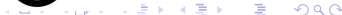
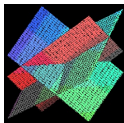


El Álgebra Lineal detrás de Google

Carlos D'Andrea

I Congreso Nacional de Estudiantes de Matemática
Corrientes, Julio 2012



Facultad de Matemáticas

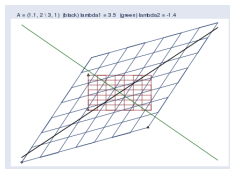
Universidad de Barcelona

- Licenciatura en Matemática
- Master en Matemática Avanzada
- Doctorado en Matemática
- Otros masters y doctorados
- Ingeniería Informática



Programa de Álgebra Lineal (Ingeniería Informática)

- Polinomios
- Números complejos
- Sistemas lineales de ecuaciones
- Matrices
- Espacios vectoriales
- Subespacios, transformaciones lineales, etc, etc, etc.
- **Valores y vectores propios – Diagonalización**

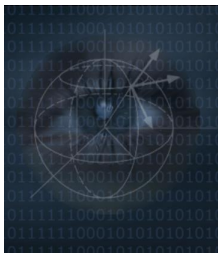


¿Para qué nos hacen estudiar todo esto???



Álgebra Lineal en la Informática

- Programación gráfica
- Teoría de grafos (redes sociales,...)
- Elaboración eficiente de filtros
- Inteligencia artificial
- Visión por ordenador
- ...



¡Todo esto lo verán después!



El álgebra lineal detrás de

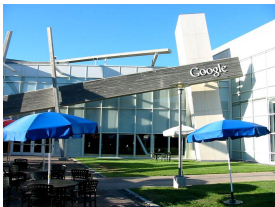


- **Google** es una variación de la palabra “googol”, que es el número 10^{100}
- Es un buscador de internet
- Fue diseñado en 1998 por dos alumnos de doctorado en informática en Stanford: Sergei Brin y Lawrence Page
- Atiende alrededor de 200.000.000 de consultas diarias, tiene más de 54.000 empleados en todo el mundo



Una gran familia

- El “campus” de Google (Googleplex) se encuentra en Menlo Park, Silicon Valley, California
- Ocupa casi 50.000 metros cuadrados
- Reclutamiento constante de jóvenes talentos en todo el mundo



Google's got Talent



Congratulations. You've made it to level 2. Go to **www.Linux.org** and enter *Bobsyouruncle* as the login and the answer to this equation as the password.

$$f(1)= 7182818284$$

$$f(2)= 8182845904$$

$$f(3)= 8747135266$$

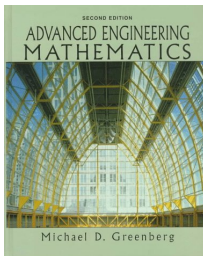
$$f(4)= 7427466391$$

$$f(5)= \underline{\hspace{2cm}}$$

¿Cómo se diseña un buscador de internet?

Es un problema de **ingeniería matemática**:

- 1 un buen conocimiento del contexto
- 2 un modelo matemático que lo explique
- 3 una cuidadosa y eficiente implementación



Trabajo básico de un buscador de internet

- “Censar” las páginas de internet de acceso público
- Indexar los datos censados de acuerdo a su importancia con respecto a las palabras claves
- Ordenar estos datos de acuerdo a su importancia con respecto a las palabras claves



¡También se requiere resistencia a la manipulación!



Web

Results 1 - 10 of about 969,000 for [miserable failure](#). (0.06 seconds)

[Biography of President George W. Bush](#)

Biography of the president from the official White House web site.

www.whitehouse.gov/president/gwbbio.html - 29k - [Cached](#) - [Similar pages](#)

[Past Presidents](#) - [Kids Only](#) - [Current News](#) - [President](#)

[More results from www.whitehouse.gov »](#)

[Welcome to MichaelMoore.com!](#)

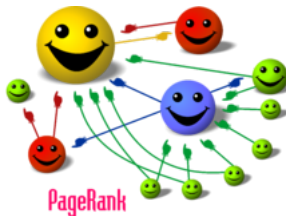
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, ...

www.michaelmoore.com/ - 35k - Sep 1, 2005 - [Cached](#) - [Similar pages](#)

[BBC NEWS | Americas | 'Miserable failure' links to Bush](#)

El algoritmo “PageRank”

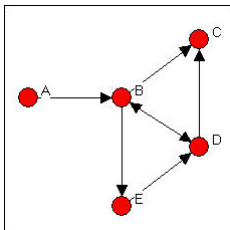
- Califica páginas indexadas de acuerdo a su “importancia” dentro de la red
- Marca registrada de Google
- Lleva su nombre debido a su inventor Larry Page



El modelo “PageRank”

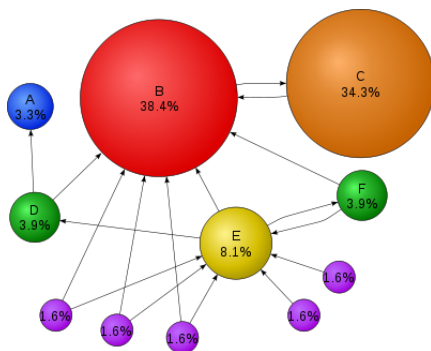
El universo de páginas de internet públicas es un gran **grafo dirigido** donde

- cada página web es un nodo
- hay una arista orientada entre páginas que citan a otras páginas



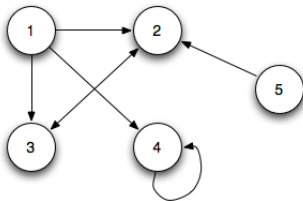
“Postulado” PageRank

La importancia x_j de la página P_j es proporcional a la suma de las importancias de las páginas que enlazan con P_j



El álgebra lineal entra en acción

M es la matriz de adyacencia del grafo de las páginas de internet



	1	2	3	4	5
1	0	1	1	1	0
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	0	1	0
5	0	1	0	0	0

El postulado Pagerank implica

$$M^t \cdot x = \lambda \cdot x$$

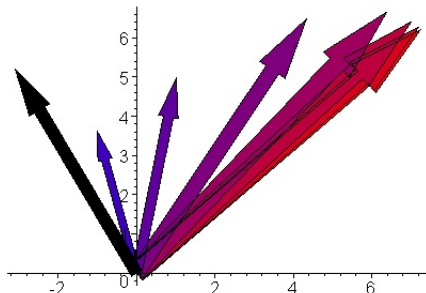
¡Vectores y valores propios!

$$\mathbf{M}^t \cdot \mathbf{x} = \lambda \cdot \mathbf{x}$$

- λ es la constante de proporcionalidad $\leftrightarrow$ un **valor propio** de $\mathbf{M}^t$
- $\mathbf{x} = (x_1, x_2, \dots, x_N)$ es el vector de “importancias” de las páginas censadas $\leftrightarrow$ un **vector propio** de $\mathbf{M}^t$ (asociado a λ)

Todo muy bonito, pero...

- ¿Por qué debería tener valores propios **reales** M^t ?
- ¿Cual de ellos elijo?
- ¿Por qué habría de haber vectores propios todos **positivos**?
- ¿Algún tipo de **unicidad**???



Teorema 1 (Perron, 1907)

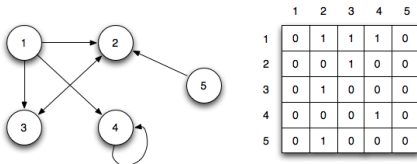
Si M tiene todas sus coeficientes **positivos**, entonces

- existe un valor propio **simple** $\lambda > 0$ tal que $M^t \cdot x = \lambda \cdot x$, con $x > 0$;
- este valor propio es mayor, en módulo, que todos los demás valores propios de la matriz;
- cualquier otro vector propio **positivo** de M^t es un múltiplo escalar de x



Pero...

Nuestra matriz M está **MUY** lejos de ser positiva



¿Qué hacemos?



Teorema 2 (Frobenius, 1908–192)

Supongamos que M tiene entradas **no negativas** y además es **irreducible**. Entonces

- existe un valor propio simple $\lambda > 0$ tal que $M^t \cdot x = \lambda \cdot x$, con $x > 0$;
- este valor propio es **mayor o igual**, en módulo, que todos los demás valores propios de la matriz;
- cualquier otro vector propio **positivo** de M^t es un múltiplo escalar de x



Matrices irreducibles

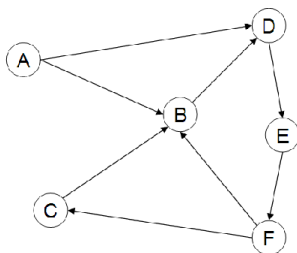
Una matriz cuadrada se dice **irreducible** si no existe ninguna permutación de sus filas y columnas que la transforme en

$$\left(\begin{array}{c|c} M_{11} & A_{12} \\ \hline \mathbf{0} & M_{22} \end{array} \right),$$

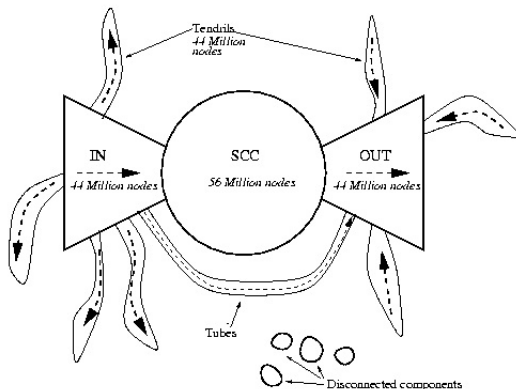
con M_{11} y M_{22} matrices cuadradas

Matrices irreducibles = grafos “fuertemente” conexos

Si se trata de la matriz de incidencia de un grafo dirigido, ser irreducible significa que puedo ir desde cualquier nodo a otro por un camino (dirigido)



¿Es el grafo de internet fuertemente conexo?



¡Ni siquiera es conexo!

Solución “a la Google” $\rightarrow$ ¡Matemática aplicada!

“Perturbamos” la matriz $\mathbb{M}$

$$c \mathbb{M} + (1 - c) \mathbb{U} \mapsto \mathbb{M}$$

donde

■ c es un parámetro entre 0 y 1 ($c_{google} \approx 0,85$)

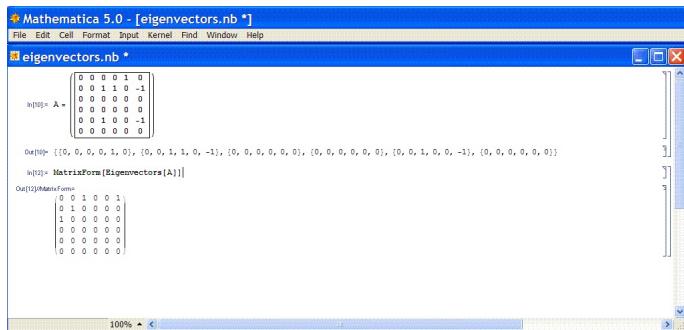
$$\blacksquare \mathbb{U} = \begin{pmatrix} \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{N} & \frac{1}{N} & \cdots & \frac{1}{N} \end{pmatrix}$$

Del existencialismo al Cálculo

No se necesitan

- Polinomios característicos
- Cálculos de raíces
- Descomposición en subespacios invariantes

¡Álgebra Lineal Numérica!



The screenshot shows the Mathematica 5.0 interface with a notebook titled "eigenvectors.nb". The input cell contains the definition of matrix A and the command to find its eigenvectors. The output cell shows the resulting list of eigenvectors and a matrix form of them.

```
In[10]:= A = 
$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

```

```
Out[10]= {{0, 0, 0, 0, 1, 0}, {0, 0, 1, 1, 0, -1}, {0, 0, 0, 0, 0, 0}, {0, 0, 0, 0, 0, 0}, {0, 0, 1, 0, 0, -1}, {0, 0, 0, 0, 0, 0}}
```

```
In[12]:= MatrixForm[Eigenvectors[A]]
```

```
Out[12]= MatrixForm[
$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$
]
```

Método de las potencias (usado por Google)

Si hay un único valor propio λ de módulo máximo entonces, consideremos la siguiente sucesión

■ $\mathbf{x}_0$ = cualquier vector de $\mathbb{R}^N$

■ $\mathbf{x}_{n+1} = \frac{\mathbf{M}^t \cdot \mathbf{x}_n}{\|\mathbf{M}^t \cdot \mathbf{x}_n\|}$

Entonces

$$\lim_{n \rightarrow \infty} \mathbf{x}_n = \mathbf{x}$$

$$\lim_{n \rightarrow \infty} \frac{\|\mathbf{M}^t \cdot \mathbf{x}_n\|}{\|\mathbf{x}_n\|} = \lambda$$

con probabilidad 1

La misma idea para otros problemas

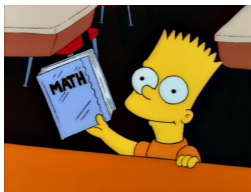
- Clasificación para las eliminatorias de la NBA
- Modelos de evolución probabilística
- Dinámica de poblaciones
- Modelos económicos



- El objetivo de Brin y Page era que al menos una de las diez primeras páginas que se muestren contenga información útil para el que consulta ... ¿Tuvieron éxito?
- En 2004 el valor de Google en el mercado era de alrededor de 25.000.000.000 USD
- El algoritmo PageRank fue patentado por la Universidad de Stanford, y Google tiene derechos exclusivos sobre esa patente.
- Stanford recibió acciones por esa patente que fueron vendidos en 2005 por 336.000.000 USD
- Desde febrero de 2011 Google utiliza combinadamente los algoritmos PageRank y Google Panda

¿Qué hemos aprendido hoy?

- Grafos y sus propiedades → Teoría de Grafos
- Matrices con entradas positivas → Matrices estocásticas
- Cálculo computacional de vectores y valores propios → Álgebra Lineal Numérica
- Teoremas de Perron y Frobenius → Análisis funcional
- PageRank y Panda → Algoritmos de búsqueda



Para saber más

- “El secreto de Google y el Álgebra Lineal”, P. Fernández, Bol. Soc. Esp. Mat. Apl. 30 (2004), 115–141
- “The \$25,000,000,000 Eigenvector: The Linear Algebra behind Google”, Kurt Bryan & Tanya Leise, Siam Review 48 (3), 569–581, 2006
- “Les Matemàtiques de Google: l’algorisme PageRank”, Joan Gimbert, Butlletí de la Societat Catalana de Matemàtiques, Vol 26, 1, 211, 29–55



Muchas gracias

<http://atlas.mat.ub.es/personals/dandrea>

