Cómo encontrar una aguja en un pajar

Lección inaugural
del curso académico 2013-2014
pronunciada por el

Prof. Dr. Luis M. Ezquerro Marín

Catedrático de Álgebra
de la Universidad Pública de Navarra

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

upna Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

Cómo encontrar una aguja en un pajar

Lección inaugural
del curso académico 2013-2014
pronunciada por el

Prof. Dr. Luis M. Ezquerro Marín
Catedrático de Álgebra
de la Universidad Pública de Navarra

Pamplona, 13 de septiembre de 2013

upna
Universidad
Pública de Navarra
Nafarroako
Unibertsitate Publikoa

Edita: Universidad Pública de Navarra / Nafarroako Unibertsitate Publikoa

Coordinación: Servicio de Comunicación

Fotocomposición: Pretexto. pretexto@pretexto.es

Imprime: Ona Industria Gráfica

Depósito Legal: NA 1286-2013

Distribución: Sección de Publicaciones
Universidad Pública de Navarra
Campus de Arrosadia
31006 Pamplona
Fax: 948 169 300
Correo: publicaciones@unavarra.es

Al recibir el encargo del señor rector para pronunciar esta lección inaugural del curso 2013-14 se acumularon en mí una plétora de sentimientos difíciles de expresar: el agradecimiento hacia quien deposita en mi persona esta nobilísima tarea, el recuerdo de quienes, a veces con su propio sacrificio, facilitaron y facilitan el recorrido para llegar a este atril, la consciencia del paso inexorable de los años, la distinción, la responsabilidad y la satisfacción por exponer a mis compañeros, y en general a toda la sociedad aquí representada, unas pinceladas de mi quehacer cotidiano...

En aquellos días del encargo explicaba a mis alumnos de grado de Ingeniería Informática algunas nociones introductorias de la *Teoría de Grafos*. Les indicaba que este año 2013 se celebra el centenario del nacimiento en Budapest del insigne matemático Paul Erdős. Me parece un buen punto de arranque para contarles lo que deseo. El catálogo de contribuciones de Erdős a las matemáticas es abrumador: trabajó en cuestiones relacionadas con la combinatoria, la teoría de grafos, la teoría de números, el análisis clásico, la teoría de conjuntos, la teoría de probabilidades... Sin duda es uno de los matemáticos más prolíficos de todos los tiempos.

Paul Erdős fue un científico de carácter singular, alguien diría que excéntrico, al que se atribuye una curiosa definición de matemático como máquina que transforma café en teoremas. Aunque dudaba de la existencia de Dios, al que llamaba «Fascista Supremo», afirmaba que Dios había escrito EL LIBRO y en este libro

imaginario había registrado todos los teoremas con sus demostraciones más bellas o más ilustrativas; sostenía que *no es necesario creer en Dios, pero es imprescindible creer en EL LIBRO*.

EL LIBRO de Erdős no es sino un ejemplo de archivo imaginario que recoge un saber universal. La literatura nos ofrece ejemplos de ficciones protagonizadas por bibliotecas universales. A la excitante idea de un archivo que contiene todos los libros posibles, pasados, presentes y futuros, se añade el estupor que produce su número que, aun siendo finito, excede en mucho la comprensión humana. *¿Cómo puede ser finito lo inagotable?*[1] se pregunta el escritor alemán Kurd Laßwitz en el cuento «La biblioteca universal»[2] publicado en 1904. En tal biblioteca, si los libros que la conforman tuvieran dos centímetros de grosor y se colocasen en fila, un rayo de luz tardaría aproximadamente de $2 \times 10^{1.999.982}$ años en atravesarla. *No se puede* –dice– *concebir ni el número de años que necesita la luz para recorrer la biblioteca, ni el número de los volúmenes*[3]; pensemos que la edad del universo se cifra en unos 14 mil millones de años, o sea alrededor de $1{,}4 \times 10^{10}$ años.

Menos de cuarenta transcurrieron hasta que Jorge Luis Borges presentó una colección de relatos cortos titulada «El jardín de los senderos que se bifurcan» en el que incluyó uno, el célebre *La biblioteca de Babel*, inspirado en la desmesura de Laßwitz[4]. En la biblioteca borgiana se almacenan no sólo las obras perdidas de Tácito, como en la de Laßwitz, sino *la historia minuciosa del porvenir*, incluyendo *la relación verídica de tu muerte* o *las autobiografías de los arcángeles*. Y lo que en el alemán era un diálogo ameno entre dos amigos bebedores de cerveza, quizás producto de tan benéfica bebida, en el argentino se convierte en una visión inquietante, turbadora y laberíntica.

Tanto Laßwitz como Borges se enfrentan a la cuestión de cómo buscar un volumen concreto en esta infinitud aparente. Ambos coinciden en que la probabilidad de encontrarlo es, en palabras de Borges, *computable en cero*. En castellano tenemos un refrán que describe con chispa esta dificultad prácticamente insuperable: *es como*

1. Wie soll das Unerschöpfliche endlich sein?

2. K. Laßwitz: Die Universalbibliothek. Publicado el 18.12.1904 en *Ostdeutschen Allgemeinen Zeitung* de Breslau. <http://gutenberg.spiegel.de/buch/3130/1>. Versión castellana de A. Hanke-Schaefer y A. Fernández Ferrer en <http://www.letraslibres.com/revista/convivio/borges-y-sus-precursores>.

3. Man kann sich die Zahl der Jahre, die das Licht braucht, an der Bibliothek entlangzulaufen, ebensowenig vorstellen, wie die Zahl der Bände selbst.

4. J. L. Borges: *La biblioteca de Babel*, extraída de «El jardín de los senderos que se bifurcan» (1941). *Obras Completas*. Círculo de Lectores, 1992.

encontrar una aguja en un pajar. Y así, mientras el germano, al llegar a este escollo, sentencia con optimismo:

> «Lo sensorial es, con el tiempo, efímero. Lo lógico es independiente del tiempo y universal. Y como lo lógico no significa otra cosa que el pensamiento de la humanidad misma, por eso tenemos este don intemporal mediante el cual compartimos las leyes perennes de lo divino, compartimos también el destino del infinito poder creativo. En ello radica la ley fundamental de la Matemática»[5].

el argentino afirma:

> «A la desaforada esperanza, sucedió, como es natural, una depresión excesiva. La certidumbre de que algún anaquel en algún hexágono encerraba libros preciosos y de que esos libros preciosos eran inaccesibles, pareció casi intolerable».

Laßwitz falleció en 1910; Borges, en 1986. Ignoramos qué hubieran pensado ambos si hubieran conocido el archivo de los archivos de nuestro tiempo, aquello que los angloparlantes denominan, con acierto, *web*, esto es, «telaraña». A pesar de que su magnitud es notoriamente inferior a la de las bibliotecas forjadas por la imaginación de los literatos, la *web* ya ha alcanzado los límites de lo que Laßwitz califica de *inagotable*. También la *web* se enfrentó en su momento con la cuestión de los buscadores pero, al contrario de lo que sucede en sus antecedentes literarios, el resultado hoy nos deja perplejos por su rapidez y su eficacia. Un matemático, y en particular un algebrista, se enorgullece de que el secreto de estos brillantes procedimientos de búsqueda *de una aguja en un pajar* sea un puñado de teoremas del Álgebra Lineal. Me propongo exponerles en esta intervención un esbozo del fundamento algebraico de los buscadores de la *web*.

1. ¿Qué ocurre cuando se realiza una búsqueda en la *web*?

El capítulo de los buscadores es dilatado y diverso; dejaré para especialistas más documentados todo lo referente al multiforme mundo de la computación. Aunque centraré mi intervención en aquellos aspectos matemáticos que constituyen el núcleo

5. Das Sinnliche ist vergänglich mit der Zeit, das Logische ist unabhängig von aller Zeit, ist allgemeingültig. Und weil dieses Logische nichts anderes bedeutet als das Denken der Menschheit selbst, so haben wir in diesem zeitlosen Gut einen Anteil an den unwandelba- ren Gesetzen des Göttlichen, an der Bestimmung der unendlichen Schöpfermacht. Darauf beruht das Grundrecht der Mathematik.

del funcionamiento de un buscador, estimo que merece la pena que describamos sucintamente algunos detalles previos.

Lo primero que hay que hacer notar es que los buscadores de las diversas compañías no realizan su labor en la *web* sino en un índice de toda la red, o al menos de la parte de la red a la que la compañía puede acceder. La primera labor de un buscador consiste en enviar unos robots denominados *arañas* (no olvidemos que estamos en una «telaraña») que exploran diversos grupos de páginas *web*, siguen los enlaces que contienen, exploran las páginas a las que estos enlaces se dirigen, y repiten su labor[6]. La información obtenida por las arañas se almacena temporalmente en un repositorio central. El módulo de indización extrae una serie de datos, los comprime y los almacena en varios índices.

Es obvio que el usuario utiliza el lenguaje humano. Si el texto de una consulta no se transformase los resultados serían muy pobres, quizás nulos. Esta transformación persigue el objetivo de que el buscador entienda lo que pedimos y nos conteste con la mayor cantidad de información pertinente. Por esta razón desaparecen del texto tildes, mayúsculas y toda palabra que no altere sustancialmente el sentido de la búsqueda: preposiciones, artículos, pronombres, conjunciones... Técnicamente estas palabras se denominan con el término inglés de *stopwords*[7]. Asímismo se realiza un proceso lingüístico denominado *lematización*[8] que consiste en extraer la raíz de una palabra y tomarla como representante de todas sus formas flexionadas.

Tras estos procesos de transformación, el software busca las páginas que se adecuan a la consulta. Con toda probabilidad encontrará cientos de miles, incluso millones, de páginas que contengan estos términos.

Cualquier búsqueda literal se debe enfrentar con dos problemas: la polisemia y la sinonimia. ¿Cómo se pueden salvar estas dificultades? Sin duda necesitamos una búsqueda por significado más que una búsqueda literal. El Álgebra Lineal proporciona el modelo de espacio vectorial. Simplificando, cada documento se codifica como un vector en el que cada coordenada, cada componente, refleja la importancia de un término particular, un concepto o una palabra clave, en la semántica de tal documento. El valor asignado o ponderación suele ser una función de la frecuencia con la que aparece el concepto en el documento o en la colección de documentos pero también tiene en cuenta si el término aparece en el título, en texto escrito con letras destaca-

6. Este primer trabajo en el proceso de búsqueda se denomina en inglés *web crawling*.

7. Puede consultarse una lista de cerca de 400 *stopwords* para el español en <http://droope.org/2011/02/28/stopwords-para-espanol-castellano/>.

8. En inglés, *stemming*.

das (negritas, bastardillas, versalitas, subrayadas), en textos asociados a enlaces con otras páginas, etc.

Con este lenguaje de vectores la idea de hallar documentos pertinentes a una consulta determinada, se transforma en la detección de los vectores más cercanos al vector que codifica la consulta. En definitiva se trata de medir la amplitud del ángulo que forman esos vectores. El producto escalar es la herramienta típica que pone en nuestras manos el Álgebra Lineal. No obstante, el «ruido» producido por la sinonimia y la polisemia es tan acusado que este método resulta un tanto tosco. Es menester refinarlo y para ello se acude –¡cómo no!– a otro resultado, mucho más elaborado, de Álgebra Lineal: la *Descomposición de Valores Singulares* de una matriz.

Sin entrar en detalles, constataremos aquí que la descomposición de valores singulares de una matriz de cualquier tamaño, nos proporciona una construcción de manera que ese «ruido» al que nos referimos se filtra sin una pérdida significativa de información.

El uso de estas técnicas algebraicas nos devuelve ya una lista ordenada de documentos. Pero esta ordenación depende de nuestra consulta y no tiene en cuenta la relevancia propia de cada documento. No todos los documentos son igualmente prestigiosos y mientras unos pueden ser referencia obligada en un ámbito hay otros que solo tiene un limitado alcance local. Esta diferencia de importancia asigna un peso independiente de nuestra consulta que no se ha tenido en cuenta en la exposición anterior.

Pero ¿cómo medimos la relevancia de un documento? En el ambiente científico nos hemos acostumbrado ya a una serie de parámetros e índices de impacto que tratan de medir la calidad de las revistas científicas y que se basan, esencialmente, en el número de citas. Su uso, y sobre todo su abuso, es un asunto controvertido.

Inspirándose en estos modelos, apareció el algoritmo *PageRank* para la asignación de importancia a las páginas *web*. Su implantación y desarrollo constituye una de las más pasmosas exhibiciones del poder de los buscadores informáticos: Google[9].

En lo que a nosotros concierne, el modelo *PageRank* está basado en un uso inteligente de ciertos teoremas cruciales del Álgebra Lineal conocidos desde principios del siglo XX: los teoremas de Perron-Frobenius.

Dedicaremos el resto de la lección a una exposición sucinta de los fundamentos matemáticos del modelo *PageRank* de Google.

9. Como curiosidad apuntamos que el nombre Google proviene del nombre *googol* que alguien, al parecer un sobrino del matemático Edward Kasner, impuso al número 10^{100}.

2. ¿Cómo mide Google la importancia de un documento en la *web*?

Cada vez que efectuamos una búsqueda, Google nos informa de tres cosas: el número total de páginas encontradas (que suele ser vastísimo), el tiempo que ha tardado en realizar esta búsqueda (que suele ser brevísimo) y una lista ordenada de las primeras páginas *web* encontradas por el buscador que tienen alguna relación con nuestra consulta. ¿Cómo es posible?

Sergey Brin (2010)

Lawrence Page (2009)

El secreto está en el algoritmo de búsqueda y de ordenación. Este método fue diseñado en 1998 en la universidad de Stanford por un matemático, Sergey Brin y un informático, Lawrence Page, cuando ambos eran estudiantes de doctorado de Informática[10].

10. S. Brin y L. Page: *The anatomy of a large-scale hypertextual web search engine*. Technical Report. Stanford. In: Seventh International World-*Wide web* Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia. <http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf>.

Hay otro documento, un poco posterior, que amplía, y corrige, al anterior: se trata del artículo de L. Page, S. Brin, R. Motwani y T. Winograd: *The PageRank citation ranking: Bringing order to the web*. Technical report, Stanford InfoLab. 1999. <http://dbpubs.stanford.edu/pub/1999-66>.

Llamaron a su algoritmo *PageRank*, al parecer para mencionar el apellido de uno de sus creadores.

Precisemos nuestro objetivo. Cuando planteamos una consulta al buscador de internet, en particular a Google, deseamos que nos proporcione una lista de las páginas *web* pertinentes con nuestra consulta y que tal lista esté ordenada según un criterio de importancia[11].

En Álgebra Llineal una lista ordenada de datos es un sinónimo de vector. A cada página *web P* se asocia una medida de su alcance a la que llamamos *importancia de P* y que designamos con *I* (*P*). Buscamos un vector, que llamaremos *vector de importancias*[12]:

$$\begin{array}{lccccc} \text{Conjunto de páginas:} & \{P_1, & P_2, & \cdots, & P_n\} \\ & \downarrow & \downarrow & & \downarrow \\ \text{Vector de importancias } I\text{:} & (I(P_1), & I(P_2), & \cdots, & I(P_n))' \end{array}$$

¿Cómo se determina *I* (*P*)? Tomando como ejemplo la elaboración de índices de impacto a partir de las citas recibidas por los artículos científicos, una primera idea, un tanto grosera, de la importancia de una página *web* nos la da el número de enlaces que recibe de otras páginas. Ahora bien, puede suceder que una página sea citada en unas pocas páginas pero muy relevantes, muy citadas por otros: «muy importantes». Parece pues que debemos asignar más importancia no sólo a las páginas más citadas sino también a las páginas que estén citadas en «páginas muy importantes». Enunciamos ya el postulado principal de *PageRank*:

Postulado *PageRank*. *La importancia de una página* web *es proporcional a la suma de las importancias de las páginas vinculadas a ella*.

11. Para la elaboración de estas notas, navegando por la *web*, hemos encontrado algunos textos en castellano que merecen ser nombrados. Se trata, en primer lugar, del artículo de P. Fernández Gallardo: *El secreto de Google y el álgebra lineal*. Bol. Soc. Esp. Mat. Apl. 30 (2004) 115-141; también destacamos la presentación de J. M. Gracia: *Álgebra Lineal tras los buscadores de Internet*, Univ. País Vasco, Dpto. Mat. Aplicada, 2002. <http://www.vc.ehu.es/campus/ centros/farmacia/deptos-f/depme/profesor/gracia/buscap.pdf>. Es destacable también el trabajo en catalán de J. Gimbert: *Les matemàtiques de GOOGLE: l'algorisme PageRank*, Butll. Soc. Cat. Mat., Vol. 26, 1 (2011) 29-55. DOI: 10.2436/20.2002.01.33. Finalmente no queremos dejar de mencionar una amena presentación, ya en inglés, que nos ha inspirado, entre otras cosas, para la elección del título de esta lección; se trata del texto *How Google Finds Your Needle in the web's Haystack,* de D. Austin, que apareció en Amer. Math. Soc. Feature Column Monthly essays on mathematical topics. <http://www.ams.org/samplings/feature-column/fcarc-page rank>.

12. Consideraremos los vectores en columna pero, para facilitar la redacción de este texto, se escriben en filas acompañadas de una prima que indica la *transposición*, esto es, la transformación de filas en columnas y viceversa.

La aplicación simple del postulado *PageRank* nos proporciona el siguiente sistema de ecuaciones lineales en el que K es la constante de proporcionalidad, que tomamos positiva:

$$I(P_i) = K \left[\sum_{P_j \text{ tiene enlace a } P_i} I(P_j) \right] \qquad (K > 0) \tag{1}$$

Pero, ¡atención!: este postulado adolece de un defecto grave: es autorrecurrente; lo definido no puede aparecer en la definición.

Es prioritario resolver este problema.

2.1. El grafo dirigido asociado a la *web*

El conjunto de páginas de *internet* públicas, o simplemente «páginas *web*», tiene estructura de grafo dirigido. Consta de:

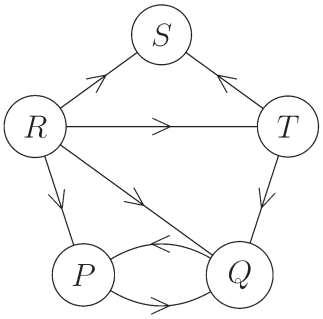
- un conjunto de *vértices* cuyos elementos son las páginas *web*: cada página es un vértice;
- un conjunto de *flechas* o *aristas dirigidas*: si en la página P hay un enlace a la página Q establecemos una flecha desde P hasta Q.

¿Cómo transformamos este ente gráfico en un objeto apto para el cálculo? Ordenamos los vértices, P_1..., P_n, y les asignamos a cada uno de ellos una fila y una columna en una matriz de forma que en la columna del vértice P_j, al llegar a la fila correspondiente a vértice P_i escribimos un uno si hay flecha de P_j a P_i y un cero si no la hay. Hemos construido la denominada *matriz de adyacencia* del grafo dirigido.

$$\begin{array}{c} \\ P_1 \\ \vdots \\ P_i \\ \vdots \\ P_n \end{array} \begin{array}{c} \begin{array}{ccccc} P_1 & \cdots & P_j & \cdots & P_n \\ \downarrow & & \downarrow & & \downarrow \end{array} \\ \begin{pmatrix} \vdots & \cdots & \vdots & \cdots & \vdots \\ \vdots & & \vdots & & \vdots \\ \vdots & \cdots & b_{ij} & \cdots & \vdots \\ \vdots & & \vdots & & \vdots \\ \vdots & \cdots & \vdots & \cdots & \vdots \end{pmatrix} \end{array} \qquad b_{ij} = \begin{cases} 1 & \text{si hay enlace de } P_j \text{ a } P_i, \\ 0 & \text{si no hay enlace de } P_j \text{ a } P_i. \end{cases}$$

Consideremos, por ejemplo, una red diminuta con cinco páginas, *R*, *S*, *T*, *P*, *Q* tales que:

- La página *P* cita la página *Q*.
- La página *Q* cita la página *P*.
- La página *R* cita las páginas *P*, *Q*, *S*, *T*.
- La página *S* no cita ninguna página.
- La página *T* cita las páginas *S*, *Q*.



La matriz de adyacencia de este grafo dirigido es

$$M = \begin{array}{c} \\ R \\ S \\ T \\ P \\ Q \end{array} \begin{array}{c} \begin{array}{ccccc} R & S & T & P & Q \end{array} \\ \left(\begin{array}{ccc|cc} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{array}\right) \end{array} = \left(\begin{array}{c|c} M_{11} & (0) \\ \hline M_{12} & M_{22} \end{array}\right)$$

No es difícil plantear el sistema de ecuaciones que deben satisfacer las importancias de las cinco páginas de acuerdo con el postulado *PageRank* y su expresión en forma matricial:

$$\begin{cases} I(R) = 0 \\ I(S) = K(I(R) + I(T)) \\ I(T) = K \times I(R) \\ I(P) = K(I(R) + I(Q)) \\ I(Q) = K(I(R) + I(T) + I(P)) \end{cases} \qquad \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} I(R) \\ I(S) \\ I(T) \\ I(P) \\ I(Q) \end{pmatrix} = \frac{1}{K} \begin{pmatrix} I(R) \\ I(S) \\ I(T) \\ I(P) \\ I(Q) \end{pmatrix}$$

Reconocemos que la matriz del sistema es *M*, esto es ¡la matriz de adyacencia del grafo dirigido!, y que si $I = (I(R), I(S), I(T), I(P), I(Q))'$ es el vector de importancias se tiene

$$MI = \frac{1}{K} I$$

es decir, el vector de importancias es un vector propio de la matriz *M* con valor propio $\frac{1}{K}$.

> «El problema de la recursión se transforma en un cálculo de vectores propios de una cierta matriz; el vector de importancias es un vector propio».

Mantendremos este gran hallazgo en todo el estudio.

Ahora bien el uso de la matriz de adyacencia M del grafo dirigido no resulta conveniente. En efecto, el sistema de ecuaciones lineales anterior cuando $K \neq 1$ tiene una única solución: la solución trivial, esto es, la importancia de las cinco páginas es 0; y si $K = 1$, la solución es $I(R) = I(S) = I(T) = 0$, e $I(P) = I(Q)$. Ninguna de las dos posibilidades es satisfactoria.

Debemos afinar nuestro análisis en busca de una matriz más adecuada a nuestros propósitos y que proporcione información sobre el grafo dirigido, es decir sobre M.

2.2. El navegante aleatorio

Reflexionamos sobre la conducta de un *navegante aleatorio*[13] por la red. Supongamos que nuestro navegante está consultando una página P_j con n_j enlaces. Al cabo de un tiempo proporcional a la importancia $I(P_j)$ de la página P_j, el navegante elige al azar un enlace para salir de P_j; tal enlace le lleva a la página P_i; la probabilidad de ir a P_i es $\frac{1}{n_j}$. Interpretamos $\frac{I(P_j)}{n_j}$ como una medida de la importancia que la página P_j trasmite a P_i. La nueva matriz formada es una alteración de la matriz M:

$$\begin{array}{c} \\ P_1 \\ \vdots \\ P_i \\ \vdots \\ P_n \end{array} \begin{array}{c} \begin{array}{ccccc} P_1 & \cdots & P_j & \cdots & P_n \\ \downarrow & & \downarrow & & \downarrow \end{array} \\ \rightarrow \begin{pmatrix} \vdots & \cdots & \vdots & \cdots & \vdots \\ \vdots & & \vdots & & \vdots \\ \vdots & \cdots & \frac{b_{ij}}{n_j} & \cdots & \vdots \\ \vdots & & \vdots & & \vdots \\ \vdots & \cdots & \vdots & \cdots & \vdots \end{pmatrix} \end{array} \quad b_{ij} = \begin{cases} 1 & \text{si hay enlace de } P_j \text{ a } P_i \\ 0 & \text{si no hay enlace de } P_j \text{ a } P_i \end{cases}$$

$$n_j = \text{nº de enlaces en la página } P_j \quad (n_j \neq 0)$$

Si llamamos a la nueva matriz H, la ecuación que define la importancia asignada a P_i es

$$I(P_i) = K \left[\sum_{P_j \text{ tiene enlace a } P_i} \frac{I(P_j)}{n_j} \right] \qquad (K > 0) \tag{2}$$

13. Traducimos así la expresión *random surfer* que aparece en los textos en inglés.

La expresión matricial del sistema de ecuaciones lineales (2) es del tipo $H\,I = \lambda I$ que nos indica que el vector de importancias es un vector propio de la matriz H.

Las columnas de H correspondientes a páginas sin enlaces son nulas. De hecho en la *web* hay muchas páginas sin enlaces. En el lenguaje propio de la *web* estas páginas de las que no sale ningún enlace se dice que son *páginas colgadas*[14]. ¿Qué puede hacer el navegante aleatorio al llegar a una página sin enlaces? Establecemos que en este momento elige una página al azar. De esta forma las páginas sin enlaces actúan como si tuvieran enlaces a todas las páginas. Así, lo que en la matriz de adyacencia M era una columna nula se trasforma en una columna donde todos los coeficientes son $\frac{1}{n}$ (n = número total de páginas *web*), una distribución de probabilidad uniforme.

Esto provoca una nueva alteración de la matriz objeto de nuestra atención. Llamamos $S = (s_{ij})$ a la nueva matriz:

$$s_{ij} = \begin{cases} \frac{1}{n_j} & \text{si la página } P_j \text{ tiene } n_j \neq 0 \text{ enlaces y hay un enlace de } P_j \text{ a } P_i, \\ 0 & \text{si la página } P_j \text{ tiene } n_j \neq 0 \text{ enlaces y no hay enlace de } P_j \text{ a } P_i, \\ \frac{1}{n} & \text{si } n_j = 0\text{: } P_j \text{ es una página colgada,} \end{cases}$$

e imponemos de nuevo que el vector de importancias I sea un vector propio de S:

$$SI = \lambda I.$$

¿Nos damos por satisfechos con este resultado? Por desgracia la matriz S no cumple los requisitos básicos para nuestro estudio. Exigimos que el vector de importancias I sea un vector propio de S para que se resuelva el problema de la recursión. No sólo esto; el vector I, además,

- debe ser un vector cuyas componentes sean todas positivas, porque así entendemos que deben ser las importancias de las páginas,
- ha de estar asociado a un valor propio positivo de S.
- nos interesa que este vector propio positivo sea único, para que no haya ambigüedad en las importancias asignadas a cada página *web*.

Nada permite afirmar que S posea tal vector propio. Es el momento oportuno para que el Álgebra Lineal entre en escena.

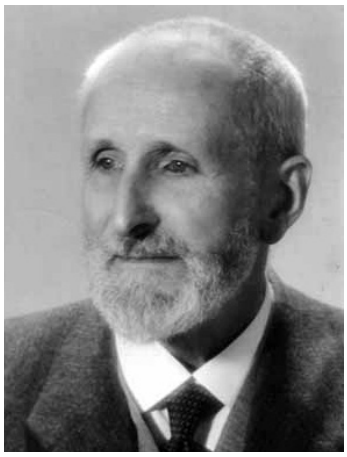
14. En inglés las páginas sin enlaces se denominan *dangling nodes*. Corresponden a los *pozos* (en inglés, *sinks*) de la teoría general de grafos dirigidos.

2.3. El auténtico protagonista: el teorema de Perron-Frobenius

Un par de teoremas formulados a principios del siglo XX por dos matemáticos alemanes proporciona la base teórica para que este algoritmo funcione: los teoremas de Perron y Frobenius.

Oskar Perron[15] en 1907 demostró uno de los teoremas más brillantes del Álgebra Lineal que nos ilumina sobre cómo son los vectores propios y los valores propios de una matriz en la que todas sus entradas son positivas. Con esta hipótesis el Teorema de Perron daría respuesta afirmativa a todos nuestros requisitos. Ciertamente ninguna entrada de S es negativa pero las entradas nulas son abundantísimas. No podemos aplicar el Teorema de Perron.

No es fácil ni trivial extender el Teorema de Perron a matrices cuyos términos sean números reales positivos o nulos. No obstante, Ferdinand Georg Frobenius[16] acometió esta tarea. Trabajó en ella entre 1908 y 1912 y como resultado obtuvo este eminente teorema. Se advierte que esta presentación es muy resumida; se puede consultar un enunciado completo y una demostración rigurosa en muchos textos de Álgebra Lineal, por ejemplo el de C. D. Meyer[17]. En el texto de B. Huppert[18] aparece la demostración de H. Wielandt.



Oskar Perron

Ferdinand George Frobenius

15. Oskar Perron fue un matemático alemán nacido el 7 de mayo de 1880 en Frankenthal y fallecido el 22 de febrero de 1975 en Munich.

16. Ferdinand Georg Frobenius, nacido en Charlottemburg el 26 de octubre de 1849 y fallecido en Berlín el 3 de agosto 1917, es uno de los matemáticos más excelsos del siglo XX, con aportaciones fundamentales en la Teoría de Grupos, en el Álgebra Lineal y en la teoría de ecuaciones diferenciales.

17. C. D. Meyer: *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, 2000.

18. B. Huppert: *Angewandte Lineare Algebra*, Walter de Gruyter & Co., Berlin, 1990.

Teorema 2.1 (Perron-Frobenius). *Si A es una matriz cuadrada irreducible cuyos términos son números reales positivos o nulos, se verifican las siguientes afirmaciones.*

1. *La matriz A posee un valor propio λ –denominado* raíz de Perron– *tal que*

   a) *λ es un número real positivo,*

   b) *si μ es otro valor propio de A (posiblemente complejo) entonces $|\mu| \leq \lambda$,*

   c) *si existen k valores propios de módulo máximo, entonces son las soluciones de $x^k - \lambda^k = 0$,*

   d) *λ es un valor propio simple de A: esto implica que todos los vectores propios de A asociados a λ son múltiplos escalares de un único vector básico.*

2. *Existe un único vector propio $X = (x_1, ..., x_n)'$ de A con valor propio λ, esto es $AX = \lambda X$, cuyas coordenadas son todas positivas y tal que*

   $$\sum_{i=1}^{n} x_i = 1.$$

   Este vector recibe el nombre de vector propio de Perron.

3. *Si Y es otro vector propio no negativo de A, entonces $Y = tX$ para algún escalar real positivo t.*

El vector de Perron sería la solución a nuestro problema. Pero dos dificultades esenciales se interponen en nuestro camino. El Teorema de Perron-Frobenius postula la existencia del vector propio de Perron para matrices irreducibles. Por otro lado, el hecho de que puedan existir *k* valores propios del mismo módulo no es deseable.

¿Qué es una matriz irreducible? Decimos que una matriz cuadrada es *reducible* si tras una permutación de las filas y de las columnas, la misma en ambos casos, se descompone en cajas

$$\begin{pmatrix} A_{11} & 0 \\ A_{12} & A_{22} \end{pmatrix}$$

donde A_{11}, A_{22} son matrices cuadradas. Si no existe tal permutación, diremos que *A* es *irreducible*.

Un simple vistazo a la matriz *M* de nuestro ejemplo, y consiguientemente a las matrices *H* y *S* construidas a partir de ella, muestra que son reducibles. La caja de ceros que aparece en *M*, y que heredan *H* y *S*, se debe a que no hay ningún camino dirigido que comience en uno de los vértices *P* o *Q* y que termine en uno de los restantes vértices. En términos de la teoría de grafos, el grafo de nuestro ejemplo no es

fuertemente conexo. Un grafo es conexo si dos vértices cualesquiera están unidos por un camino de arcos y un grafo dirigido es *fuertemente conexo* si cada vértice es accesible desde cualquier otro vértice mediante un camino dirigido. De hecho se prueba el siguiente resultado.

Teorema 2.2. *Un grafo dirigido es fuertemente conexo si y sólo si su matriz de adyacencia es irreducible.*

Esto significa que para poder aplicar el Teorema de Perron-Frobenius al grafo dirigido asociado a la red, y por tanto para que todo funcione, la condición necesaria y suficiente es que desde cualquier página se pueda alcanzar cualquier otra navegando a través de enlaces. Pero esto no es posible. ¡La *web* ni siquiera es conexa!

2.4. Modificación final: la matriz de Google

Volvemos a reflexionar sobre la manera en que nuestro navegante aleatorio se mueve por la red. La conducta de nuestro navegante queda descrita por S: o bien sigue por uno de los enlaces de la página en la que se encuentra, o bien sigue por otra página escogida al azar si se encuentra en una página colgada. Variamos un punto este proceder.

Escogemos un parámetro α entre 0 y 1 y consideramos que en cada página P el navegante aleatorio se aburre con una probabilidad igual a $1 - \alpha$ y cuando esto ocurre el navegante elige una página al azar, tanto si está vinculada a la página P como si no lo está, y sigue su navegación[19].

Notemos que, como antes, si P es una página colgada, el navegante escoge una página al azar de entre todas las de la red.

En otras palabras, el navegante aleatorio:

- con probabilidad α se guía por la matriz S, y
- con probabilidad $1 - \alpha$ escoge una página al azar.

19. En algunos textos en inglés el parámetro α recibe el nombre de *dumping factor*.

La matriz de Google es

$$G = \alpha S + (1 - \alpha) E_n$$

donde designamos con E_n la matriz en la que todas las entradas son iguales a $\frac{1}{n}$ y n es el número total de páginas *web*.

$$E_n = \frac{1}{n} \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{pmatrix}$$

¿Hemos alcanzado nuestro objetivo? Si $\alpha < 1$ entonces todas las entradas de G son estrictamente positivas; por tanto G es una matriz irreducible y cumple las hipótesis del teorema de Perron-Frobenius (de hecho, cumple incluso las del Teorema de Perron) y necesariamente se verifican todas las conclusiones. Los maravillosos resultados teóricos que hemos presentado nos aseguran la existencia de ese vector de importancias que hemos buscado con tanto ahínco: es el vector de Perron de la matriz G de Google.

Pero ¿cómo lo calculamos? Consignemos que el número total de páginas *web* es descomunal; en la actualidad se calcula o, mejor, se intuye que el número de páginas *web* «visibles» ronda los cincuenta mil millones[20] (o sea $n = 5 \times 10^{10}$). Buscar vectores propios directamente, por el método teórico, en una matriz tan descomunal es tarea imposible. La capacidad de cálculo de nuestros ordenadores es, como diría Borges, vastísima, pero es finita.

Afortunadamente la misma estructura de la matriz G nos proprociona el método de cálculo elegante y, sobre todo, factible: la matriz G es *estocástica*.

20. Es difícil establecer el número exacto de páginas *web*. En <http://stooge.hubp ages.com/hub/How-many-webpages-do-you-think-actually-exist-on-the-Internet> hay algunas estimaciones de las que hemos escogido la del texto. En esta misma referencia el autor especula, sin apoyo documental, que la «*internet* invisible», esto es, las páginas no indizadas, es, al menos, 100 veces mayor.

Es ilustrativo comprobar que a la consulta «1» en el buscador Google, la respuesta nos informa que contiene aproximadamente 25.270.000.000 de resultados.

2.5. Un proceso de Markov para obtener el vector de importancias

Si escribimos $G = (g_{ij})$, se tiene

$$g_{ij} = \begin{cases} \frac{\alpha}{n_j} + \frac{1-\alpha}{n} & \text{si hay un enlace de } P_j \text{ a } P_i, \\ \frac{1-\alpha}{n} & \text{si } P_j \text{ no es página colgada pero no hay enlace de } P_j \text{ a } P_i, \\ \frac{1}{n} & \text{si } P_j \text{ es una página colgada,} \end{cases}$$

Un cálculo sencillo nos muestra que la suma de los componentes de todas las columnas es 1. Por tanto interpretamos cada columna como una distribución de probabilidad de forma que nuestro navegante aleatorio tiene una probabilidad g_{ij} de trasladarse desde la página P_j hasta la página P_i. Si todas las columnas de una matriz son vectores de componentes no negativas que suman uno la matriz recibe el nombre de *matriz estocástica*. Es decir, la matriz G es una matriz estocástica.

Ese razonamiento puede formularse en términos de las llamadas *cadenas de Markov*. En cada momento el navegante aleatorio

1. elige a qué página se traslada teniendo en cuenta únicamente la página en la que se encuentra y no las páginas anteriores que ha ido visitando, es decir sin memoria, y
2. la probabilidad de pasar de la página P_j a la página P_i permanece siempre constante.

A. A. Markov

Estas son precisamente las características que definen las *cadenas de Markov*[21].

Es decir, la matriz G es una matriz estocástica y es la *matriz de transición* de nuestra cadena de Markov. Según esta interpretación el escalar que ocupa el lugar (i, j) en la matriz potencia G^m se interpreta como la probabilidad de que el navegante aleatorio pase desde la página P_j a la página P_i en m pasos.

En consecuencia las páginas más importantes según esta interpretación son las que tienen una probabilidad mayor de ser visitadas por un navegante aleatorio.

21. *Andrei Andreyevich Markov* es un matemático ruso nacido en Ryazan, Rusia, el 14 de junio de 1856 y fallecido en Petrogrado (ahora San Petersburgo), Rusia, el 20 de julio de 1922.

Al aplicar el Teorema de Perron-Frobenius a una matriz estocástica con todas sus entradas positivas encontramos una propiedad muy significativa.

Teorema 2.3. *Si G es una matriz estocástica con todas sus entradas positivas, la raíz de Perron de G es 1.*

Además la raíz de Perron es el único valor propio de G de módulo 1.

Para aplicar el método de las potencias o método de Markov procedemos de la siguiente manera:

1. Se escoge un vector de probabilidades I_0 como «candidato» a I.

2. Se calculan sucesivamente:

$$\begin{aligned} I_1 &= GI_0, \\ I_2 &= GI_1 = G^2 I_0, \\ I_3 &= GI_2 = G^3 I_0, \\ &\ldots \end{aligned}$$

y, en general, $I_{k+1} = GI_k = G^{k+1}I_0$.

Para que el método de las potencias sea efectivo y nos aporte el resultado que buscamos esta sucesión debe ser convergente al vector de Perron de G, que es lo que buscamos, independientemente de cuál sea el vector I_0 tomado para iniciar el proceso. El siguiente teorema nos proporciona la respuesta completa.

Teorema 2.4. *Si G es una matriz estocástica con todas sus entradas positivas, entonces*:

1. *la sucesión $\{G^k\}$ es convergente y si $\lim_{k\to\infty} G^k = L$, entonces L es una matriz estocástica;*

2. *cada columna de L es el vector de Perron I de G;*

3. *para cualquier vector de probabilidades I_0 se verifica que*

$$\lim_{k\to\infty} (G^k I_0) = I.$$

Para que nos entendamos, hemos alcanzado el objetivo que nos habíamos propuesto: el vector I, el vector de Perron de G, que aparece como límite de los vectores que van apareciendo al aplicar el método de las potencias de Markov, es el vector de importancias que buscamos desde el principio.

Merece la pena que resaltemos la clave que nos permite probar que la sucesión de potencias G^m es convergente: Los valores propios de la matriz potencia G^m son potencias de los valores propios de G. Como, a excepción del valor de Perron que es 1, todos los demás valores propios tienen módulo estrictamente menor que 1, sus potencias sucesivas van haciéndose más y más pequeñas, esto es, van tendiendo a cero. Como consecuencia, además, se obtiene que la rapidez de la convergencia se acentúa cuanto menor sea el segundo valor propio de G.[22] Procede apuntar aquí que si los valores propios de la matriz estocástica S son $\{1, \lambda_2, ..., \lambda_n\}$, entonces los valores propios de la matriz estocástica $G = \alpha S + (1 - \alpha)E_n$ son $\{1, \alpha\lambda_2, ..., \alpha\lambda_n\}$[23]. De hecho, el error que se comete en k iteraciones es menor o igual que $2\alpha^k$. Por tanto cuanto más pequeño sea α, más rápida será la convergencia.

Pero por otro lado, si α se acerca a 0 interpretamos que el navegante aleatorio se aburre con una probabilidad más alta y se olvida de la estructura de la red. De hecho, si $\alpha = 0$, entonces G es la matriz es la de un navegante en la que en cada página escoge al azar la próxima tanto si hay enlace como si no lo hay.

Tras muchos experimentos Google adoptó

$$\alpha = 0{,}85$$

Con este valor, en un número de entre 50 y 100 iteraciones obtenemos una buena aproximación de I.

Queda, por tanto, determinada la llamada *matriz de* Google:

$$G = 0{,}85S + 0{,}15E_n$$

22. Para ilustrar esta cuestión de la importancia del segundo valor propio, y no sólo para ello, es muy recomendable el trabajo de, K. Bryant y T. Leise: *The $25,000,000,000 eigenvector, The Linear Algebra behind Google*. SIAM Rev., 48(3) (2006) 569-581.

También merece la pena consultar el interesante trabajo de T. H. Haveliwala y S. D. Kamvar: *The Second Eigenvalue of the Google Matrix*. Technical Report, Stanford, 2003. <http://ilpubs.stanford.edu:8090/582/1/2003-20.pdf>.

23. Se puede encontrar una demostración de este hecho en el Theorem 4.7 del excelente, y ameno, libro de A. Langville y C. Meyer: *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

El vector de Perron de G es precisamente el vector de importancias de *PageRank*. Su cálculo efectivo requiere una cadena de Markov cuya matriz de transición sea G.[24]

Despejemos finalmente algunas dudas sobre la computación efectiva con una matriz de tamaño tan colosal mediante este método. La matriz S que apareció en un estado intermedio de este razonamiento, es una suma: $S = H + A$ donde A es una matriz tal que:

- las columnas no nulas de H son columnas nulas en A;
- las columnas nulas de H son columnas donde todos los coeficientes son $\frac{1}{n}$ en A.

Por tanto se tiene

$$GI_k = 0{,}85HI_k + 0{,}85AI_k + 0{,}15E_nI_k$$

- La mayoría de las entradas de H son nulas. De hecho la media de enlaces de las páginas *web* es de 10; es decir cada columna tiene sólo alrededor de 10 entradas no nulas. Por tanto el cálculo HI_k requiere sólo de un número de cálculos que ronda los 10.

24. Como curiosidad apuntemos que la matriz de Google asociada a la diminuta red tomada como ejemplo es:

$$G = \begin{pmatrix} 3/100 & 1/5 & 3/100 & 3/100 & 3/100 \\ 97/400 & 1/5 & 91/200 & 3/100 & 3/100 \\ 97/400 & 1/5 & 3/100 & 3/100 & 3/100 \\ 97/400 & 1/5 & 3/100 & 3/100 & 22/25 \\ 97/400 & 1/5 & 91/200 & 22/25 & 3/100 \end{pmatrix}$$

Debido al pequeño tamaño de esta matriz podemos calcular directamente el vector de Perron mediante el uso de MATHEMATICA.8. El resultado es:

$$I' = (I(R), I(S), I(T), I(P), I(Q))' = \left(\frac{9600}{226007}, \frac{16587}{226007}, \frac{11640}{226007}, \frac{3431860}{8362259}, \frac{3530800}{8362259} \right)'$$

que provoca la siguiente ordenación de las cinco páginas:

$$I(Q) > I(P) > I(S) > I(T) > I(R) \Rightarrow Q \succ P \succ S \succ T \succ R.$$

Tras sesenta y ocho iteraciones de un proceso de Markov con matriz de transición G encontramos la siguiente aproximación del vector de Perron:

$$I' \approx (0{,}0424766,\ 0{,}0733915,\ 0{,}0515028,\ 0{,}410399,\ 0{,}42223)'$$

El error cometido al tomar los valores aproximados está acotado por

$$2 \times 0,\ 8568 \approx 3{,}17 \times 10^{-5}.$$

De hecho se puede calcular directamente que el error es menor que $4{,}1 \times 10^{-7}$.

- El cálculo de AI_k nos proporciona el sumando con la importancia de las «páginas colgadas». Como todas las filas de A son iguales para evaluar este sumando sólo es necesario un cálculo.

- Algo similar ocurre con E_nI_k; su cálculo nos proporciona el sumando con la importancia de todas las páginas *web*. Como en el caso anterior, dado que todas las filas de E_n son iguales, para evaluar este sumando sólo es necesario un cálculo.

3. Epílogo

La fijación del valor 0,85 para el parámetro α es una decisión puramente empírica. No es la única: la elección de la matriz E_n también lo es: podía utilizarse otra matriz estocástica de rango 1 cuyas columnas fuesen una distribución de probabilidad no homogénea. Por el contrario, la manera de soslayar el problema de la recursión inherente al postulado principal de *PageRank* mediante el uso de vectores propios no es empírica. Ni lo es el éxito de las cadenas de Markov, fundamentado en el uso de los Teoremas de Perron-Frobenius, para determinar el vector de importancias. Los creadores de este pasmoso, y lucrativo, algoritmo conocían con profundidad las matemáticas involucradas y pudieron por tanto sortear las dificultades y afinar su razonamiento para que su construcción fuese eficaz. No se contentaron con usar un par de fórmulas sino que, entendiendo su génesis y su fundamentación, fueron capaces de confeccionar un modelo útil que cumpliese estrictamente las hipótesis teóricas para tener garantizado el resultado.

En ocasiones se destina a las Matemáticas, y en particular al Álgebra, un papel ancillar en el mundo de las tecnologías. A veces se escuchan afirmaciones hinchadas de soberbia académica que proclaman aquello de «ya explicaremos nosotros las matemáticas pues nosotros sabemos qué hace falta» seguido de aquello otro de «es esencial que las asignaturas tengan una componente práctica». No seré yo quien niegue la importancia de la praxis pero de nada sirve saber cómo se hace algo que no se sabe qué es.

Las auténticas prácticas de una titulación durarán toda la vida profesional. Si en estos años de aprendizaje no proporcionamos una sólida base teórica que pueda ser aplicada en los múltiples desafíos con los que nuestros alumnos se enfrentarán en su devenir profesional, nuestros egresados tal vez sean capaces de leer los manuales de instrucciones, pero serán incapaces de escribirlos. Renunciar al conocimiento puro es condenar nuestro futuro intelectual al raquitismo.

He dicho.