

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 Lawrence 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 grafo, donde cada sitio es un nodo. Existe una arista dirigida entre dos nodos si desde el primer sitio hay un hipervínculo al segundo. A cada uno de los nodos se le asigna un valor representando la probabilidad de que una cierta persona navegando por Internet esté observando dicho sitio, denominado el *PageRank* del sitio. En cada

época o iteración, dicha persona puede decidir quedarse en ese sitio definitivamente o desplazarse a otro sitio por medio de un hipervínculo (con probabilidad d , llamado factor de amortiguamiento). Nótese que el factor de amortiguamiento no puede ser representado como un lazo, ya que una vez que se decide quedarse en un sitio, el proceso finaliza.

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; si un sitio no tiene hipervínculos, se asume que un usuario podría ingresar el nombre de un nuevo sitio en el navegador, por lo que se generan aristas desde dicho estado absorbente a todos los demás estados del grafo. Debido a que un usuario podría hacer esto incluso en estados que no son absorbentes, se suele agregar dichas aristas en todos los nodos, con una probabilidad residual igual al factor de amortiguamiento.

El vector de probabilidades al que converge dicha cadena de Markov representa el peso o importancia de cada sitio.

III. MARCO ALGEBRAICO

Para empezar, se debe calcular el valor del *PageRank* de un nodo p_k . Dicho valor se calcula con la suma de la probabilidad de que el usuario decida quedarse indefinidamente en el nodo, y aquella asignada a que el usuario haya llegado al sitio a través de un hipervínculo. La probabilidad de haber navegado por un hipervínculo, es decir el factor de amortiguamiento d , vale aproximadamente 0,85 (establecido empíricamente [1]).

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

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 directamente alcanzables desde p_i .

Aplicando la ecuación (??) a todos los nodos y exigiendo que el *PageRank* de cada nodo converja a un valor, 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 es la matrix de adyacencia del grafo:

$$M = \begin{pmatrix} l(p_1, p_1) & l(p_1, p_2) & \cdots & l(p_1, p_N) \\ l(p_2, p_1) & \ddots & & \vdots \\ \vdots & & l(p_i, p_j) & \\ l(p_N, p_1) & \cdots & & l(p_N, p_N) \end{pmatrix} \quad (3)$$

Con $l(p_i, p_j)$ siendo el peso de la arista desde nodo p_i hacia el nodo p_j , definido como 0 si dicha arista no existe.

IV. SOLUCIONES PARA EL PROBLEMA DE CALCULAR EL PAGERANK

Se define la matriz \widehat{M} como:

$$\widehat{M} = \frac{1-d}{N}E + dM \quad (4)$$

Donde E es una matriz con todos unos.

Fácilmente se puede ver que la ecuación (??) es equivalente a la siguiente:

$$R = \widehat{M}R \quad (5)$$

$$R = \frac{1-d}{N}ER + dMR$$

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

($R \cdot \mathbf{1} = 1$, pues R es una distribución de probabilidades)

La cadena de Markov con matriz de transición \widehat{M} es irreducible y ergódica (todas las transiciones posibles tienen probabilidad mayor que 0), por lo tanto R existe y es único.

A. Método de las potencias

De la ecuación (??) se puede deducir que la solución R no es más que el autovector dominante con autovalor 1 de la matriz \widehat{M} . 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, por lo que el autovector es dominante. El autovalor será único si existe una única distribución que satisface la ecuación (??), lo cual es equivalente a exigir que $Rg(\widehat{M} - I) = N - 1$ y no menor. Esto trae a

la mesa el interesante trade-off de tiempo versus factor de amortiguamiento, que es un dato empírico. Sería interesante analizar 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 $v(0)$ inicializado arbitrariamente e iterar según la regla:

$$v(t+1) = \widehat{M}v(t) \quad (6)$$

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

$$|v(t+1) - v(t)| < \epsilon \quad (7)$$

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 versus factor de amortiguamiento, que es un dato empírico. Habría que analizar cuánto cambia la solución al variar d .

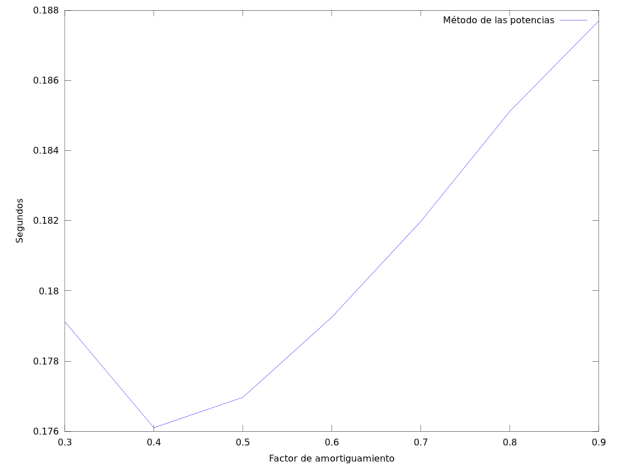


Figura 1. Gráfico del tiempo de ejecución del método de las potencias en función de el factor de amortiguamiento, para un grafo de 100 sitios.

B. Sistema lineal

De (??) se deduce:

$$R(I - dM) = \begin{pmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{pmatrix} \quad (8)$$

Sistema lineal de la forma $Ax = b$ con $A = (I - dM)$, $x = R$, $b = (1 - d)/N * \mathbf{1}$, 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 \widehat{M} es mucho más densa que la matriz M , que es, en general, algo deseable.

V. COMPARACIÓN Y CONCLUSIONES

Los grafos sobre los que se probaron los algoritmos fueron generados proceduralmente, con factores aleatorios, teniendo cuidado en conservar las características de un grafo web real (a saber, la cantidad total de links es $\Theta(\text{cantidad de sitios})$).

Se implementaron 3 algoritmos: el método de las potencias, la resolución del sistema lineal con el método de Jacobi, y la resolución del sistema lineal con el método que usa nativamente octave.

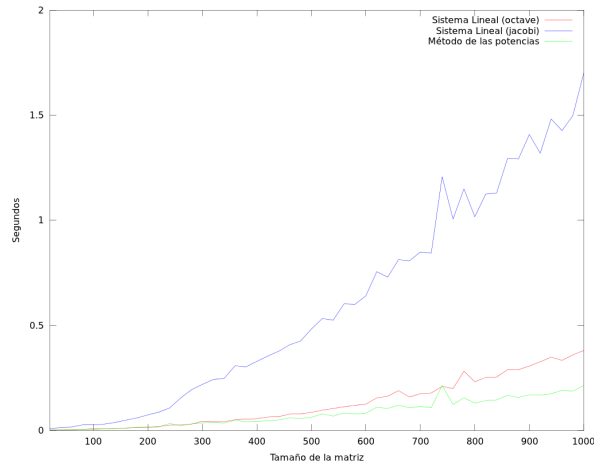


Figura 2. Gráfico del tiempo de ejecución de los tres algoritmos para grafos web de 20 a 1000 sitios.

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 implementación de los algoritmos básicos), por lo que no hay un ganador claro. Siendo así muy prometedor el nuevo enfoque 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*.