

CÁLCULO DEL VECTOR *PAGERANK* DE GOOGLE MEDIANTE EL MÉTODO ADITIVO DE SCHWARZ

Rafael Bru*, Francisco Pedroche* y Daniel B. Szyld**

* Institut de Matemàtica Multidisciplinar
Universitat Politècnica de València.
Camí de Vera s/n. 46022 València. Spain.
email: rbru,pedroche@mat.upv.es, web: <http://personales.upv.es/rbru,pedroche>

** Department of Mathematics
Temple University.
Philadelphia, PA 19122-6094, USA.
email: szyld@math.temple.edu, web: <http://www.math.temple.edu/~szyld/>

Palabras clave: Motores de búsqueda, PageRank, Matriz de Markov, Métodos iterativos de Schwarz, Descomposición de dominios.

Resumen. *El vector PageRank, introducido por los creadores del buscador Google, sirve para medir la importancia o popularidad de las páginas web. El PageRank es un vector real que da un orden de prioridad de unas páginas sobre otras. También se suele llamar PageRank al propio algoritmo de cálculo de este vector. Si imaginamos un dispositivo que pasea aleatoriamente por internet, el valor que asigna el vector PageRank a una página web en particular puede entenderse como la probabilidad de que, pasado un tiempo infinito, dicha página sea visitada por el dispositivo. El cálculo del vector PageRank equivale al cálculo del vector estacionario de una cadena de Markov ergódica. Dado que la matriz que aparece es muy grande, con millones de elementos, es importante disponer de un algoritmo numérico eficiente, siendo los de tipo iterativo los utilizados. Dentro de esta categoría se enmarcan los métodos aditivo y multiplicativo de Schwarz, cuya convergencia ha sido estudiada recientemente. En este trabajo analizamos la aplicación del método aditivo de Schwarz con solapamiento para el cálculo del vector PageRank.*

1 INTRODUCCIÓN

Los creadores del buscador Google [17] sentaron las bases para clasificar las páginas web atendiendo a su *importancia*: básicamente una página es importante si muchas páginas tienen enlaces hacia ella. El modelo matemático se reducía al cálculo del vector estacionario de una cadena de Markov ergódica. Este vector se denomina *PageRank* y sus componentes dan un valor relativo a cada página web indexada. Este cálculo era realizado por el conocido método de la potencia (*power method*). Desde entonces han aparecido

diferentes variaciones del modelo así como del método de resolución; ver, por ejemplo, [1], [11], [12].

El objetivo principal del presente trabajo es analizar la adaptación del método aditivo de Schwarz para el cálculo del vector PageRank.

La comunicación se estructura de la manera siguiente: en la sección 2 mostramos un sistema de ecuaciones cuya resolución permite el cálculo del vector PageRank. En la sección 3 exponemos los fundamentos del método aditivo de Schwarz y en la sección 4 analizamos su uso para el cálculo del vector PageRank, estudiando su puesta en práctica utilizando una matriz de tipo Google, esto es, obtenida a partir de un grafo dirigido correspondiente a un conjunto de páginas web reales.

2 Cálculo del vector PageRank

Dado un conjunto de n páginas web se define su matriz de conectividad $G = (g_{ij}) \in \mathbb{R}^{n \times n}$ de la manera siguiente: $g_{ij} = 1$ si hay un enlace de la página j a la página i , y $g_{ij} = 0$ en otro caso. El número de enlaces salientes de una página j viene dado por $c_j = \sum_i g_{ij}$, que se denomina *out-degree* de la página j .

Se define $P = (p_{ij}) \in \mathbb{R}^{n \times n}$ tal que:

$$p_{ij} = \begin{cases} g_{ij}/c_j & \text{si } c_j \neq 0 \\ 0 & \text{en otro caso.} \end{cases} \quad (1)$$

A partir de la matriz P se puede construir una matriz estocástica por columnas Q asociada a una cadena de Markov ergódica. El vector PageRank x es aquel que verifica $Qx = x$. Dado que Q es densa y enorme, una alternativa práctica se da en [7], [11] donde se demuestra que el vector PageRank x puede calcularse a partir del sistema de ecuaciones:

$$Ay = v \quad (2)$$

con v un vector positivo de distribución de probabilidad y $A = I - \alpha P$, $0 < \alpha < 1$. El vector PageRank se obtiene al normalizar en la forma $x = y/(e^T y)$, siendo e el vector columna con todos sus elementos igual a la unidad. Como la matriz P es poco densa, también lo es A . El vector v es el llamado vector de personalización o de teleportación y se introduce para dar cuenta de las páginas que no tienen enlaces salientes (*dangling pages*, ver [1]). Se suele tomar $v = e/n$. El parámetro α se suele tomar como $\alpha = 0.85$; ver [17].

3 Formulación algebraica del método aditivo de Schwarz con subdominios solapados

Los métodos iterativos de tipo Schwarz con solape [2], [8], [15], [18], [19], [20], están asociados con un dominio físico que se ha dividido en subdominios solapados descritos por matrices cuadradas A_i , $i = 1, 2, \dots$, obtenidas a partir de la matriz de coeficientes A que modela el problema. Cuando no hay solape entre los subdominios, el método aditivo de Schwarz coincide con el método de Jacobi por bloques.

Sea un sistema de ecuaciones lineales compatible de la forma:

$$Ax = b, \quad (3)$$

donde $A \in \mathbb{R}^{n \times n}$ y x y b son vectores de $V = \mathbb{R}^n$. La convergencia de los métodos aditivo y multiplicativo de Schwarz aplicado a (3) ya ha sido estudiada en los casos en que A es a) simétrica y definida positiva, b) M matriz invertible, c) H -matriz, d) matriz de rango $n - 1$ y todas sus submatrices principales M -matrices no singulares, y e) matriz positiva semidefinida; ver [2], [4], [5], [6], [8], [13], [15], [16].

Para describir los subdominios, llamemos V_i cada uno de los subespacios de $V = \mathbb{R}^n$ de dimensión n_i , $i = 1, \dots, p$, ($p > 1$) tal que su suma cubre todo V . Estos subespacios se intersectan de forma que $\sum_{i=1}^p n_i > n$. Sean los operadores de restricción:

$R_i = [I_i | O] \pi_i$, $i = 1, \dots, p$, donde I_i es la matriz identidad de orden n_i y π_i es una matriz de permutación de orden n . Se definen también las matrices diagonales $E_i = R_i^T R_i$, $i = 1, \dots, p$. Sea q el valor máximo de la diagonal de $\sum_{i=1}^p E_i$. Este valor q sirve como *medida del solape* y se cumple que: $\sum_{i=1}^p E_i \leq qI$.

La restricción de la matriz A a cada subespacio V_i viene dada por:

$$A_i = R_i A R_i^T, \quad (4)$$

que es una permutación simétrica de una submatriz principal de A de orden n_i . Cada una de estas matrices A_i puede identificarse con un bloque solapado.

En el caso de que todas las matrices A_i sean invertibles, como ocurre aquí, las iteraciones de tipo aditivo Schwarz con amortiguamiento son de la forma:

$$x^{k+1} = x^k + \theta \sum_{i=1}^p R_i^T A_i^{-1} R_i (b - Ax^k), \quad (5)$$

donde $0 < \theta \leq 1$ es el factor de amortiguamiento. Denominamos T_θ a la matriz de iteración de este esquema, que viene dada por:

$$T_\theta = I - \theta \sum_{i=1}^p R_i^T A_i^{-1} R_i A \quad (6)$$

El esquema iterativo para la solución de (3) adopta entonces la forma:

$$x^{k+1} = T_\theta x^k + c, \quad k = 0, 1, \dots, \quad (7)$$

con $c = \theta \sum_{i=1}^p R_i^T A_i^{-1} R_i b$.

4 Uso del método aditivo de Schwarz para el cálculo del PageRank

De acuerdo con (7), las iteraciones del método aditivo de Schwarz para resolver (2) se escriben:

$$y^{k+1} = T_\theta y^k + c \quad k = 0, 1, \dots, \tag{8}$$

donde T_θ viene dado por (6) y $c = \theta \sum_{i=1}^p R_i^T A_i^{-1} R_i v$.

Como A es una M -matriz no singular, es sabido que si $\theta \leq 1/q$, el método aditivo de Schwarz converge [8]. Por otra parte, es conocido que cuando A es una M -matriz no singular el método de Jacobi por bloques es convergente [3].

Del Corso *et al* [7] resolvieron (2) mediante diferentes métodos iterativos mostrando que una permutación adecuada de la matriz A junto con un método iterativo por bloques producía un ahorro sustancial respecto al método de la potencia.

Por lo que sabemos, no hay resultados publicados respecto al comportamiento numérico del método aditivo de Schwarz para resolver el vector PageRank. Este trabajo representa un primer estudio del problema.

4.1 Resultados numéricos

En este apartado consideramos una matriz P de tipo Google obtenida considerando las 50.000 primeras filas y columnas de la matriz de conectividad usada en [10]. Esta matriz, que denominaremos *stanford50000*, tiene 83.634 elementos no nulos y 14.313 *dangling nodes*.

En la figura 1 se muestra el patrón de la matriz *stanford50000* (izquierda) y el que se obtiene después de realizar una permutación simétrica RCM (Reverse Cuthill McKee, ver, por ejemplo, [9]).

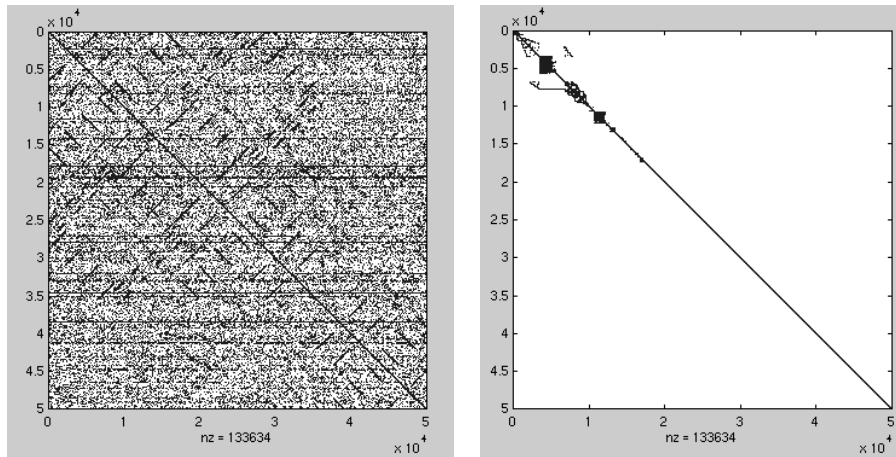


Figura 1: Estructura de la matriz P stanford50000 antes (izquierda) y después de permutar.

En la tabla 1 se muestra la evolución del radio espectral de la matriz de iteración T_θ en función del parámetro de amortiguamiento θ correspondiente a los casos sin solape (*Block-Jacobi*: BJ) y con solape (aditivo Schwarz: AS) tanto para la matriz inicial como para la permutada. El número de subdominios considerado fue de $p = 9$. La estructura de los subdominios es la siguiente: hay un subdominio de tamaño 10.000, y el resto son de tamaño 5.000. Cuando se ha considerado solape, éste se ha tomado de la manera siguiente: el primer subdominio se solapa con el segundo formando un bloque solapado de tamaño 500. El resto de solapes son de tamaño 50.

θ	Sin Permutar		Permutando	
	BJ	AS	BJ	AS
0.1	0.9850	0.9850	0.9746	0.9000
0.2	0.9700	0.9700	0.9491	0.8000
0.3	0.9550	0.9550	0.9237	0.7000
0.4	0.9400	0.9400	0.8982	0.6000
0.5	0.9250	0.9250	0.8728	0.5000
0.6	0.9100	> 1	0.8474	0.4000
0.7	0.8950	> 1	0.8219	0.4000
0.8	0.8800	> 1	0.7965	0.6000
0.9	0.8650	> 1	0.7711	0.8000
1.0	0.8500	> 1	0.7456	1.0000

Tabla 1: Evolución del radio espectral de T_θ .

De la tabla 1 se observa que el hecho de permutar la matriz se traduce en una disminución considerable del radio espectral, alcanzándose los valores menores para el caso de aditivo Schwarz y un valor del parámetro de amortiguamiento de $\theta = 0.6$ ó 0.7 . Cuando no producimos una permutación se observa que el efecto del solape en este caso empeora los valores del radio espectral. Es de destacar que los valores de convergencia del método *AS* llegan hasta la cota superior que ofrece el teorema de convergencia dado en [8]: podemos asegurar que hay convergencia hasta $\theta \leq 0.5$. Los cálculos se han llevado a cabo usando un procesador del computador IBM-1350 de la Universitat Politècnica de València ejecutando MATLAB 6.0 .

En la tabla 2 se resumen los mejores resultados obtenidos, atendiendo al número de iteraciones, cuando se ha calculado el vector PageRank resolviendo el sistema (2) mediante BJ y AS. Se muestran los casos sin permutar, y permutando, la matriz *stanfordd50000*. También se indica el efecto que produce tomar el vector inicial de la iteraciones como $y_0 = e$ ó $y_0 = v = e/n$. El criterio utilizado para detener las iteraciones ha sido en todos los casos el mismo: la norma del residuo correspondiente $Ay - v$ ha de ser menor o igual que 10^{-8} .

		Sin Permutar		Permutando			
Método	y_0	θ	iteraciones	t(s)	θ	iteraciones	t(s)
BJ	e	1.0	146	20.33	1.0	72	16.97
	e/n	1.0	91	12.51	1.0	36	8.85
AS	e	0.5	304	45.31	0.7	26	6.79
	e/n	0.5	189	28.23	0.7	14	3.60

Tabla 2: Resultados para la matriz *stanford50000*.

De la tabla 2 se observa que los mejores resultados tanto en número de iteraciones como en tiempo empleado para realizar dichas iteraciones corresponden al método aditivo de Schwarz utilizando como vector inicial $y_0 = e/n$ y tras realizar una permutación simétrica de la matriz inicial.

		Sin Permutar		Permutando			
Método	y_0	θ	iteraciones	t(s)	θ	iteraciones	t(s)
BJ	r	1.0	142	9.22	1.0	70	13.19
	r_n	1.0	91	5.70	1.0	36	6.77
AS	r	0.5	295	19.69	0.7	26	4.97
	r_n	0.5	189	12.69	0.7	14	2.65

Tabla 3: Resultados para la matriz *stanford50000* con y_0 aleatorio.

En la tabla (3) se han repetido los cálculos utilizando como vector inicial el vector aleatorio r obtenido mediante las instrucciones `rand('state',0)` y `r=rand(n,1)` ejecutando MATLAB 6.0. El vector r_n se obtiene normalizando el anterior: $r_n = r/(e^T r)$. En esta tabla se observa que los mejores resultados se obtienen cuando el vector inicial y_0 se toma normalizado. Diversas ejecuciones con vectores aleatorios sugieren que es preferible tomar y_0 como un vector aleatorio normalizado que la elección $y_0 = e/n$.

5 Conclusiones

Se ha mostrado cómo utilizar el método aditivo de Schwarz para el cálculo del vector PageRank. El método se aplica a un sistema de ecuaciones lineales no homogéneo con matriz poco densa. Los resultados numéricos, basados en una matriz de tipo Google, muestran que el método aditivo de Schwarz con solapamiento es preferible al método de Jacobi por bloques, atendiendo tanto al número de iteraciones como al tiempo empleado para llevarlas a cabo. En general, los mejores resultados se obtiene cuando se realiza una permutación previa de la matriz Google de manera que los elementos queden agrupados en la diagonal.

6 Agradecimientos

Los dos primeros firmantes gozan de la ayuda estatal DGI número MTM2004-46022, y la ayuda de la Oficina de Ciència i Tecnologia de la Presidència de La Generalitat Valenciana denominada GRUPOS03/062. El tercer firmante tiene el convenio de investigación de la National Science Foundation, número DMS-0207525.

REFERENCIAS

- [1] Arasu, A., Novak, J., Tomkins, A. y Tomlin, J. (2002) PageRank computation and the structure of the Web: experiments and algorithms. In *The Eleventh International WWW Conference*, ACM Press. Nueva York, Estados Unidos.
- [2] Benzi, M., Frommer, A., Nabben, R. y Szyld, D. B. (2001). Algebraic theory of multiplicative Schwarz methods, *Numerische Mathematik*, 89, 605-639.
- [3] Berman, A. y Plemmons, R. J. (1979). *Nonnegative matrices in the mathematical sciences*, Academic Press, Nueva York. Estados Unidos.
- [4] Bru, R., Pedroche, F. y Szyld, D. B. (2004) Additive Schwarz Iterations for Markov Chains, Research Report 04-10-08, Department of Mathematics, Temple University, Filadelfia, Estados Unidos.
- [5] Bru, R., Pedroche, F. y Szyld, D. B. (2004). Overlapping Additive and Multiplicative Schwarz Iterations for H -matrices. *Linear Algebra and its Applications*. 393, 91-105.
- [6] Bru, R., Pedroche, F. y Szyld, D. B. (2004). Sobre la convergencia del método iterativo de Schwarz para sistemas singulares. En *Métodos Computacionais em Engenharia. Proceedings of the VI Congresso de Métodos Numéricos em Engenharia*. C.A. Mota Soares et al (ed.). APMTAC-SEMNI. Lisboa, Portugal.
- [7] Del Corso, G. M., Gulli, A. y Romani, F. (2004). Fast PageRank Computation Via a Sparse Linear System. *Lecture notes in computer science*, 3243, 118-130.
- [8] Frommer, A. y Szyld, D. B. (1999). Weighted Max Norms, Splittings, and Overlapping Additive Schwarz Iterations, *Numerische Mathematik*, 83, 259-278.
- [9] George, A. y Liu, J. W. (1981). *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall. Englewood Cliffs, Nueva Jersey, Estados Unidos.
- [10] Kamvar, S. D., Haveliwala, T. H. y Golub, G. H. (2004). Adaptive methods for the computation of PageRank. *Linear Algebra and its Applications*, 386, 51-65.
- [11] Langville, A. N. y Meyer, C. D. (2005). A reordering for the PageRank problem. Technical report 04-16, North Carolina State University, Mathematics Department. Carolina del Norte, Estados Unidos.

- [12] Langville, A. N. y Meyer, C. D. (2004) . Deeper inside PageRank. *Internet Mathematics*. En preparación.
- [13] Marek, I. y Szyld, D. B. (2004). Algebraic Schwarz methods for the numerical solution of Markov chains, *Linear Algebra and its Applications*, 386, 67-81.
- [14] Moler, C. B. (2004). *Numerical Computing with MATLAB*. SIAM. Filadelfia, Estados Unidos.
- [15] Nabben, R. (2003). Comparisons between multiplicative and additive Schwarz iterations in domain decomposition methods. *Numerische Mathematik*, 95, 145-162.
- [16] Nabben, R. y Szyld, D. B. (2005). *Schwarz Iterations for Symmetric Positive Semidefinite Problems*. En preparación.
- [17] Page, L., Brin, S., Motwani, R. y Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Library Technologies Project. Accesible en: <http://dbpubs.stanford.edu:8090/pub/1999-66>.
- [18] Quarteroni, A. y Valli, A. (1999). *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science Publications, Clarendon Press, Oxford, Reino Unido.
- [19] Smith, B. F., Bjørstad, P. E. y Gropp, W. D. (1996). *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, Reino Unido.
- [20] A. Toselli y O. B. Widlund. (2005). *Domain Decomposition Methods - Algorithms and Theory*. Springer, Heidelberg, Alemania.