Clasificación de páginas web: Google

.

Rafael Bru

Institut de Matemàtica Multidisciplinar
Univ. Politècnica de València

http://personales.upv.es/rbru/

Título

- Surfista aleatorio
- Pagerank
- The world’s largest matrix computation

Buscadores web

Dos fases:

1. Obtener la información pedida
2. Clasificarla

Año 1998 800.10^6 páginas web

Año 2004 3000.10^6 páginas web

IMPORTANCIA DE PRESENTAR LA INFORMACIÓN

El inicio de Google

- Sergey Brin y Larry Page (1995)
- Estudiantes informáticos en Stanford Univ.
- Congreso de 1998 “The PageRank citation ranking: Bringing order to the web”
- IDEA: asignar a cada página web la probabilidad de ser buscada por una persona.
- Paseo aleatorio de un surfista de una página a otra.

Surfista aleatorio

- £Cómo se organiza ese paseo?
- Se pasa de una página a otra que tenga conexión de forma aleatoria.
- Esto es parte de un proceso de Markov.
- Ejemplo: tres páginas web: v_1, v_2, v_3
  - v_1 tiene conexión con v_2 y v_3
  - v_2 tiene conexión con v_2
  - v_3 no tiene conexión
  - Matricialmente

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Surfista aleatorio: Modelo Brin y Page

v_1 tiene conexión con v_2 y v_3

v_2 tiene conexión con v_1 v_3 y v_4

v_3 tiene conexión con v_2 y v_5

v_4 tiene conexión con v_3

v_5 tiene conexión con v_1 y v_4

Misma probabilidad para todas las posibles conexiones.

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

Modelo de Markov

Definición 1 *Un modelo de Markov es un sistema que evoluciona aleatoriamente con el tiempo pasando por diferentes estados.*

- Grafo formado por las páginas web y sus conexiones
- G(WWW)={V,C} siendo
  - Conjunto de vértices: V={v_i: i es la $i-sima$ página web}
  - Conjunto de arcos: C={(i,j): v_i unida con v_j }

Modelo de Markov

- El elemento p_{ij} representa la probabilidad de pasar del estado i al estado j.
- Inicialmente partimos de $\pi^{(0)} = [\pi_1^{(0)}, \pi_2^{(0)}, \ldots, \pi_n^{(0)}]$.
- La probabilidad de que el proceso este en el estado j en la siguiente etapa será

$$\pi_j^{(1)} = \pi_1^{(0)} p_{1j} + \pi_2^{(0)} p_{2j} + \cdots + \pi_n^{(0)} p_{nj}, \qquad j = 1, 2 \ldots, n$$

- Matricialmente

$$\pi^{(1)} = \pi^{(0)} P$$

- P se conoce como matriz de transición de estados.

Modelo de Markov

- La componente $j-esima$ del vector fila $\pi^{(k)}$ denota la probabilidad de que el proceso esté en el estado j, después de k etapas.
- En general,

$$\boxed{\pi^{(k)} = \pi^{(k-1)} P}$$

- El vector $\pi^{(k)}$ se conoce como el vector de distribución o de probabilidad en la etapa k.
- Solución del proceso Markov en función vector inicial

$$\boxed{\pi^{(k)} = \pi^{(0)} P^k}$$

Vector estacionario

Definición 2 *Un vector de distribución π se dice que es* ***estacionario*** *si satisface*

$$\pi = \pi\bar{P}$$

- El vector π tiene sus componentes no negativas.
- Éstas indican la probabilidad de que el sistema se encuentre en un estado i después de trascurrir un largo periodo de pasos o etapas.
- Notar que encontrar el vector estacionario es equivalente a encontrar el vector propio a la izquierda asociado al valor propio 1.

Vector estacionario

- Existencia
- Unicidad

Existencia vector estacionario

Teorema 1 *Sea $\bar{P}$ una matriz estocástica. Entonces:*

- *la matriz $\bar{P}$ tiene como valor propio el uno (1).*
- *Al valor propio 1 le corresponde un vector propio a la izquierda no negativo.*

Existencia del vector estacionario

Teorema 2 *Sea A una matriz cuadrada no negativa. Entonces:*

- *$\rho(A)$, el radio espectral de A, es un valor propio.*
- *A tiene un vector propio no negativo correspondiente a $\rho(A)$.*
- *A^T tiene un vector propio no negativo correspondiente a $\rho(A)$.*

Unicidad del vector estacionario

- El teorem anterior no asegura la unicidad ya que la matriz

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

- tiene dos vectores propios no negativos

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{y} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Surfista aleatorio

Volvamos a estudiar la matriz del modelo de Brin y Page

- La matriz del primer ejemplo

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

- no es estocástica.
- Ocurre cuando una página web no tiene conexiones a otras.
- Se llaman DANGLING nodes (colgar, suspender)

Surfista aleatorio

- Posible solución: Suponer que desde esa página se puede pasar a cualquier otra con misma probabilidad,
- Es decir, reemplazar los ceros por $1/3$,
- Quedaría

$$\overline{P} = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 1 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

que es una matriz estocástica.

Surfista aleatorio

- Tiene un único vector estacionario?
- Veamos los valores propios de una matriz estocástica
- Cómo $\sigma(A) = \sigma(A^T)$, trabajamos con la traspuesta,
- $$e^T \bar{P} = 1 \cdot e^T, \quad \text{donde} \quad e^T = (1, 1, \ldots, 1)$$
- En nuestro ejemplo
  - Valores propios $\sigma(\bar{P}^T) = \{1, 0.6076, -.2743\}$
  - Vectores propios

    $u_1 = (0.5774, 0.5774, 0.5774)$

    $u_2 = (-0.6354, 0, -0.7722)$

    $u_3(-0.8767, 0, 0.4810)$

Surfista aleatorio: reducibilidad

La matriz estocástica $\bar{P}$ es semejante a la matriz

$$Q^T \bar{P} Q = \left[\begin{array}{cc|c} 0 & 1/2 & 1/2 \\ 1/3 & 1/3 & 1/3 \\ \hline 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{c|c} A & B \\ \hline O & C \end{array}\right]$$

donde $Q = [e_1, e_3, e_2]$

Definición 3 *Se dice que una matriz P es **reducible** si existe una matriz de permutación Q tal que $Q^T PQ$ se escribe por bloques como anteriormente.*

Surfista aleatorio: irreducibilidad

Definición 4 *Una matriz que no es reducible se dice* ***irreducible***

Ejemplo: La matriz

$$\left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \end{array}\right]$$

es irreducible.

Valores propios $\{1,\ 1/2+i\sqrt{3}/2,\ 1/2-i\sqrt{3}/2\}$

Surfista aleatorio: unicidad

Teorema 3 (Frobenius, 1912) *Sea $\overline{P} \geq 0$ una matriz irreducible. Entonces*

- *$\overline{P}$ tiene un valor propio positivo real, llamado raíz Perron, igual a su radio espectral.*
- *A la raíz Perron le corresponde un vector propio positivo $x > 0$, vector Perron.*
- *El radio espectral aumenta cuando cualquier elemento de la matriz aumenta.*
- *La raíz Perron es simple.*
- *Cualquier valor propio de módulo igual a su radio espectral es simple.*
- *Cualquier vector propio positivo es un múltiplo de x.*

Surfista aleatorio: modelo de Brin y Page

Observación: La matriz

$$\overline{\overline{P}} = \alpha \overline{P} + (1-\alpha) e e^T / n,$$

donde $e^T = (1, 1, \ldots, 1)$, es:
(i) estocástica
(ii) irreducible
(iii) Si $\{1, \lambda_1, \lambda_2, \ldots, \lambda_n\}$ es el espectro de $\overline{P}$, entonces el espectro de $\overline{\overline{P}}$ es $\{1, \alpha\lambda_1, \alpha\lambda_2, \ldots, \alpha\lambda_n\}$

Esta matriz $\overline{\overline{P}}$ es la matriz de transición de Brin y Page

Bibliografía

A. Berman, y R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*, SIAM, Filadelfia, Pensilvania, Estados Unidos. Reimpresión, 1994.

S. Brin y L. Page.The anatomy of a large-scale hypertextual Web search engine, Computer *Networks and ISDN Systems*, 33: 107-117, 1998.

S. Brin y L. Page. The PageRank citation ranking: bringing order to the web, *Technical report* 1999-0120, Computer Science deptartment, Stanford Ubiversity, 1999.

A. N. Langville y C. D. Meyer. Deeper inside PageRank. *Internet Mathematics, Vol. 1(3):335-380. 2005.*

C. B. Moler. The World's Largest Matrix Computation Google's PageRank is an eigenvector of a matrix of order 2.7 billion. *MATLAB News & Notes* . 2002.

W. J. Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, Princeton, Nueva Jersey, Estados Unidos. 1994.

R. S. Varga. *Matrix Iterative Analysis*. Springer. Berlín, Alemania. 2^a edición, 2000.

Cálculo del vector PageRank

- Pagerank
- The world's largest matrix computation

PageRank: técnicas de cálculo

Valor de cada página web = componentes del vector estacionario: PageRank.

- Problema de valor y vector propio: $\pi^T = \pi^T \overline{\overline{P}}$.
- Problema de sistema lineal: $\pi^T \left(I - \overline{\overline{P}}\right) = 0^T$.

NOTAS:

1. Recordar que la suma de las componentes de $\pi^{(0)T}$ es 1 (probabilidades).
2. Entonces, la suma de las componentes de $\pi^{(k)T} = \pi^{(k-1)T})\overline{\overline{P}}$ es 1.
3. Por tanto, en el cálculo de π hay que exigir que $\pi^T e = 1$.

Método de la potencia

Como problema de valor y vector propio se reduce al **vector propio asociado al valor propio más grande**

Para ello, se usa el método de la potencia.

Algoritmo. *Método iterativo de la potencia para la matriz* M. Entrada: vector inicial $v^{(0)}$ y matriz M. Salida: vector propio v^k.

$v^{(0)T}$:= *vector inicial* $||v^{(0)T}|| = 1$

For $k = 1, 2 \ldots$ *hasta convergencia*

$w = Mv^{(k-1)}$

$v^{(k)} = w/||w||$

Método de la potencia

Teorema 4 *Sea M una matriz diagonalizable de tamaño $n \times n$. Supongamos que sus valores propios satisfacen*

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|,$$

y están asociados a los vectores propios $\{v_1, v_2, \ldots, v_n\}$. Entonces, la iteración

$$w^{(k)} = Mw^{(k-1)} \qquad k = 1, 2, \ldots$$

satisface

$$||w^{(k)} - (\pm)v_1|| = O\left(\left\|\frac{\lambda_2}{\lambda_1}\right\|^k\right)$$

cuando $k \to \infty$, donde w^0 es un vector arbitrario tal que la componente respecto del vector propio v_1, de la base, es diferente de cero.

Método de la potencia

Observación: EL método anterior converge si el valor propio es estrictamente dominante.

Nuestra matriz $\overline{\overline{P}}$ es irreducible pero podría tener valores propios $\lambda \neq 1$ pero $|\lambda| = 1$.
Recordemos la matriz del ejemplo de la definición 4.

$$\left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \end{array}\right]$$

es irreducible. Los valores propios son
$\{\lambda_1 = 1,\ \lambda_2 = 1/2 + i\sqrt{3}/2,\ \lambda_3 = 1/2 - i\sqrt{3}/2\}$.
Pero $|\lambda_2| = |\lambda_3| = 1$.

Primitividad

Definición 5 *Sea $M \geq O$ una matriz cuadrada irreducible. y sea k el número de valores propios de módulo $\rho(M)$. Si $k = 1$, entonces se dice que M es **primitiva**. Si $k > 1$ se dice que M es **cíclica** de índice k.*

- La matriz anterior es cíclica de índice 3.
- $M \geq O$ es primitiva $\Leftrightarrow M^m > O$, m entero positivo.
- Si $M > O$ (positiva), entonces es primitiva.

Teorema 5 (Perron, 1907) *Sea $M > O$ una matriz positiva. Entonces:*

- *EL radio espectral, raíz Perron, domina estrictamente a todos los demás valores propios en valor absoluto, es decir, $\rho(M) > |\lambda|$, siendo $|\lambda|$ cualquier otro valor propio.*
- *A la raíz Perron le corresponde un vector propio positivo $x > 0$, el vector Perron.*

Método de la potencia: convergencia

Recapitulamos

La matriz Google

$$\overline{\overline{P}} = \alpha \overline{P} + (1-\alpha) e e^T / n$$

Por construcción es positiva ya que ee^T es la matriz cuyos elementos son todos iguales a uno. Luego es primitiva y por tanto irreducible.

En consecuencia:

- además de tener único vector estacionario,
- el método de potencia es convergente.

Método de la potencia para la matriz Google

Algoritmo. Método iterativo de la potencia para la matriz $\bar{\bar{P}}$.

$\pi^{(0)T} = e^T/n$ Notar que $||\pi^{(0)}||_1 = 1$

For $k = 1, 2 \ldots$ hasta convergencia

$$\pi^T = \pi^{(k-1)T}\bar{\bar{P}}$$

$$\pi^{(k)T} = \pi^T/||\pi^T||_1$$

En realidad el paso principal del algoritmo se escribe,

$$\begin{aligned}\pi^T &= \pi^{(k-1)T}\bar{\bar{P}} = \alpha\pi^{(k-1)T}\bar{P} + (1-\alpha)\pi^{(k-1)}ee^T/n \\ &= \alpha\pi^{(k-1)T}\bar{P} + (1-\alpha)e^T/n \\ &= \alpha\pi^{(k-1)T}P + (\alpha\pi^{(k-1)T}\mathbf{a} + (1-\alpha))e^T/n.\end{aligned}$$

Vector $\mathbf{a}$: $\mathbf{a}_i = 1$ si la fila i de P es nula, $\mathbf{a} = 0$ en otro caso.

Método de la potencia

- La matriz P tiene mas de 4000 millones de filas (páginas).
- Implementación basada con matrix-vector.
- Las matrices llenas $\overline{P}$ y $\overline{\overline{P}}$ no se forman nunca.
- La matriz P es vacía, $nnz(P)$ está entre 3 y 10.
- en cada iteración sólo hay que almacenar un vector.

La convergencia se obtiene entre 50 y 100 iteraciones.

Factor de Convergencia

Tengamos en cuenta las trasparencias 21 y 26.

- Llamemos μ_2 al segundo valor propio de $\overline{\overline{P}}$.
- El factor asintótico de convergencia depende del valor propio μ_2.
- $|\mu_2| \leq \alpha$.

- El radio o factor de convergencia del método de la potencia aplicado a la matriz Google es

$$\alpha^k \longrightarrow 0.$$

Factor de Convergencia

- Sea τ (potencia de 10) el nivel de tolerancia. Es decir, queremos que $\alpha^k \leq \tau$.
- Tomando logarítmos

$$k < \frac{\log_{10} \tau}{\log_{10} \alpha}$$

- Por ejemplo: $\tau = 10^{-6}$ y $\alpha = 0,85$,

$$k < \frac{-6}{\log_{10} 0,85} \approx 85$$

- Es decir, hace falta 85 iteraciones para converger con tolerancia $\tau = 10^{-6}$.

Factor de Convergencia

- Google puede marcar el radio de convergencia.
- α pequeño, entonces mayor factor de convergencia, pero menos verdadera es la estructura de Internet.
- Brin y Page usan $\alpha = 0,85$.

Criterios de convergencia

- Residuo: $\|\pi^{(k)} - \pi^{(k-1)}\| \leq \tau$
- A veces algunas componentes convergen mucho mas pronto que otras. Se fijan ya esas.
- En realidad lo que importa es el orden de las páginas no el valor que se le asigna.
- Cómo evaluar la diferencia entre dos órdenes?

Algoritmo Google

Algoritmo. *Método iterativo de la potencia para la matriz* $\overline{\overline{P}}$.

Entrada: vector inicial $\pi^{(0)}$, matriz $\overline{\overline{P}}$ y tolerancia τ.

Salida: vector estacionario PageRank π^{∞}.

- $\pi^{(0)} = (1/n, 1/n, \ldots, 1/n), \ ||\pi^{(0)}|| = 1$
- WHILE $r \geq \tau$
  1. $\pi^T = \pi^{(k-1)}\overline{\overline{P}}$
  2. $r = ||\pi^T - \pi^{(k-1)}||$
  3. $\pi^{(k)} = \pi/||\pi||$
- END WHILE