PageRank

Darío Sneidermanis, Martín Sturla

19 de Abril de 2012

***Resumen*—Se analizan y compara el desempeño de la implementación clásica del algoritmo de PageRank, con el método de las potencias, y una implementación alternativa que se empezó a estudiar recientemente, usando sistemas lineales.**

***Palabras clave*—PageRank, motores de búsqueda, cadenas de Markov, método de las potencias, autovectores.**

I. Introducción

Con el inicio de las telecomunicaciones masivas por Internet a principio de los 90' y su rápido crecimiento, nació la necesidad de crear un índice de los sitios web. Estos índices eran usados por los motores de búsqueda, y acompañaron el crecimiento del tráfico de Internet. En 1994, World Wide Web Worm poseía un índice de unos 110.000 sitios, y era consultado unas 1500 veces diariamente. Ya en 1997, los motores de búsqueda contaban con índices con una cantidad de sitios que variaba desde 2 hasta 100 millones, y algunos eran consultados unas 20 millones de veces diariamente. Estos índices eran mantenidos por humanos y por lo tanto, además de ser posiblemente subjetivos, no parecían ser lo suficientemente escalables como para poder acompañar el rápido crecimiento del tráfico por Internet.

En 1998, habiendo previsto índices en el orden de los miles de millones de sitios para el fin del milenio, dos estudiantes de Stanford, Sergey Brin y Lawrance Page, publicaron un paper titulado *"The Anatomy of a Large-Scale Hypertextual Web Search Engine"* que explora el problema mencionado anteriormente y ofrece una solución; un algoritmo computable para indexar sitios web según su importancia, conocido como *PageRank*. Estos estudiantes crearon el conocido motor de búsqueda llamado *Google* utilizando este algoritmo.

Se comienza el paper con una exposición y modelado del problema en la sección II, seguido del desarrollo matemático básico en la sección III. En la sección IV se exponen los dos métodos de solución y se comparan experimentalmente en la sección V.

II. Modelado del problema

El conjunto de los sitios web puede ser representado como un digrafo, donde cada sitio es un nodo. Existe una arista dirigida entre dos nodos si desde el primer sitio hay un hipervínculo al segundo. El *PageRank* de cada sitio (nodo) es la probabilidad de que una persona apretando hipervínculos al azar llegue a ese sitio (nodo). En cada momento o iteración, dicha persona esta en un sitio y puede desplazarse a otro sitio por un hipervínculo al azar (con probabilidad d, llamado factor de amortiguamiento), o ingresar a un sitio al azar de toda la web (usar la barra de direcciones del navegador).

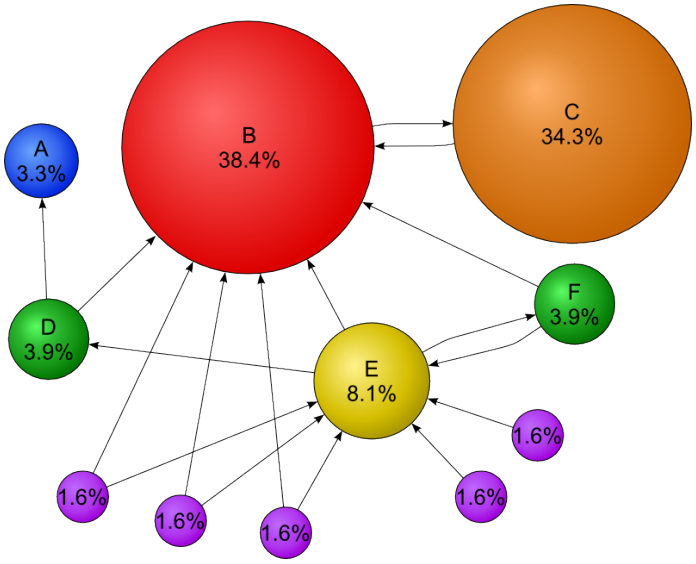

Figura 1.PageRanks, representados como porcentajes, de una red simple. Nótese que el sitio C tiene un PageRank alto, a pesar de tener solo una arista entrante, por ser esta de un sitio importante.

En vista de dichas propiedades, este grafo puede ser visto como una cadena de Markov: se toman las probabilidades iniciales de estar en cualquier nodo uniformemente, y las probabilidades de abrir un hipervínculo de un mismo nodo también uniformemente. Además no existen estados absorbentes; se asume que en cualquier momento un usuario podría ingresar (con probabilidad $1-d$) el nombre de cualquier otro sitio en el navegador, con distribución uniforme, nuevamente.

El vector de probabilidades al que converge dicha cadena de Markov, representa el peso o importancia de cada sitio: el PageRank, que es una distribucción de probabilidades. Por lo tanto, el vector debe estar normalizado.

III. Marco algebraico

Se debe calcular $PR(p_i)$, el valor del *PageRank* de un nodo p_i. Dicho valor se calcula como la suma de la probabilidad de haber llegado por la barra de direcciones (con probabilidad $(1-d)/N$), y haber venido por un hipervínculo de un sitio o nodo p:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p \in M_{p_i}} \frac{PR(p)}{L(p)} \quad (1)$$

Donde $L(p)$ es el grado de salida del nodo p, N es la cantidad total de nodos, y M_{p_i} es el conjunto de los nodos que alcanzan directamente a p_i.

Nota: La probabilidad de haber llegado a través de un hipervínculo, es decir el factor de amortiguamiento d, vale aproximadamente $0,85$ (establecido empíricamente [1]).

Aplicando la ecuación (1) a todos los nodos, se obtiene una nueva ecuación en forma matricial:

$$R = \begin{pmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{pmatrix} + dMR \quad (2)$$

Donde R es el vector de *PageRanks* y $M_{i,j}$ se define como $L(p_j)^{-1}$ si existe un hipervínculo en p_j hacia p_i, y 0 sino.

$$R_i = \frac{1-d}{N} + d \sum_{j=1}^{N} M_{i,j} R_j$$

IV. Soluciones para el problema de calcular el PageRank

Se define la matriz $\widehat{M}$ como:

$$\widehat{M} = \frac{1-d}{N} E + dM \quad (3)$$

Donde E es una matriz con todos unos.

Usando la ecuación (2):

$$\begin{aligned} R &= \begin{pmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{pmatrix} + dMR \\ &= \frac{1-d}{N} ER + dMR \\ &= \widehat{M}R \end{aligned} \quad (4)$$

($E_i \cdot R = 1$, pues R es una distribución de probabilidades)

La cadena de Markov con matriz de transición $\widehat{M}$ es ergódica y regular (ya que todas las transiciones son no nulas) y, por lo tanto, R existe y es único (teorema ergódico).

A. Método de las potencias

Dado que las filas y columnas de $\widehat{M}$ suman 1 (es estocástica), por círculos de Gershgorin el módulo de todos los autovalores deberá ser menor o igual a 1. De la ecuación (4) se puede deducir que la solución R no es más que el autovector dominante con autovalor 1 de la matriz $\widehat{M}$. El autovalor será único si existe una única distribución que satisface la ecuación (4), lo cual es equivalente a exigir que $Rg\left(\widehat{M} - I\right) = N - 1$ y no menor. Si se cumple, se puede calcular R con el método de las potencias. Cabe destacar que como R es una distribución de probabilidades, el autovector debe estar normalizado.

El método de las potencias consiste en tomar un vector $v(0)$ e iterar según la regla:

$$v(t+1) = \widehat{M} v(t) \quad (5)$$

Normalizando $v(t+1)$ en cada paso, hasta que se cumpla:

$$|v(t+1) - v(t)| < \epsilon \quad (6)$$

Donde ϵ es un error adecuado. El método de las potencias converge al autovalor dominante siempre y cuando la proyección del vector $v(0)$ a este último no sea nula.

El ritmo asintótico de convergencia de este método está gobernado por el autovalor subdominante. Este autovalor está acotado superiormente por d, el factor de amortiguamiento [4]. En la figura 1, se puede ver que para $0,5 \leq d \leq 0,9$, el factor de amortiguamiento afecta linealmente el tiempo de ejecución del algoritmo. Esto trae a la mesa el interesante trade-off de tiempo de ejecución versus factor de amortiguamiento d, que es un dato empírico. Habría que analizar cuánto cambia la solución al variar d.

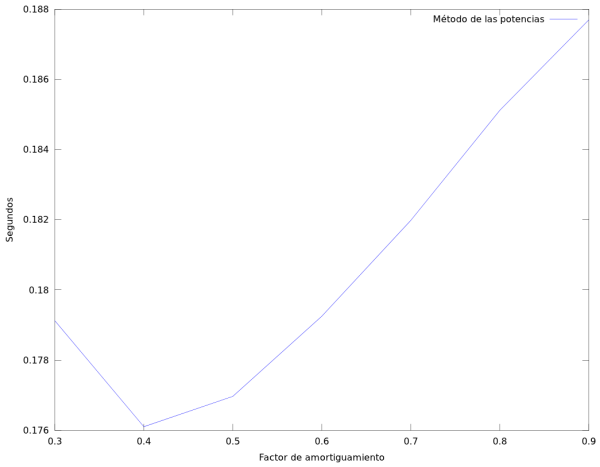

Figura 2.Gráfico del tiempo de ejecución del método de las potencias en función de el factor de amortiguamiento, para un grafo de 100 sitios.

B. Sistema lineal

De (2) se deduce:

$$(I - dM)R = \begin{pmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{pmatrix} \quad (7)$$

Sistema lineal de la forma $Ax = b$ con $A = I - dM$, $x = R$, $b = (\frac{1-d}{N}, ..., \frac{1-d}{N})^T$, que puede ser resuelto con, por ejemplo, eliminación de Gauss-Jordan o el método de Jacobi.

Nota: se podría usar también $R(I - d\widehat{M}) = 0$, pero la matriz M es mucho menos densa que la matriz $\widehat{M}$, que es, en general, algo deseable.

V. Comparación y conclusiones

Los grafos sobre los que se probaron los algoritmos fueron generados proceduralmente, aleatoriamente, teniendo cuidado en conservar las características de un grafo web real (a saber, la cantidad total de links es Θ(cantidad de sitios)):

$$\mathrm{M} = \mathrm{N} * \mathbf{rand}(\mathrm{N}) < \mathrm{AvgLinksPerPage}$$

Con AvgLinksPerPage = 10 [5].

Se implementaron 3 algoritmos: el clásico método de las potencias, y la alternativa que ha empezado a estudiar más recientemente: la resolución del sistema lineal, con el método de Jacobi y con el método que usa nativamente octave.

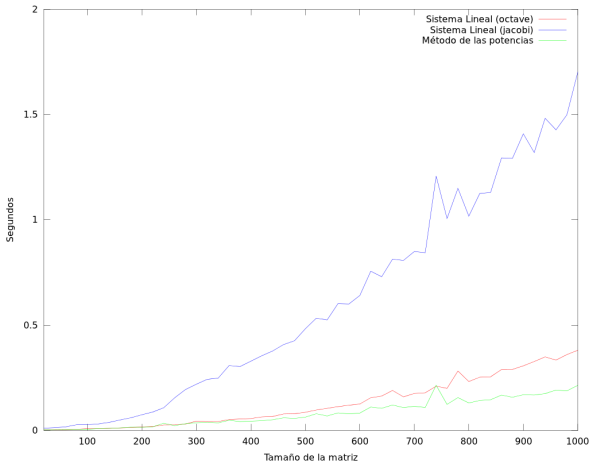

Figura 3.Gráfico del tiempo de ejecución de los tres algoritmos para grafos web de 20 a 1000 sitios.

Es interesante ver (figura 2) que la forma clásica de resolverlo y la resolución por sistema lineal (octave), tienen tiempos bastante similares; variando, aparentemente, por tan solo una constante multiplicativa pequeña. Además, se pueden añadir varias mejoras a ambos algoritmos aún [4] (esta fue una implementacion de los algoritmos básicos), por lo que no hay un ganador claro. Siendo así muy prometedor el nuevo encare por sistemas lineales de ecuaciones, que todavía no está tan explorado.

VI. References

[1] Brin, S.; Page, L. (1998). *The anatomy of a large-scale hypertextual Web search engine.*

[2] Page, L. (1997). *PageRank: Bringing Order to the Web.*

[3] Bianchini, M.; Gori, M.; Scarselli, F. (2005). *Inside PageRank.*

[4] Langville, A.; Meyer, C. (2004). *Deeper Inside PageRank.*

[5] Levering, R.; and M. Cutler. *The Portrait of a Common HTML Web Page*, in DocEng 2006: "Found that the average web page contained 474 words, 281 HTML tags, and 41 links, 10 of which pointed outside the domain"