DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Trabájo Práctico 3

PageRank

12 de noviembre de 2013 — Métodos Numéricos

Integrante	LU	Correo electrónico
Escalante, José	822/06	joe.escalante@gmail.com
Osinski, Andrés	405/07	andres.osinski@gmail.com
Raskovsky, Iván Alejandro	57/07	iraskovsky@dc.uba.ar

Instancia	Docente	Nota
Primera entrega		
Segunda entrega		

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Índice

1. Abstract

En este trabajo nos concentraremos en analizar la teoría detrás de cómo hace Google para indexar los sitios web.

Dado a conocer a través de un paper en 1998, el algoritmo Page Rank, se convirtió en una de las claves del suceso del motor de búsqueda Google. Su implementación se basa en la creación de un ranking en el cual se pondera con un cierto criterio cada una de las páginas.

En el armado del ranking recaen conceptos de Algebra Lineal, los cuales veremos en detalle a lo largo de este informe. En particular veremos un método iterativo para calcular el ranking y una variación (extraída de un paper), la cual empíricamente optimiza la cantidad de iteraciones.

Al final presentaremos resultados de experimentaciones que nos parecieron pertinentes, mostrando que efectivamente la variación del método iterativo sí optimiza la cantidad de iteraciones.

Palabras clave:

- Page Rank
- Metodo de Potencia
- Extrapolación Cuadrática
- QR

2. Introducción Teórica

Veamos primero cómo utilizando el algoritmo de Kamvar podemos optimizar el espacio requerido en memoria para el almacenamiento de la matriz P_2 y el tiempo de ejecución requerido para hacer la multiplicación entre matrices y vectores.

Queremos ver que el algoritmo propuesto por [**?**, Algoritmo 1] es equivalente a la operación $\vec{y} = A\vec{x}$, para $A = (cP' + (1-c)E)^t$, donde P' es la matriz estocástica por filas de transiciones de links ajustada para considerar saltos aleatorios en páginas sin outlinks, y E es la matriz uniforme de teletransportación con valor $\frac{1}{n}$ en cada celda.

Para ello, expandimos las ecuaciones de ambos y veremos que las mismas producen el mismo cálculo.

Primero, la matrix P^t se desarrolla como

$$(cP + (1-c)E)^t\vec{x}$$

Y la matrix de [**?**, Algoritmo 1] como

$$cP^t\vec{x} + (\|\vec{x}\|_1 - \|\vec{y}\|_2)\vec{v}$$

donde $\vec{y}$ es el vector resultante de $cP^t\vec{x}$ y $\vec{v}$ es el vector de probabilidad uniforme de valor $\frac{1}{n}$ en cada elemento. Luego planteamos la equivalencia

$$\begin{aligned}(cP + (1-c)E)^t\vec{x}v &= cP^t\vec{x} + (\|\vec{x}\|_1 - \|\vec{y}\|_2)\vec{v}\\ cP^t\vec{x} + (1-c)E^t\vec{x} &= cP^t\vec{x} + (\|\vec{x}\|_1 - \|\vec{y}\|_2)\vec{v}\\ (1-c)E^t\vec{x} &= (\|\vec{x}\|_1 - \|\vec{y}\|_2)\vec{v}\end{aligned}$$

2.1. Lema: $P^t\vec{x}$ preserva la Norma 1 de $\vec{x}$

Sea P^t una matriz estocástica por columnas, luego los elementos de cada columna suman 1.

Luego P^t describe una transformación lineal de $\vec{x}$ donde la suma de los valores de cada x_i se reparte en los y_i resultantes (por ser cada y_i una combinación lineal de los x_i.) Como cada columna de P suma 1, y cada elemento de x se termina multiplicando por los elementos de una columna, y además los valores de P y x son positivos, entonces la ecuación

$$\sum_{i=1}^{n} |x_i|$$

es equivalente a

$$\sum_{i=1}^{n} |y_i|$$

Luego P^t preserva norma 1.

Volviendo al problema, si observamos que la norma 1 de y es

$$\begin{aligned}\|\vec{y}\|_1 &= \left\|cP^t\vec{x}\right\|_1\\ &= c\,\|\vec{x}\|_1\end{aligned}$$

entonces podemos ver que

$$\begin{aligned}\|\vec{x}\|_1 - \|\vec{y}\|_1 &= \|\vec{x}\|_1 - c\,\|\vec{x}\|_1 \\ &= (1-c)\,\|\vec{x}\|_1 \\ &= (1-c)\,\|\vec{x}\|_1\end{aligned}$$

por ende

$$(\|\vec{x}\|_1 - \|\vec{y}\|_1)\vec{v} = (1-c)\,\|\vec{x}\|_1\,\vec{v}$$

entonces el método de algortimo 1 tiene la forma

$$cP^t\vec{x} + (1-c)\,\|\vec{x}\|_1\,\vec{v}$$

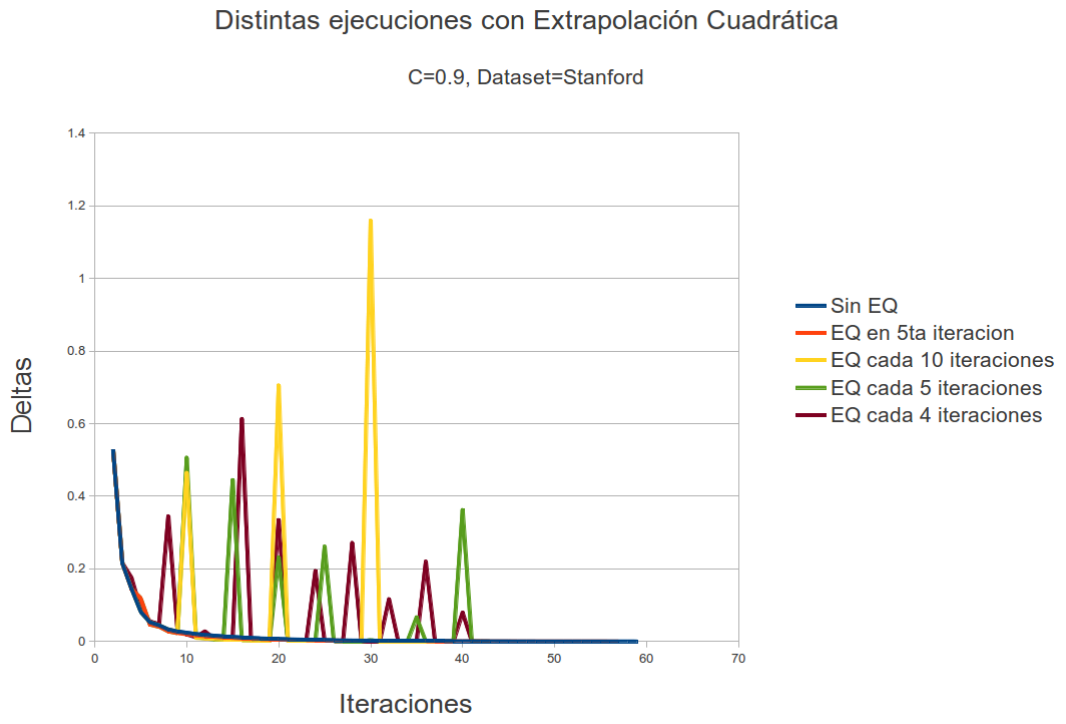
Si observamos la segunda mitad de la definición de P^t, es decir, $(1-c)E^t$, veremos que el producto a la izquierda por $\vec{x}$ resulta en una matrix con la forma

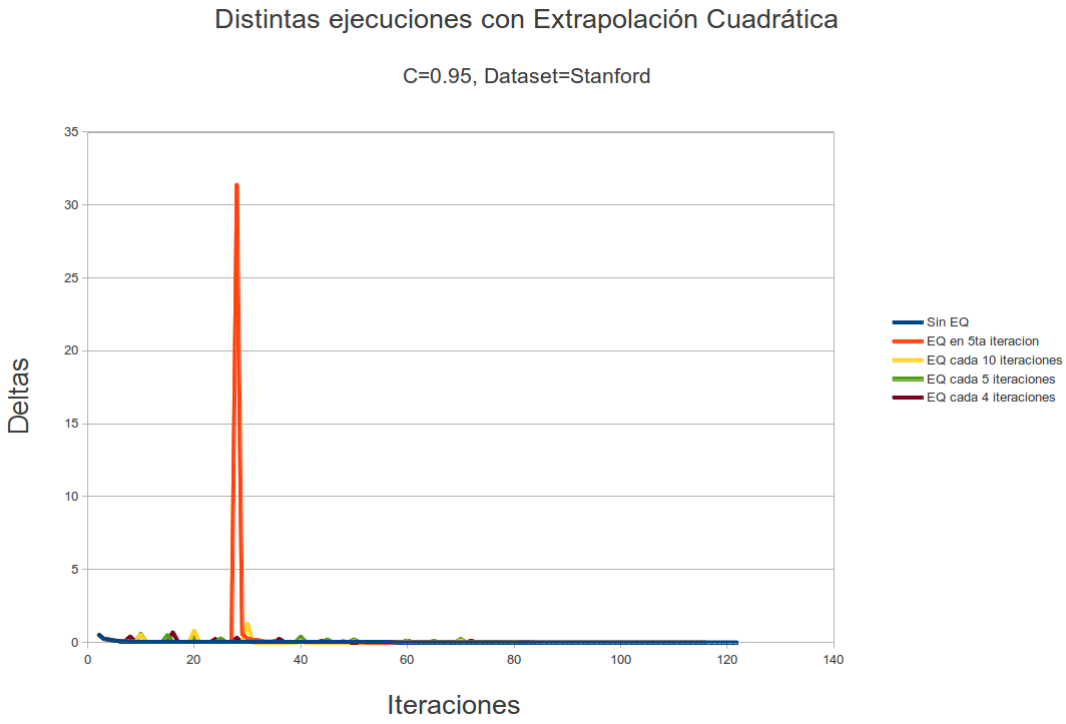
$$E^t\vec{x} = \begin{bmatrix} \frac{1-c}{n}\sum_{i=1}^{n}|x_i| \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \frac{1-c}{n}\sum_{i=1}^{n}|x_i| \end{bmatrix} = \begin{bmatrix} \frac{1-c}{n}\|\vec{x}\|_1 \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \frac{1-c}{n}\|\vec{x}\|_1 \end{bmatrix} = (1-c)\frac{1}{n}\|\vec{x}\|_1$$

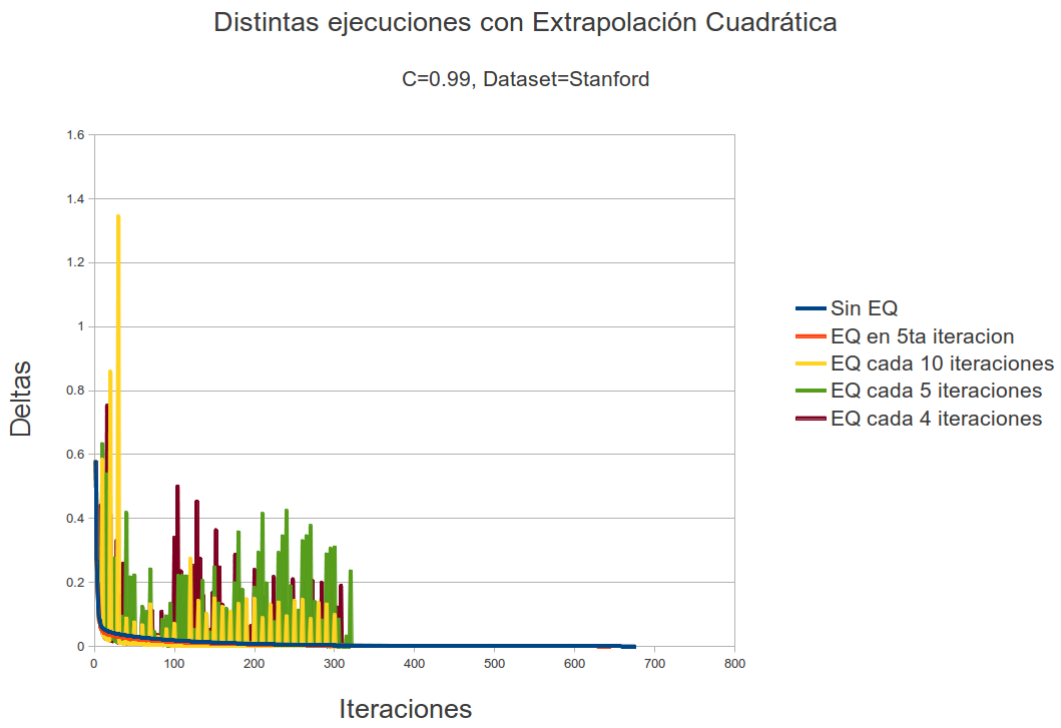
Con ello concluimos que los dos términos del algoritmo de Kamvar son equivalentes a la matriz A de transiciones. ■

3. Desarrollo

4. Resultados






Distintas ejecuciones con Extrapolación Cuadrática
C=0.99, Dataset=Stanford
Deltas
Iteraciones
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
600
700
800
Sin EQ
EQ en 5ta iteracion
EQ cada 10 iteraciones
EQ cada 5 iteraciones
EQ cada 4 iteraciones

5. Discusión y Conclusiones

A. Referencias

Wikipedia Burden