DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Ranking web

11 de octubre de 2014

Métodos Numéricos
Trabajo Práctico Nro. 2

Integrante	LU	Correo electrónico
Martin Carreiro	45/10	martin301290@gmail.com
Kevin Kujawski	459/10	kevinkuja@gmail.com
Juan Manuel Ortíz de Zárate	403/10	jmanuoz@gmail.com

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. Resumen

2. Introducción teórica

2.1. Matriz Dispersa

Se define una matrix dispersa aquella a la que la mayoría de sus elementos son cero.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & a_{04} \\ 0 & a_{11} & a_{12} & 0 & 0 \\ 0 & 0 & 0 & a_{23} & 0 \\ 0 & 0 & 0 & a_{33} & 0 \\ a_{40} & 0 & 0 & 0 & 0 \end{bmatrix}$$

2.2. DOK vs CRS vs CSC

La matriz dispersa al tener la propiedad de tener muy pocos valores no−cero es conveniente solo guardar estos y asumir el resto como cero. Existen varias estructuras como Dictionary of Keys (dok), Compressed Sparse Row (CSR) o Compressed Sparse Column (CSC). En el desarrollo de este TP, utilizamos DOK por facilidad en el uso del mismo. Tanto CSR o CSC se basan en la estructura Yale y se diferencian en como guardan los mismos valores, uno priorizando las columnas y otro las filas respectivamente.
La estructura Yale consiste en a partir de la matriz original obtener tres vectores que contengan

- A = los elementos no−cero de arriba-abajo,izquierda-derecha
- IA = los indices para cada fila i del primer elemento no-cero de dicha fila
- JA = los indices de columna para cada valor de A

Si bien en caso de que haya en una fila muchos números no-ceros es más beneficioso la utilización de esta estructura, la facilidad con DOK permite hacer pruebas más rápido.

3. Desarrollo

3.1. Page Rank

El algoritmo de PageRank lo dividimos en dos etapas, primero la inicialización en donde se crea la matriz estocástica y luego la corrida en donde se itera y calcula el pagerank hasta que la diferencia de norma entre los vectores sea menor que la tolerancia establecida.

Inicialización:

```
Genero un vector inicial.
Para cada nodo:
        - Si tiene salidas, inserto en cada celda de la columna correspondiente
           de los nodos de salida
        - Si no, guardo el nodo en el vector de desconectados.
```

Calculo del PageRank:

```
Hasta que converja:
        Multiplico la matriz por el vector actual.
        Aplico el algoritmo para tener en cuentas los nodos desconectados.
        Aplico el algoritmo para tener en cuenta el navegante aleatorio.
        Guardo el vector actual.
```

3.2. HITS

3.3. Indeg

Indeg utiliza plenamente la información de los links de las páginas que los apuntan y arma un promedio en la cantidad de estos sobre el total de links existentes en la red

```
Inicializo el vector de resultados con ceros.
Para cada conjunto de referencias de una pagina:
        Para cada referencia:
                Al vector resultados le sumo 1 / cantidad total de los links
```

4. Experimentación Y Resultados

4.1. Casos de prueba

A continuación se listarán los casos utilizados y después se compararán los resultados.

- MOVIES: Este caso incluye 5797 páginas
- ABORTION: Este caso incluye 2293 páginas
- GENETIC: Este caso incluye 3468 páginas
- STANFORD: Este caso incluye 281903 páginas
- GOOGLE: Este caso incluye 916428 páginas

4.2. Comparación de Normas

En esta sección vamos a mostrar como evoluciona la norma Manhattan (también conocida como distancia L1) entre dos vectores a medida que se suceden las iteraciones. La norma Manhattan es la distancia entre dos vectores, o en otras palabras, la suma de la diferencia coordenada a coordenada en modulo:

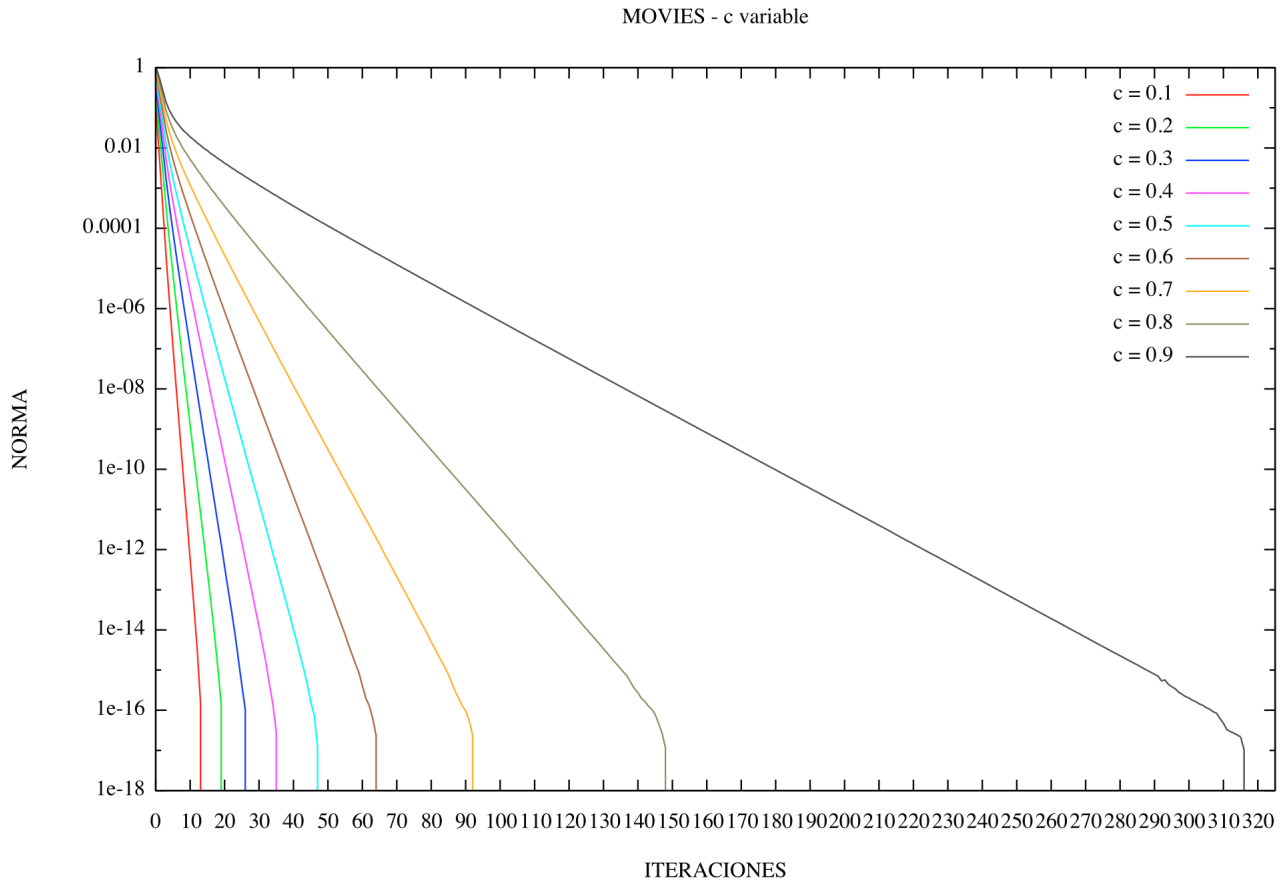
$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

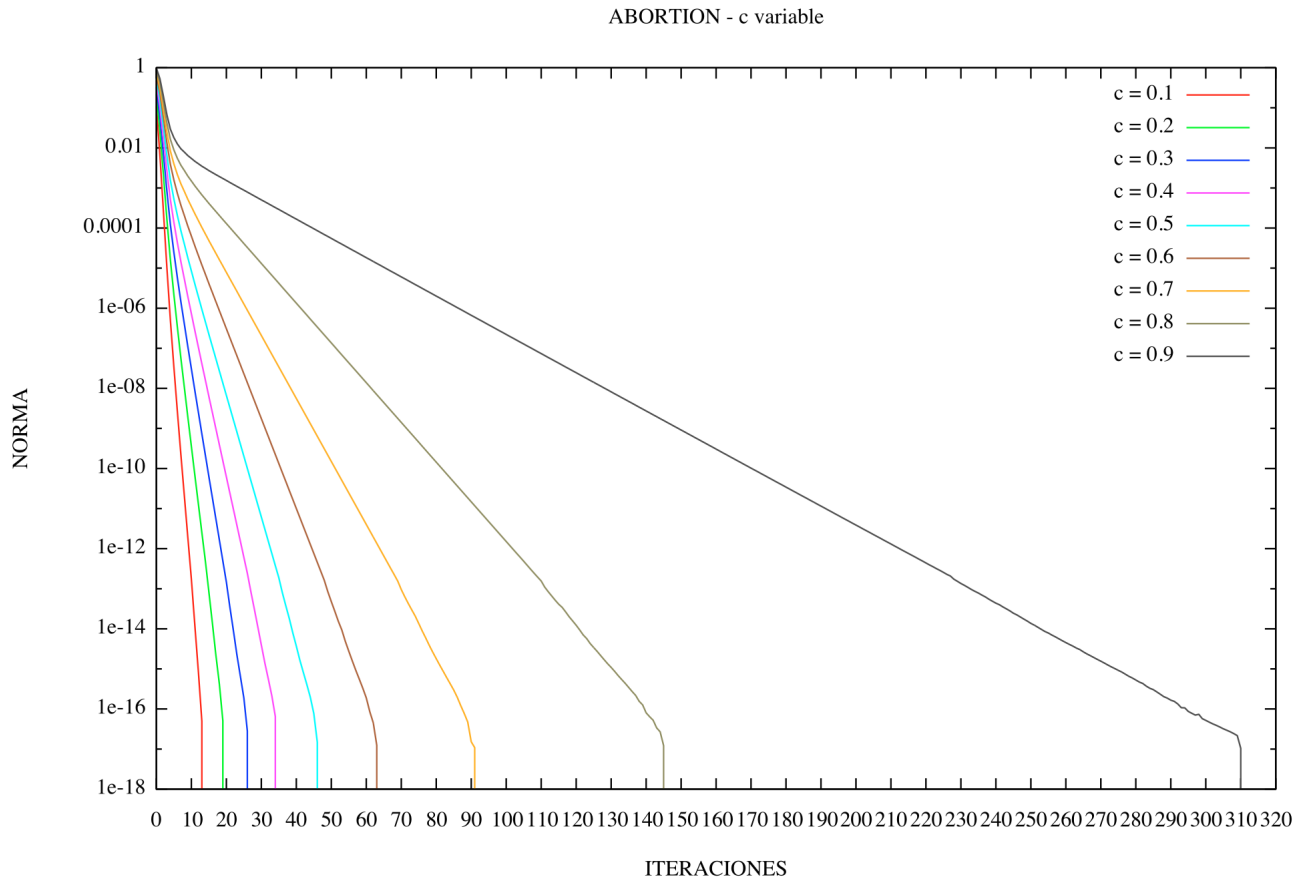
4.2.1. PageRank

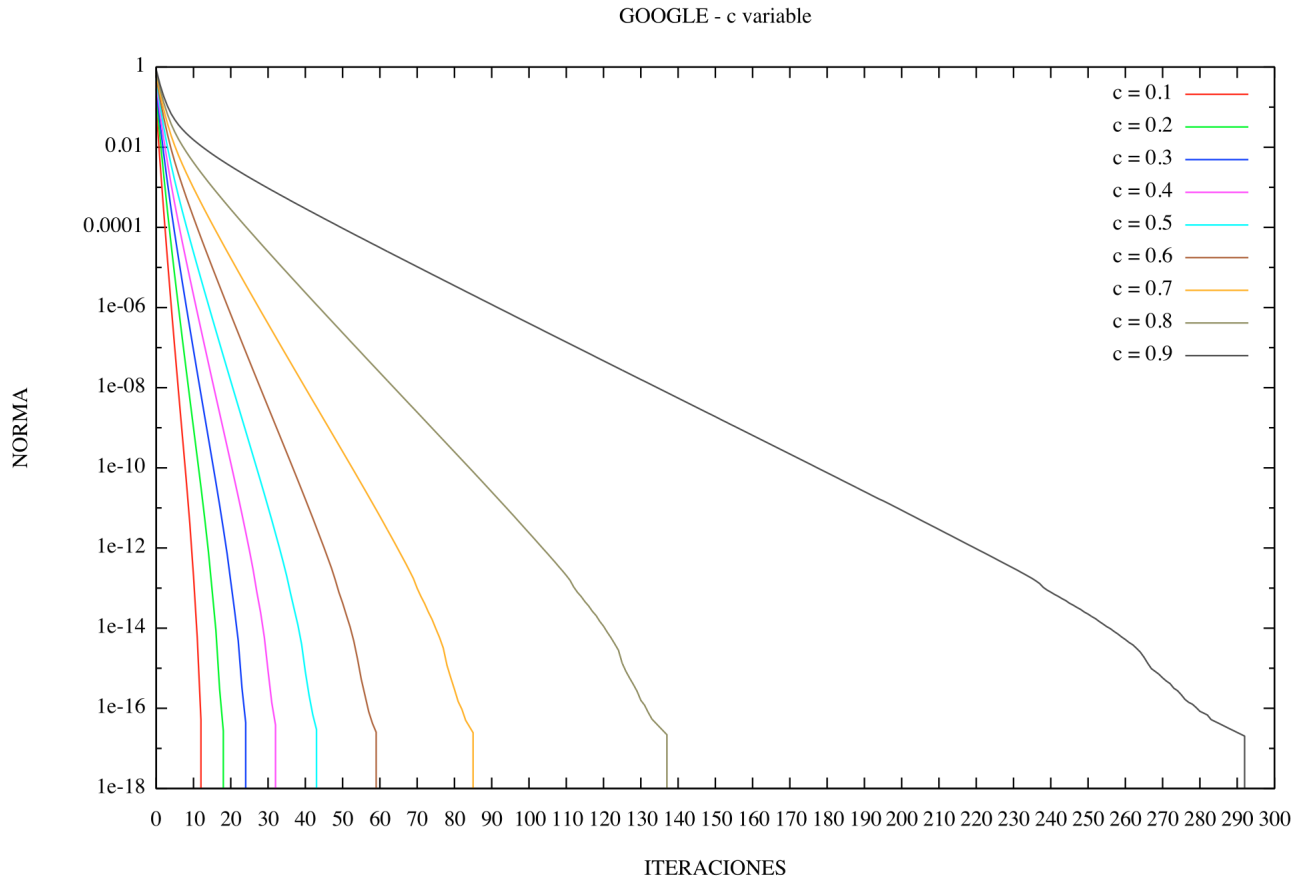
Para evaluar el comportamiento de la norma manhattan variando la probabilidad del navegante aleatorio, el cual de ahora en más lo denotaremos como el parámetro **c**

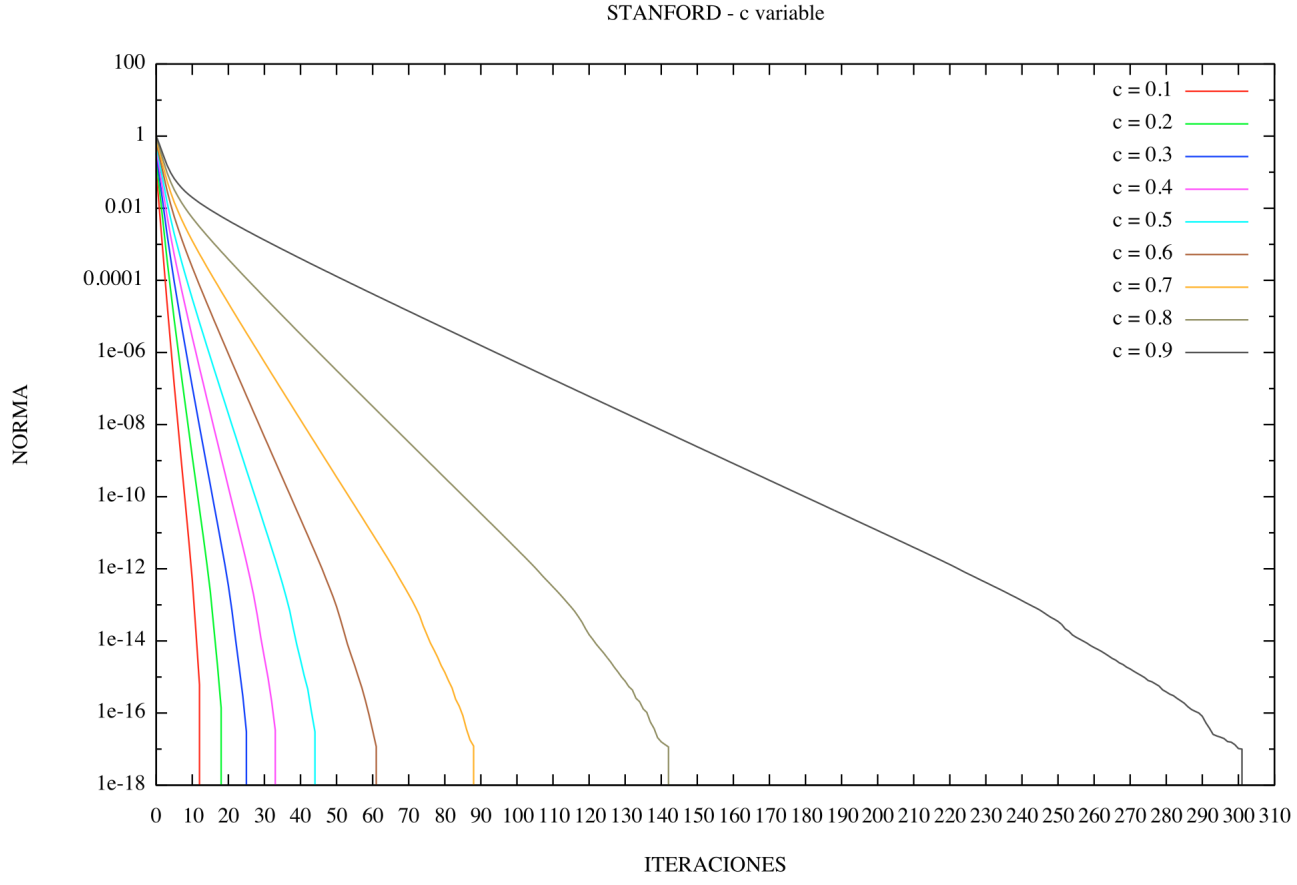
Los casos de prueba se corrieron sin un limite entre normas pero si con un limite de 1000 iteraciones, ya que, según lo que investigamos, con un c $\approx$ 0.15 la matriz suele converger en un máximo de 50 iteraciones, que por lo que se puede observar suele ser proporcional esta relación a medida que aumenta el c.

A continuación se muestran los resultados para cuatro tests de como evoluciona la norma a lo largo de las iteraciones y como varía la misma con distintos c, y que luego discutiremos más adelante:


MOVIES - c variable
NORMA
ITERACIONES
c = 0.1
c = 0.2
c = 0.3
c = 0.4
c = 0.5
c = 0.6
c = 0.7
c = 0.8
c = 0.9
1
0.01
0.0001
1e-06
1e-08
1e-10
1e-12
1e-14
1e-16
1e-18
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320


ABORTION - c variable
NORMA
ITERACIONES
c = 0.1
c = 0.2
c = 0.3
c = 0.4
c = 0.5
c = 0.6
c = 0.7
c = 0.8
c = 0.9
1
0.01
0.0001
1e-06
1e-08
1e-10
1e-12
1e-14
1e-16
1e-18
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320


GOOGLE - c variable
NORMA
ITERACIONES
1
0.01
0.0001
1e-06
1e-08
1e-10
1e-12
1e-14
1e-16
1e-18
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300
c = 0.1
c = 0.2
c = 0.3
c = 0.4
c = 0.5
c = 0.6
c = 0.7
c = 0.8
c = 0.9


STANFORD - c variable
NORMA
ITERACIONES
100
1
0.01
0.0001
1e-06
1e-08
1e-10
1e-12
1e-14
1e-16
1e-18
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310
c = 0.1
c = 0.2
c = 0.3
c = 0.4
c = 0.5
c = 0.6
c = 0.7
c = 0.8
c = 0.9

4.2.2. HITS

4.3. Comparación de Tiempos

5. Discusión

5.1. PageRank

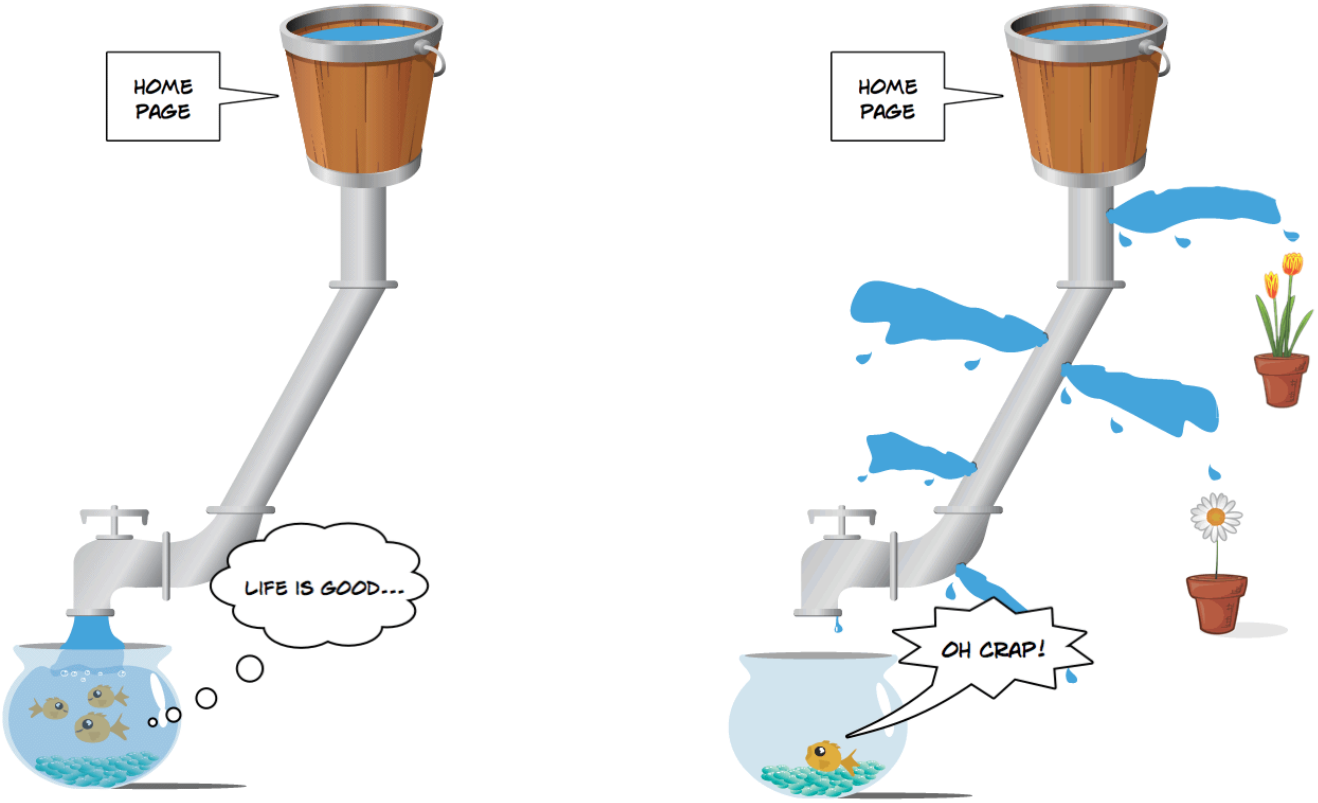
Claramente podemos notar que a medida que el C crece, el algoritmo toma más iteraciones en achicar la norma. Esto se debe a que el grado de aleatoriedad elimina el peso de la unión entre los sitios e indica una uniformidad en el comportamiento, entonces la matriz si bien estocástica ahora se encuentra distribuida esa suma = 1 por columna en varias filas. Esto produce mayor cantidad de iteraciones en el método de la potencia ya que la mayor uniformidad de la matriz provoca que ninguna 'zona' de la matriz absorba más que las demás. [1] También es bastante notorio que a pesar de que los distintos casos de prueba sean muy diferentes entre si y hasta cientos de veces más grandes, la evolución de la norma converge de formas casi idénticas y lo mismo sucede para las iteraciones requeridas hasta llegar a la norma variando el parámetro c.

La norma Manhattan de la diferencia del vector resultado entre una iteración y la siguiente disminuye tan rápidamente que para visualizarla fue mejor utilizar una escala logarítmica. En las primeras 5 iteraciones la norma disminuye de forma más dramática y a partir de la iteración 10 se puede ver una velocidad de disminución predecible. La diferencia entre una iteración y la otra tiende a disminuirse de forma exponencial.

6. Conclusiones

6.1. PageRank

A medida que fuimos investigando y probando el algoritmo del PageRank nos fue quedando cada vez más claro como es que funciona y que se necesita para obtener un buen resultado para un sitio en particular. Lo que nos pareció interesante es explicar el algoritmo con la siguiente interpretación metafórica:



Tomando como al sitio a analizar en cuestión como la pecera, y a otro sitio que tiene un link a nuestro sitio como el balde se puede observar que cuando el *recurso*, en este caso el agua, se reparte equitativamente a todos los destinatarios, por lo tanto, si mi pecera es la única que recibe agua voy a obtener más que si tiene otras bocas la canilla con la cual compartir. Esto es lo mismo que sucede en la web y tiene el cuenta el PageRank, cada sitio le distribuye equitativamente una probabilidad a cada salida, cuya suma total es 1. Por lo tanto, me conviene más que me linkee un sitio con pocas salidas que uno con gran cantidad, pero suponiendo que sus respectivos PageRank son similares, ya que mi PageRank también va a depende del de mis entradas, por lo tanto también hay que tener esto en cuenta, ya que es un factor bastante influyente. Por consiguiente, no solo depende la cantidad de sitios que apuntan a si no también el PageRank de cada uno (la *calidad*)

Referencias

[1] $http://personales.upv.es/\ pedroche/inv/_docs/fpedrochev4(sema).pdf$