PageRank

Darío Sneidermanis, Martín Sturla

19 de Abril de 2012

***Resumen*—Este documento busca analizar distintas implementaciones de PageRank y comparar su desempeño. (bla, terminemoslo despues).**

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

I. Introducción

Con el inicio de las telecomunicaciones masivas por Internet a principio de los 90s y su rápido crecimiento, nace la necesidad de crear un índice de los sitios web. Estos índices son conocidos como 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 este último 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 suficientemente escalables 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, publican un documento llamado *"The Anatomy of a Large-Scale Hypertextual Web Search Engine"* que explaya el problema mencionado anteriormente y ofrece una solución; un algoritmo computable para indexar sitios web según su importancia, conocido como *PageRank*. Estos estudiantes crearon un motor de búsqueda llamado *Google* utilizando este algoritmo.
El motivo de este documento, que busca analizar distintas implementaciones de *PageRank* y comparar su desempeño, nace del interés de analizar teoremas y conceptos antiguos de la matemática, como lo son los autovalores y autovectores, aplicados en un área moderna y con aún mucho por investigar como es la informática. (something else?)

II. PageRank

A. Modelado del problema

El conjunto de los sitios web puede ser representado como un grafo, donde cada sitio es un nodo. Existe un arco guiado 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, donimado el *PageRank* del sitio. En cada época o iteración, dicha persona puede decidir quedarse y jamás volver a usar un hipervínculo o desplazarse a otro sitio por medio de un hipervínculo. La probabilidad de que no suceda el primer suceso se denomina factor de *damping*. Nótese que como el primer suceso implica que la persona no navegará nuevamente, el fenómeno de *damping* no puede ser representado como un arco al mismo nodo.
En vista de dichas propiedades, el grafo puede ser visto como una cadena de Markov, con ciertas propiedades. Se asume que la probabilidad iniciales de estar en cualquier nodo es uniforme. Asimismo, se asume que la probabilidad de abrir cualquier hipervínculo de un mismo nodo también es uniforme. 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 arcos desde dicho estado absorbente a todos los estados del grafo. Debido a que un usuario podría hacer esto incluso en estados que no son abosrbentes, se suele agregar dichos arcos en todos los nodos, con una probabilidad residual igual al factor de *damping*.
A medida que se itera dicha cadena de Markov y se aproxima al infinito, los valores asociados a cada nodo convergen a un cierto valor. Dicho valor representa el peso o importancia de la página. (Necesita gráficos)

B. Marco algebraico

Para empezar, se debe calcular el valor *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 damping, es representado con una d y empíricamente se ha establecido un valor de $0,85$ como adecuado. El grado de un nodo p_i se denomina $L(p_i)$ y es la cantidad de hipervínculos que posee. El valor N representa la cantidad total de nodos. Se denomina M_{p_i} al conjunto de nodos directamente alcanzables desde p_i. Es decir: (deberiamos derivar mas esto?)

19 de Abril, 2012.

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

Aplicando la ecuación (1) 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 M es la matrix de adyacencia del grafo:

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

Donde $l(p_i,p_j)$ es el peso del arco del nodo p_i al nodo p_j, definido como 0 si dicho arco no existe.
La solución R, vector representando los valores a los cuales han convergido los nodos, es el *PageRank* final de cada nodo.

C. Método de las potencias en PageRank

El problema de *PageRank* puede ser resuelto utilizando el método de las potencias. Considérese $\widehat{M}$ definida como:

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

Donde E es una matriz con todos sus valores iguales a 1. Es fácil (es? quizás habría que deducirlo un poco) ver que combinando (2) y (4) se obtiene:

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

De la ecuación (5) se puede deducir que la solución R no es más que el autovector dominante con autovalor 1 de la matriz de adyacencia modificada $\widehat{M}$. Este vector existirá siempre y cuando $det\left|\widehat{M} - I\right| \neq 0$ (pasa siempre?dem?). 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 asumiendo que existe una única distribución que satisface la ecuación , lo cual es equivalente a exigir que $\widehat{M}$ sea inversible. (estaria bueno tirar por aca tipo asumiendo que la cadena de markov es logica bajo los conceptos de page rank bla bla estas cosas raras no pueden pasar etc etc). Por lo tanto, se puede calcular 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 en este caso (dado que el autovalor dominante es 1 y $\widehat{M}$ es una matriz probabilística) consiste en tomar un vector $v(0)$ inicializado arbitrariamente e iterar según la regla:

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

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.