Algoritmo PageRank

La web (o una parte de ella) puede representarse como un grafo dirigido en el que cada nodo corresponde a una página web y las flechas los links entre ellas.

Por ejemplo, en la Figura 1 se representa una web de 5 páginas, en donde por ejemplo la página 3 tiene links hacia las páginas 4 y 5, y es linkeada desde 2 y 5.

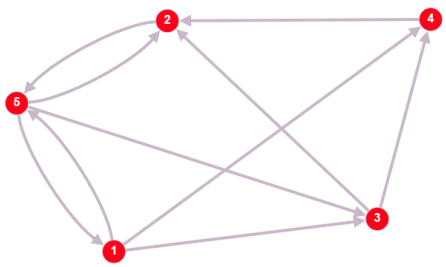

Figura 1: ejemplo

Con el fin de hacer un ranking de la web, queremos definir la *importancia* una página j. En principio podría definirse como el número de páginas paginas que refieren ("linkean") a ella. Pero si nos quedáramos con esto, no estaríamos teniendo en cuenta la importancia de esas otras páginas web, o sea, una página importante y otra menos importante, ambas con links a n, tendrían el mismo peso en esta definición. Otra posibilidad consiste en asumir que cada página web distribuye su importancia equitativamente entre las páginas a las cuales linkea. De esta manera podemos asignar un peso r_j a cada nodo j (la importancia de la j-ésima página web) dado por

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

donde r_i es el peso del nodo i, d_i es la cantidad de flechas que salen del nodo i (o cantidad de páginas linkeadas desde la página i), y $i \to j$ significa que hay un link desde la página i a la página j, por lo que la suma es sobre todos los nodos i que tienen una flecha hacia el nodo j.

Así, para el ejemplo de la Figura 1 tenemos

$$\begin{aligned} r_1 &= \frac{r_5}{3} \\ r_2 &= \frac{r_3}{2} + r_4 + \frac{r_5}{3} \\ r_3 &= \frac{r_1}{3} + \frac{r_5}{3} \\ r_4 &= \frac{r_1}{3} + \frac{r_3}{2} \\ r_5 &= \frac{r_1}{3} + r_2 \end{aligned}$$

En general, para un grafo de n nodos, si introducimos el vector

$$\bar{r} = \begin{pmatrix} r_1 & r_2 & \cdots & r_n \end{pmatrix}^t$$

que puede considerarse como *ranking* de páginas (siendo que r_i es la *importancia* de la página i), tenemos un sistema de ecuaciones de la forma

$$\bar{r} = P\bar{r} \tag{1}$$

con P matriz $n \times n$. En el caso del ejemplo tenemos

$$P = \begin{pmatrix} 0 & 0 & 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & 1 & \frac{1}{3} \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Podemos pensar que la entrada (i, j) de P es la probabilidad de que estando en la página j pasemos a la página i (o, estando en el nodo j pasemos al nodo i).

De la ecuación (1) vemos que $\bar{r}$ es un autovector de P con autovalor 1. Entonces nos podemos preguntar si, en este contexto, P siempre tiene a 1 como autovalor, si $\lambda = 1$ siempre tiene un autovector con componentes no negativas (ya que quisiéramos que las r_i sean no negativas), y en este caso, cómo calcular $\bar{r}$ en un caso real.

Primero notemos que P es una matriz de Markov.

Definición 1 *Una matriz de Markov es una matriz con todas sus entradas no negativas y cuyas columnas suman* 1*. Las matrices de Markov también suelen llamarse matrices estocásticas.*

Definición 2 *Un vector v se dice estocástico si sus entradas son no negativas y suman* 1*.*

Ya podemos contestar entonces algunas de las preguntas.

Teorema 1 *Sea P una matriz de Markov. Entonces* 1 *es autovalor de P, y además los autovalores restantes tienen módulo ≤ 1.*

Ayuda para la demostración. Ver que 1 es autovalor de P^t.

Proposición 1 *Sea A una matriz de Markov y v un vector estocástico. Entonces Av también es estocástico.*

Queremos hallar o aproximar $\bar{r}$. Siendo que $\bar{r}$ es un autovector asociado a 1, y que todos los autovalores de P tienen módulo menor o igual 1, podemos estar tentados a aplicar el método de potencias. Sin embargo este método, que aproxima el autovalor dominante (de mayor valor absoluto), requiere que este sea único. Para asegurar trabajar con una matriz con un único autovalor dominante, vamos a tener en cuenta el siguiente Teorema. Decimos que una matriz (o un vector) es *positiva* (resp. *no negativa*) si todas sus entradas son positivas (resp. *no negativas*).

Teorema 2 (Perron–Frobenius, para matrices positivas) *Sea A una matriz positiva. Entonces $\lambda_1 = \rho(A)$ es un autovalor simple de A. Además asociado a λ_1 existe un autovector positivo.*

Una demostración puede encontrarse en [1, Teorema 2.7]. Ver también [2, pág. 271].

Claramente la matriz P es no negativa, pero como vemos en el ejemplo puede tener entradas nulas, por lo que no podemos aplicar el Teorema de Perron–Frobenius. Por eso introducimos una modificación en la matriz P, que va a corresponder con conectar todos los nodos en el grafo. Sean $\beta, \varepsilon > 0$, con $\beta < 1$. Vamos a pensar que, estando en el nodo j, con probabilidad β pasamos a otro nodo de acuerdo con la matriz P, y con probabilidad $1 - \beta$ saltamos de manera uniformemente aleatoria. Esto es, la probabilidad de, estando en el nodo j, pasar al nodo i es

$$a_{ij} = \beta P_{ij} + (1 - \beta)\varepsilon.$$

Para que la matriz A siga siendo de Markov necesitamos que

$$1 = \sum_{i=1}^{n} a_{ij} = \sum_{i=1}^{n} [\beta P_{ij} + (1 - \beta)\varepsilon] = \beta + n(1 - \beta)\varepsilon.$$

Por lo tanto debe ser $\varepsilon = \frac{1}{n}$. Entonces la matriz $A = (a_{ij})$ puede escribirse como

$$A = \beta P + (1 - \beta)E$$

siendo E la matriz con todas entradas iguales a $\frac{1}{n}$.

Proposición 2 *Tenemos $\rho(A) = 1$, 1 es autovalor simple de A y los restantes autovalores tienen módulo menor que 1. Además, existe un (único) autovector $\bar{r}$ asociado a $\lambda = 1$ positivo que es además un vector estocástico.*

El método de potencias para aproximar el autovalor de mayor módulo para la matriz A puede entonces escribirse como

```
r = 1/n * ones(n,1);
r0 = zeros(n,1)
while (norm(r-r0)>1e-6)
  r0 = r;
  r = A*r;
endwhile
```

Notar que no interesa en este caso aproximar el autovalor.

Ejercicio 1 *Verificar que, si bien A es una matriz densa, las operaciones de tipo $A\bar{r}^{(k)}$ de cada paso del algoritmo anterior pueden realizarse con la misma complejidad que si se trabajara con la matriz P (que puede ser rala).*

Ejercicio 2 *Para un grafo dado, implementar el método del ejercicio 1 mostrando gráficamente el ranking de los nodos que se obtiene en cada paso y la convergencia a $\bar{r}$.*

Referencias

[1] Varga, R.S. Matrix Iterative Analysis. Springer-Verlag, Berlin, 2009.

[2] Strang, G. Linear Algebra and its Applications. Third Edition. Brooks/Cole Thompson Learning, Boston, 1988.