Laboratorio de Métodos Numéricos - Segundo cuatrimestre 2010
Trabajo Práctico Número 2: Ranking de Page (¿y Plant?)

---

1. Introducción

El motor del buscador de Google utiliza el denominado ranking de Page[1] como uno de los criterios para ponderar la importancia de los resultados de cada búsqueda. Calcular este ranking requiere simplemente resolver un sistema de ecuaciones lineales... donde la cantidad de ecuaciones e incógnitas del sistema es igual al número de páginas consideradas. ¿Simplemente?

Para un determinado conjunto de n páginas web definamos la *matriz de conectividad* $\mathbf{W}$ poniendo $w_{ij} = 1$ si la página j tiene un link a la página i y $w_{ij} = 0$ si no. Además $w_{ii} = 0$ pues ignoramos los "autolinks". De esta forma, la matriz $\mathbf{W}$ puede resultar extremadamente rala y muy grande de acuerdo al tamaño del conjunto. Por ejemplo, cada página j tiene $c_j = \sum_i w_{ij}$ links (salientes), que típicamente es un número mucho menor que n.

Se busca que el ranking sea mayor en las páginas "importantes". Heurísticamente, una página es importante cuando recibe muchos "votos" de otras páginas, es decir, links. Pero no todos los links pesan igual: los links de páginas más importantes valen más. Pocos links de páginas importantes pueden valer más que muchos links de páginas poco importantes. Y los links de páginas con muchos links valen poco, por ser "poco específicos".

Un enfoque alternativo es considerar al "navegante aleatorio". El navegante aleatorio empieza en una página cualquiera del conjunto, y luego en cada página j que visita elige con probabilidad p si va a seguir uno de sus links, o (con probabilidad $1-p$) si va a pasar a una página cualquiera del conjunto. Usualmente $p = 0{,}85$. Una vez tomada esa decisión, si decidió seguir un link elige uno al azar (probabilidad $1/c_j$), mientras que si decidió pasar a una página cualquiera entonces elige una al azar (probabilidad $1/n$). Cuando la página j no tiene links salientes ($c_j = 0$) elige al azar una página cualquiera del conjunto. Heurísticamente, luego de muchos pasos el navegante aleatorio estará en páginas importantes con mayor probabilidad.

Formalmente, la probabilidad de pasar de la página j a la página i es

$$a_{ij} = \begin{cases} (1-p)/n + (p\, w_{ij})/c_j & \text{si} \quad c_j \neq 0 \\ 1/n & \text{si} \quad c_j = 0 \end{cases},$$

y sea $\mathbf{A}$ a la matriz de elementos a_{ij}. Entonces el **ranking de Page** es la solución del sistema

$$\mathbf{A}\,\mathbf{x} = \mathbf{x} \tag{1}$$

que cumple $x_i \geq 0$ y $\sum_i x_i = 1$. Si pensamos que x_j es la probabilidad de encontrar al navegante aleatorio en la página j, tenemos que $(\mathbf{Ax})_i$ es la probabilidad de encontrarlo en la página i luego de un paso. Y el ranking de Page es aquella distribución de probabilidad que resulta "estable".

[1] Por Larry Page, uno de los fundadores de Google, otrora joven científico actualmente devenido multimillonario.

La matriz $\mathbf{A}$ puede reescribirse como

$$\mathbf{A} = p\ \mathbf{W}\ \mathbf{D} + \mathbf{e}\ \mathbf{z}^T,$$

donde $\mathbf{D}$ es una matriz diagonal de la forma

$$d_{jj} = \begin{cases} 1/c_j & \text{si} \quad c_j \neq 0 \\ 0 & \text{si} \quad c_j = 0 \end{cases},$$

$\mathbf{e}$ es un vector columna de unos de dimensión n y $\mathbf{z}$ es un vector columna cuyos componentes son

$$z_j = \begin{cases} (1-p)/n & \text{si} \quad c_j \neq 0 \\ 1/n & \text{si} \quad c_j = 0 \end{cases}.$$

Así, la ecuación (1) puede reescribirse como

$$(\mathbf{I} - p\ \mathbf{W}\ \mathbf{D})\ \mathbf{x} = \gamma\ \mathbf{e}, \tag{2}$$

donde $\gamma = \mathbf{z}^T\ \mathbf{x}$ funciona como un factor de escala.

De esta manera, un procedimiento para calcular el ranking de Page consiste en:

1. Suponer $\gamma = 1$.
2. Resolver la ecuación (2).
3. Normalizar el vector $\mathbf{x}$ de manera que $\sum_i x_i = 1$.

2. Enunciado

El objetivo de este trabajo es programar el cálculo del ranking de Page según el procedimiento descripto anteriormente. Para la resolución del sistema de ecuaciones resultante deberán implementar por lo menos

- un método directo y
- un método iterativo,

que resulten apropiados.

Previamente deberán estudiar las características de la matriz involucrada. ¿Cómo se garantiza la aplicabilidad de cada método? En el caso directo, la inversibilidad de $(\mathbf{I} - p\ \mathbf{W}\ \mathbf{D})$. ¿Está bien condicionada? Y en el caso iterativo, la convergencia del método.

Deberán proponer 3 casos de prueba (conjuntos de páginas y sus links) de acuerdo a criterios establecidos por el grupo, para probar las implementaciones.

Los programas deberán leer los archivos con los datos de los conjuntos de páginas (según el formato que se describe más abajo) y calcular el ranking. Las diferentes implementaciones deberán utilizarse para calcular el ranking de páginas de varios conjuntos disponibles en la página de la materia propuestos por la cátedra y por los demás grupos. Podrán, además, analizarse otros conjuntos elegidos por el grupo.

En base a los ensayos realizados:

- Interpretar los resultados. ¿Cómo es el ranking obtenido en cada caso de acuerdo a las estructuras de los conjuntos? ¿Qué conclusiones pueden sacar de la interpretación de los resultados?
- Respecto del ranking de Page. ¿Funciona cómo era esperado? ¿Hubo sorpresas? ¿Qué pueden concluir sobre su significado?
- Respecto de los 2 ó más métodos implementados para la resolución de los sistemas de ecuaciones lineales ¿Cómo es el desempeño de cada uno? Considerar error numérico, costo computacional en tiempo, cantidad de operaciones, memoria y cantidad de iteraciones (cuando corresponda).

Datos de entrada

Cada conjunto de páginas será descripto por 2 archivos de texto plano, a saber:

archivo de páginas: conteniendo en la primera línea la cantidad de páginas, y luego para cada página una línea con: el número y la *url* de la misma, separados por espacios; y

archivo de links: conteniendo en la primera línea la cantidad total de links, y luego para cada link una línea con: el número de la página de origen y el número de la página de destino (en ese orden), separados por espacios.

Entregas parciales en papel, no más de una carilla de texto:

1 de octubre: ideas y soluciones propuestas, plan de implementación, casos de prueba

8 de octubre: implementación y plan de experimentación.

Fecha de entrega final: 22 de octubre de 2010