que mejor reproducen los datos reales son, $m = 3$ y $p = 1/4$, lo que da un tamaño medio de $c = 12$ referencias. Los datos del ISI indican que este número es ligeramente superior, en torno a 15. Sin embargo, el acuerdo sigue siendo bueno, especialmente si se tiene en cuenta la simplicidad del modelo.

Aunque es cierto que los científicos muchas veces citan artículos que no han leído, no lo hacen realmente a ciegas, por supuesto. Sin embargo, el modelo sí pone de manifiesto que la red de citas, que puede considerarse como la red en la que se crea y por la que se transmite el conocimiento científico, tiene una estructura bastante homogénea y sorprendentemente sencilla.